Genomic evidence reveals intraspecific divergence in the Hot-spring snake Thermophis baileyi, an endemic reptile of the Qinghai-Tibet Plateau

Chaochao Yan[1], Meng-Huan Song[1], Dechun Jiang[1], Jin-Long Ren[1], Yunyun Lv[2], Hussam Zaher[3], and Jiatang Li[4]

[1]Chengdu Institute of Biology
[2]Chengdu Institute of Biology, Chinese Academy of Sciences,
[3]Universidade de São Paulo
[4]Chengdu Institute of Biology, Chinese Academy of Sciences

March 30, 2022

Abstract

Understanding how and why species evolve often requires knowledge of intraspecific divergence. In this study, we examine intraspecific divergence in the endangered hot spring snake Thermophis baileyi, an endemic species of the Qinghai-Tibet Plateau. Genomic analyses using a hybrid assembly strategy resulted in a revised, high-quality genome. Whole-genome re-sequencing of 31 sampled individuals from 15 sites served to identify drivers of intraspecific divergence, and explore the potential role gene selection plays in divergence. Our analyses resolved three groups, with inter-group admixture occurring in regions of contact. Divergence seems to have occurred during the Pleistocene because of glacial climatic oscillations and geomorphological changes. Highly diverged regions (HDRs) that distinguish the groups most likely owe to gene sorting. Inter-group HDRs involve genes under positive selection that putatively relate functionally to ecological divergence, and especially reproduction. Our findings reveal the need to integrate multiple aspects to distinguish evolutionary processes potentially involved in speciation.

Genomic evidence reveals intraspecific divergence in the Hot-spring snake *Thermophis baileyi*, an endemic reptile of the Qinghai-Tibet Plateau

Chaochao Yan[1]+, Meng-Huan Song[1,2]+, Dechun Jiang[1]+, Jin-Long Ren[1,2], Yunyun Lv[1], Hussam Zaher[4], Jia-Tang Li[1,2,3]*

[1] Chengdu Institute of Biology, Chinese Academy of Sciences, CAS Key Laboratory of Mountain Ecological Restoration and Bioresource Utilization & Ecological Restoration and Biodiversity Conservation Key Laboratory of Sichuan Province, Chengdu, China;

[2] University of Chinese Academy of Sciences, Beijing, China;

[3] Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming, China;

[4] Universidade de São Paulo, Museu de Zoologia, São Paulo, SP, Brazil.

+ These authors contributed equally to this work.

* Correspondence: lijt@cib.ac.cn

Abstract

Understanding how and why species evolve often requires knowledge of intraspecific divergence. In this study, we examine intraspecific divergence in the endangered hot spring snake *Thermophis baileyi* , an endemic species of the Qinghai-Tibet Plateau. Genomic analyses using a hybrid assembly strategy resulted in a revised, high-quality genome. Whole-genome re-sequencing of 31 sampled individuals from 15 sites served to identify drivers of intraspecific divergence, and explore the potential role gene selection plays in divergence. Our analyses resolved three groups, with inter-group admixture occurring in regions of contact. Divergence seems to have occurred during the Pleistocene because of glacial climatic oscillations and geomorphological changes. Highly diverged regions (HDRs) that distinguish the groups most likely owe to gene sorting. Inter-group HDRs involve genes under positive selection that putatively relate functionally to ecological divergence, and especially reproduction. Our findings reveal the need to integrate multiple aspects to distinguish evolutionary processes potentially involved in speciation.

