博士論文

Statistical modeling of species and population level genomic evolution

(種と集団レベルのゲノム進化の統計モデリング)

呉佳齊

(Jiaqi Wu)

Statistical modeling of species and population level genomic evolution

By

Jiaqi Wu

A thesis submitted for the degree of Doctor of Philosophy

in the Department of Agricultural and Environmental Biology of

the Graduate School of Agricultural and Life Sciences

The University of Tokyo

Supervisor: Prof. Kishino Hirohisa

March 2017

Abstract

Descripting and understanding the differences of the living organisms surrounding us is one of the central activities of human through histories. From era of Charles Darwin, especially owing to his publication of ***On the origin of species*** (1859), evolution became a basis that explains the diversity of living organisms. Darwin generated his theory based on large amount of observations of phenotypic variations, and emphasized natural selection is the main power for species evolution. In 1968, Motoo Kimura calculated the rate of genomic evolution and found it is much faster than Haldane's limit on the speed of beneficial evolution. This finding gave birth to the neutral theory of molecular evolution. The neutral theory declares that mutations at the molecular level are largely neutral and it became the central idea of molecular evolution and population genetics. In genomic era, it is possible to develop a new approach to link the molecular evolution with the phenotypic variations of living organisms.

A variant of a gene or an allele can be considered as a state of a trait of a genome. Summary statistics with careful statistical modeling can be applied to study the genomic variations. In this thesis, I summarize the current inference methods of evolutionary biology (Chapter 2.1-2.4) and propose a new model, which extends the neutral theory of molecular evolution to describe genome evolution (Chapter 2.5). Also I report the result of population genomic analysis based on the model of the joint allele frequency spectrum (Chapter 5). The former investigates species level genomic evolution and the latter investigates population level genomic evolution. Both studies describe models of the summary statistics to infer genomic evolution. The new model of species level genomic evolution enables robust estimation of the divergence times and links the gene variations and the life history traits evolution of mammals.

In Chapters 3 and 4, I formulate the models in more detail by analysing the evolutionary rates of 1,185 genes on a phylogeny of 89 mammals. In Chapter 3, I show the branch effect-based divergence time estimation approach provides robust estimate of divergence times. Remarkably, DNA, codon or protein-level analyses give the same result. Also this measures firstly the variation of genomic evolutionary rate among species. In Chapter 4, I show the efficacy of the new rate-based ancestral state reconstruction approach. By using this new approach,

I reconstructed the history of 10 discrete traits related to activity, diet and social behaviors. The results indicate that the ancestor of placental mammals was solitary, seasonally breeding, insectivorous and likely nocturnal. The predictor genes of the traits can be automatically selected from genomic data, without relying on the pre-knowledge of gene annotations. This approach has a potential to link the Darwin's natural selection theory of phenotypic variations and Kimura's neutral theory of molecular evolution. The method with its case study introduced in this thesis will hopefully inspire the field of genomic study.

Table of contents

Chapter 1. General Introduction

Descripting and understanding the differences of the living organisms surrounding us is one of the central activities of human through histories. The earliest systematic work can be traced back to Aristotle (350 B. C.), who firstly described life histories of many animals and tried to explain the differences [1-5]. From era of Charles Darwin, especially owing to his publication of ***On the origin of species*** (1859) [6], evolution became a basis that explains the diversity of living organisms. The main ideas of Darwin's Theory of evolution include [7]: 1, all living organisms are descendant of one or a few common ancestors (common ancestor theory); 2, at each generation, individuals reproduce offspring that can survive; 3, phenotypic variations of individuals are heritable; 4, only the individuals who can adapt to the environment survive (natural selection). These ideas were generated gradually based on a large amount of observations during Darwin's survey voyage, and each idea has solid evidences summarized in his book. Darwin was unaware of the genetic material of living organisms, and his conclusions were based on the observation of phenotypic traits including anatomic structure and life histories.

In 1866, Gregor Mendel published ***Experiments in plant hybridization*** and reported his work on the pea-plant experiments [8], as one of the first quantitative biological study. In his experiment, the proportion of each phenotypic variation can be precisely predicted based on the rule now we call Mendel's laws of inheritance. Another pioneer in quantitative biology is Darwin's cousin, Francis Galton, who found that phenotypic variation could be approximated by Gaussian distributions [9]. Same with Darwin, both Gregor Mendel and Francis Galton were unaware of the genetic material of living organisms, however, they employed mathematics novel to biology, and found interesting rules under the phenotypic variations, which gave hint to the attributes of the "genetic material". This made a great step though the development of modern biology.

Mendel's laws were unvalued by his contemporary scientists until it was rediscovered in 1900 [10]. The word "genetics", "gene" and "allele" also appeared to describe the study of inheritance and its main concepts around this time [11]. Mendel's laws were integrated with the "chromosome theory of inheritance" later by Thomas Hunt Morgan in 1915 and became core of genetics [12]. Afterward, population genetics

was founded by Ronald Fisher, John B.S. Haldane and Sewall Wright. Ronald Fisher published ***The Genetical Theory of Natural Selection*** in 1930 [13]. In this book, he combined Mendel's ideas with natural selection theory and put evolution onto a mathematical footing [12]. Sewall Wright formulated the stochastic processes of genetic diversity and population structure [14]. After the series of studies in population genetics, John B.S. Haldane calculated the cost of natural selection and proposed that there is a limit on the speed of advantageous substitutions [15].

In 1950s, DNA structure was discovered and was proved to be the "genetic material" of living organisms [16], also some technique to study protein variation became available. In 1962, Zuckerkandl and Pauling published their results on a new technique to identify haemoglobin protein variants, and firstly found the molecular clock, that is, the pace of molecular evolution is constant as a first approximation [17]. In 1968, Motoo Kimura calculated the rate of genomic evolution and found it is much faster than Haldane's limit on the speed of beneficial evolution. This finding gave birth to the neutral theory of molecular evolution [18]. Neutral theory that declares mutations in molecular level are largely neutral became central idea of molecular evolution and population genetics.

Evolution has timescale. Molecular evolutionary examines the difference among species, while population genetics studies the variation among individuals within a species. The pattern of molecular evolution is largely expressed by molecular phylogenetic tree. Its topology describes the order of speciation and the branch lengths describe the amounts of evolution between the successive speciation events. In population genetics, a genealogy describes the relation of genes sampled from a population. However, the information content of the genetic data on the genealogy is scarce, because genetic variation within a single species is much smaller than genetic variation between species. Therefore, the information on genealogy is largely summarized by a set of coalescent times and allele frequencies.

Data available for evolutionary study evolved from alleles to sequences and now large amount of genes and genomes are being registered in database day by day. Statistical modeling in molecular evolution and population genetics evolves with the technological advancement of data acquisition. For example, the species trees were previously based on single gene analysis [19-21], while now many studies use

genomic data to infer the species tree [22-24]. With big data, the results of evolutionary analysis became more stable and new methods are also developing [25].

Based on a large amount of observations of phenotypic variations, Darwin emphasized natural selection is the main force of species evolution. In fact, a recent study tested the neutral hypothesis of phenotypic evolution using 210 character traits of yeast, and found that morphological variations are largely adaptive [26], supports Darwin's natural selection theory. Kimura was fully aware of the importance of natural selection on phenotypic evolution. He focused on molecular evolution, and emphasized that the variations observed in molecular level are mainly due to the random fixation of mutations. In genomic era, it is possible to develop a new approach to link the molecular evolution with the phenotypic variations of living organism.

A variant of a gene or an allele can be considered as a state of a trait of a genome. Summary statistics with careful statistical modeling can be applied to study the genomic variations. In Chapter 2, after describing briefly the current well-accepted methods in phylogenetics, I propose a new model of branch lengths, which extends the neutral theory of Kimura. In Chapter 5, I introduce a study based on the model of the joint allele frequency spectrum. Both Chapter 2.5 and Chapter 5 describe models of the summary statistics to infer genomic evolution. Chapter 3 and 4 formulate the models in more detail by analysing the evolutionary rates of 1,185 genes on a phylogeny of 89 mammals. In Chapter 3, I show that the branch effect-based divergence time estimation approach provides robust estimate of divergence times. Remarkably, DNA, codon or protein-level analysis give the same result. Also this measures firstly the variation of genomic rate among species. In Chapter 4, I show the efficacy of the new rate-based ancestral state reconstruction approach. By using this new approach, I reconstructed the history of 10 discrete traits related to activity, diet and social behaviours. The results indicate the ancestor of placental mammals was solitary, seasonally breeding, insectivorous and likely nocturnal. The predictor genes of the traits can be automatically selected from genomic data, without relying on the pre-knowledge of gene annotations. This approach has a potential to link the Darwin's natural selection theory of phenotypic

variations and Kimura's neutral theory of molecular evolution. The method with its case study introduced in this thesis will hopefully inspire the field of genomic study.

Chapter 2. Methods of phylogenetic inference and an extension of the neutral theory to multiple-gene molecular evolution

Phylogenetics is a research field that studies the evolutionary history and relationships among organisms. Its major topics include 1, inference of the phylogenetic relations among species, 2, estimation of the divergence times, 3, inference of the evolutionary history of morphology and life histories of species, and 4, detection of positive selection and functional constraints behind adaptive evolution. This chapter first summarizes the core ideas of the inference methods and some of the well accepted and widely used approaches.

In Section 2.5, I propose a new model of molecular evolutionary rates that extends the neutral theory of Motoo Kimura, and describes the evolution of multiple genes in the genomes. This model provides the basis of new alternative approaches of divergence time estimation and reconstruction of the ancestral traits. In later chapters, I will show its efficiency in comparison with the existing methods.

2.1 Distance based and trait based methods of phylogenetic tree inferences

The relation among species is represented by a phylogenetic tree. The origin of this idea goes back to Charles Darwin (***the original of species***), who considered that all organisms on earth share a common ancestor. A phylogenetic tree has branches and nodes. The branch length expresses the amount of evolution along the branch. In molecular evolution, it is the expected number of nucleotide (or amino acid) change per site. It is the product of the evolutionary rate (the expected number of nucleotide or amino acid change per site per unit time) and the time duration along the branch (normally in million years, Mya). The methods for phylogenetic tree inference can be largely classified into two groups: the distance-based methods and the character state-based methods. The trait-based methods can be further classified into three groups: maximum parsimony, maximum likelihood and Bayes methods [27, 28].

2.1.1 Distance-based methods

One of the first distance-based method of phylogenetic inference was proposed by Fitch and Margoliash in 1967 [29]. It starts with calculation of evolutionary distances between the pairs of homologous sequences, which are defined as the expected numbers of substitutions per site. The simplest estimator is the p-distance, the

proportion of different sites. The p-distance expresses well the amount evolution during short time such as evolution within species, because the chance of multiple substitutions at an identical site is low. As for the between-species comparison, the correction of multiple substitutions becomes indispensable. Markov substitution models such as JC69, HKY85 or GTR are used to correct the multiple substitutions [30-32] (see figure 1 in [28]). Furthermore, the effect of heterogeneity of the substitution rates among sites can be incorporated by a gamma distribution (Γ) of rates among sites and/or the the proportion of invariable sites (I) [33].

The most highly established approach is Neighbour Joining method (NJ), which was created by Saitou and Nei in 1987 [34]. Starting from a star tree, it joins, at each step, the pair of groups of sequences that minimizes the tree length, and finally results in a tree that is expected to have globally minimum tree length (see figure 2 in [28]).

2.1.2 Character state-based methods

Character state-based phylogenetic tree inference methods have three major branches: maximum parsimony, maximum likelihood and Bayes methods. Maximum Parsimony methods are based on the principle of "Ockham's razor" that states "the simplest hypothesis proposed as an explanation of phenomena is more likely to be true one than is any other available hypothesis [35]". In this principle of the parsimony, a phylogenetic tree that can be explained by the minimum numbers of total character-state changes is regarded as the best tree. Because it counts the "character-state changes", it uses only the sites that have at least two character-states, and at least two of them are shared by at least two sequences, the so-called parsimony informative sites. This approach makes it possible to trace back the states of the sequences to the past. However, in the presence of multiple changes at sites, the estimates result may not be reliable regarding with the relationship of deep nodes in a phylogenetic tree [36].

Maximum likelihood method was firstly developed by R. A. Fisher as a statistical method to estimate the unknown parameters in a model. The likelihood function is a function of parameters given the data. Maximum likelihood method for tree inference was firstly proposed by Cavalli-sforza and Edwards 1967 [37]. The computer program with efficient algorithms traces back to Felsenstein in 1981 [38]. The

parameters to be estimated are the tree topology, branch lengths and the model of substitution rate matrix. The log likelihood of aligned sequences is the sum of the log likelihood of alignment columns. With the model that incorporates the gamma-distributed rate-heterogeneity among sites and the proportion of invariant sites, it is expressed as:

$$l = \log L = \sum_{1}^{n} \log\{f(\text{alignment}_i \mid T, B, \theta, \Gamma, \mathrm{I})\}. \quad (1)$$

Here T is the tree topology, B is the branch lengths, θ is the parameters for substitution matrix, Γ is the parameter for gamma distribution of rate heterogeneity, and I is the parameter for the proportion of invariable sites. Models of molecular evolution can be compared by AIC (Akaike's information criterion [39]) or BIC (Bayesian information criterion [40]). It is important to note that the maximum parsimony approach can be interpreted in the framework of maximum likelihood approach [41].

Bayes methods consider the uncertainty of the parameters as the probability distributions of parameters. Prior distributions are the distribution of parameters before data are analyzed, while the conditional probabilities (the posterior probabilities) are calculated by given the data [42]. Bayesian method for phylogenetic analysis was firstly introduced in 1990s by Rannala and Yang [43, 44], Mau and Newton [45], and Li *et al*., [46]. The early methods assumed molecular clock, while later methods allowed independent branch lengths on the un-rooted trees [47]. Normally, it is difficult to calculate the posterior probabilities of trees directly, thus Markov chain Monte Carlo (MCMC) algorithm is used to generate a sample from the posterior distribution of the tree [27, 28].

2.2 Divergence time estimation by relaxed molecular clock

In 1962, Zuckerkandl and Pauling firstly reported that the molecular rate is constant though time and lineages [17]. If evolution in molecular level follows a constant rate and we know the time of some nodes form fossil records, we could obtain the absolute divergence times of species. As the growing number of sequence data became available, it was recognized that tick of molecular clock is not strictly

constant but fluctuates stochastically. This observation provoked the modeling of relaxed molecular clock.

2.2.1 Variation of molecular rates and relaxed molecular clock

The variation of molecular rates had been long time under concern [48]. The molecular rates differ among branches, and also among genes. The molecular evolutionary rate is affected by multiple factors such as generation time [49-51]. The model of local molecular clock classifies the branches into the categories of evolutionary rates. The constant rate of molecular evolution is applied to the branches of each category [52]. The other type of relaxed molecular clock models assigns the different values of evolutionary rate to the branches by setting the penalty against the departure from molecular clock [53]. There are two kinds of penalty: independent rate model and correlated rate model. The former assumes the variations of the evolutionary rates are independent among branches, while the latter assumes that a rate along a branch is correlated with the rate along its ancestral branch.

2.2.2 Bayesian estimation of divergence times

Bayesian method for divergence time estimations was firstly introduced by Thorne and his colleagues [54-56], and was developed extensively since then [44, 57-59]. Its strength is integrating multiple sources of information such as sequence information and fossil calibrations. The posterior of times and rates given by Bayes theorem is [42]:

$$f(t,r \mid D) = \frac{1}{z} f(t) f(r \mid t) L(D \mid t,r). \qquad (2)$$

Here D is the sequence data, t is the divergence times, r is the molecular rates, and z is the proportionality constant. $f(t)$ refers to the prior of divergence times, which is usually integrated with fossil calibrations [55, 58]; $f(r|t)$ is the prior on the molecular rates on a tree [54, 57, 60]; while $L(D|t,r)$ is the likelihood of tree and rates based on the sequence data [38]. The setting of priors and the uncertainty of fossil calibrations may affect the estimated times, thus is of major concern in the world of Bayesian time estimation. A detailed review can be found in [42] with well-designed examples.

2.3 Inference of trees and times by multiple genomic loci: partition or coalescent

2.3.1 Concatenate and partition approaches

In the early stage of phylogenetic analysis, single or limited genetic loci (e.g. mitochondrial genes or rRNA) were used in the analysis [19-21]. When genomic data accumulate, it became possible to estimate the phylogenetic tree based on multiple loci. As mentioned in Chapter 2.2.1, different genes may have different evolutionary rates. Another problem is that sometime, the topology of a gene tree may conflict with that of the species tree [61]. When data grows, it is necessary to consider the effect of different evolutionary rates and topologies of individual genes. The earliest method to handle this problem is to concatenate all aligned sequences into a long and big alignment. This concatenate approach extracts the average pattern of the evolution among genes and eliminating the residuals from the average pattern, possibly resulting in robust estimation of species tree and divergence times [62, 63]. To take account of the different evolutionary rates and substitution matrix of genes, partition method splits the data into the categories of different pattern of molecular evolution [64-66]. Each partition can have independent parameter sets, while the final tree and times can be integrated by the results of each partition. The information criteria such as AIC or BIC are used for the selection of the partitions. This approach can improve the maximum likelihood of tree drastically [36], and result in more reliable estimation of time trees [60, 67], especially when strict molecular clock is seriously violated [68].

2.3.2 Coalescent approaches

Coalescent approaches interpret the inconsistency between gene trees and the species tree as the result of ancestral polymorphism. A pair of genes is traced back to the common ancestor. For a population in equilibrium, the mean coalescent time in generations is equal to the effective population size. Because of ancestral polymorphism, the divergence times of a gene tree are older than the speciation times. If the species tree includes nodes of successive speciation in short time, the topology of the gene tree can differ from the topology of species tree due to incomplete lineage sorting. Coalescent approaches introduce a population genetic model to describe the probability distribution of a gene tree given the phylogeny of species tree [69]. It

outperforms the concatenate approach when many of gene trees conflict with the species tree, as far as reliable divergence times of gene trees are available [70].

2.4 Reconstructing ancestral states of traits

Traits are the features of species in biology. Character traits of animals include morphological traits, life histories, ecotypes, behaviors, etc. Some of the traits are continuous traits, such as lifespan and bodyweight, while others are discrete traits, such as sociality/ solitary behavior, exists/ absence of some anatomic structures. The methods for ancestral state reconstruction for continuous and discrete traits are conceptually the same, but technically differ each other.

2.4.1 Ancestral state reconstruction of traits as a mimic of sequence evolution

Reconstructing ancestral states of traits, including morphological and life history traits, is difficult in phylogenetic analysis. One reason is due to the incomplete fossil records, and also the difficulties to connect fossil information with the extent taxa, while the other reason is the lack of clear idea and modeling of the evolution of those traits.

Current mainstream methods for ancestral state reconstruction mimic the methods of trait-based phylogenetic tree inference, so there are also three major methods for ancestral state reconstruction: maximum parsimony, maximum likelihood and Bayes methods [71-74]. These approaches rely solely on the phylogenetic information for ancestral state reconstruction, so I call it “phylogenetic inertia-based approach” in the following chapters. Same with tree inferences, maximum parsimony methods for ancestral state reconstruction also assumes the minimum numbers of total character-state changes is the best solution for ancestral states, and can be applied for both continuous and discrete traits [74]. Maximum likelihood method for ancestral state reconstruction also has a similar likelihood function to the likelihood of tree inference. It uses the substitution matrix in the form of JC69 for discrete traits, assuming the same rate for each character state change into other states [73]. Bayesian method for ancestral state reconstruction is also similar to the Bayesian tree inference method, using MCMC process to estimate the distributions of the probabilities of ancestral states [75, 76]. All of those methods assume the evolution of character traits

follows the phylogeny of species, and reconstructs the state of ancestral nodes broadly as the mean of the states of its two offspring nodes.

It should be noted, however, that reconstructing ancestral states analyzes the data of a single or a few "homologous" trait(s), whereas the inference of molecular phylogeny utilizes the information of a large number of homologous sites. Therefore, it is possible to estimate the parameters of the substitution rate matrix by counting the number of changes of each type. On the other hand, in the case of ancestral states reconstruction, the information on the substitution rate matrix is scarce.

2.4.2 Neutral theory validates molecular phylogenetics but caution is needed for traits evolution

Neutral theory of molecular evolution validates inference of phylogenetic relation based on DNA sequences. This hypothesis asserts that majority of the mutations that were fixed in the populations and leads to variation among sequences at present are selectively neutral. In other words, among frequent mutations on the genomes, adaptive mutations that enhance the chance of succession to the next generation are rare. As a result, most of the mutations are either neutral or deleterious compared with the existing genes that comprise the population at the time. Since deleterious mutations are eliminated from the population in the long term, evolution in molecular level is driven by the random fixation of neutral mutations during genetic drift process [18]. Therefore, we can expect that convergent evolution is rare in molecular level.

The rate of molecular evolution depends on the mutation rate and the proportion of neutral mutations. If a gene is under strong functional constraints, mutations are mostly deleterious and the proportion of neutral mutations is small. For example, human and chimpanzee diverged 8.2 Mya, and the pairwise difference between the two genomes is ~1% in amino acid level and ~4% for DNA level [77]. Homologous genes can be found even between human and yeast [78]. Highly conserved genes, such as housekeeping genes, are under strong purifying selection and less sensitive to the change of surrounding environment. Therefore, they were regarded as ideal data to reconstruct the phylogeny of distant species by the criterion of minimum evolution [79, 80]. In the contrary, genes under strong adaptive selection may mislead the inferred tree [81].

In the case of trait evolution, the situation is different. Charles Darwin wrote "The ears through their movements are highly expressive in many animals; but in some, such as man, the higher apes, and many ruminants, they fail in this respect" in ***The Expression of the Emotions in Man and Animals*** [82]. So "the ears through their movements" can be considered as a neutral trait of higher apes, but it is a selective trait in other species such as ruminants. He wrote many other examples in his books as well. Based on the observations of the trait evolutions, Darwin emphasized that natural selection is the main power for species evolution. A recent study tested the neutral hypothesis of phenotypic evolution using 210 character traits of yeast, and found that morphological variations are largely adaptive [26].

As for the adaptive traits, the variable environments promote the change in traits. A recent study on Aves showed the flight ability had evolved multiple times independently in Aves [83]. Furthermore, they have more chance of convergent evolution as a result of adaptation to the shared environment. For example, hedgehog, ternec and echidna are similar in shape, even though they are phylogenetically distant. Hyrax and rabbit look similar, even though they are not phylogenetically close each other. Microbats and whales have the common function of echolocation. Without sufficient caution, the ancestral state reconstruction methods that rely solely on the criterion of minimum evolution may provide misleading estimates for adaptive traits.

2.4.3 Methods based on the evolution of the genes controlling the traits

If the mechanism of a certain trait is known and the information on the key genes controlling the traits is available, it is possible to reconstruct the ancestral states of this trait based on the evolutionary history of these genes. An example is the reconstruction of colour visions of ancestral mammals based on the preservation or loss of opsin genes [84, 85].

2.4.4 Methods that utilize the rate/trait correlations

Another direction is to reconstruct the ancestral states of some traits by utilizing the information on the correlation between the trait value and the rate of molecular evolution. It is known that ω value (dN/dS, which dN refers to the non-synonymous substitution per site, and dS refers to synonymous substitution per site) is related to the body size of animals, possibly due to the negative correlation of fixation rate with

the effective population size [86, 87]. Lartillot *et al*., used this correlation and reconstructed the ancestral states of bodyweight of mammals [88, 89].

2.5 An extension of neutral theory to multiple-gene molecular evolution

Kimura calculated the rate of genome evolution and found that the selectively advantageous mutations are negligible in proportion and the substitutions that were fixed to the population and lead to molecular evolutions were nearly neutral [18]. In this framework, the hypothesis of molecular clock could be easily explained as a consequence of constant mutation rate. Both molecular clock hypothesis and the neutrality hypothesis have since been evolving. Responding to the observations of variable molecular evolutionary rates, Kimura described the rate of molecular evolution, r, as the product of the total mutation rate v and the fraction p of the molecular mutants that are selectively neutral [90]:

$$r = vp. \tag{3}$$

He noticed that p varies among genes and differs between the types of mutations. Since he concluded that deleterious mutations are eliminated from the population and do not contribute to molecular evolution, he regarded the variation of p as the variable constraints on the genes. The weaker the functional constraint is, the larger the probability that the mutations on the genes are selectively neutral, which results in a larger substitution rate (r) [90]. Some factors such as generation time, metabolic rates, exposure to UV radiation, and the efficiency of DNA repair mechanisms, etc., profoundly impact on the molecular evolutionary rates [91-93], thus mainly conduct though v. For example, shorter generation time generally results in higher substitution rates in some case studies [94, 95].

Equation (3) is naturally extended to multiple gene molecular evolution:

$$r_{ij} = c v_j \bar{p}_i \tilde{p}_{ij}, \tag{4}$$

where c is a proportionality constant, $\bar{p}_i$ corresponds to an effect from gene *i* on the proportion of neutral mutations, and $\tilde{p}_{ij}$ is included to reflect the among-branch variation of functional constraint on gene *i* of branch *j*. Here, I assume that the

mutation rate can vary among branches but is constant over the genome. By multiplying both sides of equation (4) by the time duration t_j of branch j, I express the branch length, b_{ij}, as a product:

$$b_{ij} = c \times \overline{p}_i \times (t_j \times v_j) \times \tilde{p}_{ij} . \qquad (5)$$

Here, I consider the case where the tree topology is already well established. By applying a multiplicative two-way ANOVA-type model to the estimated branch lengths, I obtain the two main effects: the gene effect and the branch effect. As is seen from equation (5), the gene effect is proportional to $\overline{p}_i$, the mean proportion of neutral mutations on the gene. Genes with low values exhibit evolutionary conservation. The branch effect is proportional to $t_j \times v_j$, and represents the expected amount of genomic evolution along the branch. The gene-branch interactions, which I obtained as residuals, are proportional to $\tilde{p}_{ij}$, and describe the pattern of variation among branches and among genes of purifying selection due to functional constraints.

In the next two chapters, I show the potential to apply this model into multiple biological topics, including divergence time estimation, ancestral state reconstruction and gene mappings, etc. Branch effect contain the information of genomic rates and genomic times, the new branch effect based approach for divergence time estimation offers an unique approach to obtain the speciation times, and firstly reported the absolute genome rates of 89 mammals. Gene-branch interaction contain the information of historical changes of functional constraint of each gene, thus can be regressed to the observed trait values of terminal taxa. Applying the obtained relationship between rates and traits to the gene-branch interactions of the internal branches, I obtained the ancestral states of 10 discrete life history and behavior traits. With the power and accuracy of the estimates carefully reported in Chapter 4, this is a new powerful approach for ancestral state reconstruction. The estimates do not rely on the functional annotation of gene, thus this is also an un-biased approach to identify genes contributed to the trait evolution.

Chapter 3. The branch effect of the extended neutral theory and a new approach of divergence time estimation

3.1 Summary

I extended neutral theory to multiple-gene molecular evolution and developed a new evolutionary framework for divergence time estimation, ancestral state reconstruction of life history traits, and gene mapping using phylogeny. I will introduce the new approach to estimate the genomic mutation rates together with the divergence times in this chapter. The branch lengths of individual genes are described as the products of branch effect, gene effect and gene-branch interactions. Branch effect is the product of genomic mutation rate and evolutionary time, thus can be used to calculate the speciation times in the Bayesian framework. The gene-branch interactions contain the information on the temporal variation of functional constraints on each gene, thus may be related to the changing states of the relevant traits in the evolutionary history. The new approach for divergence time estimation is a two-stage procedure: First, estimate the branch effect and its variances from individual gene trees, and second, estimate the divergence times by the Bayesian framework using fossil calibrations [54, 56, 58]. The estimated divergence times by using 1185 gene trees of 89 mammals was consistent with the well-accepted previous study [96]. Importantly, DNA, codon and protein-level divergence time estimation gave almost exactly the same time tree. As large as 82 percent of the variation of the branch lengths at all individual branches were explained by the projected variation among genes and the projected variation among branches. This implies that the variations of molecular evolutionary rates were largely synchronized among genes. It is also noted that the new approach reported the estimated variation of the genomic mutation rate of Mammalia for the first time.

3.2 Introduction

The branch effect of the extended theory of molecular evolution (Chapter 2.5 and /equation 4), which is estimated from the branch lengths of gene trees, is proportional to the product of genomic mutation rate and evolutionary time along a branch. By excluding the gene-specific effects of rate variation due to the change in functional constraints, it may provide the robust estimation of divergence times that are insensitive to the complex pattern of adaptive evolution to the natural environments. In this chapter, I show its potential for divergence time estimation by analyzing the sequences of 1185 genes of 89 mammalian species.

As one of most well studied Class, Class Mammalia have around 5,550 described species[97]. For last several decades, a large amount of the efforts have been made to understand mammalian phylogeny and speciation times. Class Mammalia has two subclasses: Yinotheria and Theriiformes. One subclass Yinotheria contains only one Order (Order Monotreme) with two suborders: platypoda and echidna. The other subclass of Class Mammalia is Theriiformes, with its only surviving branch Theria, contains two infraclasses: Eutheria and Metatheria. Both Eutheria and Metatheria have extinct lineages. Metatheria's only surviving lineage is Marsupials, while Eutheria's only surviving lineages is Placentalia. These terms such as Eutheria and Placentalia are exchangeable to each other, but not equal. Placentalia contains 93% of all extant mammalian species, and is with most attention in the study of mammals.

The earliest fossil of ancestral mammals, the earliest known synapsid (a clade include mammal, and mammal-like ancient animals), *Tikitherium*, dated 225 Mya [98]. So generally it is considered that the appearance of mammal should be Late Triassic [99]. The earliest fossil of crown Eutheria is *Juramaia*, which dated 160 Mya, thus the appearance of Eutheria should be earlier [100]. The fossils of crown placental orders appear in the 16 Mya intervals after Cretaceous–Paleogene boundary (K-Pg boundary, 65 Mya) indicates the repaid diversifications of crown placental mammals soon after the K-Pg extinction event [101, 102]. Concerning the diversifications of crown Placentia, there are three models regarding to the time of appearance of the Placentia, as well as the time of appearance of crown orders [101]: The explosive

model, long-fuse model, short-fuse model. The explosive model is based on the fossil records, while other two models are based on the analysis of molecular data. O'leary *et al*., 2012 [103] reconstructed tree based on the morphological characters of fossil and extant species, which supports the explosive model. Because the divergence of lineages had occurred earlier than the oldest fossil records, O'leary *et al*., 2012's time tree can be considered as the minimum bound of the times of divergence. The confliction between the long-fuse model and the short-fuse model is mostly caused by the difference of the data analyzed or the difference in the treatments of molecular clock. The short-fuse model declares that ancestral Placental appeared in the Cretaceous, and the appearance of new families and genera of each order happened before K-Pg event. The long-fuse model declares that ancestral Placentalia appeared in the Cretaceous, and that the appearance of new families and genera of each order occurred after K-Pg event. The long-fuse model is supported by fossil records and the integrated analysis of nuclear and mitochondrial genomes [96].

3.3 Materials and Methods

3.3.1 Identification of shared single-copy genes

In this study, I focused on genes without paralogues. I downloaded the homologue-type information of 43 mammalian species from the Ensembl Genome Browser [104], with the human genome used as a reference. Three homologue-type labels were available for each species: ortholog_one2one, ortholog_one2many and ortholog_many2many. I used genes labelled "ortholog_one2one" in the list of single-copy genes found between each species and humans and then constructed a presence–absence matrix for the 43 genomes. Taking into account the possibility of misannotation or DNA sequencing failure, I retained single-copy genes shared by more than 40 species. My final list contained 6,366 single-copy genes in Class Mammalia.

3.3.2 Data preparation

I downloaded 89 mammalian complete genomes from NCBI and extracted protein-coding sequences of each species using a custom Perl script. Alignments of all 6,366 single-copy genes were generated. Out of 6,366 single-copy genes, 1,202 genes were shared by all 89 species. Alignments at the amino acid level were performed in MAFFT v7.294 [105], with codons rearranged according to the amino acid alignment. All alignments were carefully checked by eye. Ambiguous regions, gaps and sites with less than 70% coverage among all species were removed. The total aligned sequence length of the 1,202 genes was 2,260,665 bp corresponding to 753,555 amino acid residues.

3.3.3 The species tree and branch lengths of protein trees

Song *et al.*, 2012 [70] had estimated a species tree of 37 species by applying a multispecies coalescent model to 447 nuclear genes. In my paper, I increased the size of the data set (89 species with 1202 single-copy nuclear genes) and attempted to obtain the species tree with improved resolution. I applied MP-EST and STAR/Njst, the maximum pseudo-likelihood procedure for estimating relationships under the coalescent model [106-108]. I estimated the nucleotide tree for each gene by the maximum likelihood method using RaxML v8.0.0 [109] under the

GTR+Γ+I model with 100 bootstrap replicates [31, 32, 109, 110]. Prior to the analysis, sequences were partitioned into three codon positions. Because the program assumes that input gene trees are precise, I excluded all maximum likelihood gene trees in which at least 30% of branches were characterized by bootstrap support values less than 30%. The uncertainty in those trees was mostly due to the short sequence lengths. The remaining 829 gene trees were used as the input for MP-EST and STAR/Njst. Because Scandentia+Glires+Primatomorpha as well as Carnivora+Perissodactyla+ Cetartiodactyla formed trifurcated clades in the estimated species tree, the positions of Scandentia and Perissodactyla could not be resolved in my analysis. Mason *et al*., 2016 [111] published a genome-wide indels-based phylogenetic tree that supports the ((Scandentia, Primatomorpha), Glire) and ((Carnivora, Perissodactyla), Cetartiodactyla) relationships. Indels may provide more accurate phylogenetic signals due to their robustness to the homoplasy compared with the information of the nucleotide substitutions, thus I followed [111] for the phylogenetic positions of Scandentia and Perissodactyla (which is also consistent with [70]). For testing the effect of the phylogenetic uncertainties on the reconstructed ancestral states, I further analyzed an alternative species tree following [112]. The estimated divergence times, together with ancestral states assuming the alternative phylogenetic positions of Scandentia and Perissodactyla [112] were very similar.

After constraining gene tree topologies to the species tree [70, 111], I estimated branch lengths of amino acid, codon and nucleotide trees of each of the 1,202 genes using PAML v4.8 [113]. I used two different models of amino-acid substitution: the LG matrix [114] and Mam matrix, the maximum likelihood estimate of the amino acid substitution matrix based on the randomly sampled 19,766 amino acid from 1202 genes of the 89 mammals ([115] Supplemental source data, PAML format), with gamma-distributed rate variation among sites [110]. Because the Mam matrix was a better fit to the data according to maximum log-likelihood values ($lnL_{mam} = -11843775$; $lnL_{LG} = -12325103$), I used the Mam matrix for protein sequence analysis (Mam matrix, see Table 3.1). For the codon analysis, I adopted the branch model [116]. For the nucleotide analyses, I applied JC69 [30] and JC69+Γ models [30, 117]. To minimize the potential bias due to the saturated information of long branches, I excluded any protein trees that had a maximum branch length larger

than 1.7 and/or a sum over all branch lengths that was larger than 17. The remaining 1,185 protein trees generated under the Mam model were used in further analyses.

3.3.4 Common and specific rates of protein evolution

In section 2.5, the branch length, b_{ij}, was expressed as a product:

$$b_{ij} = c \times \bar{p}_i \times (t_j \times v_j) \times \tilde{p}_{ij} . \quad (5)$$

The multiplicative model corresponding to equation (5) is:

$$b_{ij} = c \times a_i \times \beta_j \times \gamma_{ij} . \quad (6)$$

where α_i is proportional to $\bar{p}_i$, β_j is proportional to $t_j \times v_j$, and γ_{ij} is proportional to $\tilde{p}_{ij}$. By taking the log transformation, I obtain:

$$\log b_{ij} = C + A_i + B_j + \Gamma_{ij} . \quad (7)$$

I estimated C, A_i and B_j based on the above maximum likelihood estimates of b_{ij}, namely, $\hat{b}_{ij}$. Specifically, I treated the predicted number of substitutions, $\tilde{N}_j^{(i)} = \hat{b}_j^{(i)} \times L_i$ (where L_i is the sequence length of gene *i*), as Poisson random variables and applied Poisson regression with log link:

$$\log E\left[\tilde{N}_j^{(i)}\right] = \log L_i + C + A_i + B_j . \quad (8)$$

Even though the $\tilde{N}_j^{(i)}$ values are non-negative real numbers and are not necessarily integers, the log likelihood of the Poisson regression is a smooth function of the response variables and its domain of definition can be naturally extended to real positive values. The gene-branch specific effect Γ_{ij} was estimated as the ratio of $\tilde{N}_j^{(i)}$ to its predicted value. To account for over-dispersion, I also conducted negative binomial regression.

The predictive values of branch lengths, $\exp\left(\hat{B}_j\right)$, represent the expected amount of genomic evolution along the branch, $t_j \times v_j$. On the basis of the estimated

values of the commonality of branch lengths (i.e. the branch effect of the multiplicative two-way ANOVA model) and their variances, the time tree and the variable rate of evolution common to all genes were estimated by the Bayesian framework [54, 56, 58]. In other words, I estimated the divergence times based on the branch effect that integrates the branch lengths of gene trees rather than the branch lengths themselves. Fossil calibrations used in this study are summarized in Table 3.2.

3.4 Results

I estimated the branch effect by applying multiplicative ANOVA to the inferred branch lengths of the 1185 gene trees. Notably, the branch lengths implied by the main effects of this ANOVA-type model, except for constants of proportionality, were relatively insensitive to the choice of the model of sequence evolution and to whether sequence analyses were performed at the level of codons, nucleotides or amino acids (Figure 3.1b, Figure 3.2). In contrast, branch lengths inferred by protein, codon or nucleotide sequences for individual genes showed no strong correlation among each other (Figure 3.3). Since the branch effect represents the products of the genomic mutation rates and the time durations along branches of the species tree, this may imply that the chance of mutation is fairly homogeneous over a genome.

I constructed a species tree, which was mostly consistent with previous studies. As for the phylogenetic positions of Scandentia and Perissodactyla, I followed previous studies [70, 111]. Assuming this species tree topology, I applied a Bayesian relaxed clock method [54, 58] to infer divergence times (Figure 3.1a). The branch effects from protein-, codon- and DNA-based models yielded almost identical time trees (Figure 3.1b). The proposed inference of the time tree is robust, evidently because my approach removes the effect of variable functional constraints.

