Model-Based Genomic Studies of Protein Sequence Evolution: Convergence, Epistasis, and Amino Acid Acceptance Rates

by

Zhengting Zou

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in The University of Michigan
2017

Doctoral Committee:

Professor Jianzhi George Zhang, Chair
Assistant Professor Alan P. Boyle
Professor Daniel M. Burns Jr.
Assistant Professor Stephen A. Smith
Professor Patricia Wittkopp

Zhengting Zou

ztzou@umich.edu

ORCID iD: 0000-0003-1716-5090

DEDICATION

I dedicate this thesis to my parents.
For their everlasting love, understanding and support.

ACKNOWLEDGEMENTS

I would like to give my sincere gratitude to Dr. Jianzhi Zhang for his guidance on my doctoral researches and academic development. He has been a great mentor and scientist ever since my first interview, by providing provident and practical suggestions in all aspects. His wisdom, rigor and enthusiasm in scientific researches, as well as his optimistic and considerate personality, have profoundly supported my graduate study and inspired me to be a dedicated researcher. This is a most important and unforgettable component of my graduate life. I would also like to thank all past and current members of the Zhang Lab for creating a great lab environment and being professional colleagues as well as good friends, specifically Jian-Rong Yang, Nagarjun Vijay and Wei-Chin Ho for their discussions, suggestions and guidance on problems I encountered in daily researches; Xinzhu Wei for being a wise, ever-thinking and inspiring cohort; Chuan Li, Chuan Xu and Zhen Liu for all the inspiring discussions in daily lab life. I am very grateful to the essential help from my dissertation committee: Drs. Trisha Wittkopp, Alan Boyle, Dan Burns and Steve Smith.

I want to thank Drs. Margit Burmeister and Dan Burns as directors of the Bioinformatics Ph.D. program for supervising our graduate study and actively supporting students with funding opportunities. My gratitude should also go to all my friends, for being interesting and helpful and for making these years a great life experience.

Finally, I would like to thank my parents. Although my parents may not understand the content or significance of my research, they always believe in my choices in life and keep being supportive in every sense. Without my family, I can by no means achieve what I have today.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

ABSTRACT

Protein sequence changes are a major contributor to phenotypic evolution and biodiversity. While the genomic revolution has drastically increased the available amount of protein sequence data for comparative studies, development of analytic tools lags behind. In particular, current mathematical models of sequence evolution are over-simplified and typically ignore many heterogeneities in evolutionary processes. As a result, they often provide inadequate descriptions of evolution, leading to misleading conclusions. My thesis uncovers some of these heterogeneities and demonstrates that incorporating them into mathematical models of protein sequence evolution offers new insights into evolutionary mechanisms. For instance, convergent evolution of morphological traits has long interested biologists because it is a strong indicator of common natural selections in independent evolutionary lineages. Similarly, convergent evolution of protein sequences is commonly thought to have resulted from natural selection. In Chapter 2 of this thesis, however, I show that such interpretations are problematic, because sequence convergence can be explained by neutral evolution as long as among-site variations in amino acid composition are considered. I also find that the convergence level reduces with genetic distance. In Chapter 3, I evaluate two hypotheses that could explain the diminishing convergence with genetic distance: (i) divergent epistasis in distantly related organisms and (ii) gene tree discordance. I demonstrate that both hypotheses are at work, but their contributions vary depending on how closely related the species of interest are. In Chapter 4, I revisit a high-profile

claim of genome-wide adaptive protein sequence convergence for echolocation in three lineages of mammals. I discover that the amount of convergence observed is no more than those in proper negative controls, suggesting that these sequence convergences are largely neutral and unrelated to echolocation. A widely believed but never critically tested hypothesis in phylogenetics is that morphological data contain more convergence and hence are less suitable for phylogenetic inference than molecular data. Analyzing a large dataset including thousands of morphological traits and thousands of molecular traits, I find unequivocal evidence for this hypothesis and uncover its underlying cause in Chapter 5. I subsequently design a method to identify and remove highly convergent traits, leading to higher phylogenetic accuracies. In Chapter 6, I report a new type of evolutionary heterogeneity that potentially contributes to phylogenetic error: between-species variation in the probability with which a mutation between a specific pair of amino acids is fixed. In Chapter 7, I find that this heterogeneity leads to another previously unknown heterogeneity among species: the fitness disadvantage of nonsynonymous transversions relative to that of nonsynonymous transitions, a subject that has been studied since the dawn of the field molecular evolution. These six chapters, along with the introductory and concluding chapters, provide an integrative study of previously unknown or neglected heterogeneities in protein sequence evolution. Together, they correct misconceptions in molecular evolution, help improve phylogenetic inference, and deepen our understanding of evolutionary mechanisms.

CHAPTER 1

General Introduction

Protein sequence evolution is one of the core processes in the history of evolution. Nonsynonymous mutations in coding sequences cause amino acid changes in protein sequences, leading to protein function alteration and phenotypic changes, thus conferring fitness effects subject to selection. In this sense, certain amino acid substitutions fixed across species can result from positive selection during adaptation on one hand, making analyzing protein sequence evolution fundamental for studying organismal adaption processes. On the other hand, a proportion of the amino acid substitutions are likely to be fixed because they are neutral or nearly neutral (Kimura 1983; Ohta 1992), on which the basis of a major part of the molecular phylogenetics are build. My dissertation focuses on elucidating the important patterns in protein sequence evolution, and disentangling contributions of different factors or forces underlying these patterns with analyses based on major sequence evolution models.

In this general introduction, I will discuss the backgrounds of my studies in protein sequence evolution. First, I will summarize the previous knowledge and studies about convergence as an indication of adaptive evolution. I will then introduce the studies focusing on the neutral proportion of observed convergence, as well as possible underlying forces such as epistasis and gene tree discordance, both of which render current molecular evolution models insufficient. As a consequence of convergence, the long-held debate about whether sequence

data are more reliable than morphology in phylogenetics will be revisited. Next, I will revisit the model insufficiency issue by summarizing the major models for describing protein sequence evolution and model assumptions that is questionable given some recent studies. As a potential reflection of the model insufficiency, current understanding of transition/transversion bias will be discussed in the last section.

Convergence as evidence of adaptive evolution

Convergence is the phenomenon that identical or similar states of a feature evolve independently in multiple lineages of species (Losos 2011). Based on whether the ancestral states are different among these lineages, the phenomena of convergence can be further classified as parallelism or narrow-sense convergence, of which the former starts with similar ancestral states and the latter starts with different ancestral states (Zhang and Kumar 1997). Here in this introduction, the term "convergence" will include both categories, unless specifically indicated.

The convergence pattern was first described at phenotypic, or morphological level. It was already mentioned by Charles Darwin in 1859, as "analogical or adaptive resemblances", examples including the forelimbs of dugongs and whales, similarity between certain homopterous insects and moths, etc (Darwin 1859). Notably, Darwin believed that convergence is the result of adaptation to similar environmental conditions. Since most phenotypic changes are directly under selection, this view of phenotypic convergence as an indication of adaptation is well accepted. Many more cases have been reported, e.g. similar web architectures of spiders

occupying the same habitat types on different Hawaiian islands (Blackledge and Gillespie 2004), similar bill shape shifts of tidal marsh sparrows in North America (Grenier and Greenberg 2005), morphological similarity among trunk-ground dwelling anoles on multiple Greater Antillean islands (Langerhans, et al. 2006), intercontinental pairs of desert iguana species with matching habitats (Melville, et al. 2006). This list goes on with more examples virtually across the whole tree of life (Nevo 1979; Moore and Willmer 1997; Wittkopp, et al. 2003; Fong, et al. 2005).

With the availability of molecular sequence data, it is intriguing to trace the phenotypic adaptive convergence down to genotype level and elucidate the genetic basis of functional adaptation. In recent years, convergent amino acid substitutions discovered in lineages with convergent phenotypes are usually interpreted as results of positive selection during adaptation to similar niches (Zhang 2006; Christin, et al. 2008; Jost, et al. 2008; Castoe, et al. 2009; Shen, et al. 2010; Liu, et al. 2011; Davies, et al. 2012; Feldman, et al. 2012; Liu, et al. 2012; Shen, et al. 2012; Zhen, et al. 2012; Ujvari, et al. 2015). In these studies, certain candidate genes related to the adaptive scenario are shown to experience convergent sequence evolution. One significant case of adaptive protein sequence convergence is the correlation between the hearing gene *prestin* and echolocation in bats and toothed whales. The N7T amino acid substitution in the prestin protein turns out to be not only a typical sequence convergence unique to all echolocating mammals, but also sufficient to cause functional shift of a non-echolocating version prestin to mimic biophysical properties of the echolocating version (Li, et al. 2010; Liu, et al. 2014).

Similar experimental evidences for adaptive sequence convergence was also reported in haemoglobins of high-altitude Andean hummingbirds (Projecto-Garcia, et al. 2013).

Perhaps a more important portion of the whole picture, adaptive convergence definitely appears at more levels of the life processes than just individual morphology or amino acid sites. High-order life histories of different fauna can show convergence due to similar climatic conditions on different continents (Mazel, et al. 2017). Apart from individual protein sequence substitutions, functional adaptation of different lineages can be also achieved by convergence at many molecular levels (Arendt and Reznick 2008; Manceau, et al. 2010; Losos 2011). Examples include different residues within the same protein (Protas, et al. 2006; Rosenblum, et al. 2010; Linnen, et al. 2013; Zhou, et al. 2015; Chikina, et al. 2016), different proteins within the same pathways (Aminetzach, et al. 2009), expression levels of the same gene or genes in the same network modules (Shapiro, et al. 2006; Pfenning, et al. 2014; Berens, et al. 2015), etc. In this sense, strictly defined sequence convergence underlies probably only a small proportion of the functional convergence we observe in nature.

Nonetheless, the success of the aforementioned candidate gene approach has triggered genome-wide scale analyses, aiming to investigate the possible sequence convergence signal in genomes as a signature of adaptation (Bazykin, et al. 2007; Rokas and Carroll 2008; Parker, et al. 2013; Foote, et al. 2015; Xu, et al. 2017). However, at protein sequence level, whether most observed convergence events are due to adaptation is questionable. At phenotypic level, having independent origins of the same complex trait due to genetic drift alone is likely to be low

(Stayton 2008). But during neutral sequence evolution, substitutions at each site can be modeled as a continuous-time Markov chain process, so each type of state transition (e.g. from amino acid I to L) has certain probability to occur in any lineage. Hence a convergence event between two independent lineages may well happen by chance (Zhang and Kumar 1997; Stayton 2008) rather than due to selection. This makes the null hypothesis in adaptive convergence detection important: to claim the existence of adaptive convergence, one has to establish that the observed convergence events cannot be explained by neutral evolution. For example, the study by Foote, et al. (2015) found that the observed convergence levels between three lineages of marine mammals are no higher than those between comparable control lineages without apparent common adaptation. Thus no genome-wide sequence convergence for marine environment adaptation could be claimed. Zhang (2006) proposed four criteria for proving adaptive parallel sequence evolution, i.e. similar protein function changes in independent lineages, parallel substitutions observed in the protein, insufficiency of neutral evolution to explain the parallel changes, and causal relationship between the parallel substitution and the functional change as result. Till recently, there are only limited number of studies claiming adaptive sequence convergence that can fulfill all criteria. Hence, proper modeling of neutral protein sequence convergence is a basis for investigating adaptive convergence, which is not a trivial task given the discussion in the next section.

Convergence by constraint in neutral sequence evolution

In fact, other than adaptation, constraints during the evolutionary process of a trait has been proposed by many as an important driving force of convergence (Arendt and Reznick 2008; Elmer and Meyer 2011; Losos 2011; McGhee 2011; Wake, et al. 2011). Notably, constraints are within the scope of neutral evolution. When the state space of a trait is constrained to limited number of states, probability of convergence increases under both neutral and adaptive scenarios. For example, it is likely that a strong developmental constraint exists for vertebrates to have only four limbs, resulting in the convergent usage of forelimbs for flight in pterosaurs, birds and bats (Losos 2011). At sequence level, epistasis is a genetic constraint. Epistasis is defined as the interaction between a focal residue and surrounding environment. The consequence of epistasis is that the possible amino acid states allowed at the focal protein sequence position are constrained. As a major genetic constraint, epistasis has been one focus of studies aiming at describing protein sequence evolution, and is probably a most important factor in answering the question of whether evolution is predictable or not at sequence level. Numerous studies have shown the prevalence of sequence level epistasis (Breen, et al. 2012; Gong and Bloom 2014; Xu and Zhang 2014; Podgornaia and Laub 2015; Li, et al. 2016; Dungan and Chang 2017), indicating the contingency in the history of protein sequence evolution.

The implication of prevalent epistasis in sequence evolution is also important for model-based molecular evolution analyses. Each residue in each protein is interacting with a virtually unique set of environment, including spatially adjacent residues in the same protein or on an interacting surface of another protein, interacting nucleic acid molecules, ligands, membrane

components, etc. The difference in epistatic interactions and presumed functions of each residue suggests different constraints on which amino acids could be accepted. This site-specificity of amino acid composition, or site-specific genetic constraint, is seldom considered in major sequence evolution models. The model insufficiency here may cause underestimation of the expected level of chance convergence as previously mentioned. For example, Rokas and Carroll (2008) claimed the observation of frequent parallel changes across large sets of genes in many species clades, which is not fully predicted by neutral models, but it is pointed out that this pattern could be due to either positive selection or purifying constraints.

Convergence in morphological and molecular phylogenetics

As Darwin stated, convergence "are almost valueless to the systematist … such resemblances will not reveal—will rather tend to conceal their blood-relationship to their proper lines of descent"(Darwin 1859). In phylogenetics, evolutionary trajectories of multiple lineages are inferred by the divergence of traits, or characters, in which process closely related lineages are marked by identical-by-descent states. However, the actual identical-by-state used in inference may well be the result of convergence events, thus leading to erroneous clustering of lineages. In this sense, characters with less convergence carry less noise for inference of phylogeny. Phylogeneticists traditionally use morphological traits for tree inference, while nowadays molecular DNA or protein sequence data are more widely used. Additionally, approaches of combining morphology with molecular data in tree inference exist (Lee, et al.

2013; Bieler, et al. 2014). Till recently, the long-lasting debate about whether the morphology or molecular sequence is more suitable for phylogenetic inference still persists, and incongruence of the phylogenies inferred by the two types of data is considerable (Livezey and Zusi 2007; Legg, et al. 2013; O'Leary, et al. 2013; Springer, et al. 2013; Jarvis, et al. 2014; Pyron 2015). For example, a mammalian morphological tree in O'Leary, et al. (2013) groups armadillo, pangolin and aardvark together as an "ant and termite-eating" group, while molecular phylogeny has shown that they belong to different superorders (Meredith, et al. 2011). Many researchers seem to believe that morphological characters are more susceptible to convergence than sequence data, hence less desirable as input of phylogenetic inference (Givnish and Sytsma 1997; Page and Holmes 1998; Gaubert, et al. 2005; Wiens, et al. 2010; Wake, et al. 2011; Springer, et al. 2013; Davalos, et al. 2014; Jarvis, et al. 2014). However, the previously mentioned surging discoveries of sequence convergence in recent years, regardless of the underlying cause, cast doubt on whether this belief on molecular data still holds.

Homoplasy versus hemiplasy

The observed convergence patterns in sequence evolution do not necessarily result from true convergence events in history. While homoplasy, the true convergence and reversal events, can cause a site to show discordant state pattern with the species tree, one other possible cause is hemiplasy, the discordance between character states and true species tree caused by incomplete lineage sorting (ILS), introgression or horizontal gene transfer (HGT) (Hahn and Nakhleh 2016).

Hemiplasy events essentially cause gene tree to be discordant with species tree (GTD, gene tree discordance). Hence although the states of affected lineages are *bona fide* identical-by-descent, the pattern will be interpreted as "homoplasy" on the species tree if no GTD is assumed in the analyses. Being a violation to the traditional phylogenetic model of clearly sorted taxa and genotypes, hemiplasy is gaining more attention in recent studies. Evidence of hemiplasy events has been found in many prokaryotic and eukaryotic species clades (Mallet, et al. 2016), one significant example being the fruit flies (Ballard 2000; Bachtrog, et al. 2006; Pollard, et al. 2006). The awareness of hemiplasy has invoked application of coalescence-based species tree inference methods (Degnan and Rosenberg 2009; Jarvis, et al. 2014), and it has been suggested that some proposed functional convergences may actually be hemiplastic events (Hahn and Nakhleh 2016). Nonetheless, it may not be straightforward to distinguish between homoplasy and hemiplasy. For example, ILS is more likely to happen when an internal branch of a tree is short, but meanwhile a short branch also carries less divergence signal. In this case, given relatively long external branches, both ILS and true convergence may cause a gene to support a discordant gene tree by chance. Hence, it remains unclear whether homoplasy or hemiplasy contributes more to the discordant patterns in different phylogenetic datasets. Consequently, the significance of applying coalescent methods is yet to be decided (Scornavacca and Galtier 2017). To elucidate factors underlying observed sequence convergence and related patterns, GTD is apparently an important confounding factor to consider.

Incorporation of amino acid acceptance rate into mechanistic evolutionary models

Molecular evolution approaches are largely model-based. Specifically, sequence evolution is modeled by continuous time Markov process (Yang 2006), the core component being a transition matrix describing conditional probabilities of different types of substitutions. The aforementioned topics about sequence convergence are usually investigated under protein substitution models, which have been developed and refined ever since the accumulation of protein sequence data. The first protein substitution models were summarized statistically from existing orthologous protein sequence alignments, thus called empirical models, popular examples including the Dayhoff matrix (Dayhoff, et al. 1978), the JTT matrix (Jones, et al. 1992), the WAG matrix (Whelan and Goldman 2001) and the LG matrix (Le and Gascuel 2008), etc. Each of these matrices is developed as product of a symmetric amino acid exchangeability matrix S and a diagonal matrix of equilibrium frequencies (Yang 2006). Although based on different sets of protein data, the empirical models are highly correlated (Jones, et al. 1992; Whelan and Goldman 2001; Le and Gascuel 2008), suggesting that the determinants of amino acid exchangeability are common among different nuclear genes and different species. This further supported the feasibility to represent evolutionary processes in multiple phylogenetic lineages by a single matrix. With increasing amount of available data, mechanistic substitution models at the level of codon sequences were developed, incorporating parameters separately reflecting mutational biases and the effects of selection. In the commonly used codon substitution model implemented in the PAML package (Goldman and Yang 1994; Yang 1998,

2007), only codon pairs separated by one nucleotide substitution step have non-zero transition probability. Mutational effect of transition-transversion bias is represented by the substitution rate ratio κ. Effect of selection is reflected by a separate parameter ω, which corresponds to the d_N/d_S value, or average fixation probability of a nonsynonymous mutation.

Although more realistic than the empirical models, this widely-used codon model inevitably makes assumptions to reduce heterogeneity in the codon substitution process it tries to model. One assumption here is that different types of amino acid changes share a single selection parameter ω, while it is intuitive that different amino acid changes should be accepted with different probabilities during evolution. Previous studies have adopted different approaches to investigate the fixation probability, or acceptance rate, of changes between individual pairs of amino acids, e.g. by fitting functions of physiochemical properties (Grantham 1974; Miyata, et al. 1979), by experimental measurement (Yampolsky and Stoltzfus 2005), or by summarizing evolutionary sequence data (Tang, et al. 2004). A maximum likelihood model that relaxes the assumption was also applied to infer acceptance rates from sequence data (Yang, et al. 1998). Although the acceptance rates derived from these different approaches have been shown to correlate significantly (and also correlate with exchangeabilities in empirical models) (Yang, et al. 1998; Tang, et al. 2004; Stoltzfus and Norris 2016), there is evidence suggesting that for a particular set of sequences, the acceptance rates might be unique (Yang, et al. 1998). These studies indicated two layers of heterogeneity: acceptance rate heterogeneity among different types of amino acid changes, and acceptance rate heterogeneity among different clades of

species. There has been no study specifically focusing on the latter lineage-specific heterogeneity. If the heterogeneities prevalently exists, the current evolutionary models could be over-simplified and insufficient as null models when certain hypotheses are to be tested. An example of one of these hypotheses is introduced in the next section.

Causes of transition/transversion bias in coding sequence alignments

Transition/transversion bias has long been observed in sequence data and is also incorporated in codon substitution models. Compared with null distributions where all substitutions happen at the same rate, transitions, or nucleotide substitutions within the category of purines or pyrimidines, are usually more likely to be observed than transversions, or substitutions between a purine and a pyrimidine. In coding sequences, both the mutational process and the selection process can theoretically affect this bias. The mutational bias has been observed broadly. For example, mutation accumulation experiments found that with the existence of minimal selection, transitions accumulate at a higher rate than transversions in many species (Haag-Liautard, et al. 2008; Lynch, et al. 2008; Ossowski, et al. 2010; Schrider, et al. 2013; Zhu, et al. 2014). Natural population variation data and comparative sequence data have also been used to demonstrate this layer of transition/transversion bias (Freudenberg-Hua, et al. 2003; Rosenberg, et al. 2003; Cutter 2006; Jiang and Zhao 2006; Hershberg and Petrov 2010).

The second possible layer of bias may exist if transitions are less deleterious than transversions. Let us assume that synonymous mutations are not subject to purifying selection,

which is consistent with the major codon substitution models. First, if transitions tend to be synonymous, then they are more likely to be fixed than transversions, since nonsynonymous mutations are subject to prevalent purifying selection. This composition difference has been found in multiple species (Zhang 2000; Freudenberg-Hua, et al. 2003; Schrider, et al. 2013). Second, considering only nonsynonymous mutations, if nonsynonymous transitions are less deleterious than nonsynonymous transversions, there will also be fixation probability difference between two categories, and result in more transitions being retained in coding sequences. However, whether this second layer of bias exist or not is unclear, as different studies draw different conclusions (Zhang 2000; Freudenberg-Hua, et al. 2003; Stoltzfus and Norris 2016). To answer this question, the previously mentioned Goldman and Yang model with a single transition/transversion bias factor κ and a single selection factor ω is apparently insufficient. Ideally, a model harboring two separate selection factors for transitions and transversions should be used to investigate the nonsynonymous layer of transition/transversion bias. Furthermore, neither nonsynonymous transitions nor nonsynonymous transversions are directly under selection, because selection acts on amino acid changes resulted from transitions or transversions. In this sense, more sophisticated models categorizing selection effects according to different amino acid changes (Yang, et al. 1998) should be used when we want to disentangle the selection bias of nonsynonymous transitions or transversions to a more basic and mechanistic level.

Preview of research work

Given this background and current advances in our understanding of protein sequence evolution, my studies in this dissertation visit the following unresolved topics in six research projects: the prevalence, phylogenetic consequence and driving force of genome-wide protein sequence convergence; contribution of site epistasis to sequence convergence; and the effect of amino acid acceptance rates on patterns of coding sequence evolution.

In Chapter 2, I ask the question of whether there is prevalent adaptive convergence in genome-scale protein sequences. I compared observed level of sequence convergence in fruit flies and mammals with corresponding expectations calculated from different amino acid substitution models. Specifically, when calculating expected level of convergence, I incorporated the site-specific constraint on amino acid composition, so as to fully count the contribution of neutral sequence evolution to the observed convergence level.

In Chapter 3, I revisited the analyses in the previous chapter, being aware of the possible confounding effects of gene tree discordance (GTD). Controlling the effect of GTD by multiple approaches, I aimed to test whether the diminishing convergence pattern can be fully explained by GTD rather than true convergence under epistasis.

In Chapter 4, I revisited a previous study claiming genome-wide adaptive convergence for echolocation in three lineages of mammalian echolocators. With the question of whether the observed convergence can actually be explained without invoking adaptation to echolocation, I

repeated the analyses in properly chosen control lineages without apparent phenotypic convergence.

In Chapter 5, to answer the long-lasting question of whether molecular sequence data is more reliable than morphology in phylogenetic inference, I compared the amount of convergence between molecular sequence data and morphological data contained in a phylogenetic tree reconstruction dataset in a fair and straightforward manner. The underlying reason of convergence level difference between the two data types was then investigated. I also proposed a method to identify and remove highly convergent traits when combining morphological and molecular data in tree inference is necessary.

In Chapter 6, I focused on the more mechanistic codon substitution model, and infer the amino acid acceptance rates, or fixation probability of amino acid changes, in a broadly sampled set of species clades by maximum likelihood. I compared the pattern of relative acceptance rates in different clades to answer the question of whether there is among-clade heterogeneity. I designed statistical tests to check the significance of possible acceptance rate difference between clades, as well as simulations for positive/negative controls.

In Chapter 7, I investigated the question of whether nonsynonymous transitions are on average less deleterious than nonsynonymous transversions by incorporating this bias into a maximum likelihood model framework. ML inference of this bias was conducted in multiple species clades to show a whole picture across the tree of life. I also utilized a mechanistic codon

substitution model incorporating amino acid acceptance rates to explore the underlying mechanism of the possible bias by sequence evolution simulations.

Altogether, questions I ask in this dissertation address important but overlooked heterogeneities in protein sequence evolution. By answering these questions, we can establish better understanding of how different factors contribute to the phenomenon of sequence convergence and the overall protein sequence evolution. The findings could also inform us about potential model insufficiency or over-simplification regarding protein sequence evolution.

REFERENCES

Aminetzach YT, Srouji JR, Kong CY, Hoekstra HE. 2009. Convergent Evolution of Novel Protein Function in Shrew and Lizard Venom. Curr Biol 19:1925-1931.

Arendt J, Reznick D. 2008. Convergence and parallelism reconsidered: what have we learned about the genetics of adaptation? Trends Ecol Evol 23:26-32.

Bachtrog D, Thornton K, Clark A, Andolfatto P. 2006. Extensive introgression of mitochondrial DNA relative to nuclear genes in the Drosophila yakuba species group. Evolution 60:292-302.

Ballard JW. 2000. When one is not enough: introgression of mitochondrial DNA in Drosophila. Mol Biol Evol 17:1126-1130.

Bazykin GA, Kondrashov FA, Brudno M, Poliakov A, Dubchak I, Kondrashov AS. 2007. Extensive parallelism in protein evolution. Biol Direct 2:20.

Berens AJ, Hunt JH, Toth AL. 2015. Comparative transcriptomics of convergent evolution: different genes but conserved pathways underlie caste phenotypes across lineages of eusocial insects. Mol Biol Evol 32:690-703.

Bieler R, Mikkelsen PM, Collins TM, Glover EA, Gonzalez VL, Graf DL, Harper EM, Healy J, Kawauchi GY, Sharma PP, et al. 2014. Investigating the bivalve Tree of Life - an exemplar-based approach combining molecular and novel morphological characters. Invertebr Syst 28:32-115.

Blackledge TA, Gillespie RG. 2004. Convergent evolution of behavior in an adaptive radiation of Hawaiian web-building spiders. Proc Natl Acad Sci U S A 101:16228-16233.

Breen MS, Kemena C, Vlasov PK, Notredame C, Kondrashov FA. 2012. Epistasis as the primary factor in molecular evolution. Nature 490:535-538.

Castoe TA, de Koning AP, Kim HM, Gu W, Noonan BP, Naylor G, Jiang ZJ, Parkinson CL, Pollock DD. 2009. Evidence for an ancient adaptive episode of convergent molecular evolution. Proc Natl Acad Sci U S A 106:8986-8991.

Chikina M, Robinson JD, Clark NL. 2016. Hundreds of genes experienced convergent shifts in selective pressure in marine mammals. Mol Biol Evol 33:2182-2192.

Christin PA, Salamin N, Muasya AM, Roalson EH, Russier F, Besnard G. 2008. Evolutionary switch and genetic convergence on rbcL following the evolution of C4 photosynthesis. Mol Biol Evol 25:2361-2368.

Cutter AD. 2006. Nucleotide polymorphism and linkage disequilibrium in wild populations of the partial selfer Caenorhabditis elegans. Genetics 172:171-184.

Darwin C. 1859. On the origin of species by means of natural selection. London,: J. Murray.

Davalos LM, Velazco PM, Warsi OM, Smits PD, Simmons NB. 2014. Integrating incomplete fossils by isolating conflicting signal in saturated and non-independent morphological characters. Syst Biol 63:582-600.

Davies KTJ, Cotton JA, Kirwan JD, Teeling EC, Rossiter SJ. 2012. Parallel signatures of sequence evolution among hearing genes in echolocating mammals: an emerging model of genetic convergence. Heredity 108:480-489.

Dayhoff MO, Schwartz RM, Orcutt BC. 1978. A model of evolutionary changes in proteins. In: Dayhoff MO, editor. Atlas of protein sequence and structure. Silver Spring, MD: National Biomedical Research Foundation. p. 345-352.

Degnan JH, Rosenberg NA. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol Evol 24:332-340.

Dungan SZ, Chang BSW. 2017. Epistatic interactions influence terrestrial - marine functional shifts in cetacean rhodopsin. Proc R Soc B 284.

Elmer KR, Meyer A. 2011. Adaptation in the age of ecological genomics: insights from parallelism and convergence. Trends Ecol Evol 26:298-306.

Feldman CR, Brodie ED, Brodie ED, Pfrender ME. 2012. Constraint shapes convergence in tetrodotoxin-resistant sodium channels of snakes. Proc Natl Acad Sci U S A 109:4556-4561.

Fong SS, Joyce AR, Palsson BO. 2005. Parallel adaptive evolution cultures of Escherichia coli lead to convergent growth phenotypes with different gene expression states. Genome Res 15:1365-1372.

Foote AD, Liu Y, Thomas GW, Vinar 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.

Freudenberg-Hua Y, Freudenberg J, Kluck N, Cichon S, Propping P, Nothen MM. 2003. Single nucleotide variation analysis in 65 candidate genes for CNS disorders in a representative sample of the European population. Genome Res 13:2271-2276.

Gaubert P, Wozencraft WC, Cordeiro-Estrela P, Veron G. 2005. Mosaics of convergences and noise in morphological phylogenies: what's in a viverrid-like carnivoran? Syst Biol 54:865-894.

Givnish TJ, Sytsma KJ. 1997. Consistency, characters, and the likelihood of correct phylogenetic inference. Mol Phylogenet Evol 7:320-330.

Goldman N, Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 11:725-736.

Gong LI, Bloom JD. 2014. Epistatically interacting substitutions are enriched during adaptive protein evolution. PLoS Genet 10:e1004328.

Grantham R. 1974. Amino acid difference formula to help explain protein evolution. Science 185:862-864.

Grenier JL, Greenberg R. 2005. A biogeographic pattern in sparrow bill morphology: parallel adaptation to tidal marshes. Evolution 59:1588-1595.

Haag-Liautard C, Coffey N, Houle D, Lynch M, Charlesworth B, Keightley PD. 2008. Direct estimation of the mitochondrial DNA mutation rate in Drosophila melanogaster. PLoS Biol 6:e204.

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

Hershberg R, Petrov DA. 2010. Evidence that mutation is universally biased towards AT in bacteria. PLoS Genet 6:e1001115.

Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C, Ho SY, Faircloth BC, Nabholz B, Howard JT, et al. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science 346:1320-1331.

Jiang C, Zhao Z. 2006. Mutational spectrum in the recent human genome inferred by single nucleotide polymorphisms. Genomics 88:527-534.

Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 8:275-282.

Jost MC, Hillis DM, Lu Y, Kyle JW, Fozzard HA, Zakon HH. 2008. Toxin-resistant sodium channels: parallel adaptive evolution across a complete gene family. Mol Biol Evol 25:1016-1024.

Kimura M. 1983. The neutral theory of molecular evolution. Cambridge Cambridgeshire ; New York: Cambridge University Press.

Langerhans RB, Knouft JH, Losos JB. 2006. Shared and unique features of diversification in Greater Antillean Anolis ecomorphs. Evolution 60:362-369.

Le SQ, Gascuel O. 2008. An improved general amino acid replacement matrix. Mol Biol Evol 25:1307-1320.

Lee MSY, Soubrier J, Edgecombe GD. 2013. Rates of phenotypic and genomic evolution during the Cambrian explosion. Curr Biol 23:1889-1895.

Legg DA, Sutton MD, Edgecombe GD. 2013. Arthropod fossil data increase congruence of morphological and molecular phylogenies. Nat Commun 4:2485.

Li C, Qian WF, Maclean CJ, Zhang JZ. 2016. The fitness landscape of a tRNA gene. Science 352:837-840.

Li Y, Liu Z, Shi P, Zhang JZ. 2010. The hearing gene Prestin unites echolocating bats and whales. Curr Biol 20:R55-R56.

Linnen CR, Poh YP, Peterson BK, Barrett RDH, Larson JG, Jensen JD, Hoekstra HE. 2013. Adaptive evolution of multiple traits through multiple mutations at a single gene. Science 339:1312-1316.

Liu Y, Han N, Franchini LF, Xu H, Pisciottano F, Elgoyhen AB, Rajan KE, Zhang S. 2012. The voltage-gated potassium channel subfamily KQT member 4 (KCNQ4) displays parallel evolution in echolocating bats. Mol Biol Evol 29:1441-1450.

Liu Z, Li S, Wang W, Xu D, Murphy RW, Shi P. 2011. Parallel evolution of *KCNQ4* in echolocating bats. PLoS One 6:e26618.

Liu Z, Qi FY, Zhou X, Ren HQ, Shi P. 2014. Parallel sites implicate functional convergence of the hearing gene *prestin* among echolocating mammals. Mol Biol Evol 31:2415-2424.

Livezey BC, Zusi RL. 2007. Higher-order phylogeny of modern birds (Theropoda, Aves: Neornithes) based on comparative anatomy. II. Analysis and discussion. Zool J Linn Soc 149:1-95.

Losos JB. 2011. Convergence, adaptation, and constraint. Evolution 65:1827-1840.

Lynch M, Sung W, Morris K, Coffey N, Landry CR, Dopman EB, Dickinson WJ, Okamoto K, Kulkarni S, Hartl DL, et al. 2008. A genome-wide view of the spectrum of spontaneous mutations in yeast. Proc Natl Acad Sci U S A 105:9272-9277.

Mallet J, Besansky N, Hahn MW. 2016. How reticulated are species? Bioessays 38:140-149.

Manceau M, Domingues VS, Linnen CR, Rosenblum EB, Hoekstra HE. 2010. Convergence in pigmentation at multiple levels: mutations, genes and function. Philos Trans R Soc Lond B Biol Sci 365:2439-2450.

Mazel F, Wuest RO, Gueguen M, Renaud J, Ficetola GF, Lavergne S, Thuiller W. 2017. The geography of ecological niche evolution in mammals. Curr Biol 27:1369-1374.

McGhee GR. 2011. Convergent Evolution: Limited Forms Most Beautiful. Vienna Ser Theor Bio:1-322.

Melville J, Harmon LJ, Losos JB. 2006. Intercontinental community convergence of ecology and morphology in desert lizards. Proc Biol Sci 273:557-563.

Meredith RW, Janecka JE, Gatesy J, Ryder OA, Fisher CA, Teeling EC, Goodbla A, Eizirik E, Simao TL, Stadler T, et al. 2011. Impacts of the Cretaceous Terrestrial Revolution and KPg extinction on mammal diversification. Science 334:521-524.

Miyata T, Miyazawa S, Yasunaga T. 1979. Two types of amino acid substitutions in protein evolution. J Mol Evol 12:219-236.

Moore J, Willmer P. 1997. Convergent evolution in invertebrates. Biol Rev Camb Philos Soc 72:1-60.

Nevo E. 1979. Adaptive convergence and divergence of subterranean mammals. Annu Rev Ecol Syst 10:269-308.

O'Leary MA, Bloch JI, Flynn JJ, Gaudin TJ, Giallombardo A, Giannini NP, Goldberg SL, Kraatz BP, Luo ZX, Meng J, et al. 2013. The placental mammal ancestor and the post-K-Pg radiation of placentals. Science 339:662-667.

Ohta T. 1992. The nearly neutral theory of molecular evolution. Annu Rev Ecol Syst 23:263-286.

Ossowski S, Schneeberger K, Lucas-Lledo JI, Warthmann N, Clark RM, Shaw RG, Weigel D, Lynch M. 2010. The rate and molecular spectrum of spontaneous mutations in Arabidopsis thaliana. Science 327:92-94.

Page RDM, Holmes EC. 1998. Molecular evolution : a phylogenetic approach. Oxford ; Malden, MA: Blackwell Science.

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.

Pfenning AR, Hara E, Whitney O, Rivas MV, Wang R, Roulhac PL, Howard JT, Wirthlin M, Lovell PV, Ganapathy G, et al. 2014. Convergent transcriptional specializations in the brains of humans and song-learning birds. Science 346:1333-+.

Podgornaia AI, Laub MT. 2015. Pervasive degeneracy and epistasis in a protein-protein interface. Science 347:673-677.

Pollard DA, Iyer VN, Moses AM, Eisen MB. 2006. Widespread discordance of gene trees with species tree in *Drosophila*: evidence for incomplete lineage sorting. PLoS Genet 2:e173.

Projecto-Garcia J, Natarajan C, Moriyama H, Weber RE, Fago A, Cheviron ZA, Dudley R, McGuire JA, Witt CC, Storz JF. 2013. Repeated elevational transitions in hemoglobin function during the evolution of Andean hummingbirds. Proc Natl Acad Sci U S A 110:20669-20674.

Protas ME, Hersey C, Kochanek D, Zhou Y, Wilkens H, Jeffery WR, Zon LI, Borowsky R, Tabin CJ. 2006. Genetic analysis of cavefish reveals molecular convergence in the evolution of albinism. Nat Genet 38:107-111.

Pyron RA. 2015. Post-molecular systematics and the future of phylogenetics. Trends Ecol Evol 30:384-389.

Rokas A, Carroll SB. 2008. Frequent and widespread parallel evolution of protein sequences. Mol Biol Evol 25:1943-1953.

Rosenberg MS, Subramanian S, Kumar S. 2003. Patterns of transitional mutation biases within and among mammalian genomes. Mol Biol Evol 20:988-993.

Rosenblum EB, Rompler H, Schoneberg T, Hoekstra HE. 2010. Molecular and functional basis of phenotypic convergence in white lizards at White Sands. Proc Natl Acad Sci U S A 107:2113-2117.

Schrider DR, Houle D, Lynch M, Hahn MW. 2013. Rates and genomic consequences of spontaneous mutational events in Drosophila melanogaster. Genetics 194:937-954.

Scornavacca C, Galtier N. 2017. Incomplete lineage sorting in mammalian phylogenomics. Syst Biol 66:112-120.

Shapiro MD, Bell MA, Kingsley DM. 2006. Parallel genetic origins of pelvic reduction in vertebrates. Proc Natl Acad Sci U S A 103:13753-13758.

Shen YY, Liang L, Li GS, Murphy RW, Zhang YP. 2012. Parallel evolution of auditory genes for echolocation in bats and toothed whales. PLoS Genet 8:e1002788.

Shen YY, Liu J, Irwin DM, Zhang YP. 2010. Parallel and convergent evolution of the dim-light vision gene *RH1* in bats (Order: Chiroptera). PLoS One 5:e8838.

Springer MS, Meredith RW, Teeling EC, Murphy WJ. 2013. Technical comment on "The placental mammal ancestor and the post-K-Pg radiation of placentals". Science 341:613.

Stayton CT. 2008. Is convergence surprising? An examination of the frequency of convergence in simulated datasets. J Theor Biol 252:1-14.

Stoltzfus A, Norris RW. 2016. On the causes of evolutionary transition:transversion bias. Mol Biol Evol 33:595-602.

Tang H, Wyckoff GJ, Lu J, Wu CI. 2004. A universal evolutionary index for amino acid changes. Mol Biol Evol 21:1548-1556.

Ujvari B, Casewell NR, Sunagar K, Arbuckle K, Wuster W, Lo N, O'Meally D, Beckmann C, King GF, Deplazes E, et al. 2015. Widespread convergence in toxin resistance by predictable molecular evolution. Proc Natl Acad Sci U S A 112:11911-11916.

Wake DB, Wake MH, Specht CD. 2011. Homoplasy: from detecting pattern to determining process and mechanism of evolution. Science 331:1032-1035.

Whelan S, Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol 18:691-699.

Wiens JJ, Kuczynski CA, Townsend T, Reeder TW, Mulcahy DG, Sites JW, Jr. 2010. Combining phylogenomics and fossils in higher-level squamate reptile phylogeny: molecular data change the placement of fossil taxa. Syst Biol 59:674-688.

Wittkopp PJ, Williams BL, Selegue JE, Carroll SB. 2003. Drosophila pigmentation evolution: divergent genotypes underlying convergent phenotypes. Proc Natl Acad Sci U S A 100:1808-1813.

Xu JR, Zhang JZ. 2014. Why human disease-associated residues appear as the wild-type in other species: genome-scale structural evidence for the compensation hypothesis. Mol Biol Evol 31:1787-1792.

Xu S, He Z, Guo Z, Zhang Z, Wyckoff GJ, Greenberg A, Wu CI, Shi S. 2017. Genome-wide convergence during evolution of mangroves from woody plants. Mol Biol Evol 34:1008-1015.

Yampolsky LY, Stoltzfus A. 2005. The exchangeability of amino acids in proteins. Genetics 170:1459-1472.

Yang Z. 2006. Computational molecular evolution. Oxford: Oxford University Press.

Yang Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol 15:568-573.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24:1586-1591.

Yang Z, Nielsen R, Hasegawa M. 1998. Models of amino acid substitution and applications to mitochondrial protein evolution. Mol Biol Evol 15:1600-1611.

Zhang J. 2006. Parallel adaptive origins of digestive RNases in Asian and African leaf monkeys. Nat Genet 38:819-823.

Zhang J. 2000. Rates of conservative and radical nonsynonymous nucleotide substitutions in mammalian nuclear genes. J Mol Evol 50:56-68.

Zhang J, Kumar S. 1997. Detection of convergent and parallel evolution at the amino acid sequence level. Mol Biol Evol 14:527-536.

Zhen Y, Aardema ML, Medina EM, Schumer M, Andolfatto P. 2012. Parallel molecular evolution in an herbivore community. Science 337:1634-1637.

Zhou XM, Seim I, Gladyshev VN. 2015. Convergent evolution of marine mammals is associated with distinct substitutions in common genes. Sci Rep 5.

Zhu YO, Siegal ML, Hall DW, Petrov DA. 2014. Precise estimates of mutation rate and spectrum in yeast. Proc Natl Acad Sci U S A 111:E2310-2318.

CHAPTER 2

Are Convergent and Parallel Amino Acid Substitutions in Protein Evolution more Prevalent than Neutral Expectations?[1]

2.1 ABSTRACT

Convergent and parallel amino acid substitutions in protein evolution, collectively referred to as molecular convergence here, have small probabilities under neutral evolution. For this reason, molecular convergence is commonly viewed as evidence for similar adaptations of different species. The surge in the number of reports of molecular convergence in the last decade raises the intriguing question of whether molecular convergence occurs substantially more frequently than expected under neutral evolution. We here address this question using all one-to-one orthologous proteins encoded by the genomes of 12 fruit fly species and those encoded by 17 mammals. We found that the expected amount of molecular convergence varies greatly depending on the specific neutral substitution model assumed at each amino acid site and that the observed amount of molecular convergence is explainable by neutral models incorporating site-specific information of acceptable amino acids. Interestingly, the total number of convergent and parallel substitutions between two lineages, relative to the neutral expectation, decreases with the genetic distance between the two lineages, regardless of the model used in computing the neutral expectation. We hypothesize that this trend results from differences in the amino acids

[1] This chapter is published as: Zou Z, Zhang J. 2015. Are convergent and parallel amino acid substitutions in protein evolution more prevalent than neutral expectations? Mol Biol Evol, 32: 2085-2096.

acceptable at a given site among different clades of a phylogeny, due to prevalent epistasis, and provide simulation as well as empirical evidence for this hypothesis. Together, our study finds no genomic evidence for higher-than-neutral levels of molecular convergence, but suggests the presence of abundant epistasis that decreases the likelihood of molecular convergence between distantly related lineages.

2.2 INTRODUCTION

Convergence refers to the evolutionary phenomenon that identical or similar traits emerge independently in two or more lineages such as the origins of wings in birds and bats (Stern 2013). Phenotypic convergence is widespread and has long been viewed as evidence for independent adaptations of different species to a common environmental challenge, because the probability of multiple independent origins of the same complex trait by genetic drift alone is likely to be extremely low (McGhee 2011). Convergence can also occur at the protein sequence level, and such molecular convergences are often separated into two types: convergent and parallel amino acid substitutions (Zhang and Kumar 1997). Convergent substitutions at an amino acid position of a protein refer to changes from different ancestral amino acids to the same descendant amino acid along independent evolutionary lineages. They are distinguished from parallel substitutions where the independent changes have occurred from the same ancestral amino acid. For simplicity, we refer to both types as molecular convergence in this paper, unless otherwise noted. Similar to phenotypic convergence, molecular convergence is widely believed

to reflect common adaptations of different organisms. But, because of the limited number of amino acids acceptable at any position, molecular convergence may occur by chance without the involvement of positive selection (Zhang and Kumar 1997).

The last decade has seen a surge in the number of reports of molecular convergence, virtually all of which were interpreted as results of positive selection (Castoe, et al. 2009; Christin, et al. 2010; Christin, et al. 2008; Davies, et al. 2012; Feldman, et al. 2012; Jost, et al. 2008; Li, et al. 2010; Liu, et al. 2010; Liu, et al. 2012; Liu, et al. 2011; Shen, et al. 2012; Stern 2013; Zhang 2006; Zhen, et al. 2012), although rigorous demonstrations of the involvement of adaptive selection are not easy and thus have been rare (Zhang 2006). For example, seven hearing-related proteins are known to exhibit various degrees of molecular convergence among two groups of bats and toothed whales that independently acquired echolocation (Davies, et al. 2012; Li, et al. 2008; Li, et al. 2010; Liu, et al. 2010; Liu, et al. 2012; Liu, et al. 2011; Shen, et al. 2012). But, only in prestin, the motor protein of the outer hair cells of the inner ear of the mammalian cochlea, is there evidence that the number of observed parallel amino acid substitutions in echolocators significantly exceeds the chance expectation (Li, et al. 2010) and that these parallel substitutions are responsible for parallel functional changes of the protein (Liu, et al. 2014). Despite these caveats, the growing number of molecular convergences discovered raises the intriguing question of whether adaptive molecular convergence is a common phenomenon in protein evolution.

Rokas and Carroll (2008) addressed the above question by examining eight genome-scale

gene sets, each including four species. They showed that, in each gene set, the observed number of parallel amino acid substitutions significantly exceeds the random expectation under a neutral model of amino acid substitution. They suggested that this excess arose from common positive selection in two lineages and/or purifying selection constraining the number of amino acids acceptable at a site that was not incorporated into their neutral model. A similar conclusion was reached by Bazykin, et al. (2007). Castoe et al. (2009) also reported a larger amount of molecular convergence as well as divergence in vertebrate mitochondrial genes than expected from a neutral model. However, none of the studies investigated whether the observed amount of molecular convergence can be fully explained without invoking positive selection. As such, the prevalence of adaptive molecular convergence remains unclear.

In this study, we address the above question using genome-wide datasets of protein sequence alignments of fruit flies and mammals, respectively. We compare the inferred numbers of molecular convergences between a pair of lineages with the neutral expectations derived from several different substitution models, including those incorporating site-specific amino acid compositions. We found that the neutral expectations vary substantially depending on the model used and that some neutral models are capable of explaining the large numbers of molecular convergences observed. Interestingly, the observed number of molecular convergences relative to the neutral expectation decreases with the genetic distance between the two evolutionary lineages concerned, regardless of the specific neutral model assumed. We propose and provide

evidence that this phenomenon is a result of prevalent epistasis in protein evolution, which renders the amino acids acceptable at a position different in different species.

2.3 RESULTS

Observed and expected numbers of molecular convergences

Let us use an alignment of seven orthologous protein sequences, whose phylogenetic relationships are depicted by the tree in **Fig. 2.1**, as an example to illustrate our analysis. Suppose we are interested in molecular convergence along the two thick branches (**Fig. 2.1**). We first infer the ancestral amino acids at all interior nodes of the tree for each site of the protein (see MATERIALS AND METHODS). Let X_i be the amino acid at node *i* for a given site. By definition, convergent substitutions on the thick branches occur at those sites where $X_1 \neq X_2$, $X_3 = X_4$, $X_3 \neq X_1$, and $X_4 \neq X_2$. Similarly, parallel substitutions on the thick branches occur at those sites where $X_1 = X_2$, $X_3 = X_4$, $X_3 \neq X_1$, and $X_4 \neq X_2$. This way, the numbers of sites that have experienced convergent and parallel substitutions in the thick branches are respectively counted and referred to as the "observed" numbers of convergent and parallel substitutions.

Under the assumption that amino acid substitutions at a site follow a Markov process, we can compute the probability that a site experiences convergent (or parallel) substitutions along the thick branches, given the amino acid substitution rate matrix, the substitution rate at the site relative to the average rate of the protein considered, amino acid equilibrium frequencies, and all branch lengths measured by the expected numbers of substitutions per site for the protein

considered (see MATERIALS AND METHODS). For a protein, the expected number of sites with convergent (or parallel) substitutions is the sum of these probabilities across all sites of the protein.

Using this framework, we compared the observed and expected numbers of convergent and parallel substitutions, respectively, in 5935 orthologous protein alignments of 12 *Drosophila* species, totaling 2,028,428 amino acid sites after the removal of gaps and ambiguous sites. For each alignment, we used PAML (Yang 2007) to infer the branch lengths, substitution rate of each site relative to the average rate of the entire protein, and ancestral sequences under the known topology of the species tree (**Fig. 2.2a**). We first focused on the two exterior branches that respectively lead to *D. yakuba* and *D. mojavensis*. In computing the expected numbers of convergent and parallel substitutions in a protein, we used the JTT-f_{gene} model of amino acid substitution. This model is based on the average substitution patterns of many proteins known as the JTT model (Jones et al. 1992), with the equilibrium frequencies of the 20 amino acids replaced by the observed amino acid frequencies in the protein concerned.

The total number of observed convergent sites in the 5935 fly proteins is 292 for the pair of branches considered, whereas the expected number is only 194.2 (**Table 2.1**); the difference is statistically significant ($P < 10^{-10}$, Poisson test). The total number of observed parallel sites is 650, also significantly greater than the expected number of 388.6 ($P < 10^{-122}$). The ratio (R) between the observed and expected numbers of sites is 1.50 for convergent substitutions and 1.67 for parallel substitutions (**Table 2.1**). Rokas and Carroll (2008) were unable to study convergent

substitutions due to their use of four-taxon trees. For parallel substitutions, our result is similar to what Rokas and Carroll reported.

Considering that the amino acids acceptable at a site likely differ from those acceptable at another site because of differences in structural and functional roles of different sites, we used a second model termed JTT-f_{site} to compute the expected numbers of convergent and parallel sites, respectively. That is, for each site, the equilibrium amino acid frequencies in the JTT model are replaced with the observed amino acid frequencies at the site across all sequences in the alignment. We found that the number of observed amino acids at a site averages 1.56 across all sites and 2.74 across all variable sites. Obviously, considering this small number of acceptable amino acids at a site should increase the expected number of molecular convergence. Indeed, the expected numbers of convergent (475.2) and parallel (2125.7) sites both increase substantially, compared with those under the JTT-f_{gene} model (**Table 2.1**). As a result, *R* reduces to 0.61 for convergent sites and 0.31 for parallel sites, respectively (**Table 2.1**). Thus, if the amino acid frequencies observed at a site across the 12 *Drosophila* species truly reflect the equilibrium frequencies at the site, molecular convergence has occurred not more but less frequently than what the neutral model predicts. One caveat in the above analysis is that, because the number of taxa used is smaller than 20 and because the total branch length (0.796 substitutions per site) of the *Drosophila* tree is also much smaller than 20, the observation of a limited number of different amino acids at a site may not mean that only the observed amino acids are acceptable but could be due to insufficient evolutionary time and taxon sampling for all acceptable amino acids to

appear.