Keywords:

Speciation; population genomics; genome divergence; gene flow; selection

Introduction

A fundamental question in evolutionary biology involves how speciation occurs. An essential part of studying speciation entails identification of what drives the delimitation of species, subspecies, and populations. Traditionally, speciation is the process by which a lack of recombination or gene flow occurs between populations due to allopatry, reproductive isolation, or adaptive divergence (Ortiz-Barrientos, Reiland, Hey, & Noor, 2002). The richness of speciation theories complicates understanding how the many nonexclusive mechanisms drive the process (Turelli, Barton, & Coyne, 2001). Investigations that rely on few molecular markers may not paint a complete picture, but genome-level data can document genetic changes and in doing so potentially identify the drivers of evolution. Most genomic studies of speciation have revealed a pattern of intraspecific divergence (Brawand et al., 2014; Lawniczak et al., 2010; Martin et al., 2013; Riesch et al., 2017). The genetic structure of species usually indicates two fundamental processes: population dynamics subjects to paleoclimate and paleogeological changes and evolutionary ability of species under selection pressure (Avise, 2004). Levels of intraspecific divergence and population genetics dynamics in response to geological and climate changes is essential for understanding species' recent evolution and demography. Such data may also inform how past environmental factors affected species and may predict future responses to impending perturbations.

The diversity and adaptions of species in specific environments has attracted much recent attention (Li et al., 2018; Qu et al., 2020; Wang et al., 2018; Zhou et al., 2016). These factors could result from lineage and gene sorting, recent ecological selection, and/or background selection and selective sweeps. These nonexclusive alternatives are difficult to disentangle, but genome-level analyses may identify the drivers of intraspecific divergence within an endemic reptile on the Qinghai-Tibet Plateau (QTP).

The QTP, which is known as the "third polar region" (Wingfield et al., 2011), comprises one of biodiversity hotspots on the earth. There are a large number of endemic species distributed on the QTP (Mittermeier et al., 2004; Myers, Mittermeier, Mittermeier, da Fonseca, & Kent, 2000). The region has an average elevation of 4,500 m above sea level (a.s.l.). Severe environmental conditions, such as high UV radiation, very cold temperatures, and hypoxia, make the QTP a natural laboratory for investigating how species adapt to their environments. These conditions also seem to drive speciation. Take hypoxia for example, physiological adaptations, including increasing hemoglobin-oxygen affinity, capillarization, and muscle fiber number, have been observed in several species (Cheviron et al., 2014; McCracken et al., 2009; Qu et al., 2020). In the last decade, several investigations have studied high-elevation adaptation using genomes (Ge et al., 2013; Li et al., 2013; Zhu et al., 2018). However, studies focusing exclusively on the evolutionary pattern of species on the QTP and the factors that drive differentiation and speciation are rarely studied, even though high-elevation selection and adaptation contribute to speciation (Wang et al., 2018).

The Tibetan hot-spring snake, *Thermophis baileyi* , is endemic to the QTP and it occurs at elevations exceeding 3,600 m a.s.l. (Fig. 1A). Corresponding to many high-elevation species and populations (Ge et

al., 2013; Peng et al., 2011; Qiu et al., 2012; Simonson et al., 2010), the genomic adaptations of *T. baileyi* were found to be responses to UV radiation and hypoxia (Li et al., 2018). The species, like some others (such as *Laudakia sacra* , *Phrynocephalus theobaldi* ,*Cyrtodactylus tibetanus,* and *Scutiger boulengeri*) (Che, Jiang, Yan, & Zhang, 2020), lives only on the QTP; low elevation populations do not exist (Dorge et al., 2007). The speciation of *T. baileyi* is of great interest because its divergence does not involve low elevation counterparts. This facilitates explorations into the key drivers of its evolution within a high-elevation species. Few studies have investigated divergence of QTP species based on molecular markers (Hofmann, 2012; Hofmann et al., 2014; Liu et al., 2015), and they have few explanations for the drivers of divergence. Therefore, to better investigate the processes of divergence and speciation of the Tibetan hot-spring snake, it is important to evaluate genomic data from additional high-elevation species (Campbellstaton et al., 2017; Ellegren et al., 2012; Lamichhaney et al., 2015), including *T. baileyi* .