My analysis indicates that placental mammals originated 82.7–98.5 Mya. Although Afrotheria, Xenarthra and Boreoeutheria are inferred to have diverged before the K–Pg boundary, most extant orders diversified within a 20-Ma window of K-Pg extinction event, consistent with a previous well-accepted study [96] (Figure 3.1, Table 3.3). I also found that the rate of genomic evolution had accelerated in Rodentia and Eulipotyphla. My credibility intervals were generally wider than those of the genome-based divergence time estimation study cited above. This discrepancy is possibly due to the fact that my time tree was based on a single set of branch effects, whereas divergence times in the preceding studies were estimated using multiple sets of branch lengths of partitioned data. The authors in the earlier study assumed independent variation of evolutionary rates among partitions, which may have reduced the range of their credibility intervals [22]. In my Poisson regression analysis, the proportional reduction in deviance due to the model—an extended

measure of R^2—was as high as 0.82, indicating that the variation of molecular evolutionary rates was mostly globally synchronized. A negative binomial (NB) regression analysis gave almost the same estimates, with a proportional reduction in deviance of 0.74. The estimated divergence times assuming the alternative phylogenetic positions of Scandentia and Perissodactyla [112] were very similar (Table 3.3).

Notably, this new approach uses only one set of branch effects extracted from 1185 gene trees for divergence time estimation. Its computer time is same as the time required for a single gene analysis. After I obtain the branch effects and their variances, the calculation of time tree of 89 mammals can finish within a single day. dos Reis *et al*., 2012 [96] used concatenate approach to estimate the genomic-level time trees and they divided their nucleotide sequences into 20 partitions. If I use the same approach and number of partitions with dos Reis *et al*., 2012 (20 partitions) for the divergence time estimation, it will be 20 times slower. My branch effect based divergence time estimation approach is also currently fastest multiple-locus approach for divergence time estimation.

3.5 Discussion

Traditional concatenate-partition based approach and the new branch effect-based approach for divergence time estimation were apparently similar. However, their underlying hypothesis and principle are different. The concatenate-partition based approach clusters genes by their evolutionary rates and substitution matrix. If the genes show similar evolutionary rates and have similar substitution matrix, they are clustered into a single partition. By concatenating the gene sequences in the same partition, the tree parameters including branch lengths and their variation are estimated for each partition. As mentioned in Chapter 2.3.1, the Bayesian relaxed-clock approach can obtain reliable estimation of time trees [60, 67], even when strict molecular clock is violated [68]. However, the temporal variations of evolutionary rates are complex. For example, changes in the environment along some lineages may have decelerated the molecular evolution of a gene contributing a trait affected by the environment. The rate of molecular evolution of the same gene may have been elevated along some other lineages because of shortened generation length. The correspondence of variation of environments and variation of the rates of gene molecular evolutions is not a simple one-to-one but a complex many-to-many correspondence. As a result, the rates of a pair of partitions may have positive correlation along some lineages, while they have negative correlation along some other lineages. It is practically impossible to incorporate this complex structure as a prior for the distribution of variable rates among partitions.

Branch effect is obtained by applying multiplicative ANOVA to the inferred branch lengths of the individual gene trees, and is the product of evolutionary times and genomic mutation rates. The variations of gene trees caused by different models or data were excluded as residuals. Branch effects based on the analyses of DNA, codon and protein, were almost exactly proportional among others. By adding the fossil calibrations, I can distinguish the evolutionary times and genomic mutation rates. The change of the genomic mutation rates may be caused by the changes of generation time, metabolic rates, exposure to UV radiation, and the efficiency of DNA repair

mechanisms [91, 92]. With clear theory and biological explanations, this approach is also important to understand the evolutionary history of mammals.

3.6 Conclusion

Branch length of each individual genes is the product of branch effect, gene effect and gene-branch interactions. Branch effects contain information of genomic mutation rates and the time durations along branches of the species tree. By adding the fossil calibrations, I could estimate the divergence times of species, and it gives consistent estimations of speciation times no matter which data and model I use. Also this work measures firstly the variation of genomic rate among species. The estimation of variance of branch effect is precise in this approach, thus I could obtain reliable 95% confidence interval of estimated times.

Branch effect-based approach uses only one branch effect extracted from 1185 gene trees for divergence time estimation, thus its computer time is the same with a single gene analysis. Branch effect-based approach is also currently the fastest method for genome-level divergence time estimation.

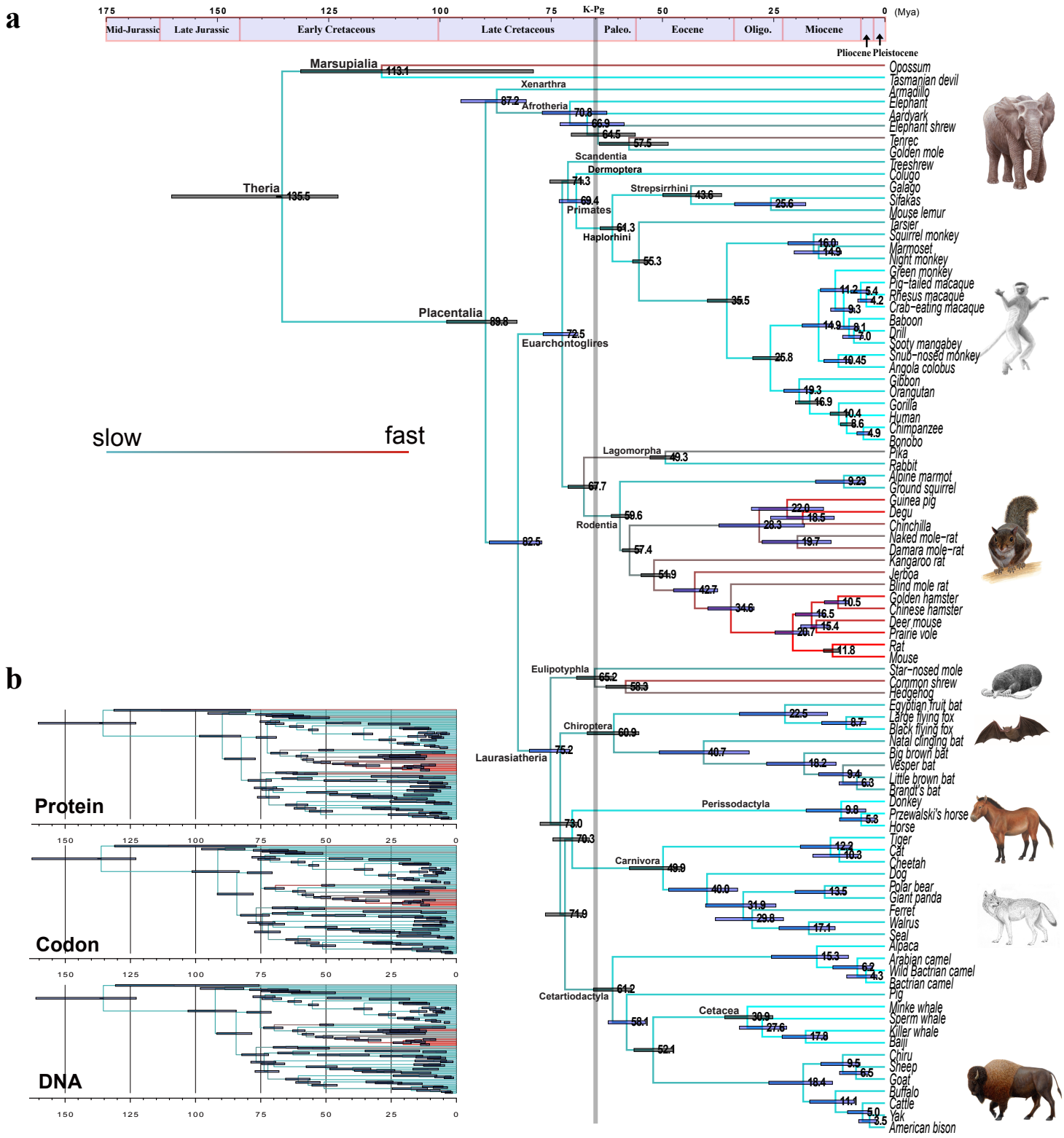

a
175
150
125
100
75
K-Pg
50
25
0
(Mya)
Mid-Jurassic
Late Jurassic
Early Cretaceous
Late Cretaceous
Paleo.
Eocene
Oligo.
Miocene
Pliocene
Pleistocene
Marsupialia
113.1
Xenarthra
87.2
Afrotheria
70.8
66.9
64.5
57.5
Scandentia
Dermoptera
74.3
Strepsirrhini
43.6
25.6
69.4
Primates
64.3
Haplorhini
55.3
35.5
26.8
Theria
135.5
Placentalia
89.8
72.5
Euarchontoglires
82.5
Lagomorpha
49.3
67.7
9.23
59.6
Rodentia
57.4
54.9
42.7
34.6
Eulipotyphla
65.2
58.3
75.2
Laurasiatheria
Chiroptera
60.9
40.7
73.0
70.3
Perissodactyla
Carnivora
49.9
40.0
31.9
29.8
74.0
Cetartiodactyla
61.2
58.1
52.1
Cetacea
30.9
27.6
17.8
18.4
Opossum
Tasmanian devil
Armadillo
Elephant
Aardvark
Elephant shrew
Tenrec
Golden mole
Treeshrew
Colugo
Galago
Sifakas
Mouse lemur
Tarsier
Squirrel monkey
Marmoset
Night monkey
Green monkey
Pig-tailed macaque
Rhesus macaque
Crab-eating macaque
Baboon
Drill
Sooty mangabey
Snub-nosed monkey
Angola colobus
Gibbon
Orangutan
Gorilla
Human
Chimpanzee
Bonobo
Pika
Rabbit
Alpine marmot
Ground squirrel
Guinea pig
Degu
Chinchilla
Naked mole-rat
Damara mole-rat
Kangaroo rat
Jerboa
Blind mole rat
Golden hamster
Chinese hamster
Deer mouse
Prairie vole
Rat
Mouse
Star-nosed mole
Common shrew
Hedgehog
Egyptian fruit bat
Large flying fox
Black flying fox
Natal clinging bat
Big brown bat
Vesper bat
Little brown bat
Brandt's bat
Donkey
Przewalski's horse
Horse
Tiger
Cat
Cheetah
Dog
Polar bear
Giant panda
Ferret
Walrus
Seal
Alpaca
Arabian camel
Wild Bactrian camel
Bactrian camel
Pig
Minke whale
Sperm whale
Killer whale
Baiji
Chiru
Sheep
Goat
Buffalo
Cattle
Yak
American bison
slow
fast
b
Protein
Codon
DNA
150
125
100
75
50
25
0

Figure 3.1 The Bayesian time tree of 89 mammals

I estimated the branch effect by applying the multiplicative ANOVA to the branch lengths of the 1185 gene trees. A Bayesian relaxed clock method was applied to the branch effect (see Supplementary Materials and Methods). The species tree topology matches that reported previously by others [70, 111]. (**a)** The time tree based on the protein sequences. Numbers at internal nodes are estimated divergence times (Ma), with 95% credibility intervals indicated by horizontal bars spanning nodes. Nodes with fossil calibrations are indicated by grey bars. Branches in red are associated with accelerated rates of genomic evolution. Scientific names of species are listed in Table 4.2. My credibility intervals were generally wider than those of the genome-based divergence time estimation study cited above. **(b)** Comparison of time trees estimated from protein sequences under the Mam+ Γ model, codon sequences under the branch model, and DNA sequences under the JC69 model. The tree topology is same as that of Figure 3.1a.

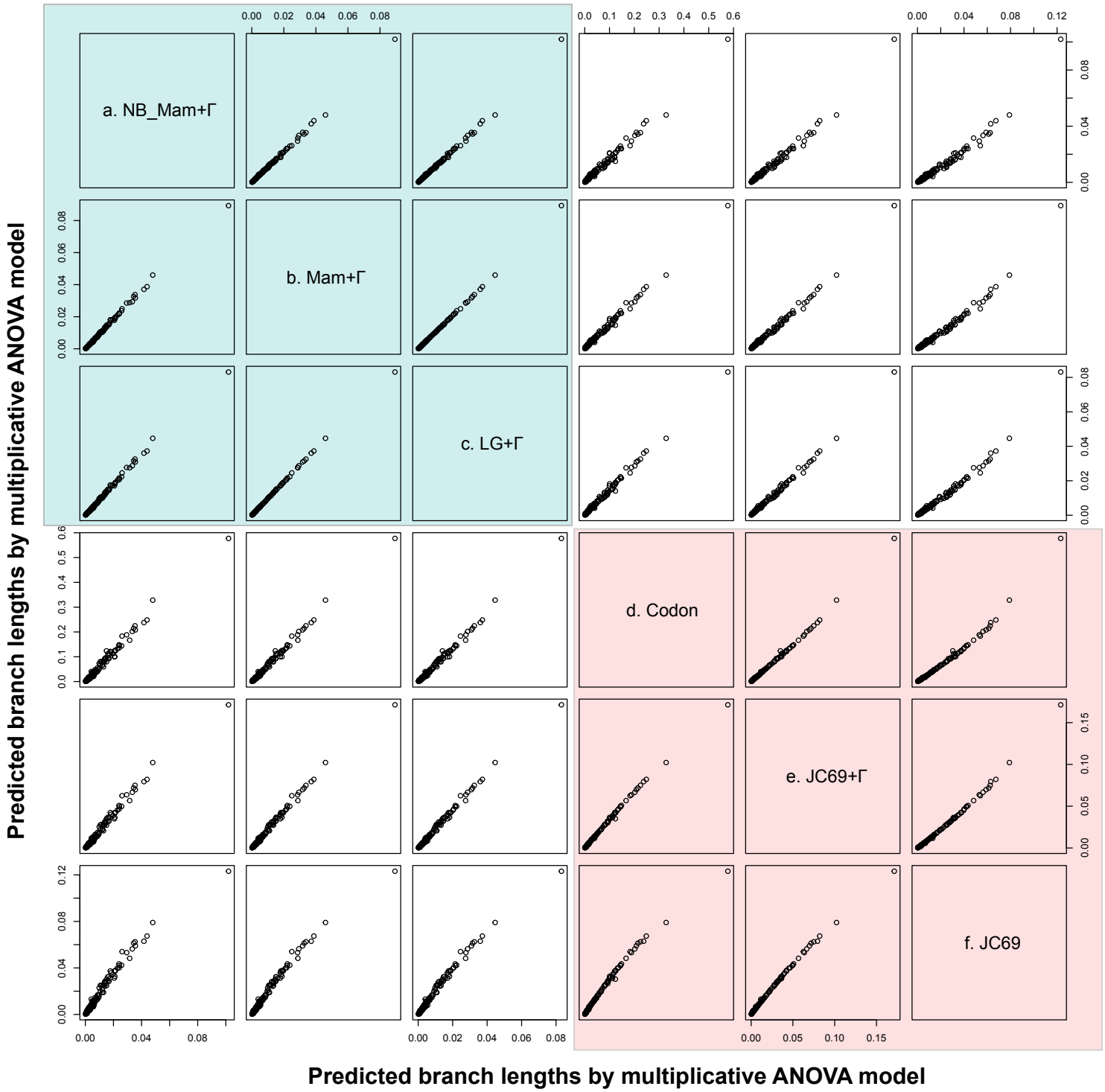

a. NB_Mam+Γ
b. Mam+Γ
c. LG+Γ
d. Codon
e. JC69+Γ
f. JC69
Predicted branch lengths by multiplicative ANOVA model
Predicted branch lengths by multiplicative ANOVA model

Figure 3.2 Comparison of branch lengths predicted by multiplicative ANOVA models

(**a)** Species-tree branch lengths predicted under negative binomial regression-based multiplicative ANOVA models (see Methods). (**b–f)** Species-tree branch lengths predicted under Poisson regression-based multiplicative ANOVA models (see Methods). For protein sequences, I applied Mam+Γ (**a–b**) and LG+Γ models (**c**) [110, 114, 115]. I used the branch model for codon sequences [116] (**d**) and JC69 (**e**) and JC69+Γ (**f**) models for DNA sequences [30, 110].

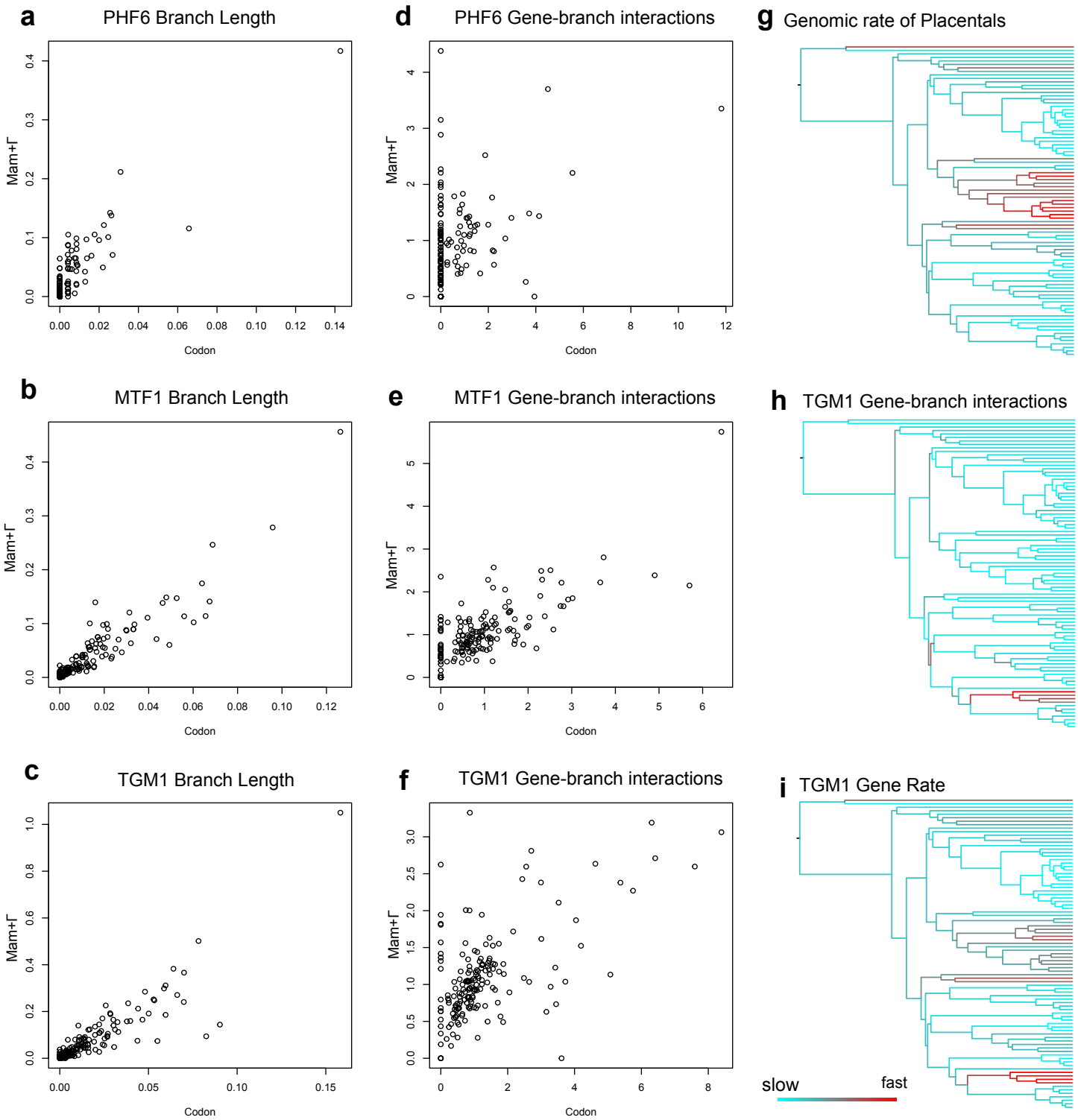

a
PHF6 Branch Length
Mam+Γ
Codon
b
MTF1 Branch Length
Mam+Γ
Codon
c
TGM1 Branch Length
Mam+Γ
Codon
d
PHF6 Gene-branch interactions
Mam+Γ
Codon
e
MTF1 Gene-branch interactions
Mam+Γ
Codon
f
TGM1 Gene-branch interactions
Mam+Γ
Codon
g
Genomic rate of Placentals
h
TGM1 Gene-branch interactions
i
TGM1 Gene Rate
slow
fast

Figure 3.3 Comparison of branch length, branch effect and gene-branch interactions

(**a–f)** Three genes are shown as examples: PHF6, MTF1 and TGM1. The x-axis shows branch lengths (BL) or gene-branch interactions (Spec) estimated under the codon model, while the y-axis refers to these two items estimated under the Mam + Γ protein model. (**g)** Genomic rate of each branch. (**h)** Gene-branch interactions of TGM1. (**i)** Molecular rate of TGM1. The rate of each branch is indicated based on the colour gradient scale, with red branches corresponding to regions of the tree in which a gene has evolved at an accelerated rate. The tree topology of **(g-i)** is same as that of Figure 3.1a

Table 3.1. Mam substitution matrix

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	0.101																			
Arg	0.009	0.184																		
Asn	0.597	0.057	2.787																	
Asp	0.133	3.880	0.146	0.114																
Cys	0.056	5.497	0.044	0.047	0.048															
Gln	0.756	0.213	0.058	4.446	0.025	1.350														
Glu	1.839	2.972	0.280	1.969	2.163	0.073	1.769													
Gly	0.074	8.177	3.601	1.093	2.737	6.614	0.077	0.175												
His	0.037	0.201	0.354	0.024	0.079	0.000	0.000	0.033	0.041											
Ile	0.151	0.722	0.014	0.032	0.255	0.634	0.047	0.063	0.893	2.275										
Leu	0.102	8.028	1.437	0.024	0.000	1.321	1.507	0.092	0.073	0.117	0.032									
Lys	0.359	0.712	0.082	0.008	0.135	0.036	0.057	0.018	0.043	5.340	3.946	0.515								
Met	0.111	0.025	0.012	0.023	3.709	0.036	0.041	0.066	0.278	1.494	6.662	0.027	0.033							
Phe	1.825	0.647	0.033	0.054	0.107	1.408	0.086	0.034	1.915	0.013	2.070	0.057	0.026	0.131						
Pro	1.825	1.159	3.953	0.087	2.844	0.080	0.025	2.489	0.179	0.258	0.712	0.064	0.030	1.193	2.438					
Ser	7.254	0.664	1.228	0.042	0.127	0.052	0.079	0.097	0.130	3.024	0.073	0.649	3.767	0.058	1.497	1.970				
Thr	0.000	9.548	0.047	0.000	10.392	1.784	0.241	1.243	0.012	0.000	2.692	0.000	0.373	0.588	0.112	0.561	0.046			
Trp	0.046	0.226	0.541	0.419	11.073	0.006	0.029	0.104	13.577	0.056	0.095	0.024	0.121	8.647	0.046	0.597	0.394	0.735		
Tyr	4.865	0.091	0.022	0.190	0.111	0.051	0.319	1.261	0.015	11.543	2.448	0.045	7.266	1.086	0.052	0.057	0.223	0.177	0.061	
Val	0.068	0.037	0.056	0.063	0.014	0.056	0.076	0.053	0.021	0.050	0.070	0.068	0.024	0.020	0.061	0.109	0.071	0.004	0.020	0.061

Table 3.2. Fossil calibrations used in this study

	node	maximum (Ma)		minimum (Ma)	
1	Theria	171.2	Absence of therians in the Middle Jurassic [118]	124.0	*Eomaia scansoria* (The oldest stem Eutheria) [118, 119]
2	Marspialia	130.0	*Sinodelphys szalayi* (Basal Metatheria) [103, 120]	64.5	*Pucadelphys andinus* (The oldest crown Marspial Sister to Didelphimorphia) [103, 121, 122]
3	Eutheria	112.0	*Sasayamamylos kawaii* (The oldest Eutheria with the common dental formula with Placental Mammal) [123]	65.0	*Protungulatum donnae*(The oldest Boreoeutheria) [103, 124]
4	Afroinsectivora (Afrosoricida+ Macroscelididae)			65.0	*Prodiacodon crustulum* (Sister to Macroscelididae) [103, 125]
5	Afrosoricida			33.9	*Eochrysochloris tribosfenus* (The oldest crown Afrosoricida. Chrysochloridea) [103, 126]
6	Euarchonta			65.0	*Purgatorius coracis* (The oldest crown Euarchonta) [103]
7	Primate	65.0	*Purgatorius coracis* (The oldest Euarchonta) [103]	55.8	*Teilhardina brandti* (The oldest crown Primate) [103, 127]
8	Strepsirrhini	55.8	*Cantius torresi* (The oldest stem Primate) [103, 128]	33.7	*Karanisia clarki* (The oldest Strepsirrhini. Lorisiformes) [96, 129]
9	Haplorhini	55.8	*Cantius torresi* (The oldest stem Primate) [103, 128]	45.0	*Tarsius eocaenus* (The oldest crown Haplorhini. Tarsiiformes) [103, 130]
10	Simiiformes			33.7	*Catopithecus browni* (The oldest crown Simiiformes. Catarrhini) [96, 129]
11	Catarrhini	33.7	Absence of hominoids in the Late Eocene [96, 129]	23.5	*Proconsul* (The oldest crown Catarrhini. Hominoidea) [96, 131]
12	Hominidae	33.7	Absence of pongines in the Late Eocene [96, 129]	11.2	*Sivapithecus* (The oldest crown Homidae. Ponginae) [96, 132]
13	Homininae			7.3	*Chororaphithecus* (The oldest crown Hominae. Stem gorilla) [96, 133]
14	Hominin	10.0	Absence of hominines in the Middle Miocene [96, 134]	5.7	*Orrorin tugenensis* (The oldest Hominini) [96, 135]
15	Glires			65.5	*Mimotona wana* (The oldest crown Glires. Lagomorpha) [103, 136]
16	Lagomorpha	65.8	Absence of crown lagomorphs in the Early Paleocene [96, 137]	48.6	Vastan calcanei (The oldest crown Lagomorpha. Laporidae) [96, 138]
17	Rodent	65.8	Absence of rodents in the Early Paleocene [96, 137]	56.8	*Sciuravus* sp. (The oldest crown Rodent fossil)
18	Myomorpha vs. Hystricomorpha	58.9	Absence of cavimorpha in the Late Paleocene [96, 139]	52.5	*Birbalomys* (The oldest Hystricomorpha) [96]
19	Myomorpha	56.8	*Sciuravus* sp. (The oldest Rodent. Sister to Myomorpha+Hystricomorpha) [103, 140]	46.2	*Simimys simplex* (The oldest crown Myomorpha. Muridae) [103, 141]
20	Muridae	14.0	Absence of crown murines in the Early Miocene [96, 142]	10.4	*Karnimata* (The lineage leading to *Rattus*) [96] [142]
21	Eulypotyphla	65.0	*Leptacodon proserpinae* (The oldest stem Eulypotyphla) [119]		
22	*Erinaceus vs. Sorex*			61.7	*Litolestes ignotus* (The oldest Erinaceidae) [103, 143]
23	Scrotifera (Laurasiatheria excluding Eulypotyphla)			65.0	*Protungulatum donnae*(The oldest Laurasiatheria) [103, 124]
24	Zooamata (Perissodactyla+Carnivora) vs. Cetartiodactyla	65.0	*Protungulatum donnae*(The oldest Boreoeutheria) [103, 124]		
25	Zooamata (Perissodactyla+Carnivora)			62.5	*Lambdotherium* (The oldest Perissodactyla) [96] [130]
26	Carnivora	63.8	*Protictis haydenianus* (The oldest stem Carnivomorpha) [103, 144]	46.2	*Hesperocyon gregarius* (The oldest crown Carnivora. Canidae) [103] [145]
27	Cetartiodactyla			55.8	*Cainotherium* sp. (The oldest crown Cetartiodactyla. Tylopoda) [103] [146]
28	Cetacea vs. Ruminant			52.4	*Himalayacetus subathuensis* (The oldest Cetacea) [96, 147]
29	Cetacea			32.0	*Llanocetus denticrenatus* (The oldest Mysticeti) [96] [148]
30	Chiroptera			55.5	*Archaeonycteris praecursor* (The oldest crown Chiroptera. Microchiropteramorpha) [103] [149]

Table 3.3 Comparison of estimated divergence times of this study and a well-accepted previous study [96] of selected nodes

Node	Time tree for dos Reis *et al.*, 2012 [96]		Species tree		Alternative Species tree [112]	
Theria	185.0	(174.5, 191.8)	135.5	(122.9,160.3)	135.0	(122.7, 159.6)
Marsupialia	66.7	(50.7, 83.7)	113.1	(79.0, 131.3)	110.3	(71.5, 131.2)
Placentalia	89.9	(88.3, 91.6)	89.8	(82.7, 89.5)	90.0	(83.0, 99.4)
Atlantogenata	87.5	(85.9, 89.1)	87.2	(80.6, 95.3)	87.5	(80.9, 96.3)
Afrotheria	70.4	(68.5,72.4)	70.8	(67.8, 75.3)	71.1	(62.8, 78.6)
Paenungulata	59.8	(57.7, 61.8)	-	-	-	-
Afroinsectiphilia	-	-	66.9	(62.5, 77.0)	67.2	(58.8, 74.5)
Afrosoricida	-	-	57.5	(48.7, 64.2)	57.7	(48.8, 65.6)
Boreotheria	82.4	(81.1, 83.8)	82.5	(77.1, 88.9)	82.9	(77.6, 89.5)
Laurasiatheria	76.0	(74.8, 77.1)	75.2	(70.9, 79.9)	75.6	(71.6, 80.8)
Eulipotyphla	61.3	(60.6, 61.8)	65.2	(61.0, 69.3)	65.5	(61.5, 70.0)
cow/horse	73.1	(72.0, 72.4)	71.9	(67.9, 76.3)	70.6	(67.1, 74.9)
Cetartiodactyla	61.4	(60.7, 62.3)	61.2	(57.0, 65.5)	60.7	(56.6, 65.0)
pig/cow	58.0	(57.4, 58.8)	58.1	(53.8, 62.2)	57.5	(53.2, 61.7)
dolphin/cow	52.7	(52.2, 53.7)	52.1	(47.5, 57.4)	51.5	(46.8, 55.8)
horse/cat/bat	72.2	(71.2, 73.3)	73.0	(68.9, 77.5)	73.5	(69.7, 77.9)
horse/cat	70.1	(69.1, 71.1)	70.3	(66.4, 74.6)	72.3	(68.7, 76.7)
Carnivora	54.1	(52.0, 55.9)	49.9	(45.6, 57.4)	50.9	(45.8, 58.8)
Chiroptera	59.3	(57.6, 60.8)	60.9	(55.3, 66.9)	61.4	(55.5, 67.4)
Euarchontoglires	75.8	(74.6, 77.0)	72.5	(69.0, 76.8)	72.8	(69.2, 77.1)
Glires	70.7	(69.6, 71.8)	67.7	(67.4, 71,2)	67.0	(64.2, 70.5)
Lagomorpha	47.8	(45.8, 49.3)	49.3	(47.1, 52.8)	49.3	(47.0, 52.8)
Rodentia	64.5	(63.4, 65.5)	59.6	(57.0, 61.5)	67.0	(64.2, 70.5)
guinea pig/rat	61.3	(60.3, 62.2)	57.4	(54.8, 59.0)	59.2	(54.6, 59.0)
kangaroo rat/rat	55.6	(54.4, 56.5)	51.9	(48.3, 54.8)	51.9	(48.3, 54.9)
mouse/rat	13.9	(13.2, 14.3)	11.8	(10.3, 13.8)	11.8	(10.3, 13.8)
human/ tree shrew (Euarchonta)	74.2	(73.0, 75.3)	71.3	(67.8, 75.3)	-	-
human/ colugo	-	-	69.4	(66.1, 73.2)	70.5	(67.1, 74.5)
(glires, tree shrew)	-	-	-	-	71.1	(67.8, 75.2)
Primates	69.0	(67.8, 70.1)	61.3	(58.6, 64.0)	61.8	(59.1, 64.5)
Strepsirrhini	54.3	(52.3, 55.8)	43.6	(36.7, 49.9)	43.7	(36.7, 50.2)
human/tarsier	65.0	(63.9, 66.0)	55.3	(53.0, 56.7)	55.4	(53.2, 56.8)
Anthropoidea	36.6	(34.9, 38.3)	35.5	(32.1, 39.9)	35.7	(32.1, 40.3)
Catarrhini	25.6	(24.4, 26.8)	25.8	(23.3, 29.7)	26.0	(23.4, 32.0)
human/orang	17.3	(16.2, 18.4)	16.9	(14.2, 20.1)	16.8	(13.9, 20.1)
human/gorilla	10.2	(9.6, 11.0)	10.4	(8.2, 12.3)	8.3	(8.2, 12.7)
human/chimp	8.7	(8.1, 9.4)	8.6	(6.7, 10.0)	8.3	(6.3, 10.0)

Chapter 4. Rates of molecular evolution suggest natural history of life history traits and a post-K-Pg nocturnal bottleneck of Placentals

4.1 Summary

Life history and behavioral traits are often difficult to discern from the fossil record, but evolutionary rates of genes and their changes over time can be inferred from extant genomic data. Under the neutral theory, molecular evolutionary rate is a product of mutation rate and the proportion of neutral mutations [18, 90]. Mutation rates may be shared across the genome, whereas proportions of neutral mutations vary among genes because functional constraints vary. By analysing evolutionary rates of 1,185 genes on a phylogeny of 89 mammals, I extracted historical profiles of functional constraints on these rates in the form of gene-branch interactions. By applying a novel statistical approach to these profiles, I reconstructed the history of 10 discrete traits related to activity, diet and social behaviors. My results indicate that the ancestor of placental mammals was solitary, seasonally breeding, insectivorous and likely nocturnal. The results suggest placental diversification began 10–20 million years before the K–Pg boundary (66Mya), with some ancestors of extant placental mammals becoming diurnal and adapted to different diets. However, from the Paleocene to the Eocene–Oligocene transition (EOT, 33.9Mya), I detect a post-K–Pg nocturnal bottleneck where all ancestral lineages of extant placentals were nocturnal. While diurnal placentals may have existed during the elevated global temperatures of the Paleocene–Eocene Thermal Maximum [150], I hypothesize that diurnal placentals were selectively extirpated during or after the global cooling of the EOT whereas some nocturnal lineages survived due to preadaptations to cold environments.

4.2 Introduction

Mammals, a diverse group occupying numerous ecological niches, comprise 5,550 described species, of which 93% are in Placentalia [97]. Placentals exhibit remarkable radiation including aquatic, fossorial and flight adaptations, and are distributed in all continents and their peripheral islands. My understanding of the evolution of life histories and ecological niches of early placentals is rather poor. Morphology of fossil species may provide some information on their life histories. For example, dental morphologies may indicate the diet [151], while relative orbit sizes may indicate diurnality of fossil species [152]. However, information about life histories of ancestral placentals is still limited due to incompleteness of the fossil record, especially for the Mesozoic placentals [103].

Traits shared by extant placentals give hints about their ancestral states. For example, most extant mammals have limited color vision and share traits such as acute auditory, tactile and olfactory senses as well as the presence of fur or brown adipose tissue to assist with thermo-regulation in cold environments [84, 153, 154]. These common characteristics suggest that mammals of the Mesozoic era were mostly nocturnal, the so-called nocturnal bottleneck hypothesis (NBH) [153, 155]. The NBH can explain many mammalian traits and is highly relevant to the behavioral evolution of mammals. On the other hand, no consensus exists for early mammals regarding other life-history traits (e.g., reproductive seasonality, mating system, and social behaviour).

The ancestral states of life-history traits on a phylogeny can be reconstructed via minimum evolution criteria. The generally governing principle of these criteria is that descendants tend to resemble their ancestors. This “phylogenetic inertia-based approach” has two problems. First, life histories, especially behaviors, can evolve rapidly, and sometimes do not show a strong phylogenetic correlation among species. Second, animals occupying similar ecological niches can show remarkable convergent evolution, even among phylogenetically distant lineages, such as seen between hedgehogs and tenrecs [36].

4.3 Materials and Methods

4.3.1 Sequence data and species tree

Sequence data and species tree is same with those used in Chapter 3.

4.3.2 Collection of life history traits

I collected information, mainly from the Animal Diversity Web (http://animaldiversity.org), about ecologically important discrete life history traits including diet, diurnal activity, mating system, social behaviour and sexual dimorphism. Details of each trait with references to original sources are summarized in Table 4.2 [156]). The polygenic nature of complex ecological traits, such as those examined in my study, hinder their quantification for comparative phylogenomic analysis. I therefore employed a conservative strategy, in which these discrete traits were treated as binary values (0 or 1; Table 4.1). Some species possessed both states of a trait. For example, tigers are active both day and night, while naked mole rats (*Heterocephalus glaber*) live underground; their activities thus do not follow a circadian rhythm. I treated these cases as missing data in the analysis. For diet, there were three states. Because my model can only handle two-state discrete traits, I used a one-to-others approach and transformed these states into three separate traits: carnivory (carnivorous or not), herbivory (herbivorous or not) and omnivory (omnivorous or not). In regards to insectivory, animals eating insects occasionally or as only a small part of their diet were not considered to be strictly insectivorous; these cases were treated as missing data in the analysis.