For the above reason, we tried the third model, JTT-CAT (Lartillot and Philippe 2004), in estimating the expected numbers of convergent and parallel sites. Instead of having one set of equilibrium amino acid frequencies for all sites of a protein (JTT-f_{gene}) or one set per site (JTT-f_{site}), CAT uses a Bayesian mixture model for among-site heterogeneities in amino acid frequencies. It estimates the total number of classes of sites and their respective amino acid frequencies, as well as the affiliation of each site to a given class. We expect that the JTT-CAT model will produce results that are between those from JTT-f_{gene} and JTT-f_{site}. However, because JTT-CAT is highly computationally intensive, we analyzed 1081 relatively long proteins from the entire set of 5935 proteins in an attempt to acquire the most information with the least amount of time. The expected numbers of convergent and parallel sites under JTT-CAT (**Table 2.1**), after being extrapolated to all 5935 proteins, are 306.6 and 480.2, respectively. As predicted, these numbers are between the corresponding values under JTT-f_{gene} and JTT-f_{site}. Consequently, *R* values under JTT-CAT (0.79 for convergent sites and 1.18 for parallel sites) are between those under JTT-f_{gene} and JTT-f_{site} (**Table 2.1**).

To examine if the above patterns are specific to the pair of branches considered, we also analyzed molecular convergence for the two exterior branches respectively leading to *D. melanogaster* and *D. yakuba*. Similar patterns were found, although both observed and expected numbers of molecular convergences are much lower for this branch pair (**Table A.1.1**), likely due to the much shorter *D. melanogaster* branch compared with the *D. mojavensis* branch (**Fig.**

2.2a). Together, these results show that the observed numbers of convergent and parallel substitutions are no longer greater than their respective neutral expectations when among-site heterogeneities in equilibrium amino acid frequencies are considered.

Lower rates of molecular convergence in more distantly related lineages

In addition to the two pairs of exterior branches in the *Drosophila* tree, we analyzed all other pairs of branches in the tree that are unconnected and do not belong to the same evolutionary path. We excluded connected branch pairs because of the difficulty in inferring molecular convergence in these branch pairs. For example, parallel substitutions in the exterior branches respectively leading to *D. yakuba* and *D. erecta* will almost always be inferred as a substitution in the interior branch leading to the common ancestor of these two species (**Fig. 2.2a**). Pairs of branches in the same evolutionary path were excluded, because in such cases the node at the beginning of one branch is a descendant of the node at the end of the other branch, violating the requirement for independent evolution in the definition of convergence. Note that although the tree in **Fig. 2.2a** is unrooted, we treated the deepest node as the root when deciding the evolutionary direction, which should not affect our analysis under the Markov process of amino acid substitution assumed here. For each pair of branches considered, we calculated the aforementioned ratio (R) between the number of molecular convergences (i.e., the total number of convergent and parallel sites) observed and that expected under a neutral substitution model. Under the same substitution model, we compared R of different branch pairs. Interestingly, R

declines with the increase of the genetic distance between the two branches compared, where the genetic distance is measured by the total length of the branches connecting the young ends of the two branches concerned (**Fig. 2.2b**). Because the same branch is used in multiple branch pairs, branch pairs are not independent from one another. We thus tested the statistical significance of Pearson's correlation between *R* and genetic distance using a Mantel test that controls such non-independence (Mantel 1967). We found the correlation significant when the neutral expectations were computed under the JTT-f_{gene} model (Pearson's correlation coefficient r = - 0.426, P = 0.0006) or JTT-f_{site} model (r = - 0.274, P = 0.017). When the neutral expectations were computed under JTT-CAT (based on the subset of 1081 genes analyzed), the correlation was marginally significant (r = - 0.174, P = 0.069).

To examine the generality of the above observation, we repeated the analysis in a set of 17 mammals, including 14 placental mammals, two marsupials, and the monotreme platypus (**Fig. 2.3a**). The dataset consists of 2759 one-to-one orthologous proteins, with a total length of 1,079,696 amino acid sites after the removal of gaps and ambiguous sites. We used either the JTT-f_{gene} model or JTT-f_{site} model to compute the expected numbers of molecular convergences, but did not use the JTT-CAT model due to its high demand for computing time. The results obtained from the mammalian proteins are highly similar to those from the *Drosophila* proteins. First, *R* generally exceeds 1 under JTT-f_{gene} but is lower than 1 under JTT-f_{site} (**Fig. 2.3b**). Second, regardless of the substitution model used in computing the expected numbers of molecular convergences, *R* declines with the increase of the genetic distance between the pair of

evolutionary lineages compared (**Fig. 2.3b**). Mantel test of the negative correlation between genetic distance and *R* showed statistical significance under each model applied (r = - 0.747, P = 0.0002 under JTT-f_{gene}; r = - 0.618, P = 0.0004 under JTT-f_{site}).

Epistasis reduces the probability of molecular convergence between divergent lineages

What makes the observed number of molecular convergences relative to the neutral expectation decrease as the two lineages compared diverge? One likely scenario is that, at a given site, amino acids that are acceptable in one clade of a phylogeny become unacceptable in another clade, resulting in a decrease in the probability of convergence. In other words, if equilibrium amino acid frequencies at a site gradually change in evolution, branch pairs with higher genetic distances should show lower probabilities of molecular convergence, which is not considered in the current computation of the neutral expectation and hence results in lower *R* values.

To test the hypothesis that changing site-specific equilibrium amino acid compositions in evolution could generate a negative correlation between *R* and the genetic distance between the branches under study, we first conducted a computer simulation using a simple tree of four taxa (**Fig. 2.4a**), in which the two thick branches being investigated for molecular convergence have the same length of b_2, whereas the two interior branches have the same length of b_1. Thus, the genetic distance between nodes 2 and 4 is $B = 2(b_1+b_2)$. We simulated the evolution of 500,000 sites using a modified JTT-f_{site} model, where the equilibrium amino acid frequencies at each site

gradually change in a random-walk fashion from the initial values taken from the original JTT model (Jone et al. 1992). For the two thick branches, we counted the number of molecular convergences that occurred and computed the expected number of molecular convergences assuming that the equilibrium amino acid frequencies were constant during evolution and equaled the average equilibrium frequencies in nodes 2 and 4. As predicted, the simulation showed that the number of observed molecular convergences relative to the expected number decreases with the rise in B (r = -0.51, P = 0.019), demonstrating that our hypothesis of changing site-specific equilibrium amino acid frequencies in evolution can in principle explain the reduction in the probability of molecular convergence between distantly related lineages. As a negative control, we repeated the above simulation with constant equilibrium amino acid frequencies in evolution. As expected, the number of observed molecular convergences relative to the expected number is no longer correlated with B (r = -0.13, P = 0.57).

To examine if acceptable amino acids at a site indeed differ between clades of organisms, we analyzed 16 proteins that have orthologous sequences from hundreds to thousands of species (Breen, et al. 2012). They include 13 mitochondrial genome-encoded proteins, one chloroplast genome-encoded protein (Rubisco), and two nuclear genome-encoded proteins (elongation factor and histone). For each protein alignment, two mutually exclusive monophyletic clades were chosen, and the presence/absence of each of the 20 amino acids at each site in each clade was recorded. Because the number of taxa within each clade is large and the total branch length within each clade is >>20 for most of the 16 proteins (**Fig. 2.5a**), the amino acids allowed at a

particular site can be approximated by the observed amino acids. For each site, we used the number of amino acids present in one clade but absent in the other clade (i.e., Hamming distance) as a measure of their amino acid compositional distance. For comparison, we computed the compositional distances after 1000 random separations of all sequences from the two clades into two groups that are of the same sizes as the original clades. We calculated the *P*-value as the proportion of times in which the randomized compositional distance equals or exceeds the observed distance. Because one test was conducted for each site in a protein, we corrected for multiple testing by converting the *P*-values to corresponding *Q*-values using the Benjamini–Hochberg method (Benjamini and Hochberg 1995). For all but one protein, the observed compositional distance is significantly (i.e., Q-value < 0.05) greater than the random expectation for a considerable number of sites (**Fig. 2.5a**). The exception is the highly conserved histone H3.2, for which none of the 120 sites show significant between-clade differences in acceptable amino acids.

In the above analysis, the two clades defined in the analysis of each protein tend to be old (e.g., ray-fined fishes and tetrapods) such that the two clades are relatively distantly related. Between such distantly related clades, it may not be surprising that acceptable amino acids are significantly different. To examine if the same phenomenon exists between relatively closely related clades, we examined COX2 and CYTB, for which sufficient numbers of sequences are available for this analysis. We found that between the *Drosophila* and *Sophophora* subgenera, 4 of 229 sites in COX2 show significant amino acid compositional differences (**Fig. 2.5b**).

Similarly, between the sister families of Muridae and Cricetidae, 50 of 381 sites in CYTB show significant compositional differences (**Fig. 2.5b**).

These results demonstrate that acceptable amino acids at a site change significantly between sister families of mammals or even within an insect genus during evolution. It is likely that epistasis, or interactions between amino acid residues within or between proteins, is the cause of this change. In the presence of epistasis, the amino acids acceptable at a site depend on the amino acids at its interacting sites. Consequently, when the amino acids at the interacting sites change in evolution, the amino acids acceptable at the focal site also change, resulting in an alteration of site-specific equilibrium amino acid frequencies. In essence, the microenvironment of the focal site changes in evolution, rendering the same amino acid different in functional effect, which reduces the probability of molecular convergence.

An alternative explanation of a decreasing R with an increasing genetic distance is a genome-wide change in amino acid content during evolution. To evaluate this possibility, for a pair of branches in the *Drosophila* or mammalian dataset, we computed the amino acid frequency vector for each younger end of the two branches for sites that differ between the two younger ends, and then calculated the Euclidian distance between the two vectors. We conducted a partial Mantel test of the correlation between R and genetic distance among branch pairs, after the control of the above Euclidian distance. We found that the negative correlation between R and genetic distance remains largely unchanged even after the control (*Drosophila*: $r = -0.471$, $P = 5\times10^{-5}$ under JTT-f_{gene}; $r = -0.297$, $P = 0.013$ under JTT-f_{site}; $r = -0.187$, $P = 0.062$ under JTT-

CAT. Mammals: $r = -0.752$, $P = 5\times10^{-5}$ under JTT-f_{gene}; $r = -0.622$, $P = 1.5\times10^{-4}$ under JTT-f_{site}). Hence, potential genome-wide changes in amino acid content cannot explain the negative correlation.

2.4 DISCUSSION

To examine the prevalence of adaptive molecular convergence in protein sequence evolution, we calculated the ratio (R) between the number of observed molecular convergences in a pair of branches and the expected number under various neutral substitution models. We found that R generally exceeds 1 when all sites in a protein are assumed to have the same equilibrium amino acid frequencies (**Figs. 2.2 and 2.3**). But when the among-site heterogeneity in equilibrium amino acid frequencies is considered by either the CAT model or the observed amino acid frequencies at each site, R is generally close to or smaller than 1 (**Figs. 2.2 and 2.3**). As shown previously, models considering the among-site heterogeneity in equilibrium amino acid frequencies almost always fit actual protein sequence data better than comparable models assuming among-site homogeneity (Lartillot and Philippe 2004; Lartillot and Philippe 2006). Because the amount of molecular convergence observed in both fruit flies and mammals can be largely explained by chance under a reasonable neutral substitution model, we conclude that there is no evidence for prevalent adaptive molecular convergence at the genomic scale. In this context, it is worth mentioning a recent study that claimed the detection of genomic signatures of adaptive molecular convergence in echolocating mammals (Parker, et al. 2013). Two

subsequently analyses, however, found no evidence supporting this claim (Thomas and Hahn 2015; Zou and Zhang 2015).

The lack of genome-wide excess of molecular convergence relative to the neutral expectation is not incompatible with adaptive molecular convergence in a small number of proteins. But, what is notable is the relatively large number of molecular convergences at the proteome level that are expected under realistic neutral models. For instance, between the exterior branches respectively leading to *D. yakuba* and *D. mojavensis* (**Fig. 2.2a**), 2601 molecular convergences are expected among ~2.03 million sites under the JTT-f_{site} model (**Table 2.1**). This frequency means that 1.28 molecular convergences are expected for an average protein of 1000 amino acids. The chance probability of observing three or more molecular convergences in this protein would be 0.138. It is thus almost guaranteed to find a protein with this amount of molecular convergence from a sizeable set of proteins surveyed. In other words, observations of molecular convergence, especially through a search in multiple proteins, should not be automatically interpreted as evidence for adaptation. Zhang and Kumar previously proposed a statistical test for adaptive molecular convergence in a protein (Zhang and Kumar 1997), but the substitution model they used was JTT-f_{gene}. We suggest that JTT-f_{site} or JTT-CAT be used in Zhang and Kumar's test to guard against false positives caused by the use of neutral models with inadequate among-site heterogeneity in equilibrium amino acid frequencies. When multiple proteins are searched, a correction for multiple testing should also be applied. Castoe et al. (2009) previously showed that the number of sites experiencing convergence substitutions

relative to the number of sites experiencing divergent substitutions (*C*/*D*) is approximately constant across different branch pairs in a tree. They suggested that adaptive convergence can be detected for a branch pair if its *C*/*D* substantially exceeds those of other branch pairs. Unfortunately, this signal simply indicates a variation in *C*/*D* among branch pairs; it does not prove or disprove adaptive convergence. For instance, a uniform *C*/*D* among branch pairs could mean widespread adaptive molecular convergence throughout the tree. Conversely, variation in *C*/*D* among branch pairs could arise from non-adaptive processes (Goldstein, et al. 2015).

We found in both *Drosophila* proteins and mammalian proteins that *R* decreases with the increase of the genetic distance between the two lineages where molecular convergence is examined. Notably, a related phenomenon was reported by Rogozin and colleagues in the analysis of highly conserved (but not invariable) amino acid sites that they used for reconstructing the metazoan phylogeny (Rogozin et al. 2008). These authors noted that when parallel substitutions were observed at such sites, the substitutions were more likely to occur in interior branches of the tree rather than exterior branches. Because exterior branches tend to be relatively distant from one another compared with interior branches, their observation is broadly consistent with ours, although these authors did not explicitly consider the expected number of parallel sites. While our paper was under review, Goldstein et al. (2015) published a similar finding in mitochondrial genes. They showed that *C*/*D* decreases when the genetic distance between the branch pair under investigation increases. However, because this trend of *C*/*D* is expected even under simple neutral models without epistasis (Goldstein et al. 2015), the

biological meaning of their finding is less clear than that of our *R*-based result.

We hypothesize that the negative correlation between *R* and the genetic distance of the two lineages considered is caused by changes in acceptable amino acids at individual sites of a protein during evolution. Indeed, the negative correlation could be recapitulated by a simulation of protein sequence evolution with gradual, random changes in site-specific equilibrium amino acid frequencies. Furthermore, the sequences of 16 proteins with hundreds to thousands of orthologs revealed widespread among-clade differences in amino acid compositions at individual sites. Our hypothesis is also consistent with previous case studies where the same amino acid substitutions show similar functional effects in relatively closely related homologs, but show different and even opposite functional effects in relatively distantly related homologs (Zhang 2003).

If the equilibrium amino acid frequencies at a site change in evolution, the equilibrium frequencies for a *Drosophila* species or lineage would differ from the average equilibrium frequencies calculated from all sequences in the *Drosophila* tree. Consequently, the number of acceptable amino acids at the site for the species is likely smaller than predicted from the average equilibrium frequencies. This bias causes an underestimation of the expected number of molecular convergences and hence an overestimation of *R*. Hence, positive selection need not be invoked even when *R* exceeds 1 under JTT-CAT or JTT-f_{site}, as observed in some closely related lineages (**Figs. 2.2b and 2.3b**). Note that this bias does not affect the comparison of *R* among different branch pairs in **Figs. 2.2b and 2.3b**, because the bias applies to all branch pairs

similarly.

Given the exclusion of impacts from a potential genome-wide change in amino acid content (**Figs. 2.2b and 2.3b**), we believe that epistasis is the best explanation of why the equilibrium amino acid frequencies at a site change during evolution. Because of epistasis, what amino acids are acceptable at a site depends on what amino acids are present at its interacting sites. Thus, amino acid replacements in evolution at these interacting sites alter the equilibrium amino acid frequencies at the focal site. Our results are consistent with the covarion model of protein evolution (Fitch and Markowitz 1970) and many studies that reveal or suggest the prevalence of epistasis in protein evolution (Breen, et al. 2012; Harms and Thornton 2013; Parera and Martinez 2014; Xu and Zhang 2014; Zhang and Rosenberg 2002). Notably, Breen et al. (2012) reported that the observed ratio between the nonsynonymous (d_N) and synonymous (d_S) substitution rates for a gene is substantially lower than predicted based on the mean number of observed amino acids per site in a large phylogeny across all sites of the protein. They explained this phenomenon by a difference in the amino acids acceptable in different parts of the phylogeny as a result of epistasis. McCandlish et al. (2013) pointed out that not all amino acids at a site are equally fit and a nearly neutral model can explain Breen et al.'s observation without invoking epistasis, because, if most amino acids are acceptable but suboptimal, d_N/d_S would be lower than predicted from the number of acceptable amino acids. However, the nearly neutral hypothesis cannot explain the negative correlation between R and the genetic distance between two lineages. In other words, our observation provides additional evidence for the prevalence of

epistasis in protein evolution. It is worth noting that changes in the equilibrium amino acid frequencies at a site could also be caused by differential adaptations of different species to their respective environments. However, because the genetic distance between two species is not expected to correlate with their environmental difference at the phylogenetic scale examined in the present study, the adaption hypothesis cannot explain why *R* declines with the genetic distance.

That the equilibrium amino acid frequencies at a site change in evolution makes it difficult to quantify these frequencies, rendering the expected amount of molecular convergence hard to estimate. Ideally, the equilibrium frequencies at a site should be estimated from an alignment of many sequences such that all acceptable amino acids are observed multiple times. But the changing equilibrium frequencies also require that only closely related sequences be used in the estimation. In the study of molecular convergence between two closely related lineages, the use of a few closely related sequences in the estimation of equilibrium amino acid frequencies may upward bias the neutral expectation of the number of molecular convergences under neutrality, which will lead to a more conservative test of adaptive convergence. To guard against false positives, we advocate this strategy as opposed to the use of many distantly related sequences in estimating equilibrium amino acid frequencies, until the advent of a better test. The study of molecular convergence between two distantly related lineages is more complex due to potential changes in equilibrium frequencies. Using average equilibrium frequencies from multiple distantly related sequences will already lead to an overestimation of the neutral

expectation of the number of molecular convergences. Until further research for a better strategy, we suggest that this strategy be used so that the test of adaptive convergence will also be conservative. Although molecular convergence is the exclusive subject of this study, the finding of among-site and among-clade heterogeneities in protein sequence evolution has implications for other evolutionary analyses of protein sequences (Blanquart and Lartillot 2008; Jayaswal, et al. 2014). Further investigations of this subject are thus highly recommended.

2.5 MATERIALS AND METHODS

Protein sequence data

The protein sequence alignments of 12 species in the genus *Drosophila* were downloaded from the FlyBase FTP (St Pierre, et al. 2014), whereas those of 17 species in the class Mammalia were downloaded from OrthoMaM (Douzery, et al. 2014). Protein isoforms corresponding to the same gene and translated from alternatively spliced transcripts were identified and only one was randomly chosen and retained in the *Drosophila* dataset. No such problem existed for the mammalian data. In each alignment, any site with gap or ambiguous amino acid in any taxon was removed. Alignments with only one remaining site after the removals were excluded from further analysis. Amino acid frequencies at each site and the average amino acid frequencies across all sites were computed.

Parameter estimation

Each protein alignment was analyzed using the codeml program in PAML v4.7 (Yang 2007). The Empirical+F model coupled with the JTT matrix implemented in the software was used. A discrete gamma model with eight rate categories was used to account for among-site rate variation. For the *Drosophila* data, we used the tree topology (**Fig. 2.2a**) provided in a previous study (*Drosophila* 12 Genomes Consortium, et al. 2007). For the mammalian data, the 17 taxa were chosen such that ambiguous nodes in Romiguier, et al. (2013) could be avoided; the unambiguous phylogeny resulted (**Fig. 2.3a**) was then used. Potential discordances between gene trees and species trees were ignored due to low probabilities ($P < 3\times10^{-5}$ based on parameters pertaining to the species used) (Nei and Kumar 2000). From the PAML output, all branch lengths of the tree, relative substitution rate of each site, and inferred ancestral sequences were obtained. In addition, we concatenated all alignments and analyzed it by codeml with the same setting. The output branch lengths were used in the trees of **Figs. 2.2a and 2.3a** and were used to calculate the genetic distances presented in **Figs. 2.2b and 2.3b**. The genetic distance between a pair of branches is the sum of the lengths of all branches connecting the two younger ends of the two branches.

All 1100 alignments with over 500 residues were chosen from the *Drosophila* dataset and were subject to analysis by PhyloBayes 3.3f (Lartillot, et al. 2009). The JTT-CAT model with the discrete gamma distribution of among-site rate variation (with eight rate categories) was used. The tree topology was fixed as in the PAML analysis. The MCMC process was set to run for 6000 steps. After 1000 burn-in steps, the remaining steps were sampled once every five steps

to estimate parameters, following a recent analysis (Parker et al. 2013) of datasets comparable in size to ours. Results were obtained for 1081 alignments, because PhyloBayes was not able to analyze the other 19 alignments. Branch lengths, site-specific substitution rates, class-specific amino acid frequencies, and the class affiliations of all sites were obtained and used in downstream analysis.

Expected number of molecular convergences

Let us use **Fig. 2.1** as an example to explain how we computed the expected number of molecular convergences. For a site, let the probabilities for the 20 amino acids to occupy node *i* be $P(X_i)$, a vector of length 20. We use $I^{(j)}$ to denote a vector with the *j*th element equal to 1 and all other 19 elements equal to 0. The inferred most likely amino acid for the common ancestor (X_0) at the site concerned was extracted from the PAML output. If the most likely amino acid is *k*, $P(X_0) = I^{(k)}$. The equilibrium amino acid frequency vector π for a site was estimated from the observed amino acid frequencies among all taxa at the site for JTT-f_{site} and across all sites of the protein for JTT-f_{gene}, respectively. The substitution matrix was then derived from the JTT matrix M_0 provided in PAML. Given the original equilibrium frequency vector π_0 determined by the JTT matrix, the new substitution matrix was derived as $M = \left(M_{ij}\right) = (M_{0ij} \cdot \pi_j/\pi_{0j})$. We then calculated the amino acids at node 1 and node 2 by $P(X_1) = P(X_0) \cdot M^{rb_1}$ and $P(X_2) = P(X_0) \cdot M^{rb_2}$, respectively, where b_1 and b_2 are branch lengths measured by the expected numbers of substitutions per site for the protein concerned for the relevant sets of

branches, respectively (**Fig. 2.1**), and *r* is the substitution rate of the site considered, relative to the average for the protein. Conditional on the amino acid appearing at node 1, $P(X_3|X_1 = j) = I^{(j)} \cdot M^{rb_3}$. Similarly, $P(X_4|X_2 = j) = I^{(j)} \cdot M^{rb_4}$. Thus, the joint probability of having amino acids A, B, C, and D at nodes 1 to 4 can be calculated as $P(A, B, C, D) = P(X_1 = A) \cdot P(X_3 = C|X_1 = A) \cdot P(X_2 = B) \cdot P(X_4 = D|X_2 = B)$. For a given site, the probability of occurrence of convergent substitutions in the thick branches equals $P_{convergent} = \sum_{A\neq C,B\neq D,C=D,A\neq B} P(A, B, C, D)$, whereas the probability of occurrence of parallel substitutions equals $P_{parallel} = \sum_{A\neq C,B\neq D,C=D,A=B} P(A, B, C, D)$. The total probability of molecular convergence at the site is $P = P_{convergent} + P_{parallel}$. The expected number of molecular convergence for all proteins is the sum of *P* over all sites of all alignments.

Mantel tests

Mantel tests and partial Mantel tests were conducted using the R package "ncf". In the matrix containing pairwise *R* values, entries were set as "NA" if *R* values do not exist. The partial Mantel test used method 1 of permutation, which permutes the entire matrix of *R* values (Legendre 2000).

Simulation of sequence evolution when the equilibrium amino acid frequencies change

Simulation of sequence evolution followed the tree in **Fig. 2.4a**. For a site, $P(X_0)$ was set to be the equilibrium amino acid frequencies specified in the JTT model (π_0), and the initial

state was chosen randomly according to π_0. The equilibrium frequencies change as in a random walk. At each step of frequency changes, two entries in the frequency vector were randomly chosen, with one subtracted by 0.01 and the other added by 0.01. When a frequency is 0 (or 1), it can only increase (or decrease). The frequency vector changes for five steps before each step of sequence evolution. The changed frequency vector is then used to derive a new JTT-f substitution matrix as described above. The site with the initial probability vector $I^{(j)}$ then evolves by a Markov process for one step to $I^{(j)} \cdot M$, corresponding to 0.01 substitutions per site, or 1 PAM. The new state k is chosen multinomially according to the evolved probability vector, and the simulation continues. The rate of change in amino acid frequencies assumed in the simulation approximates the observed values in the fly data. The branch lengths (b_1 and b_2), in the unit of 1 PAM, were set as shown in **Fig. 2.4a**. We kept b_2 constant as 10 PAM and varied b_1 between 10 and 30 PAM. The observed number of molecular convergences was counted in 500,000 simulations for each b_1 value. For each parameter set, the expected number of molecular convergences was calculated as mentioned above, with the average of the equilibrium amino acid frequency vectors at nodes 2 and 4 used as equilibrium frequencies.

Amino acid compositions in 16 proteins with huge numbers of sequences

We obtained the 16 protein alignments from a previous study (Breen et al. 2012). For each protein, two taxonomic units representing two mutually exclusive monophyletic clades (e.g. ray-finned fishes vs. tetrapods; arthropods vs. chordates) were chosen. For each site in the

protein alignment, the presences/absences (1 for presence and 0 for absence) of the 20 amino acids were recorded as a vector of length 20 for each of the two clades. Sites with gaps in all species of a clade were excluded. The compositional distance between the two clades was measured by the Hamming distance between the two presence vectors of the site. To test if the observed compositional distance is significantly larger than the random expectation, for each protein, we mixed the sequences from the two clades, randomly divided them into two groups with the actual sizes of the two original clades, and calculated the Hamming distance. This process was repeated 1000 times to obtain the null distribution of the Hamming distance for each site. A site-wise *P*-value was computed by the proportion of times in which the randomized distance from the null distribution equals or exceeds the observed distance for the site. *Q*-value was then derived for each site within a protein using the Benjamini–Hochberg method (Benjamini and Hochberg 1995). Note that we compared the observed amino acids between two clades rather than the frequencies of observed amino acids, because of the possibility that the latter differ from the equilibrium frequencies. Comparing only the observed amino acids made our results more conservative.

ACKNOWLEDGEMENTS

We thank Wei-Chin Ho, Bryan Moyers, Jinrui Xu, and Jian-Rong Yang for valuable comments. This work was supported in part by U.S. National Institutes of Health research grant R01GM103232 to J.Z.

REFERENCES

Bazykin GA, Kondrashov FA, Brudno M, Poliakov A, Dubchak I, Kondrashov AS. 2007. Extensive parallelism in protein evolution. Biol Direct 2:20.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate - a practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol 57:289-300.

Blanquart S, Lartillot N. 2008. A site- and time-heterogeneous model of amino acid replacement. Mol Biol Evol 25:842-858.

Breen MS, Kemena C, Vlasov PK, Notredame C, Kondrashov FA. 2012. Epistasis as the primary factor in molecular evolution. Nature 490:535-538.

Castoe TA, de Koning AP, Kim HM, Gu W, Noonan BP, Naylor G, Jiang ZJ, Parkinson CL, Pollock DD. 2009. Evidence for an ancient adaptive episode of convergent molecular evolution. Proc Natl Acad Sci U S A 106:8986-8991.

Christin P-A, Weinreich DM, Besnard G. 2010. Causes and evolutionary significance of genetic convergence. Trends Genet 26:400-405.

Christin PA, Salamin N, Muasya AM, Roalson EH, Russier F, Besnard G. 2008. Evolutionary switch and genetic convergence on rbcL following the evolution of C4 photosynthesis. Mol Biol Evol 25:2361-2368.

Davies KT, Cotton JA, Kirwan JD, Teeling EC, Rossiter SJ. 2012. Parallel signatures of sequence evolution among hearing genes in echolocating mammals: an emerging model of genetic convergence. Heredity 108:480-489.

Douzery EJP, Scornavacca C, Romiguier J, Belkhir K, Galtier N, Delsuc F, Ranwez V. 2014. OrthoMaM v8: a database of orthologous exons and coding sequences for comparative genomics in mammals. Mol Biol Evol 31:1923-1928.

Drosophila 12 Genomes Consortium, Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, Markow TA, Kaufman TC, Kellis M, Gelbart W, et al. 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. Nature 450:203-218.

Feldman CR, Brodie ED, Jr., Brodie ED, 3rd, Pfrender ME. 2012. Constraint shapes convergence in tetrodotoxin-resistant sodium channels of snakes. Proc Natl Acad Sci U S A 109:4556-4561.

Fitch WM, Markowitz E. 1970. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. Biochem Genet 4:579-593.

Goldstein RA, Pollard ST, Shah SD, Pollock DD. 2015. Non-adaptive amino acid convergence rates decrease over time. Mol Biol Evol, 32:1373-1381.

Harms MJ, Thornton JW. 2013. Evolutionary biochemistry: revealing the historical and physical causes of protein properties. Nat Rev Genet 14:559-571.

Jayaswal V, Wong TK, Robinson J, Poladian L, Jermiin LS. 2014. Mixture models of nucleotide sequence evolution that account for heterogeneity in the substitution process across sites and across lineages. Syst Biol 63:726-742.

Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 8: 275-282.

Jost MC, Hillis DM, Lu Y, Kyle JW, Fozzard HA, Zakon HH. 2008. Toxin-resistant sodium channels: parallel adaptive evolution across a complete gene family. Mol Biol Evol 25:1016-1024.

Lartillot N, Lepage T, Blanquart S 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 25: 2286-2288.

Lartillot N, Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol 21:1095-1109.

Lartillot N, Philippe H. 2006. Computing Bayes factors using thermodynamic integration. Syst Biol 55:195-207.

Legendre P. 2000. Comparison of permutation methods for the partial correlation and partial Mantel tests. J Stat Comput Sim 67:37-73.

Li G, Wang J, Rossiter SJ, Jones G, Cotton JA, Zhang S. 2008. The hearing gene Prestin reunites echolocating bats. Proc Natl Acad Sci U S A 105:13959-13964.

Li Y, Liu Z, Shi P, Zhang J. 2010. The hearing gene Prestin unites echolocating bats and whales. Curr Biol 20:R55-56.

Liu Y, Cotton JA, Shen B, Han X, Rossiter SJ, Zhang S. 2010. Convergent sequence evolution between echolocating bats and dolphins. Curr Biol 20:R53-54.

Liu Y, Han N, Franchini LF, Xu H, Pisciottano F, Elgoyhen AB, Rajan KE, Zhang S. 2012. The voltage-gated potassium channel subfamily KQT member 4 (KCNQ4) displays parallel evolution in echolocating bats. Mol Biol Evol 29:1441-1450.

Liu Z, Li S, Wang W, Xu D, Murphy RW, Shi P. 2011. Parallel evolution of KCNQ4 in echolocating bats. PLOS One 6:e26618.

Liu Z, Qi FY, Zhou X, Ren HQ, Shi P. 2014. Parallel sites implicate functional convergence of the hearing gene prestin among echolocating mammals. Mol Biol Evol 31:2415-2424.

Mantel N. 1967. The detection of disease clustering and a generalized regression approach. Cancer Res 27:209-220.

McCandlish DM, Rajon E, Shah P, Ding Y, Plotkin JB. 2013. The role of epistasis in protein evolution. Nature 497: E1-2.

McGhee GR. 2011. Convergent Evolution: Limited Forms Most Beautiful. Cambridge, MA: Massachusetts Institute of Technology Press.

Nei M, Kumar S. 2000. Molecular Evolution and Phylogenetics. New York: Oxford University Press.

Parera M, Martinez MA. 2014. Strong epistatic interactions within a single protein. Mol Biol Evol 31:1546-1553.

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.

Rogozin IB, Thomson K, Csürös M, Carmel L, Koonin EV. 2008. Homoplasy in genome-wide analysis of rare amino acid replacements: the molecular-evolutionary basis for Vavilov's law of homologous series. Biol Direct 3:7.

Rokas A, Carroll SB. 2008. Frequent and widespread parallel evolution of protein sequences. Mol Biol Evol 25:1943-1953.

Romiguier J, Ranwez V, Delsuc F, Galtier N, Douzery EJP. 2013a. Less is more in mammalian phylogenomics: AT-rich genes minimize tree conflicts and unravel the root of placental mammals. Mol Biol Evol 30:2134-2144.

Shen YY, Liang L, Li GS, Murphy RW, Zhang YP. 2012. Parallel evolution of auditory genes for echolocation in bats and toothed whales. PLOS Genet 8:e1002788.

St Pierre SE, Ponting L, Stefancsik R, McQuilton P. 2014. FlyBase 102--advanced approaches to interrogating FlyBase. Nucleic Acids Res 42: D780-788.

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

Thomas GW, Hahn MW. 2015. Determining the null model for detecting adaptive convergence from genomic data: a case study using echolocating mammals. Mol Biol Evol, 32: 1232-1236.

Xu J, Zhang J. 2014. Why human disease-associated residues appear as the wild-type in other species: genome-scale structural evidence for the compensation hypothesis. Mol Biol Evol 31:1787-1792.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24:1586-1591.

Zhang J. 2006. Parallel adaptive origins of digestive RNases in Asian and African leaf monkeys. Nat Genet 38:819-823.

Zhang J. 2003. Parallel functional changes in the digestive RNases of ruminants and colobines by divergent amino acid substitutions. Mol Biol Evol 20:1310-1317.

Zhang J, Kumar S. 1997. Detection of convergent and parallel evolution at the amino acid sequence level. Mol Biol Evol 14:527-536.

Zhang J, Rosenberg HF. 2002. Complementary advantageous substitutions in the evolution of an antiviral RNase of higher primates. Proc Natl Acad Sci U S A 99:5486-5491.

Zhen Y, Aardema ML, Medina EM, Schumer M, Andolfatto P. 2012. Parallel molecular evolution in an herbivore community. Science 337:1634-1637.

Zou Z, Zhang J. 2015. No genome-wide protein sequence convergence for echolocation. Mol Biol Evol, 32:1237-1241.

Table 2.1. Observed numbers of convergent and parallel sites and the corresponding numbers expected under various neutral models of amino acid substitution. Results presented are for the two exterior branches leading to *D. yakuba* and *D. mojavensis*, respectively, in Fig. 2.2a.

Type of sites	Number of sites examined	Observed number of sites	Expected number of sites		R^a	P-value[b]
			Substitution model	Number of sites		
Convergent sites						
	2,028,428	292	JTT-f_{gene}	194.2	1.50	3.8E-11
	2,028,428	292	JTT-f_{site}	475.2	0.61	9.4E-20
	780,615	93	JTT-CAT	118.0	0.79	1.0E-3
Parallel sites						
	2,028,428	650	JTT-f_{gene}	388.6	1.67	3.2E-34
	2,028,428	650	JTT-f_{site}	2125.7	0.31	8.8E-309
	780,615	218	JTT-CAT	184.8	1.18	9.4E-3

[a]Ratio between the observed number and expected number.

[b]A statistical test is conducted under the assumption that the number of convergent (or parallel) sites follows a Poisson distribution with the mean equal to the expected number. When the observed number is smaller than the expected, the lower tail probability is given; when the observed number is larger than the expected, the upper tail probability is given.

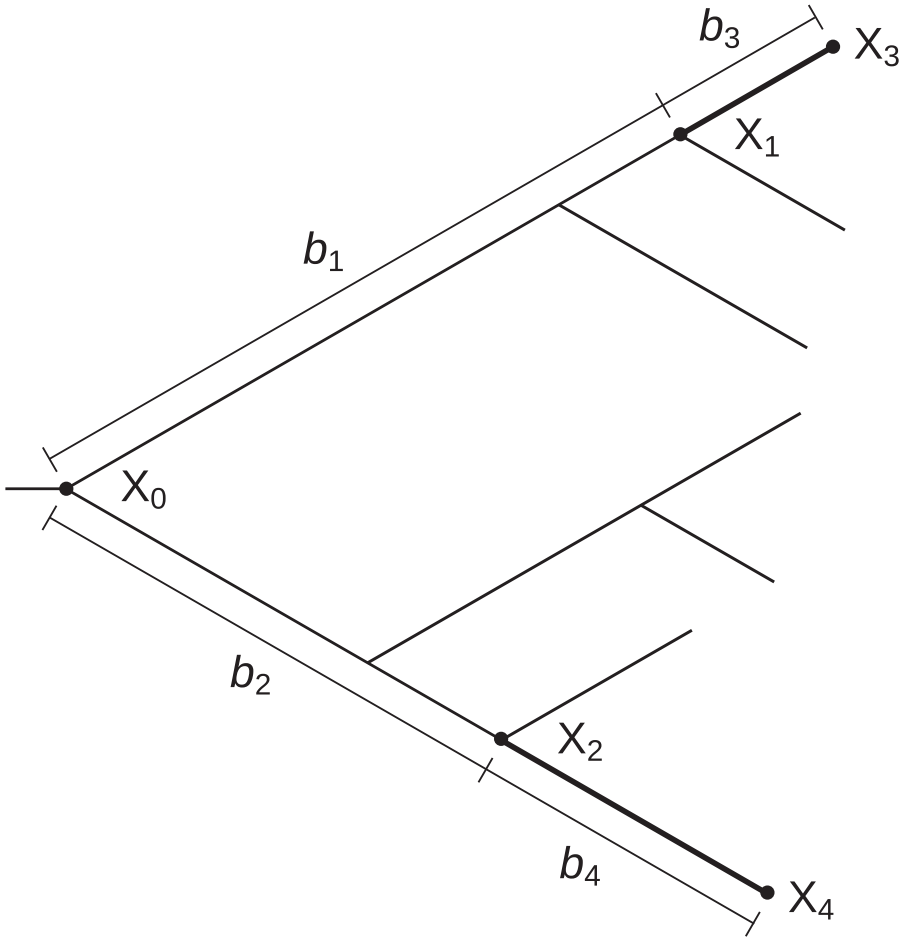


Figure 2.1 A tree illustrating the counting of the numbers of observed and expected molecular convergences between two thick branches. For a given position, the amino acids at nodes 0 to 4 are indicated by X_1 to X_4, respectively. The relevant branch lengths are indicated by the *b* values.

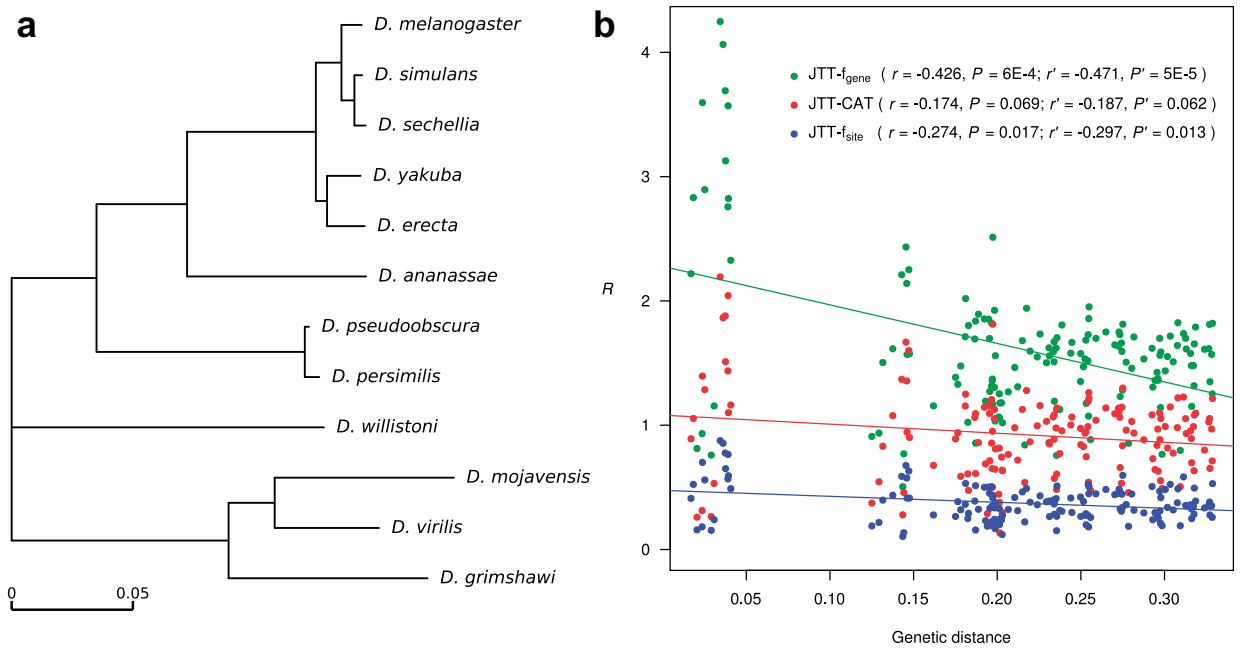


Figure. 2.2 The observed numbers of molecular convergences, relative to the expected numbers, in *Drosophila* proteins. (**a**) Phylogeny of the 12 *Drosophila* species. The topology follows *Drosophila* 12 Genomes Consortium et al. (2007), and the branch lengths are inferred using the concatenated sequences of all 5935 proteins, under the JTT-f model, where f refers to the overall amino acid frequencies from all 5935 proteins. (**b**) Negative correlation between the observed number of molecular convergences relative to the expected number (*R*) and the genetic distance between the two branches concerned. Each dot represents one branch pair, and different colors show the results under different substitution models. The *R* values under JTT-f_{gene} and JTT-f_{site} are based on all 5935 proteins, whereas those under JTT-CAT are based on a subset of 1081 proteins. Genetic distance is the number of amino acid substitutions per site between the two younger ends of the two branches considered. Lines show linear regressions. The *r* values are Pearson's correlation coefficients. Both *r* and *P*-values are from Mantel tests, and *r'* and *P'* are from partial Mantel tests controlling the between-node amino acid content difference.

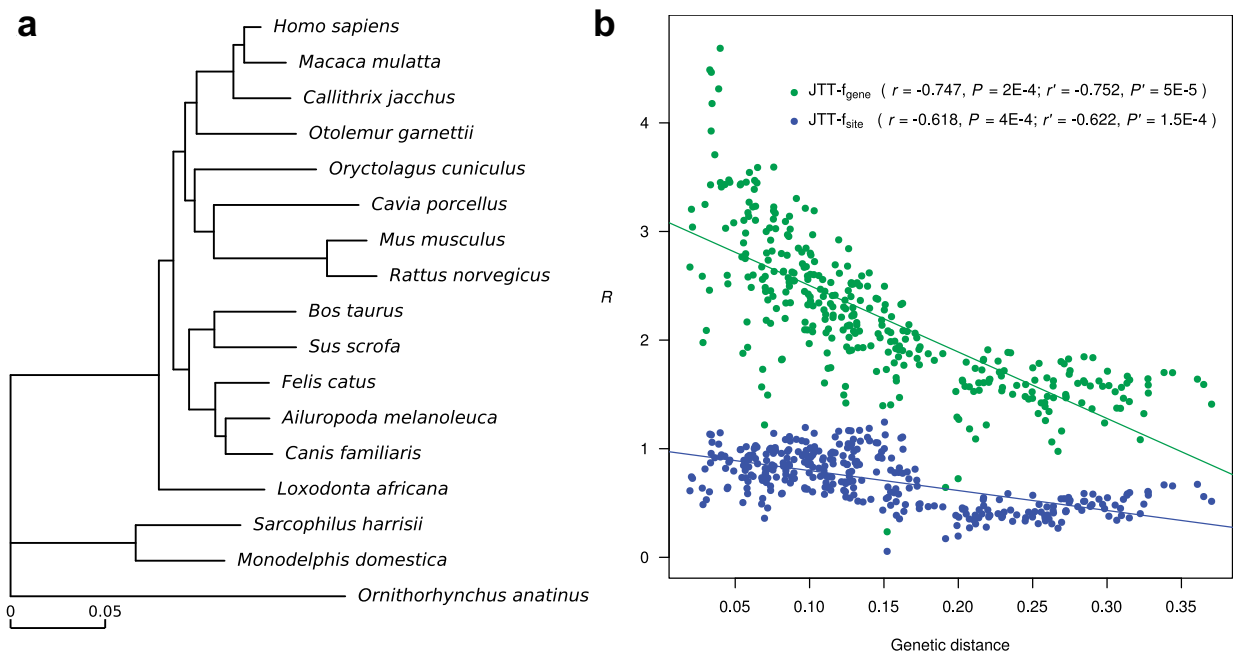


Figure. 2.3 The observed numbers of molecular convergences, relative to the expected numbers, in mammalian proteins. (**a**) Phylogeny of the 17 mammalian species. The topology follows Romiguier, et al. (2013), and the branch lengths are inferred using the concatenated sequences of all 2759 proteins, under the JTT-f model, where f refers to the overall amino acid frequencies from all 2759 proteins. (**b**) Negative correlation between the observed number of molecular convergences relative to the expected number (*R*) and the genetic distance between the two branches concerned. Each dot represents one branch pair, and different colors show the results under different substitution models. Genetic distance is the number of amino acid substitutions per site between the two younger ends of the two branches considered. Lines show linear regressions. The *r* values are Pearson's correlation coefficients. Both *r* and *P*-values are from Mantel tests, and *r'* and *P'* are from partial Mantel tests controlling the between-node amino acid content difference.