Herein, we provide a revised, high-quality reference genome for *T. baileyi* and the resequencing of 31 whole-genomes to: (1) discern its population structure and evolutionary history of *T. baileyi* and test the hypothesis of no historical gene flow among populations since their isolation; (2) examine the fine-scale genomic landscapes of diversity and divergence across populations; (3) clarify the distributions of the species in response to past geological and climate changes, and (4) identify how positive selection has impacted levels of divergence during speciation. To accomplish these goals, we test hypotheses on the potential drivers of intraspecific divergence, including genetic and environmental aspects, and genes potentially related to speciation. These actions disentangle how the multiple evolutionary forces worked in concert to shape intraspecific patterns of divergence, and highlight the need to integrate information of multiple aspects to resolve the speciation for other researches.

Materials and Methods

Reference Genome Sequencing and Assembly

Based on previous studies (Li et al., 2018), we used blood samples acquired from a female *T. baileyi* collected in Yangbajing, Xizang, China to sequence the whole genome. We isolated genomic DNA from the same sample (sample number 1-13), constricted CLR DNA libraries with insert-size lengths ˜30Kb and sequenced these libraries on the PacBio Sequel II sequencer. Firstly, we assembled initial contigs using 280 bp and 500 bp Illumina sequencing data (about 120x) by Platanus (version 1.2.4) (Kajitani et al., 2014) with optimized parameters (-k 31 -t 8 –d 0.3 -m 200). Subsequently, the initial contigs were aligned onto those long reads from PacBio (˜108x) to construct updated contigs using of DBG2OLC (default version) (Ye, Hill, Wu, Ruan, & Ma, 2016). Then, the base errors in the contigs were polished based on sequencing depth of short reads using of Nextpolish (default version) (Hu, Fan, Sun, & Liu, 2019). Assembled genome then was annotated using the same transcriptomic data and followed the same pipeline described previously (Li et al., 2018). BUSCO (version 3.1.0) (Simao, Waterhouse, Ioannidis, Kriventseva, & Zdobnov, 2015; Waterhouse et al., 2018) was used to align the ortholog genes to the reference genome used in previous study (Jiang et al., 2019). Assembly qualities of this genome were evaluated both by orthologous gene alignment and transcriptomic data mapping obtained from previous study (Li et al., 2018).

Sampling, genome resequencing, and mapping

Blood samples representing 31 individuals of *T. baileyi* spanning the geographic ranges of hot-spring snakes were collected from wild populations on the QTP and two individuals of *T. zhaoermii* and *T. shangerila* were used as outgroups (Li et al., 2018) (Table S24). For each individual of *T. baileyi* , genomic DNA was first extracted with a standard protocol, and then sequenced on the Illumina HiSeq2500 PE 150 platforms. The generated raw reads were trimmed 10–15 bp in 5' terminator and then subjected to quality control. Low-quality reads were removed if they met the criteria that a quality < 30 for $> 20\%$ read length. High-quality clean reads were mapped to our newly acquired reference genome before employing BWA-MEM (version 0.7.17) (Li & Durbin, 2009) with default option parameters, and alignment results and marked duplicate reads were sorted by SAMtools (version 1.9) (Li, 2011; Li et al., 2009). Subsequently, the Picard package

(*http://picard.sourceforge.net/*) was used to assign read group information that contains library, lane, and sample identity. The Genome Analysis Toolkit (GATK) (version 4.0.11.0) (Brouard, Schenkel, Marete, & Bissonnette, 2019) was used to execute local realignment of reads to further improve the quality of alignments in region around putative indels.

Filtering and SNP Calling

Before SNP calling, we removed low quality alignments which met either of the following criteria to avoid their effects on SNP detection and subsequent analysis: (i) alignments with a mapping quality score lower than 20; (ii) reads that have multiple best hits; (iii) reads with a flag 4; and (iv) adjust mapping quality lower than 50. We obtained 22.5Mb of SNPs and 5M of indels through SNP-calling. Using bcftools (version 1.9) (Li, 2011; Li et al., 2009) and VCFtools (version 0.1.17) (Danecek et al., 2011) to minimize the influence of sequencing and mapping bias, low quality sites were removed according to the following principles: (i) sites with minor allele frequency (MAF) less than 0.1; (ii) percentage of missing sites to all individuals above 0.75; (iii) bases with a quality below 30; and (iv) site with extremely low (< 4) coverage per individual. These filters obtained a data set comprising ˜1.06 Mb SNPs.