4.3.3 Rate-based life history trait prediction

To investigate the predictive power of evolutionary rates, I conducted logistic regression, with trait values at the terminal nodes serving as the response variables ($p(X_{term})$) and gene-branch interactions at the terminal branches as the explanatory variables (X_{term}).

$$p(X_{term}) = \frac{\exp^{\beta_0+\beta_1 X_{term}}}{1+\exp^{\beta_0+\beta_1 X_{term}}}. \quad (9)$$

The estimation of $\beta_0 + \beta_1 X$ can be found by minimum the negative log likelihood function as:

$$L = -\log(\prod_{i:Y_i=+} p(X_i) \prod_{j:Y_j=-} (1 - p(X_j))). \quad (10)$$

Because the number of explanatory variables far exceeded the sample size, I applied lasso penalized logistic regression by using glmnet [157, 158]. The penalty term to the log likelihood function by Lasso is defined by:

$$L + \lambda \sum |\beta_1|. \quad (11)$$

The lambda coefficient of the lasso penalty (λ) was selected by minimizing deviance via leave-one-out cross validation. To increase efficiency, the first stage of my analysis followed a previously described procedure of pre-screening genes to remove those least likely to usefully predict the trait values [159]. The number of genes retained for the second-stage analysis was decided by maximizing the value of the AUC using the ROCR package in R [160]. Ancestral states were obtained by the rate-trait regression model fitted to the terminal taxa:

$$p(X_{anc}) = \frac{\exp^{\beta_0+\beta_1 X_{anc}}}{1+\exp^{\beta_0+\beta_1 X_{anc}}}. \quad (12)$$

Given the result of lasso-logistic regression, I used the genes left as predictors for ancestral state reconstruction of life history traits.

The idea underlying my cross validation approach is to predict the character states at a tip of the tree without telling the inference procedure what are the actual values of the states. Because I know the true values for the tip states, I can compare the performance of alternative inference methods. In the future, modifications of my cross-validation approach could be applied to fossil information rather than extant character data.

4.3.4 Surplus confidence of rate-based prediction

For rate-based ancestral prediction, I measured the surplus confidence compared with the inference solely based on the frequency of the trait-value at the terminal nodes. When 70% of the terminal nodes have the values of 1, I may expect roughly that with 70% on average, the trait at an internal node will have the value of 1. If the predicted probability is 70%, there is no surplus. But, if the predicted probability is much larger than 70%, then relating trait values with rates can be interpreted as adding information regarding the ancestral state prediction. This surplus confidence can be formulated as the *p*-value under the null-hypothesis, which assumes independence between trait values and gene-branch interactions. To set the *p*-values on the reconstructed ancestral states, I trained my rate-based procedure using permuted trait values at the terminal nodes. For each replication, I estimated the regression coefficients and predicted ancestral states by using the gene-branch interactions of all 1185 genes. Based on 1000 analyses, I thereby obtained the null distribution of the predicted probabilities of the ancestral states. By contrasting the observed values with reference to the null distributions, I set the two-sided p-values.

4.2.5 Weighted nearest-neighbour method

I compared the performance of the above rate-based prediction method with a prediction method based on phylogenetic information. Phylogenetic inertia-based prediction assumes that closely correlated species have similar life-history trait values. For simplicity, I predicted life history traits as the weighted average of nearest neighbours. Based on the pairwise distance matrix of the commonality of branch lengths of the species tree, the trait value of each species j, $\hat{y}_j$, was predicted as:

$$\hat{y}_j = \frac{1}{\sum_{s \neq j} 1/d_{sj}^{\alpha}} \sum_{s \neq j} \frac{y_s}{d_{sj}^{\alpha}} = \frac{\sum_{s \neq j} y_s / d_{sj}^{\alpha}}{\sum_{s \neq j} 1/d_{sj}^{\alpha}}. \quad (13)$$

Here, y_s is the trait value of species s, and d_{sj} is the phylogenetic distance between species s and species j. For each trait, the value of α was chosen to maximize the AUC.

To compare the predictive powers of rate- and phylogenetic inertia-based procedures, I calculated AUC values and accuracy by leave-one-out cross validation. As cut-off points, I chose AUC values that maximized the accuracy.

4.3.6 Phylogenetic inertia-based ancestral prediction

For phylogenetic inertia-based procedures, I reconstructed ancestral states of each trait in a Bayesian framework using BayesTraits [75, 76]. I applied the multistate model to discrete traits. The 1202 gene trees obtained above were used as guide trees. I obtained predicted probabilities for discrete traits and used them to calculate correlations between traits and rates.

I used FigTree (http://tree.bio.ed.ac.uk/software/figtree/) and the ape package in R to draw trees [161, 162].

4.4 Result

4.4.1 Regressing the trait values on the gene-branch interactions

Gene–branch interactions express the relative among-branch variation of molecular evolutionary rate of a gene, and contain information on variation of functional constraints. By assuming the relationship at terminal nodes between the states of a trait and gene-branch interactions also applies to the rest of the phylogeny, it may be possible to reconstruct ancestral trait states based on the values of gene-branch interactions at the internal nodes. I applied lasso penalized logistic regression [157, 158] to relate trait values at terminal nodes to gene-branch interactions. Using the estimated relationships, I predicted unobserved values of ancestral node states of life history traits. As an alternative to this rate-based approach, I also designed a phylogenetic inertia-based weighted nearest-neighbour method (WNN) to reconstruct the terminal nodes. Prediction power of these two approaches was evaluated by leave-one-out cross validation (LOOCV).

I considered 10 discrete traits related with sociality, dirunality, seasonal breeding etc. The performances of rate-based and WNN prediction methods for each trait are summarized in Table 4.1 and Figure 4.1. In general, predictions relying on evolutionary rates were similar to or better than those obtained by WNN. In particular, rate-based prediction yielded the area under the receiver operating characteristic curve (AUC) values above 0.90 for diurnality, reproductive seasonality and insectivory. These AUC values were substantially higher than those obtained by WNN. Except for insectivory, the performance of the WNN was mostly consistent with Pagel's λ, a measure of phylogenetic correlation [72]. Although insectivory showed a strong phylogenetic signal ($\lambda = 1$), the WNN method failed to give a solid prediction (AUC = 0.648). Insectivorous clades are basal in several placental superorders, such as Afrotheria and Laurasiatheria, and their long branches reduced the performance of N-N. In contrast, rate-based prediction was robust against long branches. I tested the effects of branch length on the prediction performance of both N-N and rate-based methods. Both approaches performed well on short branches, but rate-based prediction handled long branches better than WNN (Figure 4.2).

4.4.2 Evolutionary history of insectivory

I also reconstructed the ancestral states of all 10 discrete traits using the aforementioned rate-trait regression models fitted to terminal taxa (Figures 4.3-4.7). Notably, rate-based prediction and phylogenetic inertia-based prediction (BayesTraits, multiple states) generated different inferred patterns of evolution of life-history traits. For example, Figure 4.3a-b shows the evolutionary history of insectivory as predicted by rate-based and phylogenetic inertia-based approaches. For rate-based prediction, I calculated the p-values under the null-hypothesis of independence between traits and evolutionary rates (Figure 4.3a). Rate-based prediction indicated that early placentals were mainly insectivores, especially before the K–Pg boundary. After the K–Pg boundary, non-insectivorous lineages evolved independently from their insectivorous ancestors. Accordingly, extant insectivorous clades such as Afroinsectphilia (tenrecs, golden moles, elephant shrews and aardvarks), Eulipotyphla (moles, shrews, and hedgehogs) and Chiroptera (bats) may have retained the ancestral-styled diet from Mesozoic placentals. Phylogenetic inertia-based prediction suggested the opposite scenario: the ancestral placentals were mainly non-insectivores, with insectivorous clades, including tenrecs, moles and bats, evolving from non-insectivorous ancestors before the K–Pg boundary.

My findings are harmonious with the morphometric analysis of the fossil therian mammals (eutherian-placentals and metatherian-marsupials) [151]. These analyses indicate the predominant therian diet was insectivory in the Late Cretaceous and it diversified in the Paleocene. A reconstruction of the hypothetical placental ancestor based on the morphology of multiple fossils further suggests that the common ancestor of Placentalia was insectivorous [103]. Therefore, I conclude that the rate-based prediction of insectivory is more reasonable than the phylogenetic inertia-based prediction and that ancestors of placental mammals were predominantly insectivorous.

4.4.3 Evolutionary history of diurnality, behaviour and diet

The reconstructed ancestral states of traits with high AUCs are likely to be reliable. Besides insectivory, rate-based prediction was also precise for diurnality (diurnal VS nocturnal, AUC = 0.919) and reproductive seasonality (year-round

breeding VS seasonal breeding, AUC = 0.965). The evolutionary histories of these two traits, together with sociality and diet, are summarized in Figure 4.4a–d. Predicted probabilities and the null-distributions of representative nodes are shown in Figure 4.4e-f.

Although the reconstructed states of the common ancestors (or stem taxa) of Placentalia is accompanied with uncertainty (predicted probability to be diurnal is 0.495), the rate-based inference procedure implied that the early crown Placentalia were primarily nocturnal During the Late Cretaceous, some lineages (e.g., Euarchonta and Ferungulata) became diurnal. After the K–Pg boundary (66 Mya), all ancestral lineages of existing placentals were nocturnal until the EOT (Eocene–Oligocene transition 33.9, Mya). Later, from the Oligocene to the Neogene, several lineages independently became diurnal Extant placentals share several common traits that indicate their ancestors experienced a nocturnal lifestyle for a long duration.

Reproductive seasonality is generally very difficult to infer from paleontological evidence. Furthermore, Table 4.1 shows that the phylogenetic inertia of this trait is not high ($\lambda = 0.764$). In contrast, the rate-based prediction method had high power to predict reproductive seasonality at terminal taxa (AUC = 0.965, Table 4.1), implying the strong potential of this approach to reconstruct the ancestral states of this trait. The common ancestors of Placentalia, Euarchontoglires and Laurasiatheria were predicted to be seasonal breeders, while those of Atlantogenata and Boreoeutheria were inferred to breed year-round (Figure 4.4b). Similar to the distribution of terminal states, reconstructed ancestral states for reproductive seasonality show little phylogenetic inertia, but varied rapidly throughout evolutionary history.

Concerning social versus solitary lifestyles, evolutionary histories reconstructed by rate- and phylogenetic inertia-based prediction methods were consistent with each other (Figure 4.5c–d), which implies that sociality is phylogenetically stable. Although the common ancestor of Placentalia was predicted to be solitary, most ancestral species within Placentalia were predicted to be social (Figure 4.4c). The evolutionary history of diet in placental mammals is shown in Figure 4.4d, with all reconstructed states of insectivory, carnivory, herbivory and omnivory having predicted probabilities larger than 0.8 being indicated (Figure 4.3a,

4.4d). Although Mesozoic placentals were basically insectivorous, ancestors of Euarchonta became omnivory, thus suggesting that placental mammal diversification of this trait started before the K–Pg boundary [151].

4.4.4 Genes selected as predictors

Lasso penalized logistic regression can deal with a large number of explanatory variables. By penalizing the size of coefficients, it shrinks the coefficients of non-significant variables to the value of zero. As a result, only a small proportion of variables are left as significant predictors. Table 4.3 summarizes the coefficients of the genes left as significant predictors for the trait values. These predictors may provide hints for understanding the mechanisms underlying these traits. Some of these predictors indicate a possible direct functional relationship with the trait. For example, genes related to the brain or neural system were detected as predictors of sociality (e.g. *PRICKLE1, PHF6, CPEB4,* and *RNF19A*) [163-166]. In addition, genes involved in meiosis (e.g. *ACTR8* and *INO80D*) [167, 168], embryonic stem cell plasticity (e.g. *EAH1*) [169] and sexual hormone synthesis (e.g. *STAR*) [170] were detected as predictors of reproductive seasonality, while genes associated with mating type, spermatogenesis and male fertility (e.g. *MTF1, SPERT* and *CTCF*) [171, 172] were detected as predictors of mating system (monogamous/ polygamous). The selected predictors also included genes with indirect correlations (e.g. genes functioning in development, the cell cycle or gene expression regulation) or unclear functions (e.g. *MCMDC2, OLFM3* and *AMMECR1L,* see Table 4.3).

4.4.5 Post-K–Pg nocturnal bottleneck

Mammals, especially placental mammals, began to diversify 10–20 Ma before the K–Pg boundary ([151, 173], Figure 4.4). Some lineages became diurnal (e.g. Euarchonta and Ferungulata; Figure 4.4a) and adapted to a diet broader than insects (e.g. Euarchonta and Glires; Figure 4.3a, 4.4d). Reproductive seasonality is related to nutrition supply [174]. Although common ancestors of Placentalia were seasonal breeding, ancestors of several lineages (e.g. Atlantogenata and Boreoeutheria) became year-round breeding before the K–Pg boundary (Figure 4.4b), indicating a relatively favourable nutritional situation. This evidence suggests the Mesozoic placentals had already possessed diversified life historical traits, and such diversification may have been the driving force of the rapid diversification after the K-Pg boundary.

Because dinosaurs were basically diurnal [155], it has been suggested that the mammal nocturnal bottleneck lasted until the K–Pg boundary. My study provides a new insight on the placental nocturnal bottleneck, especially on its timing. Although nocturnality/diurnality tendencies of the Mesozoic placental mammals evidently varied, I conclude that all ancestral lineages of placentals became nocturnal while passing though the K–Pg boundary and remained so until the EOT (~33.9 Mya). Therefore, this suggests that a nocturnal bottleneck occurred after the K–Pg boundary rather than during the Mesozoic era.

However, I note that, my rate-based approach only reconstructs ancestral states of surviving lineages. It does not infer the states of extinct orders. Therefore, my finding (Figure 4.4a) does not necessarily mean that all placentals were restricted to nocturnal niches during the Paleocene–Eocene Thermal Maximum (PETM). Fossil records have revealed a major turnover of many organisms and several new placental orders appeared during this PETM. Therefore, restriction of all placental lineages to nocturnal niches during the PETM seems unlikely. Whereas average global temperatures were elevated by 5–8°C during the PETM [150], global temperatures abruptly decreased at the Eocene-Oligocene boundary and this low-temperature period lasted for about 10 Ma [175]. It is possible the nocturnal traits that assist thermoregulation in cold served as preadaptations that aided survival during the global cooling event of the EOT. In contrast, diurnal species may not be as good as nocturnal species at keeping active in cold environments. In other words, the selective extinction of diurnal species in the EOT might explain why ancestors of the extant Placentalia lineages were nocturnal After the EOT, ancestors of extant placental mammals probably began to occupy diurnal niches.

Currently, 89 mammalian genomes were available for the rate-based ancestral trait prediction. This procedure will become even more accurate with increasing genomic data and with better understanding of genotype-phenotype relationships. The phylogenetic-inertia based procedure is powerful, when the criterion of minimum evolution works. The rate-based procedure will be effective, when the states of the traits are related with the strength of purifying selection on the genes. In this sense, the two procedures are complementary and potentially could be combined. In this

post-genomic era, rate-based prediction should thus aid my understanding of biodiversity and its underlying mechanisms.

4.5 Discussion

4.5.1 Additional comparisons of two approaches

Additional comparisons of discrete traits predicted by rate-based versus phylogenetic inertia-based approaches are summarized in Figure 4.6. Because phylogenetic inertia-based methods assume that descendants are similar in state to their ancestors, their predictive power depends on the strength of phylogenetic inertia of the trait values. The predictive power of the rate-based method does not depend on the phylogenetic inertia. Instead, it depends on the precision of the two-way ANOVA type Poisson regression, and the stability of the relation between the trait values and the gene-branch interaction profile. When a trait evolves slowly, the functional constraints on the associated genes may also evolve slowly. In such a case, rate-based and phylogenetic inertia-based methods will predict similar ancestral states (e.g., sociality; Figure 4.6a–b). On the other hand, if a trait evolves rapidly (e.g., reproductive seasonality; Figure 4.6e-f) or experiences convergent evolution (e.g., insectivory and diurnality; Figures 4.3 and 4.6c–d), the phylogeny becomes less informative and can mislead the prediction. In this case, rate-based prediction will give a more reasonable answer. I note that the ancestral state reconstruction assuming an alternative species tree [112] yielded very similar results (Figure 4.7).

Soon after my work was formally published [156], Maor *et al*., [176] reported the estimated timing when mammal went out of darkness and occupied the diurnal niches. These two papers presented different scenarios of the evolutionary history. My work suggests that ancestors of extant mammals were nocturnal between K-Pg boundary (65 Mya) and Eocene–Oligocene transition (EOT, 33.9 Mya), and only nocturnal mammals had passed the Eocene–Oligocene extinction event, 33.9 Mya. Maor *et al*., suggests that mammals began to occupy diurnal niches after the extinction of non-avian dinosaurs, the K-Pg extinct event, 65 Mya.

Wu *et al*., [156] asserts that only nocturnal lineages have passed the Eocene–Oligocene extinction event due to the global cooling during this period. Nocturnal mammals have better ability to adapt to the cold environment, while diurnal mammals may not. As a consequence, diurnal Placentals were selectively extirpated during or after the global cooling of the EOT, whereas some nocturnal lineages survived due to

preadaptations to cold environments. On the other hand, Maor *et al*. infer that mammals had lived in nocturnal niches to avoid the risk of dinosaurs, but got chance to enter diurnal niches after the extinction of dinosaurs at K-Pg boundary.

These two results analyzed different data. Also the methods of inference were different. Wu *et al*. analyzed 1185 genes of 89 mammals using a new rate-based prediction approach. On the other hand, Moar *et al*. assumed a phylogenetic tree of 2,415 species of extant mammalian species and applied the phylogenetic inertia-based ancestral states reconstruction method to their data of this trait. Therefore, it is not surprising that the results of Maor *et al*. contradict the result of the rate-based prediction approach (figure 4.4a, 4.6a). The key question is that: whether diurnal/nocturnal trait is phylogenetically conservative trait. It is also unclear whether dense taxa sampling can diminish the bias of the phylogenetic inertia-based method.

Extra evidences support my result that mammal already diversified before K-Pg. *Repenomamus,* a mammalian species that lived 125-123.2 Mya in China*,* were discovered with fragmentary skeleton of a juvenile *Psittacosaurus* (dinosaur) preserved in its stomach [177, 178]. Besides *Repenomamus,* the dental morphology of several other Cretaceous mammals (145 -75 Mya) indicates they were also carnivorous [179]. These fossil mammals were extinct lineages and left no offspring, however, they offered evidence that mammal before K-Pg may already possessed the ability to compete with dinosaurs. Because they were predators rather than preys of "diurnal" dinosaurs, they were likely diurnal animal as well. It will not be surprising that ancestors of extant placental mammals became diurnal before K-Pg. Another evidence comes from the eye shape of Primates. Although most lineages in Mammalia have a nocturnal eye shape [180], it is known that Primates, especially Apes are different [181]. It is also not surprising that ancestral Euarchonta (around 70 Mya, see figure 3.3) were diurnal, due to basal lineages of Euarchonta generally have big eye. A basal lineage of Euarchonta is the tree shrews, and they are diurnal Further evidence may be necessary to understand the time when ancestral mammals went out of darkness.

Phylogenetic inertia-based approaches assume that the ancestral states resembles with the states of their offspring to some extent. Because the supporting information from the extant species decays with the evolutionary times back to the past, the predicted

probability of the states at deep nodes may become similar to the frequencies at the terminal nodes. This is especially the case when trait evolves rapidly (e.g. for reproductive seasonality, see Figure 4.6f). When traits show high phylogenetic inertia (e.g. Pagel's λ is close to 1, see Table 4.1), the information on the states will be transcended far towards deep nodes (e.g. Arboreality, Figure 4.5a). Insectivory had the high phylogenetic inertia (Pagel's λ =1), but had experienced convergent evolution at several places of the mammalian tree, such as the lineages toward hedgehog and tenrec. As a result, the LOOCV of WNN shows limited prediction power, and phylogenetic inertia-based ancestral state reconstruction of insectivory provided a misleading result.

For diurnality, WNN also performs well for this trait (AUC = 0.879), so it is unclear whether phylogenetic inertia-based approach works well for the deep nodes. The time when placental mammal began to occupy diurnal niches is still under debate.

4.5.2 Integrating the two approaches

Phylogenetic-inertia prediction of ancestral states performs well when the trait evolves slowly. The prediction may be biased, when the trait evolves fast or experiences convergent evolution. The rate-based prediction method does not use the assumption of phylogenetic inertia of trait evolution but utilizes the estimated branch lengths of gene trees. The performance depends on the uncertainty of the estimated gene-branch interaction. I can integrate the two predictive procedures as follows. Let π_R and π_{PI} be the rate-based and the phylogenetic inertia-based posterior probabilities of the trait-value being 1 respectively. I note that the phylogenetic inertia-based posterior odds is the product of the prior odds and the likelihood odds:

$$\frac{\pi_{PI}}{1-\pi_{PI}} = \frac{\pi}{1-\pi} \times \frac{L(y \mid y_0 = 1)}{L(y \mid y_0 = 0)}. \qquad (4)$$

Here π is the prior probability that the trait value at the internal node is 1. $L(y \mid y_0 = 1)$ and $L(y \mid y_0 = 0)$ are the conditional likelihoods of the trait values, y, given the value at the internal node to be 1 and 0 respectively. By replacing the π in equation (4) by π_R, I obtain the rate combined with phylogenetic inertia posterior odds as

$$\frac{\pi_{R+PI}}{1-\pi_{R+PI}} = \frac{\pi_R}{1-\pi_R} \times \frac{L(y \mid y_0 = 1)}{L(y \mid y_0 = 0)}. \quad (5)$$

By taking the odds ratio of (4) and (5), I have

$$\frac{\frac{\pi_{PI}}{1-\pi_{PI}}}{\frac{\pi_{R+PI}}{1-\pi_{R+PI}}} = \frac{\frac{\pi}{1-\pi}}{\frac{\pi_R}{1-\pi_R}},$$

from which obtain

$$\frac{\pi_{R+PI}}{1-\pi_{R+PI}} = \frac{\frac{\pi_{PI}}{1-\pi_{PI}} \times \frac{\pi_R}{1-\pi_R}}{\frac{\pi}{1-\pi}}.$$

These two approaches could be integrated in the future. However, I emphasize the difference in the performance of the two predictive procedures in this paper due to mybelief that it is first important to understand their individual properties. I note that additional future evaluations of these procedures would be desirable (e.g., via LOOCV or additional information from the fossil record).

4.6 Conclusion

My rate-based ancestor state reconstruction method is an independent alternative approach that has higher predictive power than phylogenetic inertia-based approaches depending on the trait that were analysed. It is also an unbiased approach to identify genes contributing to the evolution of traits without relying on functional annotations. With genomic data growing up, rate-based ancestor state reconstruction can be even more accurate. I hope this approach can help the understanding of mammal evolution and find the mechanisms under the biological traits.

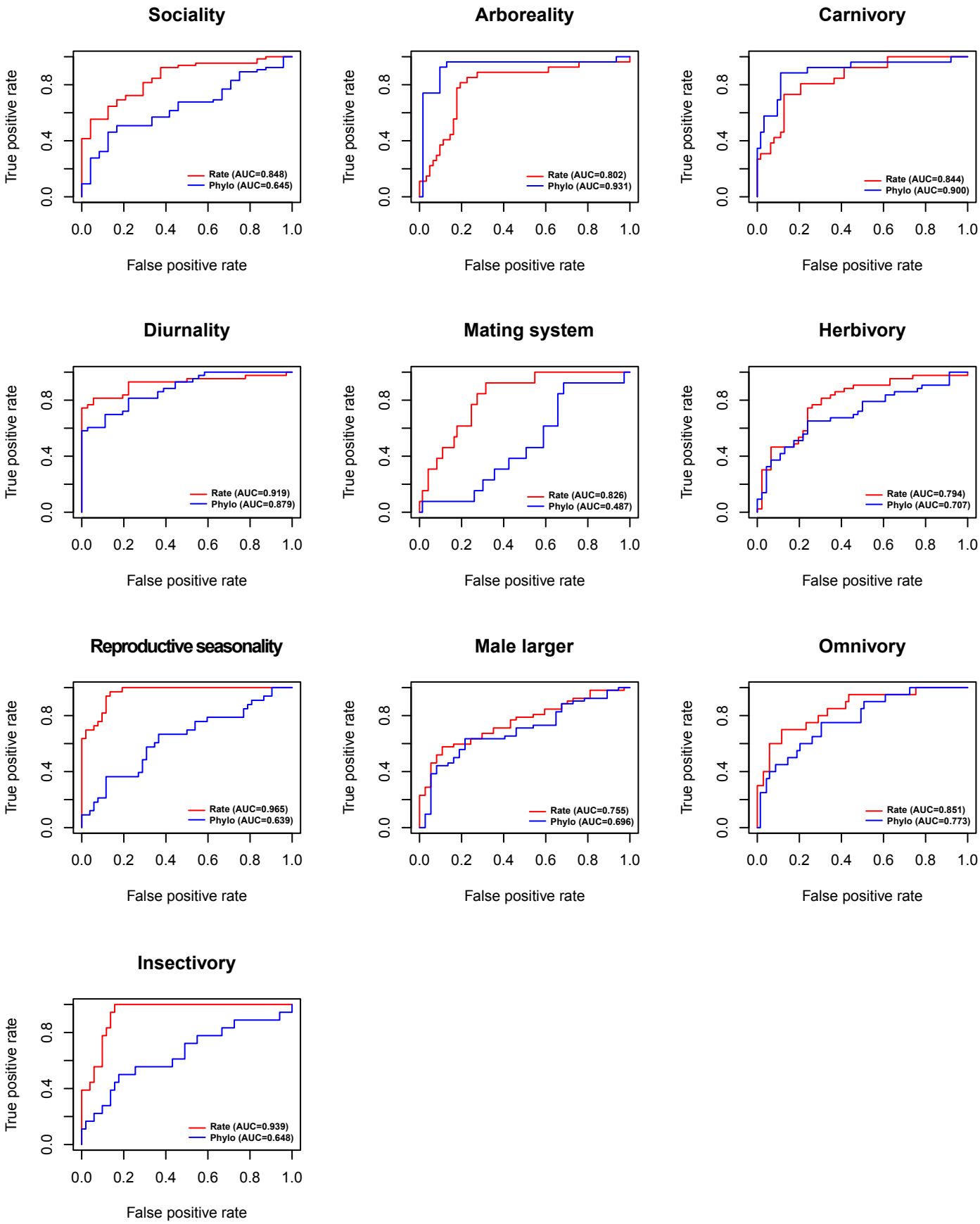


Figure 4.1 Receiver operating characteristic curves (AUCs) obtained from rate- and phylogenetic inertia- (weighted nearest-neighbour) based prediction methods using terminal trait states.

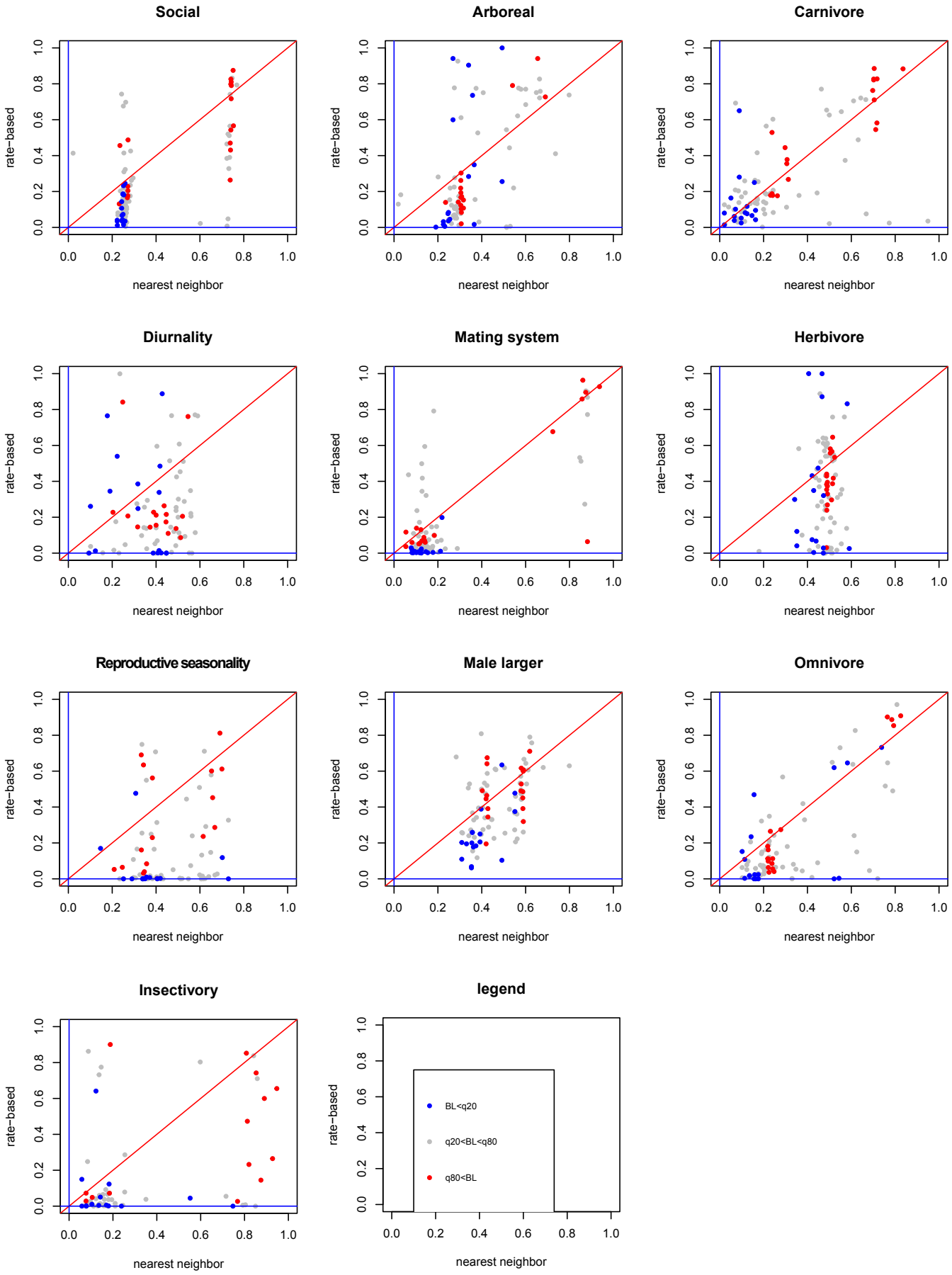
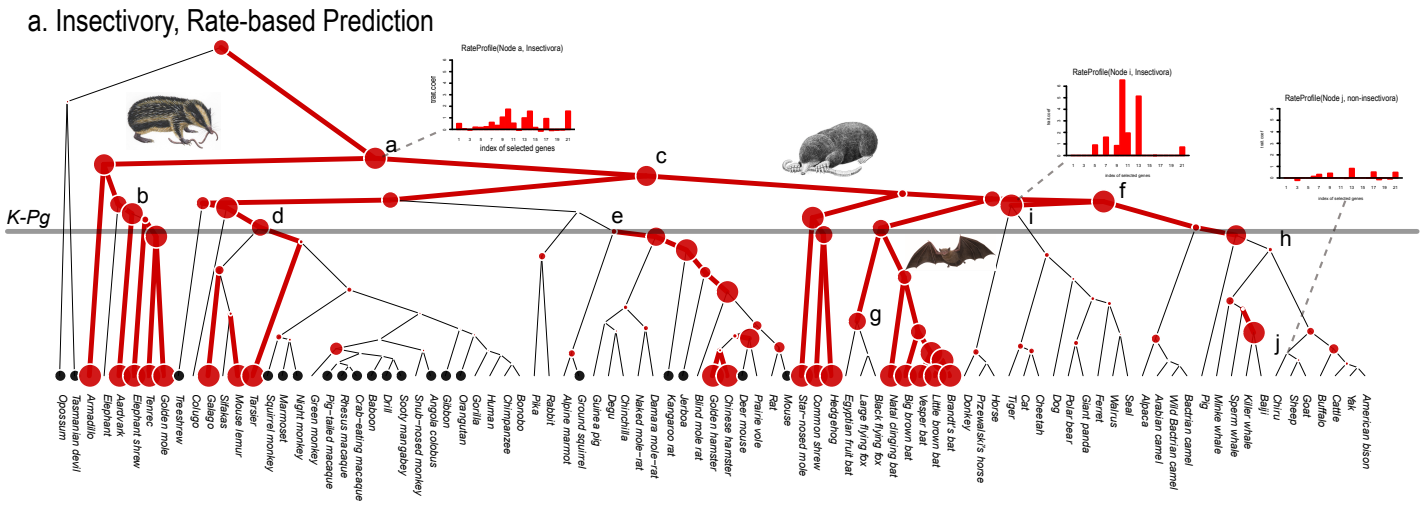

Social
Arboreal
Carnivore
Diurnality
Mating system
Herbivore
Reproductive seasonality
Male larger
Omnivore
Insectivory
legend
rate−based
nearest neighbor
0.0
0.2
0.4
0.6
0.8
1.0
BL<q20
q20<BL<q80
q80<BL

Figure 4.2. The effect of branch length on prediction error

Prediction error was calculated by taking the absolute value of predicted posterior probability minus either trait states for rate-based prediction (y-axis) or nearest-neighbour (N-N; x-axis) methods, respectively, under leave-one-out cross-validation (LOOCV). Long branches were defined as those with lengths larger than the 80% quantile of all branches (red dots in the figure), while branches with lengths less than the 20% quantile were defined as short branches (blue dots). The x = y diagonal line is also shown in each figure. A dot falling under the diagonal indicates that rate-based prediction gave a smaller prediction error and thus outperformed WNN; conversely, if a dot is above the diagonal, WNN gave a better prediction. With respect to diurnality, reproductive seasonality and insectivory traits, rate-based prediction outperformed WNN for both long and short branches.


a. Insectivory, Rate-based Prediction
RateProfile(Node a, Insectivora)
RateProfile(Node i, Insectivora)
RateProfile(Node j, non-insectivora)
index of selected genes
K-Pg
a
b
c
d
e
f
g
h
i
j
Opossum
Tasmanian devil
Armadillo
Elephant
Aardvark
Elephant shrew
Tenrec
Golden mole
Treeshrew
Colugo
Galago
Slilakas
Mouse lemur
Tarsier
Squirrel monkey
Marmoset
Night monkey
Green monkey
Pig−tailed macaque
Rhesus macaque
Crab−eating macaque
Baboon
Drill
Snub−nosed monkey
Scoty mangabey
Angola colobus
Gibbon
Orangutan
Gorilla
Human
Chimpanzee
Bonobo
Pika
Rabbit
Alpine marmot
Ground squirrel
Guinea pig
Degu
Chinchilla
Naked mole−rat
Damara mole−rat
Jerboa
Kangaroo rat
Blind mole rat
Golden hamster
Chinese hamster
Deer mouse
Prairie vole
Rat
Mouse
Star−nosed mole
Common shrew
Hedgehog
Egyptian fruit bat
Large flying fox
Black flying bat
Natal clinging bat
Big brown bat
Vesper bat
Little brown bat
Brandt's bat
Donkey
Przewalski's horse
Horse
Tiger
Cat
Cheetah
Dog
Polar bear
Giant panda
Ferret
Walrus
Seal
Alpaca
Arabian camel
Wild Bactrian camel
Bactrian camel
Pig
Minke whale
Sperm whale
Killer whale
Baiji
Chiru
Sheep
Goat
Buffalo
Cattle
Yak
American bison

