Chinese Science Bulletin

Bayesian mixed models and divergence time estimation of Chinese cavefishes (Cyprinidae: *Sinocyclocheilus*)

LI ZhiQiang[1,2], GUO BaoCheng[1,2], LI JunBing[1,2], HE ShunPing[1†] & CHEN YiYu[1]

[1] Institute of Hydrobiology, Chinese Academy of Sciences, Wuhan 430072, China;
[2] Graduate University of the Chinese Academy of Sciences, Beijing 100049, China

The genus *Sinocyclocheilus* is distributed in Yun-Gui Plateau and its surrounding region only, within more than 10 cave species showing different degrees of degeneration of eyes and pigmentation with wonderful adaptations. To present, published morphological and molecular phylogenetic hypotheses of *Sinocyclocheilus* from prior works are very different and the relationships within the genus are still far from clear. We obtained the sequences of cytochrome *b* (cyt *b*) and NADH dehydrogenase subunit 4 (ND4) of 34 species within *Sinocyclocheilus*, which represent the most dense taxon sampling to date. We performed Bayesian mixed models analyses with this data set. Under this phylogenetic framework, we estimated the divergence times of recovered clades using different methods under relaxed molecular clock. Our phyloegentic results supported the monophyly of *Sinocyclocheilus* and showed that this genus could be subdivided into 6 major clades. In addition, an earlier finding demonstrating the polyphyletic of cave species and the most basal position of *S. jii* was corroborated. Relaxed divergence-time estimation suggested that *Sinocyclocheilus* originated at the late Miocene, about 11 million years ago (Ma), which is older than what have been assumed.

Sinocyclocheilus, phylogeny, relaxed molecular clock

Cave animals have attracted much attention for their distinct troglomorphic characters, including the enlargement of some sensory organs and appendages, and the reduction and/or loss of eyes and pigmentation[1,2]. Cave animals can serve as attractive systems for studying the role of natural selection and adaptation in evolution[1]. At present, many studies utilize a single species, Mexican cavefish (*Astyanax mexicanus*) populations as model organisms to acquire an understanding of how these animals and their troglomorphic features have evolved[3−5]. In teleosts, in addition to Mexican cavefish, a primarily freshwater fish genus *Sinocyclocheilus* (Cyprinidae, Barbinae)—golden-line fish, in which more than 10 species with different degrees of troglomorphic characters, including different degrees of eyes and pigment degeneration and well-developed projection of frontal and parietal bones, can serve as another model system. The genus *Sinocyclocheilus* is endemic to China, and only distributes in karst cave waters and surface rivers or lakes in Yun-Gui Plateau, including eastern Yunnan Province, southern Guizhou Province and northern Guangxi Zhuang Autonomous Region. As of 2004, 51 valid (53 nominal) species have been described in this genus[6].

To better understand the role of natural selection and adaptation in cave animals using *Sinocyclocheilus* as a model system, it is indispensable to be clear about phylogenetic relationships among species within this genus. However, despite considerable studies, the phylogenetic relationships within *Sinocyclocheilus* remain controversial[7−9], especially for the cave species. Morphological

Received January 7, 2008; accepted May 4, 2008
doi: 10.1007/s11434-008-0297-2
†Corresponding author (email: clad@ihb.ac.cn)
Supported by the National Basic Research Program of China (Grant No. 2007CB411600) and the National Natural Science Foundation of China (Grant No. 30530120)

results of Shan and Yue showed that the cave species in their study were diphyletic[7]. However, Wang et al.[8] re-examined the phylogeny within *Sinocyclocheilus* based upon 28 morphological and osteological characters. The resulting cladogram suggested that the cave species formed a monophyletic group. Some species of *Sinocyclocheilus* are highly adaptive to cave environment, so some morphological or osteological characters might not be informative. Therefore, the origin of cave species and phylogeny of *Sinocyclocheilus* need to be tested with independent evidence. Molecular data appears useful in resolving problematic phylogeny because many molecular changes are less related to adaptive evolution than are morphological characters, such as those sites evolved neutrally. The phylogenetic results of Xiao et al.[9] suggested that cave species occurred in five major monophyletic clades based on analysis of mitochondrial DNA cyt *b* and ND4 gene sequences, which resembles little to the results of morphological studies[7,8]. Their molecular data shed more light on the phylogenetic relationships of *Sinocyclocheilus*. Howerver, Xiao et al. selected inappropriate outgroup taxa[10,11] and did not resolve the relationships among the five major clades, and the latter was ascribed to the early rapid radiation events by Xiao et al.[9]. Their divergence time estimation may be tentative since there was no calibration point and the substitution rates were from other fish mitochondrial protein-coding genes[9].

This study takes advantage of Bayesian mixed models. Bayesian mixed inference is an attractive alternative to other model-based methods because it allows separate modeling of gene partitions concurrent with exploring tree space, and at the same time allows extensive variation in model parameters for each partition, thus facilitating combined data analysis. The use of Bayesian mixed inferences has facilitated the exploration of partition-specific evolutionary models and should reduce systematic error, thus providing more accurate posterior probability estimates[12,13]. Here we sequenced cyt *b* and ND4 gene sequences of some species that were not be included in Xiao et al.'s work[9]. The goals of this study were: (i) to re-examine the phylogenetic relationships of *Sinocyclocheilus* using more outgroups; (ii) to date the origin time of the recovered clades using several relaxed molecular clock methods, including nonparametric rate smoothing, penalized likelihood and Bayesian analysis.

1 Materials and methods

1.1 Materials examined

Two mtDNA gene fragments (cyt *b*, ND4), except for those of *S. xunlensis*, *S. yimenensis*, *S. purpureus*, *S. maculates*, *S. qiubeinsis* and ND4 of *Gymnocypris eckloni*, were retrieved from GenBank. The five additional species of *Sinocyclocheilus* were collected from Guangxi Zhuang Autonomous Region and Yunnan Province. Muscle tissues used for DNA extraction in this study were preserved in 95% ethanol, and the specimens were deposited in the Fish Collection of the Institute of Hydrobiology of the Chinese Academy of Sciences. The ingroup samples were identified following the classification system of Zhao[14] because he has given us a revision taxonomic relationships within *Sinocyclocheilus* based upon detailed morphology work. In the light of the molecular phylogenetic relationships with Cyprinidae[15], we used *Danio rerio*, which belong to Danioninae (basal position in Cyprinidae) as a remote outgroup, and included members of two Cyprininae tribes, seven species of Cyprinini (*Cyprinus carpio*, *Carassius carassius*, *Gymnocypris przewalskii*, *Gymnocypris eckloni*, *Barbus barbus*, *Barbodes laticeps*, *Puntius ticto*) and *Labeo batesii* (Labeonini)[15].