We used the realSFS implemented in analysis of next generation sequencing data (ANGSD) (Korneliussen, Albrechtsen, & Nielsen, 2014) to estimate site frequency spectrum (SFS) at population level. The major/minor state was determined by estimating the minor allele frequency (MAF) based on genotype likelihoods. We used a likelihood ratio test statistic for population allele frequency and P < 1e-6 as criteria for SNP discovery. SNPs would be retained only if they were detected in more than 75% of sampled individuals over all populations.

Population Structure Analyses

We compiled an artificial nucleotide sequence from the nuclear genome data comprising all above SNPs. Meanwhile, we randomly selected 1,000 neutral loci, which are 20Kb away from coding regions in genome. We used trimAl(Capella-Gutierrez, Silla-Martinez, & Gabaldon, 2009) to convert whole-genome SNPs into nexus format. We then constructed phylogenetic network using the NeighborNet method implemented in SplitsTree (version 4.16.2) (Huson & Bryant, 2005) based on the nexus file. Principal component analysis (PCA) was executed on all SNPs by R package LEA (version 2.0) (Frichot & Francois, 2015) and package plink (version 1.90b6.10) (Chang et al., 2015; Purcell et al., 2007). Significance of eigenvectors were determined by a Tracy-Widom test. In addition, we used ADMIXTURE (version 1.3.0) (Alexander, Novembre, & Lange, 2009) and STRUCTURE (version 2.3.4) (Evanno, Regnaut, & Goudet, 2005; Pritchard, Stephens, & Donnelly, 2000) to estimate individual admixture proportions directly from next generation sequencing data. The results revealed that our samples had similar population genetic structure. To gain further insights into relationships, patterns of splits, and admixture between these populations, we produced a maximum likelihood drift tree using the software TreeMix (Pickrell & Pritchard, 2012) and inferred migration events. In addition, Hardy-Weinberg equilibrium test for each SNP site (Wigginton, Cutler, & Abecasis, 2005) was calculated in VCFtools (version 0.1.17) (Danecek et al., 2011) in each, pair, and all populations.

Demographic Analysis

G-PhoCS (Gronau, Hubisz, Gulko, Danko, & Siepel, 2011) was used to infer *T. baileyi* 's entire demographic history of comprising divergence times, ancestral population size, and migration rates, based on 1000 neutral loci. Bayesian Markov Chain Monte Carlo was used to infer parameters to jointly sample model parameters and genealogies of the input loci. Divergence time was referenced from the time estimated by a previous study (Kumar, Stecher, Suleski, & Hedges, 2017). Because G-PhoCS has a restricted capability to characterize complex migration scenarios (Wang et al., 2018), results of D-statistic (Martin, Davey, & Jiggins, 2014) test and TreeMix (Pickrell & Pritchard, 2012) were added to confirm the migration scenarios. Two post-divergence migration bands were added to test if gene flow occurred between groups W and E since they separated. Each Markov chain was executed for 2,000,000 generations while sampling parameter values

every 20,000 iterations. Burn-in and convergence of each run were identified using TRACER 1.7 (Rambaut, Drummond, Xie, Baele, & Suchard, 2018). Runs were repeated for four times to ensure the reliability and stability of the demographic parameters.

Pairwise Sequentially Markovian Coalescent (PSMC) model (Li & Durbin, 2011) was used to reconstruct demographic history with the parameters -N25 -t15 -r5 -b -p '4+25*2+4+6'. Results were scaled using a calculated mutation rate of $4x10^{-9}$ per base per generation (Li et al., 2018) and an assumed generation time of 4 years (Hofmann et al., 2014). A bootstrapping approach with 100 replicates was performed to evaluate the variation in the inferred *Ne* trajectories.