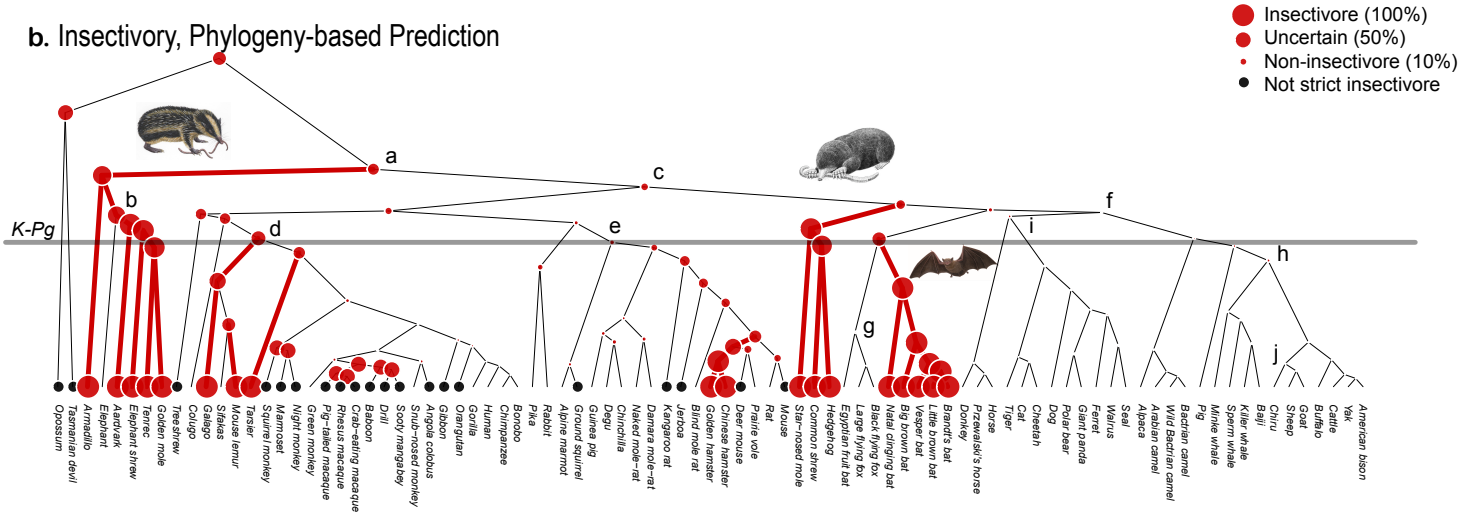
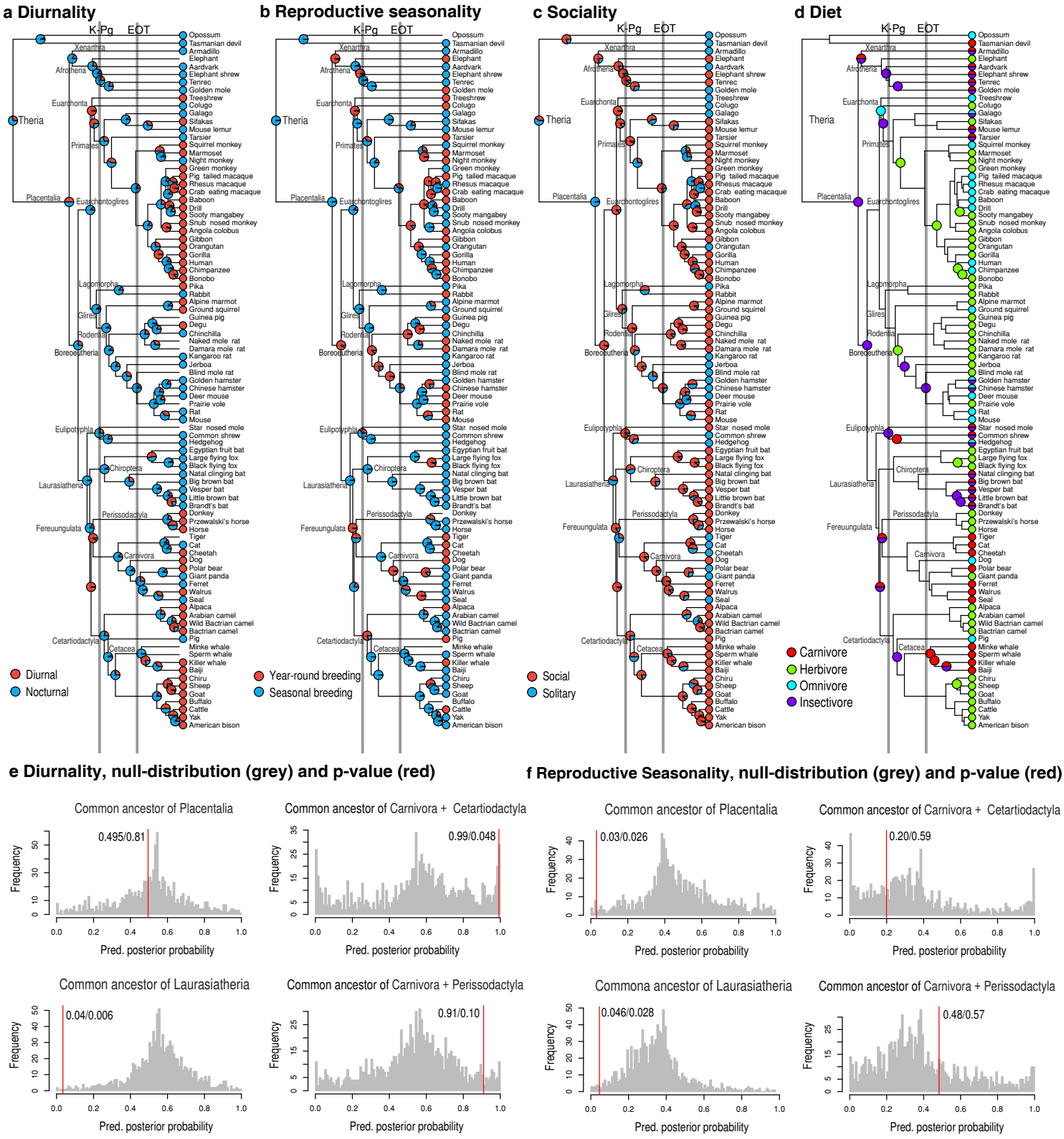

b. Insectivory, Phylogeny-based Prediction
Insectivore (100%)
Uncertain (50%)
Non-insectivore (10%)
Not strict insectivore
K-Pg
a
b
c
d
e
f
g
h
i
j
Opossum
Tasmanian devil
Armadillo
Elephant
Aardvark
Elephant shrew
Tenrec
Golden mole
Treeshrew
Colugo
Galago
Slilakas
Mouse lemur
Tarsier
Squirrel monkey
Marmoset
Night monkey
Green monkey
Pig−tailed macaque
Rhesus macaque
Crab−eating macaque
Baboon
Drill
Scoty mangabey
Snub−nosed monkey
Angola colobus
Gibbon
Orangutan
Gorilla
Human
Chimpanzee
Bonobo
Pika
Rabbit
Alpine marmot
Ground squirrel
Guinea pig
Degu
Chinchilla
Naked mole−rat
Damara mole−rat
Jerboa
Kangaroo rat
Blind mole rat
Golden hamster
Chinese hamster
Deer mouse
Prairie vole
Rat
Mouse
Star−nosed mole
Common shrew
Hedgehog
Egyptian fruit bat
Large flying fox
Black flying bat
Natal clinging bat
Big brown bat
Vesper bat
Little brown bat
Brandt's bat
Donkey
Przewalski's horse
Horse
Tiger
Cat
Cheetah
Dog
Polar bear
Giant panda
Ferret
Walrus
Seal
Alpaca
Arabian camel
Wild Bactrian camel
Bactrian camel
Pig
Minke whale
Sperm whale
Killer whale
Baiji
Chiru
Sheep
Goat
Buffalo
Cattle
Yak
American bison

Figure 4.3. Evolutionary history of insectivory

(a) Evolutionary history of insectivory reconstructed by rate-based prediction. Examples of the lasso-penalized logistic regression input function (gene-branch interaction × coefficients for each gene selected as predictor) are shown at nodes predicted as insectivorous (nodes a and i) or non-insectivorous (node j). The null-distributions of predicted probability were obtained by 1000 replications of training the rate-based predictions on permuted trait values at the terminal nodes and then predicting ancestral states. Here, the null-hypothesis assumes independence between trait values and gene-branch interactions (see Materials and Methods in Chapter 4). The two-sided p-value at each node was obtained by comparing the predicted ancestral state probabilities to their null-distributions. Examples of null-distributions and the setting of two-sided p-values are shown in Fig. 4.4e-f. The pairs of numbers at the labelled nodes are the predicted probabilities (left) and the two-sided p-values (right). **(b)** Evolutionary history of insectivory reconstructed by phylogenetic inertia-based prediction. The area of the circle at each node is proportional to the predicted probability that the animal is insectivorous, with the absence of a circle indicating a predicted probability of 0. Animals eating insects occasionally or as only a small part of their diet were not considered to be strict insectivores and were treated as missing data in the analysis. Nodes are labelled as follows: a, Placentalia; b, Afroinsectiphilia; c, Boreoeutheria; d, Primates; e, Rodents; f, Laurasiatheria; g, Ferungulata; h, Megachiroptera; i, Cetartiodactyla; j, Carnivora + Perissodactyla; k, Carnivora ; l, Goats+Chirus.


a Diurnality
b Reproductive seasonality
c Sociality
d Diet
K-Pg
EOT
Opossum
Tasmanian devil
Armadillo
Elephant
Aardvark
Elephant shrew
Tenrec
Golden mole
Treeshrew
Colugo
Galago
Sifakas
Mouse lemur
Tarsier
Squirrel monkey
Marmoset
Night monkey
Green monkey
Pig tailed macaque
Rhesus macaque
Crab eating macaque
Baboon
Drill
Sooty mangabey
Snub nosed monkey
Angola colobus
Gibbon
Orangutan
Gorilla
Human
Chimpanzee
Bonobo
Pika
Rabbit
Alpine marmot
Ground squirrel
Guinea pig
Degu
Chinchilla
Naked mole rat
Damara mole rat
Kangaroo rat
Jerboa
Blind mole rat
Golden hamster
Chinese hamster
Deer mouse
Prairie vole
Rat
Mouse
Star nosed mole
Common shrew
Hedgehog
Egyptian fruit bat
Large flying fox
Black flying fox
Natal clinging bat
Big brown bat
Vesper bat
Little brown bat
Brandt's bat
Donkey
Przewalski's horse
Horse
Tiger
Cat
Cheetah
Dog
Polar bear
Giant panda
Ferret
Walrus
Seal
Alpaca
Arabian camel
Wild Bactrian camel
Bactrian camel
Pig
Minke whale
Sperm whale
Killer whale
Baiji
Chiru
Sheep
Goat
Buffalo
Cattle
Yak
American bison
Theria
Afrotheria
Xenarthra
Euarchonta
Primates
Placentalia
Euarchontoglires
Lagomorpha
Glires
Rodentia
Boreoeutheria
Eulipotyphla
Chiroptera
Laurasiatheria
Perissodactyla
Fereuungulata
Carnivora
Cetartiodactyla
Cetacea
Diurnal
Nocturnal
Year-round breeding
Seasonal breeding
Social
Solitary
Carnivore
Herbivore
Omnivore
Insectivore

e Diurnality, null-distribution (grey) and p-value (red)
f Reproductive Seasonality, null-distribution (grey) and p-value (red)
Common ancestor of Placentalia
0.495/0.81
Common ancestor of Carnivora + Cetartiodactyla
0.99/0.048
Common ancestor of Placentalia
0.03/0.026
Common ancestor of Carnivora + Cetartiodactyla
0.20/0.59
Common ancestor of Laurasiatheria
0.04/0.006
Common ancestor of Carnivora + Perissodactyla
0.91/0.10
Commona ancestor of Laurasiatheria
0.046/0.028
Common ancestor of Carnivora + Perissodactyla
0.48/0.57
Frequency
Pred. posterior probability

Figure 4.4 Evolutionary histories of diurnality, reproductive seasonality, sociality and diet (caption page)

(a–c) Evolutionary histories of diurnality, reproductive seasonality and sociality reconstructed by rate-based prediction. The pie chart shows the predicted probability of each trait state. (**d)** Evolutionary history of diet including reconstructed states of insectivory, carnivory, herbivory and omnivory. To avoid a misleading picture due to uncertainty, pie charts indicating the traits with predicted probabilities larger than 0.8 are shown at nodes. **(e-f)** Null-distribution of predicted probabilities for selected nodes and two-sided p-values by 1000 replications of analysis with permuted terminal trait states. The pairs of numbers at the labelled nodes are the predicted probabilities (left) and the two-sided p-values (right).

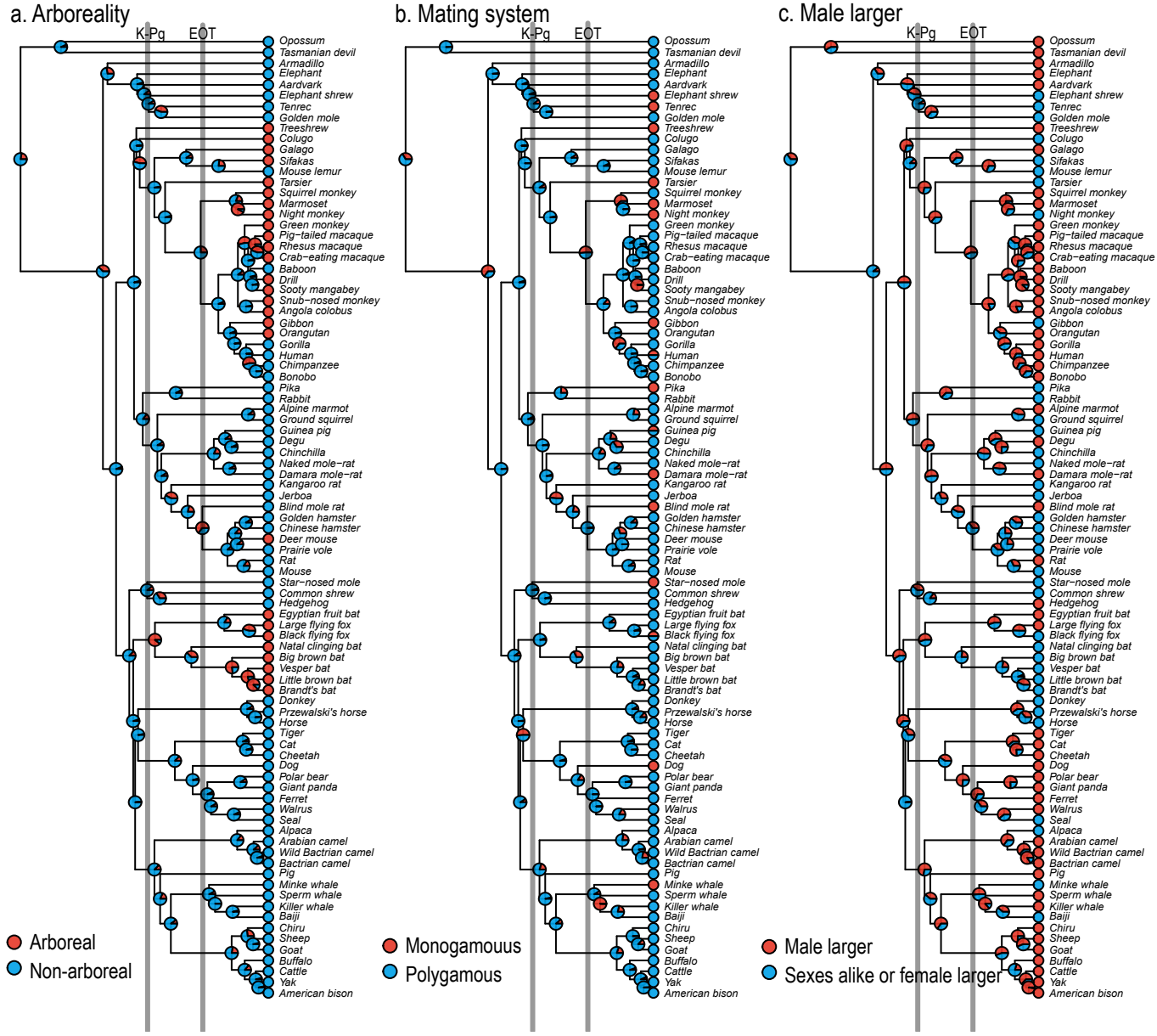


Figure 4.5 Evolutionary histories of arboreality, monogamy and male-biased sexual size dimorphism

Evolutionary histories of arboreality, monogamy and male-biased sexual size dimorphism reconstructed by rate-based prediction are shown in **a–c**, respectively. Pie charts show the posterior probability of each trait state.

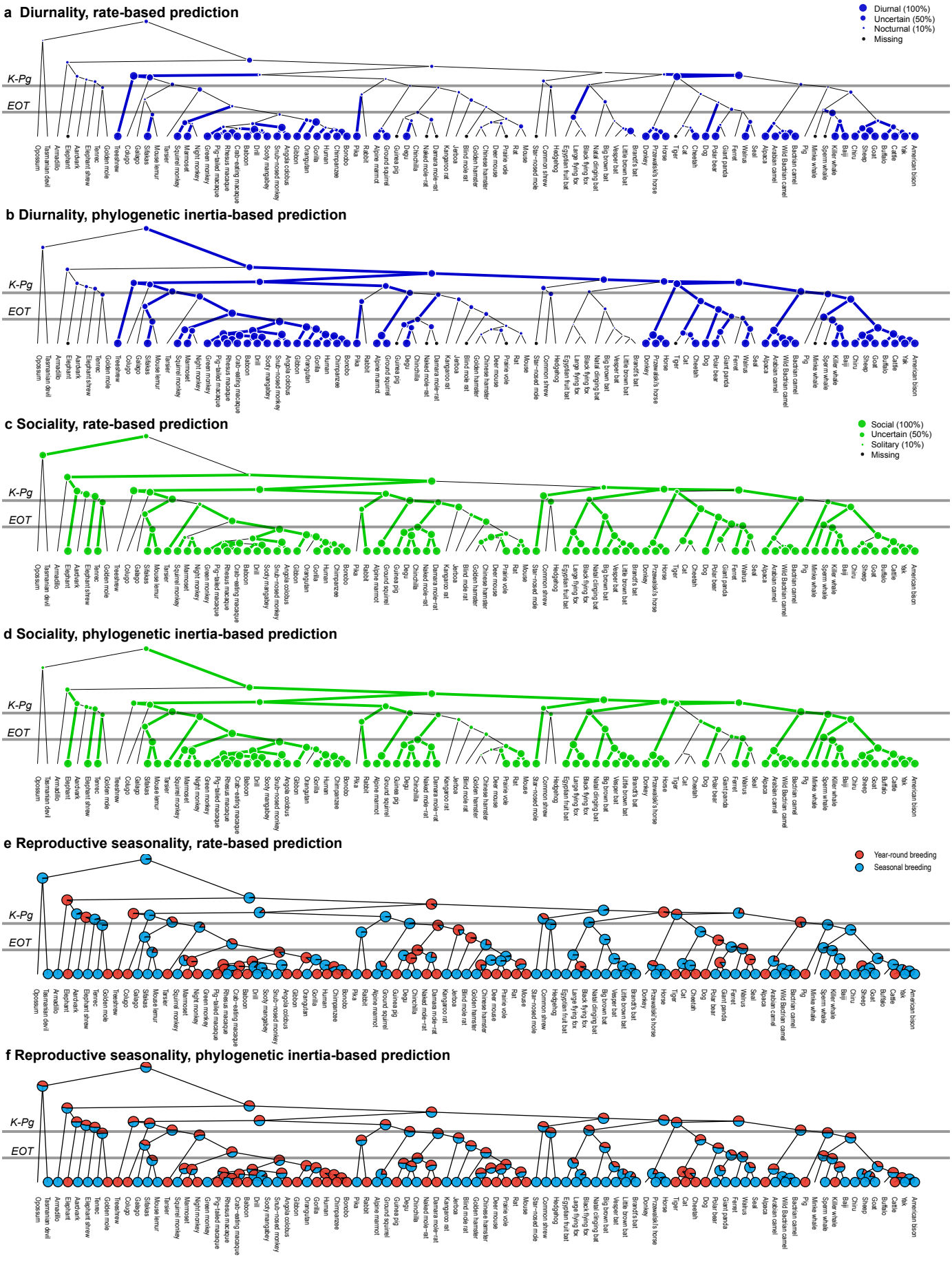
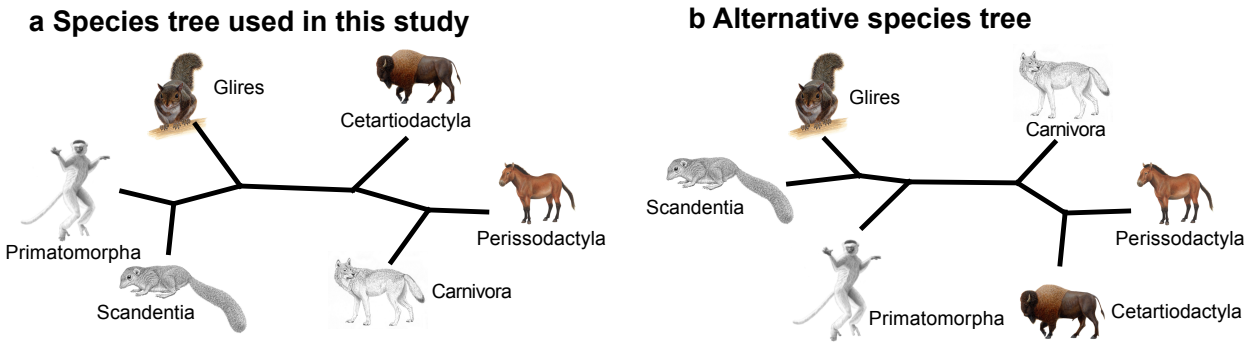
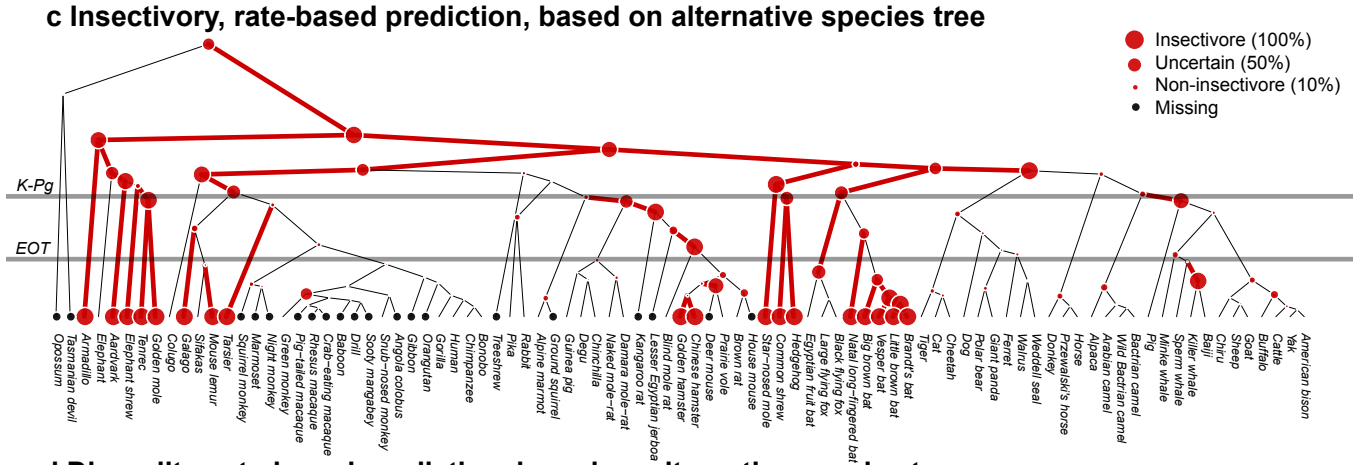
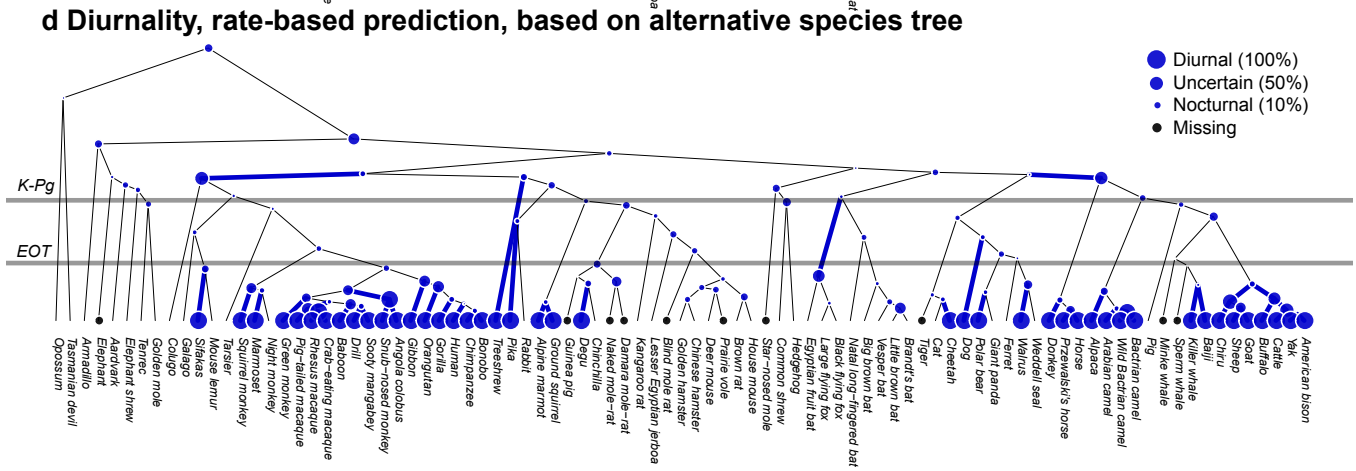

a Diurnality, rate-based prediction
Diurnal (100%)
Uncertain (50%)
Nocturnal (10%)
Missing
b Diurnality, phylogenetic inertia-based prediction
c Sociality, rate-based prediction
Social (100%)
Uncertain (50%)
Solitary (10%)
Missing
d Sociality, phylogenetic inertia-based prediction
e Reproductive seasonality, rate-based prediction
Year-round breeding
Seasonal breeding
f Reproductive seasonality, phylogenetic inertia-based prediction
K-Pg
EOT
Opossum
Tasmanian devil
Armadillo
Elephant
Aardvark
Elephant shrew
Tenrec
Golden mole
Treeshrew
Colugo
Galago
Sifakas
Mouse lemur
Tarsier
Squirrel monkey
Marmoset
Night monkey
Green monkey
Pig-tailed macaque
Rhesus macaque
Crab-eating macaque
Baboon
Drill
Baboon
Angola colobus
Snub-nosed monkey
Scaly mangabey
Gibbon
Orangutan
Gorilla
Human
Chimpanzee
Bonobo
Pika
Rabbit
Alpine marmot
Ground squirrel
Guinea pig
Degu
Chinchilla
Naked mole-rat
Damara mole-rat
Kangaroo rat
Jerboa
Blind mole rat
Golden hamster
Chinese hamster
Deer mouse
Prairie vole
Rat
Mouse
Star-nosed mole
Common shrew
Hedgehog
Egyptian fruit bat
Large flying fox
Black flying fox
Natal clinging bat
Big brown bat
Vesper bat
Little brown bat
Brandt's bat
Donkey
Przewalski's horse
Horse
Tiger
Cat
Cheetah
Dog
Polar bear
Giant panda
Ferret
Walrus
Seal
Alpaca
Arabian camel
Wild Bactrian camel
Pig
Minke whale
Sperm whale
Killer whale
Baiji
Chiru
Sheep
Goat
Buffalo
Cattle
Yak
American bison

Figure 4.6 Evolutionary history of diurnality, sociality and reproductive seasonality by both approaches

Evolutionary histories of diurnality (**a–b**), sociality (**c–d**) and reproductive seasonality (**e-f**) reconstructed by rate-based (**a, c and e**) and phylogenetic inertia-based (**b**, **d and f**) prediction methods, respectively. **a-d.** The area of the circle at each node is proportional to the predicted probability that an animal is diurnal (or social), with the absence of a circle indicating a nocturnal (or solitary) lifestyle**. e-f.** Pie charts show the predicted probability of each trait state. Animals with ambiguous trait states were treated as missing data during the analysis.


a Species tree used in this study
Glires
Cetartiodactyla
Primatomorpha
Perissodactyla
Scandentia
Carnivora
b Alternative species tree
Glires
Carnivora
Scandentia
Perissodactyla
Primatomorpha
Cetartiodactyla


c Insectivory, rate-based prediction, based on alternative species tree
Insectivore (100%)
Uncertain (50%)
Non-insectivore (10%)
Missing
K-Pg
EOT


d Diurnality, rate-based prediction, based on alternative species tree
Diurnal (100%)
Uncertain (50%)
Nocturnal (10%)
Missing
K-Pg
EOT

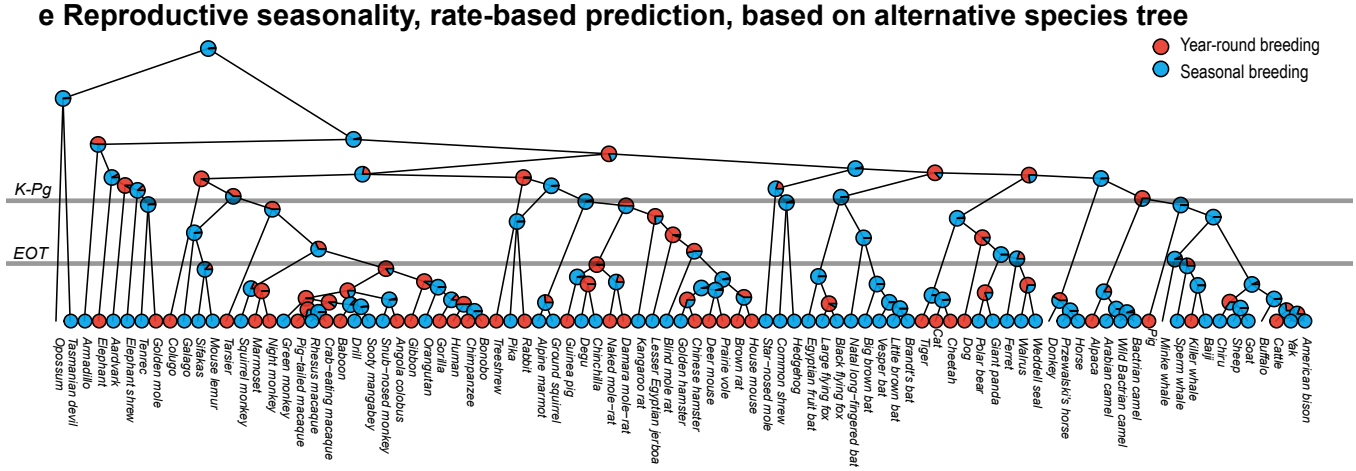

e Reproductive seasonality, rate-based prediction, based on alternative species tree
Year-round breeding
Seasonal breeding
K-Pg
EOT

Figure 4.7 Evolutionary histories of insectivory, diurnality, and reproductive seasonality by using alternative species tree

(a) Species tree used in this study. I assume ((Scandentia, Primatomorpha), Glires) for the phylogenetic position of Scandentia, while ((Perissodactyla, Carnivora), Cetartiodactyla) for the phylogenetic position of Perissodactyla follows [70, 111]. (**b)** Alternative species tree. I assume ((Scandentia, Glires), Primatomorpha) for the phylogenetic position of Scandentia, while ((Perissodactyla, Cetartiodactyla), Carnivora) for the phylogenetic position of Perissodactyla follows [112]. **c-e.** Evolutionary histories of insectivory (**c**), diurnality (**d**) and reproductive seasonality (**e**) reconstructed by rate-based prediction using alternative species tree [112]. The reconstructed evolutionary history of insectivory, diurnality, and reproductive seasonality using alternative species tree are consistent with Figures 6 and 9.

Table 4.1 Predictive power of rate-based and weighted nearest-neighbour methods.

		Rate-based Prediction		WNN***	
Trait*	**Pagel's λ**	**AUC**	**Accuracy****	**AUC**	**Accuracy***
Sociality	1.000	0.848	0.843	0.645	0.742
Diurnality	1.000	0.919	0.873	0.879	0.797
Year-round	0.764	0.965	0.906	0.639	0.682
Arboreality	0.968	0.802	0.809	0.931	0.910
Mating system	0.461	0.826	0.860	0.487	0.849
Male larger	1.000	0.755	0.708	0.696	0.697
Carnivory	1.000	0.844	0.831	0.900	0.888
Herbivory	0.963	0.794	0.753	0.707	0.708
Omnivory	0.801	0.851	0.865	0.773	0.820
Insectivory	1.000	0.939	0.884	0.648	0.768

* **Sociality: social/ solitary. Diurnality: diurnal, 1/ nocturnal, 0. Reproductive Seasonality (Year-round in table): year-round breeding, 1/ seasonal breeding, 0. Insectivory: insectivore, 1 or not, 0. Arboreality: arboreal and flying animal, 1/ terrestrial and marine animal, 0. Mating system: monogamous, 1/ polyandrous, 0. Male-biased sexual size dimorphism (Male Larger in table): male larger, 1/ sexes alike or female larger, 0. Carnivory: carnivore, 1 or not, 0. Herbivory: herbivore, 1 or not, 0, Omnivory: omnivore, 1 or not, 0.

** Accuracy depends on cut-off points. Table shows the maximum accuracy for each trait

*** Weighted nearest-neighbour method

Table 4.2 Scientific name and life history traits of 89 mammals (part 1)

	Species Names	Sociality	Diurnality	*Yearly	Insectivory	Resource
Cheetah	*acinonyx jubatus*	Solitary	diurnal	yearly	not	ADW (*acinonyx jubatus*)
Giant panda	*ailuropoda melanoleuca*	Solitary	nocturnal	seasonal	not	ADW (*ailuropoda melanoleuca*)
Night monkey	*aotus nancymaae*	Social	nocturnal	yearly	not strick	ADW (*aotus nancymaae*)
Minke whale	*balaenoptera acutorostrata*	Social	natatorial	both	not	ADW (*balaenoptera acutorostrata*)
American bison	*bison bison bison*	Social	diurnal	seasonal	not	ADW (*bison bison bison*)
Yak	*bos mutus*	Social	diurnal	seasonal	not	ADW (*bos grunniens*)
Cattle	*bos taurus*	Social	diurnal	yearly	not	ADW (*bos taurus*)
Buffalo	*bubalus bubalis*	Social	diurnal	both	not	ADW (*bubalus bubalis*)
Marmoset	*callithrix jacchus*	Social	diurnal	yearly	not strick	ADW (*callithrix jacchus*) [182]
Bactrian camel	*camelus bactrianus*	Social	diurnal	seasonal	not	ADW (*camelus bactrianus*)
Arabian camel	*camelus dromedarius*	Social	diurnal	seasonal	not	ADW (*camelus dromedarius*)
Wild Bactrian camel	*camelus ferus*	Social	diurnal	seasonal	not	ADW (*camelus ferus*), IUCN (*camelus ferus*)
Dog	*canis familiaris*	Social	diurnal	yearly	not	ADW (*canis familiaris*)
Goat	*capra hircus*	Social	diurnal	seasonal	not	ADW (*capra hircus*)
Guinea pig	*cavia porcellus*	Social	crepuscular	yearly	not	ADW (*cavia porcellus*)
Sooty mangabey	*cercocebus atys*	Social	diurnal	seasonal	not strick	ADW (*cercocebus atys*)
Chinchilla	*chinchilla lanigera*	Social	nocturnal	seasonal	not	ADW (*chinchilla lanigera*)
Green monkey	*chlorocebus sabaeus*	Social	diurnal	seasonal	not	ADW (*chlorocebus sabaeus*)
Golden mole	*chrysochloris asiatica*	Solitary	nocturnal	yearly	yes	ADW (*chrysochloris asiatica*),
Angola colobus	*colobus angolensis*	Social	diurnal	yearly	not strick	ADW (*colobus angolensis*)
Star-nosed mole	*condylura cristata*	Social	Cathemeral	seasonal	yes	ADW (*condylura cristata*)
Chinese hamster	*cricetulus griseus*	Solitary	nocturnal	yearly	yes	ADW (Cricetinae)
Armadillo	*dasypus novemcinctus*	Solitary	nocturnal	seasonal	yes	ADW (*dasypus novemcinctus*)
Kangaroo rat	*dipodomys ordii*	Solitary	nocturnal	seasonal	not strick	ADW (*dipodomys ordii*)
Tenrec	*echinops telfairi*	Social	nocturnal	seasonal	yes	ADW (Tenrecidae), IUCN
Elephant shrew	*elephantulus edwardii*	Social	nocturnal	seasonal	yes	ADW (*Elephantulus intufi*)
Big brown bat	*eptesicus fuscus*	Social	nocturnal	seasonal	yes	ADW (*eptesicus fuscus*)
Donkey	*equus asinus*	Social	diurnal	both	not	ADW (*equus asinus*)
Horse	*equus caballus*	Social	diurnal	seasonal	not	ADW (*equus caballus*)
Przewalski's horse	*equus przewalskii*	Social	diurnal	seasonal	not	ADW (*equus przewalskii*)
Hedgehog	*erinaceus europaeus*	Solitary	nocturnal	seasonal	yes	ADW (*erinaceus europaeus*)
Cat	*felis catus*	Solitary	nocturnal	yearly	not	ADW (*felis catus*)
Damara mole-rat	*fukomys damarensis*	Social	fossorial	yearly	not	[183, 184]
Colugo	*galeopterus variegatus*	Solitary	nocturnal	yearly	not	ADW (*galeopterus variegatus*)
Gorilla	*gorilla gorilla*	Social	diurnal	yearly	not	ADW (*gorilla gorilla*)
Naked mole-rat	*heterocephalus glaber*	Social	fossorial	yearly	not	ADW (*heterocephalus glaber*)
Human	*homo sapiens*	Social	diurnal	yearly	not	ADW (*homo sapiens*)
Ground squirrel	*ictidomys tridecemlineatus*	Social	diurnal	seasonal	not strick	ADW (*ictidomys tridecemlineatus*)
Lesser Egyptian jerboa	*jaculus jaculus*	Solitary	nocturnal	seasonal	not strick	ADW (*jaculus jaculus*)
Weddell seal	*leptonychotes weddellii*	Solitary	nocturnal	seasonal	not	ADW (*leptonychotes weddellii*)
Baiji	*lipotes vexillifer*	Social	diurnal	seasonal	not	ADW (*lipotes vexillifer*)

Elephant	*loxodonta africana*	Social	Cathemeral	yearly	not	ADW (*loxodonta africana*)
Crab-eating macaque	*macaca fascicularis*	Social	diurnal	yearly	not strick	ADW (*macaca fascicularis*)
Rhesus macaque	*macaca mulatta*	Social	diurnal	seasonal	not strick	ADW (*macaca mulatta*)
Pig-tailed macaque	*macaca nemestrina*	Social	diurnal	yearly	not strick	ADW (*macaca nemestrina*)
Drill	*mandrillus leucophaeus*	Social	diurnal	seasonal	not strick	ADW (*mandrillus leucophaeus*)
Alpine marmot	*marmota marmota*	Social	diurnal	seasonal	not	ADW (*marmota marmota*), ADW (*marmota bobak*)
Golden hamster	*mesocricetus auratus*	Solitary	nocturnal	seasonal	yes	ADW (*mesocricetus auratus*)
Mouse lemur	*microcebus murinus*	Social	nocturnal	seasonal	yes	ADW (*microcebus murinus*)
Prairie vole	*microtus ochrogaster*	Social	crepuscular	yearly	not	ADW (*microtus ochrogaster*)
Natal long-fingered bat	*miniopterus natalensis*	Social	nocturnal	seasonal	yes	ADW (*miniopterus australis*), ADW (*miniopterus schreibersii*)
Opossum	*monodelphis domestica*	Solitary	nocturnal	both	not strick	ADW (*monodelphis domestica*)
House mouse	*mus musculus*	Social	nocturnal	yearly	not strick	ADW (*mus musculus*)
Ferret	*mustela putorius furo*	Social	nocturnal	seasonal	not	ADW (*mustela putorius furo*)
Brandt's bat	*myotis brandtii*	Social	nocturnal	seasonal	yes	ADW (*myotis*) [185]
Vesper bat	*myotis davidii*	Social	nocturnal	seasonal	yes	ADW (*myotis*) [185]
Little brown bat	*myotis lucifugus*	Social	nocturnal	seasonal	yes	ADW (*myotis lucifugus*)
Blind mole rat	*nannospalax galili*	Solitary	fossorial	seasonal	not	ADW (spalax ehrenbergi)
Gibbon	*nomascus leucogenys*	Social	diurnal	yearly	not strick	ADW (*nomascus leucogenys*)
Pika	*ochotona princeps*	Solitary	diurnal	seasonal	not	ADW (*ochotona princeps*)
Degu	*octodon degus*	Social	diurnal	seasonal	not	ADW (*octodon degus*)
Walrus	*odobenus rosmarus*	Social	diurnal	seasonal	not	ADW (*odobenus rosmarus*)
Killer whale	*orcinus orca*	Social	diurnal	yearly	not	ADW (*orcinus orca*)
Aardvark	*orycteropus afer afer*	Solitary	nocturnal	seasonal	yes	ADW (*orycteropus afer afer*)
Rabbit	*oryctolagus cuniculus*	Social	nocturnal	yearly	not	ADW (*oryctolagus cuniculus*)
Galago	*otolemur garnettii*	Solitary	nocturnal	seasonal	yes	ADW (*otolemur garnettii*)
Sheep	*ovis aries*	Social	diurnal	seasonal	not	ADW (*ovis aries*)
Bonobo	*pan paniscus*	Social	diurnal	yearly	not	ADW (*pan paniscus*)
Chimpanzee	*pan troglodytes*	Social	diurnal	yearly	not	ADW (*pan troglodytes*)
Tiger	*panthera tigris altaica*	Solitary	Cathemeral	yearly	not	ADW (*panthera tigris altaica*)
Chiru	*pantholops hodgsonii*	Social	diurnal	seasonal	not	ADW (pantholops hodgsonii)
Baboon	*papio anubis*	Social	diurnal	yearly	not strick	ADW (*papio anubis*)
Deer mouse	*peromyscus maniculatus*	Solitary	nocturnal	yearly	not strick	ADW (*peromyscus maniculatus*)
Sperm whale	*physeter catodon*	Social	natatorial	seasonal	not	ADW (*physeter catodon*)
Orangutan	*pongo abelii*	Solitary	diurnal	seasonal	not strick	ADW (*pongo abelii*)
Sifakas	*propithecus coquereli*	Social	diurnal	seasonal	not	ADW (*propithecus coquereli*)
Black flying fox	*pteropus alecto*	Social	nocturnal	seasonal	not	ADW (*pteropus alecto*)
Large flying fox	*pteropus vampyrus*	Social	nocturnal	seasonal	not	ADW (*pteropus vampyrus*)
Brown rat	*rattus norvegicus*	Social	nocturnal	yearly	not	ADW (*rattus norvegicus*)
Snub-nosed monkey	*rhinopithecus roxellana*	Social	diurnal	seasonal	not	ADW (*rhinopithecus roxellana*)
Egyptian fruit bat	*rousettus aegyptiacus*	Social	nocturnal	seasonal	not	ADW (*rousettus aegyptiacus*)
Squirrel monkey	*saimiri boliviensis*	Social	diurnal	seasonal	not strick	ADW (*saimiri boliviensis*)
Tasmanian devil	*sarcophilus harrisii*	Solitary	nocturnal	seasonal	not strick	ADW (*sarcophilus harrisii*)
Common shrew	*sorex araneus*	Solitary	nocturnal	seasonal	yes	ADW (*sorex araneus*)
Pig	*sus scrofa*	Social	nocturnal	yearly	not	ADW (*sus scrofa*)

Tarsier	*tarsius syrichta*	Social	nocturnal	yearly	yes	ADW (*tarsius syrichta*)
Treeshrew	*tupaia chinensis*	Solitary	diurnal	yearly	not strick	ADW (*tupaia belangeri*)
Polar bear	*ursus maritimus*	Solitary	diurnal	seasonal	not	ADW (*ursus maritimus*)
Alpaca	*vicugna pacos*	Social	diurnal	yearly	not	ADW (*lama pacos*)

* Reproductive seasonality, yearly in table.