1.2 DNA extraction, PCR amplification and sequencing

Total genomic DNA was extracted from ethanol-presvered muscle tissues using phenol/chloroform extraction[16]. Target regions of the mitochondrial DNA were amplified from the total DNA extracts using the polymerase chain reaction (PCR). The complete cyt *b* gene was amplified with primers Glu-F and Thr-R[17]. The complete ND4 gene was amplified by using primers designed in this study: ND4F (5′-AAC AAG ACC TCT GAT TTC GGC TCA-3′) and ND4R (5′-TAG CTT CCA CTT GGA TTT GCA CC-3′). Reaction mixtures contained approximately 100 ng of template DNA, 1 μL of each primer (each 10 μmol/L), 5 μL of 10×reaction buffer, 2 μL dNTPs (each 2.5 mmol/L), and 2.0 U *Taq* DNA polymerase in total 50 μL volume. The PCR amplification profile consisted of an initial denaturation step at 94℃ for 3 min, followed by 35 cycles performed in the following order of denaturation at 94℃ for 30 s; annealing at 54℃ for 30 s (51℃ for ND4); and elongation at 72℃ for 1 min (90 s for ND4); and a final exten-

sion at 72℃ for 8 min. PCR amplification products were fractionated by electrophoresis through 0.8% agarose gels, recovered from the gels and purified with OMEGA (from OMEGA Bio-Tek) purification Kit according to manufacturer's instructions. Complete nucleotide sequences of ND4 gene were determined using purified PCR products with the same primers of PCR and an intermedial primer (5′-GAT GAG TAG GCA ATT AGT GA-3′). In order to get complete cyt *b* gene sequences, purified PCR products were cloned using pMD18-T vector (TAKARA) into *E. coli* Top10 strain, and sequenced using M13 universal sequencing primers. The sequences have been deposited in GenBank (Accession Nos. are listed in Table 1).

Table 1 Details of specimens and sequences used in this study

Taxon and sample	Locality [a)]	GenBank accession No.[b)]	
		cyt b	ND4
Sinocyclocheilus macrocephalus		AY854683, AY854684	AY854740, AY854741
Sinocyclocheilus oxycephalus		AY854685	AY854742
Sinocyclocheilus lunanensis		AY854686	AY854743
Sinocyclocheilus maitianheensis		AY854710	AY854767
Sinocyclocheilus anophthalmus		AY854715	AY854772
Sinocyclocheilus malacopterus		AY854697, AY854699, AY854700	AY854754, AY854756, AY854757
Sinocyclocheilus guishanensis		AY854722	AY854779
Sinocyclocheilus yangzongensis		AY854725, AY854726	AY854782, AY854783
Sinocyclocheilus qujingensis		AY854719	AY854776
Sinocyclocheilus rhinocerous		AY854720	AY854777
Sinocyclocheilus angustiporus		AY854702	AY854759
Sinocyclocheilus hyalinus		AY854721	AY854778
Sinocyclocheilus huaningensis		AY854718	AY854775
Sinocyclocheilus lateristriatus		AY854703 AY854704 AY854706	AY854760 AY854761 AY854763
Sinocyclocheilus tingi		AY854701	AY854758
Sinocyclocheilus grahami		AY854694, AY854695, AY854696	AY854751, AY854752, AY854753
Sinocyclocheilus microphthalmus		AY854687, AY854689	AY854744, AY854746
Sinocyclocheilus lingyunensis		AY854691, AY854693	AY854748, AY854750
Sinocyclocheilus anatirostris		AY854708	AY854765
Sinocyclocheilus tianeensis		AY854717	AY854774
Sinocyclocheilus furcodorsalis		AY854709	AY854766
Sinocyclocheilus halfibindus		AY854723	AY854780
Sinocyclocheilus altishoulderus		AY854724	AY854781
Sinocyclocheilus jii		AY854727, AY854728	AY854784, AY854785
Sinocyclocheilus macrolepis		AY854729	AY854786
Sinocyclocheilus macrophthalmus		AY854733	AY854790
Sinocyclocheilus jiuxuensis		AY854736	AY854793
Sinocyclocheilus bicornutus		AY854730, AY854731	AY854787, AY854788
Sinocyclocheilus cyphotergous		AY854711	AY854768
Sinocyclocheilus multipunctatus		AY854712, AY854713	AY854769, AY854770
Sinocyclocheilus longibarbatus		AY854714	AY854771
Sinocyclocheilus xunlensis	Huanjiang, Guangxi	**EU366187, EU366190**	**EU366184, EU366185**
Sinocyclocheilus yimenensis	Yimen, Yunnan	**EU366191, EU366192**	**EU366179, EU366180**
Sinocyclocheilus purpureus	Luoping, Yunnan	**EU366189, EU366194**	**EU366177, EU366178**
Sinocyclocheilus maculatus	Yiliang, Yunnan	**EU366193**	**EU366183**
Sinocyclocheilus qiubeinsis	Songming, Yunnan	**EU366188, EU366195**	**EU366181, EU366182**
Outgroup			
Cyprinus carpio		X61010	X61010
Carassius carassius		NC_006291	NC_006291
Barbus barbus		NC_008654	NC_008654
Puntius ticto		NC_008658	NC_008658
Labeo batesii		AB238967	AB238967
Gymnocypris przewalskii		AB239595	AB239595
Gymnocypris eckloni		AY463522	**EU366186**
Barbodes laticeps		AY854738	AY854795
Danio rerio		AC024175	AC024175

a) For those sequences of *Sinocyclocheilus* species without localities, please refer to Xiao et al.[9]; b) the Accession Nos. of newly obtained sequences are labeled with bold type.

1.3 Phylogenetic analysis

Multiple alignment of sequences was performed using Clustal X[18] with default parameters, and verified by eye. The DNA sequences were translated into protein sequences to determine whether nonsense mutations or indels were present using MEGA3.1[19]. Measures of nucleotide composition across all taxa were obtained using the program PAUP*4.0b10[20]. A chi-square (χ^2) test of base heterogeneity was calculated for each codon position and for all codon positions, as implemented in PAUP*4.0b10[20].