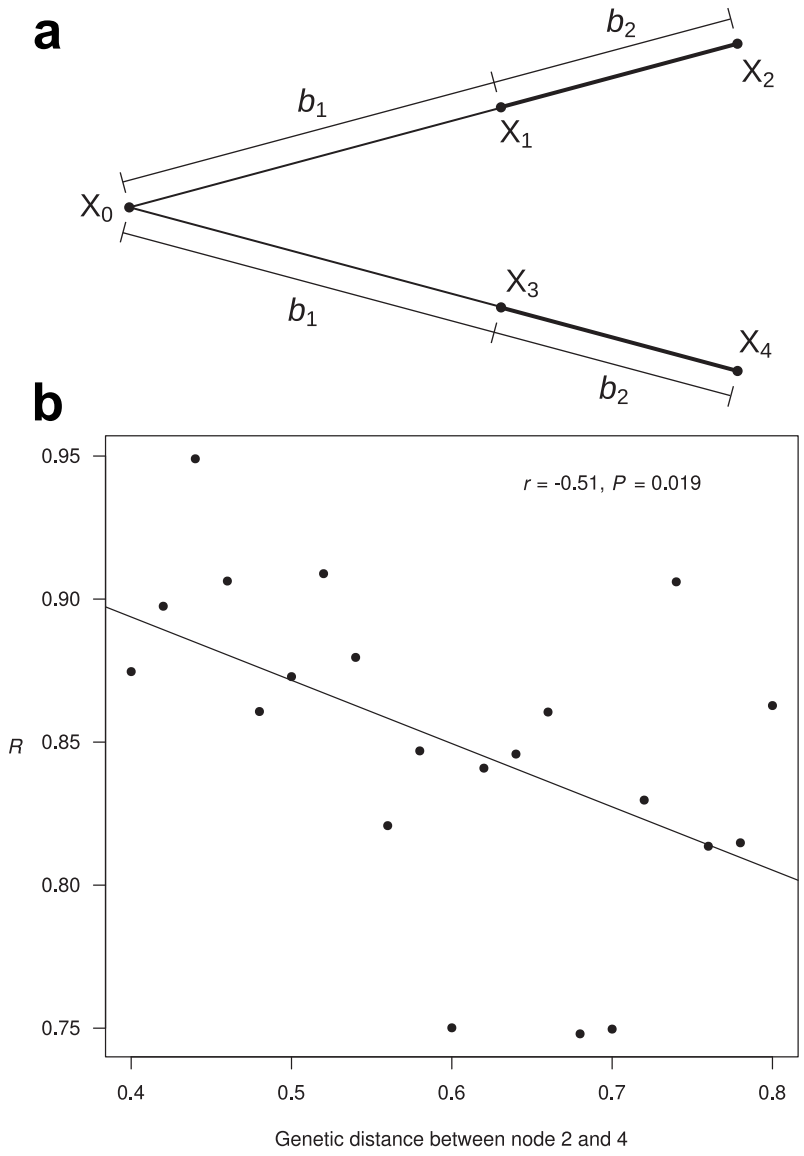
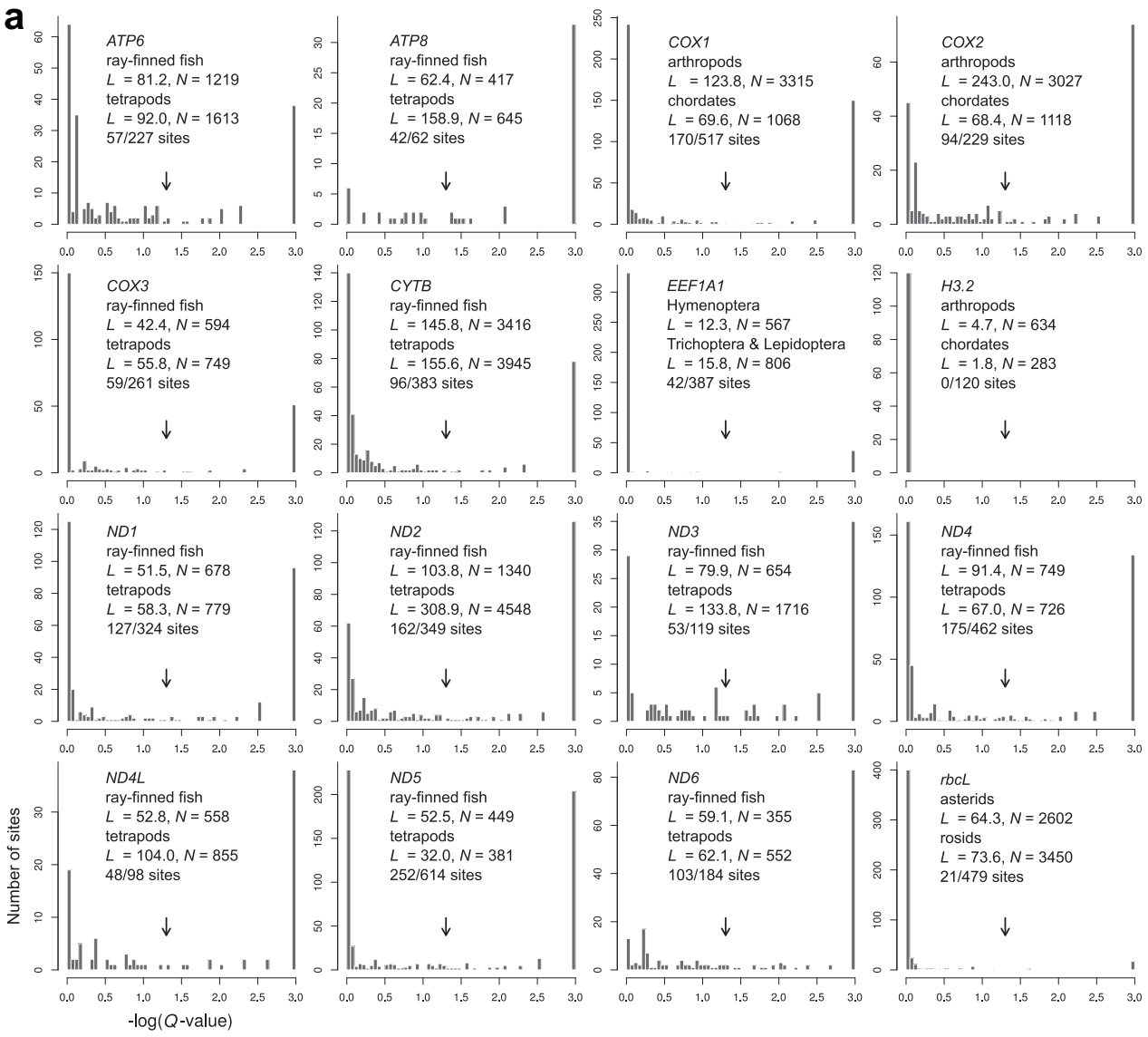


Figure. 2.4 Simulation of protein sequence evolution with changing equilibrium amino acid frequencies at each site. (**a**) The tree used in the simulation. Molecular convergence is examined for the two thick branches. Amino acids at nodes 0 to 4 are indicated by X_0 to X_4, respectively. The branch lengths are indicated by the *b* values. (**b**) The number of observed molecular convergences relative to the expected (*R*) decreases with the genetic distance between nodes 2 and 4. Each dot represents a simulation of 500,000 sites. For different dots, b_1 varies but b_2 is the same.


a
ATP6
ray-finned fish
L = 81.2, N = 1219
tetrapods
L = 92.0, N = 1613
57/227 sites
ATP8
ray-finned fish
L = 62.4, N = 417
tetrapods
L = 158.9, N = 645
42/62 sites
COX1
arthropods
L = 123.8, N = 3315
chordates
L = 69.6, N = 1068
170/517 sites
COX2
arthropods
L = 243.0, N = 3027
chordates
L = 68.4, N = 1118
94/229 sites
COX3
ray-finned fish
L = 42.4, N = 594
tetrapods
L = 55.8, N = 749
59/261 sites
CYTB
ray-finned fish
L = 145.8, N = 3416
tetrapods
L = 155.6, N = 3945
96/383 sites
EEF1A1
Hymenoptera
L = 12.3, N = 567
Trichoptera & Lepidoptera
L = 15.8, N = 806
42/387 sites
H3.2
arthropods
L = 4.7, N = 634
chordates
L = 1.8, N = 283
0/120 sites
ND1
ray-finned fish
L = 51.5, N = 678
tetrapods
L = 58.3, N = 779
127/324 sites
ND2
ray-finned fish
L = 103.8, N = 1340
tetrapods
L = 308.9, N = 4548
162/349 sites
ND3
ray-finned fish
L = 79.9, N = 654
tetrapods
L = 133.8, N = 1716
53/119 sites
ND4
ray-finned fish
L = 91.4, N = 749
tetrapods
L = 67.0, N = 726
175/462 sites
ND4L
ray-finned fish
L = 52.8, N = 558
tetrapods
L = 104.0, N = 855
48/98 sites
ND5
ray-finned fish
L = 52.5, N = 449
tetrapods
L = 32.0, N = 381
252/614 sites
ND6
ray-finned fish
L = 59.1, N = 355
tetrapods
L = 62.1, N = 552
103/184 sites
rbcL
asterids
L = 64.3, N = 2602
rosids
L = 73.6, N = 3450
21/479 sites
Number of sites
-log(Q-value)

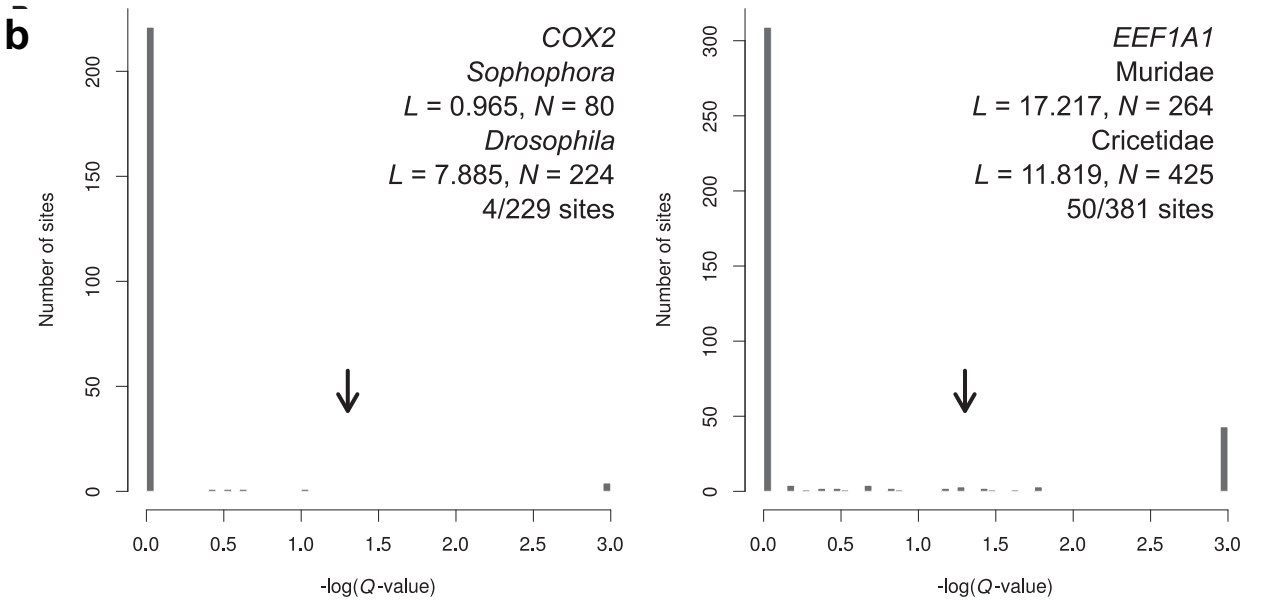

b
COX2
Sophophora
L = 0.965, N = 80
Drosophila
L = 7.885, N = 224
4/229 sites
EEF1A1
Muridae
L = 17.217, N = 264
Cricetidae
L = 11.819, N = 425
50/381 sites
Number of sites
-log(Q-value)

Figure. 2.5 Site-specific differences in acceptable amino acids between different clades of organisms. (**a**) Frequency distribution of $-\log_{10}(Q\text{-value})$ measuring the significance of difference in acceptable amino acids at a site between two distantly related large clades. Arrows correspond to Q-value = 0.05. For each plot, the name, total branch lengths (L), and number of species (N) of each of the two clades compared are indicated. The number of sites with Q-value < 0.05 is indicated, followed by the total number of sites examined. (**b**) Frequency distribution of $-\log_{10}(Q\text{-value})$ measuring the significance of difference in acceptable amino acids at a site between two closely related clades.

CHAPTER 3

Gene Tree Discordance does not Explain away the Temporal Decline of Convergence in Mammalian Protein Sequence Evolution[1]

3.1 ABSTRACT

Several authors reported lower frequencies of protein sequence convergence between more distantly related evolutionary lineages and attributed this trend to epistasis, which renders the acceptable amino acids at a site more different and convergence less likely in more divergent lineages. A recent primate study, however, suggested that this trend is at least partially and potentially entirely an artifact of gene tree discordance (GTD). Here we demonstrate in a genome-wide dataset from 17 mammals that the temporal trend remains (1) upon the control of the GTD level, (2) in genes whose genealogies are concordant with the species tree, and (3) for convergent changes, which are extremely unlikely to be caused by GTD. Similar results are observed in a comparable dataset of 12 fruit flies in some but not all of these tests. We conclude that, at least in some cases, the temporal decline of convergence is genuine, reflecting an impact of epistasis on protein evolution.

3.2 INTRODUCTION

Protein sequence convergence refers to independent amino acid substitutions at the same site in two or more evolutionary lineages that result in the same end state. It may signal molecular adaptation and therefore has long been of interest (Stewart, et al. 1987; Doolittle 1994; Zhang and Kumar 1997; Christin, et al. 2010; Storz 2016). Recent genomic analyses found that

[1] This chapter is published as: Zou Z, Zhang J. 2017. Gene tree discordance does not explain away the temporal decline of convergence in mammalian protein sequence evolution. Mol Biol Evol, 34: 1682-1688.

protein sequence convergence is widespread (Bazykin, et al. 2007; Rokas and Carroll 2008; Castoe, et al. 2009; Parker, et al. 2013; Foote, et al. 2015; Zou and Zhang 2015a), but the vast majority appears to have occurred by chance rather than by adaptive selection (Foote, et al. 2015; Thomas and Hahn 2015; Zou and Zhang 2015b, a). Interestingly, a number of authors reported that the frequency of sequence convergence between two lineages decreases as their genetic distance increases (Rogozin, et al. 2008; Povolotskaya and Kondrashov 2010; Naumenko, et al. 2012; Goldstein, et al. 2015; Zou and Zhang 2015a). It was proposed that epistasis, or interaction among amino acid residues within or between proteins, renders the selective constraint at the same site vary among species depending on the genetic background. Consequently, the probability of convergence between two lineages declines as they become more divergent from each other (Goldstein, et al. 2015; Zou and Zhang 2015a). Indeed, the same amino acid sites were found to be subject to different selective constraints in different lineages, and a computer simulation confirmed that epistasis can produce diminishing convergence over time (Zou and Zhang 2015a). Nevertheless, Mendes and colleagues recently proposed an alternative explanation of the temporal decline of convergence (Mendes, et al. 2016). Specifically, incomplete lineage sorting (ILS) and introgression can cause a gene genealogy to differ from the underlying species tree, a phenomenon called gene tree discordance (GTD). GTD creates artificial signals of convergence when sequence changes are inferred using the species tree. Because the probability of ILS and introgression declines with species divergence, the amount of artificial convergence created by GTD is expected to reduce as the two lineages compared become more distant from each other. Indeed, Mendes et al. found several lines of evidence supporting their hypothesis, including the disappearance of the temporal decline of convergence in a 12-primate dataset of 5264 genes when the influence of GTD is excluded

(Mendes, et al. 2016). It is clear that GTD cannot explain the temporal decline of convergence observed in mitochondrial genes (Goldstein, et al. 2015) due to the unique features of mitochondrial inheritance (Mendes et al. 2016). What is unclear, however, is whether GTD is fully responsible for the temporal declines of convergence in other nuclear gene datasets, because the relative contributions of GTD and epistasis to the temporal decline of convergence likely depend on the level of species divergence, which varies among datasets. It is important to clarify the above question, because if the temporal trend is always fully explainable by GTD in nuclear genes, there would be no genuine diminishing convergence over time for these genes and the role of epistasis in protein evolution might be substantially smaller than is currently thought. We therefore reanalyzed the two nuclear gene datasets (17 mammals and 12 fruit flies, respectively) where we previously discovered the temporal decline of convergence (Zou and Zhang 2015a).

3.3 RESULTS

Convergence measures

Hereinafter, independent amino acid substitutions at the same site that have the same ancestral state and the same end state are referred to as parallel changes while those with different ancestral states are referred to as convergent changes (Zhang and Kumar 1997). These two categories are collectively referred to as convergence. The distinction between parallel and convergent changes is important, because GTD is expected to create artificial parallel changes but not artificial convergent changes (Mendes, et al. 2016). The reason for the latter notion is that for ILS to create artificial convergent changes at a site, the site must be polymorphic with at least three distinct high-frequency alleles, which is extremely unlikely. Similarly, for

introgression to create artificial convergent changes at a site, the site must experience at least two different amino acid substitutions within a time that is sufficiently short to allow introgression, which is improbable.

Mendes et al. used the ratio between the observed numbers of convergences and divergences (*C*/*D*) as a measure of convergence level between two lineages. Note that divergences can be separated into two types: those starting from the same ancestral states as in parallel changes and those starting from different ancestral states as in convergent changes. It is known that, when *C* is the total number of parallel and convergent changes and *D* is the total number of the two types of divergence events, *C/D* decreases with the divergence time between the two lineages concerned even when neither epistasis nor GTD exists, because the probability of convergent changes relative to that of parallel changes rises as the genetic distance between the two lineages increases (Goldstein, et al. 2015). Mendes et al. suggested that *C*/*D* no longer declines with the divergence time when only parallel or convergent (but not both) changes are considered in *C* and only the corresponding type of divergence events is considered in *D*; these two *C*/*D* ratios are respectively referred to as $(C/D)_s$ and $(C/D)_d$, where the subscript "s" stands for the same ancestral states and "d" stands for different ancestral states. Our computer simulation in the absence of epistasis and GTD confirmed that *C*/*D*, but not $(C/D)_s$ or $(C/D)_d$, decreases with time (**Fig. A.2.1**).

In addition to $(C/D)_s$ and $(C/D)_d$, we used the ratio (*R*) between the observed and expected numbers of convergences to measure the convergence level, because *R* has a clear biological meaning and, in the absence of epistasis and GTD, is not expected to correlate with the genetic distance between lineages, as was previously demonstrated by simulation (Zou and Zhang 2015a). An amino acid substitution model is needed in computing *R*, and we used two

models employed in the original study: JTT-f_{site} and JTT-f_{gene} (Zou and Zhang 2015a). Both models assume the JTT substitution matrix (Jones, et al. 1992) except that the former uses the observed amino acid frequencies at a site as its equilibrium amino acid frequencies whereas the latter uses the observed amino acid frequencies of an entire protein as the equilibrium frequencies at each site of the protein. To examine the robustness of our results, we used two different distances between evolutionary lineages: (1) the total length of branches linking the descendant nodes of the two branches compared (Zou and Zhang 2015a) and (2) the total length of branches linking the ancestral nodes of the two branches compared (Mendes, et al. 2016). They are referred to as d_1 and d_2, respectively. Using d_1 and d_2 yielded qualitatively similar results in most cases (**Table 3.1**). We therefore describe only the results with d_1 in the main text unless otherwise mentioned.

Does GTD fully explain the temporal decline of convergence in mammals: Test I

A straightforward statistical test of the null hypothesis that GTD fully explains the temporal decline of convergence is to conduct a partial correlation between genetic distance and convergence level after controlling the GTD level. The partial correlation should be zero under the null hypothesis. We first inferred the maximum likelihood gene tree for each protein. For each independent branch pair in the species tree, we sampled four species whose tree includes the two focal branches and their respective sister branches (**Fig. A.2.2**), and estimated the GTD level for the focal branch pair by the proportion of genes whose gene trees differ from the species tree of these four species (see MATERIALS AND METHODS).

We started with the mammalian data, composed of 2759 protein sequence alignments of 14 placentals, two marsupials, and one monotreme (Zou and Zhang 2015a). Of 342 branch pairs

for which convergence can be measured, we found 208 independent branch pairs for which the GTD level can be estimated. For these branch pairs, we found R to be negatively correlated with d_1 even after the control of the GTD level (r = -0.51, P = 0.03 under JTT-f_{site}; r = -0.64, P = 0.001 under JTT-f_{gene}; **Table 3.1**), suggesting that GTD does not explain away the diminishing convergence over time.

Because GTD could result in apparent parallel changes, we further tested the null hypothesis by correlating $(C/D)_s$ with d_1 after the control of the GTD level. This partial correlation is significantly negative (r = -0.66, P = 0.0003; **Table 3.1**), consistent with the result based on R.

If GTD is the primary cause of the temporal decline of protein sequence convergence as Mendes et al. proposed, $(C/D)_s$ for synonymous sites is also expected to decline with d_1 (Mendes, et al. 2016). But this trend is not statistically significant (**Table 3.1**), suggesting at most a minor influence of GTD on convergence level in our data. Note that the significant negative correlation between $(C/D)_s$ for synonymous sites and d_1 after the control of GTD (**Table 3.1**) is due to the unexpected negative correlation between $(C/D)_s$ and GTD (e.g., r = -0.38, P = 0.009 upon the control of d_1), which does not conform to Mendes et al.'s hypothesis.

Together, test I demonstrates that, in the mammalian data, GTD is not the primary cause of the temporal decline in protein convergence. In the original study (Zou and Zhang 2015a), we rejected the hypothesis that potential genome-wide changes in amino acid frequencies cause the observed temporal decline of convergence. Hence, we no longer consider this possibility here. A potential source of error in our analysis arises from ancestral sequence inference. Analyzing 2759 protein sequence alignments generated by an evolutionary simulation with realistic parameters for the species tree, branch lengths, site-specific evolutionary rates, and JTT-f_{gene}

model with gene-specific amino acid frequencies, we found no significant correlation between $(C/D)_s$ and genetic distance, confirming that ancestral sequence inference and other steps in our analysis does not create artificial diminishing convergence over time.

Does GTD fully explain the temporal decline of convergence in mammals: Test II

The null hypothesis that GTD fully explains the temporal decline of convergence can be further tested by examining genes whose gene trees are concordant with the species tree, because the temporal pattern of convergence caused by GTD should disappear when only the concordant genes are analyzed. In the mammalian data, only 77 gene trees are concordant with the presumptive species tree. Nonetheless, the negative correlation between R and d_1 remains significant for these concordant genes ($r = -0.53$, $P = 5\times10^{-4}$ under JTT-f_{site}; $r = -0.60$, $P = 5\times10^{-5}$ under JTT-f_{gene}; **Fig. 3.1a**). Similarly, $(C/D)_s$ decreases with d_1 ($r = -0.54$, $P = 0.0005$; red dots in **Fig. 3.1b**). Note that removing all genes with discordant gene trees renders the above test conservative, because true convergence, which can also cause GTD, may have been removed too. Although the presumptive mammalian species tree may differ from the true species tree, the fact that we count convergence in all genes having the same gene tree ensures that gene tree variation does not affect our analysis. While recombination within genes may cause a seemingly concordant gene to harbor a discordant segment of DNA, there is no correlation between d_1 and $(C/D)_s$ for synonymous sites of concordant genes ($r = -0.02$, $P = 0.48$; grey dots in **Fig. 3.1b**), suggesting no impact of potential residual GTD in concordant genes.

Does GTD fully explain the temporal decline of convergence in mammals: Test III

Because it is very unlikely for GTD to create artificial convergent changes (Mendes, et al. 2016), a negative correlation between R and d_1 for convergent changes would not be explainable by GTD. Indeed, this correlation is significant (r = -0.45, P = 0.02 under JTT-f_{site}; r = -0.51, P = 0.003 under JTT-f_{gene}; **Fig. 3.1c**). The same trend is found between $(C/D)_d$ and d_1 (r = -0.31, P = 0.02; red dots in **Fig. 3.1d**). Similar to using only concordant genes, using only convergent changes renders our test conservative, because all parallel changes are excluded despite that only a fraction of them may be artifacts of GTD. As expected, no negative correlation is observed between d_1 and $(C/D)_d$ for synonymous sites (r = 0.27, P = 0.04; grey dots in **Fig. 3.1d**).

Taken together, the three tests support that the temporal decline of convergence in the mammalian data is not fully attributable to GTD. This finding, in conjunction with the previously published evidence for epistasis (Zou and Zhang 2015a), strongly implicates epistasis in causing diminishing convergence over time in mammals.

Does GTD fully explain the temporal decline of convergence in fruit flies?

We next analyzed the fruit fly data, composed of 5935 protein alignments from 12 *Drosophila* species (Zou and Zhang 2015a). For this dataset, GTD level can be evaluated for 84 independent branch pairs out of 150 ones for which convergence can be measured. The null hypothesis that GTD fully explains the temporal decline of convergence in fruit flies is refuted by some but not all of the three tests. First, upon control of the GTD level, no significant negative partial correlation was observed between genetic distance and R or $(C/D)_s$ (**Table 3.1**), failing to reject the null hypothesis. Second, for concordant genes, a significant negative correlation was detected between d_1 and R (r = -0.42, P = 0.04 under JTT-f_{site}; r = -0.56, P = 0.01 under JTT-f_{gene}; **Fig. 3.2a**) or $(C/D)_s$ (r = -0.42, P = 0.02; red dots in **Fig. 3.2b**) but not $(C/D)_s$ for

synonymous sites ($r = 0.10$, $P = 0.34$; grey dots in **Fig. 3.2b**), refuting the null hypothesis. Finally, when only convergent changes are considered, the null hypothesis is rejected when d_2 is used (**Table 3.1**), but is rejected in some but not all analyses when d_1 is used (**Fig. 3.2c**, **3.2d**; **Table 3.1**). These inconsistent results suggest that GTD is more important than epistasis in creating the pattern of diminishing convergence over time in the *Drosophila* data.

3.4 DISCUSSION

The difference in the relative contributions of GTD and epistasis to the temporal decline of convergence among the three datasets (primates in Mendes, et al. (2016); mammals and flies in this study) is at least in part caused by different frequencies of GTD in the three groups of organisms analyzed. For three species, the probability of GTD due to ILS is $P = \frac{2}{3}e^{-\frac{T}{2N}}$, where T is the number of generations between the two relevant speciation events and N is the effective population size (Hudson 1983; Pamilo and Nei 1988). Let us assume that for mammals (M), $N_M = 10^4$ and generation time $t_M = 5$ years, and for *Drosophila* fruit flies (D) $N_D = 10^6$ and $t_D = 0.1$ year (Charlesworth 2009). Given the same time interval of T' million years between relevant speciation events, $P_D / P_M = e^{\frac{T'}{2N_M t_M} - \frac{T'}{2N_D t_D}} = e^{5T'}$. Hence, the probability of GTD is expected to be higher in fruit flies than in mammals given equal speciation frequencies between the two groups. For example, when $T' = 0.5$ million years, the probability of GTD due to ILS is 5.5% in fruit flies but only 0.45% in mammals. Introgression occurs in both mammals and flies (Ballard 2000; Bachtrog, et al. 2006; Mallet, et al. 2016), although their rates are unclear. Consequently, the impact of GTD is expected to be higher in fruit flies than in mammals if ILS is an important contributor to GTD. The primate data have relatively short speciation intervals compared with the mammalian data and are thus expected to be influenced more by GTD. The impact of

epistasis should depend on sequence divergence; datasets with larger ranges of sequence divergence are expected to be more influenced by epistasis. This factor may render epistasis more influential in the mammalian data than in the primate data. In the mammalian data, depending on the distance (d_1 or d_2) and convergence (R under JTT-f_{site}, R under JTT-f_{gene}, or $(C/D)_s$) measures used, the partial correlation between d and convergence after the control for GTD (in the Mantel test) is on average 76% of the corresponding correlation without the control for GTD (**Table 3.1**), suggesting that the contribution of epistasis is at least as important as GTD.

In conclusion, we showed that, at least for the mammalian data analyzed, GTD cannot fully explain the temporal decline of convergence, which implicates the contribution of epistasis. The different results obtained from three datasets (primates, mammals, and fruit flies) demonstrate that the relative roles of GTD and epistasis in creating diminishing convergence over time depend on speciation intervals and sequence divergences and are thus data-dependent.

3.5 MATERIALS AND METHODS

In a species tree, let branch X connect an interior node X_0 and one of its two immediate descendants X_1 and let branch Y connect an interior node Y_0 and one of its two immediate descendants Y_1 (**Fig. A.2.2**). We estimated the GTD level for all independent branch pairs, where Y_0 is not on the path from X_0 to the tree root and X_0 is not on the path from Y_0 to the tree root. Let X_2 be the other immediate descendant of X_0 and let Y_2 be the other immediate descendant of Y_0. Let exterior nodes X_1', X_2', Y_1', and Y_2' be randomly picked descendants of X_1, X_2, Y_1, and Y_2, respectively. The four exterior nodes have a phylogenetic relationship of $((X_1', X_2'), (Y_1', Y_2'))$ in the species tree. For a gene, if the topology of X_1', X_2', Y_1', and Y_2' in

the gene tree is inconsistent with that in the species tree, this gene is defined as showing GTD for branch pair (X, Y). The overall GTD level for the branch pair (X, Y) is the proportion of genes that show GTD for (X, Y). Gene trees were inferred using RAxML v8.2.4 under the JTT-f_{gene} model with substitution rate variation following a gamma distribution (Stamatakis 2014). The species trees of the mammals and fruit flies analyzed here respectively follow those in Fig. 2A and Fig. 3A of Zou and Zhang (2015a). Mantel tests and partial Mantel tests were conducted using the R package "ncf". In all matrices used for these tests, entries that do not correspond to a branch pair with an available GTD level were set as "NA". The partial Mantel test used method 1 of permutation, which permutes the entire matrix of *R* or *C*/*D* values (Legendre 2000). Protein sequences were acquired from Zou and Zhang (2015a), who obtained them from OrthoMaM v8 (Douzery, et al. 2014) and Flybase (in October 2013). The corresponding coding DNA sequences were retrieved from OrthoMaM v9 and Flybase (in September 2016). The protein sequences and nucleotide sequences have consistent lengths after the removal of ambiguous sites as described in Zou and Zhang (2015a), and can be accessed from http://www.umich.edu/~zhanglab/download/Zou_201702/index.htm. The 2759 alignments of mammalian proteins have a median length of 315 amino acids, while the 5935 alignments of fly proteins have a median length of 289 amino acids.

ACKNOWLEDGEMENTS

We thank Matt Hahn for valuable comments. This work was supported in part by U.S. National Institutes of Health research grant R01GM089827 to J.Z.

REFERENCES

Bachtrog D, Thornton K, Clark A, Andolfatto P. 2006. Extensive introgression of mitochondrial DNA relative to nuclear genes in the *Drosophila yakuba* species group. Evolution 60:292-302.

Ballard JW. 2000. When one is not enough: introgression of mitochondrial DNA in *Drosophila*. Mol Biol Evol 17:1126-1130.

Bazykin GA, Kondrashov FA, Brudno M, Poliakov A, Dubchak I, Kondrashov AS. 2007. Extensive parallelism in protein evolution. Biol Direct 2:20.

Castoe TA, de Koning AP, Kim HM, Gu W, Noonan BP, Naylor G, Jiang ZJ, Parkinson CL, Pollock DD. 2009. Evidence for an ancient adaptive episode of convergent molecular evolution. Proc Natl Acad Sci U S A 106:8986-8991.

Charlesworth B. 2009. Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. Nat Rev Genet 10:195-205.

Christin PA, Weinreich DM, Besnard G. 2010. Causes and evolutionary significance of genetic convergence. Trends Genet 26:400-405.

Doolittle RF. 1994. Convergent evolution: the need to be explicit. Trends Biochem Sci 19:15-18.

Douzery EJ, Scornavacca C, Romiguier J, Belkhir K, Galtier N, Delsuc F, Ranwez V. 2014. OrthoMaM v8: a database of orthologous exons and coding sequences for comparative genomics in mammals. Mol Biol Evol 31:1923-1928.

Foote AD, Liu Y, Thomas GW, Vinar 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.

Goldstein RA, Pollard ST, Shah SD, Pollock DD. 2015. Nonadaptive amino acid convergence rates decrease over time. Mol Biol Evol 32:1373-1381.

Hudson RR. 1983. Testing the constant-rate neutral allele model with protein sequence data. Evolution 37:203-217.

Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 8:275-282.

Legendre P. 2000. Comparison of permutation methods for the partial correlation and partial Mantel tests. J Stat Comput Simul 67:37-73.

Mallet J, Besansky N, Hahn MW. 2016. How reticulated are species? Bioessays 38:140-149.

Mendes FK, Hahn Y, Hahn MW. 2016. Gene tree discordance can generate patterns of diminishing convergence over time. Mol Biol Evol 33:3299-3307.

Naumenko SA, Kondrashov AS, Bazykin GA. 2012. Fitness conferred by replaced amino acids declines with time. Biol Lett 8:825-828.

Pamilo P, Nei M. 1988. Relationships between gene trees and species trees. Mol Biol Evol 5:568-583.

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.

Povolotskaya IS, Kondrashov FA. 2010. Sequence space and the ongoing expansion of the protein universe. Nature 465:922-926.

Rogozin IB, Thomson K, Csuros M, Carmel L, Koonin EV. 2008. Homoplasy in genome-wide analysis of rare amino acid replacements: the molecular-evolutionary basis for Vavilov's law of homologous series. Biol Direct 3:7.

Rokas A, Carroll SB. 2008. Frequent and widespread parallel evolution of protein sequences. Mol. Biol. Evol. 25:1943-1953.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312-1313.

Stewart CB, Schilling JW, Wilson AC. 1987. Adaptive evolution in the stomach lysozymes of foregut fermenters. Nature 330:401-404.

Storz JF. 2016. Causes of molecular convergence and parallelism in protein evolution. Nat Rev Genet 17:239-250.

Thomas GW, Hahn MW. 2015. Determining the null model for detecting adaptive convergence from genomic data: a case study using echolocating mammals. Mol Biol Evol 32:1232-1236.

Zhang J, Kumar S. 1997. Detection of convergent and parallel evolution at the amino acid sequence level. Mol Biol Evol 14:527-536.

Zou Z, Zhang J. 2015a. Are convergent and parallel amino acid substitutions in protein evolution more prevalent than neutral expectations? Mol Biol Evol 32:2085-2096.

Zou Z, Zhang J. 2015b. No genome-wide protein sequence convergence for echolocation. Mol Biol Evol 32:1237-1241.

Table 3.1 Pearson's correlations between genetic distance and various convergence levels

	Mammals				Fruit flies			
	Amino acid			Synonymous	Amino acid			Synonymous
	R (JTT-f_{site})	R (JTT-f_{gene})	C/D	C/D	R (JTT-f_{site})	R (JTT-f_{gene})	C/D	C/D
Mantel test (all genes)								
d_1	-0.73** [a]	-0.79****	-0.74**** [b]	-0.19 [b]	-0.49*	-0.63**	-0.44* [b]	0.12 [b]
d_2	-0.88****	-0.74****	-0.57*** [b]	-0.0052 [b]	-0.75****	-0.72****	-0.41** [b]	0.19 [b]
Partial Mantel test (controlling GTD)								
d_1	-0.51*	-0.64**	-0.66*** [b]	-0.38* [b]	0.55	0.18	-0.015 [b]	0.049 [b]
d_2	-0.72***	-0.48***	-0.40* [b]	-0.21 [b]	-0.015	0.23	0.20 [b]	0.18 [b]
Mantel test (concordant genes)								
d_1	-0.53***	-0.60****	-0.54*** [b]	-0.015 [b]	-0.42*	-0.56*	-0.42* [b]	0.094 [b]
d_2	-0.68****	-0.55****	-0.38** [b]	0.12 [b]	-0.71****	-0.69****	-0.43** [b]	0.20 [b]
Mantel test (convergent changes)								
d_1	-0.45*	-0.51**	-0.31* [c]	0.27 [c]	-0.32	-0.44*	-0.21 [c]	0.46 [c]
d_2	-0.52***	-0.47***	-0.33** [c]	0.17 [c]	-0.47**	-0.50***	-0.31* [c]	0.43 [c]

[a] Significance is shown only when $r < 0$. *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$; ****, $P \leq 0.0001$.
[b] $(C/D)_s$.
[c] $(C/D)_d$.
d_1 = total length of branches linking the descendant nodes of the two branches compared.
d_2 = total length of branches linking the ancestral nodes of the two branches compared.

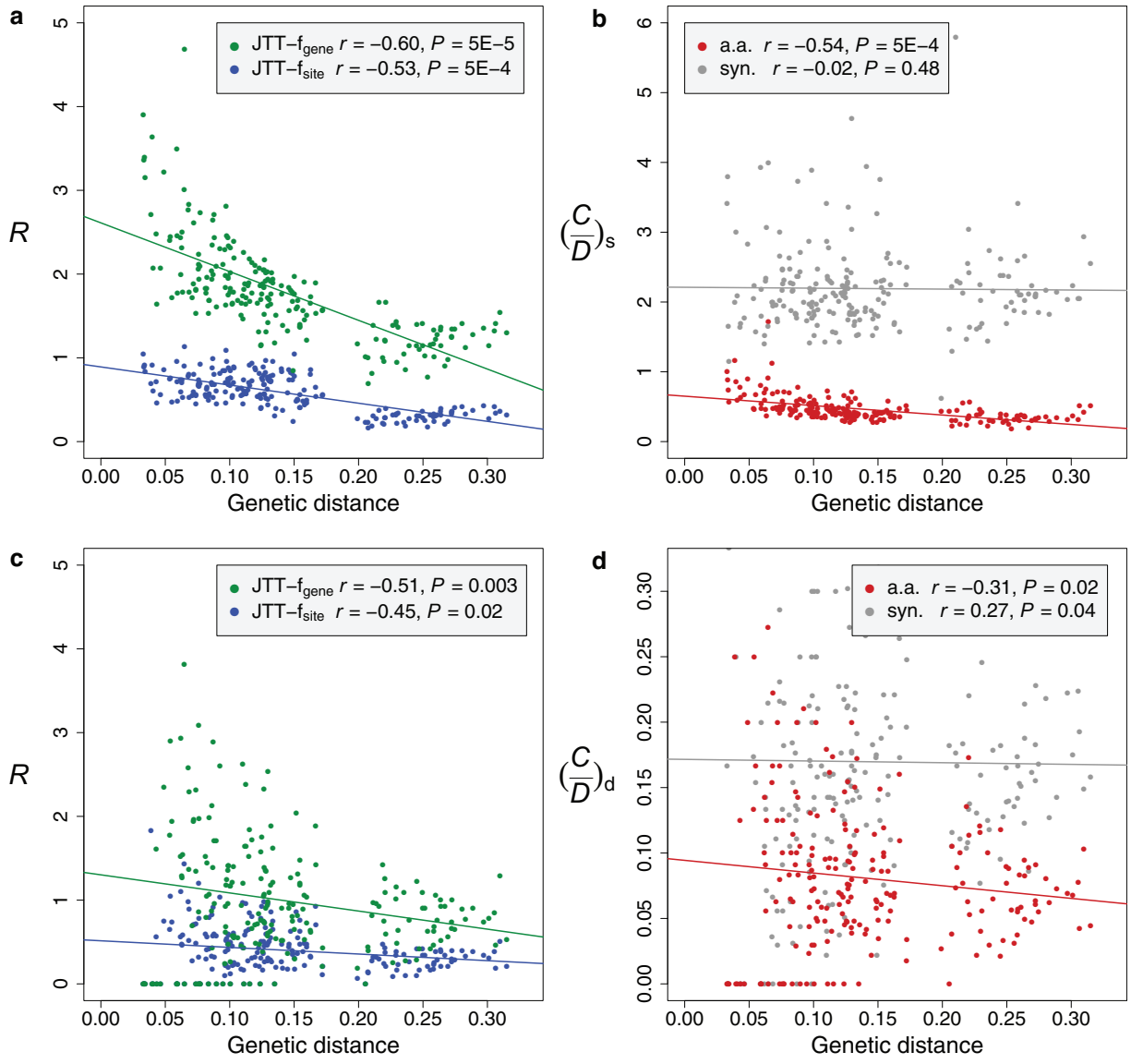


Figure 3.1 Correlation between convergence level and genetic distance in mammals. (**a**) Scatter plot of R against the genetic distance d_1 for genes having gene trees concordant with the presumptive species tree ("concordant genes"). (**b**) Scatter plot of $(C/D)_s$ against d_1 for concordant genes. (**c**) Scatter plot of R for convergent changes in all genes against d_1. (**d**) Scatter plot of $(C/D)_d$ for all genes against d_1. Each dot represents a branch pair and different colors show results under different substitution models or for different types of substitutions, as indicated in inset legends. d_1 is the number of amino acid substitutions per site between the descendant nodes of the two branches considered. The r values are Pearson's correlation coefficients. Both r and P-values are from Mantel tests. Colored lines show linear regressions from data points of the same color. a.a.: amino acid substitutions; syn.: synonymous nucleotide substitutions.

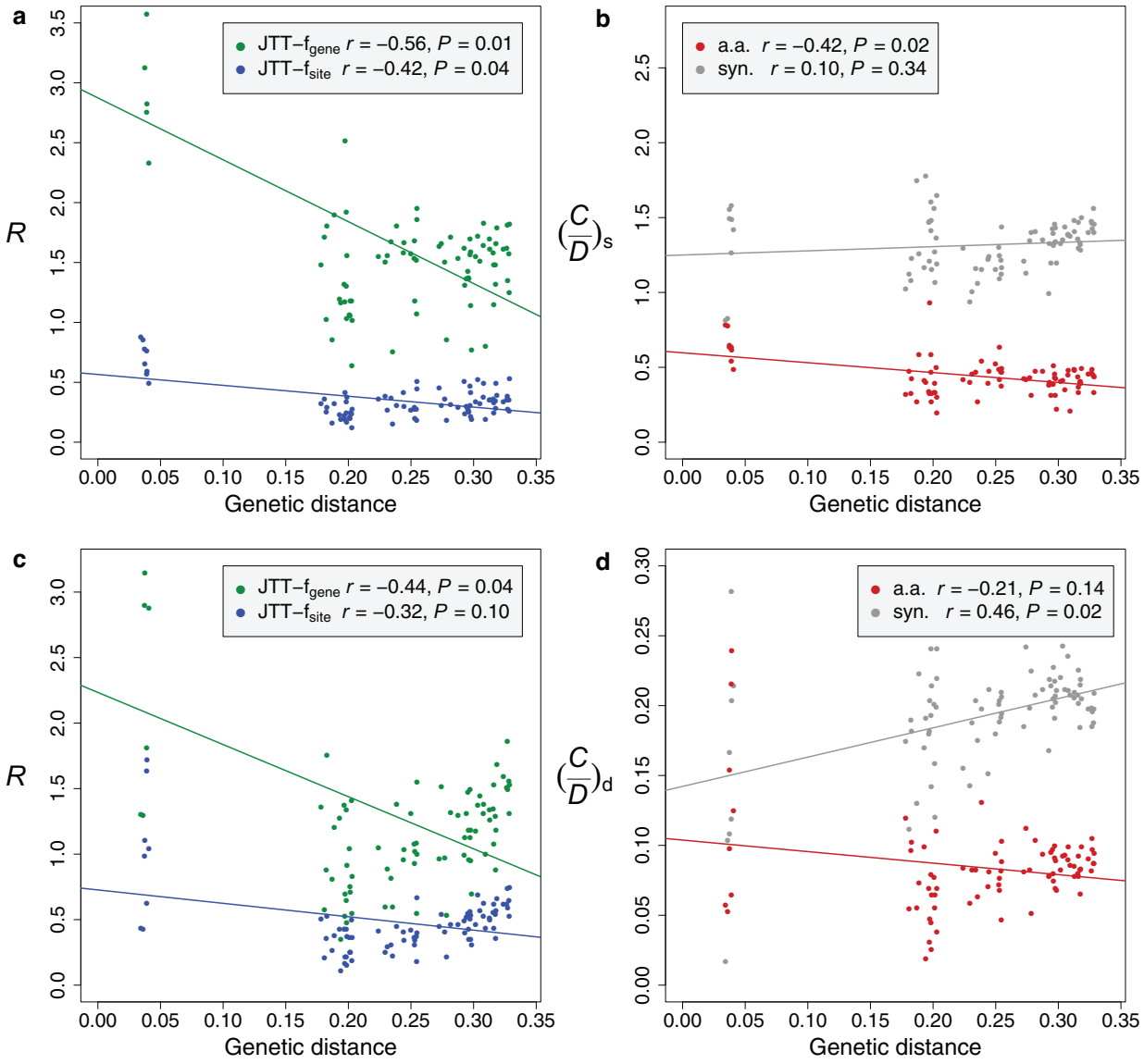


Figure 3.2 Correlation between convergence level and genetic distance in fruit flies. (**a**) Scatter plot of R against the genetic distance d_1 for genes having gene trees concordant with the presumptive species tree ("concordant genes"). (**b**) Scatter plot of $(C/D)_s$ against d_1 for concordant genes. (**c**) Scatter plot of R for convergent changes in all genes against d_1. (**d**) Scatter plot of $(C/D)_d$ for all genes against d_1. Each dot represents a branch pair and different colors show results under different substitution models or for different types of substitutions, as indicated in inset legends. d_1 is the number of amino acid substitutions per site between the descendant nodes of the two branches considered. The r values are Pearson's correlation coefficients. Both r and P-values are from Mantel tests. Colored lines show linear regressions from data points of the same color. a.a.: amino acid substitutions; syn.: synonymous nucleotide substitutions.

CHAPTER 4

No Genome-Wide Protein Sequence Convergence for Echolocation[1]

4.1 ABSTRACT

Toothed whales and two groups of bats independently acquired echolocation, the ability to locate and identify objects by reflected sound. Echolocation requires physiologically complex and coordinated vocal, auditory, and neural functions, but the molecular basis of the capacity for echolocation is not well understood. A recent study suggested that convergent amino acid substitutions widespread in the proteins of echolocators underlay the convergent origins of mammalian echolocation. Here we show that genomic signatures of molecular convergence between echolocating lineages are generally no stronger than those between echolocating and comparable non-echolocating lineages. The same is true for the group of 29 hearing-related proteins claimed to be enriched with molecular convergence. Reexamining the previous selection test reveals several flaws and invalidates the asserted genome-wide evidence for adaptive convergence. Together, these findings indicate that the reported genomic signatures of convergence largely reflect the background level of sequence convergence unrelated to the origins of echolocation.

4.2 INTRODUCTION

Echolocation originated independently in toothed whales and two groups of bats (**Fig. 4.1a**). Understanding the molecular basis of this complex evolutionary innovation is of substantial interest. Comparing 22 mammalian genome sequences, Parker et al. reported

[1] This chapter is published as: Zou Z, and Zhang J. 2015. No genome-wide convergence for echolocation. Mol Biol Evol, 32: 1237-1241.

hundreds of convergently evolving proteins among echolocators and suggested that genome-wide molecular convergence underlay the origins of echolocation and associated phenotypes (Parker et al. 2013). However, protein convergence could also occur by chance (Zhang and Kumar 1997; Castoe et al. 2009). Here we show that the reported genomic signatures of convergence largely reflect such chance events that are unrelated to the origins of echolocation.

4.3 RESULTS

Parker et al. assembled and aligned the orthologous sequences of 2326 proteins. For each protein, they estimated the log-likelihood differences per site (ΔL) between the known mammalian species tree (H0) and each of two alternative trees (H1 and H2). In H1, the two groups of echolocating bats are clustered, whereas in H2, the echolocating bats and the bottlenose dolphin, an echolocating whale, are grouped (**Fig. 4.1a**). A negative $\Delta L_{H0\text{-}H1}$ (or $\Delta L_{H0\text{-}H2}$) indicates that the evolution of the protein favors H1 (or H2) over H0, which Parker et al. regarded as a signature of molecular convergence of echolocators.

However, because the null distribution of $\Delta L_{H0\text{-}H1}$ is unknown, it is necessary to set a negative control against which $\Delta L_{H0\text{-}H1}$ is compared. By exchanging in H1 the phylogenetic positions of non-echolocating (orange) and echolocating (purple) bats belonging to Yinpterochiroptera, we created H1' (**Fig. 4.1a**), which does not cluster the two groups of echolocating bats but otherwise exhibits the same amount of phylogenetic distortion from H0 as does H1. Significantly more negative $\Delta L_{H0\text{-}H1}$ values compared with $\Delta L_{H0\text{-}H1'}$ values across the 2326 proteins would be consistent with Parker et al.'s claim of a genome-wide signature of protein convergence associated with bat echolocation. However, we found that the frequency distribution of $\Delta L_{H0\text{-}H1}$ is superimposed on that of $\Delta L_{H0\text{-}H1'}$ (**Fig. 4.1b**). Similarly, we created H2'

by exchanging the phylogenetic positions of cow and dolphin in H2 (**Fig. 4.1a**). Again, the frequency distribution of $\Delta L_{H0\text{-}H2}$ is superimposed on that of $\Delta L_{H0\text{-}H2'}$ (**Fig. 4.1c**), suggesting just as much convergence between bats and cow as was observed between bats and dolphin. Note that, in the species tree, the branch length measured by the number of amino acid substitutions per site across all proteins analyzed is greater for the exterior branch leading to cow than that leading to dolphin (see Fig. 1a of Parker et al. 2013). We predicted and verified by computer simulation that this branch length difference results in a slightly more positive $\Delta L_{H0\text{-}H2'}$ than $\Delta L_{H0\text{-}H2}$ on average, rendering our comparison between $\Delta L_{H0\text{-}H2'}$ and $\Delta L_{H0\text{-}H2}$ conservative. The branches leading to echolocating (purple) and non-echolocating (orange) bats of Yinpterochiroptera have similar lengths in the species tree (see Fig. 1a of Parker et al. 2013) and therefore the comparison between $\Delta L_{H0\text{-}H1'}$ and $\Delta L_{H0\text{-}H1}$ is fair.

Earlier work identified seven hearing-related proteins that underwent convergent evolution in echolocators (Li et al. 2008; Li et al. 2010; Liu et al. 2010; Liu et al. 2011; Davies et al. 2012; Liu et al. 2012; Shen et al. 2012). Parker et al. mentioned 22 additional proteins annotated as "hearing" or "deafness" in the data analyzed (Parker et al. 2013). We found that $\Delta L_{H0\text{-}H1}$ is smaller than $\Delta L_{H0\text{-}H1'}$ for 59% of the 29 hearing proteins, not significantly more than the random expectation of 50% ($P > 0.2$, one-tail binomial test). When the seven known convergent proteins are excluded, $\Delta L_{H0\text{-}H1}$ is smaller than $\Delta L_{H0\text{-}H1'}$ for only 45% of the 22 proteins ($P > 0.7$). Similarly, $\Delta L_{H0\text{-}H2}$ is smaller than $\Delta L_{H0\text{-}H2'}$ for 59% of the 29 proteins ($P > 0.2$) and 50% of the 22 proteins ($P > 0.5$). Qualitatively identical results were obtained by paired *t*-tests. Hence, as a group, hearing proteins show no significant enrichment of phylogenetic signals for echolocator-specific convergence.

Next, we inferred ancestral protein sequences for interior nodes in H0 and counted the number of sites with convergent amino acid substitutions along relevant sets of branches as a direct measure of protein convergence. In total, 2270 proteins each with available sequences from at least dolphin, cow, and the six bat species in **Fig. 4.1a** were analyzed. For example, under the hypothesis of a single origin of bat echolocation, protein convergence associated with echolocation should occur between branches I and II but not between I and III (**Fig. 4.1a**). However, we observed no more convergent sites in the former than the latter (**Table 4.1**). It was previously shown that, under no adaptive convergence, the number of convergent sites in a branch set is expected to be proportional to the number of divergent sites in the same set (Castoe et al. 2009), where divergent sites are those at which divergent amino acid substitutions have occurred along the branches of interest. We thus asked if the number of convergent sites in branches I and II significantly exceeds that in branches I and III, given their numbers of divergent sites. The answer is clearly no, and in fact the opposite is true (**Table 4.1**). Consistent observations were made in comparisons between branch sets (IV and II) and (IV and III) and between branch sets (V and II) and (V and III) (**Table 4.1**). A similar result was found in an independent analysis of 6400 genes from 10 mammals (Thomas and Hahn 2015), although it is unknown why there are more convergent sites between bats and cow than between bats and dolphin, given their respective numbers of divergent sites. In two comparisons, however, we observed more convergent sites in echolocators than the controls. First, in the comparison between the three-branch sets (II, IV, and V) and (III, IV, and V) under the hypothesis of dual origins of bat echolocation, the branch set representing echolocators has more convergent sites than the control set, although the difference is not significant (**Table 4.1**). Notably, of the 14 convergent sites among branches II, IV, and V, 12 fall in six of the seven known convergently

evolving hearing proteins (prestin is not included in this analysis due to missing data), indicating that at most a few proteins were subject convergent evolution in all three echolocation lineages. Second, the comparison between branch sets (IV and V) and (IV and VI) shows that the former set, representing echolocating bats, has significantly more convergent sites than the control set (**Table 4.1**). But after the removal of the six known convergently evolving hearing proteins, the two branch sets are no longer significantly different (**Table 4.1**). Together, these direct comparisons in the number of convergent sites between echolocating lineages and control lineages offer no evidence for genome-wide convergence in echolocating lineages beyond the background level.

Parker et al. assumed that a significant negative correlation between site-wise ΔL and ω (nonsynonymous/synonymous rate ratio) within a gene indicates adaptive convergence. Although adaptive convergence may lead to a negative correlation between ΔL and ω, there is no proof that neutral evolution cannot. Compared with low-ω sites, high-ω sites are more likely to experience convergence by chance (as well as divergence). Thus, one cannot exclude the possibility that a negative correlation results from neutral evolution, especially when ω is not significantly greater than 1. Furthermore, to estimate ω, Parker et al. used H1 or H2 instead of the species tree. Because the true evolutionary histories of all genes considered here are described by the species tree, the ω estimates based on the wrong trees are biologically meaningless. We thus reanalyzed the two genes (*Rapgef1* and *Cdkl5*) presented in the insets of Fig. 2a of Parker et al. (2013) that were reported to show adaptive convergence for H1. Under the species tree and using the branch-site likelihood method (Zhang et al. 2005), we tested the action of positive selection in the foreground branches of IV and V (**Fig. 4.1a**) while all other branches in the tree were treated as background branches (see MATERIALS AND METHODS).

But, in neither gene did we find signal for positive selection (**Table 4.2**). We similarly analyzed the two genes (*Bnipl* and *Nubp2*) in the insets of Fig. 2b of Parker et al. (2013) that were claimed to show adaptive convergence for H2. We tested positive selection in branches I and II (under the model of a single origin of bat echolocation) or branches II, IV, and V (under the model of dual origins of bat echolocation). Again, there was no significant signal of positive selection (**Table 4.2**). Regardless, for each of the four genes, we identified those sites that have a Bayes Empirical Bayes (BEB) probability of >0.5 to be in class 2a or 2b, meaning that these sites likely have higher ω in the foreground branches than in the background branches. Only in two genes did we find more than one such site, but in neither gene was the correlation negative between ΔL and foreground ω for these sites ($r = 0.30$, $P = 0.38$ for *Cdkl5* and $r = 0.44$, $P = 0.33$ for *Bnipl* with I and II being the foreground branches). Thus, even on the basis of Parker et al.'s own criterion, proper analysis reveals no adaptive convergence in these genes.

4.4 DISCUSSION

In summary, our re-analyses of Parker et al.'s data showed that the genome-wide phylogenetic signal of molecular convergence is no stronger for echolocators than for comparable non-echolocators. Note that the phylogenetic test employed by Parker et al. is not a formal test of molecular convergence, because convergence does not necessarily result in a wrong phylogeny and a wrong phylogeny is not necessarily caused by convergence (Zhang and Kumar 1997; Castoe et al. 2009). Nonetheless, it is clear that Parker et al.'s conclusion is not supported even on the basis of this phylogenetic test. Furthermore, we found that echolocators experienced no more molecular convergence than non-echolocators when the few hearing genes known to be subject to convergent evolution were removed. Thus, the reported genomic

signatures of protein convergence must largely reflect the background chance convergences that are unrelated to the independent origins of echolocation in mammals. This conclusion, however, does not preclude the possibility that the convergent substitutions previously identified from the case studies of a few hearing proteins are important for echolocation. But given the non-negligible chance occurrence of molecular convergence, proof of adaptive convergence of a protein should include proper statistical tests and functional assays (Zhang 2006). In this regard, it is worth mentioning that both these requirements have been fulfilled for prestin, the motor protein of the outer hair cells in the mammalian cochlea (Li et al. 2010; Liu et al. 2014). Specifically, the Asn to Thr change at position 7 of prestin occurred three times, in branches II, IV, and V, respectively, and increased a key parameter of prestin function that is associated with high-frequency hearing in echolocating mammals. At the genomic scale, the rapid accrual of gene sequences has stimulated genome-wide detections of molecular convergence (Bazykin et al. 2007; Rokas and Carroll 2008; Parker et al. 2013), but efforts are also needed to establish the expected level of neutral molecular convergence, against which the observed levels can be compared such that adaptive molecular convergence may be inferred. In the absence of such neutral expectations, it is imperative to use appropriate negative controls in the study of potential adaptive molecular convergence.