Demographic parameters based on site frequency spectrum were obtained using coalescent simulations implementing the composite likelihood method in fastsimcoal2 software (Excoffier, Dupanloup, Huertasanchez, Sousa, & Foll, 2013). All parameter estimates were ML estimates from 55 independent fastsimcoal2 runs, with 500,000 simulations per likelihood estimation (-n 500,000) and 100 cycles of the likelihood maximization algorithm. Confidence intervals were accessed by parametric bootstrapping, with 100 bootstrap replicates per model. The best model was identified by using the maximum value of likelihoods and Akaike information criterion (AIC); simulated datasets were compared with the observed site frequency spectra to evaluate the fit of the best demographic model.

Genomic Diversity and Divergence

Population genetic summary statistics were direct functions of the site frequency spectrum (SFS). We estimated folded SFS using our own R script referenced and modified from the R scripts (*https://github.com/vsousa/EG_cE3c/tree/master/CustomScripts/Fastsimcoal_VCFtoSFS*). As an exception, Tajima's D was calculated using the software VCFtools (version 0.1.17) (Danecek et al., 2011). Nucleotide diversity (π), $\vartheta\pi$, population scaled mutation rate $\vartheta\omega$, quantified population genetic differentiation (*Fst*), and the number of nucleotide substitutions (*Dxy*) between each pair of groups were calculated using the python script from*https://github.com/simonhmartin/genomics_general*.*Fst* and *Dxy* were calculated using a sliding window approach (50 Kb sliding window and 10 Kb step-size) and the windows were discarded if minSites was less than 500. We measured and compared patterns of linkage disequilibrium (LD) for each group using PopLDdecay (Zhang, Dong, Xu, He, & Yang, 2019). In addition, all genetic parameters above (e.g., Tajima's D, $\vartheta\omega$, and π) were estimated also for different genomic regions (high divergence regions and the residue genomic regions).

Genes under positive selection

To identify potential positively selected genes (PSGs) in the hot-spring snake, seven other species (*Python bivittatus* , *Boa constrictor* , *Ophiophagus hannah* , *Thamnophis elegans* ,*Anolis carolinensis* , *Pogona vitticeps* , *Ophisaurus gracilis*) were retrieved from NCBI and GigaScience databases (Alföldi et al., 2011; Castoe et al., 2011; Daren et al., 2019; Georges et al., 2015; Song et al., 2015; Vonk et al., 2013; https://vgp.github.io/genomeark/Thamnophis_elegans/) (Table S1). Combined with our new genome, single gene families were obtained by OrthoFinder (version 2.2.7) (Emms & Kelly, 2015, 2019). Potential positively selected genes and quickly evolved genes were tested using the codeml script implemented in PAML (version 4.9i) (Yang, 2007), with *T. baileyi* set as the foreground branch and the others as background branches.

Among groups, we applied the population branch statistic (PBS) (Yi et al., 2010) to verify that recent selective sweeps had acted specifically in each group. We compared the levels of polymorphism and divergence among the three groups (W, C, and E) to identify genes under positive selection in each group. We compared the three pairwise *Fst*values between these groups and used the classical transformation by Cavalli-Sforza, $T = -\log(1-Fst)$ to obtain estimates of group divergence time T in units scaled by group size (Yi et al., 2010). The length of the branch leading to group W was estimated as follows:

$$\mathrm{PBSW}=\frac{T(W,\ C)+T(W,\ E)-T(C,\ E)}{2}$$

Calculations for branch lengths leading to groups C and E were similar to the equation for W. We defined he upper 2.5% of each PBS distribution as the high divergence regions (HDRs).