Bayesian mixed inferences were run in MrBayes 3.1.1[21], employing partition specific modeling. MODELTEST 3.7[22] was used to choose models for each data partition in Bayesian mixed inferences. We preferred to use Bayes information criterion (BIC) for model selection because it has several advantages over widely used hierarchical likelihood- ratio tests (hLRTs). Among the pitfalls of hLRTs are: (i) need for an arbitrary choice between sequential addition or removal of parameters; (ii) election of parameter addition or removal order, and (iii) inability to address model selection uncertainty[23]. We employed different partition strategies and used the Bayes factor to select among *a priori* partition strategies. Because different genes and different codon positions may evolve according to very different rules, we chose eight partition strategies similar to those of Brandley et al.'s[24] (Table 2). All partition strategies were denoted with a capital P and a numerical subscript identifying the number of data partitions. Additional subscript letters identify multiple partitioning strategies that have the same number of partitions but partition the data differently. A Bayes factor (B_{01}) is equal to the ratio of the marginal likelihoods of H_0 and H_1. As these are difficult to compute directly, one can use the harmonic means as a valid approximation[24]. We used the *sump* command in MrBayes to get the log-transformed harmonic means. Therefore, the Bayes factor can be calculated as the ratio of the harmonic means of likelihoods of the two analyses being tested. In this study, 2lnBayes factor$\geqslant$10 was considered as very strong evidence supporting the alternative hypothesis based on hypothesized cut-off values[25]. Preliminary analysis indicated that the default temperature parameter (Temp = 0.2) resulted in infrequent or complete absence of swapping of states between adjacent chains. We adjusted Temp = 0.04 so that the values of adjacent chains swapping frequencies were 10% to 70% (an appropriate target value according to the MrBayes 3.1.1 manual) and used this value for all Bayesian analyses. All Bayesian inferences consisted of 5000000 generations with a random starting tree, default priors and four Markov chains (with Temp = 0.04) sampled every 1000 generations. All the parameters of partitions were unlinked, except for branch lengths and topology. All partitioned analyses accommodated among-partition rate variation (APRV) by using the "prset ratepr = variable" option. We plotted –ln*L* scores against generation time to detect stationarity in MCMC analyses. MCMC convergence was also explored by plotting all parameters in the model against generation time using the *cumulative* function of program AWTY online[26]. To decrease the chance of trapping on local optima, two separate analyses were performed for each partition strategy, mean –ln*L* scores were compared for the two runs, and posterior probability estimates for each clade were compared between the two analyses using scatter-plot created by the *compare* command of the program AWTY online[26]. If posterior probability estimates for clades were similar in both analyses, then the post burn-in trees for the two analyses were combined.

Table 2 Partitioning strategies used in this study.

Partition strategy	Partition identity
P_1	all data combined
P_{2A}	one partition for the 3rd positions of combined data; the other partition for the 1st and 2nd positions of combined data
P_{2B}	cyt *b*; ND4
P_3	codon positions for combined data
P_{4A}	cyt *b*; ND4 codon positions
P_{4B}	cyt *b* codon positons; ND4
P_{4C}	1st and 2nd positions of cyt *b*; 3rd positions of cyt *b*; 1st and 2nd positions of ND4; 3rd positions of ND4
P_6	cyt *b* and ND4 codon position respectively

1.4 Molecular clock test and divergence time estimation

To estimate divergence times, we first performed likelihood-ratio test to investigate if substitution rate constancy fitted the cyt *b* and ND4 dataset respectively. We used PAUP*4.0b10[20] to compare the likelihood of the most likely tree with and without a molecular clock enforced. We also used the program BEAST v1.4.5[27] to investigate the behaviour of rates throughout the tree with the rate for each branch following an uncorrelated relaxed lognormal clock for the combined data. Posterior estimates were

obtained by sampling every 1000 MCMC steps from a total of 3000000 steps under GTR+I+Γ model. The results were analyzed using the software Tracer v1.3[28], with burn-in being 1500000 steps.

There are no known golden-line fish fossils or related geological events, so the tree could not be calibrated internally. Therefore, two calibration points in the outgroups were used to calibrate the tree. The C1 calibration point is based upon the departure of the Upper Yellow River and Qinghai Lake-0.15 Ma[29], used as a minimum age. The C2 calibration point is the recent literature-based separation of the tribes of Cyprinini and Labeonini-15.96 Ma[15]. Because the species within *Tor* were not included in our study, this point was used as a maximum age. The best tree topology selected by Bayes factor was used to conduct molecular dating. Nonparametric rate smoothing (NPRS)[30] and penalized likelihood (PL)[31] implemented in r8s[32] and Bayesian method implemented in Multidivtime[33,34] (available from http://statgen.ncsu.edu/thorne/) were used to estimate divergence times. For NPRS and PL estimation, Powell and TN methods for optimizing the objective function were used respectively; the log penalty function were used for PL estimations; cross-validation was used to select the optimal values of smoothing in PL.

For Bayesian molecular dating, in order to reduce computation time, the tree was pruned to include only a single haplotype for each species. This method allows for evolutionary rate variation among lineages and among genes in muli-gene datasets. We used two partitions partitioned by gene in this analysis. We followed the method of Rutschmann[34] and Yang and Yoder[35] in estimating divergence time. This method is implemented in three steps. First, Baseml program in PAML v.3.14[36] was used to estimate the following model parameters under the F84+Γ model of nucleotide substitution for each partition from the sequence data: nucleotide frequencies, transition/transversion parameter κ, and the shape parameter α of the Γ distribution with four rate categories of rates among sites. Then, these parameters were used to estimate the maximum likelihood of branch lengths and a variance-covariance matrix of branch lengths, using the estbranches program. Next, the species *Danio rerio*, used as the outgroup of all other species, was pruned from the tree, and the output data of estbranches was used as input for the Multidivtime program to estimate the posterior of divergence times, with their standard deviations (SD) and the 95% credibility intervals (CI) via Markov chain Monte Carlo. The Markov chain was run for 2000000 generations and sampled every 100 generations after an initial burn-in period of 200000 cycles. To ensure the convergence of the MCMC algorithm, two independent runs were performed for the same data and same prior distributions with different starting points. Prior gamma distribution on the three parameters of the relaxed clock model were assumed and specified through the mean and SD of the root age, root rate and rate autocorrelation; 18 Ma (SD = 1.8 Ma) for the expected time between tips and root without node time constraints (the emergence of Cyprininae[15]), 0.007 (SD = 0.0007)substitutions per site per million years for the rate at the root node; and 0.08 (SD = 0.08) for the parameter (v) that controls the degree of rate autocorrelation per Myr along the descending branches of the tree. We adjusted the value and SD of v according to the recommendation of Thorne and Kishino[33] so that the value when multiplied by the approximate time from the root to the present the product is between 1 and 2; all remaining parameters were set at the default values.