Table 4.2 Scientific name and life history traits of 89 mammals (part 2)

	Arboreality	Mating System	*Male larger	Carnivory	Herbivory	Omnivore	Resource
Cheetah	No	polygynous	male larger	Carnivore	no	no	ADW (*acinonyx jubatus*)
Giant panda	No	polygynous	male larger	no	Herbivore	no	ADW (*ailuropoda melanoleuca*)
Night monkey	Yes	monogamous	sexes alike	no	Herbivore	no	ADW (*aotus nancymaae*)
Minke whale	No	monogamous	female larger	Carnivore	no	no	ADW (*balaenoptera acutorostrata*)
American bison	No	polygynous	male larger	no	Herbivore	no	ADW (*bison bison bison*)
Yak	No	polygynous	male larger	no	Herbivore	no	ADW (*bos grunniens*)
Cattle	No	polygynous	male larger	no	Herbivore	no	ADW (*bos taurus*)
Buffalo	No	polygynous	male larger	no	Herbivore	no	ADW (*bubalus bubalis*)
Marmoset	Yes	monogamous	male larger	no	Herbivore	no	ADW (*callithrix jacchus*) [182]
Bactrian camel	No	polygynous	male larger	no	Herbivore	no	ADW (*camelus bactrianus*)
Arabian camel	No	polygynous	male larger	no	Herbivore	no	ADW (*camelus dromedarius*)
Wild Bactrian camel	No	polygynous	male larger	no	Herbivore	no	ADW (*camelus ferus*), IUCN (*camelus ferus*)
Dog	No	monogamous	male larger	no	no	Omnivore	ADW (*canis familiaris*)
Goat	No	polygynous	male larger	no	Herbivore	no	ADW (*capra hircus*)
Guinea pig	No	both	sexes alike	no	Herbivore	no	ADW (*cavia porcellus*)
Sooty mangabey	Yes	polygynous	male larger	no	Herbivore	no	ADW (*cercocebus atys*)
Chinchilla	No	polygynous	female larger	no	Herbivore	no	ADW (*chinchilla lanigera*)
Green monkey	Yes	polygynous	male larger	no	Herbivore	no	ADW (*chlorocebus sabaeus*)
Golden mole	No	polygynous	sexes alike	Carnivore	no	no	ADW (*chrysochloris asiatica*), ADW(mole)
Angola colobus	Yes	polygynous	male larger	no	Herbivore	no	ADW (*colobus angolensis*)
Star-nosed mole	No	monogamous	sexes alike	Carnivore	no	no	ADW (*condylura cristata*)
Chinese hamster	No	polygynous	sexes alike	no	no	Omnivore	ADW (Cricetinae)
Armadillo	No	polygynous	male larger	Carnivore	no	no	ADW (*dasypus novemcinctus*)
Kangaroo rat	No	polygynous	sexes alike	no	Herbivore	no	ADW (*dipodomys ordii*)
Tenrec	No	monogamous	sexes alike	Carnivore	no	no	ADW (Tenrecidae), IUCN
Elephant shrew	No	monogamous	female larger	Carnivore	no	no	ADW (*Elephantulus intufi*)
Big brown bat	Yes(fly)	polygynous	female larger	Carnivore	no	no	ADW (*eptesicus fuscus*)
Donkey	No	polygynous	sexes alike	no	Herbivore	no	ADW (*equus asinus*)
Horse	No	polygynous	sexes alike	no	Herbivore	no	ADW (*equus caballus*)
Przewalski's horse	No	polygynous	sexes alike	no	Herbivore	no	ADW (*equus przewalskii*)
Hedgehog	No	polygynous	male larger	no	no	Omnivore	ADW (*erinaceus europaeus*)
Cat	No	polygynous	male larger	Carnivore	no	no	ADW (*felis catus*)
Damara mole-rat	No	monogamous	male larger	no	Herbivore	no	[183, 184]
Colugo	Yes	polygynous	sexes alike	no	Herbivore	no	ADW (*galeopterus variegatus*)
Gorilla	No	polygynous	male larger	no	Herbivore	no	ADW (*gorilla gorilla*)
Naked mole-rat	No	polyandrous	sexes alike	no	Herbivore	no	ADW (*heterocephalus glaber*)
Human	No	both	male larger	no	no	Omnivore	ADW (*homo sapiens*)
Ground squirrel	No	polygynous	sexes alike	no	no	Omnivore	ADW (*ictidomys tridecemlineatus*)
Lesser Egyptian jerboa	No	polygynous	female larger	no	Herbivore	no	ADW (*jaculus jaculus*)
Weddell seal	No	polygynous	female larger	Carnivore	no	no	ADW (*leptonychotes weddellii*)

Baiji	No	polygynous	female larger	Carnivore	no	no	ADW (*lipotes vexillifer*)
Elephant	No	polygynous	male larger	no	Herbivore	no	ADW (*loxodonta africana*)
Crab-eating macaque	Yes	polygynous	male larger	no	no	Omnivore	ADW (*macaca fascicularis*)
Rhesus macaque	Yes	polygynous	male larger	no	no	Omnivore	ADW (*macaca mulatta*)
Pig-tailed macaque	Yes	polygynous	male larger	no	no	Omnivore	ADW (*macaca nemestrina*)
Drill	Yes	polygynous	male larger	no	no	Omnivore	ADW (*mandrillus leucophaeus*)
Alpine marmot	No	polygynous	male larger	no	Herbivore	no	ADW (*marmota marmota*), ADW (*marmota bobak*)
Golden hamster	No	polygynous	sexes alike	no	no	Omnivore	ADW (*mesocricetus auratus*)
Mouse lemur	No	polygynous	sexes alike	Carnivore	no	no	ADW (*microcebus murinus*)
Prairie vole	No	polygynous	sexes alike	no	Herbivore	no	ADW (*microtus ochrogaster*)
Natal long-fingered bat	Yes(fly)	polygynous	sexes alike	Carnivore	no	no	ADW (*miniopterus australis*), ADW (*miniopterus schreibersii*)
Opossum	No	polygynous	male larger	no	no	Omnivore	ADW (*monodelphis domestica*)
House mouse	No	polygynous	sexes alike	no	no	Omnivore	ADW (*mus musculus*)
Ferret	No	polygynous	male larger	Carnivore	no	no	ADW (*mustela putorius furo*)
Brandt's bat	Yes(fly)	polygynous	female larger	Carnivore	no	no	ADW (*myotis*) [185]
Vesper bat	Yes(fly)	polygynous	female larger	Carnivore	no	no	ADW (*myotis*) [185]
Little brown bat	Yes(fly)	polygynous	female larger	Carnivore	no	no	ADW (*myotis lucifugus*)
Blind mole rat	No	monogamous	male larger	no	Herbivore	no	ADW (spalax ehrenbergi)
Gibbon	Yes	monogamous	sexes alike	no	Herbivore	no	ADW (*nomascus leucogenys*)
Pika	No	monogamous	sexes alike	no	Herbivore	no	ADW (*ochotona princeps*)
Degu	No	polygynous	male larger	no	Herbivore	no	ADW (*octodon degus*)
Walrus	No	polygynous	male larger	Carnivore	no	no	ADW (*odobenus rosmarus*)
Killer whale	No	polygynous	male larger	Carnivore	no	no	ADW (*orcinus orca*)
Aardvark	No	polygynous	male larger	Carnivore	no	no	ADW (*orycteropus afer afer*)
Rabbit	No	polygynous	sexes alike	no	Herbivore	no	ADW (*oryctolagus cuniculus*)
Galago	Yes	polygynous	male larger	no	no	Omnivore	ADW (*otolemur garnettii*)
Sheep	No	polygynous	male larger	no	Herbivore	no	ADW (*ovis aries*)
Bonobo	No	polygynous	male larger	no	Herbivore	no	ADW (*pan paniscus*)
Chimpanzee	No	polygynous	male larger	no	no	Omnivore	ADW (*pan troglodytes*)
Tiger	No	polygynous	male larger	Carnivore	no	no	ADW (*panthera tigris altaica*)
Chiru	No	polygynous	male larger	no	Herbivore	no	ADW (pantholops hodgsonii)
Baboon	No	polygynous	male larger	no	no	Omnivore	ADW (*papio anubis*)
Deer mouse	Yes	polygynous	sexes alike	no	no	Omnivore	ADW (*peromyscus maniculatus*)
Sperm whale	No	polygynous	male larger	Carnivore	no	no	ADW (*physeter catodon*)
Orangutan	Yes	polygynous	male larger	no	Herbivore	no	ADW (*pongo abelii*)
Sifakas	Yes	polygynous	sexes alike	no	Herbivore	no	ADW (*propithecus coquereli*)
Black flying fox	Yes(fly)	both	sexes alike	no	Herbivore	no	ADW (*pteropus alecto*)
Large flying fox	Yes(fly)	polygynous	male larger	no	Herbivore	no	ADW (*pteropus vampyrus*)
Brown rat	No	polygynous	male larger	no	no	Omnivore	ADW (*rattus norvegicus*)
Snub-nosed monkey	Yes	polygynous	male larger	no	Herbivore	no	ADW (*rhinopithecus roxellana*)
Egyptian fruit bat	Yes(fly)	polygynous	male larger	no	Herbivore	no	ADW (*rousettus aegyptiacus*)
Squirrel monkey	Yes	polygynous	male larger	no	no	Omnivore	ADW (*saimiri boliviensis*)
Tasmanian devil	No	polygynous	male larger	Carnivore	no	no	ADW (*sarcophilus harrisii*)
Common shrew	No	polygynous	sexes alike	Carnivore	no	no	ADW (*sorex araneus*)

Pig	No	polygynous	male larger	no	no	Omnivore	ADW (*sus scrofa*)
Tarsier	Yes	monogamous	sexes alike	Carnivore	no	no	ADW (*tarsius syrichta*)
Treeshrew	Yes	monogamous	male larger	no	no	Omnivore	ADW (*tupaia belangeri*)
Polar bear	No	polygynous	male larger	Carnivore	no	no	ADW (*ursus maritimus*)
Alpaca	No	polygynous	sexes alike	no	Herbivore	no	ADW (*lama pacos*)

*Male-biased Sexual Dimorphism, male larger in table

Table 4.3 Summary of genes identified as predictors for each trait

Sociality(social,1/ solitary, 0)		
Gene	**Coefficient**	**Full Name**
ZMAT3	-0.777028055	Zinc Finger Matrin-Type 3
PRICKLE1	-0.448908605	Prickle Planar Cell Polarity Protein 1
RNF19A	-0.418737217	Ring Finger Protein 19A, RBR E3 Ubiquitin Protein Ligase
SLC25A38	-0.291873789	Solute Carrier Family 25 Member 38
PHF6	-0.229256782	PHD finger protein 6
SLC25A27	-0.135066624	solute carrier family 25 member 27
TCTEX1D1	-0.106226644	Tctex1 Domain Containing 1
CPEB4	-0.095116503	Cytoplasmic Polyadenylation Element Binding Protein 4
SPCS2	-0.091144129	Signal Peptidase Complex Subunit 2
CLTA	-0.049504664	Clathrin Light Chain A
DET1	-0.016162882	De-Etiolated Homolog 1 (Arabidopsis)
KAT7	0.003581275	Lysine Acetyltransferase 7
ABI1	0.006262475	Abl Interactor 1
IFT46	0.098240927	Intraflagellar Transport 46
Diurnality (diurnal,1/ nocturnal, 0)		
Gene	**Coefficient**	**Full Name**
AMMECR1L	-0.209591048	AMMECR1 Like
CYLD	-0.096280291	Cylindromatosis
DCTN4	-0.089812415	Dynactin Subunit 4
GEM	-0.04463599	GTP-binding protein GEM
NR1H4	-0.009548895	Nuclear Receptor Subfamily 1 Group H Member 4
PMCH	0.003970611	Pro-Melanin Concentrating Hormone
CALU	0.004782391	Calumenin
TIAL1	0.007201134	TIA1 Cytotoxic Granule Associated RNA Binding Protein Like 1
CPSF4	0.009094721	Cleavage And Polyadenylation Specific Factor 4
TADA2A	0.013955868	Transcriptional Adaptor 2A
NFYC	0.015117146	Nuclear Transcription Factor Y Subunit Gamma
CDK5RAP3	0.020902842	CDK5 Regulatory Subunit Associated Protein 3
SAR1B	0.022779385	Secretion Associated Ras Related GTPase 1B
FDFT1	0.023601995	Farnesyl-Diphosphate Farnesyltransferase 1
SRSF1	0.027218966	Serine/Arginine-Rich Splicing Factor 1
WNT2B	0.027738966	Wnt Family Member 2B
TLK2	0.034771162	Tousled Like Kinase 2
CLDN18	0.037355397	Claudin 18
CTDSPL2	0.048089425	CTD Small Phosphatase Like 2
WDTC1	0.053036046	WD And Tetratricopeptide Repeats 1
RNF145	0.056782244	Ring Finger Protein 145
PSMA4	0.064173956	Proteasome Subunit Alpha 4
POU2F3	0.075753489	POU Class 2 Homeobox 3
MAPK9	0.081774219	Mitogen-Activated Protein Kinase 9

SMC3	0.089829488	Structural Maintenance Of Chromosomes 3
CTNND1	0.108373701	Catenin Delta 1
SEMA5B	0.112911534	Semaphorin 5B
KLHL12	0.152121078	Kelch Like Family Member 12
MYO1C	0.156393302	Myosin IC
CNGA4	0.244419432	Cyclic Nucleotide Gated Channel Alpha 4
PROSC	0.255341569	Proline Synthetase Co-Transcribed Homolog (Bacterial)
DFNB59	0.266247757	Deafness, Autosomal Recessive 59
GLRB	0.271467191	Glycine Receptor Beta

Yearly (year-around, 1/ seasonal, 0 reproduction)

Gene	Coefficient	Full Name
ACTR8	-2.087408331	ARP8 Actin-Related Protein 8 Homolog
OLFM3	-1.857014667	Olfactomedin 3
CCDC126	-0.374409215	Coiled-Coil Domain Containing 126
RSRC2	-0.233598676	Arginine/Serine-Rich Coiled-Coil 2
ABHD6	-0.229213759	Abhydrolase Domain Containing 6
CNOT8	-0.213965832	CCR4-NOT Transcription Complex Subunit 8
SLC4A7	-0.147693459	Solute Carrier Family 4 Member 7
LMCD1	-0.103797055	LIM And Cysteine Rich Domains 1
DPF2	-0.065265706	Double PHD Fingers 2
CPSF3	-0.063417687	Cleavage And Polyadenylation Specific Factor 3
SCARB1	-0.051981076	Scavenger Receptor Class B Member 1
SLC7A14	-0.04236182	Solute Carrier Family 7 Member 14
ARPC5L	-0.00257765	Actin Related Protein 2/3 Complex Subunit 5-Like
PSMC2	-4.12E-05	Proteasome 26S Subunit, ATPase 2
PSMD11	0.015765148	Proteasome 26S Subunit, Non-ATPase 11
RNF20	0.022853329	Ring Finger Protein 20
SIRT5	0.02441427	Sirtuin 5
HDAC3	0.030265535	Histone Deacetylase 3
SRP68	0.034608881	Signal Recognition Particle 68
ATP2B1	0.053680135	ATPase Plasma Membrane Ca2+ Transporting 1
FAM69A	0.061151781	Family With Sequence Similarity 69 Member A
INO80D	0.10489345	INO80 Complex Subunit D
JAKMIP2	0.109585837	Janus Kinase And Microtubule Interacting Protein 2
NFYC	0.112753497	Nuclear Transcription Factor Y Subunit Gamma
PIWIL2	0.153139214	Piwi Like RNA-Mediated Gene Silencing 2
G2E3	0.166310692	G2/M-Phase Specific E3 Ubiquitin Protein Ligase
MDH2	0.202206768	Malate Dehydrogenase 2
IFT57	0.246847666	Intraflagellar Transport 57
EIF2B2	0.347871114	Eukaryotic Translation Initiation Factor 2B Subunit Beta
ROBO4	0.361089875	Roundabout Guidance Receptor 4
MYEF2	0.429368685	Myelin Expression Factor 2
ARCN1	0.48861277	Archain 1
STRADA	0.577033763	STE20-Related Kinase Adaptor Alpha
NR2C1	0.60880119	Nuclear Receptor Subfamily 2 Group C Member 1

TRMT10B	0.710725113	TRNA Methyltransferase 10B
PSMF1	0.798915758	Proteasome Inhibitor Subunit 1
EZH1	1.014036592	Enhancer Of Zeste 1 Polycomb Repressive Complex 2 Subunit
STAR	1.190417678	Steroidogenic Acute Regulatory Protein
METAP2	1.252059174	Methionyl Aminopeptidase 2
ADD3	1.285354177	Adducin 3
MCMDC2	2.677265469	Minichromosome Maintenance Domain Containing 2

Arboreality (arboreal, 1/ or not, 0)

Gene	Coefficient	Full Name
CNOT8	-0.149061035	CCR4-NOT Transcription Complex Subunit 8
DFNB59	-0.103201704	Deafness, Autosomal Recessive 59
KIF2A	-0.050959492	Kinesin Family Member 2A
ANAPC16	-0.023464429	Anaphase Promoting Complex Subunit 16
MTX3	-0.01706862	Metaxin 3
PSMD14	-0.010852085	Proteasome 26S Subunit, Non-ATPase 14
YWHAB	-0.003547718	Tyrosine 3-Monooxygenase/Tryptophan 5-Monooxygenase Activation Protein Beta
LHX6	-0.003008224	LIM Homeobox 6
ZDHHC17	-0.000729342	Zinc Finger DHHC-Type Containing 17
SAR1B	0.000791309	Secretion Associated Ras Related GTPase 1B
RNF145	0.000966699	Ring Finger Protein 145
RPRD1B	0.001895077	Regulation Of Nuclear Pre-MRNA Domain Containing 1B
CTCF	0.005109566	CCCTC-Binding Factor
TMEM50B	0.029289674	Transmembrane Protein 50B
SCARB1	0.033237107	Scavenger Receptor Class B Member 1
SEC11A	0.037630871	SEC11 Homolog A, Signal Peptidase Complex Subunit
TPST1	0.044930228	Tyrosylprotein Sulfotransferase 1
MPPE1	0.057914544	Metallophosphoesterase 1
PRKRA	0.059909829	Protein Activator Of Interferon Induced Protein Kinase EIF2AK2
ERCC8	0.0600502	Excision Repair Cross-Complementation Group 8
HOMER1	0.104147313	Homer Scaffolding Protein 1
ASB7	0.118854738	Ankyrin Repeat And SOCS Box Containing 7
GABPA	0.126121311	GA Binding Protein Transcription Factor Alpha Subunit
ARCN1	0.142768672	ARCN1 Gene
CNOT6L	0.15186007	CCR4-NOT Transcription Complex Subunit 6 Like
CNR1	0.162701677	Cannabinoid Receptor 1 (Brain)
ARIH2	0.171275225	Ariadne RBR E3 Ubiquitin Protein Ligase 2
KIF3B	0.205585681	Kinesin Family Member 3B
WNT8A	0.216923258	Wnt Family Member 8A
CYLD	0.352659793	CYLD Lysine 63 Deubiquitinase

Mating system (monogamous, 1/ ploygamous, 0)

Gene	Coefficient	Full Name
DPF2	-0.005692731	Double PHD Fingers 2
INTS12	0.002249234	Integrator Complex Subunit 12

CTCF	0.024983332	CCCTC-Binding Factor
PSMA4	0.029061235	Proteasome Subunit Alpha 4
CNR1	0.042869811	Cannabinoid Receptor 1 (Brain)
CCNG2	0.143776502	Cyclin G2
MTF1	0.179592075	Metal-responsive transcription factor 1
WNT2	0.190613183	Wnt Family Member 2
EZH1	0.219868841	Enhancer Of Zeste 1 Polycomb Repressive Complex 2 Subunit
SPERT	0.304757977	Spermatid-associated protein
WASL	0.42024451	Wiskott-Aldrich Syndrome Like
BUD13	0.446819432	BUD13 Homolog
MFAP3	0.457595364	Microfibrillar Associated Protein 3
MTSS1	0.485151937	Metastasis Suppressor 1
MRS2	0.57839873	MRS2, Magnesium Transporter
GRAMD3	0.638059543	GRAM Domain Containing 3
THG1L	0.795916023	TRNA-Histidine Guanylyltransferase 1 Like

Male Larger (male larger, 1/ or not, 0)

Gene	Coefficient	Full Name
CLK1	-0.354001413	CDC Like Kinase 1
TTK	-0.256834203	TTK Protein Kinase
ANO5	-0.24494582	Anoctamin 5
ANGEL2	-0.242372641	angel homolog 2 (Drosophila)
KBTBD8	-0.21775626	kelch repeat and BTB domain containing 8
AKTIP	-0.200297976	AKT Interacting Protein
ELOVL2	-0.189027697	ELOVL Fatty Acid Elongase 2
ADD3	-0.180510838	Adducin 3
HSF2	-0.125215156	heat shock transcription factor 2
CBL	-0.11045345	Cbl Proto-Oncogene
CPSF7	-0.094813372	Cleavage And Polyadenylation Specific Factor 7
RAB11FIP2	-0.057221091	RAB11 family interacting protein 2
GABPB1	-0.013120804	GA Binding Protein Transcription Factor Beta Subunit 1
CYLD	-0.012145546	Cylindromatosis
EFTUD2	-0.005496689	Elongation Factor Tu GTP Binding Domain Containing 2
LPIN3	0.106204973	Lipin 3
STAR	0.308253129	Steroidogenic Acute Regulatory Protein

Carnivory (Carnivore, 1/ or not, 0)

Gene	Coefficient	Full Name
ENPP2	-0.043615686	Ectonucleotide Pyrophosphatase/Phosphodiesterase 2
PROSC	-0.02663645	Proline Synthetase Co-Transcribed Homolog (Bacterial)
CTCF	-0.010073108	CCCTC-Binding Factor
CTDSPL2	-0.002734977	CTD Small Phosphatase Like 2
PHGDH	0.00736625	Phosphoglycerate Dehydrogenase
SCARB1	0.009095294	Scavenger Receptor Class B Member 1
NCKAP1	0.012867194	NCK Associated Protein 1
PRKAB1	0.017240688	Protein Kinase AMP-Activated Non-Catalytic Subunit Beta 1

DPCD	0.023544021	Deleted In Primary Ciliary Dyskinesia Homolog (Mouse)
ABL1	0.054652386	ABL Proto-Oncogene 1, Non-Receptor Tyrosine Kinase
AMMECR1L	0.108158909	AMMECR1 Like
FYCO1	0.108685914	FYVE And Coiled-Coil Domain Containing 1
LHX4	0.115787512	LIM Homeobox 4
DENND4B	0.117926833	DENN Domain Containing 4B
TIPRL	0.149362546	TOR Signaling Pathway Regulator
GEM	0.155488364	GTP-binding protein GEM
FGB	0.180305565	Fibrinogen Beta Chain
SNX10	0.21428252	sorting nexin 10
TTLL1	0.248551247	tubulin tyrosine ligase-like family, member 1
TGM1	0.281607847	transglutaminase 1 (K polypeptide epidermal type I, protein-glutamine-gamma-glutamyltransferase)
PAPD7	0.296113223	PAP Associated Domain Containing 7
LMCD1	0.328985887	LIM And Cysteine Rich Domains 1

Herbivory (Herbiovre, 1/ or not, 0)

Gene	Coefficient	Full Name
TTLL1	-0.192278776	tubulin tyrosine ligase-like family, member 1
CPSF3	-0.170472802	Cleavage And Polyadenylation Specific Factor 3
HOXD10	-0.153316052	Homeobox D10
HS3ST5	-0.109175618	Heparan Sulfate-Glucosamine 3-Sulfotransferase 5
ARIH2	-0.079102353	Ariadne RBR E3 Ubiquitin Protein Ligase 2
NR1H4	-0.058818363	Nuclear Receptor Subfamily 1 Group H Member 4
NAA15	-0.046580546	N(Alpha)-Acetyltransferase 15, NatA Auxiliary Subunit
ASB5	-0.044156172	Ankyrin Repeat And SOCS Box Containing 5
PHGDH	-0.03533651	Phosphoglycerate Dehydrogenase
G3BP2	-0.029008839	G3BP Stress Granule Assembly Factor 2
EFTUD2	-0.017873457	Elongation Factor Tu GTP Binding Domain Containing 2
PSMA4	-0.016326376	Proteasome Subunit Alpha 4
CSTF1	-0.012993188	Cleavage Stimulation Factor Subunit 1
NCKAP1	-0.008299235	NCK Associated Protein 1
HDAC3	0.00349265	Histone Deacetylase 3
TMEM98	0.006173311	Transmembrane Protein 98
CSNK1G1	0.007468903	Casein Kinase 1 Gamma 1
GRIA2	0.007799025	Glutamate Ionotropic Receptor AMPA Type Subunit 2
PSMD14	0.010333524	Proteasome 26S Subunit, Non-ATPase 14
FYN	0.027876627	FYN Proto-Oncogene, Src Family Tyrosine Kinase
MEF2C	0.030092811	Myocyte Enhancer Factor 2C
AP3M2	0.030186499	Adaptor Related Protein Complex 3 Mu 2 Subunit
DHRS3	0.043281516	Dehydrogenase/Reductase 3
SNF8	0.046587869	SNF8, ESCRT-II Complex Subunit
HNRNPH3	0.04818168	Heterogeneous Nuclear Ribonucleoprotein H3
AP1G1	0.05554078	Adaptor Related Protein Complex 1 Gamma 1 Subunit
NMT1	0.063060716	N-Myristoyltransferase 1
IFT20	0.068654333	Intraflagellar Transport 20
ALDH1A2	0.074303585	Aldehyde Dehydrogenase 1 Family Member A2

HOOK1	0.079183092	Hook Microtubule Tethering Protein 1
ZDHHC5	0.082217822	Zinc Finger DHHC-Type Containing 5
KIF3B	0.093681722	Kinesin Family Member 3B
IFT57	0.123413202	Intraflagellar Transport 57
VPS39	0.134109858	VPS39, HOPS Complex Subunit
PARP11	0.152111102	Poly(ADP-Ribose) Polymerase Family Member 11
KCNJ1	0.262550187	Potassium Voltage-Gated Channel Subfamily J Member 16
ELOVL7	0.379067567	ELOVL Fatty Acid Elongase 7

Omnivory (omnivore, 1/ or not, 0)

Gene	Coefficient	Full Name
PARP11	-0.309600906	Poly(ADP-Ribose) Polymerase Family Member 11
CCDC42	-0.241173075	Coiled-Coil Domain Containing 42
ABCD3	-0.080215854	ATP Binding Cassette Subfamily D Member 3
PLEK2	-0.065798562	Pleckstrin 2
ABLIM1	-0.048443606	Actin Binding LIM Protein 1
EIF2S1	-0.047395885	Eukaryotic Translation Initiation Factor 2 Subunit Alpha
DPCD	-0.033461558	Deleted In Primary Ciliary Dyskinesia Homolog (Mouse)
HNRNPH3	-0.028857596	Heterogeneous Nuclear Ribonucleoprotein H3
RPRD1B	-0.020546706	Regulation Of Nuclear Pre-MRNA Domain Containing 1B
AHCYL1	-0.011009125	Adenosylhomocysteinase Like 1
INTS9	-0.01072498	Integrator Complex Subunit 9
ACTL6B	-0.000497905	Actin Like 6B
CNOT6L	0.006381619	CCR4-NOT Transcription Complex Subunit 6 Like
STK38	0.009401955	Serine/Threonine Kinase 38
TPST1	0.017944984	Tyrosylprotein Sulfotransferase 1
PUS7	0.058418391	Pseudouridylate Synthase 7 (Putative)
PSMA4	0.092991286	Proteasome Subunit Alpha 4
FAF2	0.126585083	Fas Associated Factor Family Member 2
ATP6V1A	0.236637588	ATPase H+ Transporting V1 Subunit A
HS3ST5	0.271474958	Heparan Sulfate-Glucosamine 3-Sulfotransferase 5
SLC7A14	0.305132884	Solute Carrier Family 7 Member 14
ELMOD2	0.311750049	ELMO Domain Containing 2
ARIH2	0.356122287	Ariadne RBR E3 Ubiquitin Protein Ligase 2
SHOC2	0.365833305	soc-2 suppressor of clear homolog (C. elegans)

Insectivory (Insectivore, 1/ or not, 0)

Gene	Coefficient	Full Name
IFT46	-0.133232956	Intraflagellar Transport 46
HADH	-0.062972335	Hydroxyacyl-CoA Dehydrogenase
JAKMIP2	-0.058739664	Janus Kinase And Microtubule Interacting Protein 2
ACAD8	-0.042125576	Acyl-CoA Dehydrogenase Family Member 8
TSSK3	-0.032605841	Testis Specific Serine Kinase 3
KCTD6	-0.013615494	Potassium Channel Tetramerization Domain Containing 6
CNR1	0.096987526	Cannabinoid Receptor 1 (Brain)
KCNE4	0.115693818	Potassium Voltage-Gated Channel Subfamily E Regulatory Subunit 4

CPSF3	0.124108269	Cleavage And Polyadenylation Specific Factor 3
CETN3	0.138276816	Centrin 3
SHOC2	0.15591852	soc-2 suppressor of clear homolog (C. elegans)
TSG101	0.193432651	Tumor Susceptibility 101
CNOT6L	0.243456292	CCR4-NOT Transcription Complex Subunit 6 Like
LMCD1	0.280174703	LIM And Cysteine Rich Domains 1
SEC11A	0.313293348	SEC11 Homolog A, Signal Peptidase Complex Subunit
ANKRD50	0.360251361	Ankyrin Repeat Domain 50
PHTF1	0.690324558	putative homeodomain transcription factor 1
DICER1	0.8033749	Dmx-like 1
HSPH1	0.828396317	Heat Shock Protein Family H (Hsp110) Member 1
RAB11FIP2	1.18707395	RAB11 family interacting protein 2
ANO5	1.218667544	Anoctamin 5

* The genomic locations and function of each gene, please see Table S4 in [156].

** In Table S4 in [156], information on the genomic location of each gene is based on the human genome. Explanations of gene functions were collected mainly from RefSeq (http://www.ncbi.nlm.nih.gov/refseq/), GeneCard (http://www.genecards.org) and proceedings papers.

Chapter 5. Adaptation and long distance gene flow among South India, South China, and Japan of *Spodoptera litura* (Lepidoptera: Noctuidae)

5.1 Summary

Spodoptera litura (Lepidoptera: Noctuidae), normally called tobacco cutworm or cotton leafworm, is a notorious agriculture pest in Asia. Here, I obtained the SNPs of 56 individuals of *S. litura* sampled from 12 locations ranging from South India, South China to Japan, to study its population structure, migration and genomic adaptations of this moth. For Guangzhou and Hunan, I have more than one sample, thus I have 15 samples in total, and averagely 4 individuals per sample (some sample have less individuals). After screening, 46,595,432 SNPs without missing sites were used in this analysis. The result by structure and Fst clustering showed that samples from remote locations, Fujian, Zhejiang, Okinawa, Tsukuba, Hyderabad, Matsyapuri clustered together, while some samples at single locations, for example, two samples from Guangzhou and three samples from Hunan are quite different. This indicates a complex population structure and migration pattern of this moth. Based on the result by structure analysis and Fst clustering, I divided 56 individuals into 3 populations: isolated population, local population and migrating population, and conducted the analysis based on joint allele frequency spectrum of those populations using δaδi package. I found a high gene flow among India migrating, China local, China migrating and Japan migrating populations. The migration direction indicated by this analysis is consistent with the summer monsoon in Southeast Asia. I also found higher gene flow between local populations and migration populations than the gene flow among India migrating, China migrating and Japan migrating populations, indicate local population as an important source of the migration populations.

5.2 Introduction

Spodoptera litura (Lepidoptera: Noctuidae) distributes mainly in tropical and subtropical in Asia, is a destructive agriculture pest, feed on more than 100 crops and cause extensive damage each year [186]. *S. litura* can fly for more than 18 hour in laboratory conditions, and female moths are not trapped by the "oogenesis-flight syndrome" [187]. Female with ovarian development are also caught in July [188], indicating both males and females possess the ability of long-distance migration[189]. Its occurrence in several locations is coincident with the "weather-forecasting" in Asia [188], also a previous study report that massive and abrupt occurrences of *S. litura* are associated with the paths of typhoons in Japan [187]. *S. litura* may take a "weather-forecasting ships" for migration [188]. A phylogenetic analysis nuclear ITS2 of 158 larvae of *S. litura* from 11 locations in China and Korea implied a high gene flow among those sampling locations [190]. These evidences indicate *S. liutra* may use wind for long-distance migration.

Mita and his colleagues sequenced the complete genome of 56 individuals of *S. litura* from 12 locations (15 samples) ranging from South India, South China to Japan and I analyzed this data [191]. My results indicate that *S. litura* can be divided into isolated, local and migrating populations. The difference between isolated population and other two populations are sup-population level. I find high gene flow among South India, South China, and Japan, and *S. litura*'s migration direction is consistent with the direction of summer monsoon in Southeast Asia. As the first genome-wide population genetics study on a destructive agriculture pest moth, I hope my result will help understanding the nature and expansion of this moth in Asia.

5.3 Materials and Methods

5.3.1 Sampling and Sequencing

S. litura was sampled from three samples of India (Delhi, Hyderabad, and Matsyapuri), 11 samples of China, including Fujian, Guanxhi, 2 samples from Guangzhou (Guanzhou1 and Guangzhou2), Hainan, Hubei, Shanxi, Zhejiang, 3 samples from Hunan (Hunan1, Hunan2 and Hunan3), and 2 samples of Japan (Tsukuba and Okinawa). Four individuals were sampled from each location, except for Hunan1 (3 individuals) and Fujian (1 individual). In total, 56 individuals were used in this study.

5.3.2 Mapping and SNP calling

At first, mapping of reads of each individual to the reference genome was conducted. Proper mapping rate was about 70% for 56 individuals (Table 5.1). SNP calling was conducted by comparing 56 genomes together with reference genome. Finally a multiple VCF file including 56 individuals was generated. Sites with missing values or have Quality values below 20 were screened by VCFtools software [192]. In total 46,595,432 SNPs were identified and included in this analysis.

5.3.3 Genetic diversity

The nucleotide diversity, π, of 15 samples, pairwise Fst, and Tajima's D [193] were calculated by using VCFtools software with window size 5,000 bp, step 2,500 bp. The genomic nucleotide diversity was obtained by averaging over the values of windows. The weighted Fst was calculated using the Weir and Cockerham estimator [194]. Based on the pairwise Fst, hierarchical cluster analysis (complete) was conducted by using R software. Because of small sample size in each sample, interpretation of population genomic analysis needs careful evaluation of the precision. The precision of π and Fst values were evaluated by parametric bootstrap with coalescent simulation[195]. Haplotypes of windows were generated using the population-specific π values multiplied by 5,000 and 4Nm's calculated as $1/F_{ST}$-1. Two haplotypes were generated for each window. One thousand sets of haplotypes were generated independently and concatenated to make a bootstrap sample. For each of 100 bootstrap samples, the π values and pairwise Fst were calculated to estimate the

standard errors. The adopted the number of sets was less than the number of the scaffolds. Because the genome size of *S. litura* was about 4×10^8 bp, I mimicked the subsampling of windows that were separated 4×10^5 bp apart on average so that I could estimate approximate independence between the sub-sampled windows.