4.5 MATERIALS AND METHODS

To calculate the mean site-wise ΔL for each protein, we followed the procedure described in Parker et al. (2013), using the 2326 alignments provided by the authors. Soft polytomies in H2 and H2' were resolved by RAxML (version 8.0.22) (Stamatakis 2014) as previously described (Parker et al. 2013). Marginal ancestral sequences in H0 were inferred by the

Bayesian method (Yang et al. 1995) implemented in PAML (version 4.7) (Yang 2007), using the parameters that yielded the maximum likelihood of H0. Convergent substitutions at a site are those inferred substitutions that resulted in the same amino acid and occurred in all branches examined for convergence; they thus include both convergent and parallel substitutions defined in Zhang and Kumar (1997). Coding sequences were fitted to branch-site model A in PAML with site classes 2a and 2b having $\omega_2 \geq 1$ in foreground branches and $0 < \omega_0 < 1$ (class 2a) and $\omega_1 = 1$ (class 2b) in background branches. Model A was compared by a likelihood ratio test with the null model in which ω_2 was fixed at 1. We estimated the ω value of a site by the mean of ω values of all site classes weighted by the BEB posterior probabilities with which the site belongs to these classes.

ACKNOWLEDGEMENTS

We thank the authors of Parker et al. (2013) for sharing the sequence alignments and Xiaoshu Chen, Matthew Hahn, Wei-Chin Ho, Stephen Rossiter, Jian-Rong Yang for valuable comments on early drafts of this paper. This work was supported in part by research grant R01GM103232 from the U.S. National Institutes of Health to J.Z.

REFERENCES

Bazykin GA, Kondrashov FA, Brudno M, Poliakov A, Dubchak I, Kondrashov AS. 2007. Extensive parallelism in protein evolution. Biol Direct 2: 20.

Castoe TA, de Koning AP, Kim HM, Gu W, Noonan BP, Naylor G, Jiang ZJ, Parkinson CL, Pollock DD. 2009. Evidence for an ancient adaptive episode of convergent molecular evolution. Proc Natl Acad Sci U S A 106: 8986-8991.

Davies KT, Cotton JA, Kirwan JD, Teeling EC, Rossiter SJ. 2012. Parallel signatures of sequence evolution among hearing genes in echolocating mammals: an emerging model of genetic convergence. Heredity (Edinb) 108: 480-489.

Li G, Wang J, Rossiter SJ, Jones G, Cotton JA, Zhang S. 2008. The hearing gene Prestin reunites echolocating bats. Proc Natl Acad Sci U S A 105: 13959-13964.

Li Y, Liu Z, Shi P, Zhang J. 2010. The hearing gene Prestin unites echolocating bats and whales. Curr Biol 20: R55-56.

Liu Y, Cotton JA, Shen B, Han X, Rossiter SJ, Zhang S. 2010. Convergent sequence evolution between echolocating bats and dolphins. Curr Biol 20: R53-54.

Liu Y, Han N, Franchini LF, Xu H, Pisciottano F, Elgoyhen AB, Rajan KE, Zhang S. 2012. The voltage-gated potassium channel subfamily KQT member 4 (*KCNQ4*) displays parallel evolution in echolocating bats. Mol Biol Evol 29: 1441-1450.

Liu Z, Li S, Wang W, Xu D, Murphy RW, Shi P. 2011. Parallel evolution of *KCNQ4* in echolocating bats. PLoS One 6: e26618.

Liu Z, Qi FY, Zhou X, Ren HQ, Shi P. 2014. Parallel sites implicate functional convergence of the hearing gene prestin among echolocating mammals. Mol Biol Evol 31: 2415-2424.

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.

Rokas A, Carroll SB. 2008. Frequent and widespread parallel evolution of protein sequences. Mol Biol Evol 25: 1943-1953.

Shen YY, Liang L, Li GS, Murphy RW, Zhang YP. 2012. Parallel evolution of auditory genes for echolocation in bats and toothed whales. PLoS Genet 8: e1002788.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30: 1312-1313.

Thomas GWC, Hahn MW. 2015. Determining the null model for adaptive convergence from genomic data: a case study using echolocating mammals. Mol Biol Evol 32:1232-1236.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24: 1586-1591.

Yang Z, Kumar S, Nei M. 1995. A new method of inference of ancestral nucleotide and amino acid sequences. Genetics 141: 1641-1650.

Zhang J. 2006. Parallel adaptive origins of digestive RNases in Asian and African leaf monkeys. Nat Genet 38: 819-823.

Zhang J, Kumar S. 1997. Detection of convergent and parallel evolution at the amino acid sequence level. Mol Biol Evol 14: 527-536.

Zhang J, Nielsen R, Yang Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol 22: 2472-2479.

Table 4.1 Comparison in the total number of sites that have experienced convergent substitutions from 2270 proteins

Comparison	Observed number of convergent sites	Observed number of divergent sites	*P*-value[c]
(I & II) vs. (I & III)[a]	176 vs. 223	380 vs. 352	0.012
(IV & II) vs. (IV & III)	204 vs. 287	479 vs. 445	0.00022
(V & II) vs. (V & III)	152 vs. 183	325 vs. 270	0.0067
(IV & V & II) vs. (IV & V & III)	14 vs. 4	27 vs. 22	0.083
(IV & V) vs. (IV & VI)	93 vs. 207	75 vs. 273	0.0062*
(IV & V) vs. (IV & VI)[b]	66 vs. 204	67 vs. 269	0.18

[a] Two sets of branches for which the number of amino acid sites that have experienced convergent substitutions are counted and compared. Roman numbers refer to the branch labels in **Fig. 4.1a**.

[b] After the removal of six hearing genes previously reported to be subject to convergent evolution in echolocators.

[c] G-test of the hypothesis that the number of convergent sites in a branch set is proportional to the number of divergent sites in the same branch set. An asterisk is given if the case branch set has significantly more convergent sites than expected.

Table 4.2 Branch-site likelihood ratio test of positive selection in genes claimed by Parker et al. (2013) to have undergone adaptive convergence

Gene	Foreground branches[a]	Log-likelihood under branch-site model A	Log-likelihood under null model (ω_2=1)	χ^2 (df=1)	*P*-value
Rapgef1	IV and V	-3462.452928	-3462.452928	0.000	1.00
Cdkl5	IV and V	-6161.704025	-6161.704025	0.000	1.00
Bnipl	I and II	-3498.585575	-3498.600281	0.029	0.59
Bnipl	II, IV and V	-3499.990739	-3499.990739	0.000	1.00
Nubp2	I and II	-4527.964898	-4527.964898	0.000	1.00
Nubp2	II, IV and V	-4527.964898	-4527.964898	0.000	1.00

[a] Roman numbers refer to the branch labels in **Fig. 4.1a**. All other branches are background branches.

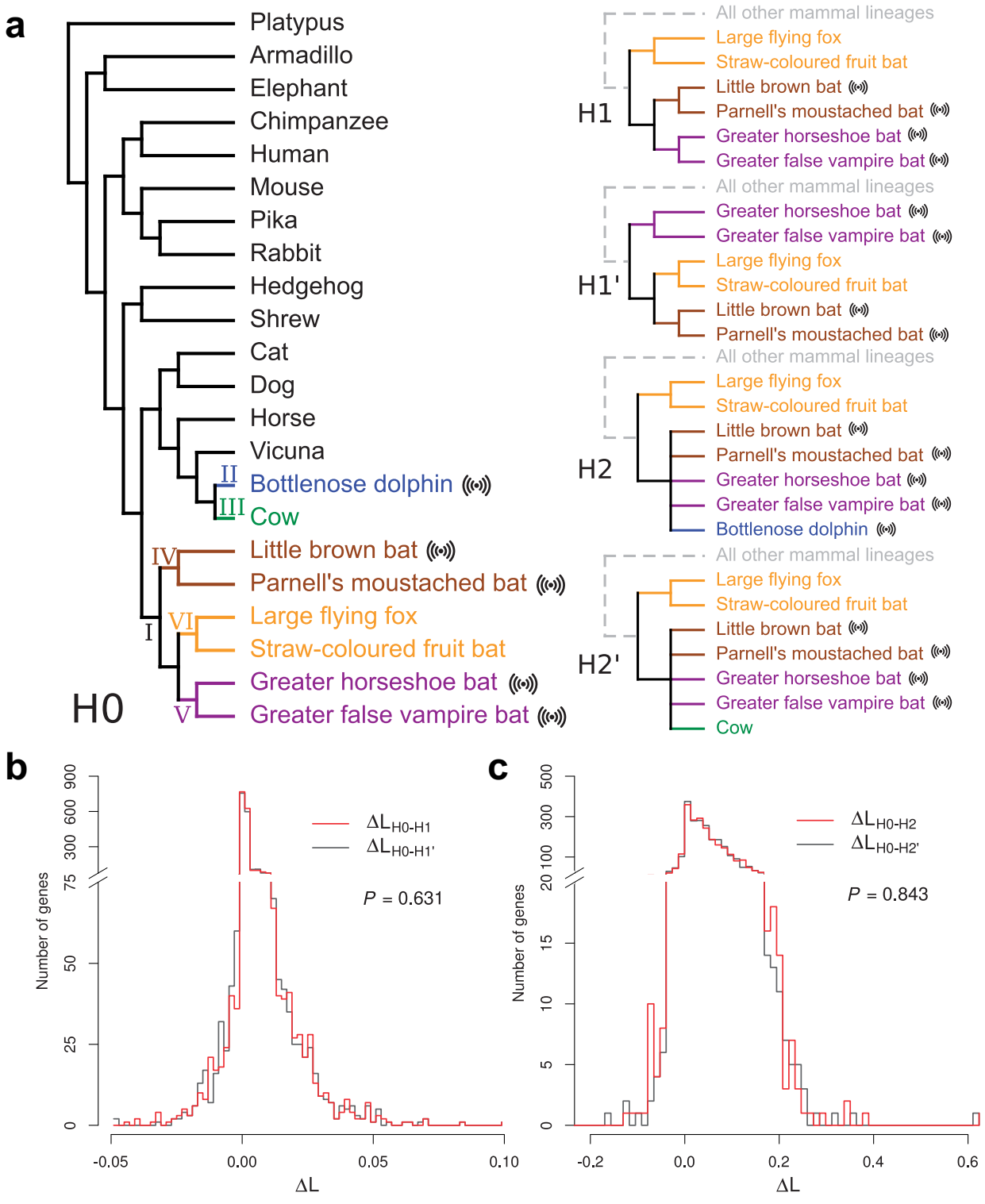


Figure 4.1 No genome-wide signatures of protein sequence convergence associated with echolocation. (**a**) Hypotheses and corresponding tree topologies. H0, species tree; H1, clustering of the two groups of echolocating bats; H1', clustering of echolocating Yangochiroptera bats and

non-echolocating Yinpterochiroptera bats; H2, clustering of echolocating bats and dolphin; H2', clustering of echolocating bats and cow. In H1, H1', H2, and H2', the tree topology for "all other mammalian lineages" is the same as in the species tree. Echolocating species are indicated with an echo symbol. The six branches where convergent sites are counted in **Table 4.1** are marked by I to VI. (**b**) Frequency distributions of $\Delta L_{H0\text{-}H1}$ and $\Delta L_{H0\text{-}H1'}$ among the 2326 proteins are not significantly different. ΔL refers to the per site logarithm of the likelihood ratio between two hypotheses for a protein. (**c**) Frequency distributions of $\Delta L_{H0\text{-}H2}$ and $\Delta L_{H0\text{-}H2'}$ among the 2326 proteins are not significantly different. In (**b**) and (**c**), the *P*-values are from Kolmogorov–Smirnov tests.

CHAPTER 5

Morphological and Molecular Convergences in Mammalian Phylogenetics[1]

5. 1 ABSTRACT

Phylogenetic trees reconstructed from molecular sequences are often considered more reliable than those reconstructed from morphological characters, in part because convergent evolution, which confounds phylogenetic reconstruction, is believed to be rarer for molecular sequences than for morphologies. However, neither the validity of this belief nor its underlying cause is known. Comparing thousands of characters of each type that have been used for inferring the phylogeny of mammals, we find that on average morphological characters indeed experience much more convergences than amino acid sites, but this disparity is explained by fewer states per character rather than an intrinsically higher susceptibility to convergence for morphologies than sequences. We show by computer simulation and actual data analysis that a simple method for identifying and removing convergence-prone characters improves phylogenetic accuracy, potentially enabling, when necessary, the inclusion of morphologies and hence fossils for reliable tree inference.

5.2 INTRODUCTION

Having a reliable species tree is prerequisite for understanding evolution, which is necessary for making sense of virtually every biological phenomenon. Traditionally, species trees are inferred using morphological, physiological, or behavioral characters, collectively called morphological characters hereinafter. The advent of molecular biology supplied

[1] This chapter is published as: Zou Z, and Zhang J. 2016. Morphological and molecular convergences in mammalian phylogenetics. Nat Commun, 7: 12758.

numerous molecular characters in the form of DNA and protein sequences, which are often (although not universally) considered more suitable than morphological characters for phylogenetic inference(Jousselin et al. 2003; Perelman et al. 2011; Wake et al. 2011; Legg et al. 2013; Springer et al. 2013; Jarvis et al. 2014). A major reason of this consideration concerns convergence, which refers to repeated origins of the same character state in multiple evolutionary lineages and is a primary source of error in phylogenetic reconstruction. Compared with morphological characters, molecular characters are believed by many (but not all) to be less susceptible to convergence(Givnish and Sytsma 1997; Page and Holmes 1998; Jousselin et al. 2003; Gaubert et al. 2005; Wiens et al. 2010; Wake et al. 2011; Davalos et al. 2012; Legg et al. 2013; Springer et al. 2013; Davalos et al. 2014). Nevertheless, this belief appears to arise in the early days of molecular systematics when morphological convergence had long been known while molecular convergence had not. Recent genetic and genomic studies, however, revealed a large number of convergences in protein sequence evolution (Zhang and Kumar 1997; Rokas and Carroll 2008; Castoe et al. 2009; Christin et al. 2010; Li et al. 2010; Zhen et al. 2012; Parker et al. 2013; Projecto-Garcia et al. 2013; Foote et al. 2015; Ujvari et al. 2015; Zou and Zhang 2015a), casting a doubt on the above belief. Determining whether morphological characters are more prone to convergence than molecular characters is important for several reasons. First, although morphological and molecular trees are often concordant with each other, this is not always the case(Pisani et al. 2007; Perelman et al. 2011; Legg et al. 2013; O'Leary et al. 2013; Jarvis et al. 2014). Knowledge of the relative prevalence of convergence in the two types of characters helps decide which tree is more trustable. Furthermore, it helps decide whether total evidence trees reconstructed jointly from the two types of characters (Wiens et al. 2010; Lee et al. 2013; Legg et al. 2013; O'Leary et al. 2013; Bieler et al. 2014; Davalos et al. 2014; Pyron

2015) are preferred over trees based on any one type. Second, phylogenetic analysis that includes fossils can help understand evolutionary relation, time, and process for fossils as well as extant species(Page and Holmes 1998; Wiens et al. 2010; Legg et al. 2013; O'Leary et al. 2013; Davalos et al. 2014; Jarvis et al. 2014; Lee and Palci 2015; Pyron 2015). Because molecular characters are inaccessible in the vast majority of fossils, knowing the frequency of morphological convergence is critical to assessing the reliability of phylogenies involving fossils. Third, convergence is caused by either repeated adaptations of different evolutionary lineages to similar environmental challenges or chance. Recent studies suggested that most molecular convergence events are attributable to chance(Foote et al. 2015; Thomas and Hahn 2015; Zou and Zhang 2015a, b). A comparison between morphological and molecular characters may provide information about the relative roles of selection and drift in morphological evolution.

Because not all morphological or molecular characters are employed by phylogeneticists, a fair comparison between the two character types in the context of phylogenetics should concentrate on characters used for phylogenetic reconstruction. To this end, we analyzed a large dataset including 3,414 parsimony informative morphological characters and 5,722 parsimony informative amino acid sites that was previously compiled for the inference of mammalian phylogeny of 46 extant and 40 fossil species(O'Leary et al. 2013). Our analysis focused on extant species because they have both types of characters. We found that morphological characters experience much more convergences than molecular characters. We devised a method to identify and remove convergence-prone characters, enabling the inclusion of morphologies and hence fossils for reliable tree inference.

5.3 RESULTS

Whole-tree analysis

Analyzing character convergence requires a species tree. Because the mammalian tree is not completely resolved, we used three trees, respectively reconstructed using the morphological characters, molecular characters, and both types of characters in the dataset. Under each tree, we inferred the ancestral states at all interior nodes for each character. For every pair of independent branches, we identified characters that showed convergence (**Fig. 5.1a**; see MATERIALS AND METHODS) and compared the mean number of convergences per character between morphological and molecular characters. For example, under the morphological tree, the exterior branches respectively leading to wolf (*Canis lupus*) and aardvark (*Orycteropus afer*) form an independent branch pair (**Fig. A.3.1a**), where 0.0072 convergences per morphological character was observed, significantly exceeding that (0.0038) per molecular character ($P = 0.03$, Fisher's exact test; see MATERIALS AND METHODS). Among 3396 pairs of independent branches in the morphological tree, 79.1% exhibit a higher convergence per morphological character than that per molecular character (**Fig. 5.1b**), significantly exceeding the chance expectation ($P < 1 \times 10^{-4}$, bootstrap test; see MATERIALS AND METHODS). There are 645 branch pairs with significantly higher per character morphological convergence than molecular convergence (Q-value < 0.05, Fisher's exact test; blue dots in **Fig. 5.1b**), whereas the opposite is true for only 61 branch pairs (orange dots in **Fig. 5.1b**). The mean number of convergence per morphological character is 1.7 times that per molecular character.

It was proposed that convergence is more fairly compared among characters or branch pairs by the ratio between the number of *c*on*v*ergence and that of *di*v*e*rgence (*Cv*/*Dv*; **Fig. 5.1a**)(Castoe et al. 2009; Thomas and Hahn 2015) because both *Cv* and *Dv* increase with the

amount of evolution. Hence, we identified divergence events for each branch pair (see MATERIALS AND METHODS) and then calculated the total number of convergence events relative to the total number of divergence events for the branch pair for each type of characters. We found that morphological characters exhibit overwhelmingly larger *Cv*/*Dv*, compared with molecular characters (**Fig. 5.1c**). The mean *Cv*/*Dv* ratio of morphological characters is 4.0 times that of molecular characters.

If the morphological tree used differs from the unknown true tree, inferring convergence under the morphological tree underestimates morphological convergence and hence the conclusion of a higher convergence for morphological characters than molecular characters should be conservative. As expected, when the above analyses were repeated under the molecular tree (**Fig. A.3.1b**) or the total evidence tree (**Fig. A.3.1c**), we found even higher convergences (**Fig. 5.1d**; **Fig. A.3.2a)** and *Cv*/*Dv* ratios (**Fig. 5.1e**; **Fig. A.3.2b**) for morphological characters than for molecular characters. Similar results were obtained using conventional measures of homoplasy such as the consistency index (*ci*) and rescaled consistency index (*rc*). That is, regardless of the tree topology used, morphological characters show lower *ci* and *rc*, thus higher homoplasy, than molecular characters (**Fig. A.3.3**).

DNA sequences instead of amino acid sequences are sometimes used as molecular characters in phylogenetics. We therefore also conducted a whole-tree analysis of the 19,227 parsimony informative nucleotide sites in the dataset, with the tree inferred from the nucleotide sequences as the molecular tree. Regardless of whether the morphological or molecular tree is used, we observed higher convergence per character and higher *Cv/Dv* ratio for morphological characters than nucleotide sites (**Fig. A.3.4a-d**).

Quartet analysis

Because the true mammalian tree is unknown, to ensure a fair comparison between morphological and molecular characters, we further examined every four species in the data that show the same phylogenetic relationship in the morphological and molecular trees, which we refer to as quartets (**Fig. 5.2a**). Given a quartet and their phylogenetic relationship, a parsimony-informative character is said to be convergent if at least two changes are required to explain the observed states (**Fig. 5.2a**). We identified all convergence events for each quartet. Averaged across 7146 quartets that can be examined, we observed 0.026 convergences per morphological character, which is three times that per molecular character (0.0085). Higher morphological convergence than molecular convergence is found in 93.9% of quartets (**Fig. 5.2b**), significantly exceeding the chance expectation ($P < 1\times10^{-4}$, bootstrap test). A total of 6087 quartets show significantly higher per character morphological convergence than molecular convergence (Q-value < 5%), while only 104 quartets show the opposite (**Fig. 5.2b**).

Given a quartet and their phylogenetic relationship, a parsimony-informative character is said to be consistent when only one change is needed to explain the observed states. Convergence provides an erroneous phylogenetic signal for the quartet, whereas consistency offers the correct signal. We thus computed, for each quartet, the ratio between the total number of *c*on*v*ergences and that of *c*on*s*istencies (*Cv*/*Cs* ratio) for each type of characters, which may be viewed as the noise/signal ratio. Again, morphological characters tend to have higher *Cv*/*Cs* ratios than molecular characters (**Fig. 5.2c**). The above results also hold when nucleotide sites instead of amino acid sites are used as molecular characters (**Fig. A.3.4e,f**).

Number of states per character

We found that 75.2% of parsimony-informative morphological characters are binary (**Fig. 5.3a**). Because binary characters can only have one kind of change given an ancestral state, it is obvious that they are susceptible to convergence once multiple changes occur. By contrast, only a small fraction (12.4%) of molecular characters are binary (**Fig. 5.3a**). The median number of states is five for molecular characters, significantly higher than that (two) for morphological characters ($P < 10^{-300}$, Mann-Whitney U test).

The probability of convergence relative to that of divergence for a character is expected to decrease with the number of states. Indeed, the *Cv*/*Dv* ratio decreases with the number of states for both types of characters (**Fig. 5.3b; Fig. A.3.5**) and this trend remains after the control of evolutionary rate (**Table A.3.1**). We estimated that the *Cv/Dv* ratio of an average morphological character is 0.89 times that of a molecular character with the same number of states, and the corresponding number is 0.55 for *Cv/Cs* (see MATERIALS AND METHODS). These results indicate that, compared with molecular characters, the higher convergence of morphological characters is caused by having fewer states rather than intrinsically higher susceptibilities to adaptive convergent evolution, because morphological characters are no more prone to convergence than molecular characters once the number of states is controlled for.

The above patterns remain unchanged even when nucleotide sites instead of amino acid sites are used as molecular characters (**Table A.3.2**). Interestingly, although there can be no more than four states at each nucleotide site, the median number of states (three) per nucleotide site is still significantly higher than that (two) per morphological character ($P < 10^{-300}$).

Removing convergence-prone characters improves phylogenetics

Because the vast majority of molecular convergences are explainable by chance(Foote et al. 2015; Thomas and Hahn 2015; Zou and Zhang 2015a, b), the fact that average morphological characters have even smaller *Cv*/*Dv* and *Cv*/*Cs* ratios than those of molecular characters of the same numbers of states suggest that most morphological convergences observed in the data analyzed are probably also attributable to chance. If convergence is owing to chance rather than lineage-specific selection, it is possible to identify and remove convergence-prone characters using species with reliable phylogenetic relations and then infer the tree for species of uncertain relations using the remaining characters. This approach would be especially beneficial to phylogenetic inference that includes morphological data because of the relatively frequent convergence in such data. We propose the following procedure when analyzing a dataset with both morphological and molecular characters. First, we infer the morphological and molecular trees separately. Second, quartets (i.e., groups of four species with the same phylogenetic relations in the two trees) are identified and the *Cv*/*Cs* ratio is calculated based on these quartets for each character. Third, we remove all characters whose *Cv*/*Cs* ratio exceeds a cutoff and infer the tree using all remaining morphological and molecular characters combined.

To investigate whether the above approach improves phylogenetic accuracy, we conducted 50 simulations of mammalian morphological and molecular characters based on their respective empirical distributions of the number of states (**Fig. A.3.6a**). Quartet analysis demonstrates that the simulated data have similar properties as the real data (**Fig. A.3.6b,c; Table A.3.3**). We measured the Robinson-Foulds distance (d_{RF}) between an inferred tree and the known true tree in simulation; d_{RF} is twice the fraction of branch partitions that differ between the two trees(Robinson and Foulds 1981); the smaller the d_{RF}, the more accurate the inferred tree. We found that d_{RF} is significantly greater for the 50 morphological trees than the 50 molecular

trees ($P = 1.6\times10^{-14}$, Mann-Whitney U test), confirming the damage of random convergence on phylogenetic accuracy. We set 10 *Cv*/*Cs* cutoffs from 5 to 0.03 and inferred 10 low-convergence total evidence trees for each simulated dataset using the above proposed procedure (see MATERIALS AND METHODS). We found that d_{RF} to the true tree is generally smaller for low-convergence trees than the original tree reconstructed using all characters (green symbols in **Fig. 5.4a**), and the improvement in phylogenetic accuracy plateaus when the cutoff reaches 0.3. By contrast, trees based on a random removal of the same number of characters do not show smaller d_{RF} when compared with the original tree (pink symbols in **Fig. 5.4a**). As expected, the mean number of states is higher for the remaining low-convergence characters than for the same number of characters randomly picked from the original simulated data (**Fig. A.3.6d**).

Removing convergence-prone characters alters the bat tree

We applied the above pipeline to the mammalian dataset including both morphological characters and amino acid sequences. The same 10 *Cv*/*Cs* ratio cutoffs as in the simulation were used in removing high-convergence characters, and low-convergence total evidence trees of all 86 extant and fossil species were inferred using the remaining morphological and molecular characters. For the 46 extant species that can be compared, the resultant low-convergence trees are generally more similar than trees based on the same numbers of randomly selected characters to the original molecular tree (**Fig. A.3.7**). The low-convergence trees are also generally more different than trees based on the same numbers of randomly selected characters from the original morphological tree (**Fig. A.3.7**). Although the true mammalian tree is unknown, these observations are consistent with our finding that convergence is less frequent in molecular characters than morphological characters.

Regarding intra-order relationships, the phylogeny of bats has been highly controversial. Specifically, all echolocating bats typically form a monophyletic group in morphological trees, suggesting a single origin of bat echolocation(O'Leary et al. 2013). But they tend to form a paraphyly in molecular trees(Teeling et al. 2000; Teeling et al. 2005; Meredith et al. 2011; Tsagkogeorga et al. 2013), suggesting the possibility of two origins of bat echolocation or one origin followed by a loss. In the original total evidence tree (**Fig. A.3.8a**) reconstructed using the data analyzed here, all five extant species of echolocating bats form a monophyly to the exclusion of the only non-echolocating extant bat *Pteropus giganteus*, with a 99.2% bootstrap support (**Fig. 5.4b**). When the 3930 characters (1007 morphological and 2923 molecular) with *Cv*/*Cs* ratio < 0.2 are used after the removal of 2407 morphological characters and 2799 molecular characters (**Fig. A.3.8b**), echolocating bats become paraphyletic; the echolocating *Rhinopoma hardwickii* and non-echolocating *P. giganteus* are grouped with a 95.0% bootstrap support (**Fig. 5.4c**). Note that using low-convergence morphological characters alone does not result in this new topology. For comparison, we generated 50 randomly subsampled datasets, each with 1007 morphological and 2923 molecular characters. Although 18 of them also yielded the same topology as in **Fig. 5.4c**, the corresponding bootstrap support ranged between 18% and 70%, suggesting that the strong support for the paraphyly of echolocators in **Fig. 5.4c** is not explained simply by subsampling of the original data. Our results are not sensitive to the *Cv*/*Cs* ratio cutoff, because the same bat relationships were recovered when any *Cv*/*Cs* cutoff of 0.3 or smaller was used.

5.4 DISCUSSION

Our analysis of comparably large numbers of morphological and molecular characters previously used in inferring the mammalian tree showed that morphological characters experienced more convergent evolution than molecular characters, confirming a long-held belief of the phylogenetics community. Nevertheless, we caution that our conclusion should be further scrutinized using additional data from additional groups of species, because they are currently based on only one, albeit very large, dataset of one group of species. There are three potential sources of error in our inference of convergence. First, use of a wrong species tree could bias our inference. But, as demonstrated, our results are robust to different species trees used. Second, our inference of convergence relies on ancestral state reconstruction by parsimony that may contain errors(Zhang and Nei 1997). But, such errors should be comparable between the two types of characters. Third, it was recently proposed that some inferred convergences may be caused by incomplete lineage sorting (ILS) rather than genuine convergent changes(Hahn and Nakhleh 2016). Similar to genuine convergence, apparent convergence owing to ILS also confounds phylogenetic inference and thus need not be separated from our estimates of convergence. Hence, the three potential errors do not affect our conclusion.

Regarding the reason behind the higher convergence of morphological characters than molecular characters, our results do not support the common view that morphological characters are intrinsically more prone to convergence because they are more frequently subject to positive selection. Instead, we found the probability of convergence for a character to decrease with the number of states and found no greater intrinsic propensities for convergence (as measured by *Cv*/*Dv* and *Cv*/*Cs* ratios) among morphological characters than molecular characters after the control of the number of states. A likely explanation for this unexpected finding is that

phylogeneticists have removed morphological characters that are subject to frequent positive selection (e.g., body size and coat color) from phylogenetic analysis, because such characters are known to lack reliable phylogenetic signals(Wiley and Lieberman 2011). As a result, the morphological characters used for phylogenetic inference have relatively low intrinsic propensities for convergence. If most convergences of the morphological characters in the data analyzed are not manifestations of repeated adaptations but pure chance, one wonders what morphological characters are responsible for the clustering of species with seemingly adaptive convergences in the morphological tree, such as the clade of the four ant- and termite-eaters: the nine-banded armadillo *Dasypus novemcinctus*, collared anteater *Tamandua tetradactyla*, Chinese pangolin *Manis pentadactyla*, and aardvark *Orycteropus afer* (**Fig. A.3.1a**). These species form three independent lineages (*Dasypus* + *Tamandua*, *Manis*, and *Orycteropus*) in the molecular tree (**Fig. A.3.1b**) as well as the total evidence tree (**Fig. A.3.1c**). We found that, even on the basis of the molecular tree, at most 14 morphological characters are inferred to have experienced convergence among the three lineages, and the actual number is likely much smaller because, for 13 of the 14 characters, convergence is but one of several equally parsimonious evolutionary scenarios. However, none of the 14 characters are apparently related to ant- and termite-eating or are specific to these four species. For instance, the only character for which the sole parsimonious reconstruction indicates convergence among the three lineages describes the shape of the medial border of humerus trochlea. The *humerus* is a long bone in the arm or forelimb that runs from the shoulder to the elbow and trochlea refers to a grooved structure reminiscent of a pulley's wheel. This character does not appear to be related to ant- and termite-eating. In fact, manatee (*Trichechus manatus*) and ring-tailed lemur (*Lemur catta*) also have the same state as the four ant- and termite-eating mammals for this character. These findings are consistent with

our conclusion that most morphological convergences observed here are caused by chance rather than repeated adaptations. Of course, we cannot exclude the possibility that a small number of morphological convergences observed in this dataset are adaptive.

Nevertheless, morphological characters experience more convergences than molecular characters, because of much fewer states in the former than the latter. The low number of states per morphological character may be related to one or both of the following reasons(Davalos et al. 2012; Davalos et al. 2014). First, curating multistate morphological characters may be more subjective and error-prone, resulting in a reduced use of such characters in phylogenetics(Wiens 2001). Second, most morphological characters may have a small state space, rendering finding multistate characters difficult(Wagner 2000).

Because of the higher prevalence of convergence among morphological characters than molecular characters and the rapid accumulation of molecular sequence data, we suggest that phylogenetic reconstruction should normally use only molecular data. In the event that molecular data are inaccessible for some taxa such as fossils, one should consider using morphological characters with relatively large numbers of states to minimize convergence in phylogenetic analysis.

Given a dataset of morphological and molecular characters, we proposed a method to reconstruct more accurate total evidence trees by identifying and removing convergence-prone characters in the dataset, and demonstrated its validity by computer simulation. Homoplasy, which interferes with phylogenetic inference, also includes reversal in addition to convergence. While our study focuses on convergence, it is worth noting that convergence-prone characters are also expected to be reversal-prone if most convergences are chance events owing to the availability of only few states, as indicated by the present data. Thus, in removing convergence-

prone characters, we effectively also take out many reversal-prone characters; the success of our method may be in part attributable to this effect. Because our method relies on the assumption that characters that are convergence-prone in the quartets analyzed are also convergence-prone in other species, it is not effective in removing characters that are convergence-prone in a few specific lineages such as those subject to adaptive convergence. In principle, one could also downweight instead of removing convergence-prone characters, but the appropriate weights are unknown. Future studies can investigate how to acquire the best weights for improving phylogenetic accuracy.

We showed that the original total evidence mammalian tree in which all echolocating bats form a monophyly is altered upon the removal of convergence-prone characters. The low-convergence tree shows a paraphyly of echolocating bats, identical to the recently published genome-based bat phylogeny(Tsagkogeorga et al. 2013). Assuming that the genome-based tree is correct, our results demonstrated the utility of our method in actual phylogenetic inference with the total evidence approach. Besides, our low-convergence tree also supports the monophyly of pangolin (*Manis pentadactyla*) and carnivores (**Fig. A.3.8b**), which is not reflected in the original total evidence tree (**Fig. A.3.8a**) but is supported by previous molecular studies(Meredith et al. 2011; Du Toit et al. 2014). As shown by our computer simulation, although removing convergence-prone characters improves phylogenetic accuracy, low-convergence trees may still contain errors. Identifying and removing convergence-prone characters is by no means a panacea for phylogenetics. While rapidly accumulating genome sequences will eventually dwarf the morphological data of any extant species, morphological data will remain useful in phylogenetic analysis that needs to contain fossils, whose value to understanding evolution is indispensable. For this reason, understanding and remedying

convergence, which is more prevalent in morphological than molecular characters, will remain an important task in phylogenetics. Of course, morphological characters that can be studied in fossils do not represent a random sample of all morphological characters. Whether this nonrandomness will bias phylogenetic inference(Sansom and Wills 2013) is also worth investigation.

5.5 MATERIALS AND METHODS

Dataset used

The original dataset is composed of 4,541 morphological characters and 11,365 amino acid sites(O'Leary et al. 2013). It includes 86 species, with 40 fossil taxa having only morphological characters and 46 extant species having both types of characters. We focused on extant species in this study because they have both types of characters for comparison. The morphological tree, molecular tree, and total evidence tree (i.e., based on both types of characters) built using the parsimony method were provided by the original study (see **Fig. A.3.1**). We removed all parsimony-uninformative characters for the 46 extant species. A parsimony-informative character has at least two states, each represented by at least two taxa. Parsimony trees of the 46 extant species based on the remaining 3,414 morphological characters or 5,722 amino acid sites agree with those based on all characters of the same types.

Whole-tree analysis

Ancestral states of each parsimony-informative character were inferred for all interior nodes in the morphological tree, molecular tree, or total evidence tree by parsimony using Mesquite (V. 3.03) (http://mesquiteproject.org/). Equal weights were given to equally

parsimonious pathways in counting convergence events, such that if only one of *n* equally parsimonious pathways for a character shows convergence for a branch pair, 1/*n* convergence events are counted. Missing extant states of a character were inferred simultaneously during the inference by parsimony, such that no additional changes are required due to the missing state assignment. Mesquite also output the number of states appearing in the 46 extant species for each character and the number of changes each character experienced along the entire tree.

An independent branch pair refers to two branches that are not ancestral to each other and contain no common node. For example, let the starting and end states of one branch (node 1 to node 3) be X_1 and X_3, and let those of another branch (node 2 to node 4) be X_2 and X_4, respectively. These two branches form an independent branch pair if (i) the four nodes are all distinct from one another, (ii) node 3 is not on the path from the tree root to node 4, and (iii) node 4 is not on the path from the tree root to node 3. For an independent pair of branches, there is a convergence if and only if $X_1 \neq X_3$, $X_2 \neq X_4$, and $X_3 = X_4$. This definition includes both parallel and convergent changes previously defined(Zhang and Kumar 1997). Similarly, there is a divergence in the independent branch pair if and only if $X_1 \neq X_3$, $X_2 \neq X_4$, and $X_3 \neq X_4$. Thus, once ancestral states are inferred, we know whether a character experiences convergence, divergence, or neither for a branch pair. For a character, the consistency index (*ci*) is the smallest minimal number of changes required to explain the observed states by any tree (Min) divided by the minimal number of changes required by the tree under evaluation (Obs). Retention index (*ri*) = (Max-Obs)/(Max-Min), where Max is the largest minimal number of changes required to explain the observed states by any tree. Rescaled consistency index (*rc*) equals consistency index multiplied by retention index(Nei and Kumar 2000). Values of *ri* and *ci* were calculated by Mesquite.

We used Fisher's exact test to compare the number of convergences per character or *Cv*/*Dv* ratio between morphological and molecular characters. For example, in the branch pair leading to wolf and aardvark, we inferred 24.67 convergences among 3414 morphological characters and 21.88 convergences among 5722 molecular characters. We rounded the decimal number of convergence to the nearest integer and tested the null hypothesis that the probability of experiencing convergence is the same for the two types of characters using the following 2×2 contingency table: 25, 3414-25, 22, and 5722-22. We obtained a two-tailed *P*-value of 0.0332 using Fisher's exact test, indicating that the frequency of convergence is significantly higher for morphologies than for sequences for the branch pair. For the same branch pair, we inferred 9.84 and 88.57 divergence events for morphological and molecular characters, respectively. We tested the null hypothesis that *Cv*/*Dv* ratio is the same for the two types of characters using the contingency table of 25, 10, 22, and 89. The obtained two-tailed *P*-value from Fisher's exact test is 3.9×10^{-8}, indicating that *Cv*/*Dv* ratio is significantly higher for morphological characters than for molecular characters for this branch pair. There were two branch pairs with no convergence and no divergence for molecular characters under the morphological tree, and three such branch pairs under either the molecular tree or the total evidence tree. These branch pairs had undefined *Cv*/*Dv* ratios for molecular characters and could not be tested in Fisher's exact test. Hence, they were excluded from the analysis and corresponding figures.

Because branch pairs (or quartets) are not independent from one another, simple parametric statistic tests cannot be used. We thus used a bootstrap method to test the null hypothesis that per character convergence is lower for morphological characters than molecular characters. First, we generated one bootstrap sample containing the same number of both morphological and molecular characters as in the original data. Second, we analyzed all branch

pairs using the bootstrap sample and examined if >50% branch pairs show a lower morphological convergence than molecular convergence. We repeated the above two steps 10,000 times and computed the fraction of bootstrap samples in which >50% branch pairs show a lower morphological convergence than molecular convergence. This fraction is an estimate of the probability that the null hypothesis is correct, hence is the *P*-value of this bootstrap test. The same bootstrap method was used to test the null hypothesis that *Cv*/*Dv* ratio and *Cv/Cs* ratio is lower for morphological characters than molecular characters in respective analyses.

Quartet analysis

Four extant taxa Y_1, Y_2, Y_3, and Y_4 are selected if they satisfy the following conditions: (i) Y_1 and Y_2 form a monophyletic group in exclusion of Y_3 and Y_4 in both the morphological and molecular trees of all extant taxa examined; (ii) Y_3 and Y_4 form a monophyletic group in exclusion of Y_1 and Y_2 in both the morphological and molecular trees; and (iii) the root of this four-species tree is located on the internal branch in both the morphological and molecular trees. Mapping a parsimony-informative character onto this quartet tree, we say that the character shows a convergence if the states of (Y_1, Y_2, Y_3, Y_4) are (A, B, A, B) or (A, B, B, A), where A and B are two observed states of the character in the four species. By contrast, we say that the character shows a consistency if (A, A, B, B) is observed. Statistical tests followed those in the whole-tree analysis, except that quartets replaced branch pairs. There were 103 quartets with zero convergence and zero consistency for molecular characters. These quartets had undefined *Cv/Cs* ratios for molecular characters and could not be tested by Fisher's exact test. Hence, they were excluded from the analysis and corresponding figures.

Comparison of *Cv*/*Dv* given the number of states

The *Cv/Dv* ratio of a character is calculated by the sum of *Cv* across all branch pairs divided by the sum of *Dv* across all branch pairs for the character. Morphological and molecular characters are divided into bins according to the number of states. For each bin, a ratio between mean morphological *Cv/Dv* and mean molecular *Cv/Dv* is calculated. Finally, this ratio is averaged across bins, weighted by the number of morphological characters in each bin. Hence, the weighted average reflects *Cv/Dv* of morphological characters relative to that of molecular characters of the same numbers of states. When the measure of *Cv/Cs* is used, the same procedure is followed except that quartets instead of branch pairs are used.

Simulation of character evolution

The evolution of morphological and molecular characters was simulated according to Markov processes, based on the tree topology and branch lengths of the nucleotide maximum likelihood tree from the original study(O'Leary et al. 2013) (**Fig. A.3.6a**). The Newick format of the tree is

(((((((((1:0.0759,(2:0.0568,3:0.0467)47:0.0234)48:0.00318,(((4:0.0448,5:0.0626)49:0.00468,((6:0.0656,7:0.0707)50:0.00570,(8:0.0634,9:0.0616)51:0.00210)52:0.0142)53:0.0150,((10:0.0602,(((11:0.0383,(12:0.0233,13:0.0128)54:0.0165)55:0.00491,14:0.0721)56:0.00919,15:0.0666)57:0.00518)58:0.0222,16:0.0559)59:0.00143)60:0.000543)61:0.00249,(17:0.1007,(18:0.0849,(19:0.1390,20:0.1468)62:0.0152)63:0.00163)64:0.00940)65:0.0110,(((21:0.0989,22:0.0676)66:0.0303,(23:0.1102,((24:0.0777,25:0.1875)67:0.00944,(26:0.0941,27:0.1660)68:0.00225)69:0.0131)70:0.00528)71:0.00149,((28:0.0618,(29:0.0913,(30:0.0414,31:0.0231)72:0.0368)73:0.00438)74:0.00775,(32:0.00806,33:0.00966)75:0.0581)76:0.00177)77:0.00997)78:0.0107,(34:0.0664,35:0.0869)79:0.

0309)80:0.00230,((36:0.0429,(37:0.1000,38:0.0439)81:0.00304)82:0.0133,(39:0.0695,((40:0.150
6,41:0.0760)83:0.00679,42:0.1275)84:0.00331)85:0.00283)86:0.0300)87:0.1278,(43:0.0834,44:0
.0739)88:0.1754)89:0.1518,(45:0.0454,46:0.0378)90:0.1518)91:0.0000. In the simulated evolution, a model equivalent to the Jukes-Cantor model assuming equal equilibrium frequencies of all states and equal exchange rates among all states was used. For each morphological character, its number of states *N* is a randomly drawn number from the empirical distribution of the number of states in the original morphological data (**Fig. 5.3a**). The 1 PAM transition matrix for this character is an *N*×*N* square matrix *M* with each non-diagonal item equal to 0.01/*N*. The relative evolutionary rate *r* of the character is randomly drawn according to a Pearson correlation of 0.64 with the number of states *n*, as was observed in the actual data. Specifically, we draw a random variable n' from the empirical distribution of the number of states and compute $r = 0.64n + n'\sqrt{1 - 0.64^2}$. We then normalize *r* such that the mean *r* from all characters equals 1.

The character evolution then starts from a random initial state at the tree root and evolves by a Markov chain along tree branches. Molecular characters were similarly simulated. Fifty simulations were conducted, each composed of 20,000 morphological characters and 40,000 molecular characters. The number of states used to generate each character and the number of substitution steps in evolution were recorded for downstream analysis. Quartet analysis based on a randomly picked simulation showed that the properties of these characters resemble those of the real data. Specifically, for almost all quartets, convergence per character and *Cv*/*Cs* ratio are higher for morphological characters than molecular character (**Fig. A.3.6b**,**c**). In addition, a significant negative partial correlation was observed between the number of states and *Cv*/*Cs* ratio when the number of steps was controlled (**Table A.3.3**).

Inference of parsimony trees

Because the evolutionary models of morphological characters have not been well established, model-based tree inference is not used here. Instead, we inferred maximum parsimony trees using PAUP4.0 (http://people.sc.fsu.edu/~dswofford/paup_test/) for both morphological and molecular data to allow fair comparisons. When analyzing the real data, 1000 replicated heuristic searches were performed with parameters from the original study(O'Leary et al. 2013). All fossil taxa were included when morphological characters were used in the inference. Consensus trees were derived when multiple equally parsimonious topologies were found, with a strict collapse of branches and equal weights of all topologies. In the analysis of simulated characters, 5000 replications were used instead of 1000. Bootstrap tests were conducted in PAUP with 1000 replicates unless otherwise mentioned. Bootstrap values were calculated and mapped by custom Python scripts; equal weights were given to all equally parsimonious trees resulting from each bootstrapped dataset.

Phylogenetic analyses of low-convergence characters

We used various *Cv*/*Cs* ratio cutoffs to remove characters whose *Cv*/*Cs* ratios are higher than the cutoffs. For example, in the real data of 9136 parsimony-informative characters, 5206 characters showed $Cv/Cs > 0.2$, according to quartet analysis. Hence, under the cutoff of $Cv/Cs = 0.2$, we retained 9136-5206 = 3930 characters for tree inference, including 1007 morphological and 2923 molecular characters. As a control, we randomly drew 1007 morphological and 2923 molecular characters from all 9136 characters and conducted a phylogenetic analysis. This control was repeated 50 times.

Data availability

All morphological and molecular data analyzed were previously published(O'Leary et al. 2013). The data matrices and related files were retrieved from MorphoBank Project 773 (http://www.morphobank.org/).

ACKNOWLEDGEMENTS

We thank Wei-Chin Ho, Bryan Moyers, Jian-Rong Yang, and especially the two anonymous reviewers for constructive comments. This work was supported in part by a research grant from the U.S. National Institutes of Health (GM103232) to J.Z.

REFERENCES

Bieler R, Mikkelsen PM, Collins TM, Glover EA, González VL, Graf DL, Harper EM, Healy J, Kawauchi GY, Sharma PP, et al. 2014. Investigating the Bivalve Tree of Life – an exemplar-based approach combining molecular and novel morphological characters. Invertebr Syst 28:32-115.

Castoe TA, de Koning AP, Kim HM, Gu W, Noonan BP, Naylor G, Jiang ZJ, Parkinson CL, Pollock DD. 2009. Evidence for an ancient adaptive episode of convergent molecular evolution. Proceedings of the National Academy of Sciences of the United States of America 106:8986-8991.

Christin PA, Weinreich DM, Besnard G. 2010. Causes and evolutionary significance of genetic convergence. Trends in genetics : TIG 26:400-405.

Davalos LM, Cirranello AL, Geisler JH, Simmons NB. 2012. Understanding phylogenetic incongruence: lessons from phyllostomid bats. Biol Rev Camb Philos Soc 87:991-1024.

Davalos LM, Velazco PM, Warsi OM, Smits PD, Simmons NB. 2014. Integrating incomplete fossils by isolating conflicting signal in saturated and non-independent morphological characters. Syst Biol 63:582-600.

Du Toit Z, Grobler JP, Kotze A, Jansen R, Brettschneider H, Dalton DL. 2014. The complete mitochondrial genome of temminck's ground pangolin (*Smutsia temminckii*; Smuts, 1832) and phylogenetic position of the Pholidota (Weber, 1904). Gene 551:49-54.

Foote AD, Liu Y, Thomas GW, Vinar 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.

Gaubert P, Wozencraft WC, Cordeiro-Estrela P, Veron G. 2005. Mosaics of convergences and noise in morphological phylogenies: what's in a viverrid-like carnivoran? Syst Biol 54:865-894.

Givnish TJ, Sytsma KJ. 1997. Consistency, characters, and the likelihood of correct phylogenetic inference. Mol Phylogenet Evol 7:320-330.

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

Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C, Ho SY, Faircloth BC, Nabholz B, Howard JT, et al. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science 346:1320-1331.

Jousselin E, Rasplus JY, Kjellberg F. 2003. Convergence and coevolution in a mutualism: evidence from a molecular phylogeny of Ficus. Evolution 57:1255-1269.

Lee MS, Palci A. 2015. Morphological phylogenetics in the genomic age. Curr Biol 25:R922-929.

Lee MS, Soubrier J, Edgecombe GD. 2013. Rates of phenotypic and genomic evolution during the Cambrian explosion. Curr Biol 23:1889-1895.

Legg DA, Sutton MD, Edgecombe GD. 2013. Arthropod fossil data increase congruence of morphological and molecular phylogenies. Nat Commun 4:2485.

Li Y, Liu Z, Shi P, Zhang J. 2010. The hearing gene Prestin unites echolocating bats and whales. Curr Biol 20:R55-56.

Meredith RW, Janecka JE, Gatesy J, Ryder OA, Fisher CA, Teeling EC, Goodbla A, Eizirik E, Simao TL, Stadler T, et al. 2011. Impacts of the Cretaceous Terrestrial Revolution and KPg extinction on mammal diversification. Science 334:521-524.

Nei M, Kumar S. 2000. Molecular Evolution and Phylogenetics. New York: Oxford University Press.

O'Leary MA, Bloch JI, Flynn JJ, Gaudin TJ, Giallombardo A, Giannini NP, Goldberg SL, Kraatz BP, Luo ZX, Meng J, et al. 2013. The placental mammal ancestor and the post-K-Pg radiation of placentals. Science 339:662-667.

Page RDM, Holmes EC. 1998. Molecular Evolution: a Phylogenetic Approach: Blackwell Science.

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.

Perelman P, Johnson WE, Roos C, Seuanez HN, Horvath JE, Moreira MA, Kessing B, Pontius J, Roelke M, Rumpler Y, et al. 2011. A molecular phylogeny of living primates. PLoS Genet 7:e1001342.

Pisani D, Benton MJ, Wilkinson M. 2007. Congruence of morphological and molecular phylogenies. Acta biotheoretica 55:269-281.

Projecto-Garcia J, Natarajan C, Moriyama H, Weber RE, Fago A, Cheviron ZA, Dudley R, McGuire JA, Witt CC, Storz JF. 2013. Repeated elevational transitions in hemoglobin function during the evolution of Andean hummingbirds. Proceedings of the National Academy of Sciences of the United States of America 110:20669-20674.

Pyron RA. 2015. Post-molecular systematics and the future of phylogenetics. Trends Ecol Evol 30:384-389.

Robinson DF, Foulds LR. 1981. Comparison of phylogenetic trees. Math Biosci 53:131-147.

Rokas A, Carroll SB. 2008. Frequent and widespread parallel evolution of protein sequences. Mol Biol Evol 25:1943-1953.

Sansom RS, Wills MA. 2013. Fossilization causes organisms to appear erroneously primitive by distorting evolutionary trees. Scientific reports 3:2545.

Springer MS, Meredith RW, Teeling EC, Murphy WJ. 2013. Technical comment on "The placental mammal ancestor and the post-K-Pg radiation of placentals". Science 341:613.

Teeling EC, Scally M, Kao DJ, Romagnoli ML, Springer MS, Stanhope MJ. 2000. Molecular evidence regarding the origin of echolocation and flight in bats. Nature 403:188-192.

Teeling EC, Springer MS, Madsen O, Bates P, O'Brien S J, Murphy WJ. 2005. A molecular phylogeny for bats illuminates biogeography and the fossil record. Science 307:580-584.