2 Results

2.1 Characteristics of individual genes

For 9 individuals, we sequenced the complete cyt *b* gene and identified 8 haplotypes. The other complete cyt *b* gene was retrived from GenBank. A total of 1140 positions were analyzed, of which 537 characters were variable, and 434 of these characters were phylogenetically informative (47.1% and 38.1%, respectively). Mean base composition of cyt *b* gene sequences was 29.5%, 27.9%, 27.9%, 14.7% (A, T, C, G, respectively), with a low G content. Nucleotide composition test showed that all taxa analyzed only exhibited heterogeneity at the third codon positions, but not at first and second positions: first position, $\chi^2 = 34.036$, d$f = 180$, $P = 1.000$; second position, $\chi^2 = 3.756$, d$f = 180$, $P = 1.000$; third position, $\chi^2 = 316.585$, d$f = 180$, $P < 0.001$. But for all codon positions, nucleotide composition was homogenous: $\chi^2 = 130.112$, d$f = 180$, $P = 0.998$.

For 9 individuals of *Sinocyclocheilus* and one individual of *Gymnocypris eckloni*, we sequenced the complete ND4 gene (1381 bp) and identified 10 haplotypes. All newly obtained ND4 sequences start with initial codon ATG and end with an incomplete T stop codon. The other ND4 gene sequences were retrieved from GenBank. A total of 1032 positions were analyzed, of

which 484 characters were variable, and 397 of these characters were phylogenetically informative (46.9% and 38.5%, respectively). Mean base composition of ND4 gene sequences was 31.0%, 27.3%, 27.1%, 14.6% (A, T, C, G, respectively), with a low G content. Nucleotide composition test showed that all taxa analyzed exhibited homogeneity at three codon positions: first position, $\chi^2 = 25.028$, d$f = 180$, $P = 1.000$; second position, $\chi^2 = 3.979$, d$f = 180$, $P = 1.000$; third position, $\chi^2 = 167.505$, d$f = 180$, $P = 0.739$. The base compositional homogeneity of cyt *b* and ND4 sequences will not adversely affect phylogenetic inference.

2.2 Data partitioning and phylogeny of *Sinocyclocheilus*

The two genes were concatenated to form a combined data set of 2172 unambiguously aligned sites. Models selected by MODELTEST 3.7 are listed in Table 3. TrN and TrNef cannot be specified in MrBayes, so they were replaced by more generalized GTR.

Table 3 Data partitions, their estimated models of sequence evolution, and total number of characters of each partition in phylogenetic analysis

Partition	Model	Number of characters in partition
ND4	TrN+I+G	1032
ND4 1st codon	TrNef+G	344
ND4 2nd codon	HKY+I+G	344
ND4 3rd codon	TrN+I+G	344
ND4 1st and 2nd codon	HKY+I+G	688
cyt *b*	TrN+I+G	1140
cyt *b* 1st codon	K80+G	380
cyt *b* 2nd codon	HKY+I	380
cyt *b* 3rd codon	GTR+I+G	380
cyt *b* 1st and 2nd codon	HKY+I+G	760
all data	TrN+I+G	2172
all data 1st codon	K80+I+G	724
all data 2nd codon	HKY+I+G	724
all data 3rd codon	GTR+I+G	724
all data 1st and 2nd codon	HKY+I+G	1448

For mixed-model Bayesian analyses, the –ln*L* values reached stationarity by generation 2×10^5; however, the *cumulative* graphs of posterior probabilities produced by AWTY[26] showed that the posterior probabilities of the splits did not get stabilized until approximately 2.5×10^6 (plots no shown) generations. The scatter-plot produced by the *compare* command of the program AWTY online[26] showed no significant variability among independent runs, so the results of both analyses were combined. Given these results, the first 2500 trees of each run were discarded as 'burn-in' and a 50% majority–rule consensus tree that summarizes topology and branch-length information was calculated based upon the remaining 5002 trees.

The estimated Bayes factor comparisons between different partition strategies and mean –ln*L* of each partition strategy are listed in Table 4. As P_3 was the best partitioning strategy and all the mixed-model Bayesian analyses recovered nearly identical topologies, all discussions of Bayesian phylogeny and clade posterior probabilities are limited to this analysis (Figure 1). The monophyly of *Sinocyclocheilus* was confirmed with very strong support (PP = 0.96) and six clades connected deeply in phylogenetic history of this genus were identified. *S. jii* was at the basal split in *Sinocyclocheilus*, the other species of *Sinocyclocheilus* formed the sister group to *S. jii* with very strong support (PP = 0.96) and could be subdivided into five clades, of which clades A, C, D and E were all strongly supported (PP = 1.00 for all clades) and clade B was with moderate support (PP = 0.74). The species of Clade D and E are all cave species, and the latter is the sister group to clades A, B, C and D with strong support (PP = 1.00). Clade A is the most species rich clade in *Sinocyclocheilus*.

Table 4 2lnBayes factor results of comparisons between each partitioning strategy and –ln*L* of each partition strategy[a)]

	Partitioning strategies							
	P_1	P_{2A}	P_{2B}	P_3	P_{4A}	P_{4B}	P_{4C}	P_6
–ln*L*	20445.47	19555.93	20437.85	19458.26	20024.37	19898.88	19542.06	19477.12
P_1	–							
P_{2A}	1761.96	–						
P_{2B}	–12.62	–1774.58	–					
P_3	1966.94	204.98	1979.56	–				
P_{4A}	822.28	–939.68	834.9	–1144.66	–			
P_{4B}	1080.22	–681.74	1092.84	–886.72	257.94	–		
P_{4C}	1788.18	26.22	1800.8	–178.76	965.9	707.96	–	
P_6	1891.90	129.94	1904.52	–75.04	1069.62	811.68	103.72	–

a) The left bottom matrix represents 2lnBayes factors.