To identify the genes under balancing selection, Tajima's D for 5,000 bp-windows was calculated. Under the null hypothesis of neutral mutations in a population in equilibrium, the distribution has the mean 0 and the variance 1. However, a value of Tajima's D is affected not only by the selection but also by the population history. Since the latter effect applies all sites in the genome, the *p*-values of the windows were calculated based on the reference normal distribution with the estimated mean and variance. Benjamini-Hochberg procedure [196] was applied to the *p*-values to select the windows by controlling the false discovery rate (FDR).

To confirm the observed population structure, I conducted the model-based structure analysis [197, 198]. Based on the allele frequency divergence among the ancestral populations, **P**, and the membership coefficients that assign the populations to the ancestral populations, **Q**, I calculated the predicted allele frequency divergence between the population $\mathbf{PQP}^t$. I also analysed individual-level membership coefficients and the allele frequency divergence.

5.3.4 Joint allele frequency spectrums and migration pattern

I further estimated the global pattern of migration by analysing the joint allele frequency spectrums in terms of the population histories and the migration patterns by ∂a∂i (diffusion approximation for demographic inference) [199]. To avoid the complex effect of selection, I analysed SNPs in introns. Out of 20M SNPs in introns, I randomly sampled 2M SNPs. Based on the multi-dimensional scaling of Fst and the assignment of the individual genomes by structure, I constructed six population groups: Indian local population (with the sample from Delhi), Indian migratory population (with the samples from Hyderabad and Matsyapuri), Chinese isolated population (with the samples from Guangzhou2 and Hunan1), Chinese local population (with the samples from Hunan3, Guangxi, Hainan, three individuals of Hunan2 and Hainan), Chinese migratory population (with the samples from Fujian, one individual of Hunan2, Hunan3, Hunan4, Zhejiang and Guangzhou1), and

Japanese migrating population (with the samples from Okinawa and Tsukuba). To each pair of population groups, I applied the IM (isolation with migration) model [200] with population expansion/shrinkage. The estimated migration rates represent number of migrating chromosomes per generation. To obtain the population sizes and the time of population splitting from the estimated relative values, I followed Zhan *et al*., 2004 [201], which assumed the generation time of 0.3 year and used the standard mutation rate of 8.4×10^{-9} (per site per generation), from *Drosophila* [202]. The standard errors were obtained by parametric bootstrap of coalescent simulation [195]. Assuming the estimated scenarios of population history, I generated 100 bootstrap samples of 2M SNPs. To reflect the correlation structure between SNP loci, I assumed that they are evenly distributed on 28 chromosomes. SNPs on different chromosomes are independent. Noting that the mean distance between the neighbouring SNP loci was $\frac{4.6\times10^8}{2.0\times10^6} = 2.3\times10^2$ bp, I set the recombination rate to be $\rho = 2.3\times10^{-5}$. I also tested two alternative values, $\rho = 0$ and $\rho = 0.01$, and obtained similar standard errors.

5.4 Results

5.4.1 Genetic diversity

A large number of SNPs enabled precise estimation of the genomic nucleotide diversity π for each surveyed local population (Table 5.2, Figure 5.1). It ranged between 0.013 and 0.016 at most local populations. The diversity was a little lower at Delhi in northern India (0.0086±0.0001), and an order of magnitude lower at two sampling locations in Hunan province in China: Hunan1 (0.0018±0.0000) and NSU (0.0019±0.0000).

5.4.2 Signal of population expansion

The distribution of the 5,000 bp-window Tajima's D had the mean value of -0.8345. Negative value of Tajima's D generally implies either the sign of population expansion or the sign of purifying selection. Out of the 86,835 windows, 84,538 (97.4%) had negative values of Tajima's D (Figure 5.2). The whole coverage of negative Tajima's D value over the genome strongly suggests the history of population expansion. When I confine the target of the analysis to the windows on the CDS, 46,272 out of 76,548 windows had the value of Tajima's D being 0. The distribution among the other 30,276 windows was a little wider, but similar to the distribution among the whole windows. Forty-two windows on the CDS had high value of Tajima's D (>2) and were suspected to be under the balancing selection (Table 5.3). Out of Forty-two windows, blast search found homologs for fifteen sequences. They were aminopeptidase, neuropeptide receptor, non-LTR retrotransposon Jockey-like reverse transcriptase gene, E3 ubiquitin-protein ligase, ATP-dependent RNA helicase, and cytochrome P450 genes, and possibly related with insecticide resistance.

5.4.3 High east-west gene flow ranging across South India, South China, and Japan

Figures 5.1 show the geographical map of the genetic diversities of the surveyed local populations and the extent of gene flow between them. An east-west high gene flow was observed ranging between South India, South China and Japan. Notably, the pairwise Fst values were below 0.01 between Hyderabad in India, Fujian in China,

Okinawa and Tsukuba in Japan (Figure 5.1): F_{ST}(Hyderabad, Okinawa) = 0.002 ± 0.004), F_{ST}(Okinawa, Tsukuba) = 0.005 ± 0.006, F_{ST}(Hyderabad, Fujian) = 0.006 ± 0.011, F_{ST}(Fujian, Tsukuba)=0.008±0.011, F_{ST}(Fujian,Okinawa)=0.008±0.011 and F_{ST}(Hyderabad,Tsukuba)=0.008±0.005. The three local populations, Delhi, Hunan1, and Guangzhou2, which had lower genetic diversity, had little genetic exchange among each other and with the other local populations. The gene flow among the other local populations in China was mild: $0.1<F_{ST}<0.4$.

Wan *et al*. [190] investigated the DNA sequence variation of *S. litura* in 11 local populations in China and Korea, and observed a high gene flow of nuclear ITS2 sequences between the two countries. Consistent with the current results, the geographical map (Figure 5.3) also implies that the east-west gene flow is even stronger than north-south gene flow within China. Tojo *et al.*[203] studied overseas migration of *S. litura* from May to mid-July in four regions in East Asia: western Japan, South Korea, China, and Taiwan. Using the world wind record and a simulation model, they estimated the trajectory of insects that resulted in the abrupt increase of catch in western Japan, and found that catch increases coincide with the arrival of southwesterly air currents from southern China and/or Taiwan. My results (Fig. 4a,b) imply an even longer distance trip from southern India to China and Japan [190].

The model-based structure analysis [198] provided the predicted population structure which is consistent with the F_{ST}-based cluster analysis (Figure 5.4). By incorporating the estimated allele frequency divergence between the ancestral populations, I obtained a very stable picture of population structure against the assumed number of the ancestral populations, *K*. Here again, extremely high gene flow between central Indea (Hyderabad and Matsyapuri), the southeast coast of mainland China (Zhejiang, Guangzhou and Fujian), and Japan (Okinawa and Tsukuba) were observed. The assignments of individual genomes to the ancestral populations a detailed picture on the gene flows (Figure 5.5, 5.6). These results are consistent with the study of DNA sequence variation of *S. litura* among populations of *S. litura* in China and Korea [190]. In addition, oversea migration from southern China to western Japan driven by typhoons was reported [203, 204]. Geographical data on Asian monsoon in July-

August [205] may support my result, an even longer distance trip from southern India to China and Japan.

To further understand the global pattern of migration route, I analyzed the joint allele frequency spectrums (Figure 5.7) by ∂a∂i (diffusion approximation for demographic inference) [199]. Based on the F_{ST}-based population structure and the model-based assignment of the individual genomes, I constructed six population groups: two population froups in India (India_local and India_migrate), three population groups in China (China_isolate, China_local, and China_migrate), and Japan. By applying the isolation with migration model [200] to each of the pairs of population groups, I identified a global route from Indian migrating population through Chinese local population, which ranges from south at Hainan to north at Hubei (Figure 5.8). This Chinese local population has a large number of migrants to and from the Chinese migrating population. Moderate numbers of migrants from China to Japan and from China to India were observed. Also, the local populations in India and China have been shrinking largely since 2000-3000 years ago. In contrast, Japanese population has been expanding for the last 5000 years (Table 5.4, Figure 5.9).

Figure 5.10a shows examples of joint allele frequency spectrums among different samples. Samples between local populations and migrating populations (e.g. Hunan3 and Okinawa) show higher gene flow compared with the samples within local populations (e.g. Hunan3 and Hubei). Together with the result shown in Figure 5.7 and Figure 5.8, we think that the migration pattern of S. litura may be a “star-like” migration, that from each local population, some migrate individuals generate and enter the migrating populations (Figure 5.10b). The allele frequency of migrating population is most close to the average allele frequencies of all local populations. This migrating pattern can also explain the result of Fst clustering and population structure analysis.

5.5 Discussion

My results firstly show the population structure and migration pattern of *S. litura*. That is, *S. litura* have three populations: isolated population, local population and migrating population. The difference between isolated populations to other populations is a sub-population level, with Fst larger than 0.5 isolated populations and local/ migrating populations (Table 5.3, Figure 5.4). Surprisingly, it was indicated that local populations were important intermediate populations to receive and generate the new migrating populations. The migration rate between migrating population and local population is stronger than that among migrating populations from India, China and Japan (Table 5.4, Figure 5.8). *S. litura*'s migration is possible continuously accompanied with the high gene flow with local populations.

The migration pattern inferred by joint allele frequency spectrums is consistent with the direction of summer monsoon in Asia. In my study, I have 10 samples from China, among which 5 of them are local populations while 5 are migrating populations. I only have 3 samples from India, with one sample belong to local populations. No samples from other South Asia countries were included in my study. I expect there are many local populations along its migration routes. Hopefully, this may be confirmed in the future. More detailed study of the differences between local populations and migrating populations may also be necessary.

5.6 Conclusion

I analyzed 46,595,432 SNPs of 56 individuals of *S. litura* sampled from 12 locations (in total 15 samples) ranging from India, China and Japan. The Fst-based clustering and structure analysis indicate that samples from remote locations, Fujian, Zhejiang, Okinawa, Tsukuba, Hyderabad, Matsyapuri clustered together, while some samples at single locations, for example, two samples from Guangzhou and three samples from Hunan are quite different. Based on this observation, I divided 56 individuals into 3 populations: isolated population, local population and migrating population, and conducted the analysis based on joint allele frequency spectrum of those populations using δaδi package. I found a high gene flow among India migrating, China local, China migrating and Japan migrating populations.

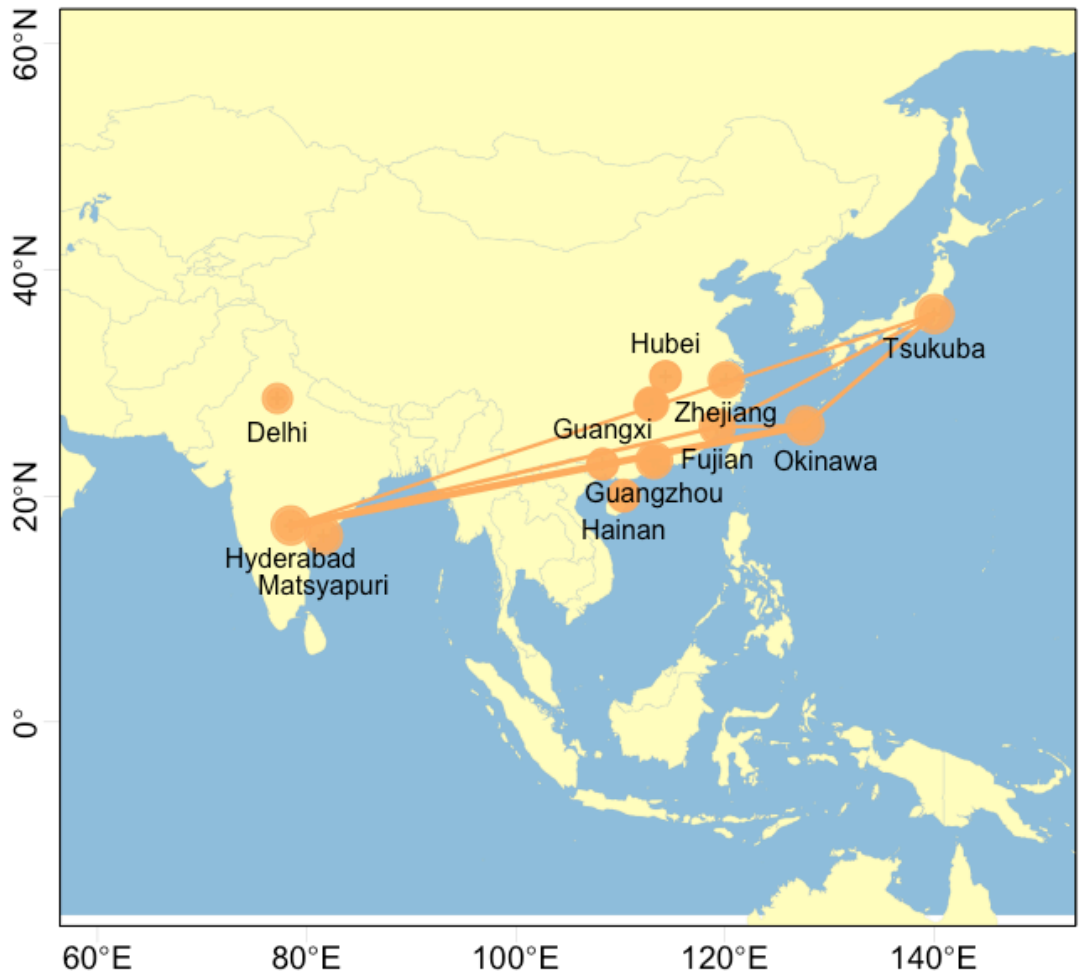


Figure 5.1 Geographical map of the genetic diversities of the surveyed local populations and the extent of gene flows in this study

The location pairs with high gene flow ($F_{ST} < 0.05$) are connected by segments. The sampling locations are as follows. Three locations of India (Delhi, Hyderabad, and Matsyapuri), 11 locations of China, including Fujian, Guanxhi, 2 locations from Guangzhou (Guanzhou and SCNU), Hainan, Hubei, Shanxi, Zhejiang, 3 locations from Hunan (Hunan1, Hunan2 and Hunan3), and 2 locations of Japan (Tsukuba and Okinawa)

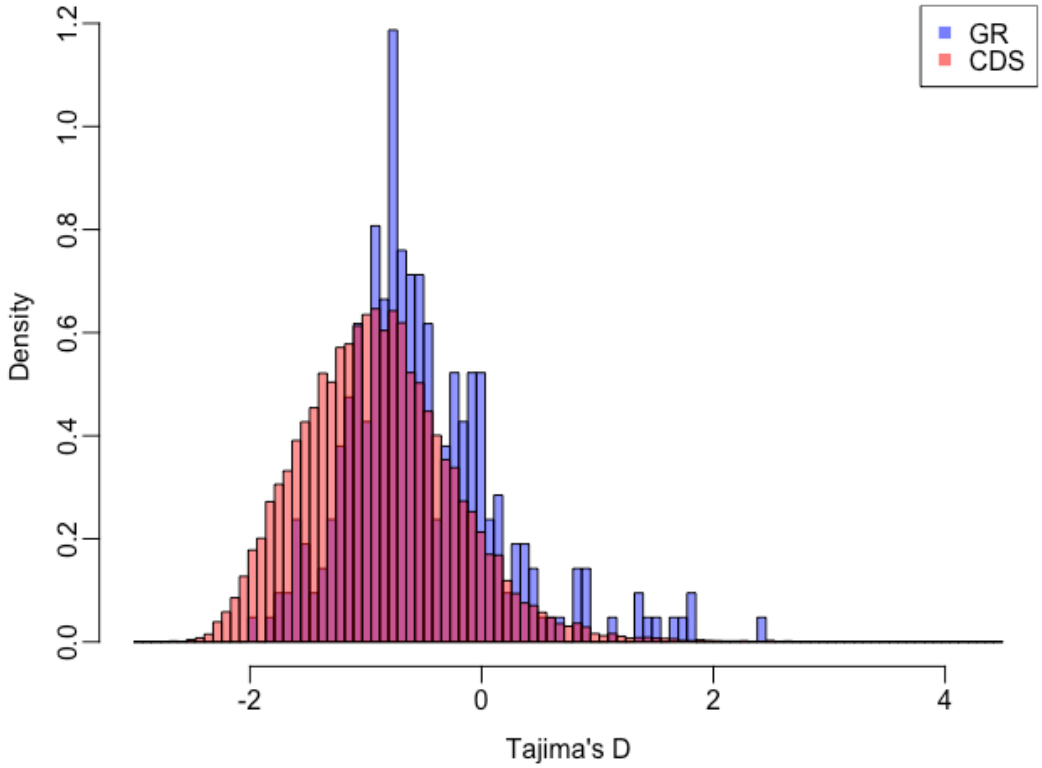


Figure 5.2 Gene flows of nuclear ITS2 sequences among local populations identified in Wan *et al.* [190]

The location pairs with high gene flow ($F_{ST} < 0.05$) are connected by segments. The sampling locations are as follows. In China (province): Nanning (Guangxi), Changsha (Hunan), Jianli (Hubei), Wuhan (Hubei), Nantong (Jiangsu). In Korea (province): Kangreung (Kangwon), Ansung (Gyeonggi), Cheongwon (Chungbuk), Milyang (Kyungnam), Noan (Chonnam), Jeju Island. CDS: coding region. GR: gustatory receptor.

Figure5.3 The distribution of Tajima's D of 5,000 bp windows

The histogram colored blue is the distribution among the windows intersecting with GRs (Gustatory Receptor), and the histogram colored red is the distribution among the windows intersecting with coding region (CDS).

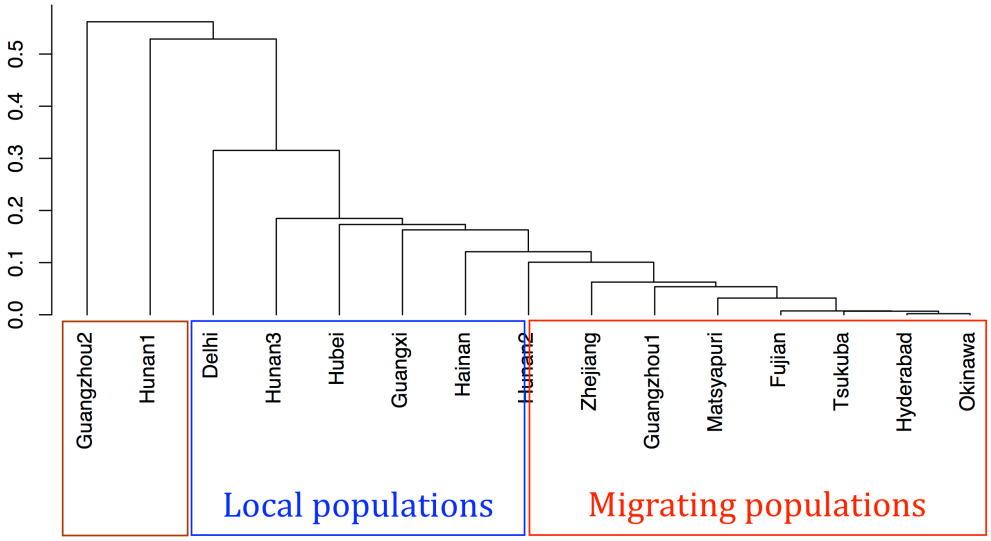


Figure 5.4 F_{ST}-based cluster analysis of local populations

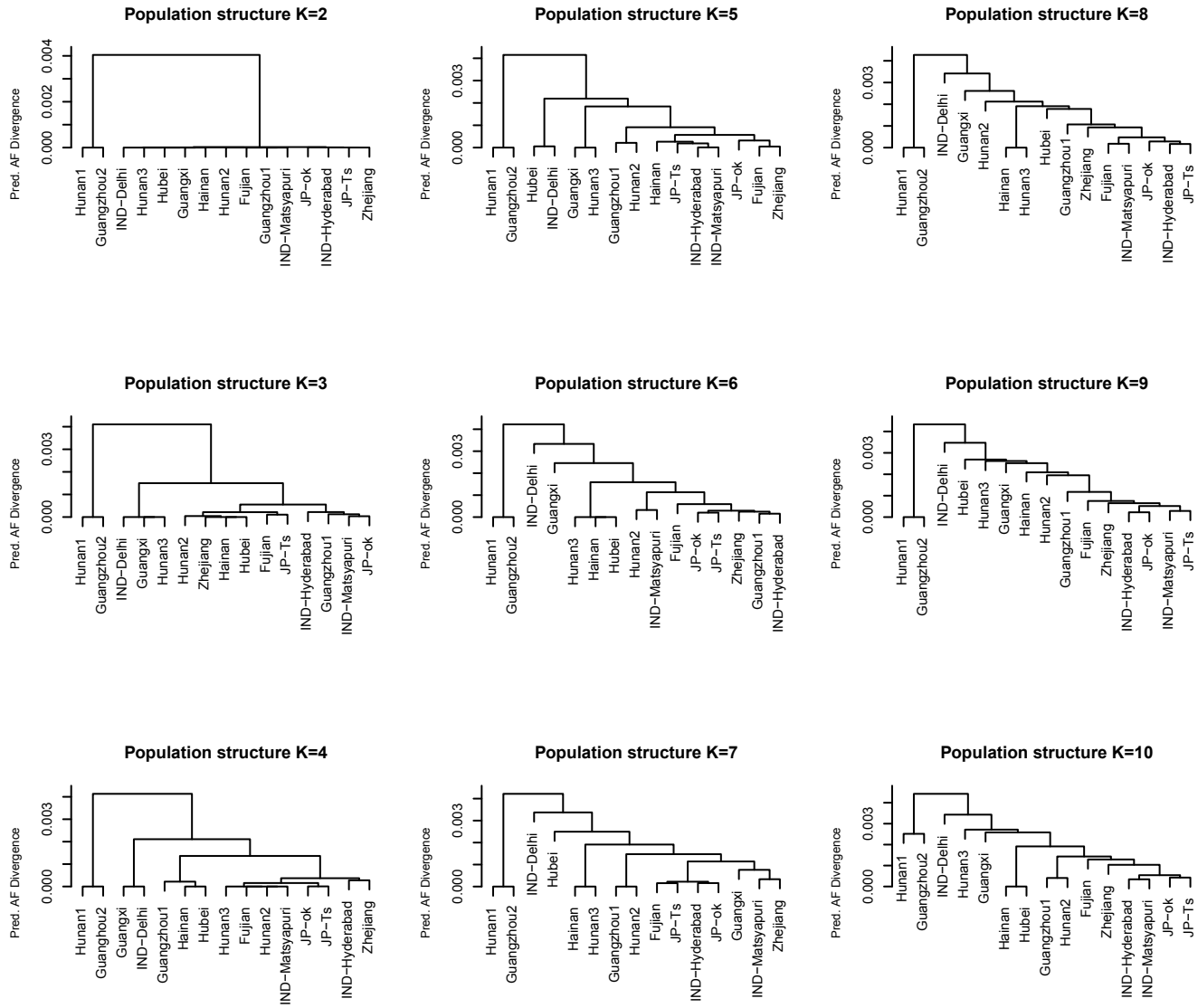


Figure 5.5 Structure analysis of population genomes: long distance gene flow and diversifying selection (K=2-10)

Cluster analysis of populations based on the predicted allele frequency divergence

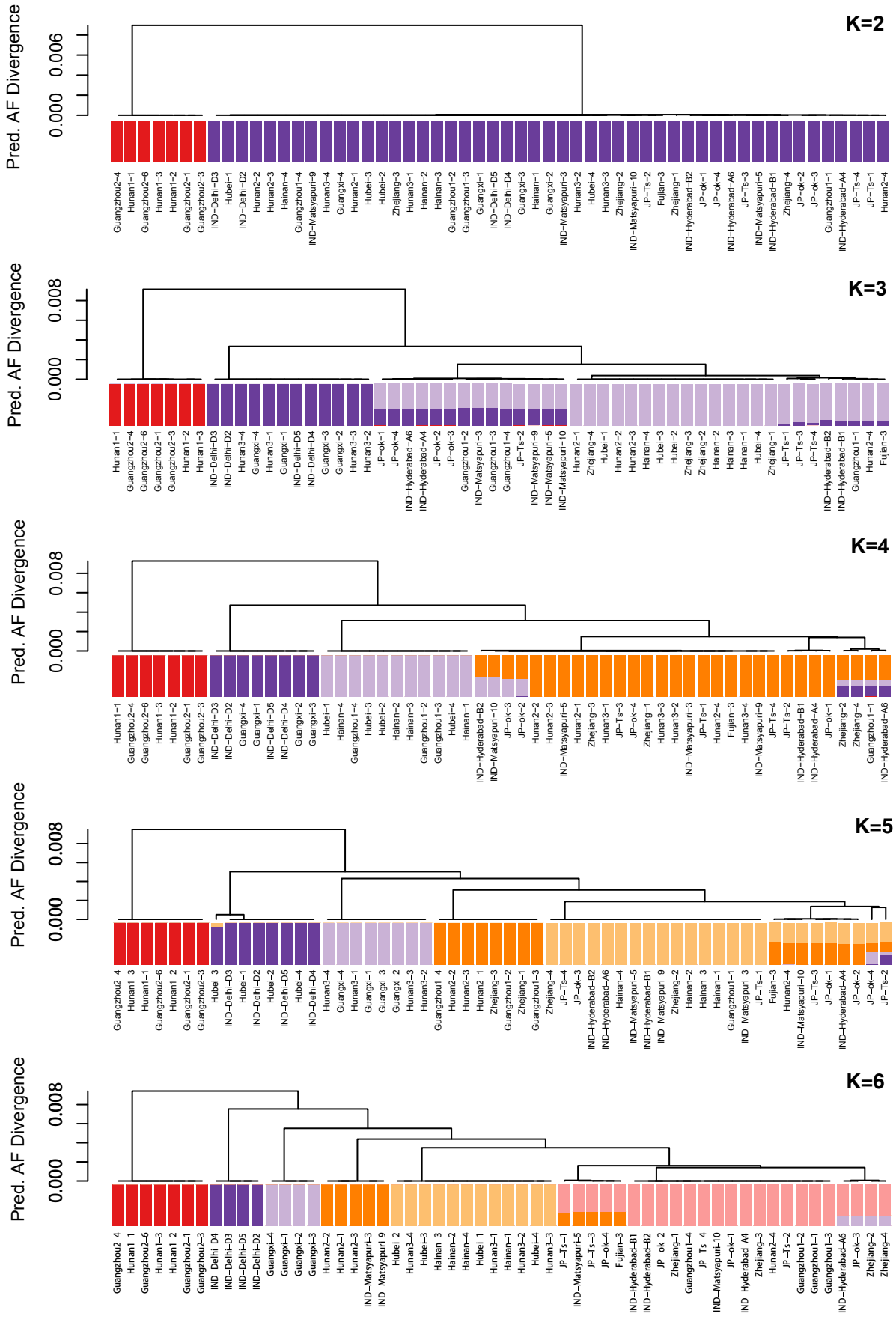


Figure 5.6a Structure analysis of individual genomes: longlistance gene flow and diversifying selection (K = 2-5)

Assignment of the individual genomes in the samples to the ancestral populations predicted by structure

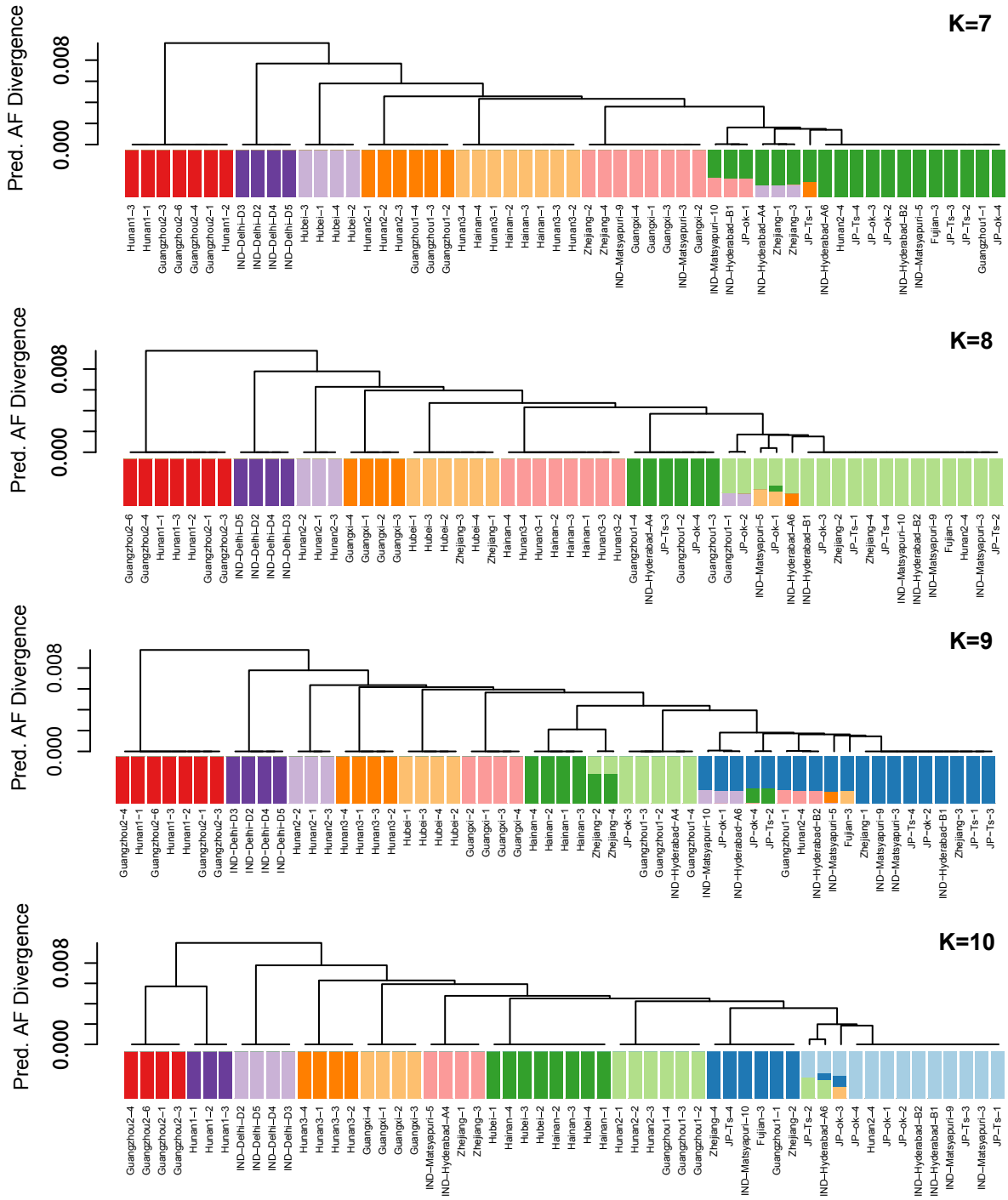


Figure 5.6b Structure analysis of individual genomes: long distance gene flow and diversifying selection (K = 6-10)

Assignment of the individual genomes in the samples to the ancestral populations predicted by structure

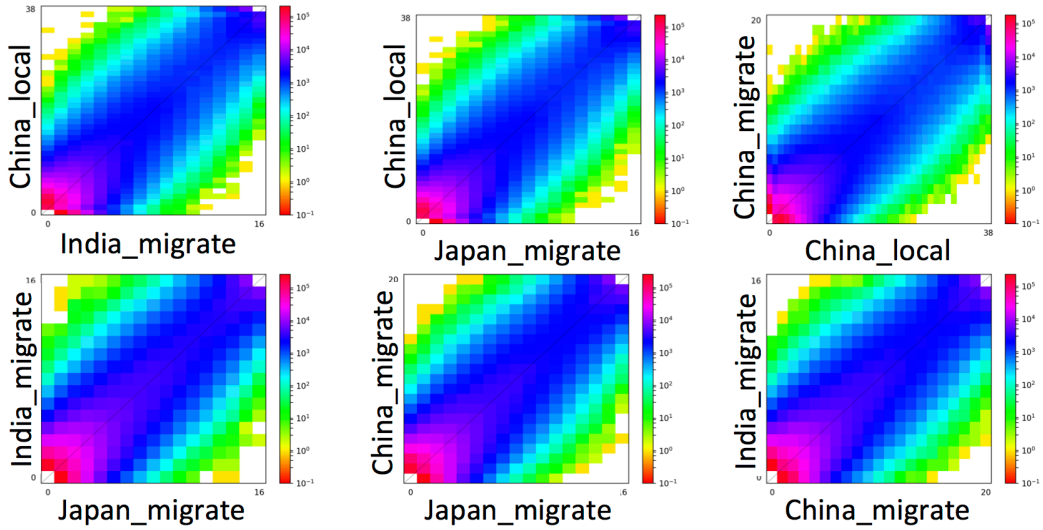


Figure 5.7 Two-dimensional allele frequency spectra in the paired population groups.

The width of bend of two dimensional allele frequency spectra indicates the similarity of the allele frequency between two populations, and narrower band indicates two populations have more similar allele frequencies. The two-dimensional allele frequency spectra of China_local to migrate population (India_migrate, China_migrate and Japan_migrate) show narrower band compared with the pairwise two-dimensional allele frequency spectra among migrate populations, indicate higher gene flow between local populations and migrate populations than the gene flow within migrate populations.

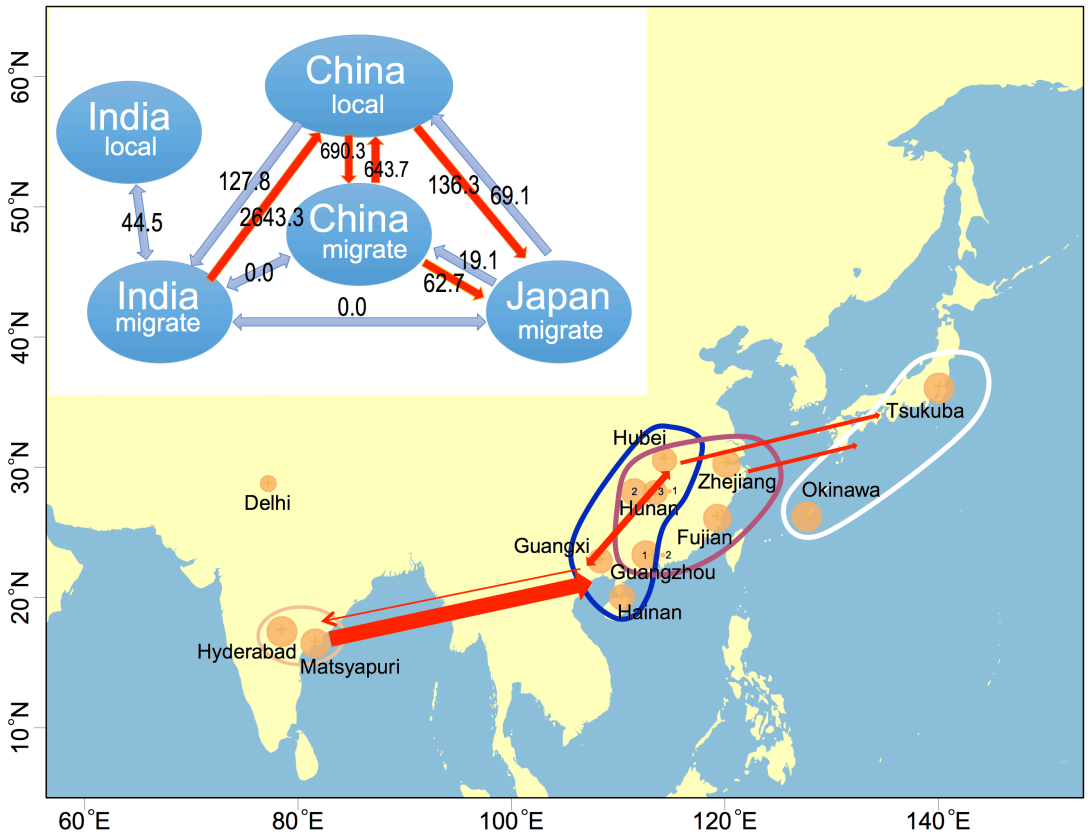


Figure 5.8 Global picture of the migration route predicted by ∂a∂i

The inset shows the number of migrating chromosomes per generation. The four closed ropes represent the migrating population in India, local populations in China, migrating populations in China, and the populations in Japan. The size of the circles represents the genetic diversity (π).

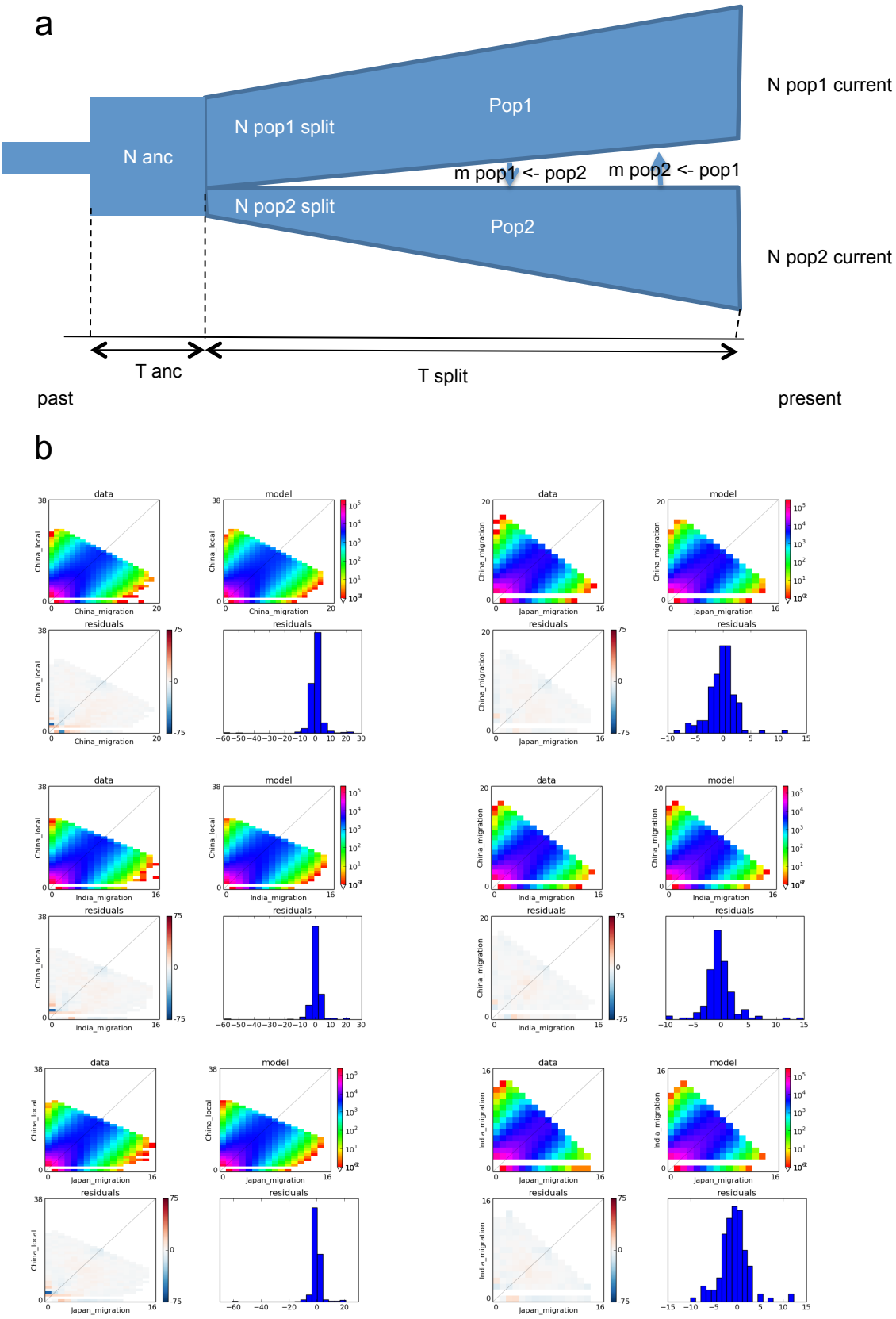


Figure 5.9 Analysis of the joint allele frequency spectrums by ∂a∂i.

a, Model of isolation wiht migration applied to pairs among the five population groups, India_local, India_migrate, China_local, China_migrate, and Japan. b, Observed and predicted allele frequency spectrums and residuals.