Thomas GW, Hahn MW. 2015. Determining the null model for detecting adaptive convergence from genomic data: a case study using echolocating mammals. Mol Biol Evol 32:1232-1236.

Tsagkogeorga G, Parker J, Stupka E, Cotton JA, Rossiter SJ. 2013. Phylogenomic analyses elucidate the evolutionary relationships of bats. Curr Biol 23:2262-2267.

Ujvari B, Casewell NR, Sunagar K, Arbuckle K, Wuster W, Lo N, O'Meally D, Beckmann C, King GF, Deplazes E, et al. 2015. Widespread convergence in toxin resistance by predictable molecular evolution. Proceedings of the National Academy of Sciences of the United States of America 112:11911-11916.

Wagner PJ. 2000. Exhaustion of morphologic character states among fossil taxa. Evolution 54:365-386.

Wake DB, Wake MH, Specht CD. 2011. Homoplasy: from detecting pattern to determining process and mechanism of evolution. Science 331:1032-1035.

Wiens JJ. 2001. Character analysis in morphological phylogenetics: problems and solutions. Syst Biol 50:689-699.

Wiens JJ, Kuczynski CA, Townsend T, Reeder TW, Mulcahy DG, Sites JW, Jr. 2010. Combining phylogenomics and fossils in higher-level squamate reptile phylogeny: molecular data change the placement of fossil taxa. Syst Biol 59:674-688.

Wiley EO, Lieberman BS. 2011. Phylogenetics: Theory and Practice of Phylogenetic Systematics. New York, NY: John Wiley & Sons.

Zhang J, Kumar S. 1997. Detection of convergent and parallel evolution at the amino acid sequence level. Mol Biol Evol 14:527-536.

Zhang J, Nei M. 1997. Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods. J Mol Evol 44 Suppl 1:S139-146.

Zhen Y, Aardema ML, Medina EM, Schumer M, Andolfatto P. 2012. Parallel molecular evolution in an herbivore community. Science 337:1634-1637.

Zou Z, Zhang J. 2015a. Are convergent and parallel amino acid substitutions in protein evolution more prevalent than neutral expectations? Mol Biol Evol 32:2085-2096.

Zou Z, Zhang J. 2015b. No genome-wide protein sequence convergence for echolocation. Mol Biol Evol 32:1237-1241.

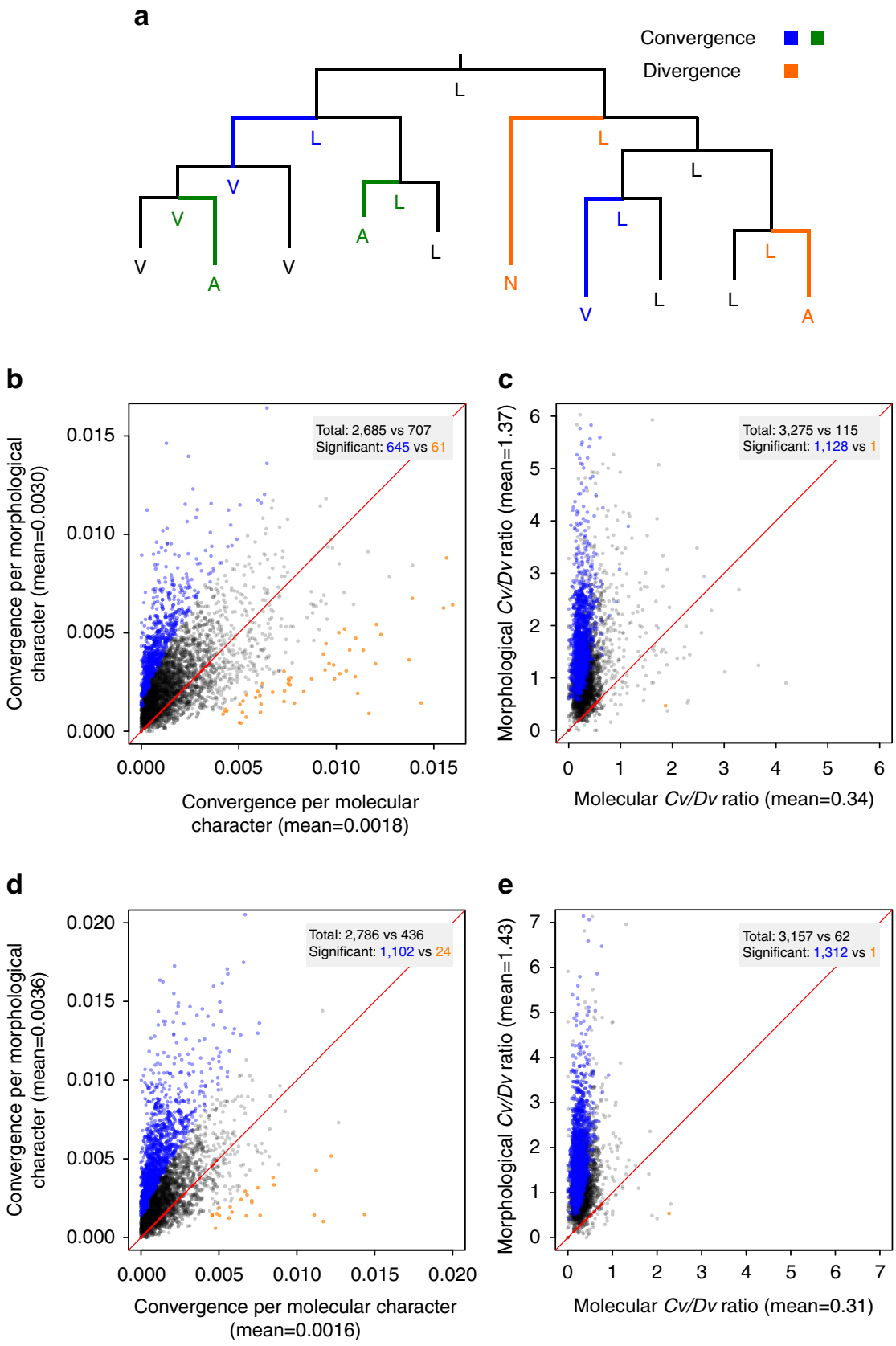

a
Convergence
Divergence
L
L
L
L
L
L
L
L
V
V
V
A
V
A
L
L
N
V
L
L
A
b
Total: 2,685 vs 707
Significant: 645 vs 61
Convergence per morphological character (mean=0.0030)
Convergence per molecular character (mean=0.0018)
0.000
0.005
0.010
0.015
c
Total: 3,275 vs 115
Significant: 1,128 vs 1
Morphological Cv/Dv ratio (mean=1.37)
Molecular Cv/Dv ratio (mean=0.34)
0
1
2
3
4
5
6
d
Total: 2,786 vs 436
Significant: 1,102 vs 24
Convergence per morphological character (mean=0.0036)
Convergence per molecular character (mean=0.0016)
0.000
0.005
0.010
0.015
0.020
e
Total: 3,157 vs 62
Significant: 1,312 vs 1
Morphological Cv/Dv ratio (mean=1.43)
Molecular Cv/Dv ratio (mean=0.31)
0
1
2
3
4
5
6
7

Figure 5.1 Whole-tree analysis of morphological and molecular convergences in mammals. (**a**) Schematic examples of convergence and divergence. Given the states of the interior and exterior nodes of the tree, the blue and green branch pairs each experienced a convergence event, while the orange branch pair experienced a divergence event. A, L, N, and V are four different states of a character. (**b**) Mean number of convergences per morphological character and that per molecular character for each branch pair examined under the morphological tree. (**c**) Convergence/divergence (*Cv*/*Dv*) ratio for each branch pair under the morphological tree. (**d**) Mean number of convergences per morphological character and that per molecular character for each branch pair examined under the molecular tree. (**e**) *Cv*/*Dv* ratio for each branch pair under the molecular tree. In panels (**b**)-(**e**), each dot represents a branch pair. In the grey box of each panel, "total" refers to the numbers of dots above and below the diagonal (dots on the diagonal are not counted), respectively, and "significant" refers to the numbers of dots significantly (at *Q*-value of 0.05) above (blue) and below (orange) the diagonal, respectively. Total number of dots above the diagonal significantly exceeds that below the diagonal in panels (**b**)-(**e**) ($P < 1\times10^{-4}$, bootstrap test). For panels (**c**) and (**e**), branch pairs with infinite *Cv/Dv* values are not plotted but included in the comparison.

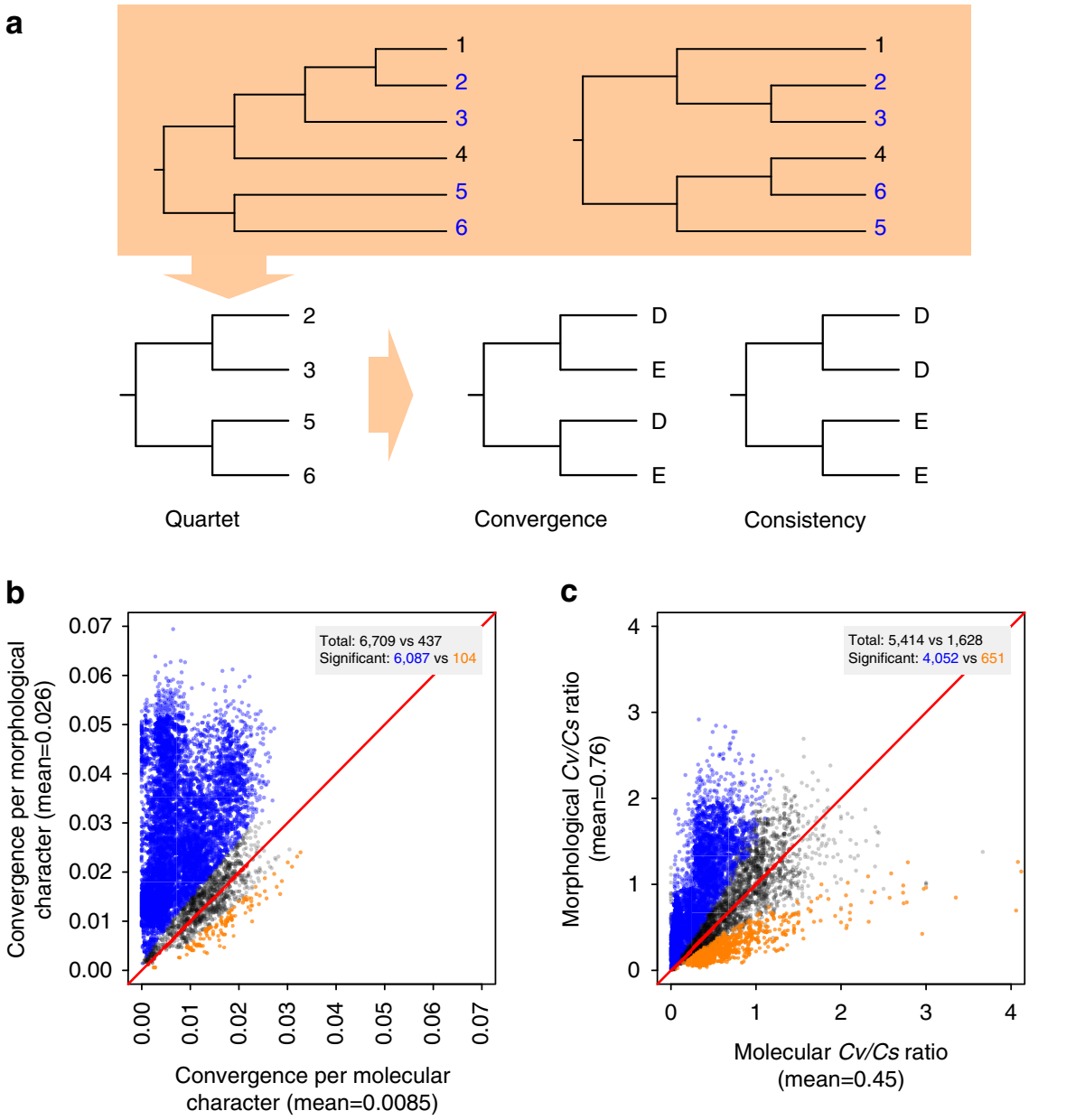


Figure 5.2 Quartet analysis of morphological and molecular convergences in mammals. (**a**) A schematic example of a quartet, which are four species (2, 3, 5, and 6) showing the same phylogenetic relationship in the morphological (left) and molecular (right) trees. Examples of character states exhibiting convergence and consistency are shown. (**b**) Mean number of convergences per morphological character and that per molecular character for each quartet examined. (**c**) Convergence/consistency (*Cv*/*Cs*) ratio for each quartet. In panels (**b**) and (**c**), each dot represents a quartet. In the grey box of each panel, "total" refers to the numbers of dots above and below the diagonal (dots on the diagonal are not counted), respectively, and "significant" refers to the numbers of dots significantly (at *Q*-value of 0.05) above (blue) and below (orange) the diagonal, respectively. Total number of dots above the diagonal significantly exceeds that

below the diagonal in panels (**b**) and (**c**) ($P < 1\times10^{-4}$, bootstrap test). In panel (**c**), quartets with infinite *Cv/Cs* values are not plotted but included in the comparison.

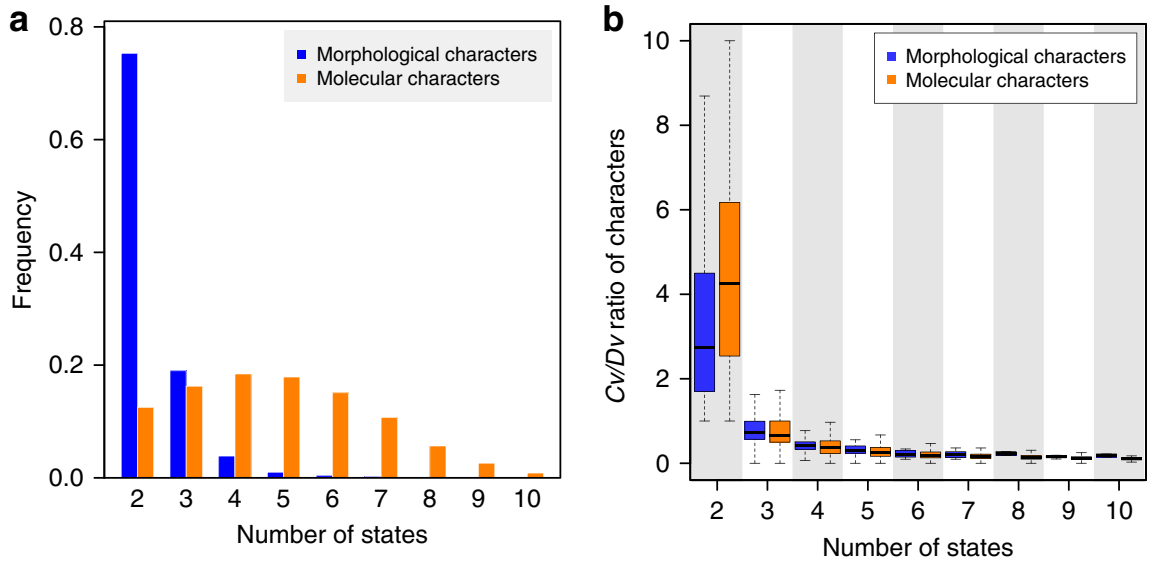


Figure 5.3 Morphological characters tend to have fewer states than molecular characters. (**a**) Frequency distribution of the number of states per character. (**b**) *Cv*/*Dv* ratio of a character decreases as the number of states increases. *Cv/Dv* ratio of a character is the sum of convergences across all branch pairs divided by that of divergences. The top and bottom edges of a box represent the first and third quartiles of the distribution, respectively, while the thick line inside the box represents the median. The two whiskers show the maximum value not greater than the 1st quartile plus 1.5 times the box height and the minimum value not smaller than the 3rd quartile minus 1.5 times the box height, respectively. *Cv*/*Dv* ratios are calculated under the morphological tree. The same pattern is observed when *Cv*/*Dv* ratios are calculated under the molecular tree (**Fig. A.3.5**).

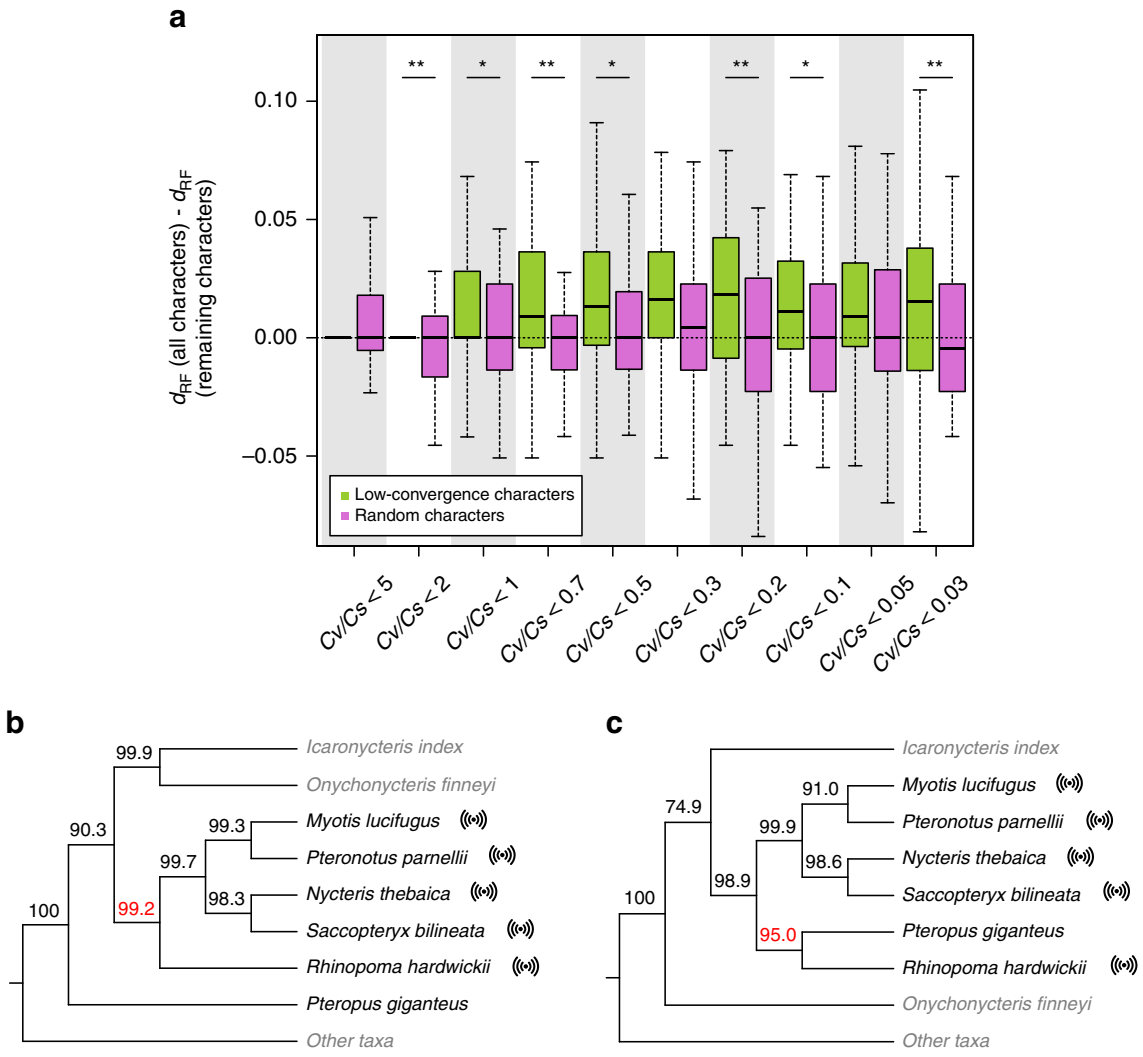


Figure 5.4 Removing convergence-prone characters improves phylogenetic accuracy. (**a**) Simulation results showing that using characters with *Cv*/*Cs* ratios below certain cutoffs reduces the Robinson-Foulds distance (d_{RF}) between the true tree and the inferred tree, while using the same number of randomly picked characters does not. The top and bottom edges of a box respectively represent the first and third quartiles of the distribution from 50 simulations, while the thick line inside the box represents the median. The two whiskers show the maximum value not greater than the 1st quartile plus 1.5 times the box height and the minimum value not smaller than the 3rd quartile minus 1.5 times the box height, respectively. *Cv*/*Cs* ratios are estimated based on quartets (sets of four species with the same phylogenetic relationships in the inferred morphological and molecular trees of the simulated data). *, $P < 0.05$, paired Mann-Whitney *U* test from 50 simulations; **, $P < 0.01$. (**b**) Bat relationships based on the parsimony tree reconstructed using 9136 informative morphological and molecular characters. Echolocating species are marked with an echo sign. Extant bats are in black, while fossil bats and other taxa are in grey. Bootstrap percentage is presented for each internal node, with the red color

highlighting the bootstrap percentage for the monophyly of echolocators. (**c**) Topology based on 3930 informative characters with *Cv*/*Cs* ratios < 0.2. The red color highlights the bootstrap percentage for the sister relationship between the non-echolocator *Pteropus giganteus* and the echolocator *Rhinopoma hardwickii*.

CHAPTER 6

Amino Acid Acceptance Rates Differ between Clades on the Tree of Life during Genome-Wide Protein Evolution

6.1 ABSTRACT

In phylogenetic and molecular evolution analyses, it is common to model protein sequence evolution by Markov process. Mechanistically, the transition probability between different amino acid or codon states is decided by mutational probability and acceptance rate during selection. The relative acceptance rate of each type of amino acid change is usually assumed to be invariable among different species in practice. However, we use maximum likelihood approach to infer the relative acceptance rates for a broad sample of 68 two-species clades, and show that the rates can differ between two clades, e.g. rodents and carnivores. We designed a shuffling test to confirm the significance of acceptance rate difference between two clades, and found that the difference exists even between orthologous genes in different clades. Our results suggest a genome-wide clade-specific factor affecting the acceptance rate, of which the cause awaits exploration. Furthermore, the application significance of this acceptance rate difference might be of importance, because currently it is widely assumed that a single largely constant matrix of amino acid acceptance rates applies in all species.

6.2 INTRODUCTION

Amino acid substitution models play essential roles in protein evolution analysis, describing the transition probabilities of changing from each amino acid state to another in a Markov process. Currently, two types of amino acid substitution models are widely used: empirical models and mechanistic models. Empirical models summarize the relative rates of

substitutions between different amino acid pairs in a set of known orthologous sequence alignments (Dayhoff et al. 1978; Jones et al. 1992; Whelan and Goldman 2001; Le and Gascuel 2008), without disentangling biological factors affecting the rates. With no explicit component describing mutational bias, empirical models cannot distinguish the effect of mutation from that of selection during sequence evolution (Yampolsky and Stoltzfus 2005; Stoltzfus and Yampolsky 2007). Mechanistic models at codon level have been developed to account for this distinction (Goldman and Yang 1994; Yang 1998, 2007), with transition-trasnversion bias reflected by ratio κ and overall selection effect on nonsynonymous substitutions reflected by a ratio ω, or d_N/d_S. While ω can differ between different codon sites or phylogenetic lineages, the common practice does not consider its variation among different amino acid pairs. For example, a single ω is used to describe the fixation probability given either an I to L change or an R to D change, etc., although I to L change is intuitively more acceptable due to similar biochemical properties. Here we follow previous literature to denote the rate that a certain type of amino acid change happens relative to synonymous expectation as the acceptance rate of this amino acid pair (Miyata et al. 1979; Yang et al. 1998).

However, it has been known that different amino acid changes have different impact on the structure and function of the protein, thus different acceptance rates. Classification of amino acid changes into "conservative" ones and "radical" ones according to the dissimilarity of biochemical properties led to mixed results about which category is more acceptable (Rand et al. 2000; Zhang 2000). The acceptance rates for individual types of amino acid changes has also been estimated by different approaches. For example, Grantham (1974) and Miyata et al. (1979) derived distance / dissimilarity measures between pairs of amino acids based on physiochemical properties such and polarity and volume. Yampolsky and Stoltzfus (2005) did meta-study on a

compiled set of studies investigating effects of artificially introduced amino acid changes, and argued that the resulted "experimental exchangeability" (EX) is naturally independent of mutational bias in evolution, hence a better formulation of acceptance rate. On the other hand, Tang et al. (2004) adopted a molecular evolution approach to count the observed amino acid changes and the respective expectation, and derive an Evolutionary Index (EI) as the ratio of the two. Mutational factors such as transition-transversion bias and number of nucleotide substitutions are taken into account when counting the expectation. Hence EI describes only the acceptance rate of each amino acid pair, and was later used for detecting positive selection (Tang and Wu 2006). The authors argued that EIs derived from different groups of species are highly correlated, and calculated a universal measure U describing the relative acceptance rates common among different groups. A codon substitution model was also proposed in Yang et al. (1998), in which each acceptance rate ω_{ij} between amino acid *i* and *j* can vary as a free parameter, so that maximum likelihood (ML) estimation of all ωs could be implemented.

Correlations has been found between different acceptance measures and also acceptance rates estimated from different datasets. The ω_{ij} values correlate with the Miyata et al. (1979) physiochemical distance, so does the EI values with Grantham's distance, PAM and experimental exchageability (Yang et al. 1998; Tang et al. 2004; Stoltzfus and Norris 2016). The experimental exchangeability EX, as well as BLOSUM matrix, Grantham distance, WAG matrix, etc., was shown to be able to serve as predictors of disease causing potential of human SNPs and independent mutational effect of artificial amino acid exchanges (Yampolsky and Stoltzfus 2005). Furthermore, EI values estimated from four independent taxonomic groups are highly correlated, and a common "severity-of-effect distribution" is suggested for experimental exchangeability measurement regardless of which focal protein is investigated (Tang et al. 2004;

Yampolsky and Stoltzfus 2005). However, all the correlations here are far from perfect. The highest EI correlation coefficient shown in Tang et al. (2004) is less than 0.91, between rodent and yeast. The general model with free ω_{ij}s fit sequence alignment data better than all other nested models mentioned by likelihood ratio test (Yang et al. 1998), indicating unique acceptance rate for a particular set of sequences.

While the existence of a common set of amino acid acceptance rates is discussed in the above studies, it is unclear whether the rates are specifically affected by which species / clade of species in the tree of life they are estimated from. In this study, we adopt molecular evolution approach to answer this question. We first show the accuracy of the ML acceptance rate inference method in PAML by simulating sequence alignments with certain sets of acceptance rate. Then, acceptance rates of 75 amino acid pairs were inferred by maximum likelihood in different groups of two-species clades. Empirical statistical tests show that correlations of acceptance rates between two clades can be significantly smaller than random expectation, indicating that the relative acceptance rates are different among different clades. With no previous phylogenetics study adopting this variation, we discuss the potential impact of our findings on the biology and application of protein and codon evolutionary models.

6.3 RESULTS

ML inference of acceptance rates in simulated sequence alignments is accurate.

We follow the codon evolution model in Yang et al. (1998) as follows:

$$q_{uv} = \begin{cases} 0, & \text{if the two codons differ at more than one position,} \\ \pi_v, & \text{for synonymous transversion,} \\ \kappa\pi_v, & \text{for synonymous transition,} \\ \omega_{ij}\pi_v, & \text{for nonsynonymous transversion,} \\ \omega_{ij}\kappa\pi_v, & \text{for nonsynonymous transition,} \end{cases} \quad (1)$$

In this equation, q_{uv} is the rate of substitution between codon u and codon v; π_v is the equilibrium frequency of the resulted codon v, summarized into vector $\boldsymbol{\pi}$; κ is the transition:transversion ratio for synonymous substitutions; ω_{ij} is the acceptance rate between amino acid i and j, which are respectively encoded by codon u and v. In this model, ω_{ij} exists only if substitution between amino acid i and j can be realized by a single nucleotide change, hence forming a vector $\boldsymbol{\omega}$ with 75 elements Parameters including κ and $\boldsymbol{\omega}$ can be inferred in a maximum likelihood manner by PAML (Yang 2007).

To validate the accuracy of this ML inference method, we extract realistic parameters from real data and simulate sequence evolution under the above codon model. Pairwise alignments of all orthologous coding sequences (CDS) between two *Escherichia coli* strains (*E. coli* clade) and between mouse and rat (rodent clade) were derived and concatenated, resulting in an *E. coli* clade alignment and a rodent clade alignment. The program codeml in PAML was used to estimate the codon equilibrium frequencies $\boldsymbol{\pi}$, transition:transversion ratio κ, and acceptance rates $\boldsymbol{\omega}$ (see MATERIALS AND METHODS). Next, we simulated four sets of CDS alignments under the codon model above: (1) with rodent $\boldsymbol{\pi}$, κ and $\boldsymbol{\omega}$; (2) with *E. coli* $\boldsymbol{\pi}$, κ and $\boldsymbol{\omega}$; (3) with rodent $\boldsymbol{\pi}$ and κ but *E. coli* $\boldsymbol{\omega}$; (4) with *E. coli* $\boldsymbol{\pi}$ and κ but rodent $\boldsymbol{\omega}$. In this case, evolution (1) and (4) are under the same rodent $\boldsymbol{\omega}$, while both (2) and (3) uses *E. coli* $\boldsymbol{\omega}$. Then these parameters were again inferred by codeml from the simulated CDS alignments. If $\boldsymbol{\omega}$ can be accurately estimated independent of other parameters during the CDS evolution, we would predict high correlation between the inferred $\boldsymbol{\omega}$ and the corresponding original $\boldsymbol{\omega}$ used in simulation; also we would expect high correlation of inferred $\boldsymbol{\omega}$ between (1) and (4), (2) and (3) while that between any other combinations are low. A third prediction is that correlation of inferred $\boldsymbol{\omega}$s between replicates within each parameter set should be high. In fact, all three

predicted patterns were observed, as shown in **Fig. 6.1**. Acceptance rates ω inferred from simulated sequence alignments have a correlation of $r \geq 0.90$ with the actual ω assigned during simulations (**Fig. 6.1a**). When inferred ωs are compared with each other, alignments simulated under the same ω have highly correlated inferred ωs (first six columns in **Fig. 6.1b**), regardless of whether the other parameters in simulation are the same. On the other hand, alignments simulated under different ωs have significantly dissimilar inferred values of ωs, even if the other parameters are identical (columns 7 – 10 in **Fig. 6.1b**). These results confirm that the maximum likelihood inference of ω in codeml is accurate. One deviation of the above simulations from real sequence evolution is the evolutionary rate variation among sites. However, even if we assume exponentially distributed site-specific relative rate during simulation, the patterns described above does not change (data not shown), validating the applicability of the ML inference to real sequence data.

Relative acceptance rates are different among different species clades.

Now that we have confirmed the accuracy of ML inference of acceptance rates, the inference was applied to a set of 68 available species clade that is widely sampled across the tree of life. Each clade includes a pair of closely related species, and equilibrium of protein sequence evolution is assumed within clade. 15 eukaryotic clades were sampled, including six pairs of vertebrates, two pairs of insects, two pairs of fungi, three pairs of plants and two pairs of protozoans belonging to outgroups of above groups. 53 prokaryotic clades including one pair of archaea and 52 pairs of bacteria were also sampled (see **Table. S1**). With the genome-wide concatenated coding sequence alignments available for each clade (see MATERIALS AND METHODS), we respectively estimated 68 ωs by codeml. Since different clades may have

different overall d_N/d_S value, this ω_{all} is used to normalize the $\boldsymbol{\omega}$ of the focal clade, so that for each of the 68 clades, relative acceptance rates $\boldsymbol{\omega' = \omega/\omega_{all}}$. **Fig. 6.2a** shows two examples of the inferred $\boldsymbol{\omega'}$ for the *E. coli* and the rodents clade. The absolute values of acceptance rates are different between the two clades. The ω'_{ij}s in *E. coli* ranges from 0.105 to 9.276, while the range in rodents is 0.218 to 3.237. Furthermore, while the correlation between the two $\boldsymbol{\omega'}$s can be as high as 0.60 (Spearman's rank correlation, P = 2E-8), there is still considerable discordance between the same ω category in different clades. For example, in *E. coli*, the amino acid pair arginine – lysine (R-K) has the highest acceptance rate, while in rodent the pair with highest ω is threonine – methionine (T-M). Comparison between symmetric items in the two halves of matrix in **Fig. 6.1a** indicates this general trend.

This discordance of acceptance rate is further reflected in **Fig. 6.2b**, where the ranks of ω'_{ij}s for all 68 clades are shown. In each row, the rank of each ω_{ij} among the 75 ωs is color-coded for a certain clade. If $\boldsymbol{\omega'}$ are closely related between different clades, we would expect the same ω category has similar rank in all clades, i.e. showing similar color in each column. Nevertheless, what we observe in **Fig. 6.2b** is a rugged landscape, where a single ω category can have drastically different ranks in different clades. For example, the acceptance rate ω_{WC} for tryptophan – cysteine is ranked as the second lowest among 75 ω_{ij}s in the clade with two strains of the bacteria *Prochlorococcus marinus*, but as second highest in the clade with two strains of *Phytoplasma asteris* or in the clade *Borrelia garinii* + *Borrelia afzelii*. On average, for a certain amino acid pair, the highest rank and the lowest rank of its acceptance rate in 68 clades can differ by 51, indicating the relative acceptance rate of each amino acid pair varies considerably among clades. On the other hand, there exist amino acid pairs that have relatively similar ranks in all

clades, e.g. the rank of serine – alanine is always high (1^{st} – 16^{th}) among 75 pairs, while the rank of valine – aspartic acid is always low (60^{th} - 75^{th}).

In addition to the rank variation of acceptance rates, we also investigated the actual variation range of each particular ω category. **Fig. 6.2c** show the percentage of each ω'_{ij} relative to the largest ω' value in the same category. As an example, the largest relative acceptance rate for alanine – aspartic acid ω'_{AD} (1.775) is in the clade with two species of *Burkholderia*, and the ω'_{AD} for other clades ranges from 21.5% to 79.5% of this value. The prevalence of blue cells in this figure indicates that the actual numerical value of ω changes considerably among clades. Together with an example $\boldsymbol{\omega}$ matrices and the rank variation in **Fig. 6.2a** and **6.2b**, so far we have shown that the ML inferred relative acceptance rates are different among different species clades.

Now that the above patterns support a significant variation of $\boldsymbol{\omega}$, the underlying reason could still be trivial. For example, the ML inference process could cause the difference between the estimated $\boldsymbol{\omega}$ due to small sample size for some amino acid pairs. As a control, we simulated sequence evolution of the same 68 clades with a single $\boldsymbol{\omega}$ inferred from the rodents clade, while all other parameters were set the same as those inferred from each clade, including length of sequence alignments and genetic distance between the two species in the clade. Inference of $\boldsymbol{\omega}$s for these simulated clades were conducted, and the ranks of each ω'_{ij} and percentages relative to the largest value in each category are shown in **Fig. 6.2d** and **Fig. 6.2e**, corresponding to **Fig 6.2b** and **6.2c**. Clearly, the ranking is more similar between clades and percentage variations are smaller for most ω categories, confirming that simulated sequence evolution with a single $\boldsymbol{\omega}$ cannot generate the variation of ω'_{ij} ranks and ranges we saw in the real case. Hence, acceptance rate variation is not likely to be caused by trivial technical artifact. For a positive control,

simulated alignments using the 68 inferred $\boldsymbol{\omega}$s for corresponding clades show similar pattern as we see in **Fig 6.2b** and **6.2c** (data not shown).

Shuffling test indicates significant dissimilarity of acceptance rates between clades

Although we have shown the dramatic variation of inferred $\boldsymbol{\omega}$s among species clades, there is no formal statistical test against randomized control. Thus, the possibility still exists that any two pieces of sequence alignments may show different acceptance rates. Since ML inference of individual ω_{ij}s primarily uses the information of variable sites, we designed a shuffling scheme to test whether uniqueness of $\boldsymbol{\omega}$s is specific for individual closely related clades. The shuffling test shuffle variable sites showing the same codon in one species of each clade between two clades. In this sense, only the amino acid substitutions and associated acceptance rates are randomized between two clades while other properties remain largely unaltered. If acceptance rate difference between the two clades is significant, we would predict the two $\boldsymbol{\omega}$s after shuffling show higher correlation with each other than before shuffling.

To check the reliability of this shuffling test, we conducted the test on the previously mentioned simulated sequence alignments (1) – (4). Comparison between (1) and (4) or comparison between (2) and (3) should show no significance, since both (1) and (4) were simulated under the rodent $\boldsymbol{\omega}$, while (2)(3) evolve under *E. coli* $\boldsymbol{\omega}$. In contrast, we use (1) – (3) and (2) – (4) as positive control, because in each clade pair, the parameters $\boldsymbol{\pi}$, κ are identical and $\boldsymbol{\omega}$s are different during simulation. The shuffling tests show conservative performance on these control comparisons as shown in **Fig. A.4.1**. For the five replicates of simulated clades (1) and (4), the correlation coefficients before shuffling (red dots) are significantly higher than those after shuffling (grey dots). The same is true between clade (2) and (3). In contrast, between (1) –

(3) or (2) – (4) when initial ωs are different, correlation increases after shuffling. If the test has maximum power, we would expect for test between (1) – (4) and (2) – (3), the red dots fall within the distribution of grey dots. The observed pattern is conservative: For clades with similar acceptance rates, randomization might even decrease the between-clade correlation, so that we could only detect large ω difference where this decrease does not outweigh the difference. Pearson correlation is used throughout the text and figures here, and Spearman's rank correlation show comparable patterns that do not vary any conclusions (data not shown).

We apply this shuffling test to the 68 clades analyzed above. To check if the presumed acceptance rate difference has a phylogenetic pattern, we fix the rodents as one clade, and conduct shuffling test between rodents with each of the other 67 clades. Despite the conservative nature of the test, 8 out of 14 comparisons with eukaryotes and 2 out of 53 comparisons with prokaryotes show significantly lower correlation of ω than 100 randomized controls (columns with red asterisks in **Fig. 6.3a**). As an example, the correlation coefficient between two ωs inferred for rodents and great apes (human vs. chimpanzee) is 0.78, but after alignment shuffling this value ranges from 0.89 to 0.92, indicating the original two ωs are significantly different. Since in **Fig. 6.3a** the clades compared with rodents clade are ranked roughly into phylogenetic order (see MATERIALS AND METHODS), there is a trend that clades that are more distantly related with rodents tend to show weaker correlation of ω with rodents. Moreover, clearly most significant differences are seen between rodents and another eukaryotic clade. As a direct negative control, we conducted the same set of shuffling tests in the previously mentioned 68 clades simulated with a single rodent clade ω. None of the 67 comparisons show significance (**Fig. A.4.2a**, compare with **Fig. 6.3a**). In contrast, positive control clades simulated using each of the 68 inferred ωs show significant ω difference in 37 of 67 comparisons (**Fig. A.4.2b**). These

results support the power of the shuffling test, as well as the fact that the codon model used for simulation is realistic enough to generate acceptance rate difference we discover in real data.

Orthologous coding sequences can show different acceptance rates between clades.

Now that we observe the acceptance rate difference between clades, the next question would be the underlying mechanism. The coding sequence alignments of each clade contains all nuclear proteins with orthologous sequences available for both species. Hence one hypothesis is that acceptance rate difference between clades stems from different gene contents on which inference of ωs is based. To test this hypothesis, we want to compare only the orthologous gene contents between two clades. For the purpose, we collected 11 pairs of mammalian clades. For each pair of clades, one clade is set as human – chimpanzee, and the other clade ranges from macaque – vervet to opossum – Tasmanian devil (see **Table. S2**). An alignment of all four species in two clades were obtained and separated into two alignments for respective clades. Thus the comparison only involves orthologous codon positions shared between two clades. If shuffling tests still support significant ω difference between clades, the previous hypothesis is falsified. It turns out that for all 11 comparisons, true correlation between two ωs is significantly lower than the distribution of all correlations after shuffling (**Fig. 6.3b**).

As orthologous sequences can show significantly different ωs, this acceptance rate difference seems to be caused by a genome-wide factor rather that gene-specific processes. To further investigate this hypothesis, we took the rodents clade (mouse vs. rat) and the avian clade (chicken vs. turkey) as an example, and separate genes in each clade into two categories: those with orthologs existing in the other clade, and those without orthologs in the other clade. Hence, we have four groups of genes forming four concatenated alignments: orthologs in rodents (RO),

non-orthologs in rodents (RN), avian orthologs (AO) and avian non-orthologs (AN). $\boldsymbol{\omega}$s were inferred for each alignment. The correlation coefficient between RO and RN (denoted as $r_{(RO\text{-}RN)}$) is 0.98, while $r_{(AO\text{-}AN)}$ is 0.96. In contrast, correlation between orthologous genes $r_{(RO\text{-}AO)}$ is only 0.85, even lower than $r_{(RN\text{-}AN)}$, which is 0.87. That different genes in the same genome has more similar $\boldsymbol{\omega}$ than the same genes in different genomes confirm the existence of a genome-wide factor affecting amino acid acceptance rates. Notably, we also checked codon frequencies of the four groups, and the same pattern is true: different genes in the same genome share more similar codon frequencies ($r_{(RO\text{-}RN)} = 0.99$, $r_{(AO\text{-}AN)} > 0.99$) than genes in different genomes ($r_{(RO\text{-}AO)} = 0.95$, $r_{(RN\text{-}AN)} = 0.90$).

6.4 DISCUSSION

Many studies have measured the "exchangeability" of amino acid pairs (Kawashima et al. 2008). In this study, we adopt a maximum likelihood inference method to estimate amino acid acceptance rates in multiple closely related species clades. The mechanistic codon substitution model separate mutational bias during sequence evolution from specific selection effect on different amino acid substitutions. The likelihood method is better than counting method such as in Tang et al. (2004) since it naturally account for multiple hits of codon substitution and is thus less affected by genetic distance between two species in one clade. Compared with the experimental exchageability by Yampolsky and Stoltzfus (2005), this method is not limited by the time and labor costs of assaying the fitness effect of individual mutation. Instead we can utilize existing large-scale comparative sequence data, which in turn avoid potential bias cast by experimental environments or selection of focal protein molecules.

With a sample of 68 clades containing both prokaryotes and eukaryotes, our result indicate the indispensable acceptance rate difference among certain different branches on the tree of life. Among 53 prokaryotic clades, only two show difference of $\boldsymbol{\omega}$ compared with rodents. On the contrary, we can see an increased level of $\boldsymbol{\omega}$ variation in prokaryotic species: In **Fig. 6.2a**, the relative acceptance rates $\boldsymbol{\omega}'$ of *E. coli* have larger range of variation than those of rodents, and eukaryotes at the top of **Fig. 6.2c** tends to have less variation of individual ω'_{ij} values than the prokaryotes. The variation of prokaryotic acceptance rates might be caused partially by smaller number of codon sites available and hence larger stochasticity of ML inference, which is also reflected by wider distributions of correlation coefficients *r* after alignment shuffling with rodents. Given the conservative nature of the shuffling test, there might be more prokaryotic clades that have different acceptance rate from the rodents clade. Importantly, the focus of acceptance rate difference we discuss here is the relative rate ($\boldsymbol{\omega}'$) difference. Each clade has an overall d_N/d_S value indicate the strength of purifying selection on coding sequence, and this value may well differ among clades. Instead, our focus is whether some amino acid pair always have higher acceptance than the others in all species. If not, then certain amino acid changes are favored in some species but no so in others, hence the relative rate difference. Since we mostly calculate correlation as a description of similarity, results based on $\boldsymbol{\omega}'$ or $\boldsymbol{\omega}$ should be equivalent.

One additionally notable pattern is the general phylogenetic trends of acceptance rates correlation and shuffling test significance. The $\boldsymbol{\omega}$ correlations between rodents clade and other mammalian clades are on average higher than those between rodents and other eukaryotes, while the latter are higher than those between rodents and prokaryotes (see **Fig. 6.3a**). For simulated clades with the same $\boldsymbol{\omega}$, we only observe this decrease of correlation with genetic distance in shuffled alignments (grey dots in **Fig. A.4.2a**). Moreover, significant $\boldsymbol{\omega}$ difference is observed

mostly in eukaryotes. These results suggest that, compared with $\boldsymbol{\omega}$ inference of the original alignments, $\boldsymbol{\omega}$s of the shuffled alignments are more affected by phylogenetic distances between the two clades compared. This is further confirmed by an additional set of simulation, where we simulate clades with an $\boldsymbol{\omega}$ that deviates from the rodents $\boldsymbol{\omega}$ with a certain level of variation, and conduct shuffling tests between these clades and a rodents clade simulated under the rodents $\boldsymbol{\omega}$. As shown in **Fig. A.4.3**, larger variation from original $\boldsymbol{\omega}$ lead to smaller *r* (red dots), but *r*s of the shuffled alignments decrease as the other parameters ($\boldsymbol{\pi}$, κ, genetic distance between species, number of codons in alignment) used in the simulation comes from clades more distant from rodents. Furthermore, in **Fig. 6.1a**, lower correlation of $\boldsymbol{\omega}$ with true values is observed as long as E. coli $\boldsymbol{\pi}$ and κ are used, indicating these parameters alone can affect the variation of $\boldsymbol{\omega}$ inference. This may explain why most comparisons between rodents and prokaryotes are not significant in **Fig. 6.3a**.

The biological significance and reason of this acceptance rate difference is interesting to explore. Each type of amino acid change can happen at many sites in the proteome of a species. Since each site has virtually unique environment (adjacent amino acid residues, interacting ligands or nucleic acids, interacting proteins, physiochemical micro-environment), it is intuitive to reason that the same type of change happening at different sites may have different acceptance rates. Nonetheless, each item in the $\boldsymbol{\omega}$ vector can be considered as an average acceptance rate summarized from all changes of one type across the genome. Since every $\boldsymbol{\omega}$ is the average within a clade, it is not trivial to explain why $\boldsymbol{\omega}$'s are different among each other: Consider I-to-L mutations happen at hundreds of sites in great apes and hundreds in rodents. Even if the overlap between the two sets of changes is minimal, certain non-random factor should still exist as a cause of the difference between the two acceptance rates averaged across these sites scattered in

the proteome of great apes and rodents. Given the observation that acceptance rate correlation is higher for closely related eukaryotes, there seems to be a genome-wide, clade-specific factor affecting the acceptance rates. One additional argument for this statement is that different parts of the genome in the same clade have higher $\boldsymbol{\omega}$ correlation than orthologous parts of the genome in different clades (see last part of RESULTS). These being clear, we currently still have no clue of such kind of factor, specifically defined for each pair of amino acids. The codon substitution model we use is time-reversible (Yang et al. 1998), assuming the codon frequencies in each clade have reached equilibria. There exists the possibility that the inference is inaccurate for real data because equilibrium is not reached. However, when we calculate the Euclidean distance between codon frequencies of the two species in one clade, we found that in all 68 clades, the true distance is no larger than a bootstrapped sample of the same alignment (**Fig. A.4.4**). In fact, the variation of codon frequencies within a clade is on average less than 5%. Hence it is unlikely that disequilibrium cause the acceptance rate difference.

Different amino acid acceptance rates have been discovered between different pairs of amino acids and between different gene categories in the genome (Jones et al. 1994; Yang et al. 1998; Castellano et al. 2009). However, the species- or clade-specific selection effect of amino acid substitutions has never been explicitly explored. Yet if true, our finding would suggest that the assumption of the current codon substitution models cannot hold any more, that $\boldsymbol{\omega}'$ does not change across the tree. This heterogeneity of acceptance rates across difference groups of species could be important, since we know that different substitution models used may lead to different conclusions in molecular evolution or phylogenetic analyses (Bruno and Halpern 1999; Keane et al. 2006; Yang 2006; Zou and Zhang 2015). As a finer-scale update on our understanding of the coding sequence evolution process, the amino acid acceptance rates difference between species

clades may be an interesting biological phenomenon, as well as a potentially important indication to the current protein evolution models, hence requiring further investigation on its biological reason and methodological significance.

6.5 MATERIALS AND METHODS

Sequence data acquisition and alignment

Sequence data used in this study are retrieved from different sources listed in **Table A.4.1**. Coding sequence alignments of four mammalian clades, the fruit flies and the yeasts were directly retrieved from respective databases. For other eukaryotic clades retrieved from Ensembl, we query a list of one-to-one orthologous genes for the two species and download their coding sequences. Then the coding sequences are translated into protein sequences using MACSE v1.02 (Ranwez et al. 2011). Local pairwise protein sequence alignments were conducted for each pair of orthologs by MAFFT v7.294b (Katoh and Standley 2013) using the L-INS-i algorithm. Alignments of coding sequences were then derived by substituting amino acids with corresponding codons by custom Python script. All prokaryotic clades were sampled from the strains available in the ATGC database (Novichkov et al. 2009). All the above derived CDS alignments were then filtered so that no gaps, missing data or ambiguous codons exist.

Inference of acceptance rates

The inference of $\boldsymbol{\omega}$ is conducted by codeml program in PAML 4.9c (Yang 2007). The inference run under a user tree of only two species, codon frequencies for each individual codon, no clock, model 0 for coding sequence (one ω), NSsites = 0, fixed alpha. Besides, omega and kappa are not fixed, and control parameter aaDist is set as 7 to infer individual ωs. Since only

codon changes involving one nucleotide substitution is considered to have non-zero rate in the rate matrix, there are 75 amino acid pairs that can interchange with this kind of one-step substitution. Hence the $\boldsymbol{\omega}$ vector has 75 items. In **Fig. 6.2** and **Fig. 6.3b**, the $\boldsymbol{\omega}$ of each clade is the average of 9 inferred $\boldsymbol{\omega}$s from replicate runs of codeml.

Simulating coding sequence evolution

All simulations follow the codon substitution model specified by equation (1). To simulate a clade with a pair of species, a transition matrix P of codons (61×61) is first derived. For each pair of codon, the instant rate of substitution q is set as in equation (1). The resulted rate matrix Q is normalized to have a total rate of 1, and the transition matrix $P = e^{Qt}$ (Yang 2006). For each codon to be simulated, an ancestral codon is randomly sampled according to the equilibrium codon frequencies, then this codon is evolved under a Markov process, based on the genetic distance of evolution and the matrix P, separately to derive two descendant codons for respective species. For simulations in **Fig. 6.1**, 20 replicate simulations were conducted for each parameter set, and in each simulation, the CDS alignment has one million codons with genetic distance = 0.1 substitution per site between two species. No site-specific variation of evolutionary rate is assumed unless mentioned. When site-specific relative evolutionary rate is specified, this rate is multiplied with genetic distance. The relative rates are sampled from an exponential distribution with mean equal to 1. For each shuffling test in **Fig. 6.3a**, **Fig. 6.3b**, **Fig. A.4.1-3**, 100 independent shuffling of the original alignments were conducted. In **Fig. A.4.3** when individual ωs are subject to variation, each ω is multiplied by a factor sampled from a Gamma distribution with mean = 1 and standard deviation = 0.05, 0.1, 0.2 or 0.5, respectively.

Shuffling test

For two clades A and B, each clade is represented by a coding sequence alignment, with two sequences from two species (species 1 and species 2) aligned. For each particular codon, we find all positions in each alignment that have this codon in species 1 and have nonsynonymous substitution, i.e. species 1 and 2 have difference codon. Next, we shuffle these codon positions jointly in A and B. For example, an CAC-CAG site in A would be subject to shuffling together with a CAC-UAC site in B, because they respectively codes for an H-Q / H-Y site at protein level. During this shuffling process, since nonsynonymous variable sites are of small number compared with the total number of sites in the alignments, the equilibrium codon frequency in each clade will not be significantly changed. Only shuffling variable sites ensures the original within-clade genetic distances are largely unchanged.