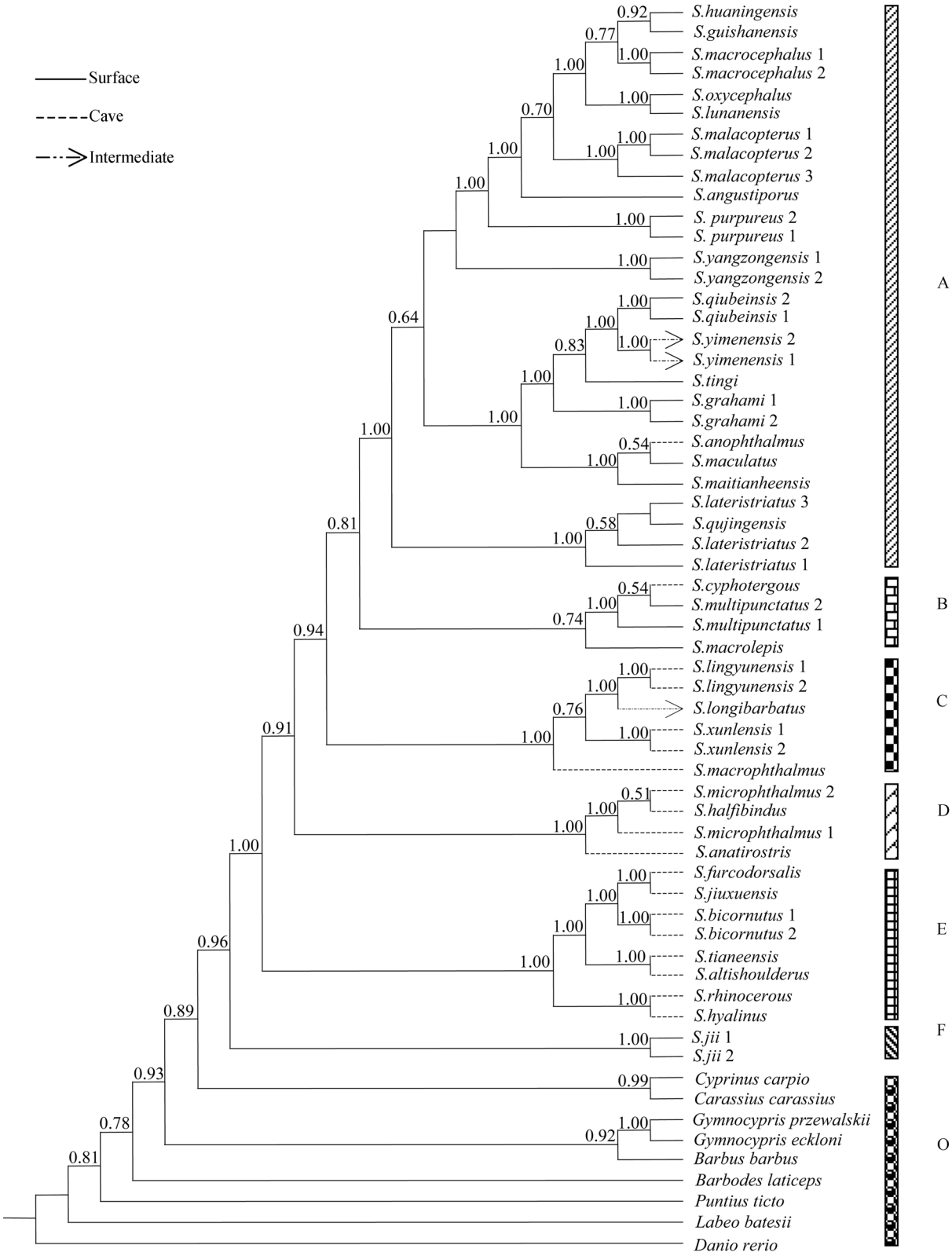

Figure 1 A majority-rule consensus tree of partitioned Bayesian analysis (P_3). Numbers above nodes represent node supports inferred from Bayesian posterior probability analysis. <50% are not shown. The habitat type of ingroup species is from Xiao et al.[9], except for the five newly added species.

2.3 Molecular dating

Significant difference was observed between the likelihood scores of clock and non-clock models for cyt *b* and ND4 dataset: $\chi^2 = 152.33558$, d$f = 59$, $P<<0.001$ for cyt *b*; $\chi^2 = 102.00113$, d$f = 59$, $P<0.001$ for ND4. The standard deviation of parameter ucld was 0.519, coefficient of variation of rates was 0.512, suggesting moderate rate heterogeneity among lineages. The optimal smoothing value selected by cross-validation analyses was 0.0001 for PL.

The tree obtained from P_3 was used to conduct molecular dating. The chronogram obtained with Multidivtime[34] is provided in Figure 2, and the divergence times obtained for ten key nodes within *Sinocyclocheilus* are listed in Table 5. The molecular date estimates produced by three different methods were broadly consistent. The divergence time for *Sinocyclocheilus* was about 11 Ma (with the 95% confidence interval ranging from 8.00 to 13.23 in Multidivtime). Nodes 2, 3, 4 and 5 occurred shortly afterwards, which represents rapid speciation events. The rapid speciation events of recovered clades began in late Miocene, except for the clade C which began to diversify from Mid to Late Pliocene.

Table 5 Divergence dates estimated from molecular data using three different methods

Node	r8s		Multidivtime	
	NPRS	PL	Mean (SD)	95% CI
1	11.51	11.41	10.61(1.35)	(8.00, 13.23)
2	9.36	9.16	8.99(1.34)	(6.42, 11.64)
3	8.79	8.58	8.73(1.33)	(6.19, 11.37)
4	8.09	7.87	8.20(1.32)	(5.68, 10.80)
5	7.68	7.45	7.88(1.32)	(5.37, 10.51)
6	5.63	5.27	5.71(1.20)	(3.50, 8.12)
7	7.12	6.90	7.38(1.31)	(4.89, 10.01)
8	2.97	2.85	3.49(1.08)	(1.69, 5.88)
9	6.29	6.10	6.80(1.36)	(4.26, 9.54)
10	6.44	5.89	6.54(1.30)	(4.15, 9.22)