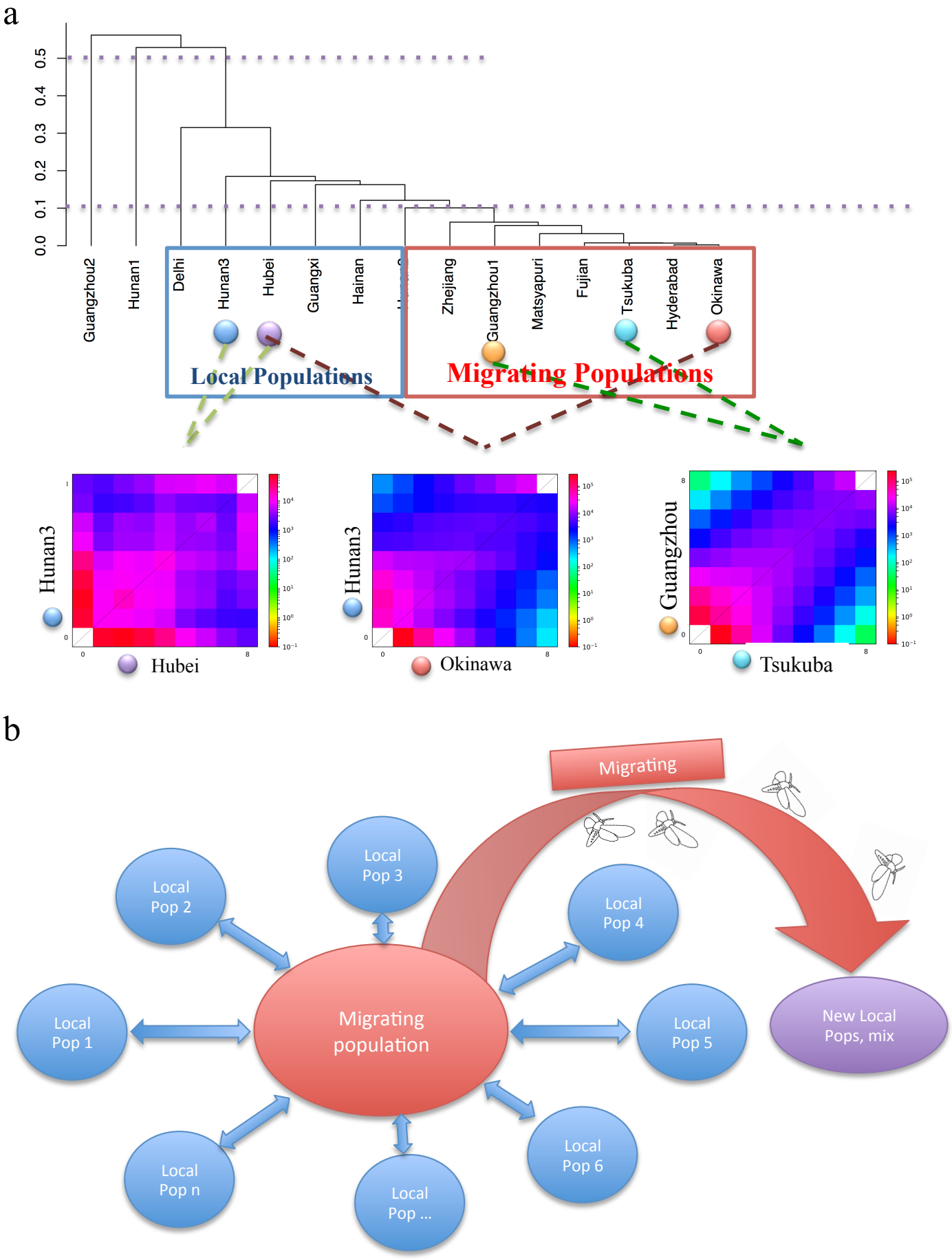


Figure 5.10 Two-dimensional allele frequency spectra in the paired samples, and a "star-like" migrating hypothesis

Table 5.1 SNP calling statistics of *S. litura* sample.

strains	raw reads (PE)	clean reads(PE)	dupplicated reads	mapped reads	mapped rated	proper mapped reads	proper mapped rate	snp number	sequencing depth	mapped depth
Fujian-3	49,948,303	44,107,681	0	35,436,430	0.80	32,327,026	0.73	8,124,993	21.00	15.39
Guangxi-1	35,750,378	32,036,579	0	25,723,709	0.80	23,132,531	0.72	7,234,759	15.26	11.02
Guangxi-2	48,723,692	42,947,177	0	33,926,158	0.79	30,922,626	0.72	8,124,265	20.45	14.73
Guangxi-3	62,134,773	47,002,718	0	36,975,608	0.79	33,685,761	0.72	8,140,494	22.38	16.04
Guangxi-4	55,131,064	42,079,721	0	32,586,493	0.77	28,870,445	0.69	7,778,236	20.04	13.75
Guangzhou-1	47,386,490	42,216,265	0	34,248,270	0.81	31,561,816	0.75	8,131,573	20.10	15.03
Guangzhou-2	47,381,820	40,157,888	0	32,436,370	0.81	29,420,791	0.73	7,956,124	19.12	14.01
Guangzhou-3	45,033,758	39,008,158	0	31,502,649	0.81	28,628,638	0.73	7,851,254	18.58	13.63
Guangzhou-4	57,568,725	40,067,496	0	31,293,992	0.78	26,690,066	0.67	7,304,105	19.08	12.71
Hainan-1	51,056,695	45,376,894	0	36,760,076	0.81	33,858,898	0.75	7,491,604	21.61	16.12
Hainan-2	54,047,603	47,904,337	0	39,094,511	0.82	36,075,498	0.75	7,552,831	22.81	17.18
Hainan-3	46,008,347	40,307,969	0	32,355,487	0.80	29,700,856	0.74	7,744,977	19.19	14.14
Hainan-4	50,603,807	46,002,285	0	36,598,359	0.80	33,433,698	0.73	7,993,360	21.91	15.92
Hubei-1	49,345,780	45,220,663	0	36,826,994	0.81	33,620,501	0.74	8,287,003	21.53	16.01
Hubei-2	54,087,204	49,506,361	0	39,501,848	0.80	36,470,734	0.74	8,362,355	23.57	17.37
Hubei-3	52,698,155	48,320,917	0	39,322,915	0.81	36,265,071	0.75	8,459,224	23.01	17.27
Hubei-4	47,512,918	44,390,732	0	36,152,586	0.81	32,843,142	0.74	8,241,445	21.14	15.64
Hunan1-1	81,700,483	72,218,263	0	62,745,560	0.87	59,757,501	0.83	3,891,026	34.39	28.46
Hunan1-2	60,686,520	53,677,051	0	46,795,734	0.87	44,421,858	0.83	3,712,569	25.56	21.15
Hunan1-3	64,703,715	54,961,001	0	47,675,134	0.87	44,211,495	0.80	3,753,051	26.17	21.05
Hunan2-1	50,470,982	44,611,111	0	35,640,396	0.80	32,101,416	0.72	8,322,765	21.24	15.29
Hunan2-2	52,404,186	46,163,099	0	36,510,372	0.79	32,745,225	0.71	8,341,661	21.98	15.59
Hunan2-3	43,402,282	38,977,494	0	30,374,820	0.78	27,299,048	0.70	7,794,440	18.56	13.00
Hunan2-4	55,912,648	47,414,470	0	30,705,799	0.65	27,931,073	0.59	7,880,633	22.58	13.30
Hunan3-1	48,990,487	41,192,802	0	33,590,574	0.82	30,291,627	0.74	8,044,493	19.62	14.42
Hunan3-2	50,229,123	44,862,146	0	36,297,763	0.81	33,108,909	0.74	8,312,814	21.36	15.77
Hunan3-3	50,716,451	44,306,245	0	34,772,829	0.78	31,237,038	0.71	8,137,196	21.10	14.87
Hunan3-4	53,245,925	45,271,935	0	35,359,429	0.78	32,089,594	0.71	8,125,632	21.56	15.28
IND-Delhi-D2	45,470,129	39,185,015	0	31,064,654	0.79	28,277,730	0.72	6,362,407	18.66	13.47
IND-Delhi-D3	50,612,852	45,357,166	0	36,054,545	0.79	33,011,849	0.73	6,572,110	21.60	15.72
IND-Delhi-D4	45,986,924	41,410,120	0	30,760,537	0.74	28,330,630	0.68	6,258,570	19.72	13.49
IND-Delhi-D5	51,582,952	47,472,715	0	37,817,715	0.80	34,802,556	0.73	6,896,841	22.61	16.57
IND-Hyderabad-A4	46,884,810	39,238,965	0	32,127,550	0.82	29,216,363	0.74	7,519,482	18.69	13.91
IND-Hyderabad-A6	47,091,801	42,418,756	0	34,503,418	0.81	31,656,569	0.75	8,032,265	20.20	15.07
IND-Hyderabad-B1	55,832,656	47,981,022	0	39,281,217	0.82	36,156,785	0.75	8,202,457	22.85	17.22
IND-Hyderabad-B2	46,099,846	40,007,539	0	32,626,711	0.82	29,915,620	0.75	7,722,991	19.05	14.25
IND-Matsyapuri-10	57,342,652	52,727,420	0	42,125,065	0.80	38,465,011	0.73	8,510,713	25.11	18.32
IND-Matsyapuri-3	50,118,978	45,345,602	0	36,159,791	0.80	33,048,725	0.73	8,247,509	21.59	15.74
IND-Matsyapuri-5	52,786,173	47,604,291	0	37,341,849	0.78	34,213,035	0.72	8,277,465	22.67	16.29
IND-Matsyapuri-9	36,295,165	32,835,667	0	25,812,178	0.79	22,565,857	0.69	7,029,474	15.64	10.75
JP-ok-1	51,654,458	46,222,609	0	38,734,303	0.84	35,898,735	0.78	7,825,348	22.01	17.09
JP-ok-2	47,838,019	42,336,323	0	31,380,499	0.74	28,892,016	0.68	7,428,300	20.16	13.76
JP-ok-3	49,965,394	44,612,911	0	31,348,557	0.70	28,999,815	0.65	7,164,361	21.24	13.81
JP-ok-4	48,707,090	43,735,977	0	36,564,393	0.84	33,864,080	0.77	7,713,099	20.83	16.13
JP-Ts-1	48,263,663	44,420,687	0	36,125,059	0.81	33,058,064	0.74	8,101,444	21.15	15.74
JP-Ts-2	52,254,680	47,779,350	0	38,849,898	0.81	35,638,552	0.75	8,331,715	22.75	16.97
JP-Ts-3	52,937,991	48,787,767	0	39,695,769	0.81	36,563,660	0.75	8,200,433	23.23	17.41
JP-Ts-4	47,096,266	43,512,206	0	35,471,372	0.82	32,595,059	0.75	8,009,079	20.72	15.52
NSU-1	50,846,850	44,192,423	0	38,202,920	0.86	36,116,233	0.82	3,839,755	21.04	17.20
NSU-3	51,563,295	42,930,224	0	36,276,947	0.85	34,010,428	0.79	3,786,539	20.44	16.20
NSU-4	53,090,372	43,010,360	0	30,423,631	0.71	28,937,163	0.67	3,459,342	20.48	13.78
NSU-6	47,565,655	39,574,037	0	34,192,979	0.86	32,200,676	0.81	3,701,912	18.84	15.33
Zhejiang-1	46,159,946	33,699,742	0	26,655,266	0.79	23,213,443	0.69	6,995,633	16.05	11.05
Zhejiang-2	43,664,509	32,686,893	0	26,158,959	0.80	23,418,477	0.72	7,079,637	15.57	11.15
Zhejiang-3	54,964,512	39,792,819	0	31,446,445	0.79	27,678,321	0.70	7,652,171	18.95	13.18
Zhejiang-4	52,198,930	36,667,317	0	28,488,732	0.78	24,353,271	0.66	7,086,138	17.46	11.60

Table 5.2 Genomic nucleotide diversities π of individuals of each sampling location.

	pi	se
Fujian	0.01489	0.00044
Guangxi	0.01308	0.00019
Guangzhou	0.01512	0.00021
Hainan	0.01351	0.00021
Hubei	0.01309	0.00020
Hunan1	0.00180	0.00003
Hunan2	0.01433	0.00022
Hunan3	0.01294	0.00021
IND-Delhi-D	0.00857	0.00013
IND-Hyderabad	0.01616	0.00026
IND-Matsyapuri	0.01566	0.00023
JP-ok	0.01604	0.00026
JP-Ts	0.01621	0.00025
NSU	0.00194	0.00003

Table 5.3 Pairwise F_{ST} values among the localities. The upper-right figures above diagonal are the F_{ST} values and the lower-left figures below diagonal are the standard errors

	Fujian	Guangxi	Guangzhou1	Hainan	Hubei	Hunan1	Hunan2	Hunan3	IND-Delhi	IND-Hyd	IND-Mat	JP-ok	JP-Ts	Guangzhou2	Zhejiang
Fujian		0.2352	0.0789	0.1447	0.2316	0.8142	0.1438	0.2444	0.4281	0.0063	0.0480	0.0084	0.0075	0.8120	0.0914
Guangxi	0.0125		0.1580	0.2085	0.2399	0.5520	0.1927	0.2445	0.3638	0.1212	0.1435	0.1203	0.1237	0.5760	0.1619
Guangzhou1	0.0125	0.0051		0.1290	0.1613	0.4693	0.1134	0.1666	0.2866	0.0413	0.0643	0.0402	0.0443	0.5001	0.0812
Hainan	0.0124	0.0075	0.0059		0.2112	0.5184	0.1648	0.2163	0.3362	0.0926	0.1151	0.0921	0.0959	0.5476	0.1333
Hubei	0.0107	0.0070	0.0086	0.0067		0.5524	0.1965	0.2480	0.3665	0.1251	0.1468	0.1245	0.1281	0.5769	0.1659
Hunan1	0.0129	0.0138	0.0106	0.0085	0.0123		0.5039	0.5580	0.6882	0.4304	0.4520	0.4309	0.4310	0.6539	0.4737
Hunan2	0.0122	0.0047	0.0063	0.0047	0.0070	0.0081		0.2006	0.3200	0.0766	0.0988	0.0762	0.0797	0.5321	0.1173
Hunan3	0.0121	0.0077	0.0062	0.0072	0.0070	0.0105	0.0039		0.3710	0.1288	0.1516	0.1292	0.1324	0.5818	0.1704
IND-Delhi	0.0125	0.0082	0.0166	0.0106	0.0107	0.0120	0.0064	0.0064		0.2467	0.2689	0.2494	0.2532	0.7036	0.2914
IND-Hyderabad	0.0108	0.0067	0.0051	0.0056	0.0062	0.0073	0.0049	0.0069	0.0062		0.0242	0.0021	0.0082	0.4641	0.0445
IND-Matsyapuri	0.0103	0.0048	0.0050	0.0051	0.0112	0.0088	0.0046	0.0080	0.0098	0.0038		0.0252	0.0309	0.4843	0.0674
JP-ok	0.0113	0.0063	0.0034	0.0079	0.0070	0.0135	0.0044	0.0187	0.0064	0.0044	0.0041		0.0054	0.4644	0.0435
JP-Ts	0.0107	0.0053	0.0044	0.0053	0.0038	0.0076	0.0045	0.0042	0.0122	0.0051	0.0038	0.0064		0.4649	0.0480
Guangzhou2	0.0101	0.0094	0.0084	0.0123	0.0093	0.0064	0.0101	0.0081	0.0080	0.0095	0.0089	0.0106	0.0087		0.5043
Zhejiang	0.0118	0.0090	0.0048	0.0062	0.0079	0.0125	0.0076	0.0074	0.0083	0.0052	0.0043	0.0033	0.0051	0.0093	

Table 5.3 Annotation of the identified high Tajima's D CDS sequences with false discovery rate of 10%.

#Scaffold	BIN_START	N_SNPS	TajimaD	P-value	Annotation (By Blast)
scaffold1	2315000	11	1.45675	0.000183341	Nothing
scaffold1	2860000	15	1.61943	6.95E-05	Nothing
scaffold5	1770000	107	1.41666	0.000230833	Nothing
scaffold7	1060000	21	1.5072	0.000136522	peroxisomal multifunctional enzyme
scaffold9	40000	31	1.42507	0.000220008	non-LTR retrotransposon Jockey-like reverse transcriptase gene, homolog
scaffold16	1895000	70	1.3099	0.000419084	Nothing
scaffold25	1985000	11	1.47782	0.000162208	Nothing
scaffold32	340000	15	2.08417	3.15E-06	Nothing
scaffold33	140000	14	1.37242	0.000296436	Nothing
scaffold34	25000	28	1.64009	6.11E-05	Amyelois transitella P protein-like (LOC106131030) homolog
scaffold43	1415000	14	1.60841	7.43E-05	Amyelois transitella phosphatidylinositol 4-kinase alpha (LOC106129982) homolog
scaffold43	165000	16	2.32401	5.29E-07	Nothing
scaffold46	390000	18	1.26148	0.000544791	Nothing
scaffold47	1320000	13	1.47369	0.000166161	Nothing
scaffold63	300000	37	1.39877	0.000255535	Papilio xuthus kelch-like protein 10 (LOC106126522) homolog
scaffold63	565000	12	2.04574	4.14E-06	Nothing
scaffold63	575000	18	1.64719	5.85E-05	Nothing
scaffold63	640000	12	1.5164	0.000129298	Nothing
scaffold68	835000	117	3.43103	2.65E-11	Spodoptera frugiperda sequence from BAC clone 96D18 homolog
scaffold72	1350000	12	2.15165	1.93E-06	Nothing
scaffold78	1070000	79	1.88987	1.22E-05	Nothing
scaffold78	1250000	33	1.87062	1.38E-05	72F1_SfBAC_fin, Spodoptera frugiperda BAC, egg DNA homolog
scaffold78	1265000	24	1.35939	0.000318841	Nothing
scaffold78	645000	25	1.62416	6.75E-05	Nothing
scaffold84	330000	32	2.65135	3.77E-08	Nothing
scaffold92	305000	41	1.50189	0.000140863	SlP450_99
scaffold92	310000	105	2.05265	3.94E-06	SlP450_99
scaffold92	315000	54	2.52847	1.04E-07	SlP450
scaffold92	325000	26	1.64025	6.11E-05	SlP450
scaffold92	335000	33	2.05501	3.88E-06	SlP450_103
scaffold92	340000	158	1.60868	7.42E-05	SlP450_103; SlP450_104
scaffold92	345000	216	1.74604	3.13E-05	SlP450_104; SltuGR177; SltuGR176
scaffold92	350000	174	2.15649	1.86E-06	SltuGR176; SlP450_105;
scaffold92	355000	187	2.64644	3.93E-08	SlP450_105; SlP450_106
scaffold92	365000	63	1.31427	0.000409176	SlP450_107
scaffold92	370000	13	1.41125	0.000238058	Nothing
scaffold92	380000	55	1.99786	5.80E-06	Spodoptera frugiperda sequence from BAC clone 17L04 homolog

scaffold94	160000	12	1.74064	3.24E-05	Bombyx mori transmembrane and TPR repeat-containing protein CG4341 (LOC101744805) homolog
scaffold94	195000	25	1.40823	0.000242182	Bombyx mori transmembrane and TPR repeat-containing protein CG4341 (LOC101744805) homolog
scaffold94	40000	59	1.66652	5.19E-05	Amyelois transitella melanoma inhibitory activity protein 3 (LOC106132358) homolog
scaffold99	0	75	2.23755	1.02E-06	Nothing
scaffold111	760000	17	1.48468	0.000155834	Papilio polytes gelsolin-like (LOC106108575) homolog
scaffold124	995000	14	1.94381	8.43E-06	Nothing
scaffold131	125000	27	2.21216	1.23E-06	Bombyx mori neuropeptide receptor A5 (NGR-A5) homolog
scaffold131	505000	29	1.75328	2.99E-05	Nothing
scaffold131	55000	107	2.42131	2.47E-07	SlituGR206
scaffold131	620000	23	1.57059	9.35E-05	Nothing
scaffold131	630000	11	2.22979	1.08E-06	Nothing
scaffold132	60000	17	1.37172	0.000297601	Nothing
scaffold140	505000	31	1.34852	0.000338726	SlABCA2
scaffold140	520000	81	1.4567	0.000183394	SlABCA2
scaffold140	530000	13	2.52022	1.12E-07	SlABCA2
scaffold140	560000	22	1.64765	5.83E-05	Nothing
scaffold140	580000	28	1.91929	9.96E-06	Nothing
scaffold143	165000	11	1.39733	0.000257627	NothinP
scaffold152	10000	60	1.40664	0.000244379	Nothing
scaffold156	45000	34	1.4939	0.000147638	Helicoverpa assulta acetyl-CoA carboxylase 2 mRNA homolog
scaffold158	105000	19	2.90308	4.20E-09	Papilio xuthus pericentrin-like (LOC106126656) homolog
scaffold158	110000	35	1.63552	6.29E-05	Papilio xuthus pericentrin-like (LOC106126656) homolog
scaffold167	825000	20	1.35275	0.000330855	Nothing
scaffold169	775000	27	1.27422	0.000508706	Nothing
scaffold179	115000	17	1.81437	2.01E-05	Amyelois transitella paired box pox-neuro protein (LOC106135272) homolog
scaffold183	105000	69	2.03097	4.60E-06	Nothing
scaffold183	195000	19	1.70485	4.08E-05	Nothing
scaffold183	200000	34	1.8947	1.18E-05	Nothing
scaffold187	120000	49	1.47193	0.000167873	Sl_COE2_ace1; 70A06_SfBAC_fin, Spodoptera frugiperda BAC, egg DNA homolog
scaffold187	590000	44	1.31438	0.000408929	Nothing
scaffold187	765000	13	1.55979	9.98E-05	Bombyx mori PAX3- and PAX7-binding protein 1 (LOC101745121) homolog
scaffold191	150000	23	1.80801	2.10E-05	Nothing
scaffold203	110000	11	1.2622	0.00054269	Nothing
scaffold215	535000	37	1.33184	0.000371494	Nothing
scaffold252	260000	12	1.88048	1.30E-05	Nothing
scaffold259	0	21	2.57028	7.41E-08	Nothing
scaffold260	180000	18	2.4618	1.79E-07	Nothing
scaffold282	180000	58	1.89661	1.16E-05	Nothing
scaffold292	30000	83	1.43581	0.000206872	Bombyx mori WD repeat and FYVE domain-containing protein 3 (LOC101746532) homolog
scaffold292	35000	35	1.41682	0.000230623	Bombyx mori WD repeat and FYVE domain-containing protein 3 (LOC101746532) homolog

scaffold312	365000	13	1.58014	8.83E-05	PREDICTED: Amyelois transitella T-cell leukemia homeobox protein 2 (LOC106130720) homolog
scaffold323	220000	117	1.32809	0.000379255	SlituGR120; SlituGR121
scaffold323	245000	113	1.82013	1.94E-05	SlituGR125; SlituGR126
scaffold323	345000	155	1.49563	0.000146145	SlituGR139; SlituGR140
scaffold323	400000	104	1.40802	0.000242471	SlituGR150; SlituGR151
scaffold363	10000	19	1.72313	3.63E-05	Nothing
scaffold383	95000	13	1.47317	0.000166665	ARP2
scaffold386	155000	12	1.30125	0.000439355	Nothing
scaffold388	335000	22	2.26698	8.18E-07	Plutella xylostella cleft lip and palate transmembrane protein 1-like protein (LOC105395101) homolog
scaffold392	10000	23	1.57124	9.31E-05	Nothing
scaffold393	175000	13	1.3342	0.000366685	Nothing
scaffold399	260000	75	1.37353	0.000294597	Nothing
scaffold418	85000	31	1.27842	0.000497304	SlituOR7
scaffold422	70000	11	1.53406	0.000116424	Nothing
scaffold424	225000	41	1.9333	9.06E-06	Papilio xuthus E3 ubiquitin-protein ligase TRIM9 (LOC106127822) homolog
scaffold424	230000	40	1.45598	0.00018416	Papilio xuthus E3 ubiquitin-protein ligase TRIM9 (LOC106127822) homolog
scaffold424	235000	30	1.33709	0.000360876	Papilio xuthus E3 ubiquitin-protein ligase TRIM9 (LOC106127822) homolog
scaffold424	245000	44	2.54577	9.07E-08	72F1_SfBAC_fin, Spodoptera frugiperda BAC, egg DNA homolog
scaffold424	270000	18	3.06562	9.44E-10	72F1_SfBAC_fin, Spodoptera frugiperda BAC, egg DNA homolog
scaffold424	280000	11	2.47044	1.67E-07	72F1_SfBAC_fin, Spodoptera frugiperda BAC, egg DNA homolog
scaffold446	190000	20	2.32992	5.05E-07	PREDICTED: Megachile rotundata lysosomal alpha-mannosidase (LOC100881716) homolog
scaffold462	140000	24	1.40361	0.000248619	Nothing
scaffold465	255000	12	1.88568	1.25E-05	Nothing
scaffold470	10000	104	1.63708	6.23E-05	SlituGR181; SlituGR180
scaffold470	185000	35	2.09293	2.96E-06	Nothing
scaffold470	195000	33	1.35975	0.000318201	Nothing
scaffold470	210000	14	1.63871	6.17E-05	Nothing
scaffold470	235000	37	1.55317	0.000103853	Nothing
scaffold470	240000	47	1.5733	9.20E-05	Nothing
scaffold470	245000	18	1.33724	0.000360576	Nothing
scaffold471	30000	33	2.11001	2.61E-06	cadherin-like receptor
scaffold479	185000	78	1.828	1.84E-05	SlituGR200
scaffold479	20000	78	1.42065	0.000225637	Nothing
scaffold479	220000	197	1.50994	0.000134332	Tremella mesenterica DSM 1558 hypothetical protein (TREMEDRAFT_16078) homolog
scaffold479	225000	90	2.52223	1.10E-07	Albugo laibachii Alem1, genomic contig CONTIG_18_Em1_cons_v4_198151_220_163917
scaffold500	210000	49	1.65442	5.59E-05	Nothing
scaffold518	185000	17	1.31203	0.000414227	Nothing
scaffold566	105000	95	1.3933	0.000263566	Nothing
scaffold566	110000	91	1.99386	5.96E-06	Nothing
scaffold575	25000	24	1.28569	0.000478125	Nothing
scaffold575	40000	22	1.65	5.75E-05	TATA-box-binding protein
scaffold575	50000	37	1.56848	9.47E-05	72F1_SfBAC_fin, Spodoptera frugiperda BAC, egg

					DNA homolog
scaffold600	100000	35	1.50391	0.000139197	Nothing
scaffold604	115000	47	2.65322	3.71E-08	Nothing
scaffold636	95000	12	1.33203	0.000371104	Nothing
scaffold640	40000	51	1.29969	0.000443106	Nothing
scaffold651	15000	20	1.3152	0.000407096	Nothing
scaffold669	60000	24	1.56916	9.43E-05	Nothing
scaffold674	80000	76	1.34199	0.00035122	Bombyx mori kettin protein (Kettin) homolog
scaffold698	0	23	2.27595	7.64E-07	Nothing
scaffold698	5000	145	1.59452	8.09E-05	Nothing
scaffold867	10000	80	1.44795	0.000192907	Spodoptera exigua midgut class 1 aminopeptidase N (apn1) homolog
scaffold2122	0	11	2.69779	2.54E-08	Nothing
scaffold2447	0	88	1.92843	9.36E-06	Nothing

Table 5.4 The estimated population histories based on the analysis of two dimensional allele frequency spectra

population 1 population 2	China_local China_migrate	China_migrate India_migrate	China_migrate Japan_migrate	India_migrate Japan_migrate	China_local India_migrate	China_local Japan_migrate	India_local India_migrate
N_{anc}	1.38×10^{7} $\pm 5.23\times10^{5}$	1.26×10^{7} $\pm 6.01\times10^{5}$	1.45×10^{7} $\pm 1.31\times10^{7}$	1.08×10^{7} $\pm 5.37\times10^{5}$	1.41×10^{7} $\pm 5.70\times10^{5}$	1.38×10^{7} $\pm 5.57\times10^{5}$	1.14×10^{7} $\pm 4.89\times10^{5}$
T_{anc}	9.49×10^{5} $\pm 4.99\times10^{4}$	9.17×10^{5} $\pm 6.62\times10^{4}$	6.99×10^{5} $\pm 2.81\times10^{6}$	9.13×10^{5} $\pm 6.80\times10^{4}$	9.28×10^{5} $\pm 5.11\times10^{4}$	9.19×10^{5} $\pm 5.06\times10^{4}$	8.42×10^{5} $\pm 6.32\times10^{4}$
T_{split}	1.16×10^{3} $\pm 4.21\times10^{1}$	8.88×10^{3} $\pm 1.20\times10^{3}$	2.30×10^{5} $\pm 1.19\times10^{5}$	2.89×10^{3} $\pm 2.82\times10^{-7}$	2.20×10^{2} ± 3.05	5.54×10^{3} $\pm 1.74\times10^{2}$	2.50×10^{3} $\pm 1.37\times10^{2}$
N_{pop1_split}	1.32×10^{7} $\pm 5.20\times10^{5}$	3.28×10^{6} $\pm 3.68\times10^{5}$	1.40×10^{7} $\pm 1.25\times10^{7}$	9.53×10^{6} $\pm 5.35\times10^{5}$	1.41×10^{7} $\pm 5.66\times10^{5}$	1.38×10^{7} $\pm 5.57\times10^{5}$	6.17×10^{6} $\pm 2.38\times10^{5}$
$N_{pop1\_current}$	8.24×10^{3} $\pm 1.47\times10^{2}$	1.04×10^{6} $\pm 1.07\times10^{4}$	5.81×10^{6} $\pm 2.93\times10^{7}$	3.37×10^{5} $\pm 9.22\times10^{3}$	1.17×10^{3} $\pm 8.18\times10^{1}$	4.92×10^{4} $\pm 8.10\times10^{2}$	6.83×10^{2} $\pm 4.93\times10^{1}$
N_{pop2_split}	5.68×10^{5} $\pm 1.23\times10^{4}$	9.31×10^{6} $\pm 2.33\times10^{5}$	4.63×10^{5} $\pm 5.95\times10^{5}$	1.22×10^{6} $\pm 2.17\times10^{3}$	3.92×10^{3} $\pm 3.97\times10^{1}$	4.86×10^{4} $\pm 7.24\times10^{2}$	5.23×10^{6} $\pm 2.52\times10^{5}$
$N_{pop2\_current}$	1.20×10^{4} $\pm 2.37\times10^{2}$	3.62×10^{5} $\pm 5.11\times10^{4}$	5.54×10^{6} $\pm 4.69\times10^{6}$	3.98×10^{5} $\pm 8.53\times10^{3}$	9.17×10^{4} $\pm 3.40\times10^{2}$	2.13×10^{8} $\pm 3.02\times10^{6}$	4.15×10^{4} $\pm 1.66\times10^{3}$
$M_{pop1\leftarrow pop2}$	6.44×10^{2} $\pm 1.55\times10^{1}$	9.00×10^{-4} ± 0.021	1.91×10^{1} ± 2.98	2.83×10^{-3} ± 0.057	2.64×10^{3} ± 5.46	6.91×10^{1} ± 1.01	4.45×10^{1} ± 0.42
$M_{pop2\leftarrow pop1}$	6.90×10^{2} ± 7.18	9.00×10^{-4} ± 0.011	6.27×10^{1} $\pm 2.28\times10^{1}$	2.83×10^{-3} ± 0.017	1.28×10^{2} $\pm 2.25\times10^{2}$	1.36×10^{2} ± 2.01	4.45×10^{1} ± 0.41

The size of a population (N) represents the number of chromosomes. The migration rate (M) is the number of migrating chromosomes per generation. The time of the population split is the number of years before present. The figures after ± are the standard errors obtained by parametric bootstrap.

Chapter 6. Conclusion

In this PhD Thesis, I presented the results of two of my works, the first of which deals with the evolution of mammalian life history and the other is the population history, migration and adaptation of agricultural pests (Spodoptera litura). The former investigates species level genomic evolution and the latter investigates population level genomic evolution. For the first work, I built a new framework of evolutionary biology that can accurately calculate species divergence times and trace the history of species life history evolution. For the second work, I applied a previous existing method to solve an actual biological problem. The main methods include analysis of branch lengths of individual genes, analysis of genome-wide distribution of Tajima's D, and fitting a formula in population genetics to joined allele frequency spectrum. The key features are statistical models of the summary statistics to infer genomic evolution.

In my new framework, genomic evolution is the context behind the evolution of each individual gene; each individual gene has its own unique rate of evolution due to its function. Genomic evolution is affected by the genome-wide mutation rate and the species divergence times; and the unique evolutionary rate of each gene is correlated with the selection pressure on the gene. The changing biological traits and behavior is the result of adaptation to the environment. Therefore, changes of gene-specific molecular evolutionary rates may well record the history of biological traits and behavioral changes. My result for ancestral state reconstruction was conducted without any pre-information on ancestral nodes. Nevertheless, the result on the evolutionary history of diurnality suggested ancient climate change, especially global cooling event after EOT (33.9 Mya), had deep impact on the evolution of modern mammals.

My work on the population history, migration and adaptation of Spodoptera litura offered another example of the interaction of genomic variation and environmental changes. S. litura could cause huge agriculture damage in large area of Asian is largely due to its polyphagy and strong ability of long-distance migration. Polyghagy facilitates its long-distance migration by enabling this insect feed different food resource that are available on the way. The analysis of genome-wide distribution of Tajima's D indicated an explosion of detoxification genes (P450 gene clusters on scaffold 93), which may be essential for acquisition of polyphagy for this insect.

Again, genomic adaptations and animal behavior have tightly linked to each other.

No matter insect or mammal, their biological traits are constrained by their genetic materials, also subject to the environmental changes. The inference of the interaction among genome evolution, environmental changes and changes in biological traits is the core target of my study. Lives on earth show high diversity, however, evolution is a continuous process and all extant species share common ancestors. Mechanisms under the same biological process maybe the same, and the functions of genes are highly conservative among different species. For example, genes involved in DNA replication, translation and innate immune system, such as DNA Polymerase, Toll-like receptors (TLRs), etc. can be found from insect to human with similar structure and function. Animal behavior is complex and may be controlled by many interacting factors. However, the animal behavior shared by different species may still be governed by the same biological mechanisms. Comparing the variations among different genomes and correlating these variations to the biological traits provides a new way to study such mechanisms behind animal behavior.

As is seen above, the extended neutral theory that was introduced in Chapters 2-4 describes the species-level genome evolution, and infers the variation of functional constraints over the long-term evolutionary time scale. On the other hand, the statistical models in Chapter 5 describes the population-level genome evolution, and infers the recent population history and adaptation after the time at the most recent common ancestor. By integrating the two approaches, it may become possible to estimate the enhanced or reduced functional constraints on the genomic loci in modern society. My future study attempts to answer the important questions such as "How did the agricultural civilization and the medical progress change the direction of the evolution of the human genome?"

Acknowledgement:

I hope to pay my special respect and thanks to my supervisor Prof. Hirohisa Kishino. When I first came to Japan, my knowledge of Japan was limited. I was wondering what I could learn from my PhD courses in Japan. Now I think I am really lucky, and feel his supervision strengthened my ability of scientific research greatly. I published two first author papers during PhD courses, one in Current Biology and one in Nature Ecology&Evolution. He also gave me many chances to visit excellent researchers in the world and join the international meetings. I enjoyed study and research in Tokyo U.

I want to pay my special thanks to Prof. Jeffrey Thorne in North Carolina State University. He helped us to revise the manuscript of Current Biology paper and gave many great suggestions. We had asked him 5 times to be co-author of this papers, but he declined 5 times because he restricts his publications to the products of the research that he is deeply involved in from the beginning. I am impressed by his great gentleness, and respect his conscientiousness to Science.

I hope to thank Prof. Kazuei Mita in Southwest University (China). He invited Prof. Kishino and me to join the genome project on *S. litura*. He tried his best to organize this international cooperation work. I am impressed that researchers from more than 5 countries had joined this work. Without his hard working, this genome project could not have been well integrated to generate valuable results.

I hope to thank all members in the Laboratory of Biometry and Bioinformatics of The University of Tokyo. I thank everyone's kindness in mylab. When I meet difficulties, they always help me patiently.