REFERENCES

Bruno WJ, Halpern AL. 1999. Topological bias and inconsistency of maximum likelihood using wrong models. Mol Biol Evol 16:564-566.

Castellano S, Andres AM, Bosch E, Bayes M, Guigo R, Clark AG. 2009. Low exchangeability of selenocysteine, the 21st amino acid, in vertebrate proteins. Mol Biol Evol 26:2031-2040.

Dayhoff MO, Schwartz RM, Orcutt BC. 1978. A model of evolutionary changes in proteins. In: Dayhoff MO, editor. Atlas of protein sequence and structure. Silver Spring, MD: National Biomedical Research Foundation. p. 345-352.

Goldman N, Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 11:725-736.

Grantham R. 1974. Amino acid difference formula to help explain protein evolution. Science 185:862-864.

Jones DT, Taylor WR, Thornton JM. 1994. A mutation data matrix for transmembrane proteins. FEBS Lett 339:269-275.

Jones DT, Taylor WR, Thornton JM. 1992. The Rapid Generation of Mutation Data Matrices from Protein Sequences. Comput Appl Biosci 8:275-282.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772-780.

Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, Kanehisa M. 2008. AAindex: amino acid index database, progress report 2008. Nucleic Acids Res 36:D202-205.

Keane TM, Creevey CJ, Pentony MM, Naughton TJ, McInerney JO. 2006. Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. Bmc Evol Biol 6.

Le SQ, Gascuel O. 2008. An improved general amino acid replacement matrix. Mol Biol Evol 25:1307-1320.

Miyata T, Miyazawa S, Yasunaga T. 1979. Two types of amino acid substitutions in protein evolution. J Mol Evol 12:219-236.

Novichkov PS, Ratnere I, Wolf YI, Koonin EV, Dubchak I. 2009. ATGC: a database of orthologous genes from closely related prokaryotic genomes and a research platform for microevolution of prokaryotes. Nucleic Acids Res 37:D448-454.

Rand DM, Weinreich DM, Cezairliyan BO. 2000. Neutrality tests of conservative-radical amino acid changes in nuclear- and mitochondrially-encoded proteins. Gene 261:115-125.

Ranwez V, Harispe S, Delsuc F, Douzery EJ. 2011. MACSE: Multiple Alignment of Coding SEquences accounting for frameshifts and stop codons. PLoS One 6:e22594.

Stoltzfus A, Norris RW. 2016. On the causes of evolutionary transition:transversion bias. Mol Biol Evol 33:595-602.

Stoltzfus A, Yampolsky LY. 2007. Amino acid exchangeability and the adaptive code hypothesis. J Mol Evol 65:456-462.

Tang H, Wu CI. 2006. A new method for estimating nonsynonymous substitutions and its applications to detecting positive selection. Mol Biol Evol 23:372-379.

Tang H, Wyckoff GJ, Lu J, Wu CI. 2004. A universal evolutionary index for amino acid changes. Mol Biol Evol 21:1548-1556.

Whelan S, Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol 18:691-699.

Yampolsky LY, Stoltzfus A. 2005. The exchangeability of amino acids in proteins. Genetics 170:1459-1472.

Yang Z. 2006. Computational molecular evolution. Oxford: Oxford University Press.

Yang Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol 15:568-573.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24:1586-1591.

Yang Z, Nielsen R, Hasegawa M. 1998. Models of amino acid substitution and applications to mitochondrial protein evolution. Mol Biol Evol 15:1600-1611.

Zhang J. 2000. Rates of conservative and radical nonsynonymous nucleotide substitutions in mammalian nuclear genes. J Mol Evol 50:56-68.

Zou Z, Zhang J. 2015. Are convergent and parallel amino acid substitutions in protein evolution more prevalent than neutral expectations? Mol Biol Evol 32:2085-2096.

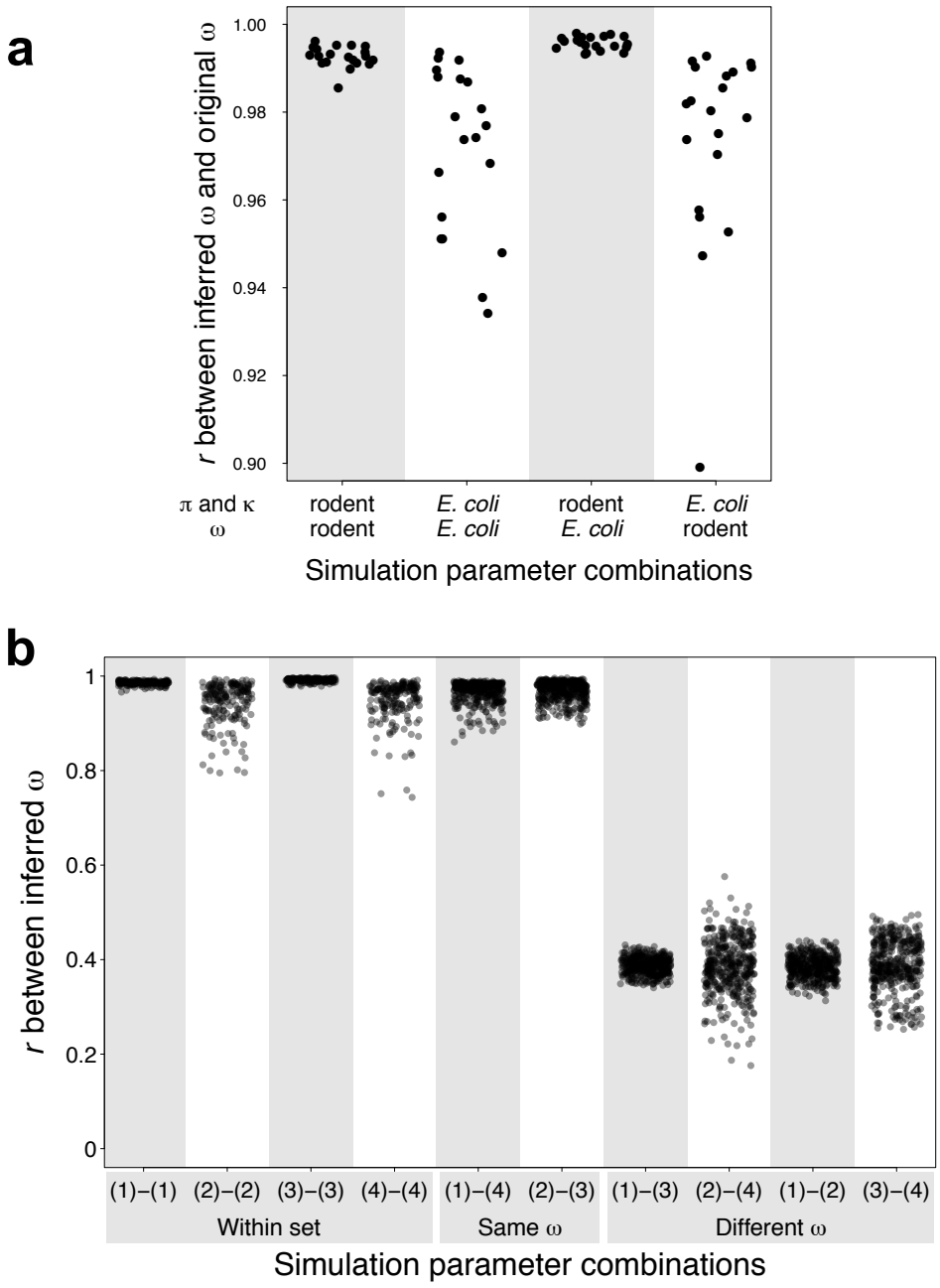


Figure 6.1 Simulated sequence alignments of species clades confirm the accuracy of ML acceptance rate (*ω*) inference. (**a**) Pearson correlation between inferred ***ω*** and the true value used for simulation. The source of parameters for each column is labeled below X axis. There are 20 replicate simulations in each column. (**b**) Pearson correlation between inferred ***ω***s of different parameter sets. X axis label for each column indicate the two parameter sets that are compared, corresponding to the main text and column 1 – 4 in (**a**). In the first four columns, there are 190 *r*'s for each pair of replicates in the same parameter set. In the other columns, there are 400 *r*'s plotted for each pair of replicates in two different parameter set. Whether the two inferred ***ω***s

correlated in the column come from simulations under the same $\boldsymbol{\omega}$ is also indicated below X axis. “Within set” denote comparisons between different replicates of the same simulation parameter set.

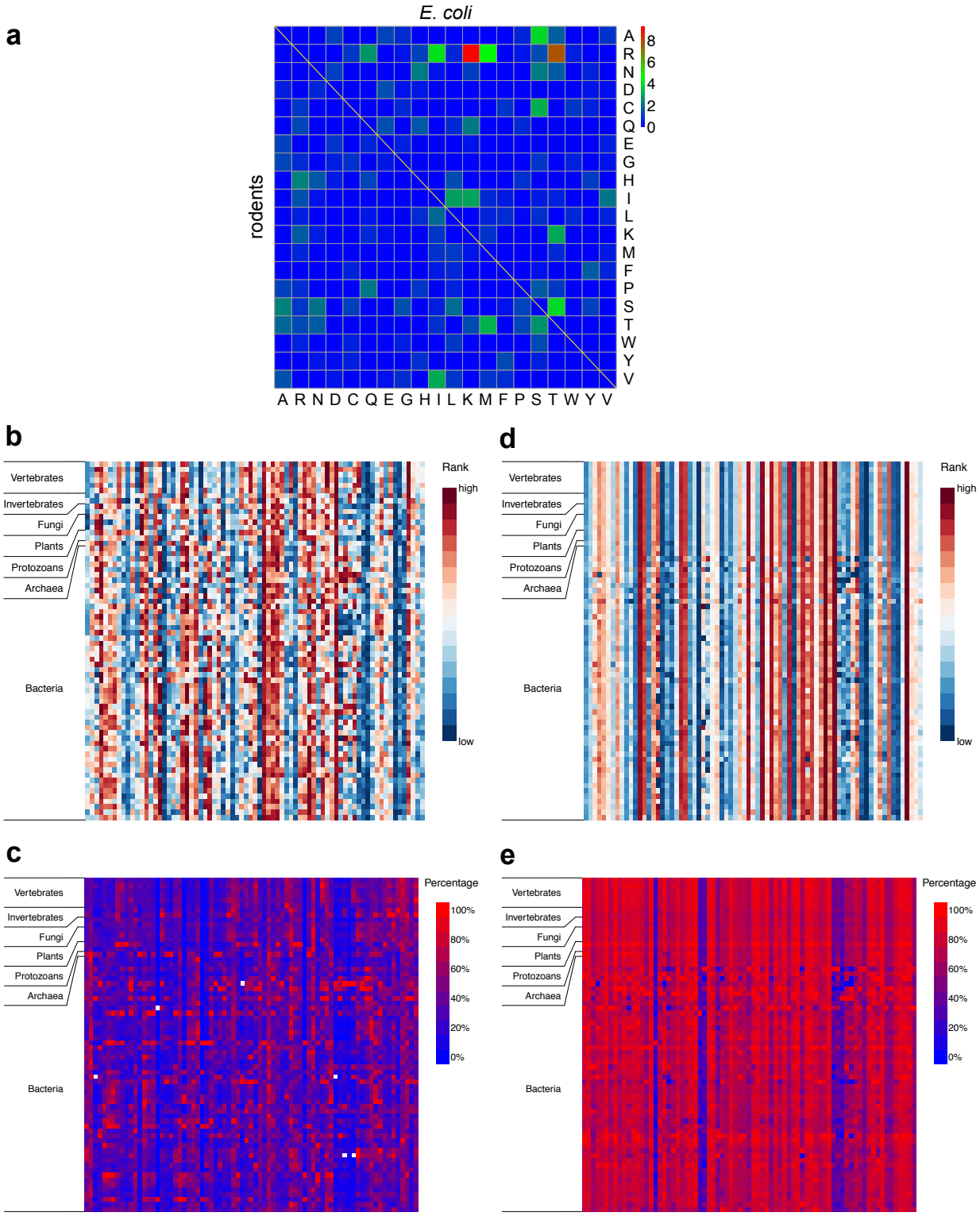


Figure 6.2 Inferred ω's for different real clades show different patterns (a - c), while inferred ω's for clades simulated with the same set of acceptance rates show highly correlated patterns (d, e). (**a**) Inferred acceptance rates for the rodents clade and *E. coli* clade.

The acceptance rates are formed into triangle matrices for all amino acid pairs, with the *E. coli* $\boldsymbol{\omega'}$ forming the upper triangle, rodents $\boldsymbol{\omega'}$ forming the lower triangle and zeros filling the diagonal items. Within each triangle, 75 cells are non-zero, while other 115 amino acid pairs have zero acceptance rates consistent with the codon substitution model. (**b**) Ranks of each ω'_{ij} among the 75 non-zero ω's for the same clade in real clades. (**c**) Percentage of each ω'_{ij} relative to the largest ω' value in the same category in real clades. (**d**) Ranks of each ω'_{ij} among the 75 non-zero ω's for the same clade in simulated clades. (**e**) Percentage of each ω'_{ij} relative to the largest ω' value in the same category in simulated clades. In (**b**) and (**d**), red color indicating large acceptance rate (high rank) and blue color indicating small value (low rank) according to the scale. In (**c**) and (**e**), red color indicating high percentage, and blue indicate low percentage. Each row represents a clade and each column is an ω'_{ij} category in (**b - e**), so each cell is a particular non-zero ω'_{ij} in a specific clade.

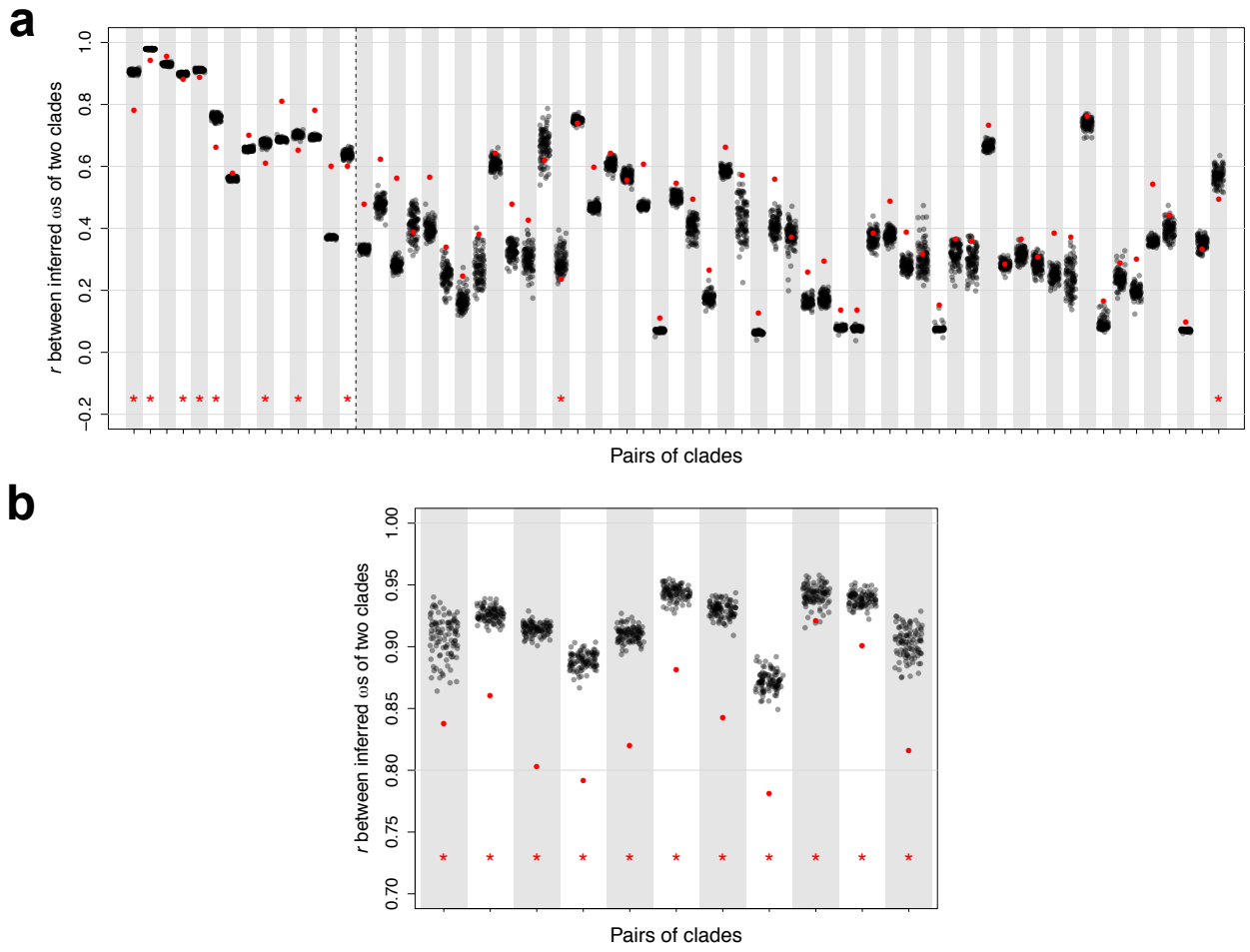


Figure 6.3 Shuffling tests (a) between the rodents clade and the other 67 clades (b) between the human – chimpanzee clade and other 11 mammalian clades. Each column is a shuffling test. Red dots indicate the Pearson correlation coefficients between ωs of two clades. 100 grey dots represent *r*'s between ωs of two clades after 100 independent alignment shuffling. If a red dot falls within 5% lower tail of the grey dots distribution, i.e., there being no more than four grey dots below it, a red asterisk is plotted to indicate significant smaller acceptance rate correlation than shuffled control.

CHAPTER 7

Are Nonsynonymous Transversions more Deleterious than Nonsynonymous Transitions during Coding Sequence Evolution?

7.1 ABSTRACT

Transitions typically have higher substitution rates than transversions in coding sequence evolution. This phenomenon has several potential causes. First, the mutation rate is higher for transitions than transversions. Second, transitional mutations are more likely than transversional mutations to be synonymous and hence have a higher rate of fixation. Third, it has been suggested that transitional nonsynonymous mutations are more likely than transversional nonsynonymous mutations to conserve amino acid physicochemical properties and so have a higher rate of acceptance. This third possibility was recently challenged by that no detectable difference in fitness effects between transitional and transversional nonsynonymous mutations in large mutagenesis experiments. However, due to the limited sensitivity of laboratory measures of fitness effects, we used evolutionary data to revisit this issue. We modified an existing codon model of sequence evolution by adding a new parameter η, which is the ratio between the fixation probability of a transitional nonsynonymous mutation and that of a transversional nonsynonymous mutation. Using a likelihood estimator of η, we examined genome-wide concatenated alignments of coding sequences from many species pairs across the tree of life. Surprisingly, η varies widely from smaller than 1 to greater than 1. Thus, in some species, transitional nonsynonymous mutations are more deleterious than transversional nonsynonymous mutations, but the opposite is true in some other species. Our extensive searches reveal that this diversity may arise from variable amino acid acceptance rates across the tree of life.

7.2 INTRODUCTION

During nucleotide sequence evolution, substitutions between two purines (A - G) or between two pyrimidines (C - T) are transitions, while those between the two categories are transversions. Since there are four types of transitions and eight types of transversions, ratio of observed transitions versus transversions (Ts/Tv) is expected to be 0.5 if we assume all types of nucleotide substitution have the same occurrence rate. In practice, a transition/transversion bias has been widely reported, that transitional substitutions tend to happen at a higher rate than transversional substitutions (Yang 2006). The effect of this bias has been incorporated into most major substitution models for describing nucleotide sequence evolution, from the two-parameter K80 model to mechanistic codon evolution models (Kimura 1980; Hasegawa et al. 1984; Tamura and Nei 1993; Goldman and Yang 1994; Yang et al. 1998).

Substitutions are practically fixed differences between orthologous sequences of existing species. Whether a substitution happens in coding nucleotide sequence is determined by both the mutation process and the selection on this mutation afterwards. Hence, the transition/transversion bias we observe can be caused by bias in either step. The mutational bias towards transitions has been reported among spontaneous mutations in mutation accumulation (MA) lines of *Drosophila melanogaster* (Haag-Liautard et al. 2008; Schrider et al. 2013), baker's yeast (Lynch et al. 2008; Zhu et al. 2014), *Arabidopsis thaliana* (Ossowski et al. 2010). Similar biases were found among intergenic, noncoding or synonymous single nucleotide polymorphisms assumed to be free of selection, e.g. in human populations (Freudenberg-Hua et al. 2003; Jiang and Zhao 2006), bacteria clones (Hershberg and Petrov 2010) and natural *Caenorhabditis elegans* populations (Cutter 2006). Inference under the HKY model based on pairs of mammalian species also suggest high bias towards transition in four-fold degenerate sites even when all CpG

hypermutable sites are excluded (Rosenberg et al. 2003). Notably, there are cases where no significant deviation of Ts/Tv from 0.5 is observed (Keller et al. 2007; Denver et al. 2009).

While the mutational bias towards transition is confirmed by above studies, it has been repeatedly mentioned that selection is also biased between transition and transversion. For example, in Zhu et al. (2014), an elevated Ts/Tv of 0.95 is observed from 867 spontaneous mutations in yeast MA lines, but the ratio is much higher (Ts/Tv = 2.96) for the fixed substitutions between the MA ancestral line and the yeast reference genome. This suggest that average fixation probability is higher for transitions compared with transversions. The same pattern is also suggested in *Caenorhabditis elegans* (Denver et al. 2009) and *Drosophila melanogaster* (Haag-Liautard et al. 2008). For coding sequences, the selection bias towards transition may be two-fold. First, because synonymous mutations are expected to be less deleterious than nonsynonymous ones, they are more likely to be retained. Hence if transitional mutations are more likely to be synonymous, Ts/Tv would increase. This has been observed in spontaneous mutations (Schrider et al. 2013), polymorphism data (Freudenberg-Hua et al. 2003) and evolutionary alignment data (Zhang 2000). The second layer of selection bias could be that nonsynonymous transitions are less deleterious than nonsynonymous transversions, thus more likely to be retained as fixed substitutions. If coding sequence evolution is a largely neutral process, radical nonsynonymous changes between physiochemically distinctive amino acids are expected to be more deleterious than conservative changes. Categorizing all 20 amino acids into different physiochemical classes, Zhang (2000) observed that nonsynonymous transitions are more likely to cause amino acid changes within the same class, i.e. conservative changes, than transversions, this supporting the existence of second layer selection bias. Freudenberg-Hua et al. (2003) classified amino acid changes as radical or conservative according to the Grantham

physiochemical distance (Grantham 1974) and showed the same trend in human polymorphism data. While these studies support the existence of a selection advantage of nonsynonymous transitions against transversion, Stoltzfus and Norris (2016) argued against this bias. Based on eight datasets containing 1,239 nonsynonymous mutations and their fitness effects measure in mutagenesis studies, the authors showed that classification of transition versus transversion has no power in predicting the fitness effect of a nonsynonymous mutation, compared with other classifications. Whether the second layer of selection bias towards transition is thus requiring further investigation.

Here, we point out that the classification of transition versus transversion cannot be seen by selection. Both classes contain a mixture of different types of amino acid changes that are directly acted on by selection, and the contribution of each type to the overall selection effect on transition or transversion depends on their frequencies, which is variable from species to species and from datasets to datasets. For example, the set of nonsynonymous transitions in mutagenesis studies may contain different proportions of amino acid changes from a set of fixed substitutions between species. Thus, knowing that nonsynonymous mutations in mutagenesis show certain transition/transversion bias does not indicate same bias in the true evolution history. Consequently, to ask whether nonsynonymous transitions are on average less deleterious than nonsynonymous transversions, we have to analyze evolutionary sequence data. In this study, we incorporate the selection bias between nonsynonymous transition and transversion as an independent parameter in a commonly used codon substitution model. We derived sequence data from a sample of clades consisting of closely related species across the tree of life, and infer the selection bias parameter under maximum likelihood framework. We found that the selection bias is not always towards nonsynonymous transition, but sometimes towards transversion.

Simulations of sequence evolution suggests that a major factor causing this variable bias direction is the difference of amino acid acceptance rates among different species clades.

7.3 RESULTS

ML inference of the selection bias η show different patterns among clades.

To parameterize the effect of nonsynonymous transition/transversion bias, we modify the Markov codon substitution model in Goldman and Yang (1994) by adding a parameter η as an additional factor for the rate of nonsynonymous transitions.

$$q_{uv} = \begin{cases} 0, & \text{if the two codons differ at more than one position,} \\ \pi_v, & \text{for synonymous transversion } (VS), \\ \kappa\pi_v, & \text{for synonymous transition } (IS), \\ \omega\pi_v, & \text{for nonsynonymous transversion } (VN), \\ \eta\omega\kappa\pi_v, & \text{for nonsynonymous transition } (IN), \end{cases} \quad (1)$$

In this case, η represents the nonsynonymous Ts/Tv normalized by synonymous Ts/Tv, i.e. $(IN/VN)/(IS/VS)$. If nonsynonymous transitions have neither selection advantage nor disadvantage upon nonsynonymous transversion, we expect η to be 1. Values higher than 1 indicate selection bias towards nonsynonymous transitions and vice versa. Since the original codon substitution model has been incorporated in the codeml program from PAML (Yang 2007), we modified the source code and realized maximum likelihood (ML) estimation of parameters in the above model with η (program named codemlz, see MATERIALS AND METHODS).

To verify the accuracy of this ML inference framework, we simulated sequence evolution under the same model as described above to generate a pairwise sequence alignment of two species (a clade) per simulation. A series of η values ranging from 0.3 to 2.5 were used to

conduct multiple simulations. To confirm that values of other parameters cannot affect η inference, i.e. η is independent, we also vary the genetic distance between the two species, the mutational transition/transversion bias κ or the overall selection effect ω. The inferred ηs show neither deviation from the underlying true value in simulation, nor correlation with the other varied parameters (**Fig. 7.1**). For example, in **Fig. 7.1a**, when true η is 1.2 in simulation, the six estimated η values have a mean of 0.902 with standard deviation of 0.020. Pooling all η estimations together, one sample *t*-test show no significant deviation of the estimated η from true value (denote as η_0, $P = 0.09$). No correlation could be found between the deviation $\eta - \eta_0$ with the true genetic distance between species (Spearman's $\rho = -0.13$, $P = 0.34$). The same pattern of no correlation is true for various κ values (**Fig. 7.1b**) and ω values (**Fig. 7.1c**), indicating the inference process is accurate when model assumptions are true.

Given this ML inference method, we want to investigate the selection bias in different species across the tree of life. We sampled a set of 68 available species clade that is widely distributed across the tree of life, with coding sequence data available from different sources (see **Table. S1**). Each clade includes a pair of closely related species, and equilibrium of protein sequence evolution is assumed within clade. There are 15 eukaryotic clades, including six pairs of vertebrates, two pairs of insects, two pairs of fungi, three pairs of plants and two pairs of protozoans belonging to outgroups of above groups. 53 prokaryotic clades including one pair of archaea and 52 pairs of bacteria were also sampled. With the genome-wide concatenated coding sequence alignments available for each clade (see MATERIALS AND METHODS), we respectively estimated ηs by the codemlz program for each clade. The resulted pattern is rather surprising (**Fig. 7.2**). Among both eukaryotes and prokaryotes, there are clades with η higher than 1 and lower than 1, meaning that the nonsynonymous transition/transversion bias exists, and

are different among different clades of species. For example, between two ant species *Atta cephalotes* and *Solenopsis invicta*, η is estimated to be 0.54, which means a nonsynonymous mutation has only half the probability to be fixed if it is a transition, compared with when it is a transversion, i.e. transitions are more deleterious than transversions in this clade. On the other hand, for the clade containing two malaria pathogen *Plasmodium vivax* and *Plasmodium knowlesi*, the inferred η is 2.0, indicating nonsynonymous transitions are twice as likely to be fixed as nonsynonymous transversion. Among 68 clades, only seven were inferred to have no significant bias (η is not significantly different from 1). Nonsynonymous transitions are less deleterious in 27 clades (η>1) and more deleterious in 34 clades (η<1).

Among-clade variation of amino acid acceptance rates can generate comparable η difference as observed

Next, we want to investigate the mechanism of the observed η difference among clades, especially what drives the η variation from lower than 1 to higher than 1. As mentioned in the introduction, the fixation probability (d_N/d_S) of a transition is a mean effect of selection on all corresponding nonsynonymous codon changes. To describe factors in this process, a modified version of a general codon model proposed in Yang et al. (1998) is used here:

$$q_{uv} = \begin{cases} 0, & \text{if the two codons differ at more than one position,} \\ \pi_v, & \text{for synonymous transversion,} \\ \kappa\pi_v, & \text{for synonymous transition,} \\ \omega'_{ij}\omega_0\pi_v, & \text{for nonsynonymous transversion,} \\ \omega'_{ij}\kappa\omega_0\pi_v, & \text{for nonsynonymous transition,} \end{cases} \quad (2)$$

In this equation, q_{uv} is the rate of substitution between codon u and codon v; π_v is the equilibrium frequency of the resulted codon v, summarized into vector $\boldsymbol{\pi}$; κ is the

transition/transversion ratio for synonymous substitutions, thus the mutational bias; ω_0 is the overall selection strength, i.e. d_N/d_S affecting all nonsynonymous mutations; ω'_{ij} is the relative acceptance rate between amino acid i and j, which are respectively encoded by codon u and v. The acceptance rate for a pair of amino acid is the fixation probability of a nonsynonymous mutation causing changes from one amino acid to the other. This set of parameter ($\boldsymbol{\omega}'$) can be either derived from empirical amino acid similarity measure, e.g. the physiochemical Grantham distance (Grantham 1974), or directly inferred by maximum likelihood (Yang et al. 1998). To identify the cause of observed η difference among clades, we simulate sequence evolution with this "variable acceptance rate" model with different parameter settings.

First, we simulated with a series of κ values while keeping other parameters constant, and then inferred η from the resulted sequence alignments. By plotting the estimated η's against the corresponding true κ's in the simulation, we could observe a significant negative correlation between κ and η (Spearman's ρ = -0.28, P = 0.006, **Fig. 7.3a**). Nevertheless, with κ changing from 1.0 to 5.5, the simulated clades show η's no higher than 0.83 and no lower than 0.77. Among 68 clades in real data, 61 clades have κ within the range [1.0, 5.5], but η estimates vary far beyond [0.77, 0.83] (**Fig. 7.2**). Hence the different κ among clades alone cannot explain the large variation of η. Similar pattern was observed when ω_0 is varied while other parameters are kept constant (**Fig. 7.3b**). Despite significant positive correlation between ω_0 and η (Spearman's ρ = 0.77, P = 4E-25), estimated η merely ranges from 0.76 to 0.86 when ω_0 varies within [0.01, 0.70]. Thus, overall d_N/d_S variation among clades is not likely to be sufficient reason of the observed η variation. Next, we simulated the 53 prokaryotic clades based on their own codon frequencies with other parameters kept identical among clades. Among these simulated clades, η estimation ranges from 0.70 to 0.93 (**Fig. 7.3c**), still much smaller than those estimated from real

data (0.31 – 1.57). Besides, no positive correlation was found between η's estimated from the simulated clades and those inferred from real clades (Pearson's r = -0.23, P = 0.08). These results indicate that different $\boldsymbol{\pi}$'s among clades cannot drive the observed η variation.

Given these negative findings, we investigate the last component in the model, relative amino acid acceptance rates $\boldsymbol{\omega}'$. First, the inverse of corresponding Grantham distance between a pair of amino acid is used as a reference value for the relative acceptance rate, forming the matrix $\boldsymbol{\omega}'$. Then a series of varied $\boldsymbol{\omega}'$s were generated through randomly increasing or decreasing each item by a certain percentage. Keeping κ, ω_0 and $\boldsymbol{\pi}$ the same, we conducted simulations of sequence evolution with these varied $\boldsymbol{\omega}'$s, to show the effect of variation in acceptance rates. Interestingly, we found that when $\boldsymbol{\omega}'$ is varied by 60% or more, the simulated clades have large η variation, ranging from 0.4 to 1.3 (**Fig. 7.4a**). This level of variation closely matches what we observed in real data, and importantly, η>1 is observed for simulated clades, meaning variation of $\boldsymbol{\omega}'$ can cause the nonsynonymous transition/transversion bias to change direction, from transition being more deleterious to transversion being more deleterious (**Fig. 7.2**). Furthermore, we used the codeml program in PAML to infer the amino acid acceptance rates, essentially $\omega_0 \cdot \boldsymbol{\omega}'$, from each of the 68 clades (see MATERIALS AND METHODS), and simulated 68 clades with corresponding acceptance rates while setting κ and $\boldsymbol{\pi}$ universal among clades. Surprisingly, the η's estimated from the simulated clades are highly matched with those estimated from the real clades (Pearson's r = 0.95, P = 4E-34, **Fig. 7.4b**), even though all simulated clades evolved with the same set of κ and $\boldsymbol{\pi}$. These findings with $\boldsymbol{\omega}'$ strongly indicate that the relative acceptance rates difference among clades is the most important underlying mechanistic cause for the observed η variation.

7.4 DISCUSSION

Being functionally important, coding sequence evolution is probably under the most direct and strongest selection force in the genome during the evolutionary history, while mutational factors prevailing the genome also affect this process. To understand protein evolution and disentangle the multiple layers of forces that shaped the currently observed coding sequences, it is important to apply realistic mechanistic models. In this study, we first formulate the debated nonsynonymous transition/transversion bias into a model framework and established reliable maximum likelihood inference of the bias parameter η. Then the bias was inferred in a collection of 68 species clades that spans different domains of life. Surprisingly, we found the whether nonsynonymous transitions are more deleterious than nonsynonymous transversion is clade-specific, indicating that there exists features unique to each clade that shapes the value of η. This surprising result is actually natural, considering that the classification of substitutions into transitions and transversions is at nucleotide level, hence not directly acted on by selection in protein evolution. Each category consists of many different amino acid changes. For example, considering single nucleotide substitutions under standard codon table, transitions contain changes between histidine and arginine (CAT/C – CGT/C), between valine and alanine (GTN – GCN), etc., while transversions include changes between threonine and lysine (ACA/G – AAA/G), between leucine and arginine (CTN – CGN), etc. The fixation probability of a nonsynymous transition or transversion is actually an average effect of the fixation probability of the corresponding codon changes (e.g. CAT to CGT, GCA to GTA etc. for transition). Hence, we use the model described by equation (2) as the mechanistic model to disentangle the effect of each factor on shaping the patterns of η. By changing the value of focal factor while controlling

other variables in simulations of sequence evolution, we found that variable transition/transversion mutational bias κ, overall selection strength ω_0 cannot explain the large range of η variation we observed in real data. Nor are clade-specific codon frequencies $\boldsymbol{\pi}$ able to reproduce the η variation among clades. Interestingly, we found when the fixation probabilities $\boldsymbol{\omega}'$, i.e. relative acceptance rates of different amino acid changes vary, the resulted η can show different directions of the bias in different clades. We then showed that when clade-specific $\boldsymbol{\omega}'$s are inferred and used in simulation, η's inferred from simulated clades match well with those inferred from real clades. Thus, among all factors included in the mechanistic model, only clade-specificity of relative acceptance rates can explain the different nonsynonymous transition/transversion biases we observed among 68 clades.

One potential restriction of our model-based approach is that the model cannot fully reflect the complexity of the actual sequence evolution process. In our codon model inference process, a strong assumption is that site-specific rate distribution is uniform, or there is no evolutionary rate variation among different codon sites. However, actual rate heterogeneity among sites has been reported to be fairly large (Yang 2006). To check if this model simplification can affect our conclusions, we simulated sequence evolution under both equation (1)("η model") and equation (2)("variable acceptance rate model") with site-specific evolutionary rates following exponential distribution. Indeed, simulations under the η model show that estimated η can be higher than the real value (**Fig. A.5.1a-c**), and there is apparent positive correlation of the estimation with within-clade genetic distance (Spearman's ρ = 0.75, P = 7E-11, **Fig. A.5.1a**) and κ (Spearman's ρ = 0.34, P = 0.001, **Fig. A.5.1b**), and negative correlation with ω (Spearman's ρ = -0.74, P < 1E-300, **Fig. A.5.1c**). However, the deviation of η estimation from true values is small (less than 0.2) in these simulations, hence not likely to be the

major cause of the observed η>1 cases. Furthermore, η of the 68 real clades are not correlated with within-clade genetic distance (Spearman's ρ = -0.19, P = 0.12), negatively correlated with κ (Spearman's ρ = -0.67, P <1E-300), and not correlated with ω (Spearman's ρ = 0.12, P = 0.33). Hence, the actual patterns in real clades at least are not caused by the biases above. By simulations under the "variable acceptance rate model", we found that the previous patterns in **Fig. 7.3** and **Fig. 7.4a** are not affected. When site-specificity exists for evolutionary rates, variation of κ, ω_0 and clade-specific codon frequency still cannot reproduce clades with the η range observed in **Fig. 7.2** (**Fig. A.5.2a-c**). Clades with varying acceptance rates, on the other hand, still show varying η's ranging from 0.5 to 1.4 (**Fig. A.5.2d**). Thus, we reckon the η inference framework is sufficient to support our main conclusions.

Despite this limitation, our model based inference method has advantages. First, the ML inference can be easily applied to genome-wide coding sequence datasets, hence is suitable for investigating the pattern in many different clades of species. Second, modeling the codon substitution model can disentangle different factors affecting the sequence evolution process, so that the result is less likely to be artefact cause by, for example, mutational bias. Although Stoltzfus and Norris (2016) claimed that each of the mutagenesis studies they analyze has the power to distinguish between fitness effect of conservative versus radical mutations, the selection bias between transitions and transversion might be a smaller effect, hence not resolvable comparison between 544 transitions and 695 transversions. Furthermore, the selected studies in this meta-analysis may represent a small group of species that by chance have weak nonsynonymous transition/transversion bias. The statement about nonsynonymous transitions being more conservative in (Zhang 2000) were derived from counting the proportion conservative changes among all transitions and transversions, based on codon frequencies from

47 genes in three mammalian species. Indeed, we observed η>1 in all mammalian clades we have (**Fig. 7.2**, first four clades), but our ML model contains more factors than codon frequencies, and found different bias direction with a broader sampling of species.

Since transition versus transversion is a nucleotide-level classification unseen by selection, our finding that different categories are less deleterious in different species is not unexpected considering the mechanism. Nevertheless, the mechanistic explanation of the pattern is itself interesting, because the relative acceptance rates of amino acid pairs are essentially the average fixation probability of a certain amino acid change across the genome. Each type of amino acid change may occur in different proteins and at different protein domain structures, so should intuitively show various fitness effects and fixation probabilities. However, the genome-wide acceptance rate is an average across many sites, yet there is clade-specificity. Whether this clade-specificity of amino acid acceptance rate truly exists is intriguing, and has been investigated by Zou and Zhang (unpublished). As we are taking steps to further understand the process of protein sequence evolution, these findings together argue for the application of more realistic and mechanistic models in molecular evolution and phylogenetic studies.

7.5 MATERIALS AND METHODS

Modification of the codeml program

We modified the codeml program in PAML 4.8 and name the modified version as codemlz. The goal and function of codemlz is limited to the following model settings according to the control file of the original codeml program: seqtype = 1, clock = 0, model = 0, NSsites = 0, Mgene = 0, fix_alpha = 1, alpha = 0. Two options are added to the control file for η estimation: "fix_eta" and "eta". Setting "fix_eta = 0" allow ML algorithm to infer η starting from the initial

value specified by "eta", while "fix_eta = 1" let the algorithm optimize with a fixed η value specified by "eta". Inferrd η value is output into the "mlc" file generated by the program. The codemlz program can be accessed as a GitHub repository (https://github.com/ztzou/codemlz).

Sequence data acquisition and alignment

Sequence data used in this study are retrieved from different sources listed in **Table A.5.1**. Coding sequence alignments of four mammalian clades, the fruit flies and the yeasts were directly retrieved from respective databases. For other eukaryotic clades retrieved from Ensembl, we query a list of one-to-one orthologous genes for the two species and download their coding sequences. Then the coding sequences are translated into protein sequences using MACSE v1.02 (Ranwez et al. 2011). Local pairwise protein sequence alignments were conducted for each pair of orthologs by MAFFT v7.294b (Katoh and Standley 2013) using the L-INS-i algorithm. Alignments of coding sequences were then derived by substituting amino acids with corresponding codons by custom Python script. All prokaryotic clades were sampled from the strains available in the ATGC database (Novichkov et al. 2009). All the above derived CDS alignments were then filtered so that no gaps, missing data or ambiguous codons exist.

Inference of η

The inference of the nonsynonymous transition/transversion bias η is conducted by codemlz. To ensure power of the ML inference, we run codemlz on the sequence alignment of a clade six times with different initial η values, from 0.2 to 10.0. Then the optimized likelihood of each run was compared with another run fixing $\eta = 1.0$ by a likelihood ratio test. Runs with test statistic $D \leq 0$ were discarded, and means of inferred parameters (genetic distance within a clade,

κ, ω, η) from other runs are used in downstream analyses and figures for this clade. For all 68 clades, genetic distance between two species are less than one nucleotide substitutions per codon. A user tree formed by the two species is used.

Simulating coding sequence evolution

Simulations in **Fig. 7.1** and **Fig. A.5.1** follow the codon substitution model specified by equation (1). Simulations in **Fig. 7.3**, **Fig. 7.4** and **Fig. A.5.2** follow equation (2). To simulate a clade with a pair of species, a transition matrix P of codons (61×61) is first derived. For each pair of codon, the instant rate of substitution q is set as in equation (1). The resulted rate matrix Q is normalized to have a total rate of 1, and the transition matrix $P = e^{Qt}$ (Yang 2006). For each codon position to be simulated, codon in one species is randomly sampled according to the equilibrium codon frequencies, then this codon is evolved under a Markov process, based on the genetic distance and the matrix P, to derive the codons in another species. For all simulations, the CDS alignment has 700,000 codons. When site-specificity of evolutionary rate is simulated, each site is assigned a relative rate randomly sampled from exponential distribution with mean = 1. This rate is multiplied with genetic distance during evolution. For simulations in **Fig. 7.1** and **Fig. A.5.1**, genetic distance = 0.1 substitution per site between two species. For simulations in in **Fig. 7.3**, **Fig. 7.4a** and **Fig. A.5.2**, except for the factor that is changing among simulations, other parameters (e.g. genetic distance within a clade, κ, ω_0) are set to the mean value across 53 prokaryotic clades, to ensure a realistic scenario. The amino acid relative acceptance rates $\boldsymbol{\omega}'$ are set to item-wise inverse of the Grantham matrix (Grantham 1974) when not varied. Simulation in **Fig. 7.4b** follow equation (2) while the acceptance rates $\omega_0 \cdot \boldsymbol{\omega}'$ was set as the values inferred by codeml with following settings: count codon frequencies for each individual codon; no clock;

model 0 for coding sequence (one ω); NSsites = 0; fixed alpha = 0; omega and kappa are not fixed; control parameter aaDist = 7 to infer individual ωs (see **CHAPTER 6**).

REFERENCES

Cutter AD. 2006. Nucleotide polymorphism and linkage disequilibrium in wild populations of the partial selfer *Caenorhabditis elegans*. Genetics 172:171-184.

Denver DR, Dolan PC, Wilhelm LJ, Sung W, Lucas-Lledo JI, Howe DK, Lewis SC, Okamoto K, Thomas WK, Lynch M, et al. 2009. A genome-wide view of *Caenorhabditis elegans* base-substitution mutation processes. Proc Natl Acad Sci U S A 106:16310-16314.

Freudenberg-Hua Y, Freudenberg J, Kluck N, Cichon S, Propping P, Nothen MM. 2003. Single nucleotide variation analysis in 65 candidate genes for CNS disorders in a representative sample of the European population. Genome Res 13:2271-2276.

Goldman N, Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 11:725-736.

Grantham R. 1974. Amino acid difference formula to help explain protein evolution. Science 185:862-864.

Haag-Liautard C, Coffey N, Houle D, Lynch M, Charlesworth B, Keightley PD. 2008. Direct estimation of the mitochondrial DNA mutation rate in *Drosophila melanogaster*. PLoS Biol 6:e204.

Hasegawa M, Yano T, Kishino H. 1984. A new molecular clock of mitochondrial-DNA and the evolution of hominoids. Proc Jpn Acad Ser B Phys Biol Sci 60:95-98.

Hershberg R, Petrov DA. 2010. Evidence that mutation is universally biased towards AT in bacteria. PLoS Genet 6:e1001115.

Jiang C, Zhao Z. 2006. Mutational spectrum in the recent human genome inferred by single nucleotide polymorphisms. Genomics 88:527-534.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772-780.

Keller I, Bensasson D, Nichols RA. 2007. Transition-transversion bias is not universal: a counter example from grasshopper pseudogenes. PLoS Genet 3:e22.

Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16:111-120.

Lynch M, Sung W, Morris K, Coffey N, Landry CR, Dopman EB, Dickinson WJ, Okamoto K, Kulkarni S, Hartl DL, et al. 2008. A genome-wide view of the spectrum of spontaneous mutations in yeast. Proc Natl Acad Sci U S A 105:9272-9277.

Novichkov PS, Ratnere I, Wolf YI, Koonin EV, Dubchak I. 2009. ATGC: a database of orthologous genes from closely related prokaryotic genomes and a research platform for microevolution of prokaryotes. Nucleic Acids Res 37:D448-454.

Ossowski S, Schneeberger K, Lucas-Lledo JI, Warthmann N, Clark RM, Shaw RG, Weigel D, Lynch M. 2010. The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. Science 327:92-94.

Ranwez V, Harispe S, Delsuc F, Douzery EJ. 2011. MACSE: Multiple Alignment of Coding SEquences accounting for frameshifts and stop codons. PLoS One 6:e22594.

Rosenberg MS, Subramanian S, Kumar S. 2003. Patterns of transitional mutation biases within and among mammalian genomes. Mol Biol Evol 20:988-993.

Schrider DR, Houle D, Lynch M, Hahn MW. 2013. Rates and genomic consequences of spontaneous mutational events in *Drosophila melanogaster*. Genetics 194:937-954.

Stoltzfus A, Norris RW. 2016. On the causes of evolutionary transition:transversion bias. Mol Biol Evol 33:595-602.

Tamura K, Nei M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol 10:512-526.

Yang Z. 2006. Computational molecular evolution. Oxford: Oxford University Press.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24:1586-1591.

Yang Z, Nielsen R, Hasegawa M. 1998. Models of amino acid substitution and applications to mitochondrial protein evolution. Mol Biol Evol 15:1600-1611.

Zhang J. 2000. Rates of conservative and radical nonsynonymous nucleotide substitutions in mammalian nuclear genes. J Mol Evol 50:56-68.

Zhu YO, Siegal ML, Hall DW, Petrov DA. 2014. Precise estimates of mutation rate and spectrum in yeast. Proc Natl Acad Sci U S A 111:E2310-2318.

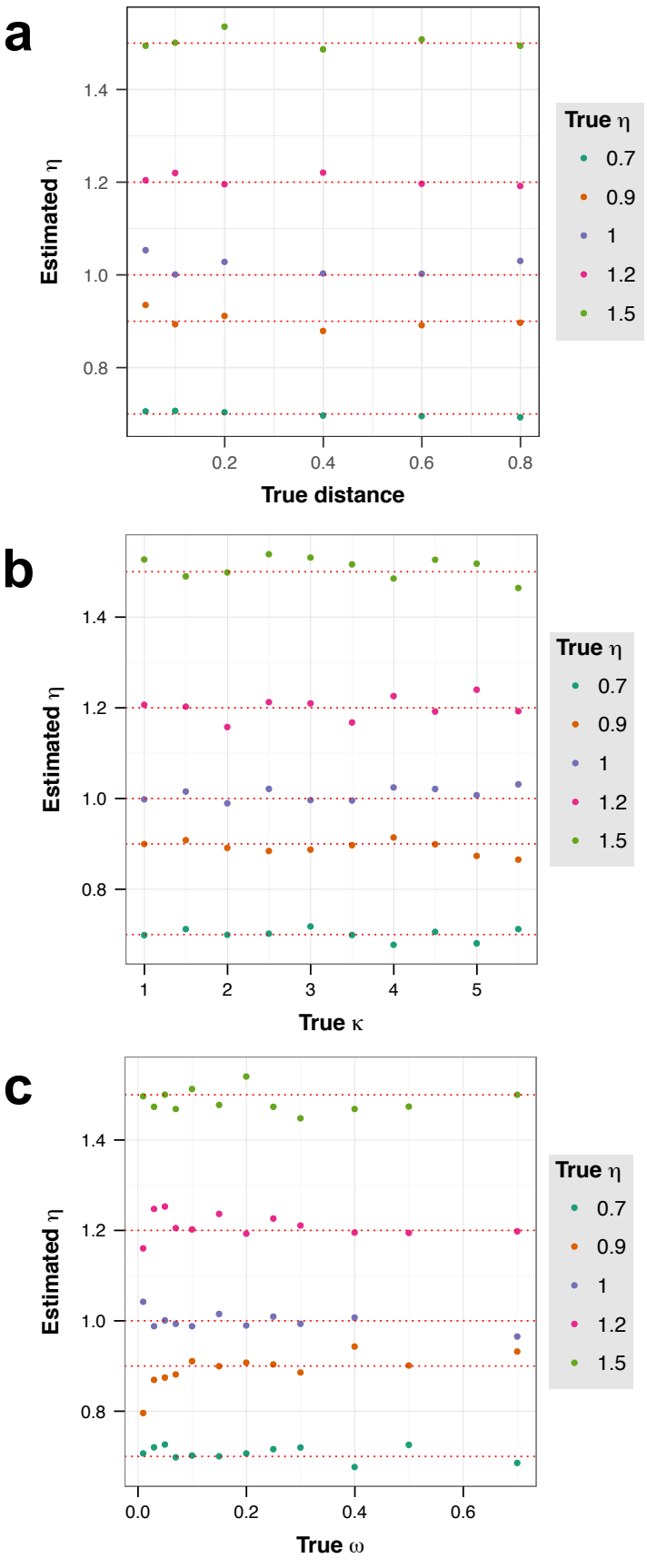

a
Estimated η
1.4
1.2
1.0
0.8
0.2
0.4
0.6
0.8
True distance
True η
0.7
0.9
1
1.2
1.5
b
Estimated η
1.4
1.2
1.0
0.8
1
2
3
4
5
True κ
True η
0.7
0.9
1
1.2
1.5
c
Estimated η
1.4
1.2
1.0
0.8
0.0
0.2
0.4
0.6
True ω
True η
0.7
0.9
1
1.2
1.5

Figure 7.1 The inferred η's show no deviation from the true values in simulation under the same model, nor do they correlate with the other varied parameters, such as (**a**) genetic distance between two species in the same clade, (**b**) transition/transversion mutational bias κ, and (**c**) overall selection ω. Each dot is one η estimation plotted against the true value of another parameter used during simulation. True value of each η is indicated by colors shown in legend, dotted lines correspond to the true value for clear comparison. In this analysis, more true values of η were used, but only five are plotted here for clarity.