3 Discussion

3.1 Performance of partitioned Bayesian analyses and partition choice

The use of mixed-model Bayesian analyses does greatly improve mean $-\ln L$ scores. However, simply adding partitions does not necessarily improve $-\ln L$. Identity of each partition is more important, which supported the results of Brandley et al.[24] and Guo and Wang[37]. For example, partitioning the combined data by gene (P_{2B}), which was adopted by Xiao et al.[9], seems worse than no partitioning; partitioning strategies P_{4A}, P_{4B}, P_{4C} and P_6 are all worse than P_3. For gene considered, partitioning the cyt *b* data by codon positions (P_{4B}) has larger effect than that of ND4 (P_{4A}) on the mean $-\ln L$. The former is 125.49 likelihood units better than the latter. Partitioning combined data by codon positions (P_{2A}, P_3, P_{4C} and P_6) has dramatic improvement on $-\ln L$ compared to partitioning by genes (P_{2B}, P_{4A} and P_{4B}), suggesting that the same codon positions of different protein-coding gene on the same strand of the mitochondrial genome have similar evolutionary dynamics. The use of 3 partitions improves mean $-\ln L$ more than any of the alternative strategies according to the Bayes factors. Using 2 ln Bayes factor$\geqslant$10 as a current criterion, all partition strategies are decisively different from each other (Table 4), but the topologies are similar, with only some nodal posterior probabilities changed, especially for some deep nodes. For example, the posterior probabilities of 0.85, 0.96, 0.51 and 0.51 in P_{2B} are replaced by 0.91, 0.94, 0.81 and 0.64 in P_3, respectively. According to Brandley et al.[24] and Guo and Wang[37], this raises the question of whether 2 ln Bayes factor$\geqslant$10 imposes enough penalty to additional partitions.

3.2 Phylogeny of *Sinocyclocheilus*

Our tree contrasts with morphology-based phylogenetic hypotheses[7,8], but is partly congruent with Xiao et al.'s[9] phylogenetic studies. As *Cyprinus carpio* and *Carassius carassius* are more closely related to *Sinocyclocheilus* than *Barbodes laticeps*, *Barbodes laticeps* is not an optimal outgroup. The cave species are neither monophyletic[8] nor diphyletic[7], but a polyphyletic assemblage. This suggests that some morphological or osteological characters may be uninformative. Clades D and E are identical with clades III and IV of Xiao et al.[9], respectively. *S. xunlensis* clusters with the species in clade II of Xiao et al.[9] to form the clade C in our analyses. Clade A is formed by clade V of Xiao et al.[9] and the other four newly added species. Clade B is a major difference between our analyses and Xiao et al.'s[9]. According to Xiao et al.'s result[9], clade I included *S. multipunctatus* and *S. cyphotergous*, while our clade B include *S. macrolepis* in addition to these two species. Although the PP of clade B is not high, all its BP are above 50% in MP and ML tree (unpublished data). The unstable position of *S. macrolepis* in Xiao et al.'s[9] study may be due to the inappropriate outgroup

taxa.

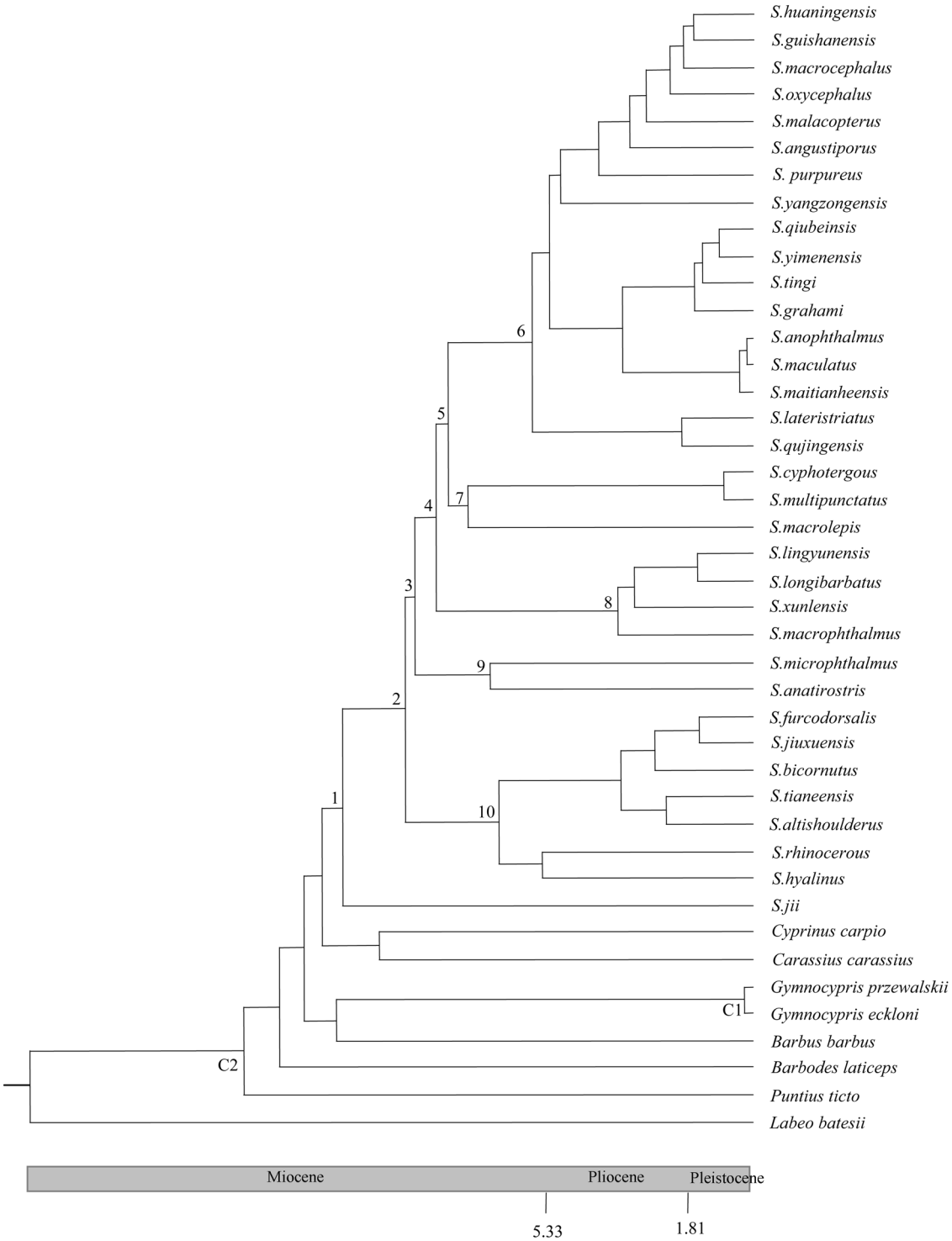

Figure 2 *Sinocyclocheilus* and outgroups chronogram based on the Bayesian relaxed clock analyses for the combined data. Branch lengths are proportional to divergence times. C1 and C2 denote nodes used for calibrating molecular date estimates.