I hope to thank Dr. Takahiro Yonezawa, my beloved husband. He was my senior of my Master courses. He taught me the most basic knowledge in this region and guided me to enter this field in my Master courses. Myfrequent discussions on Science during my PhD courses also helped my growth in this field. Without his great support both for daily life and research, my PhD courses may have resulted in failure.

I hope to pay my special respect and thanks to my Master supervisor, Prof. Yang Zhong, who passed away due to a traffic accidence in Sep. 25th 2017. He accepted me

when I hoped to enter his lab and gave me chance to study under the instruction of Prof. Masami Hasegawa and Dr. Takahiro Yonezawa. Without him, I had no chance to enter current research field. It is so sad he left us. May he rest in peace.

References

1. Aristotle (350 B.C.E). On longevity and shortness of life, (Greek).
2. Aristotle (350 B.C.E). On life and death, (Greek).
3. Aristotle (350 B.C.E.). Movement of Animals.
4. Aristotle (350 B.C.E.). History of Animals.
5. Aristotle (350 B.C.E). Progression of Animals.
6. Darwin, C. (1859). On the origin of species, (London: John Murray).
7. Huxley, T.H. (1860). ART. VIII.—Darwin on the Origin of Species, Volume 17, (London: Baldwin, Cradock, and Joy).
8. Mendel, J.G. (1866). Experiments in plant hybridization.
9. Galton, F. (1889). Natural Inheritance, (London: MacMillan).
10. Moore, R.G. (2001). The "rediscovery" of Mendel's work. Biosciene *27*, 13–24.
11. Jude, C. (2013). Genes and genetics: the language of scientific discovery, (London: Oxford English Dictionary).
12. Grafen, A.R., Mark (2006). Richard Dawkins: How A Scientist Changed the Way We Think, (New York: Oxford University).
13. Fisher, R.A. (1930). The Genetical Theory of Natural Selection (London: Oxford University Press).
14. Wright, S. (1931). Evolution in Mendelian population. Genetics *16*, 97-159.
15. Haldane, J.B.S. (1957). The cost of natural selection. . J. Genetics *55*, 511-524.
16. Watson, J.D., and Crick, F.H. (1953). The structure of DNA. Cold Spring Harb. Symp. Quant. Biol. *18*, 123–131.
17. Zuckerkandl, E., and Pauling, L.B. (1962). Molecular disease, evolution, and genetic heterogeneity, (New York: Academic Press).
18. Kimura, M. (1968). Evolutionary rate at the molecular level. Nature *217*, 624–626.
19. Arnason, U., Bodin, K., Gullberg, A., Ledje, C., and Mouchaty, S. (1995). A molecular view of pinniped relationships with particular emphasis on the true seals. J Mol Evol *40*, 78-85.

20. Meyer, A. (1994). Shortcomings of the cytochrome b gene as a molecular marker. Trends Ecol Evol *9*, 278-280.
21. Olsen, G.J. (1987). Earliest phylogenetic branchings: comparing rRNA-based evolutionary trees inferred with various techniques. Cold Spring Harb Symp Quant Biol *52*, 825-837.
22. dos Reis, M., Zhu, T., and Yang, Z. (2014). The impact of the rate prior on Bayesian estimation of divergence times with multiple Loci. Syst. Biol. *63*, 555–565.
23. Misof, B., Liu, S., Meusemann, K., Peters, R.S., Donath, A., Mayer, C., Frandsen, P.B., Ware, J., Flouri, T., Beutel, R.G., et al. (2014). Phylogenomics resolves the timing and pattern of insect evolution. Science *346*, 763-767.
24. Jarvis, E.D., Mirarab, S., Aberer, A.J., Li, B., Houde, P., Li, C., Ho, S.Y., Faircloth, B.C., Nabholz, B., Howard, J.T., et al. (2014). Whole-genome analyses resolve early branches in the tree of life of modern birds. Science *346*, 1320-1331.
25. Li, H., and Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. Nature *475*, 493-496.
26. Ho, W.C., Ohya, Y., and Zhang, J. (2017). Testing the neutral hypothesis of phenotypic evolution. Proc Natl Acad Sci U S A *114*, 12219-12224.
27. Yang, Z. (2006). Computational Molecular Evolution, (Oxford, England: Oxford University Press).
28. Yang, Z., and Rannala, B. (2012). Molecular phylogenetics: principles and practice. Nat. Rev. Genet. *13*, 303–314.
29. Fitch, W.M., and Margoliash, E. (1967). Construction of phylogenetic trees. Science *155*, 279-284.
30. Jukes, T.H., and Cantor, C.R. (1969). Evolution of Protein Molecules, (New York: Academic Press).
31. Hasegawa, M., Kishino, H., and Yano, T.A. (1985). Dating of the Human Ape Splitting by a Molecular Clock of Mitochondrial-DNA. J. Mol. Evol. *22*, 160–174.
32. Rodriguez, F., Oliver, J.L., Marin, A., and Medina, J.R. (1990). The General Stochastic-Model of Nucleotide Substitution. J. Theor. Biol. *142*, 485–501.

33. Yang, Z. (1993). Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol. Biol. Evol. *10*, 1396–1401.
34. Saitou, N., and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. *4*, 406–425.
35. Swinburne, R. (1997). The Evolution of the Soul, (Oxford, England: Oxford University Press).
36. Wu, J., Hasegawa, M., Zhong, Y., and Yonezawa, T. (2014). Importance of synonymous substitutions under dense taxon sampling and appropriate modeling in reconstructing the mitogenomic tree of Eutheria. Genes Genet. Syst. *89*, 237–251.
37. Cavalli-Sforza, L.L., and Edwards, A.W. (1967). Phylogenetic analysis. Models and estimation procedures. Am J Hum Genet *19*, 233-257.
38. Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. *17*, 368–376.
39. Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control *19*, 716–723.
40. Schwarz, G. (1978). Estimating the Dimension of a Model. Annals of Statistics *6*, 461-464.
41. Tuffley, C., and Steel, M. (1997). Links between maximum likelihood and maximum parsimony under a simple model of substitution. Bulletin of Mathematical Biology *59*, 581–607.
42. dos Reis, M., Donoghue, P.C., and Yang, Z. (2016). Bayesian molecular clock dating of species divergences in the genomics era. Nat Rev Genet *17*, 71-80.
43. Rannala, B., and Yang, Z. (1996). Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. J Mol Evol *43*, 304-311.
44. Yang, Z., and Rannala, B. (1997). Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo Method. Mol Biol Evol *14*, 717-724.

45. B., M., and A., N.M. (1997). Phylogenetic inference for binary data on dendrograms using Markov chain Monte Carlo. J. Comp. Grap. Stat. *6*, 122–131.
46. Li, S., Pearl, D., and Doss, H. (2000). Phylogenetic tree construction using Markov chain Monte Carlo. J. Am. Stat. Assoc. *95*, 493–508.
47. Huelsenbeck, J.P., and Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics *17*, 754–755.
48. Li, W.H., Tanimura, M., and Sharp, P.M. (1987). An evaluation of the molecular clock hypothesis using mammalian DNA sequences. J. Mol. Evol. *25*, 330–342.
49. Wolfe, K.H., Sharp, P.M., and Li, W.H. (1989). Mutation rates differ among regions of the mammalian genome. Nature *337*, 283-285.
50. Nabholz, B., Glemin, S., and Galtier, N. (2008). Strong variations of mitochondrial mutation rate across mammals--the longevity hypothesis. Mol Biol Evol *25*, 120-130.
51. Mooers, A.O., and Harvey, P.H. (1994). Metabolic rate, generation time, and the rate of molecular evolution in birds. Mol Phylogenet Evol *3*, 344-350.
52. Rambaut, A., and Bromham, L. (1998). Estimating divergence dates from molecular sequences. Mol Biol Evol *15*, 442-448.
53. Sanderson, M.J. (1997). A nonparametric approach to estimating divergence times in the absence of rate constancy. Mol. Biol. Evol. *14*, 1218–1232.
54. Thorne, J.L., Kishino, H., and Painter, I.S. (1998). Estimating the rate of evolution of the rate of molecular evolution. Mol Biol Evol *15*, 1647-1657.
55. Kishino, H., Thorne, J.L., and Bruno, W.J. (2001). Performance of a divergence time estimation method under a probabilistic model of rate evolution. Mol Biol Evol *18*, 352-361.
56. Thorne, J.L., and Kishino, H. (2002). Divergence time and evolutionary rate estimation with multilocus data. Syst Biol *51*, 689-702.
57. Drummond, A.J., Ho, S.Y., Phillips, M.J., and Rambaut, A. (2006). Relaxed phylogenetics and dating with confidence. PLoS Biol. *4*, e88.

58. Yang, Z., and Rannala, B. (2006). Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. Mol Biol Evol *23*, 212-226.
59. Lepage, T., Bryant, D., Philippe, H., and Lartillot, N. (2007). A general comparison of relaxed molecular clock models. Mol Biol Evol *24*, 2669-2680.
60. Rannala, B., and Yang, Z. (2007). Inferring speciation times under an episodic molecular clock. Syst Biol *56*, 453-466.
61. Maddison, W.P. (1997). Gene trees in species trees. Syst. Biol. *46*, 523–536.
62. Tonini, J., Moore, A., Stern, D., Shcheglovitova, M., and Orti, G. (2015). Concatenation and species tree methods exhibit statistically indistinguishable accuracy under a range of simulated conditions. PLoS Curr *7*.
63. Gatesy, J., and Springer, M.S. (2014). Phylogenetic analysis at deep timescales: unreliable gene trees, bypassed hidden support, and the coalescence/concatalescence conundrum. Mol Phylogenet Evol *80*, 231-266.
64. Nylander, J.A., Ronquist, F., Huelsenbeck, J.P., and Nieves-Aldrey, J.L. (2004). Bayesian phylogenetic analysis of combined data. Syst Biol *53*, 47-67.
65. Brandley, M.C., Schmitz, A., and Reeder, T.W. (2005). Partitioned Bayesian analyses, partition choice, and the phylogenetic relationships of scincid lizards. Syst Biol *54*, 373-390.
66. Kainer, D., and Lanfear, R. (2015). The effects of partitioning on phylogenetic inference. Mol Biol Evol *32*, 1611-1627.
67. Zhu, T., Dos Reis, M., and Yang, Z. (2015). Characterization of the uncertainty of divergence time estimation under relaxed molecular clock models using multiple loci. Syst Biol *64*, 267-280.
68. Angelis, K., Alvarez-Carretero, S., Dos Reis, M., and Yang, Z. (2017). An Evaluation of Different Partitioning Strategies for Bayesian Estimation of Species Divergence Times. Syst Biol.

69. Liu, L., Yu, L., Kubatko, L., Pearl, D.K., and Edwards, S.V. (2009). Coalescent methods for estimating phylogenetic trees. Mol Phylogenet Evol *53*, 320-328.

70. Song, S., Liu, L., Edwards, S.V., and Wu, S. (2012). Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. Proc. Natl. Acad. Sci. USA *109*, 14942–14947.

71. Pagel, M. (1997). Inferring evolutionary processes from phylogenies. Zoologica. Scripta. *26*, 331–348.

72. Pagel, M. (1999). Inferring the historical patterns of biological evolution. Nature *401*, 877–884.

73. Pagel, M. (1999). The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies. Syst. Biol. *48*, 612–622.

74. Cunningham, C.W., Omland, K.E., and Oakley, T.H. (1998). Reconstructing ancestral character states: a critical reappraisal. Trends Ecol Evol *13*, 361-366.

75. Pagel, M., and Meade, A. (2006). Bayesian analysis of correlated evolution of discrete characters by reversible-jump Markov chain Monte Carlo. Am. Nat. *167*, 808–825.

76. Pagel, M., Meade, A., and Barker, D. (2004). Bayesian estimation of ancestral character states on phylogenies. Syst. Biol. *53*, 673-684.

77. Varki, A., and Altheide, T.K. (2005). Comparing the human and chimpanzee genomes: searching for needles in a haystack. Genome Res *15*, 1746-1758.

78. Tugendreich, S., Bassett, D.E., Jr., McKusick, V.A., Boguski, M.S., and Hieter, P. (1994). Genes conserved in yeast and humans. Hum Mol Genet *3 Spec No*, 1509-1517.

79. T., M.M., and R., C.R. (1985). Phylogeny of the Vibrionaceae, and recommendation of two new genera, Listonella and Shewanella. Syst Appl Microbiol *6*, 171–182.

80. Lane, D.J., Pace, B., Olsen, G.J., Stahl, D.A., Sogin, M.L., and Pace, N.R. (1985). Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. Proc Natl Acad Sci U S A *82*, 6955-6959.

81. Liu, Y., Rossiter, S.J., Han, X., Cotton, J.A., and Zhang, S. (2010). Cetaceans on a molecular fast track to ultrasonic hearing. Curr Biol *20*, 1834-1839.
82. Darwin, C. (1872). The expression of the emotions in man and animals, (London: John Murray).
83. Yonezawa, T., Segawa, T., Mori, H., Campos, P.F., Hongoh, Y., Endo, H., Akiyoshi, A., Kohno, N., Nishida, S., Wu, J., et al. (2017). Phylogenomics and Morphology of Extinct Paleognaths Reveal the Origin and Evolution of the Ratites. Curr Biol *27*, 68-77.
84. Jacobs, G.H. (2009). Evolution of colour vision in mammals. Philos Trans R Soc Lond B Biol Sci *364*, 2957–2967.
85. Wu, Y., Wang, H., and Hadly, E.A. (2017). Invasion of Ancestral Mammals into Dim-light Environments Inferred from Adaptive Evolution of the Phototransduction Genes. Sci Rep *7*, 46542.
86. Popadin, K., Polishchuk, L.V., Mamirova, L., Knorre, D., and Gunbin, K. (2007). Accumulation of slightly deleterious mutations in mitochondrial protein-coding genes of large versus small mammals. Proc Natl Acad Sci U S A *104*, 13390-13395.
87. Ohta, T. (1973). Slightly deleterious mutant substitutions in evolution. Nature *246*, 96-98.
88. Lartillot, N., and Delsuc, F. (2012). Joint reconstruction of divergence times and life-history evolution in placental mammals using a phylogenetic covariance model. Evolution *66*, 1773–1787.
89. Lartillot, N., and Poujol, R. (2011). A phylogenetic model for investigating correlated evolution of substitution rates and continuous phenotypic characters. Mol. Biol. Evol. *28*, 729–744.
90. Kimura, M. (1977). Preponderance of Synonymous Changes as Evidence for Neutral Theory of Molecular Evolution. Nature *267*, 275–276.
91. Bromham, L. (2011). The genome as a life-history character: why rate of molecular evolution varies between mammal species. Philos Trans R Soc Lond B Biol Sci *366*, 2503-2513.
92. Ho, S.Y., and Duchene, S. (2014). Molecular-clock methods for estimating evolutionary rates and timescales. Mol Ecol *23*, 5947-5965.

93. Martin, A.P., and Palumbi, S.R. (1993). Body size, metabolic rate, generation time, and the molecular clock. Proc. Natl. Acad. Sci. U. S. A. *90*, 4087–4091.

94. Kumar, S., and Subramanian, S. (2002). Mutation rates in mammalian genomes. Proc. Natl. Acad. Sci. U. S. A. *99*, 803–808.

95. Thomas, J.A., Welch, J.J., Lanfear, R., and Bromham, L. (2010). A generation time effect on the rate of molecular evolution in invertebrates. Mol. Biol. Evol. *27*, 1173–1180.

96. dos Reis, M., Inoue, J., Hasegawa, M., Asher, R.J., Donoghue, P.C., and Yang, Z. (2012). Phylogenomic datasets provide both precision and accuracy in estimating the timescale of placental mammal phylogeny. Proc. Biol. Sci. *279*, 3491–3500.

97. Nowak, R.M. (1999). Walker's Mammals of the World, Sixth Edition. Edition, (Baltimore: Johns Hopkins University Press).

98. Datta, P.M. (2005). Earliest mammal with transversely expanded upper molar from the Late Triassic (Carnian) Tiki Formation, South Rewa Gondwana Basin, India. J Vertebr Paleontol *25*, 200–207.

99. Luo, Z.X., and Martin, T. (2007). Analysis of Molar Structure and Phylogeny of Docodont Genera. Bulletin of Carnegie Museum of Natural History *39*, 27–47.

100. Luo, Z.X., Yuan, C.X., Meng, Q.J., and Ji, Q. (2011). A Jurassic eutherian mammal and divergence of marsupials and placentals. Nature *476*, 442–445.

101. Archibald, J.D., and Deutschman, D.H. (2001). Quantitative analysis of the timing of the origin and diversification of extant placental orders. J. Mamm. Evol. *8*, 107–124.

102. Alroy, J. (1999). The fossil record of North American mammals: evidence for a Paleocene evolutionary radiation. Syst Biol *48*, 107-118.

103. O'Leary, M.A., Bloch, J.I., Flynn, J.J., Gaudin, T.J., Giallombardo, A., Giannini, N.P., Goldberg, S.L., Kraatz, B.P., Luo, Z.X., Meng, J., et al. (2013). The placental mammal ancestor and the post-K-Pg radiation of placentals. Science *339*, 662–667.

104. Herrero, J., Muffato, M., Beal, K., Fitzgerald, S., Gordon, L., Pignatelli, M., Vilella, A.J., Searle, S.M., Amode, R., Brent, S., et al. (2016). Ensembl comparative genomics resources. Database (Oxford) *2016*.

105. Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. *30*, 772–780.

106. Liu, L., and Pearl, D.K. (2007). Species trees from gene trees: Reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. Syst. Biol. *56*, 504–514.

107. Liu, L.A., Yu, L.L., and Edwards, S.V. (2010). A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. Bmc Evol. Biol. *10*.

108. Rannala, B., and Yang, Z.H. (2003). Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics *164*, 1645–1656.

109. Stamatakis, A., Hoover, P., and Rougemont, J. (2008). A rapid bootstrap algorithm for the RAxML Web servers. Syst. Biol. *57*, 758–771.

110. Yang, Z. (1996). Among-site rate variation and its impact on phylogenetic analyses. Trends Ecol. Evol. *11*, 367–372.

111. Mason, V.C., Li, G., Minx, P., Schmitz, J., Churakov, G., Doronina, L., Melin, A.D., Dominy, N.J., Lim, N.T.L., Springer, M.S., et al. (2016). Genomic analysis reveals hidden biodiversity within colugos, the sister group to primates. Sci Adv *2*, e1600633.

112. Tarver, J.E., Dos Reis, M., Mirarab, S., Moran, R.J., Parker, S., O'Reilly, J.E., King, B.L., O'Connell, M.J., Asher, R.J., Warnow, T., et al. (2016). The Interrelationships of Placental Mammals and the Limits of Phylogenetic Inference. Genome Biol. Evol. *8*, 330–344.

113. Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. *24*, 1586–1591.

114. Le, S.Q., and Gascuel, O. (2008). An improved general amino acid replacement matrix. Mol. Biol. Evol. *25*, 1307–1320.

115. Dang, C.C., Lefort, V., Le, V.S., Le, Q.S., and Gascuel, O. (2011). ReplacementMatrix: a web server for maximum-likelihood estimation of amino acid replacement rate matrices. Bioinformatics *27*, 2758–2760.
116. Yang, Z. (1998). Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol. Biol. Evol. *15*, 568–573.
117. Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. *39*, 306–314.
118. Benton, M.J., Donoghue, P.C.J., and Asher, R.J. (2009). The Timetree of Life, chapter Calibrating and constraining molecular clocks, (Oxford: Oxford University Press).
119. Ji, Q., Luo, Z.X., Yuan, C.X., Wible, J.R., Zhang, J.P., and Georgi, J.A. (2002). The earliest known eutherian mammal. Nature *416*, 816–822.
120. Luo, Z.X., Ji, Q., Wible, J.R., and Yuan, C.X. (2003). An Early Cretaceous tribosphenic mammal and metatherian evolution. Science *302*, 1934–1940.
121. Rougier, G.W., Wible, J.R., and Novacek, M.J. (1998). Implications of Deltatheridium specimens for early marsupial history. Nature *396*, 459–463.
122. Bi, S.D., Jin, X.S., Li, S., and Du, T.M. (2015). A new Cretaceous Metatherian mammal from Henan, China. Peerj *3*.
123. Kusuhashi, N., Tsutsumi, Y., Saegusa, H., Horie, K., Ikeda, T., Yokoyama, K., and Shiraishi, K. (2013). A new Early Cretaceous eutherian mammal from the Sasayama Group, Hyogo, Japan. Proc. Biol. Sci. *280*, 20130142.
124. Archibald, J.D. (1998). in Evolution of Tertiary Mammals of North America, vol. 1, Terrestrial Carnivores, Ungulates and Ungulatelike Mammals, Volume 1, (Cambridge: Cambridge Univ. Press).
125. Novacek, M.J. (1986). The Skull of Leptictid Insectivorans and the Higher-Level Classification of Eutherian Mammals. Bull. Am. Mus. Nat. Hist. *183*, 1–111.
126. Seiffert, E.R., Simons, E.L., Ryan, T.M., Bown, T.M., and Attia, Y. (2007). New remains of Eocene and Oligocene Afrosoricida (Afrotheria) from

Egypt, with implications for the origin(s) of afrosoricid zalambdodonty. J. Vertebr. Paleontol. *27*, 963–972.

127. Gingerich, P.D. (1993). Early Eocene Teilhardina brandti: Oldest omomyid primate from North America. . Contrib. Mus. Paleontol. Univ. Mich. *28*, 321.
128. Gingerich, P.D. (1986). Early Eocene Cantius-Torresi - Oldest Primate of Modern Aspect from North-America. Nature *319*, 319–321.
129. Seiffert, E.R., Simons, E.L., Clyde, W.C., Rossie, J.B., Attia, Y., Bown, T.M., Chatrath, P., and Mathison, M.E. (2005). Basal anthropoids from Egypt and the antiquity of Africa's higher primate radiation. Science *310*, 300–304.
130. Beard, K.C. (1998). A new genus of Tarsiidae (Mammalia: Primates) from the middle Eocene of Shanxi Province, China, with notes on the historical biogeography of tarsiers. Bull. Carnegie Mus.Nat. Hist. *1998*, 260–277.
131. Tassy, P., and Pickford, M. (1983). Un nouveau mastodonte zygolophodonte (proboscidea, mammalia) dans le mioc ˇcne inférieur d' afrique orientale: Systématique et paléoenvironnement. . Geobios *16*, 53–77.
132. Kappelman, J., Kelley, J., Pilbeam, D., Sheikh, K.A., Ward, S., Anwar, M., Barry, J.C., Brown, B., Hake, P., Johnson, N.M., et al. (1991). The earliest occurrence ofsivapithecus from the middle miocene chinji formation of pakistan. J .Hum. Evol. *21*.
133. Suwa, G., Kono, R.T., Katoh, S., Asfaw, B., and Beyene, Y. (2007). A new species of great ape from the late Miocene epoch in Ethiopia. Nature *448*, 921–924.
134. Begun, D.R. (2010). Miocene Hominids and the Origins of the African Apes and Humans. Annu Rev Anthropol *39*, 67–84.
135. Richmond, B.G., and Jungers, W.L. (2008). Orrorin tugenensis femoral morphology and the evolution of hominin bipedalism. Science *319*, 1662–1665.
136. Ting, S.Y., Tong, Y.S., William, C., and Vertebrata, P. (2011). Asian early Paleogene chronology and mammalian faunal turnover events. Vertebrat. Palasiatic. *49*, 1—28.

137. Asher, R.J., Meng, J., Wible, J.R., McKenna, M.C., Rougier, G.W., Dashzeveg, D., and Novacek, M.J. (2005). Stem Lagomorpha and the antiquity of Glires. Science *307*, 1091–1094.

138. Rose, K.D., DeLeon, V.B., Missiaen, P., Rana, R.S., Sahni, A., Singh, L., and Smith, T. (2008). Early Eocene lagomorph (Mammalia) from Western India and the early diversification of Lagomorpha. Proc. Roy. Soc. B. *275*, 1203–1208.

139. Marivaux, L., Vianey-Liaud, M., and Jaeger, J.J. (2004). High-level phylogeny of early Tertiary rodents: dental evidence. Zool. J. Linn. Soc. *142*, 105–134.

140. Meng, J., Hu, Y.M., and Li, C.K. (2003). The osteology of Rhombomylus (mammalia, glires): Implications for phylogeny and evolution of glires. Bull. Am. Mus. Nat. Hist. , 1–247.

141. Rodrigues, H.G., Marivaux, L., and Vianey-Liaud, M. (2010). Phylogeny and systematic revision of Eocene Cricetidae (Rodentia, Mammalia) from Central and East Asia: on the origin of cricetid rodents. J. Zoological Syst. Evol. Res. *48*, 259–268.

142. Jacobs, L.L., and Flynn, L.J. (2005). Interpreting the Past: Essays on Human, Primate, and Mammal Evolution in Honor of David Pilbeam, chapter of mice ... again: the Siwalik rodent record, murine distribution, and molecular clocks, (Leiden: Brill Academic Publishers).

143. McKenna, M.C., and Bell, S.K. (1997). Classification of Mammals Above the Species Level, (New York: Columbia Univ. Press).

144. Spaulding, M., and Flynn, J.J. (2012). Phylogeny of the Carnivoramorpha: The impact of postcranial characters. J. Syst. Palaeontology *10*, 653–677.

145. Spaulding, M., Flynn, J.J., and Stucky, R.K. (2010). A New Basal Carnivoramorphan (Mammalia) from the 'Bridger B' (Black's Fork Member, Bridger Formation, Bridgerian Nalma, Middle Eocene) of Wyoming, USA. Palaeontology *53*, 815–832.

146. Spaulding, M., O'Leary, M.A., and Gatesy, J. (2009). Relationships of Cetacea (Artiodactyla) Among Mammals: Increased Taxon Sampling Alters Interpretations of Key Fossils and Character Evolution. Plos One *4*.

147. Bajpai, S., and Gingerich, P.D. (1998). A new Eocene archaeocete (Mammalia, Cetacea) from India and the time of origin of whales. Proc. Natl. Acad. Sci. USA *95*, 15464–15468.
148. Fitzgerald, E.M.G. (2010). The morphology and systematics of Mammalodon colliveri (Cetacea: Mysticeti), a toothed mysticete from the Oligocene of Australia. Zool. J. Linn. Soc. *158*, 367–476.
149. Tabuce, R., Antunes, M.T., and Sige, B. (2009). A New Primitive Bat from the Earliest Eocene of Europe. J. Vertebr. Paleontol. *29*, 627–630.
150. McInerney, F.A., and Wing, S.L. (2011). The Paleocene-Eocene Thermal Maximum: A Perturbation of Carbon Cycle, Climate, and Biosphere with Implications for the Future. Annu. Rev. Earth Planet. Sci. *39*, 1–656.
151. Grossnickle, D.M., and Newham, E. (2016). Therian mammals experience an ecomorphological radiation during the Late Cretaceous and selective extinction at the K–Pg boundary. Proc. R. Soc. B. *283*, 20160256.
152. Hall, M.I. (2008). The anatomical relationships between the avian eye, orbit and sclerotic ring: implications for inferring activity patterns in extinct birds. J. Anat. *212*, 781–794.
153. Heesy, C.P., and Hall, M.I. (2010). The nocturnal bottleneck and the evolution of mammalian vision. Brain Behav Evol *75*, 195–203.
154. Cannon, B., and Nedergaard, J. (2004). Brown adipose tissue: function and physiological significance. Physiol Rev *84*, 277–359.
155. Gerkema, M.P., Davies, W.I., Foster, R.G., Menaker, M., and Hut, R.A. (2013). The nocturnal bottleneck and the evolution of activity patterns in mammals. Proc. Biol. Sci. *280*, 20130508.
156. Wu, J., Yonezawa, T., and Kishino, H. (2017). Rates of Molecular Evolution Suggest Natural History of Life History Traits and a Post-K-Pg Nocturnal Bottleneck of Placentals. Curr Biol *27*, 3025–3033 e3025.
157. Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. J. Roy. Stat. Soc. B Met. *58*, 267–288.
158. Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. J. Stat. Softw. *33*, 1–22.

159. Wu, T.T., Chen, Y.F., Hastie, T., Sobel, E., and Lange, K. (2009). Genome-wide association analysis by lasso penalized logistic regression. Bioinformatics *25*, 714–721.
160. Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005). ROCR: visualizing classifier performance in R. Bioinformatics *21*, 3940–3941.
161. Paradis, E., Claude, J., and Strimmer, K. (2004). APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics *20*, 289–290.
162. Rambaut, A. (2011). FigTree. (http://tree.bio.ed.ac.uk/software/figtree/).
163. Okuda, H., Miyata, S., Mori, Y., and Tohyama, M. (2007). Mouse Prickle1 and Prickle2 are expressed in postmitotic neurons and promote neurite outgrowth. FEBS Lett. *581*, 4754–4760.
164. Zhang, C., Mejia, L.A., Huang, J., Valnegri, P., Bennett, E.J., Anckar, J., Jahani-Asl, A., Gallardo, G., Ikeuchi, Y., Yamada, T., et al. (2013). The X-linked intellectual disability protein PHF6 associates with the PAF1 complex and regulates neuronal migration in the mammalian brain. Neuron *78*, 986–993.
165. Tsai, L.Y., Chang, Y.W., Lin, P.Y., Chou, H.J., Liu, T.J., Lee, P.T., Huang, W.H., Tsou, Y.L., and Huang, Y.S. (2013). CPEB4 knockout mice exhibit normal hippocampus-related synaptic plasticity and memory. PLoS One *8*, e84978.
166. Park, H., Yang, J., Kim, R., Li, Y., Lee, Y., Lee, C., Park, J., Lee, D., Kim, H., and Kim, E. (2015). Mice lacking the PSD-95-interacting E3 ligase, Dorfin/Rnf19a, display reduced adult neurogenesis, enhanced long-term potentiation, and impaired contextual fear conditioning. Sci. Rep. *5*, 16410.
167. Saravanan, M., Wuerges, J., Bose, D., McCormack, E.A., Cook, N.J., Zhang, X., and Wigley, D.B. (2012). Interactions between the nucleosome histone core and Arp8 in the INO80 chromatin remodeling complex. Proc. Natl. Acad. Sci. USA *109*, 20883–20888.
168. Serber, D.W., Runge, J.S., Menon, D.U., and Magnuson, T. (2016). The Mouse INO80 Chromatin-Remodeling Complex Is an Essential Meiotic Factor for Spermatogenesis. Biol. Reprod. *94*, 8.

169. Aloia, L., Di Stefano, B., and Di Croce, L. (2013). Polycomb complexes in stem cells and embryonic development. Development *140*, 2525–2534.
170. Stocco, D.M. (2001). StAR protein and the regulation of steroid hormone biosynthesis. Annu. Rev. Physiol. *63*, 193–213.
171. Hernandez-Hernandez, A., Lilienthal, I., Fukuda, N., Galjart, N., and Hoog, C. (2016). CTCF contributes in a critical way to spermatogenesis and male fertility. Sci Rep *6*, 28355.
172. Vardanyan, A., Atanesyan, L., Egli, D., Raja, S.J., Steinmann-Zwicky, M., Renkawitz-Pohl, R., Georgiev, O., and Schaffner, W. (2008). Dumpy-30 family members as determinants of male fertility and interaction partners of metal-responsive transcription factor 1 (MTF-1) in Drosophila. BMC Dev. Biol. *8*, 68.
173. Feild, T.S., Brodribb, T.J., Iglesias, A., Chatelet, D.S., Baresch, A., Upchurch, G.R., Jr., Gomez, B., Mohr, B.A., Coiffard, C., Kvacek, J., et al. (2011). Fossil evidence for Cretaceous escalation in angiosperm leaf vein evolution. Proc. Natl. Acad. Sci. USA *108*, 8363–8366.
174. Ungerfeld, R., and Bielli, A. (2012). Seasonal and social factors affecting reproduction, (Oxford: Encyclopedia of Life Support Systems (EOLSS)).
175. Royer, D.L., Montañez, I.P., Tabor, N.J., and Beerling, D.J. (2004). CO2 as a primary driver of Phanerozoic climate. GSA Today *14*, 4–10.
176. Maor, R., Dayan, T., Ferguson-Gow, H., and Jones, K.E. (2017). Temporal niche expansion in mammals from a nocturnal ancestor after dinosaur extinction. Nat Ecol Evol *1*, 1889-1895.
177. Wilson, G.P., and Riedel, J.A. (2010). New Specimen Reveals Delta Theroidan Affinities of the North American Late Cretaceous Mammal Nanocuris. J Vertebr Paleontol *30*, 872–884.
178. Hu, Y., Meng, J., Wang, Y., and Li, C. (2005). Large Mesozoic mammals fed on young dinosaurs. Nature *433*, 149-152.
179. Muizon, C.D., and Lange-Badré, B. (1997). Carnivorous dental adaptations in tribosphenic mammals and phylogenetic reconstruction. Lethaia *30*, 353–366.
180. Hall, M.I., Kamilar, J.M., and Kirk, E.C. (2012). Eye shape and the nocturnal bottleneck of mammals. Proc Biol Sci *279*, 4962-4968.

181. Ross, C.F., and Kirk, E.C. (2007). Evolution of eye size and shape in primates. J Hum Evol *52*, 294-313.
182. Rowe, N. (1996). Pictorial Guide to the Living Primates. , Volume ISBN 0-9648825-0-7, (East Hampton: Pogonias Press).
183. Bennett, N.C., and Jarvis, J.U.M. (2004). Cryptomys damarensis. Mammalian Species *756*, 1–5.
184. Oosthuizen, M.K., Cooper, H.M., and Bennett, N.C. (2003). Circadian rhythms of locomotor activity in solitary and social species of African mole-rats (family: Bathyergidae). J. Biol. Rhythms *18*, 481–490.
185. Seim, I., Fang, X.D., Xiong, Z.Q., Lobanov, A.V., Huang, Z.Y., Ma, S.M., Feng, Y., Turanov, A.A., Zhu, Y.B., Lenz, T.L., et al. (2013). Genome analysis reveals insights into physiology and longevity of the Brandt's bat Myotis brandtii. Nature Communications *4*.
186. Ferry, N., Edwards, M.G., Gatehouse, J.A., and Gatehouse, A.M. (2004). Plant-insect interactions: molecular approaches to insect resistance. Curr Opin Biotechnol *15*, 155-161.
187. Murata, M., Etoh, T., Itoyama, K., and Tojo, S. (1998). Sudden occurrence of the common cutworm, Spodoptera litura (Lep- idoptera: Noctuidae) in southern Japan during the typhoon season. . Appl. Entomol. Zool. *33*, 419–427.
188. Fu, X., Zhao, X., Xie, B., Ali, A., and Wu, K. (2015). Seasonal Pattern of Spodoptera litura (Lepidoptera: Noctuidae) Migration Across the Bohai Strait in Northern China. J Econ Entomol *108*, 525-538.
189. Murata, M., and Tojo., S. (2004). Flight capability and fatty acid level in triacylglycerol of long-distance migratory adults of the common cutworm, Spodoptera litura. Zool. Sci. *21*, 181–188.
190. Wan, X., Li, J., Kim, M.J., Park, H.C., Kim, S.S., and Kim, I. (2011). DNA sequence variation of the tobacco cutworm, Spodoptera litura (Lepidoptera: Noctuidae), determined by mitochondrial A+T-rich region and nuclear ITS2 sequences. Biochem Genet *49*, 760-787.
191. Cheng, T., Wu, J., Wu, Y., Chilukuri, R.V., Huang, L., Yamamoto, K., Feng, L., Li, W., Chen, Z., Guo, H., et al. (2017). Genomic adaptation to polyphagy

and insecticides in a major East Asian noctuid pest. Nat. Ecol. Evol. *1*, 1747-1756.

192. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al. (2011). The variant call format and VCFtools. Bioinformatics *27*, 2156-2158.
193. Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics *123*, 585-595.
194. Weir, B.S., and Cockerham, C.C. (1984). Estimating F-Statistics for the Analysis of Population Structure. Evolution *38*, 1358-1370.
195. Hudson, R.R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics *18*, 337-338.
196. Benjamini, Y., and Hochberg, Y. (1995). Controling the false discovery rate: a practical and powerful approach to multiple testing. J Royal Statist Soc *57*, 289-300.
197. Wang, B., Yim, S.Y., Lee, J.Y., Liu, J., and Ha, K.J. (2014). Future change of Asian-Australian monsoon under RCP 4.5 anthropogenic warming scenario. Climate Dynamics *42*, 83–100.
198. Raj, A., Stephens, M., and Pritchard, J.K. (2014). fastSTRUCTURE: variational inference of population structure in large SNP data sets. Genetics *197*, 573-589.
199. Gutenkunst, R.N., Hernandez, R.D., Williamson, S.H., and Bustamante, C.D. (2009). Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet *5*, e1000695.
200. Hey, J., and Nielsen, R. (2004). Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of Drosophila pseudoobscura and D. persimilis. Genetics *167*, 747-760.
201. Zhan, S., Zhang, W., Niitepold, K., Hsu, J., Haeger, J.F., Zalucki, M.P., Altizer, S., de Roode, J.C., Reppert, S.M., and Kronforst, M.R. (2014). The genetics of monarch butterfly migration and warning colouration. Nature *514*, 317-321.
202. Haag-Liautard, C., Dorris, M., Maside, X., Macaskill, S., Halligan, D.L., Houle, D., Charlesworth, B., and Keightley, P.D. (2007). Direct estimation of per

nucleotide and genomic deleterious mutation rates in Drosophila. Nature *445*, 82-85.

203. Tojo, S., Ryuda, M., Fukuda, T., Matsunaga, T., Choi, D.R., and Okuda, A. (2013). Overseas migration of the common cutworm, Spodoptera litura (Lepidoptera: Noctuidae), from May to mid-July in East Asia. Appl Entomol Zool *48*, 141-140.

204. Murata, M., and Tojo, S. (2004). Flight capability and fatty acid level in triacylglycerol of long-distance migratory adults of the common cutworm, Spodoptera litura. Zoolog Sci *21*, 181-188.

205. Wang, B., Yim, S.Y., Lee, J.Y., Liu, J., and Ha, K.J. (2014). Future change of Asian-Australian monsoon under RCP 4.5 anthropogenic warming scenario. Clim Dyn *42*, 83-100.