For genes within the HDRs that were considered as positively selected for each group, we collected those falling in the 97.5th percentile of ranked *Fst* to broaden the gene-set among groups. Functional classification of GO categories for these genes was performed using the clusterProfiler in R (Guangchuang Yu, Wang, Han, & He, 2012). Enrichment analyses were performed and Fisher's exact test was used to calculate the statistical significance of enrichment. The adjust$P$ -value cut-off was 0.05. To further investigate the potential key genes for hot-spring snake, we combined all positively selected gene-sets and quickly evolved genes to identify those essential to the evolution of *T. baileyi.*

Results

Sampling and Sequencing

To utilize the advantage of long-reads sequencing technology, we generated ˜15,319,559 PacBio long reads (totaling ˜190 Gb), with N50 read length of 21,712 and max read length of 145,876 (Table S2) to get the high quality genome of a female*T. baileyi* , representing ˜108× coverage of the previously estimated 1.76 Gb genome (Li et al., 2018) (Table S2). A hybrid strategy (combined Illumina reads from previous studies (Li et al., 2018) and PacBio long reads) obtained assemblies of 1.85 Gb, which had contig N50 values of 4 Mb (Table S3). Approximately 95.3% of the genome sequence was contained in the 1618 longest scaffolds (> 644,927), with the largest spanning 25.23 Mb. Gene annotation predicted 22,292 genes in the new reference genome, which was uploaded to Figshare database (accession number XXX).

We re-sequenced the genomes of 31 *T. baileyi* to a mean sequencing depth of over 15×, spanning the geographic distribution of the species (Fig. 1A & Table S4). Retrieved clean data of the outgroup taxa*T. zhaoermii* and *T. shangrila* (Li et al., 2018) both mapped to *T. baileyi* 's reference genome at an average mapping rate of 97% (Table S5). Sequences start with alignment to our new reference genome, and then methods based on uncertainty in the assignment of genotypes were used to estimate population genetic parameters based on the site frequency spectrum. In total, 1.06Mb of high-quality SNPs were identified in *T. baileyi* (Table S6).

Population Structure

Population structure analyses using ADMIXTURE (Fig. 1B) rejected the null hypothesis of unrestricted gene flow and indicated three geographical groups: west (W), central (C), and east (E) (Fig. S1A), even though the best K-value was 2 (Fig. S1B), as it was in STRUCTURE (Figs. S1C & S1D). Group W occurred mainly in western parts of the QTP, group C mainly in the Nyenchen Tanglha Mountains and adjacent regions, and E mainly in the eastern QTP (Fig. 1A). Gene flow between W, E, and C occurred mostly at Nyenchen Tanglha Mountains and adjacent regions (Figs. 1A, S1A & S1C). A phylogenetic network based on whole genomic SNPs also recovered three groups with minimal differentiation between C and E, with W in a deeply diverged state (Fig. 1C). A principal components analysis (PCA) of SNPs depicted a history of group differentiation across all individuals (Figs. 1D, S1E & S1F). The first principal component (PC1; variance explained = 40.18%; Tracy-Widom test, $P < 0.01$) separated W from C and E groups, while the second principal component (PC2; variance explained = 4.975%; Tracy-Widom test, $P < 0.01$) separated E from C, thus confirming the genetically distinct groups (Table S7). The results of Hardy-Weinberg equilibrium test for each SNP site in each, pair, and all of three groups also confirmed three groups (Fig. S1G).

Group W showed more extensive genome-wide linkage disequilibrium (LD) than did C and E (Fig. S2A), indicating a more recent bottleneck. Negative Tajima's D values also rejected the null hypothesis of a stable population size through time owing to a bottleneck in W. The genetic diversity (π) ranged from 0.26 to 0.36 in each group (Table S8). Group C exhibited the highest genetic diversity, which strong asymmetric genetic introgression could partially explain (Table S8). Genome-wide pairwise genetic divergence (*Fst*) between

groups ranged from 0.04 to 0.28, indicating increased divergence times among groups (Fig. S2B and table S8). The mean pairwise nucleotide difference in inter group comparisons (Dxy) (Table S8) was also evident to this pattern.