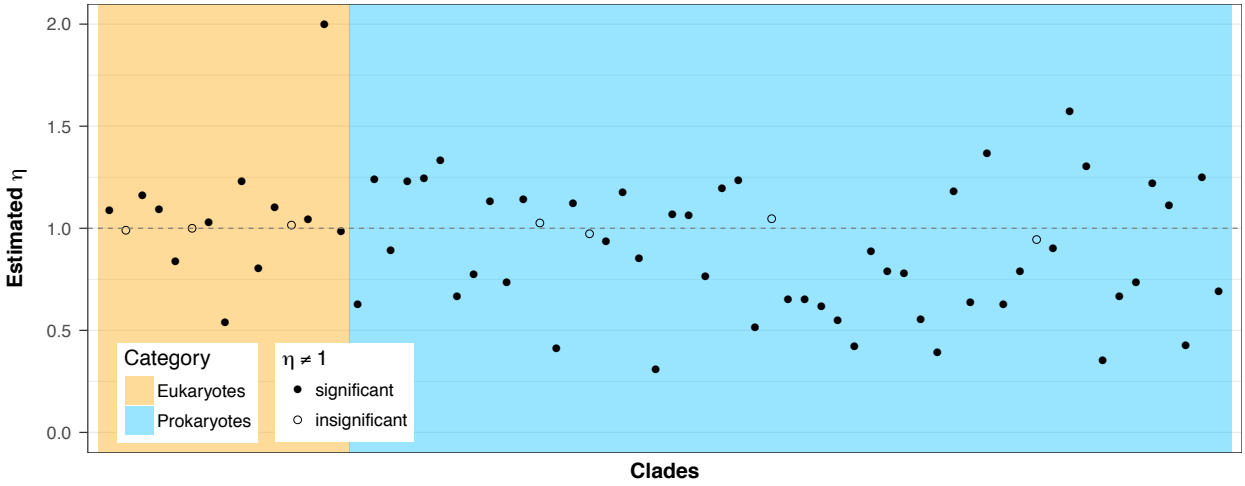


Figure 7.2 Nonsynonymous transition/transversion selectional bias η varies among different pairs (clades) of species. Inferred η from 15 pairs of eukaryotes (orange background) and 53 pairs of prokaryotes (blue background). Solid dots indicate values significantly deviating from 1 (likelihood ratio test, $P < 0.05$ for all runs). Dashed line indicate η=1.

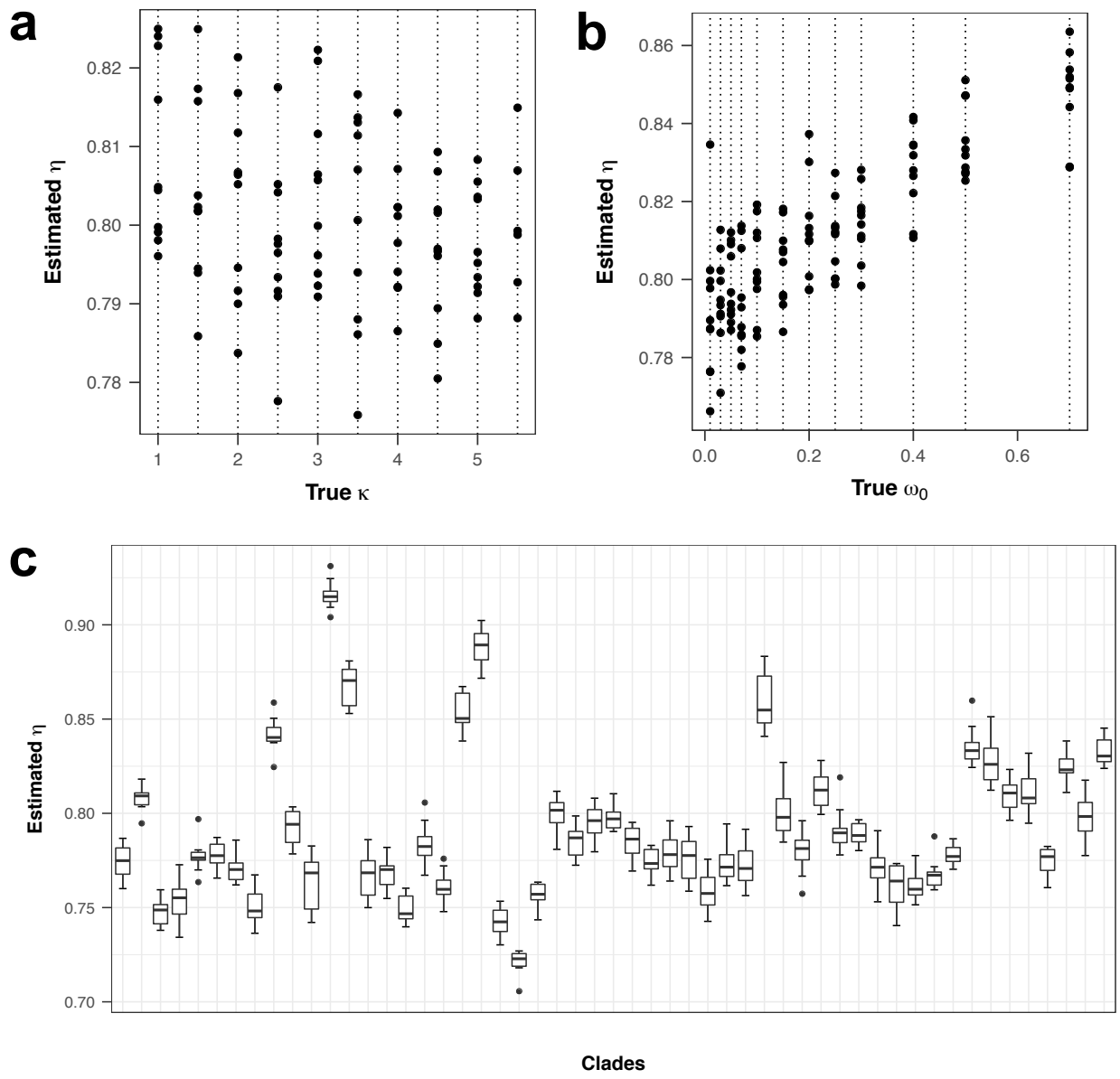


Figure 7.3 Variations in κ, ω_0, or π cannot explain the large variation in η among clades. η's inferred from simulated sequence alignments are plotted against the true (**a**) κ's, (**b**) ω_0's or (**c**) $\boldsymbol{\pi}$'s. Dashed vertical lines in (**a**) and (**b**) indicate the values of true κ's or ω's specified in simulation. For each value, η estimations from 10 replicate simulations are plotted. In (**c**), the codon frequencies of the 53 prokaryotic clades are used, and boxplots show distribution of η estimations from 10 replicate simulations.

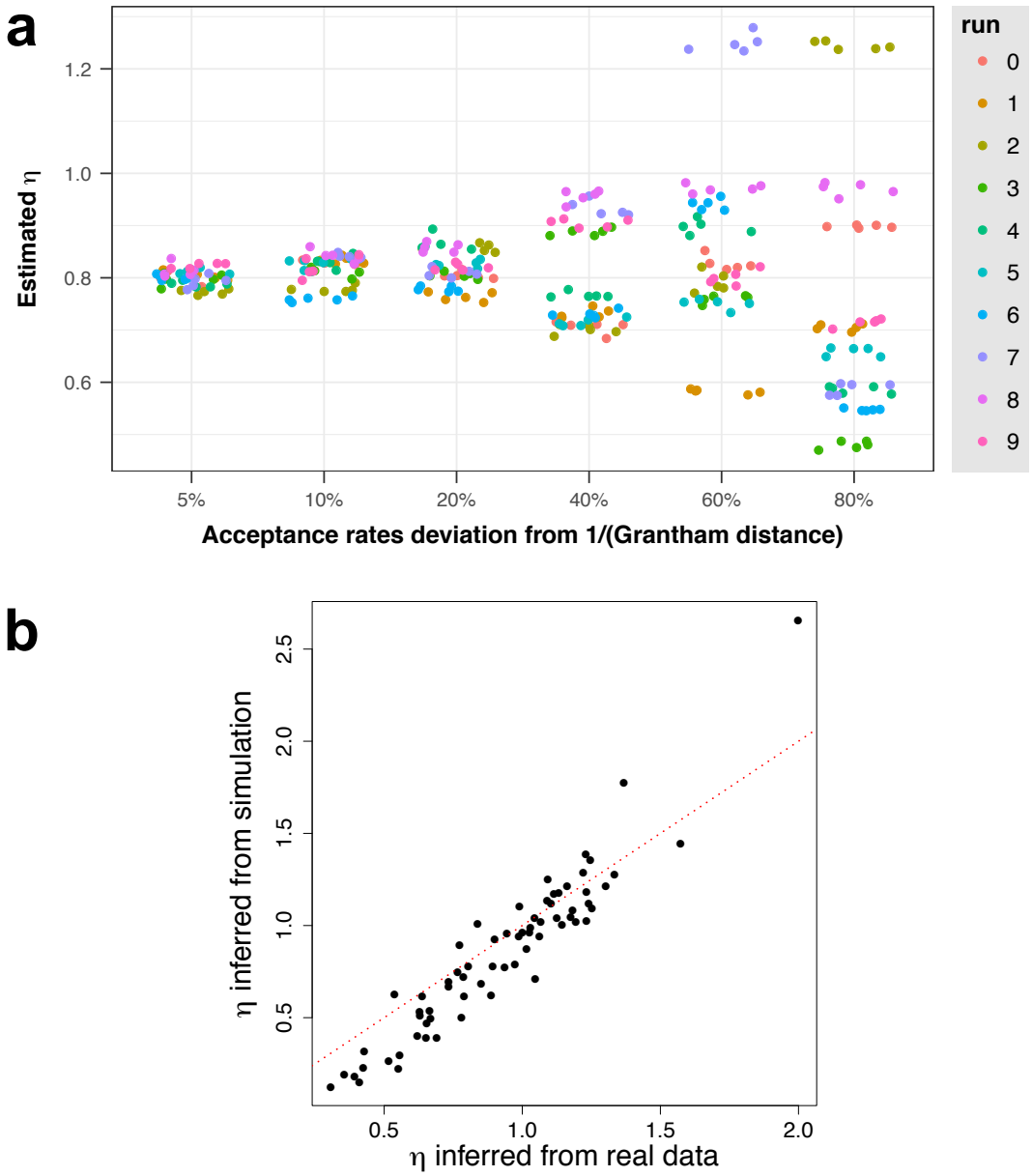


Figure 7.4 Variation of acceptance rates ω' among clades may explain η variation. (**a**) Simulations with $\boldsymbol{\omega'}$ derived from Grantham matrix. For each of the 10 runs at a certain percentage level x%, each acceptance rate $\omega'(aa_i, aa_j)$ was either added or subtracted x% of its reference value (1/Grantham distance) randomly and then used for simulation. Different runs are color-coded according to the legend. For each run, five replicate sequence evolution simulations were conducted and corresponding η plotted. (**b**) The η's estimated from the 68 simulated clades are plotted against corresponding values estimated from the real clades. Dotted red line indicate y=x.

Chapter 8

Conclusions

In this chapter, I will first summarize the conclusions of each research project in Chapter 2 – 7, and then discuss some major limitations as well as future directions.

In Chapter 2, I found the expectation of sequence convergence level is greatly affected by choice of different models. Specifically, models incorporating site-specificity of amino acid composition predict much higher level of neutral convergence than models without considering this factor. Consequently, observed genome-wide convergence can be explained by neutral evolution under the former models without invoking adaptation, showing that there is likely no prevalent adaptive convergence in the genome. Among many studies focusing on protein sequence convergence, this is the first attempt to explicitly model the site-specificity of amino acid composition to show its effect on convergence. I also found that convergence level diminishes with genetic distance, and proposed that diverged epistatic effects in different species lineages cause this pattern. I further showed evidence of such epistatic constraint and its divergence in different lineages of species using real sequence data. Thus, this chapter simultaneously shows heterogeneities among sites and among lineages, which are further elucidated in Chapter 3 and Chapter 4.

In Chapter 3, after controlling the effect of GTD by multiple approaches, I found that the diminishing convergence pattern still exists, confirming that epistasis is indeed one cause of this phenomenon in mammals. Respective analyses in fruit flies suggested that the relative contribution of epistasis and GTD to this phenomenon may vary from case to case according to the data analyzed.

In Chapter 4, By repeating the analyses in sister lineages without apparent phenotypic convergence, I showed that the observed convergence level in the three echolocating mammal lineages is no higher than the level in these control lineages, hence again showing that sequence convergence is largely neutral at genome scale, and there is no genome-wide adaptive convergence for echolocation in mammals.

In Chapter 5, I first found that more morphological convergences were observed than molecular convergences in the dataset. Moreover, I found that one important cause of the higher level of morphological convergence is likely the smaller number of morphological character states compared with sequence data. I also showed by computer simulation that the convergence filtering pipeline for combined analysis of morphological and molecular data can improve the accuracy of tree reconstruction. This is the first explicit and practical comparison of convergence level between morphology and sequence data in phylogenetics, and my results confirm that sequences should be a preferred type of phylogenetic data compared with morphology.

In Chapter 6, I showed that, unexpectedly, the genome-wide amino acid acceptance rates, i.e. fixation probability of different types of amino acid changes, have significant heterogeneity among different species clades. The difference was supported by statistical tests in multiple datasets. We proposed that there is a genome-scale, clade-specific factor driving this pattern. Discovery of this lineage-specificity is both novel and biologically interesting, which points out potential model over-simplification.

In Chapter 7, I found that the nonsynonymous transition/transversion biases are in different directions for different clades: In some clades, nonsynonymous transitions are less deleterious, while in others nonsynonymous transversions can be less deleterious. By checking the bias pattern in simulated clades, I found that the variation of bias patterns in real data is likely

driven by the acceptance rate variation previously found in Chapter 6, while other factors in codon sequence evolution have minor effects. Together, Chapter 6 and Chapter 7 address lineage-specific codon sequence evolution patterns that were not resolved previously, and provide a mechanistic explanation with potential biological significance.

Limitations and future directions are discussed below.

Modeling epistatic constraints on protein sequence evolution in phylogeny

In Chapter 2 and Chapter 4, I showed that substantial amount of convergence should be expected during neutral evolution. Both previous studies and my results indicate the role of epistasis as an important driving force of neutral convergence. I showed in Chapter 2 that only limited number of amino acids are allowed at each position in protein sequences even among a very large sample of species. It was also shown that when we assume all 20 amino acids can be allowed at each site (JTT-f_{gene} model), the calculated expectation of convergence level is much lower than when the site-wise constraint is considered (JTT-f_{site} model or JTT-CAT model). This site-specificity of amino acid composition has not been incorporated into major evolutionary models currently used in phylogenetic inferences. The CAT model is a partial realization, by assigning sites in a protein sequence into discrete amino acid frequency classes with posterior probabilities (Lartillot and Philippe 2004). Direct inference of the amino acid composition at a site is practically difficult. Since number of taxa available in most phylogenetic studies is limited, simply counting the observed amino acids at a site in the alignment is likely to cause underestimation of the actual number of amino acid allowed (Thomas et al. 2017), thus overestimating expected neutral convergence.

However, there is a second layer of complexity, regarding the divergence of epistasis that our results suggested. In a phylogeny, a direct ramification of having divergence of epistasis in different lineages is that amino acid composition in one lineage does not apply to another. Thus, to make more accurate estimation of the composition at a site in one lineage, simply augmenting the number of sampled species might cause disequilibrium within the "lineage". Ideally, one has to sample intensively within a closely related species clade to know the full distribution of amino acid allowed at each site. Nevertheless, the interplay of site-specificity and lineage-specificity may turn out to be a trade-off. Recently, models incorporating site- and lineage-heterogeneity have been developed and applied in resolving practical phylogenies (Blanquart and Lartillot 2008; Jayaswal et al. 2014; Pisani et al. 2015). In the future, calculation of expected neutral convergence level might achieve higher accuracy under this type of models.

Disentangling the effects of homoplasy and hemiplasy

The results of Chapter 3 and other recent studies (Hahn and Nakhleh 2016; Mendes et al. 2016) have suggested that the existence of hemiplasy can appear as true convergence. Although several approaches were used to exclude the possible effect of hemiplasy at genome-wide level and confirmed existence of true convergence pattern, it remains unclear how to distinguish the effect of hemiplasy from that of homoplasy from case to case. As mentioned in Chapter 1, when internal branches are short, likelihood of both events will increase. Furthermore, detection of hemiplasy uses gene tree discordance (GTD) as a signal. However, we can only observe the fact that gene tree inferred from a gene is discordant from the species tree, but this discordance may well be cause by true convergence that confounded phylogenetic signals, rather than by *bona fide* gene tree discordance events in the evolutionary history. For example, Salichos and Rokas

(2013) found that among 1,070 gene trees of 23 yeast strains, none of them is completely concordant with the species tree. It is unlikely that all these discordances are truly due to incomplete lineage sorting, introgression, or horizontal gene transfer, etc. Similarly, the fact that the hearing protein prestin support monophyly of two lineages of echolocating bats can either indicate adaptive convergence or be due to remaining ancestral polymorphism. This can even be extended to functional level, questioning whether echolocation in the two bat lineages was developed independently, or was just a phenotype that happen to be sorted into two non-sister lineages.

One difference between hemiplasy and homoplasy is that the former often causes linked changes while the latter should largely be scattered in the genome. However, the linked regions containing ancient hemiplasy may be broken by new mutations and recombinations. Hence, studying recent split of lineages might be of help to disentangle the mixed effects of homoplasy and hemiplasy in confounding true species divergence patterns.

Underlying biological factors of amino acid acceptance rate variation

Although we discovered the surprising pattern that acceptance rates differ among clades of species, no biological explanation is currently conspicuous. Based on the model we use, the underlying factor should specifically map to each pair of amino acids, because the selection towards or against certain single amino acid is counted for by the equilibrium codon frequencies in the substitution model.

Furthermore, we also showed that different genes in the same genome actually share more similar acceptance rates than orthologous genes in different genomes. This suggest the necessity of incorporating a set of clade-specific acceptance rates into evolutionary models.

Being another layer of heterogeneity among lineages, this is not trivial to handle. Importantly, the available amount of sequence data is a fundamental constraint on the level of model complexity and number of parameters. In the future, with more sequence data available for broader and denser samples of species on the tree of life, it is intriguing to incorporate more heterogeneity into phylogenetic models and elucidate protein evolution patterns in higher resolution.

REFERENCES

Blanquart S, Lartillot N. 2008. A site- and time-heterogeneous model of amino acid replacement. Mol Biol Evol 25:842-858.

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

Jayaswal V, Wong TK, Robinson J, Poladian L, Jermiin LS. 2014. Mixture models of nucleotide sequence evolution that account for heterogeneity in the substitution process across sites and across lineages. Syst Biol 63:726-742.

Lartillot N, Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol 21:1095-1109.

Mendes FK, Hahn Y, Hahn MW. 2016. Gene tree discordance can generate patterns of diminishing convergence over time. Mol Biol Evol 33:3299-3307.

Pisani D, Pett W, Dohrmann M, Feuda R, Rota-Stabelli O, Philippe H, Lartillot N, Worheide G. 2015. Genomic data do not support comb jellies as the sister group to all other animals. Proc Natl Acad Sci U S A 112:15402-15407.

Salichos L, Rokas A. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. Nature 497:327-331.

Thomas GW, Hahn MW, Hahn Y. 2017. The effects of increasing the number of taxa on inferences of molecular convergence. Genome Biol Evol 9:213-221.

APPENDICES

A.1 Supplementary table for Chapter 2

Table A.1.1. Observed numbers of convergent and parallel sites and the corresponding numbers expected under neutral models of amino acid substitution. Results presented are for the two exterior branches leading to *D. yakuba* and *D. melanogaster*, respectively, in **Fig. 2.2a**.

Type of sites	Number of sites examined	Observed number of sites	Expected number of sites		Ratio[a]	P-value[b]
			Substitution model	Number of sites		
Convergent sites						
	2,028,428	12	JTT-f_{site}	12.1	0.99	0.56
	2,028,428	12	JTT-f_{gene}	4.1	2.93	0.0011
	780,615	2	JTT-CAT	3.5	0.57	0.32
Parallel sites						
	2,028,428	479	JTT-f_{site}	620.6	0.77	1.9E-09
	2,028,428	479	JTT-f_{gene}	128.9	3.72	1.7E-123
	780,615	142	JTT-CAT	73.1	1.94	6.4E-13

[a]Ratio between the observed number and expected number.

[b]A statistical test is conducted under the assumption that the number of convergent (or parallel) sites follows a Poisson distribution with the mean equal to the expected number. When the observed number is smaller than the expected, the lower tail probability is given; when the observed number is larger than the expected, the upper tail probability is given.

A.2 Supplementary figures for Chapter 3

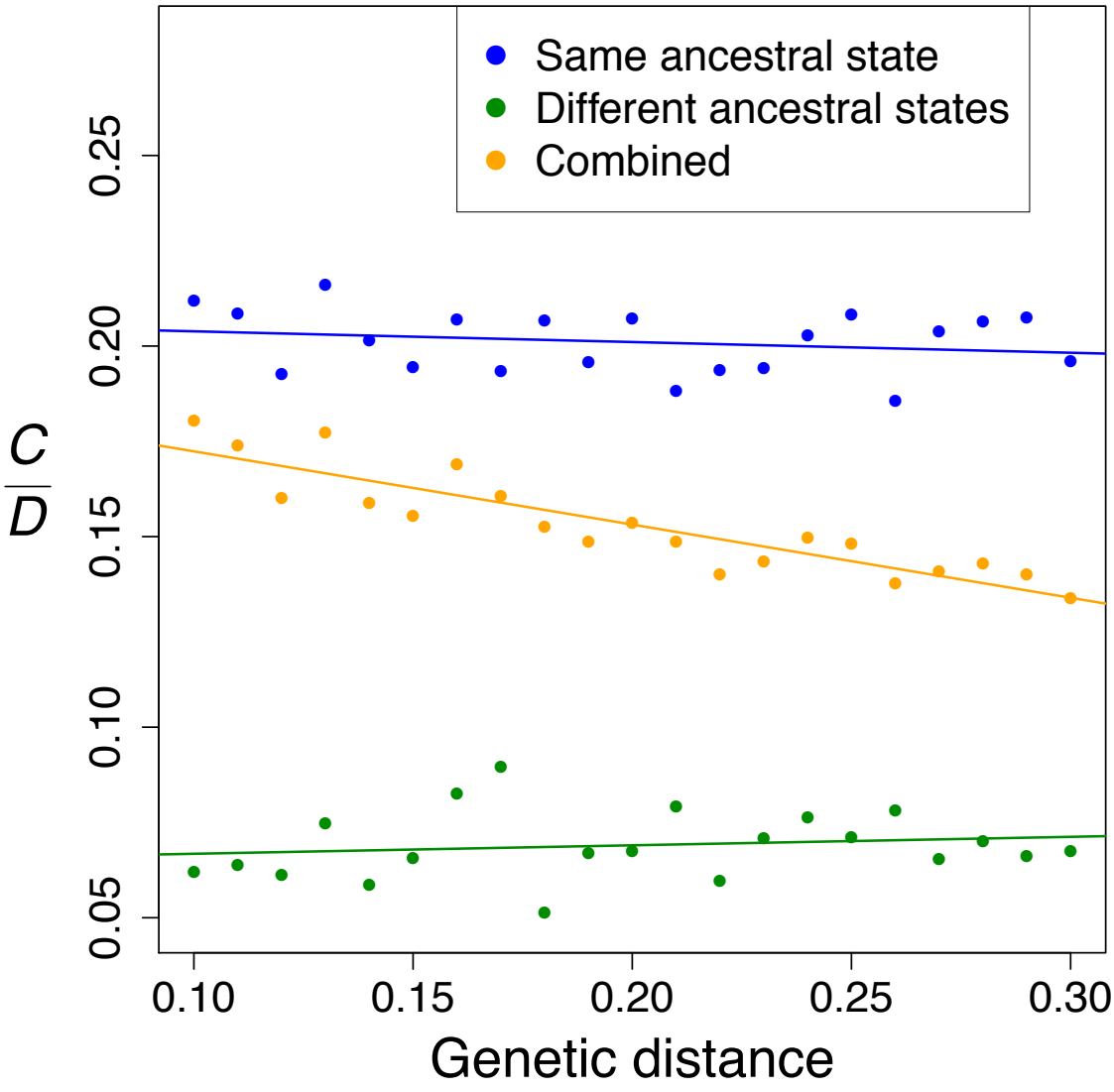


Figure A.2.1. Simulations showing the correlation between *C/D* and genetic distance in the absence of epistasis and incomplete lineage sorting. Simulation of 500,000 amino acid sites was conducted with the same method used to produce Figure 4 in Zou and Zhang (2015a) except that equilibrium amino acid frequencies at each site are maintained during sequence evolution. Blue and green dots respectively represent $(C/D)_s$ and $(C/D)_d$, while orange dots represent the ratio between the total number of parallel and convergent substitutions and that of all divergence events. Colored lines show linear regressions from data points of the same color. Only orange dots show a significant correlation with genetic distance ($r = -0.90$, $P = 4\times10^{-8}$).

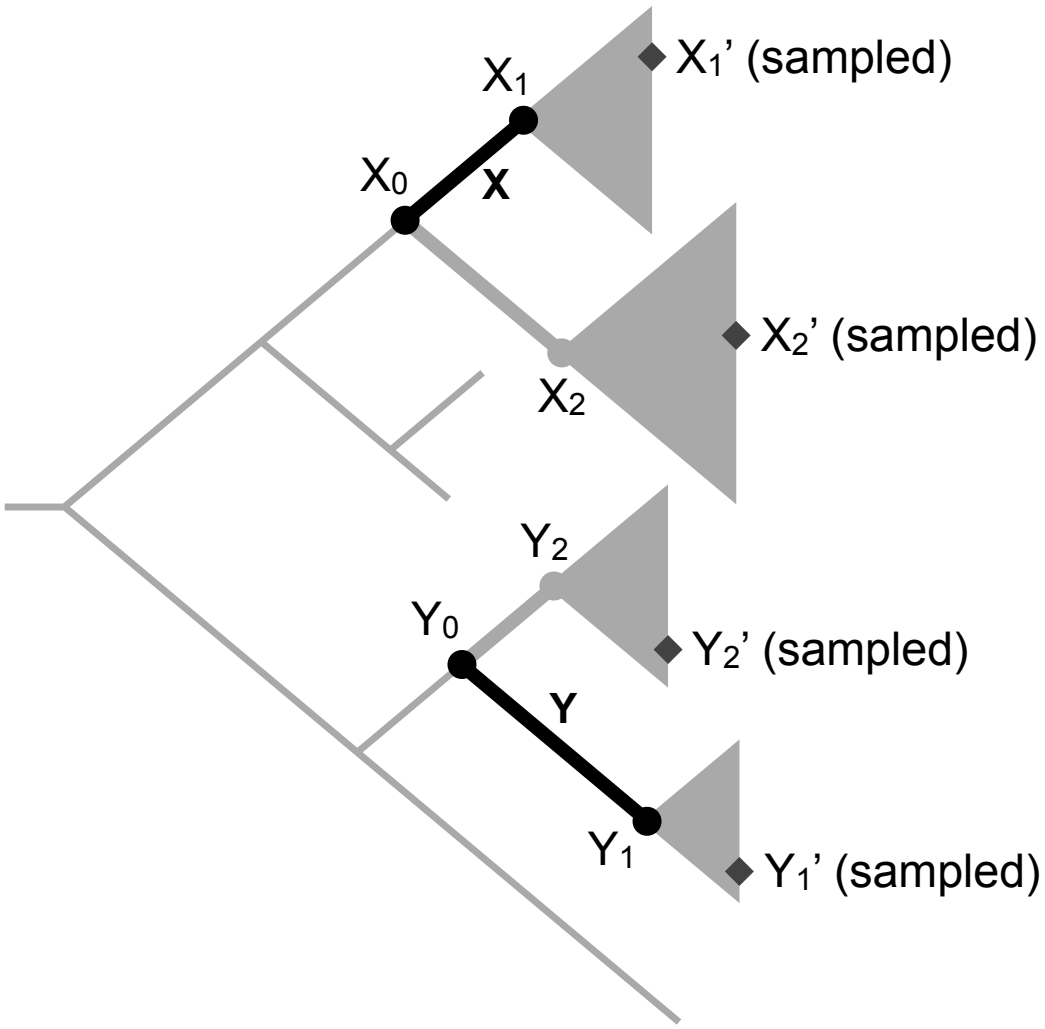


Figure A.2.2. A schematic illustration for GTD level estimation. Grey triangles represent the monophyletic groups of all descendants of X_1, X_2, Y_1 and Y_2, respectively. See MATERIALS AND METHODS for details of the estimation.

A.3 Supplementary tables and figures for Chapter 5

Table A.3.1. Convergence level negatively correlates with number of states after the control of evolutionary rate in the actual data

Spearman's rank correlation	Partial correlation with number of states (controlling for the number of steps)		
	all informative characters	morphological characters	molecular characters
Cv/Dv ratio of characters	ρ = -0.86, P < 1E-300	ρ = -0.79, P < 1E-300	ρ = -0.77, P < 1E-300
Cv/Cs ratio of characters	ρ = -0.31, P = 2E-202	ρ = -0.19, P = 1E-27	ρ = -0.22, P = 8E-63

Amino acid sites are used as molecular data. Evolutionary rates (number of steps) are inferred on the basis of the morphological tree.

Table A.3.2. Convergence level negatively correlates with number of states after the control of evolutionary rate in the actual data

Spearman's rank correlation	Partial correlation with number of states (controlling for the number of steps)		
	all informative characters	morphological characters	molecular characters
Cv/Dv ratio of characters	ρ = -0.78, P < 1E-300	ρ = -0.78, P < 1E-300	ρ = -0.76, P < 1E-300
Cv/Cs ratio of characters	ρ = -0.18, P = 3E-155	ρ = -0.19, P = 1E-28	ρ = -0.17, P = 2E-124

Nucleotide sites are used as molecular data. Evolutionary rates (number of steps) are inferred on the basis of the morphological tree.

Table A.3.3. Convergence level negatively correlates with number of states after the control of evolutionary rate in the simulated data

Spearman's rank correlation	Partial correlation with number of states (controlling for the number of steps)		
	all informative characters	morphological characters	molecular characters
Cv/Cs ratio of characters	ρ = -0.32, P < 1E-300	ρ = -0.14, P = 5E-79	ρ = -0.33, P < 1E-300

Simulated amino acid sites are used as molecular data. Evolutionary rates (number of steps) are recorded during the simulation.

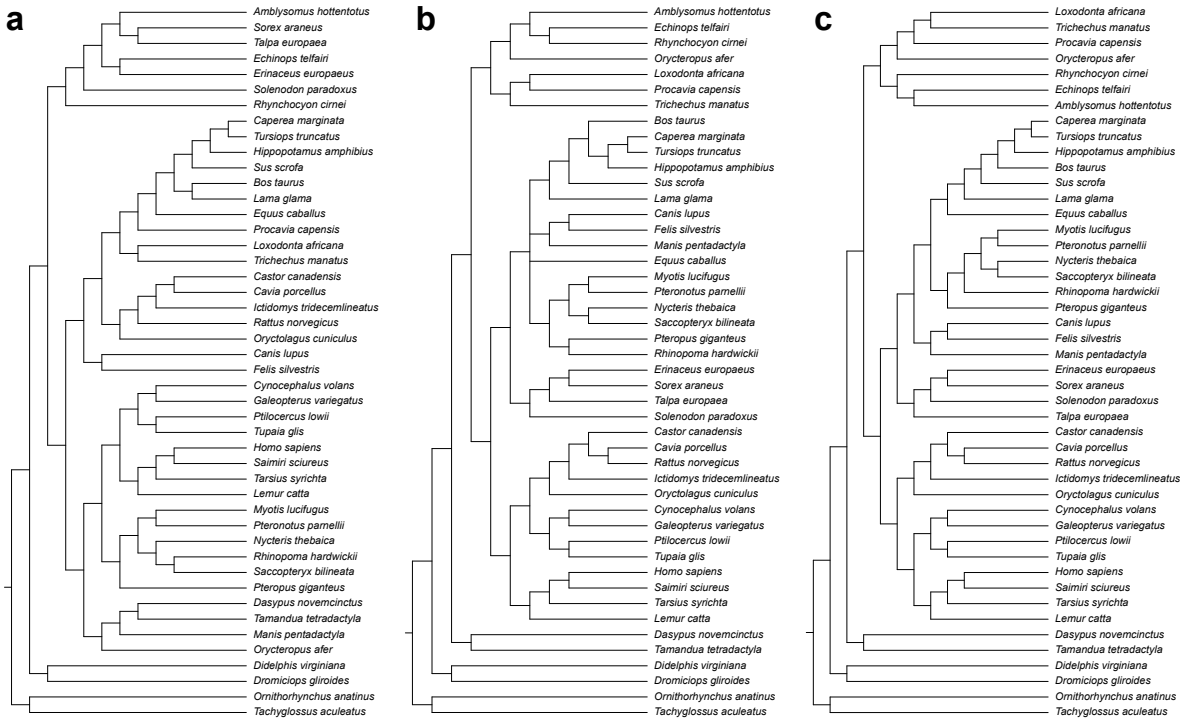


Figure A.3.1. Tree topologies used in the analysis of convergence. (**a**) The parsimony tree based on all morphological characters of all 86 species. Only the 46 extant species are shown here. (**b**) The parsimony tree based on all molecular characters of the 46 extant species. (**c**) The parsimony tree based on all morphological and molecular characters of all 86 species. Only the 46 extant species are shown here.

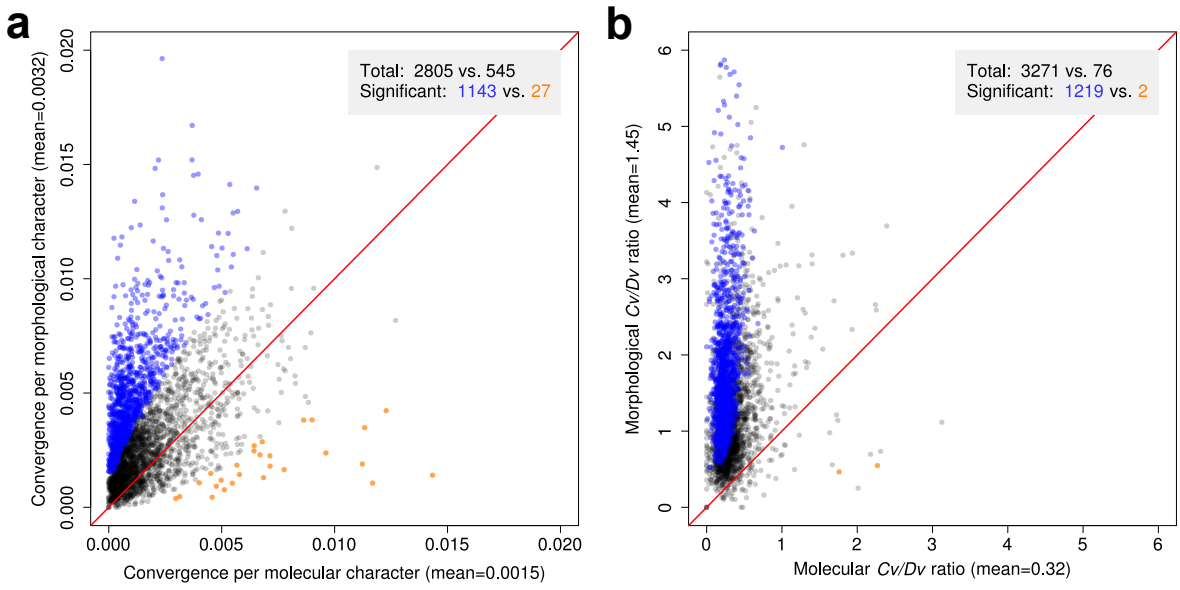


Figure A.3.2. Whole-tree analysis of convergence based on the total evidence tree. (**a**) Mean number of convergences per morphological character and that per molecular character for each branch pair examined. (**b**) Convergence/divergence (*Cv*/*Dv*) ratio for each branch pair. In (**a**) and (**b**), each dot represents a branch pair. In the grey box of each panel, 'total' refers to the numbers of dots above and below the diagonal, respectively, and 'significant' refers to the numbers of dots significantly (at *Q*-value of 0.05) above (blue) and below (orange) the diagonal, respectively (dots on the diagonal are not counted). Total number of dots above the diagonal significantly exceeds that below the diagonal in both panels ($P < 1\times10^{-4}$, bootstrap test).

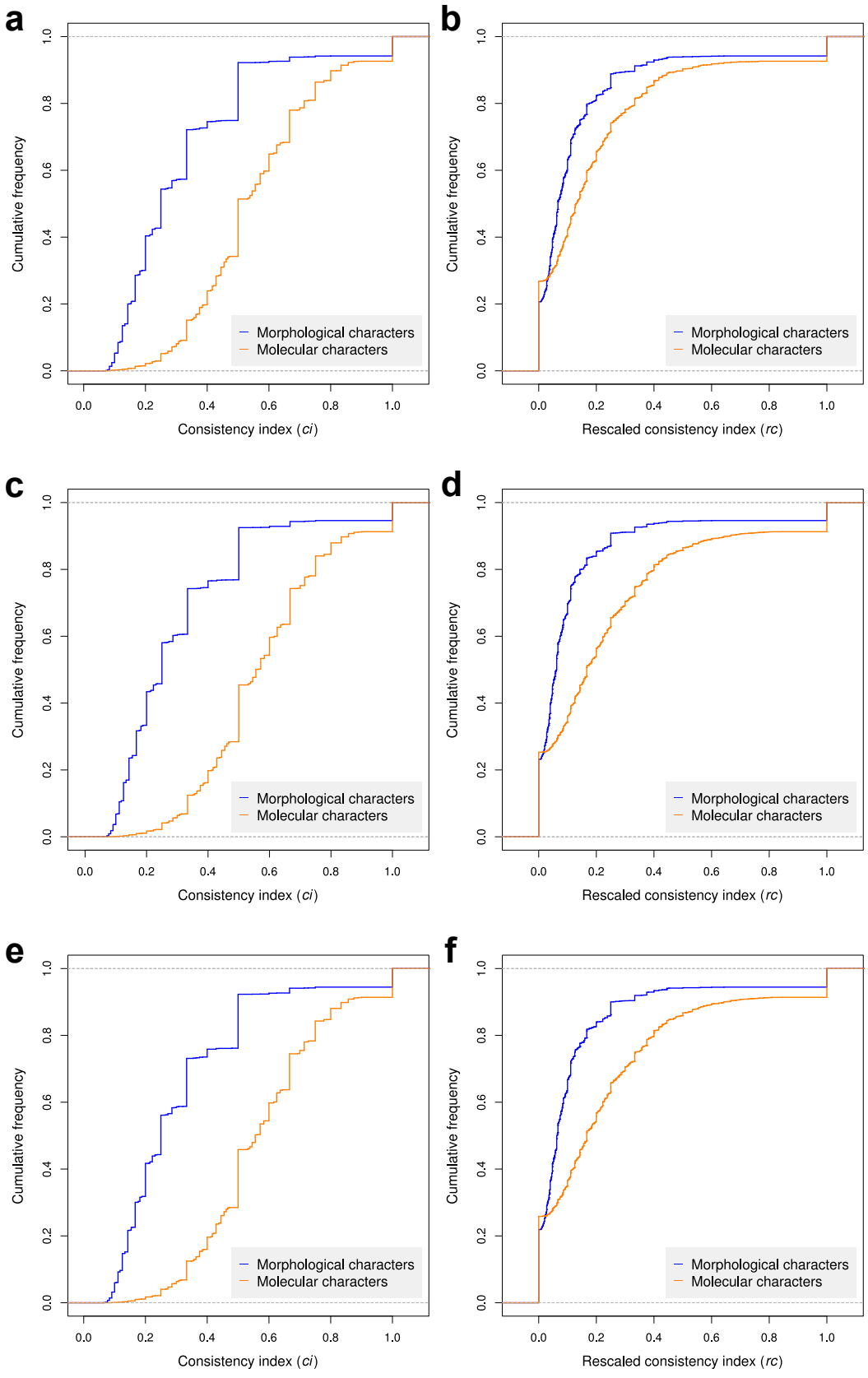

a
b
c
d
e
f
Cumulative frequency
0.0
0.2
0.4
0.6
0.8
1.0
Consistency index (ci)
Rescaled consistency index (rc)
Morphological characters
Molecular characters

Figure A.3.3. Consistency index and rescaled consistency index are generally higher for molecular characters than morphological characters. (**a**) Cumulative frequency distributions of consistency index for parsimony-informative morphological characters and molecular characters based on the morphological tree. The difference between the two distributions is significant ($P < 1\times10^{-300}$, Mann-Whitney U test). (**b**) Cumulative frequency distributions of rescaled consistency index for parsimony-informative morphological characters and molecular characters based on the morphological tree. The difference between the two distributions is significant ($P < 3\times10^{-47}$, Mann-Whitney U test). (**c**) Cumulative frequency distributions of consistency index for parsimony-informative morphological characters and molecular characters based on the molecular tree. The difference between the two distributions is significant ($P < 1\times10^{-300}$, Mann-Whitney U test). (**d**) Cumulative frequency distributions of rescaled consistency index for parsimony-informative morphological characters and molecular characters based on the molecular tree. The difference between the two distributions is significant ($P < 2\times10^{-137}$, Mann-Whitney U test). (**e**) Cumulative frequency distributions of consistency index for parsimony-informative morphological characters and molecular characters based on the total evidence tree. The difference between the two distributions is significant ($P < 1\times10^{-300}$, Mann-Whitney U test). (**f**) Cumulative frequency distributions of rescaled consistency index for parsimony-informative morphological characters and molecular characters based on the total evidence tree. The difference between the two distributions is significant ($P < 4\times10^{-111}$, Mann-Whitney U test).

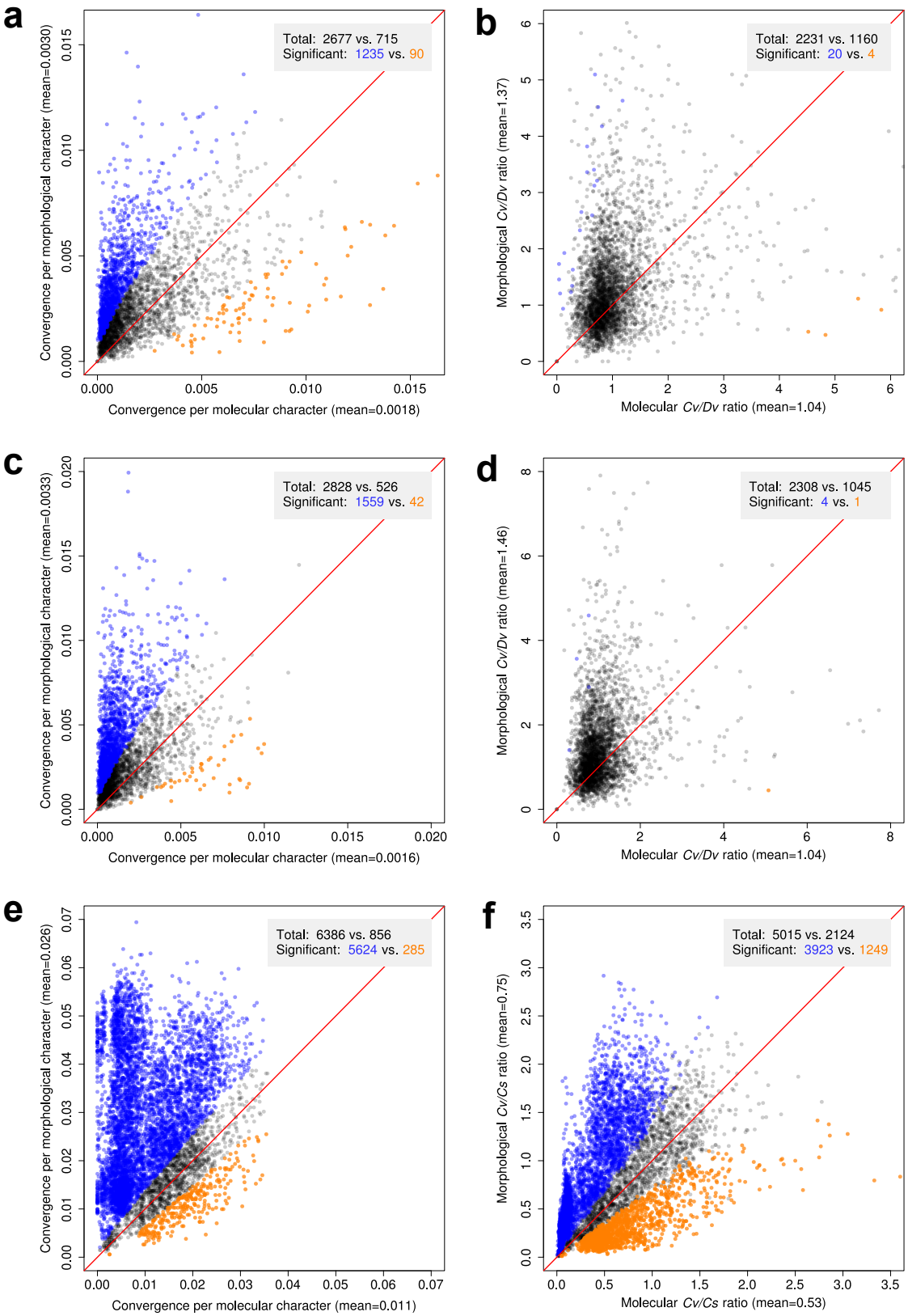

a
Total: 2677 vs. 715
Significant: 1235 vs. 90
Convergence per morphological character (mean=0.0030)
Convergence per molecular character (mean=0.0018)
b
Total: 2231 vs. 1160
Significant: 20 vs. 4
Morphological Cv/Dv ratio (mean=1.37)
Molecular Cv/Dv ratio (mean=1.04)
c
Total: 2828 vs. 526
Significant: 1559 vs. 42
Convergence per morphological character (mean=0.0033)
Convergence per molecular character (mean=0.0016)
d
Total: 2308 vs. 1045
Significant: 4 vs. 1
Morphological Cv/Dv ratio (mean=1.46)
Molecular Cv/Dv ratio (mean=1.04)
e
Total: 6386 vs. 856
Significant: 5624 vs. 285
Convergence per morphological character (mean=0.026)
Convergence per molecular character (mean=0.011)
f
Total: 5015 vs. 2124
Significant: 3923 vs. 1249
Morphological Cv/Cs ratio (mean=0.75)
Molecular Cv/Cs ratio (mean=0.53)

Figure A.3.4. Whole-tree analysis (a-d) and quartet analysis (e,f) with nucleotide sites being the molecular data. (**a**) Mean number of convergences per morphological character and that per molecular character for each branch pair examined under the morphological tree. (**b**) Convergence/divergence (*Cv*/*Dv*) ratio for each branch pair under the morphological tree. (**c**) Mean number of convergences per morphological character and that per molecular character for each branch pair examined under the molecular tree. (**d**) *Cv*/*Dv* ratio for each branch pair under the molecular tree. Labels, legends, and color schemes in (**a**)-(**d**) follow **Fig. 5.1**. (**e**) Mean number of convergences per morphological character and that per molecular character for each quartet examined. (**f**) Convergence/consistency (*Cv*/*Cs*) ratio for each quartet. Labels, legends, and color schemes in (**e**) and (**f**) follow **Fig. 5.2**.