Another major difference between Xiao et al.'s[9] and this study lies in the phylogenetic relationships among the six clades. Our MP tree is (((((A,C),B),D),E),F) and it cannot be rejected by AU, KH and SH test (unpublished data) at the 5% level of significance compared with Bayesian tree (((((A,B),C),D),E),F) in present study.

So, only the positions of clades A, B and C are not resolved. However, the branching orders among clades I, II, III, IV and V were not resolved in Xiao et al's study[9]. So the unresolved positions of III and IV of Xiao et al.[9] cannot be ascribed to rapid diversification in *Sinocyclocheilus*.

The validity of *S. lunanensis* and *S. halfibindus* are contentious[7,9,14]. The *p* distances of cyt *b* between *S. halfibindus* and two haplotypes of *S. microphthalmus* are 0.526% and 0.702% respectively; the *p* distances of ND4 were 0.388% and 0.484% respectively. The *p* distances between *S. lunanensis* and *S. oxycephalus* were 0.175% with cyt *b* and 0.097% with ND4. Given the small genetic distance and similar morphology, we supported the hypothesis that *S. lunanensis* is a synonym of *S. oxycephalus*, and *S. halfibindus* is a synonym of *S. microphthalmus*[7,14].

3.3 Origin and diversification of *Sinocyclocheilus* species

The date estimates presented here is older than Xiao et al.'s[9]. For example, the times for clade D and E at 95% CI are 4.26—9.54, 4.15—9.22 Ma, respectively. However, the estimated times for clades III and IV of Xiao et al. at 95% highest posterior density (HPD) interval were 2.254—8.177, 1.94—6.299 Ma, respectively. The time scales were consistent with the assumption that the ancestors of *Sinocyclocheilus* species were distributed in Yun-gui Plateau in the late Tertiary and had diversified to some degree before the end of the Tertiary[38], but the time of origin and diversification of this genus may be much older than previously assumed[38].

The most recent common ancestor (MRCA) of *Sinocyclocheilus* dated back to about 11 Ma, with the next four divergence (nodes 2, 3, 4, 5 in Figure 2) occurring shortly afterwards. The four divergence times are from 9-8 Ma. This divergence time estimation can be used to distinguish among competing hypotheses about the phases of uplift of the Tibetan Plateau because *Sinocyclocheilus* species have low ability to disperse, limited distribution and the uplifting of Yun-Gui Plateau is closely correlated with the uplifting of Tibet Plateau. One point of view is that rapid uplifting and un-roofing of southern Tibet began about 20 Ma, and further significant increases in altitude of the Tibetan plateau are thought to have occurred about 10—8 Ma[39]. However, an alternative view is that the Tibetan Plateau reached its maximum height before 8 Ma but was then lowered by extensional faulting, with a recent rapid uplift of the plateau occurring about 3.6 Ma accompanied by the largest glacier in the Northern hemisphere[40]. According to the former hypothesis, the onset of east Asian monsoons are about 9—8 Ma[41], and the summer monsoons can bring plenty of rain water to east Asia. So, the molecular date estimates here lend support for the former.

As we know, change in the carbon dioxide concentration of the atmosphere is commonly regarded as a likely forcing mechanism on global climate over geological time because of its large and predictable effect on temperature. Interestingly, since the early Miocene (about 24 Ma), atmospheric CO_2 concentrations (pCO_2) appear to have remained below 500 μL/L and were more stable than before[42], but there was still considerable fluctuation of pCO_2. pCO_2 estimates for the middle and late Miocene indicated that pCO_2 increased from 14 to 9 Ma, and then stabilized at preindustrial values by 9 Ma[42,43], all above 210 μL/L[42]. The age presented here for all of the ten nodes are broadly consistent with the fluctuation of pCO_2. Before the onset of major glaciation in the Northern Hemisphere, Asia was in a period of warmth and continental humidity, and bedrock weathering was probably more intense during periods of warm climate, high pCO_2, continental humidity. So, diversification of *Sinocyclocheilus* may be associated with the change of environment in this period. However, further evidence is needed.

We are grateful to LI ZaiYun, CHEN ZiMing, TAO JinNeng for assistances in collecting specimens and providing tissues. LI WeiXian provided some species. We thank GUO XianGuang and M. C. Brandley for their kind help on the partitioned Bayesian analyses. We are also grateful to WANG XuZhen whose comments greatly improved the presentation of our manuscript.

1 Culver D C, Kane T C, Fong D W. Adaptation and Natural Selection in Caves. Cambridge: Harvard University Press, 1995
2 Romero A. The Biology of Hypogean Fishes. Dordrecht: Kluwer Academic Publishers, 2001
3 Protas M E, Hersey C, Kochanek D, et al. Genetic analysis of cavefish reveals molecular convergence in the evolution of albinism. Nat Genet, 2006, 38(1): 107—111
4 Jeffery W R. Adaptive evolution of eye degeneration in the Mexican blind cavefish. J Hered, 2005, 96(3): 185—196
5 Dowling T E, Martasian D P, Jeffery W R. Evidence for multiple genetic forms with similar eyeless phenotypes in the blind cavefish, *Astyanax mexicanus*. Mol Biol Evol, 2002, 19(4): 446—455

6 Zhao Y H, Watanabe K, Zhang C G. *Sinocyclocheilus donglanensis*, a new cavefish (Teleostei: Cypriniformes) from Guangxi, China. Ichthyol Res, 2006, 53: 121－128

7 Shan X H, Yue P Q. The study on phylogeny of the *Sinocyclocheilus* fishes (Cypriniformes: Cyprinidae: Barbinae). Zool Res (in Chinese), 1994, 15(suppl): 36－44

8 Wang D Z, Chen Y Y, Li X Y. An analysis on the phylogeny of the genus *Sinocyclocheilus*. Acta Academiae Medicinae Zunyi (in Chinese), 1999, 22: 1－6