Demographic History

To investigate the three group's evolutionary history, we first tested the historical changes of effective population size (Ne) for each group sequentially employing the pairwise Markovian Coalescent (PSMC) method (H. Li & Durbin, 2011). Groups W and E underwent a period of decrease in Ne starting ˜0.5 Ma (Figs. 2A, S2C, S2D & S2E), coinciding with the Naynayxungla glaciation (0.5˜0.72 Ma). Group E exhibited the smallest historicalNe and experienced a much more severe bottleneck than groups W and C ˜0.2–0.6 Ma, which was consistent with an origin via a founder effect, while W sustained a comparatively large Nebetween 0.1–0.5 Ma that was followed by a lower Ne compared with other groups. Group C underwent an earlier decrease ˜1 Ma and then an increase ˜0.06–0.1 Ma, followed by the Last Glacial Maximum period (0.03–0.07 Ma) (Zheng, Xu, & Shen, 2002) (Fig. 2A).

The generalized phylogenetic coalescent sampler (G-PhoCS) (Gronau et al., 2011) served to investigate divergence times, gene flow, and effective population sizes. Group W split from the common ancestor of C and E at least 0.325 Ma (95% confidence interval (CI) = 0.26–0.39 Ma), while group C and E diverged at 6.2 kya (95% CI: 4.9–7.5 kya) (Fig. 2B and Table S9). The intensity of gene flow was highest from W to C (> E to C > C to E > C to W > W to E and E to W), which was also indicated through TreeMix analyses and ABBA-BABA statistics (Figs. 2C, S2F, S2G & Tables S9 & S10). As implemented in fastsimcoal2, pairwise joint site frequency spectra were indicated by using a composite likelihood approach (Excoffier et al., 2013). The best model confirmed the divergent pattern of three groups (Table S11).

Genomic Divergence of HDRs

To understand the evolutionary forces influencing genomic divergence's landscape, outlier windows were detected for each pair of groups. The upper 2.5% of each PBS distribution were defined as high divergent regions (HDRs). These were characterized by an excess of low-frequency variants as revealed by more negative Tajima's D, as well as strongly reduced levels of nucleotide diversity (π). We further tested what factors contributed to the formation of HDRs that was based onDxy 's values, we compared Dxy in HDRs to those of outside HDRs in each pair of groups. We revealed substantially elevatedDxy in HDRs between W and E, as well as between W and C (Table S8). Results suggested that haplotypes with HDRs may have become genetically isolated before the outside group-pairs under comparison.

Positively Selected Genes

To test the null hypothesis of no selection by searching for positively selected genes (PSGs), 5,050 single-copy orthologous gene-clusters and 10,452 species-specific mutations in 3599 genes were retrieved from$T.$ $bailey i$ and seven published reptilian genomes. Analyses identified 409 and 758 genes as positively selected genes and rapid evolved genes, respectively (Tables S12 & S13), among which 190 genes were newly identified compared with previous study (Li et al., 2018). These genes, which potentially related to speciation, occurred in different categories, such as $AGR2$, $GATA6$, and $NFIB$being enriched in lung secretory cell differentiation associated and$SESN1$ functioning in response to DNA damage and oxidative stress. These were considered to be related to living at high elevations. Gene ontology (GO) evaluations of the newly identified genes showed many significantly overrepresented functional classes ($P < 0.05$). These classes also included several functional categories associated with high-elevation adaptation, such as DNA repair (GO: 0006281) (Tables S14 & S15).

PBS and Fst tests explored the genetic basis of non-monophyly of groups. Identified genes had high inter-group divergence under recent natural selection. Because group W was most divergent compared with E, analyses focused on positively selected genes within W or between W and E. PBS tests identified 591 genes

under positive selection, and 292 genes were identified by *Fst* . The annotated genome was used to infer the function of all PSGs, and we annotated PSGs within HDRs along the length of the branch leading to W (PBSw) and within genes identified via *Fst* (W vs E). Eight categories were identified: reproduction, immunity, energy metabolism, development, structural materials, metabolic process, genes related to basic biological processes, and uncharacterized genes. In total, 299 and 141 genes associated HDRs and*Fst* gene sets of the basic biological processes, respectively (Tables S16 & S17). Several genes subjected to selection encoded proteins related to reproduction, such as, morphogenesis of the preimplantation embryo (*VEZT*), organization of oocytes (*ERMP1*), spermatozoon maturation and fertility (*Herc4*), and spermiogenesis and oogenesis (*TAF4B*) (Tables S18 & S19). These proteins diverged rapidly and emerged as candidates that were involved in postzygotic isolation across groups. In addition, within PBSw and *Fst* (W vs E) results identified 66 and 38 genes that may associate with high-elevation adaption, respectively (Tables S16 & S17).