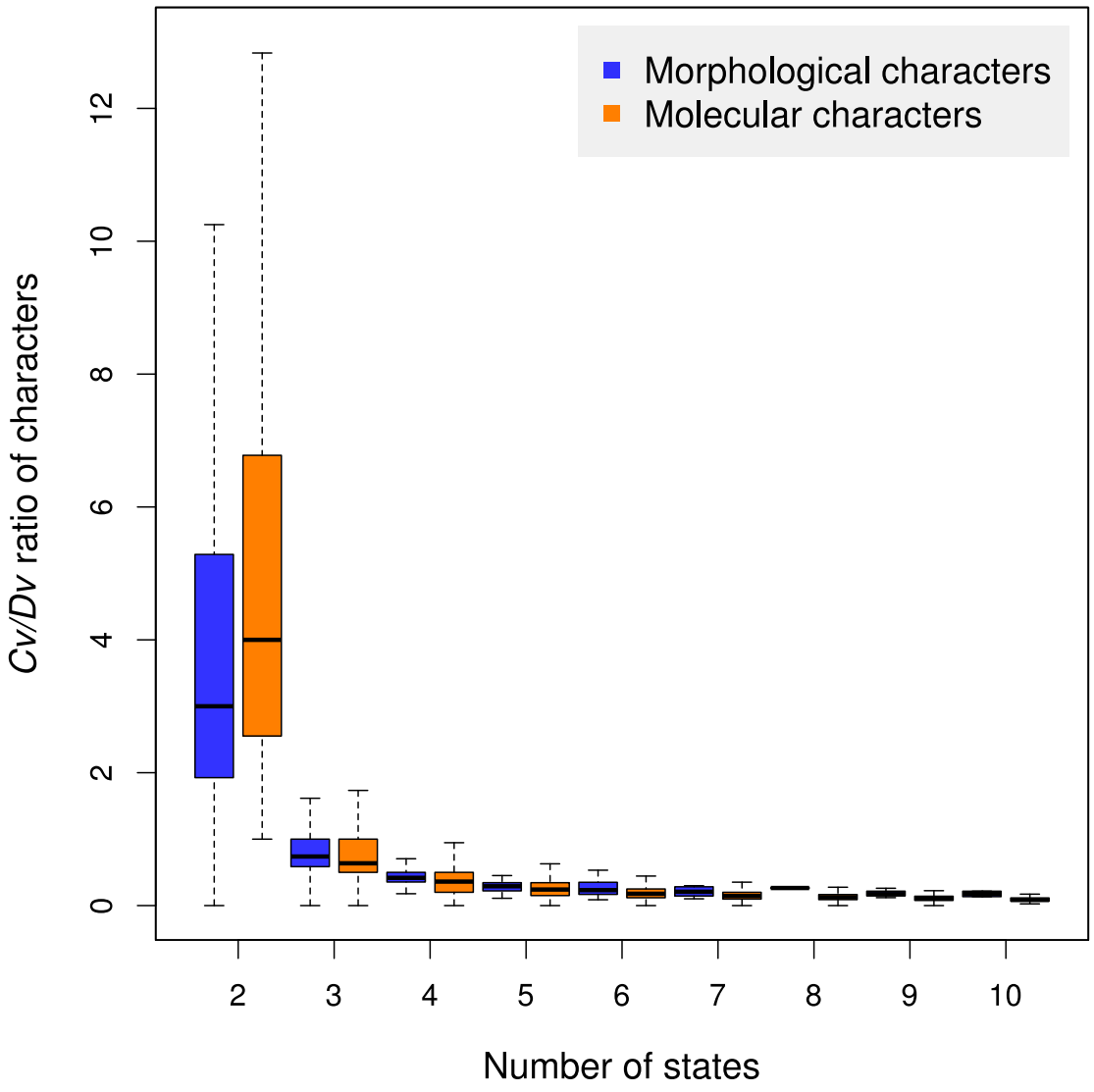
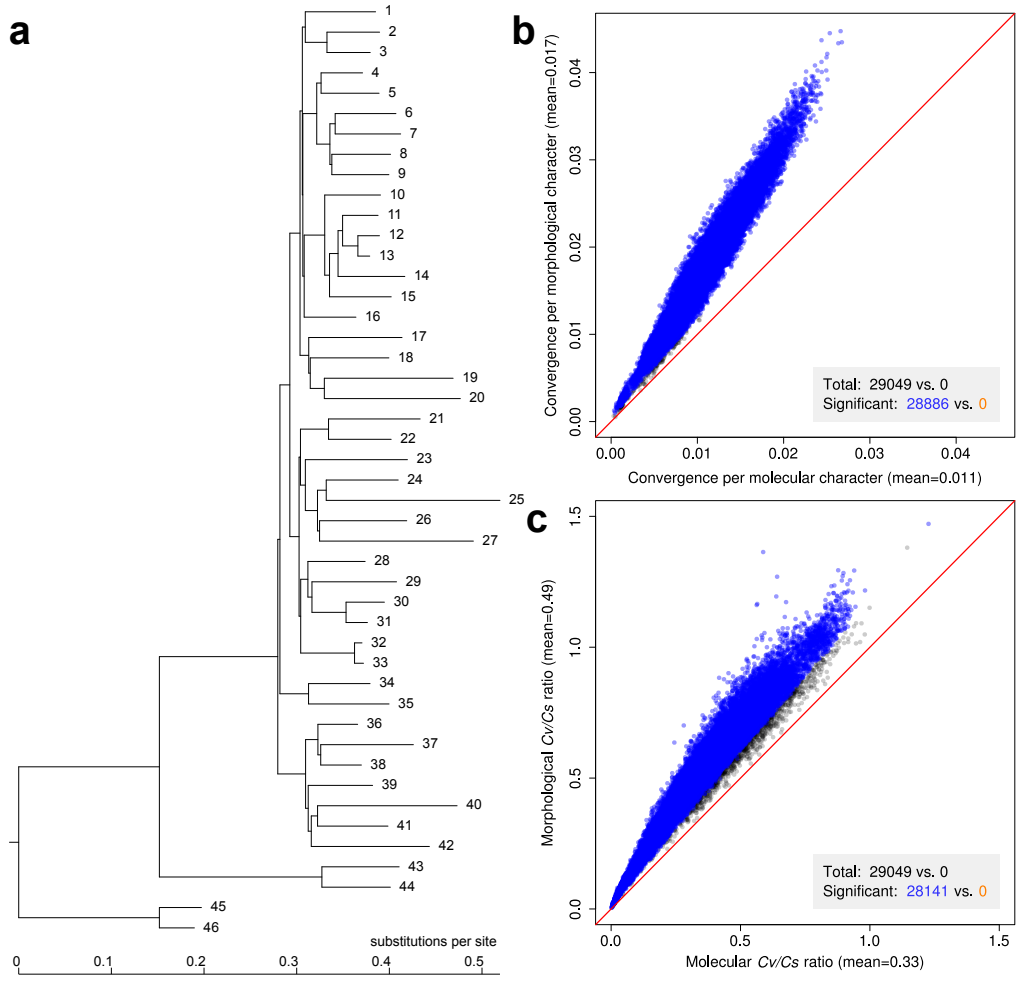


Figure A.3.5. *Cv*/*Dv* ratio decreases as the number of states increases. The *Cv/Dv* ratio of a character is the sum of convergences across all branch pairs divided by that of divergences. *Cv*/*Dv* ratios are calculated under the molecular tree.


a
b
c
substitutions per site
Convergence per morphological character (mean=0.017)
Convergence per molecular character (mean=0.011)
Total: 29049 vs. 0
Significant: 28886 vs. 0
Morphological Cv/Cs ratio (mean=0.49)
Molecular Cv/Cs ratio (mean=0.33)
Total: 29049 vs. 0
Significant: 28141 vs. 0

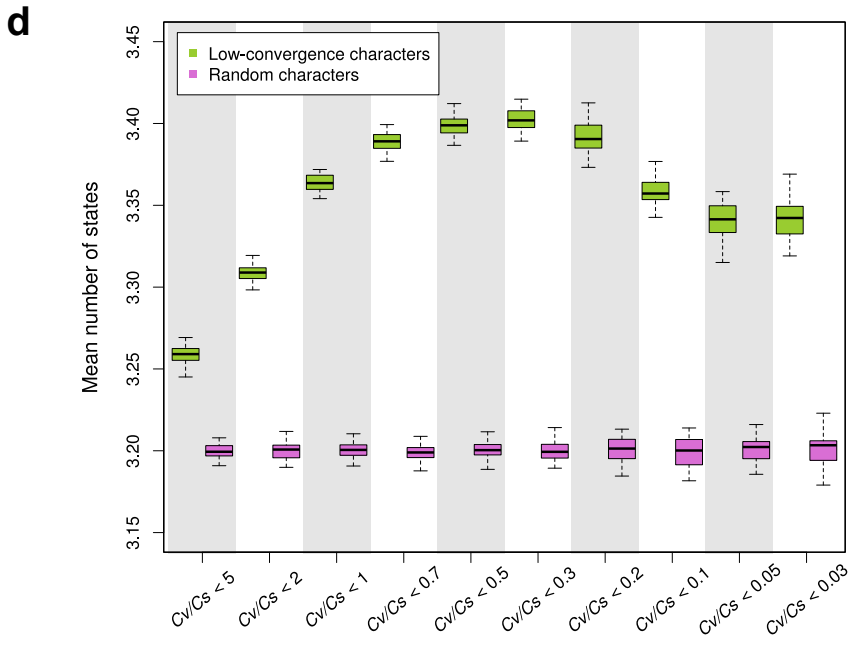

d
Low-convergence characters
Random characters
Mean number of states
Cv/Cs < 5
Cv/Cs < 2
Cv/Cs < 1
Cv/Cs < 0.7
Cv/Cs < 0.5
Cv/Cs < 0.3
Cv/Cs < 0.2
Cv/Cs < 0.1
Cv/Cs < 0.05
Cv/Cs < 0.03

Figure A.3.6. Properties of simulated morphological and molecular characters. (**a**) Tree used in the simulation, which is the nucleotide maximum likelihood tree from the original study. (**b**) Convergence per character for all quartets that show the same phylogenetic relationships in the morphological and molecular trees. (**c**) *Cv*/*Cs* ratios for the same quartets. Each dot represents a quartet. Annotations and legends follow **Fig. 5.2b** and **5.2c**. In (b) and (c), each dot represents a quartet. Convergence and consistency information was obtained from quartet analysis. Numbers of states and steps were directly recorded in the simulation. The data were from the first of the 50 simulations. Number of dots above the diagonal significantly exceeds that below the diagonal ($P < 1\times10^{-4}$, bootstrap test). (**d**) The remaining characters after the removal of high-convergence characters tend to have larger numbers of states, compared with those of the same numbers of randomly picked characters from the original data. The top and bottom edges of a box respectively represent the first and third quartiles of the distribution from 50 simulations, while the thick line inside the box represents the median. The two whiskers show the maximum value not greater than the 1st quartile plus 1.5 times the box height and the minimum value not smaller than the 3rd quartile minus 1.5 times the box height, respectively. Differences between all pairs of boxes are significant ($P < 1\times10^{-8}$) by Mann-Whitney U test.

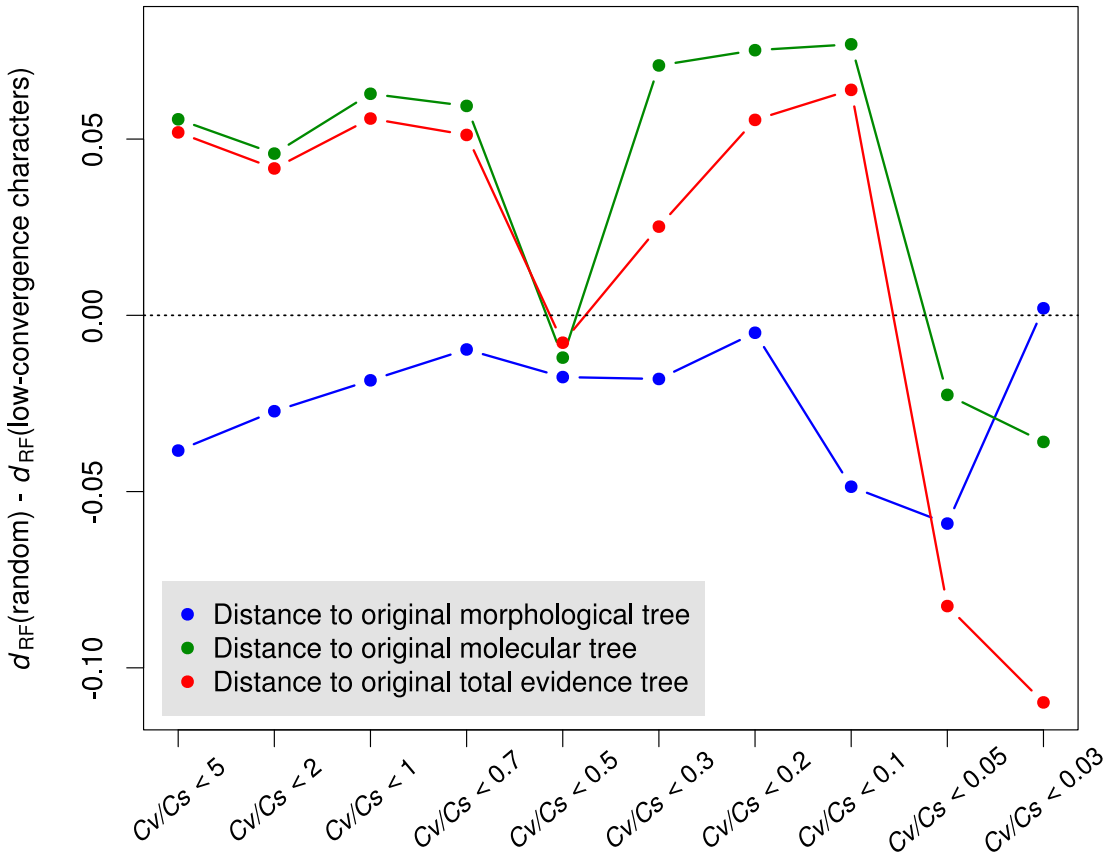


Figure A.3.7. Decrease in Robinson-Foulds distance (d_{RF}) between the inferred tree and an original tree after the removal of high *Cv*/*Cs* characters, relative to that after the removal of the same number of randomly picked characters. Positive Y-axis values show that, relative to removing random characters, removing high *Cv*/*Cs* characters makes the tree closer to the original tree being compared.

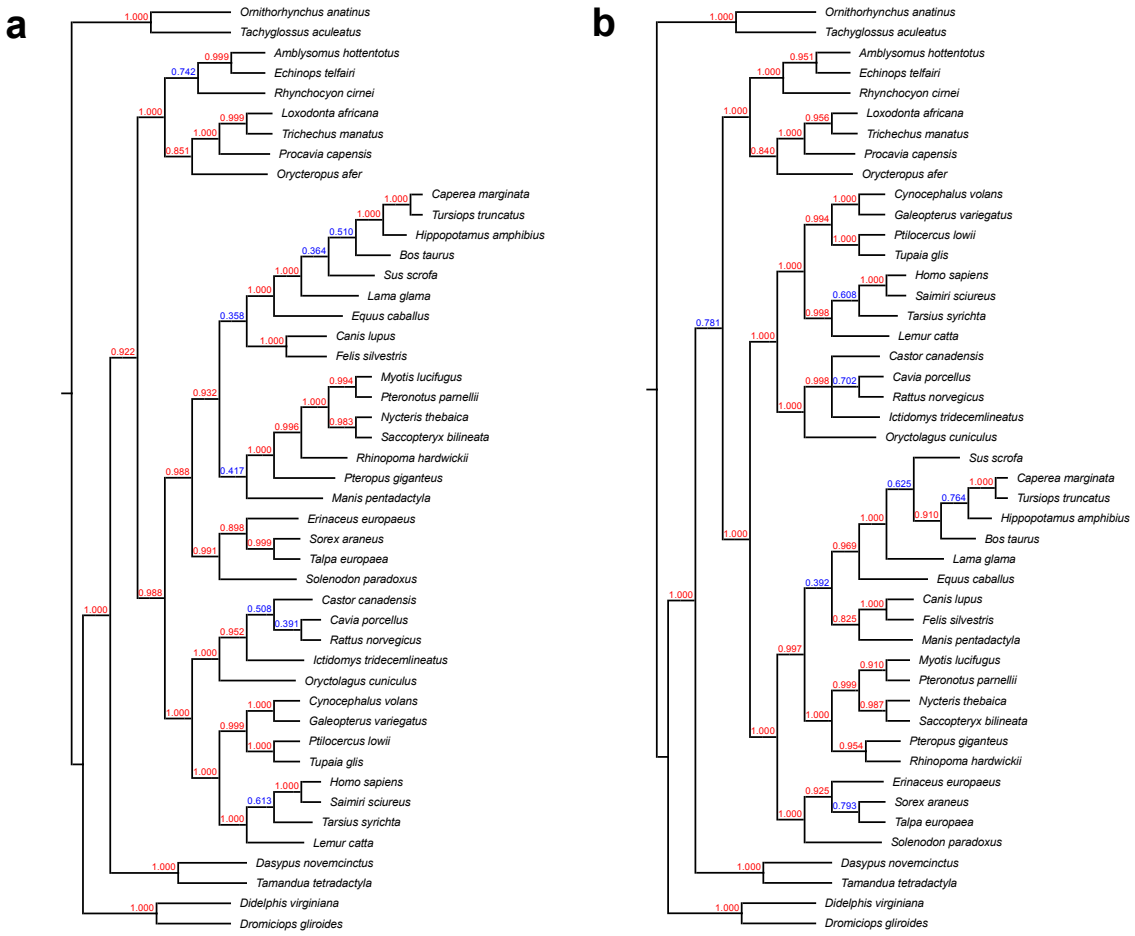


Figure A.3.8. Parsimony trees of mammals before and after the removal of high-convergence characters. (**a**) Parsimony tree of 46 extant species based on all informative morphological and amino acid characters. (**b**) Parsimony tree of 46 extant species based on informative morphological and amino acid characters with *Cv/Cs* < 0.2. Branch labels indicate the proportions of 1000 bootstrapped trees that support the subdivision of 46 species by this branch. Bootstrap values lower than 0.8 are colored in blue. All 86 species are included in the phylogenetic and bootstrap analyses, although only extant species are shown here.

A.4 Supplementary tables and figures for Chapter 6

Table A.4.1. 68 clades used for ω inference and shuffling tests.

	Clade		Data source
	Species 1	Species 2	
1	*Mus musculus*	*Rattus norvegicus*	OrthoMaM v9
2	*Homo sapiens*	*Pan troglodytes*	OrthoMaM v9
3	*Canis lupus familiaris*	*Felis catus*	OrthoMaM v9
4	*Monodelphis domestica*	*Sarcophilus harrisii*	OrthoMaM v9
5	*Gallus gallus*	*Meleagris gallopavo*	Ensembl 84
6	*Takifugu rubripes*	*Tetraodon nigroviridis*	Ensembl 84
7	*Drosophila sechellia*	*Drosophila simulans*	Flybase
8	*Atta cephalotes*	*Solenopsis invicta*	Ensembl Metazoa
9	*Saccharomyces cerevisiae*	*Saccharomyces paradoxus*	http://www.saccharomycessensustricto.org/
10	*Fusarium graminearum*	*Fusarium pseudograminearum*	Ensembl Fungi
11	*Arabidopsis thaliana*	*Arabidopsis lyrata*	Ensembl Plants
12	*Oryza sativa Japonica*	*Oryza glaberrima*	Ensembl Plants
13	*Solanum tuberosum*	*Solanum lycopersicum*	Ensembl Plants
14	*Plasmodium vivax*	*Plasmodium knowlesi*	Ensembl Protists
15	*Phytophthora infestans*	*Phytophthora parasitica*	Ensembl Protists
16	*Methanococcus maripaludis* S2	*Methanococcus maripaludis* C7	ATGC Version 1.0
17	*Lactobacillus johnsonii* NCC 533	*Lactobacillus gasseri* ATCC 33323	ATGC Version 1.0
18	*Lactococcus lactis* subsp. *lactis* Il1403	*Lactococcus lactis* subsp. *cremoris* MG1363	ATGC Version 1.0
19	*Streptococcus pyogenes* M1 GAS	*Streptococcus pyogenes* MGAS315	ATGC Version 1.0
20	*Streptococcus gordonii* str. Challis substr. CH1	*Streptococcus sanguinis* SK36	ATGC Version 1.0
21	*Streptococcus thermophilus* LMD-9	*Streptococcus thermophilus* LMG 18311	ATGC Version 1.0
22	*Clostridium perfringens* ATCC 13124	*Clostridium perfringens* SM101	ATGC Version 1.0
23	Onion yellows phytoplasma OY-M	Aster yellows witches'-broom phytoplasma AYWB	ATGC Version 1.0
24	*Bacteroides fragilis* YCH46	*Bacteroides fragilis* NCTC 9343	ATGC Version 1.0
25	*Borrelia garinii* PBi	*Borrelia afzelii* PKo	ATGC Version 1.0
26	*Corynebacterium glutamicum* ATCC 13032	*Corynebacterium glutamicum* R	ATGC Version 1.0
27	*Dehalococcoides* sp. CBDB1	*Dehalococcoides* sp. BAV1	ATGC Version 1.0
28	*Mycobacterium avium* subsp. *paratuberculosis* K-10	*Mycobacterium avium* 104	ATGC Version 1.0

29	*Anabaena variabilis* ATCC 29413	*Nostoc* sp. PCC 7120	ATGC Version 1.0
30	Prochlorococcus marinus str. AS9601	*Prochlorococcus marinus* str. MIT 9215	ATGC Version 1.0
31	*Prochlorococcus marinus* str. NATL1A	*Prochlorococcus marinus* str. NATL2A	ATGC Version 1.0
32	*Prochlorococcus marinus* str. MIT 9313	*Prochlorococcus marinus* str. MIT 9303	ATGC Version 1.0
33	*Prochlorococcus marinus* str. MIT 9515	*Prochlorococcus marinus* subsp. *pastoris* str. CCMP1986	ATGC Version 1.0
34	*Salinispora arenicola* CNS-205	*Salinispora tropica* CNB-440	ATGC Version 1.0
35	*Thermotoga petrophila* RKU-1	*Thermotoga maritima* MSB8	ATGC Version 1.0
36	*Bacillus anthracis* str. Sterne	*Bacillus cereus* E33L	ATGC Version 1.0
37	*Geobacillus kaustophilus* HTA426	*Geobacillus thermodenitrificans* NG80-2	ATGC Version 1.0
38	*Listeria welshimeri* serovar 6b str. SLCC5334	*Listeria monocytogenes* str. 4b F2365	ATGC Version 1.0
39	*Staphylococcus aureus* subsp. *aureus* MW2	*Staphylococcus aureus* subsp. *aureus* USA300	ATGC Version 1.0
40	*Bradyrhizobium* sp. BTAi1	*Bradyrhizobium* sp. ORS278	ATGC Version 1.0
41	*Candidatus Pelagibacter ubique* HTCC1002	*Candidatus Pelagibacter ubique* HTCC1062	ATGC Version 1.0
42	*Ehrlichia ruminantium* str. Welgevonden	*Ehrlichia ruminantium* str. Gardel	ATGC Version 1.0
43	*Nitrobacter winogradskyi* Nb-255	*Nitrobacter hamburgensis* X14	ATGC Version 1.0
44	*Rhizobium etli* CFN 42	*Rhizobium leguminosarum* bv. *viciae* 3841	ATGC Version 1.0
45	*Rhodobacter sphaeroides* ATCC 17025	*Rhodobacter sphaeroides* ATCC 17029	ATGC Version 1.0
46	*Rhodopseudomonas palustris* HaA2	*Rhodopseudomonas palustris* BisB5	ATGC Version 1.0
47	*Rickettsia typhi* str. Wilmington	*Rickettsia prowazekii* str. Madrid E	ATGC Version 1.0
48	*Rickettsia canadensis* str. McKiel	*Rickettsia felis* URRWXCal2	ATGC Version 1.0
49	*Sinorhizobium medicae* WSM419	*Sinorhizobium meliloti* 1021	ATGC Version 1.0
50	*Bordetella pertussis* Tohama I	*Bordetella bronchiseptica* RB50	ATGC Version 1.0
51	*Burkholderia cenocepacia* AU 1054	*Burkholderia ambifaria* AMMD	ATGC Version 1.0
52	*Neisseria meningitidis* MC58	*Neisseria meningitidis* FAM18	ATGC Version 1.0
53	*Campylobacter jejuni* subsp. *jejuni* NCTC 11168	*Campylobacter jejuni* subsp. *jejuni* 81116	ATGC Version 1.0
54	*Helicobacter acinonychis* str. Sheeba	*Helicobacter pylori* J99	ATGC Version 1.0
55	*Aeromonas hydrophila* subsp. *hydrophila* ATCC 7966	*Aeromonas salmonicida* subsp. *salmonicida* A449	ATGC Version 1.0

56	*Salmonella enterica* subsp. *enterica* serovar Choleraesuis str. SC-B67	*Salmonella enterica* subsp. *enterica* serovar Typhi str. Ty2	ATGC Version 1.0
57	*Shigella sonnei* Ss046	*Shigella flexneri* 2a str. 301	ATGC Version 1.0
58	*Escherichia coli* K12	*Escherichia coli* APEC O1	ATGC Version 1.0
59	*Haemophilus influenzae* PittGG	*Haemophilus influenzae* 86-028NP	ATGC Version 1.0
60	*Legionella pneumophila* str. Corby	*Legionella pneumophila* subsp. *pneumophila* str. Philadelphia 1	ATGC Version 1.0
61	*Pseudomonas aeruginosa* PAO1	*Pseudomonas aeruginosa* PA7	ATGC Version 1.0
62	*Pseudomonas putida* F1	*Pseudomonas putida* KT2440	ATGC Version 1.0
63	*Pseudomonas syringae* pv. *syringae* B728a	*Pseudomonas syringae* pv. tomato str. DC3000	ATGC Version 1.0
64	*Psychrobacter arcticus* 273-4	*Psychrobacter cryohalolentis* K5	ATGC Version 1.0
65	*Vibrio vulnificus* CMCP6	*Vibrio vulnificus* YJ016	ATGC Version 1.0
66	*Xanthomonas campestris* pv. *campestris* str. 8004	*Xanthomonas campestris* pv. *vesicatoria* str. 85-10	ATGC Version 1.0
67	*Xylella fastidiosa* Temecula1	*Xylella fastidiosa* 9a5c	ATGC Version 1.0
68	*Desulfovibrio vulgaris* subsp. *vulgaris* DP4	*Desulfovibrio vulgaris* subsp. *vulgaris* str. Hildenborough	ATGC Version 1.0

Table A.4.2. 11 pairs of mammalian clades with orthologous coding sequences.

	Clade 1		Clade 2	
	Species 1	Species 2	Species 1	Species 2
1	*Homo sapiens*	*Pan troglodytes*	*Macaca mulatta*	*Chlorocebus sabaeus*
2	*Homo sapiens*	*Pan troglodytes*	*Otolemur garnettii*	*Microcebus murinus*
3	*Homo sapiens*	*Pan troglodytes*	*Mus musculus*	*Rattus norvegicus*
4	*Homo sapiens*	*Pan troglodytes*	*Oryctolagus cuniculus*	*Ochotona princeps*
5	*Homo sapiens*	*Pan troglodytes*	*Bos taurus*	*Sus scrofa*
6	*Homo sapiens*	*Pan troglodytes*	*Canis familiaris*	*Felis catus*
7	*Homo sapiens*	*Pan troglodytes*	*Myotis lucifugus*	*Pteropus vampyrus*
8	*Homo sapiens*	*Pan troglodytes*	*Erinaceus europaeus*	*Sorex araneus*
9	*Homo sapiens*	*Pan troglodytes*	*Dasypus novemcinctus*	*Choloepus hoffmanni*
10	*Homo sapiens*	*Pan troglodytes*	*Loxodonta africana*	*Procavia capensis*
11	*Homo sapiens*	*Pan troglodytes*	*Monodelphis domestica*	*Sarcophilus harrisii*

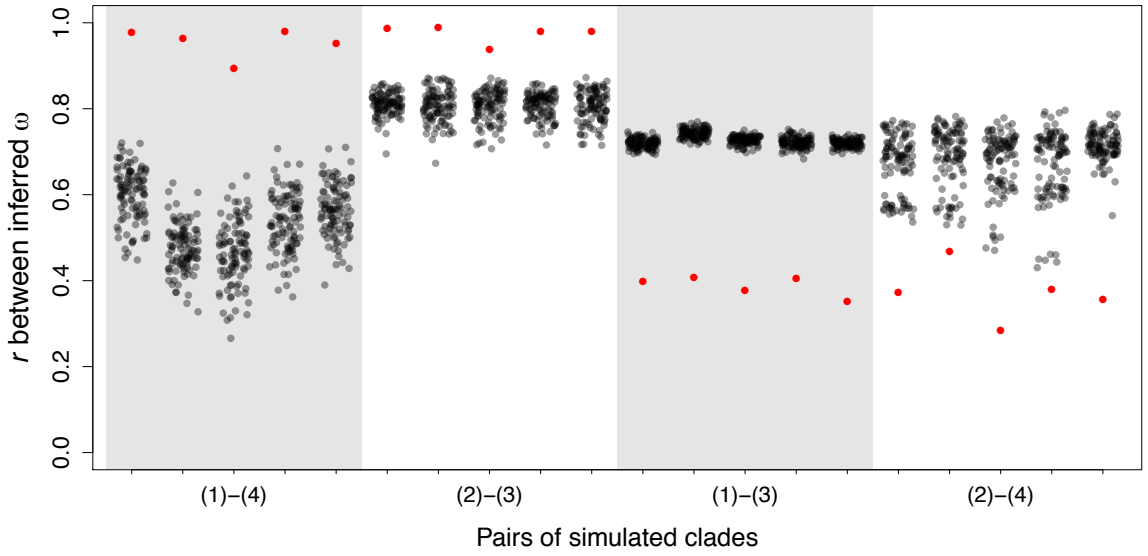


Figure A.4.1. Shuffling tests between simulated clades with different combinations of parameter sets. Each column is a shuffling test. Labels under X axis indicate the two parameter sets being compared, corresponding to (1) – (4) in the main text and the four columns in **Fig. 6.1a**. Five shuffling tests were conducted for each pair of parameter sets, between five independent replicate simulations for (1) and five for (2). Red dots indicate the Pearson correlation coefficients between $\boldsymbol{\omega}$s of two clades. 100 grey dots represent *r*'s between $\boldsymbol{\omega}$s of two clades after 100 independent alignment shuffling. If a red dot falls within 5% lower tail of the grey dots distribution, i.e., there being no more than four grey dots below it, a red asterisk is plotted to indicate significant smaller acceptance rate correlation than shuffled control.

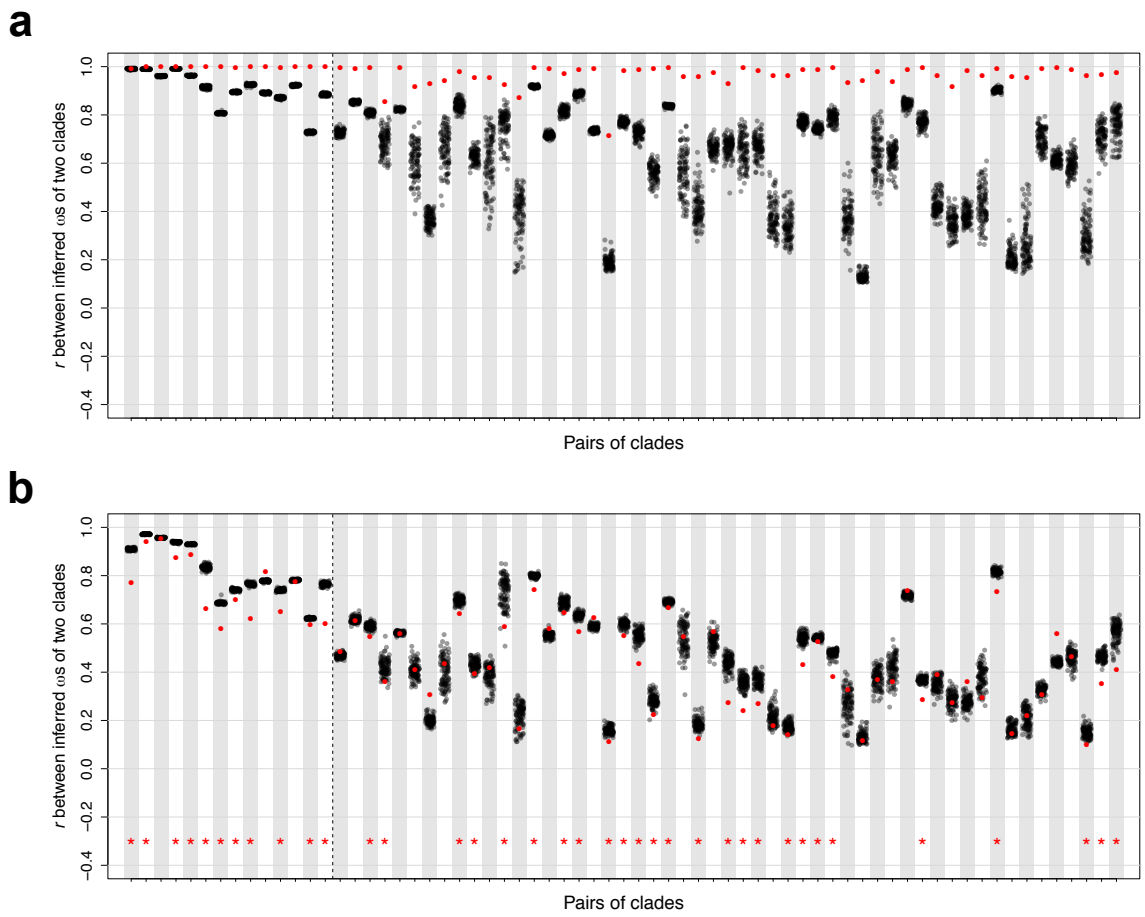


Figure A.4.2. Shuffling tests between simulated rodents clade and the other 67 simulated clades (**a**) with the same rodents ***ω*** and (**b**) with different ***ω***s inferred from real data. Axes, labels and color schemes follow those of **Fig. 6.3** and **Fig. A.4.1**.

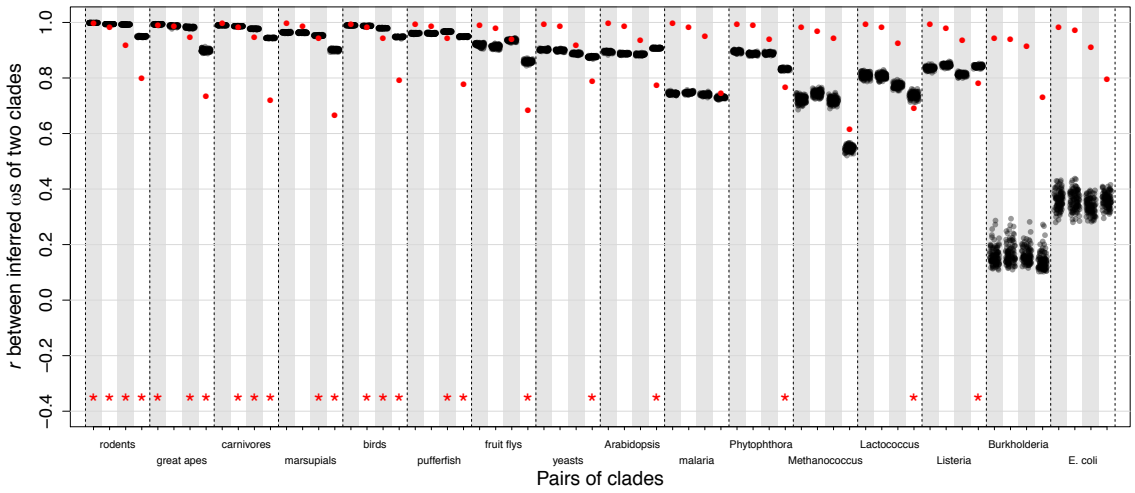


Figure A.4.3. Shuffling tests between simulated clades with *ω*s varied from the rodents *ω*, and a clade simulated with all parameters inferred from the rodents clade. Each four columns bounded by dash lines correspond to four simulated clades, with ***ω*** varying from the rodents ***ω*** by 5%, 10%, 20% and 50% (see MATERIALS AND METHODS), and other parameters from the clade indicated below the X axis. Axes, labels and color schemes follow those of **Fig. 6.3** and **Fig. A.4.1-2**.

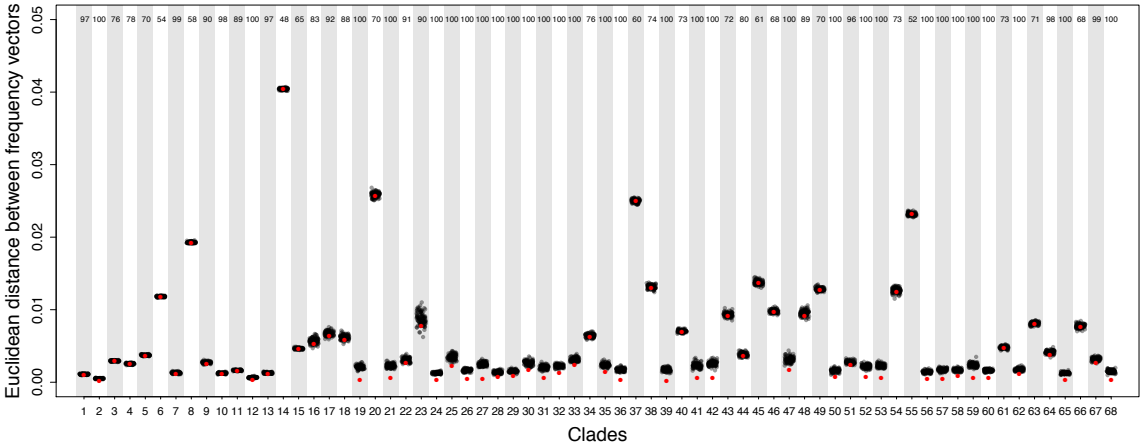


Figure A.4.4. Euclidean distance of codon frequencies between species within clade is no larger than bootstrap controls. In each column, red dot indicates the Euclidean distance between codon frequency vectors of two species in the clade. Grey dots indicate 100 Euclidean distances, each between codon frequency vectors of two species after bootstrapping all codon positions in the alignment. The number at the top of each column indicates the number of grey dots above the red dot, which supports that codon frequency difference within clade is smaller than the bootstrap control.

A.5 Supplementary table and figures for Chapter 7

Table A.5.1. 68 clades used for η inference.

	Clade		Data source
	Species 1	Species 2	
1	*Mus musculus*	*Rattus norvegicus*	OrthoMaM v9
2	*Homo sapiens*	*Pan troglodytes*	OrthoMaM v9
3	*Canis lupus familiaris*	*Felis catus*	OrthoMaM v9
4	*Monodelphis domestica*	*Sarcophilus harrisii*	OrthoMaM v9
5	*Gallus gallus*	*Meleagris gallopavo*	Ensembl 84
6	*Takifugu rubripes*	*Tetraodon nigroviridis*	Ensembl 84
7	*Drosophila sechellia*	*Drosophila simulans*	Flybase
8	*Atta cephalotes*	*Solenopsis invicta*	Ensembl Metazoa
9	*Saccharomyces cerevisiae*	*Saccharomyces paradoxus*	http://www.saccharomycessensustricto.org/
10	*Fusarium graminearum*	*Fusarium pseudograminearum*	Ensembl Fungi
11	*Arabidopsis thaliana*	*Arabidopsis lyrata*	Ensembl Plants
12	*Oryza sativa Japonica*	*Oryza glaberrima*	Ensembl Plants
13	*Solanum tuberosum*	*Solanum lycopersicum*	Ensembl Plants
14	*Plasmodium vivax*	*Plasmodium knowlesi*	Ensembl Protists
15	*Phytophthora infestans*	*Phytophthora parasitica*	Ensembl Protists
16	*Methanococcus maripaludis* S2	*Methanococcus maripaludis* C7	ATGC Version 1.0
17	*Lactobacillus johnsonii* NCC 533	*Lactobacillus gasseri* ATCC 33323	ATGC Version 1.0
18	*Lactococcus lactis* subsp. *lactis* Il1403	*Lactococcus lactis* subsp. *cremoris* MG1363	ATGC Version 1.0
19	*Streptococcus pyogenes* M1 GAS	*Streptococcus pyogenes* MGAS315	ATGC Version 1.0
20	*Streptococcus gordonii* str. Challis substr. CH1	*Streptococcus sanguinis* SK36	ATGC Version 1.0
21	*Streptococcus thermophilus* LMD-9	*Streptococcus thermophilus* LMG 18311	ATGC Version 1.0
22	*Clostridium perfringens* ATCC 13124	*Clostridium perfringens* SM101	ATGC Version 1.0
23	Onion yellows phytoplasma OY-M	Aster yellows witches'-broom phytoplasma AYWB	ATGC Version 1.0
24	*Bacteroides fragilis* YCH46	*Bacteroides fragilis* NCTC 9343	ATGC Version 1.0
25	*Borrelia garinii* PBi	*Borrelia afzelii* PKo	ATGC Version 1.0
26	*Corynebacterium glutamicum* ATCC 13032	*Corynebacterium glutamicum* R	ATGC Version 1.0
27	*Dehalococcoides* sp. CBDB1	*Dehalococcoides* sp. BAV1	ATGC Version 1.0
28	*Mycobacterium avium* subsp. *paratuberculosis* K-10	*Mycobacterium avium* 104	ATGC Version 1.0

29	*Anabaena variabilis* ATCC 29413	*Nostoc* sp. PCC 7120	ATGC Version 1.0
30	Prochlorococcus marinus str. AS9601	*Prochlorococcus marinus* str. MIT 9215	ATGC Version 1.0
31	*Prochlorococcus marinus* str. NATL1A	*Prochlorococcus marinus* str. NATL2A	ATGC Version 1.0
32	*Prochlorococcus marinus* str. MIT 9313	*Prochlorococcus marinus* str. MIT 9303	ATGC Version 1.0
33	*Prochlorococcus marinus* str. MIT 9515	*Prochlorococcus marinus* subsp. *pastoris* str. CCMP1986	ATGC Version 1.0
34	*Salinispora arenicola* CNS-205	*Salinispora tropica* CNB-440	ATGC Version 1.0
35	*Thermotoga petrophila* RKU-1	*Thermotoga maritima* MSB8	ATGC Version 1.0
36	*Bacillus anthracis* str. Sterne	*Bacillus cereus* E33L	ATGC Version 1.0
37	*Geobacillus kaustophilus* HTA426	*Geobacillus thermodenitrificans* NG80-2	ATGC Version 1.0
38	*Listeria welshimeri* serovar 6b str. SLCC5334	*Listeria monocytogenes* str. 4b F2365	ATGC Version 1.0
39	*Staphylococcus aureus* subsp. *aureus* MW2	*Staphylococcus aureus* subsp. *aureus* USA300	ATGC Version 1.0
40	*Bradyrhizobium* sp. BTAi1	*Bradyrhizobium* sp. ORS278	ATGC Version 1.0
41	*Candidatus Pelagibacter ubique* HTCC1002	*Candidatus Pelagibacter ubique* HTCC1062	ATGC Version 1.0
42	*Ehrlichia ruminantium* str. Welgevonden	*Ehrlichia ruminantium* str. Gardel	ATGC Version 1.0
43	*Nitrobacter winogradskyi* Nb-255	*Nitrobacter hamburgensis* X14	ATGC Version 1.0
44	*Rhizobium etli* CFN 42	*Rhizobium leguminosarum* bv. *viciae* 3841	ATGC Version 1.0
45	*Rhodobacter sphaeroides* ATCC 17025	*Rhodobacter sphaeroides* ATCC 17029	ATGC Version 1.0
46	*Rhodopseudomonas palustris* HaA2	*Rhodopseudomonas palustris* BisB5	ATGC Version 1.0
47	*Rickettsia typhi* str. Wilmington	*Rickettsia prowazekii* str. Madrid E	ATGC Version 1.0
48	*Rickettsia canadensis* str. McKiel	*Rickettsia felis* URRWXCal2	ATGC Version 1.0
49	*Sinorhizobium medicae* WSM419	*Sinorhizobium meliloti* 1021	ATGC Version 1.0
50	*Bordetella pertussis* Tohama I	*Bordetella bronchiseptica* RB50	ATGC Version 1.0
51	*Burkholderia cenocepacia* AU 1054	*Burkholderia ambifaria* AMMD	ATGC Version 1.0
52	*Neisseria meningitidis* MC58	*Neisseria meningitidis* FAM18	ATGC Version 1.0
53	*Campylobacter jejuni* subsp. *jejuni* NCTC 11168	*Campylobacter jejuni* subsp. *jejuni* 81116	ATGC Version 1.0
54	*Helicobacter acinonychis* str. Sheeba	*Helicobacter pylori* J99	ATGC Version 1.0
55	*Aeromonas hydrophila* subsp. *hydrophila* ATCC 7966	*Aeromonas salmonicida* subsp. *salmonicida* A449	ATGC Version 1.0

56	*Salmonella enterica* subsp. *enterica* serovar Choleraesuis str. SC-B67	*Salmonella enterica* subsp. *enterica* serovar Typhi str. Ty2	ATGC Version 1.0
57	*Shigella sonnei* Ss046	*Shigella flexneri* 2a str. 301	ATGC Version 1.0
58	*Escherichia coli* K12	*Escherichia coli* APEC O1	ATGC Version 1.0
59	*Haemophilus influenzae* PittGG	*Haemophilus influenzae* 86-028NP	ATGC Version 1.0
60	*Legionella pneumophila* str. Corby	*Legionella pneumophila* subsp. *pneumophila* str. Philadelphia 1	ATGC Version 1.0
61	*Pseudomonas aeruginosa* PAO1	*Pseudomonas aeruginosa* PA7	ATGC Version 1.0
62	*Pseudomonas putida* F1	*Pseudomonas putida* KT2440	ATGC Version 1.0
63	*Pseudomonas syringae* pv. *syringae* B728a	*Pseudomonas syringae* pv. tomato str. DC3000	ATGC Version 1.0
64	*Psychrobacter arcticus* 273-4	*Psychrobacter cryohalolentis* K5	ATGC Version 1.0
65	*Vibrio vulnificus* CMCP6	*Vibrio vulnificus* YJ016	ATGC Version 1.0
66	*Xanthomonas campestris* pv. *campestris* str. 8004	*Xanthomonas campestris* pv. *vesicatoria* str. 85-10	ATGC Version 1.0
67	*Xylella fastidiosa* Temecula1	*Xylella fastidiosa* 9a5c	ATGC Version 1.0
68	*Desulfovibrio vulgaris* subsp. *vulgaris* DP4	*Desulfovibrio vulgaris* subsp. *vulgaris* str. Hildenborough	ATGC Version 1.0

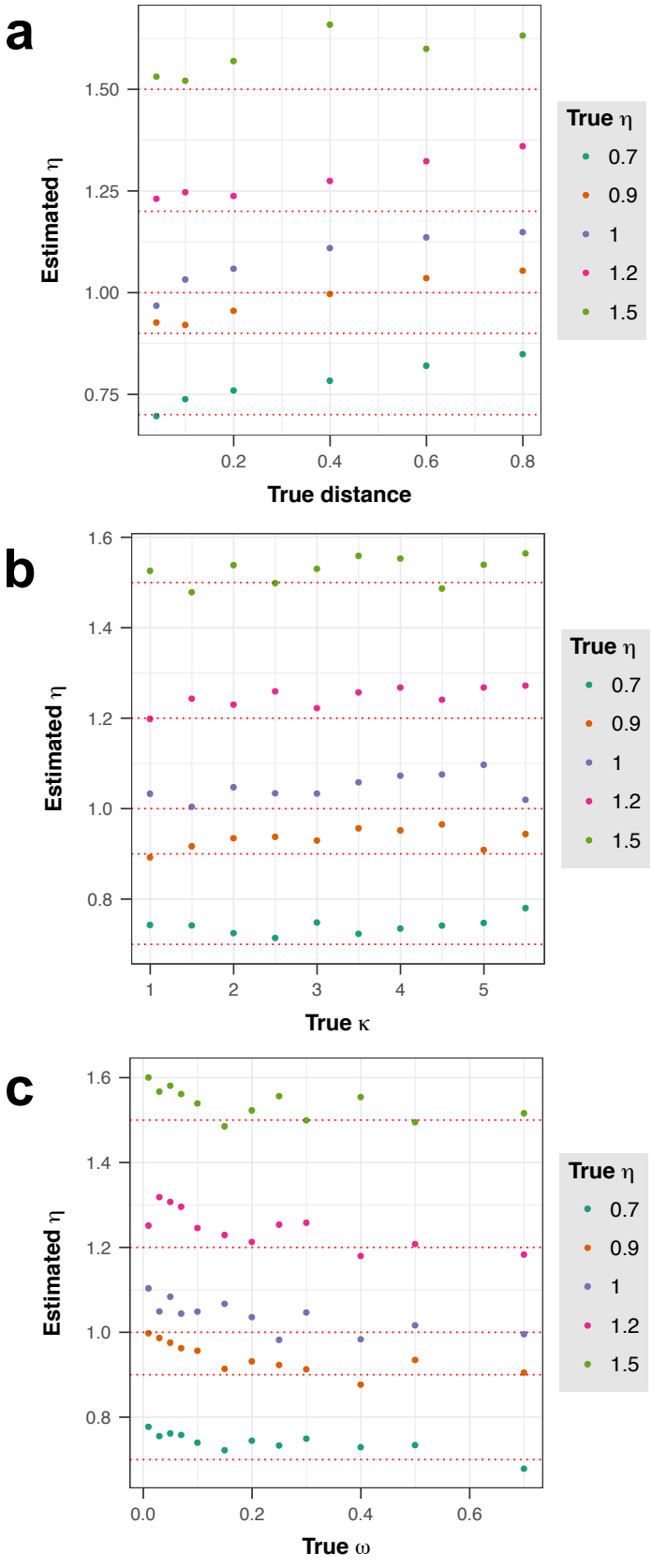

a
Estimated η
1.50
1.25
1.00
0.75
0.2
0.4
0.6
0.8
True distance
True η
0.7
0.9
1
1.2
1.5
b
Estimated η
1.6
1.4
1.2
1.0
0.8
1
2
3
4
5
True κ
True η
0.7
0.9
1
1.2
1.5
c
Estimated η
1.6
1.4
1.2
1.0
0.8
0.0
0.2
0.4
0.6
True ω
True η
0.7
0.9
1
1.2
1.5

Figure A.5.1. The inferred ηs show deviation from the true values in simulation when site-specificity of evolutionary rate is considered, together with their correlation with the other varied parameters, such as (**a**) genetic distance between two species in the same clade, (**b**) transition/transversion mutational bias κ, and (**c**) overall selection ω. Each dot is one η estimation plotted against the true value of another parameter used during simulation. True value of each η is indicated by colors shown in legend, dotted lines correspond to the true value for clear comparison. In this analysis, more true values of η were used, but only five are plotted here for clarity.

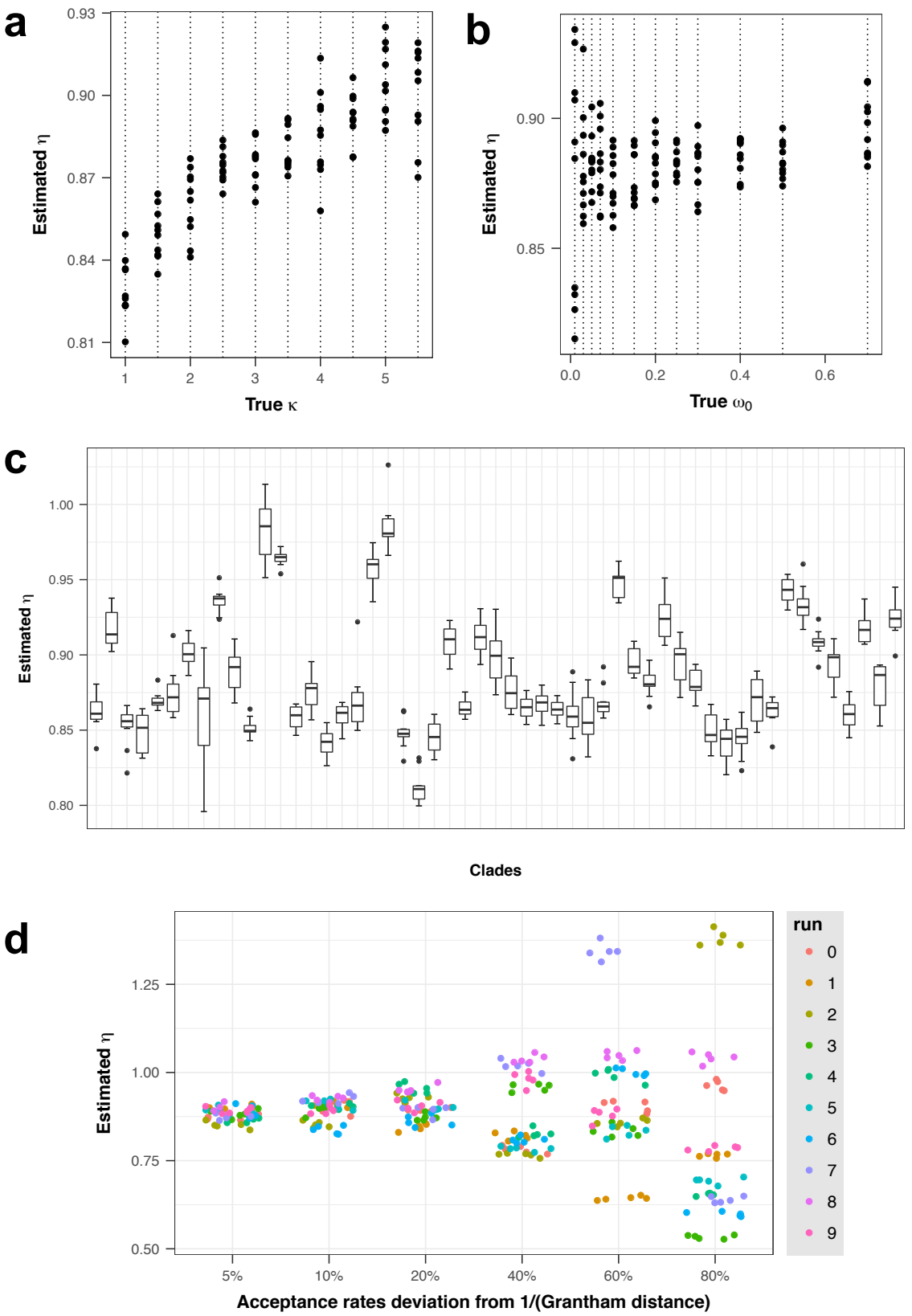

a
Estimated η
0.93
0.90
0.87
0.84
0.81
1
2
3
4
5
True κ
b
Estimated η
0.90
0.85
0.0
0.2
0.4
0.6
True ω0
c
Estimated η
1.00
0.95
0.90
0.85
0.80
Clades
d
Estimated η
1.25
1.00
0.75
0.50
5%
10%
20%
40%
60%
80%
Acceptance rates deviation from 1/(Grantham distance)
run
0
1
2
3
4
5
6
7
8
9

Figure A.5.2. When site-specificity of evolutionary rate is considered, variations in κ, ω_0, or π still cannot explain the large variation in η among clades while variation of acceptance rates ω' can. η's inferred from simulated sequence alignments are plotted against the true (**a**) κ's, (**b**) ω_0's or (**c**) π's. Dashed vertical lines in (**a**) and (**b**) indicate the values of true κ's or ω's specified in simulation. For each value, η estimations from 10 replicate simulations are plotted. In (**c**), the codon frequencies of the 53 prokaryotic clades are used, and boxplots show distribution of η estimations from 10 replicate simulations. (**b**) Simulations with ω' derived from Grantham matrix. For each of the 10 runs at a certain percentage level x%, each acceptance rate $\omega'(aa_i, aa_j)$ was either added or subtracted x% of its reference value (1/Grantham distance) randomly and then used for simulation. Different runs are color-coded according to the legend. For each run, five replicate sequence evolution simulations were conducted and corresponding η plotted.