9 Xiao H, Chen S Y, Liu Z M, et al. Molecular phylogeny of *Sinocyclocheilus* (Cypriniformes: Cyprinidae) inferred from mitochondrial DNA sequences. Mol Phylogenet Evol, 2005, 36(1): 67－77

10 Chen X Y, Yang J X. A systematic revision of "*Barbodes*" fishes in China. Zool Res, 2003, 24: 377－386

11 Kottelat M. Nomenclature of the genera *Barbodes*, *Cyclocheilichthys*, *Rasbora* and *Chonerhinos* (Teleostei: Cyprinidae and Tetraodontidae), with comments on the definition of the first reviser. Raffles Bull Zool, 1999, 47: 591－600

12 Nylander J A A, Ronquist F, Huelsenbeck J P, et al. Bayesian phylogenetic analysis of combined data. Syst Biol, 2004, 53: 47－67

13 Brandley M C, Schmitz A, Reeder T W. Partitioned Bayesian analyses, partition choice, and the phylogenetic relationships of scincid lizards. Syst Biol, 2005, 54:373－390

14 Zhao Y H. An endemic cavefish genus *Sinocyclocheilus* in China－species diversity, systematics, and zoogeography (Cypriniformes: Cyprinidae). Doctor Dissertation (in Chinese). Beijing: Institute of Zoology, Chinese Academy of Sciences, 2006

15 Wang X Z, Li J B, He S P. Molecular evidence for the monophyly of East Asian groups of Cyprinidae (Teleostei: Cypriniformes) derived from the nuclear recombination activating gene 2 sequences. Mol Phylogenet Evol, 2007, 42: 157－170

16 Sambrook J, Fritsch E F, Maniatis T. Molecular Cloning: A Laboratory Manual. 2nd ed. New York: Cold Spring Harbor Laboratory Press, 1989

17 Zardoya R, Doadrio I. Phylogenetic relationships of Iberian cyprinids: Systematic and biogeographical implications. Proc Biol Sci, 1998, 265: 1365－1372

18 Thompson J D, Gibson T J, Plewniak F, et al. The Clustal X windows interface: Flexible strategies for multiple sequences alignment aided by quality analysis tools. Nucleic Acids Res, 1997, 25: 4876－4882

19 Kumar S, Tamura K, Nei M. MEGA3: Integrated software for molecular evolutionary genetics analysis and sequence alignment. Brief Bioinform, 2004, 5: 150－163

20 Swofford D L. PAUP*: Phylogenetic analysis using Parsimony (*and other methods), Version 4. Sunderland, Sinauer Associates, 2002

21 Ronquist F, Huelsenbeck J P. MRBAYES 3: Bayesian phylogenetic inference under mixed models. Bioinformatics, 2003, 19: 1572－1574

22 Posada D, Crandall K A. Modeltest: Testing the model of DNA substitution. Bioinformatics, 1998, 14: 817－818

23 Posada D, Buckley T R. Model selection and model averaging in phylogenetics: Advantages of akaike information criterion and Bayesian approaches over likelihood ratio tests. Syst Biol, 2004, 53: 793－808

24 Newton M A, Raftery A E. Approximate Bayesian inference with the weighted likelihood bootstrap. J R Stat Stoc, 1994, 56: 3－48

25 Kass R, Raftery A. Bayes factor. J Am Stat Assoc, 1995, 90: 773－795

26 Wilgenbusch J C, Warren D L, Swofford D L. AWTY: A system for graphical exploration of MCMC convergence in Bayesian phylogenetic inference. Available at http://ceb.csit.fsu.edu/awty, 2004

27 Drummond A J, Rambaut A. BEAST v1.4.5. University of Oxford, Oxford. Available at http://evolve.zoo.ox.ac.uk/beast/, 2005

28 Rambaut A, Drummond A J. Tracer. Version 1.3, Available at http://evolve.zoo.ox.ac.uk, 2003

29 Li J J, Fang X M, Pan B T, et al. Late Cenozoic intensive uplift of Qinghai-Xizang Plateau and its impacts on environments in surrounding area. Quarter Sci (in Chinese), 2001, 21: 381－391

30 Sanderson M J. A nonparametric approach to estimating divergence times in the absence of rate constancy. J Mol Evol, 1997, 14: 1218－1231

31 Sanderson M J. Estimating absolute rates of molecular evolution and divergence times: A penalized likelihood approach. Mol Biol Evol, 2002, 19: 101－109

32 Sanderson M J. r8s: Inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. Bioinformatics, 2003, 19: 301－302

33 Thorne J L, Kishino H. Divergence time and evolutionary rate estimation with multilocus data. Syst Biol, 2002, 51: 689－702

34 Rutschmann F. Bayesian molecular dating using PAML/multidivtime. A step-by-step manual. University of Zurich, Switzerland. Available at http://www.plant.ch, 2005

35 Yang Z, Yoder A D. Comparison of likelihood and Bayesian methods for estimating divergence times using multiple gene loci and calibration points, with application to a radiation of cute-looking mouse lemur species. Syst Biol, 2003, 52: 705－716

36 Yang Z. PAML: A program package for phylogenetic analysis by maximum likelihood. Comp Appl Biosci, 1997, 13: 555－556

37 Guo X G, Wang Y Z. Partitioned Bayesian analyses, dispersal-vicariance analysis, and the biogeography of Chinese toad-headed lizards (Agamidae: Phrynocephalus): A re-evaluation. Mol Phylogenet Evol, 2007, 45: 643－662

38 Wang D Z, Chen Y Y. The origin and adaptive evolution of the genus *Sinocyclocheilus*. Acta Hydrobiol Sin, 2000, 24: 630－634

39 Harrison T M, Copeland P, Kidd W S F, et al. Raising Tibet. Science, 1992, 255: 1663－1670

40 Li J J, Fang X M, Ma H Z, et al. Geomorphological and environmental evolution in the upper reaches of the Yellow River during the late Cenozoic. Sci China Ser D-Earth Sci, 1996, 39: 380－390

41 An Z, Kutzbach J E, Prell W L, et al. Evolution of Asian monsoons and phased uplift of the Himalaya-Tibetan plateau since Late Miocene times. Nature, 2000, 411: 62－66

42 Pearson P N, Palmer M R. Atmospheric carbon dioxide concentrations over the past 60 million years. Nature, 2000, 406: 695－699

43 Pagani M, Freeman K H, Arthur M A. Late Miocene atmospheric CO_2 concentrations and the expansion of C_4 grasses. Science, 1999, 285: 876－879