Gene ontology enrichment analyses of all PSGs associated with general functional classes. Examples include DNA-binding transcription factor activity (GO:0003700), lipid metabolic process (GO:0006629), and small GTPase mediated signal transduction (GO:0007264) (Tables S20 & S21).

Combined PSGs comparative genomics and population genetics identified genes with the potential to be most essential for the evolution of hot-spring snake (Fig. 3A & Table S22). Genes *FBLN2* ,*IMMT* , *ITB2* , and *VETZ* (Fig. 3B, Tables S16, S17, S22 & S23) were positively selected and quickly evolved genes. They generally related to development, mitochondria, immunity, and reproduction, respectively.

Discussion

Our new reference genome for *T. baileyi* consists of contigs larger than any other previously published snake genome, with contig N50 of 4Mb and BUSCO 95.3% (Table 1). This genome is useful for both population genomics and selection on specific genes, especially when combined with the 31 high-quality re-sequencing genomes. The genome of*T. baileyi* yields insights into adaptations for living at high elevations. Because *T. baileyi* is an endangered reptile, the analyses are important for setting strategies for protection, and establishing legislation.

Population genomic analyses of *T. baileyi* show intraspecific divergence. Although the three identify groups W, C and adjacent regions, and E, the groups exhibit little morphological differentiation. Thus, incipient speciation may be occurring given the divergence between W and the hypothetical ancestor of C and E. Gene flow between C and E suggests a complex history of expansion and colonization with genes from E occurring in all populations of C.

Intraspecific Divergence

Few studies of speciation on the QTP have reconstructed explicitly the demographic history of the group and estimated the intensity of gene flow between divergent groups occurring (Wang et al., 2018). The timing, intensity of gene flow, and changes in effective population size revealed either large or limited intensity of gene flow between pairs of groups. Relatively high gene flow appears to have existed between W and C. Historical demographics and genetic drift best explain the level of divergence between pairs of groups. Numerous HDRs (quantified by PBS) appear scattered across the genome. In these regions, *Dxy* is mostly elevated (Table S8), which is consistent with these regions originating from ancient divergent sorting (Han et al., 2017).

The clear genetic delimitation of the three groups allows the estimation of divergence times (Fig. 2B and Table S4), and all date to late Pleistocene glaciations. The several glacial and interglacial periods appear to have great effects on the topography of rivers and mountains on the QTP (Zheng et al., 2002; S. Zhou, Wang, Wang, & Xu, 2006). The divergence time of group W coincides with the Guxiang (0.13–0.30 Ma) glaciation, one of four major glaciations on the QTP (Zheng et al., 2002; Zhou et al., 2006). Numerous studies of plants and other animals also resolved an association between glaciation and time of intraspecific divergence (Fan et al., 2012; Jin & Liu, 2010; Yu, Zhang, Rao, & Yang, 2013; Zhao et al., 2011). Thus, glaciation seems to be a driver of intraspecific divergence in the QTP.

Fragmentation can drive Pleistocene intraspecific divergence on the QTP, and the results from G-PhoCS and TreeMix analyses unambiguously indicate asymmetrical gene flow. Does apparent gene flow come from gene flow occurred during or following divergence over a continuous period of time or secondary contact and admixture after a long period of isolation? Substantial historical gene flow appears to have occurred between group W and the common ancestor of groups C and E (Fig. 2B), as well as from W to C following their divergence. This implies that divergence had at least two events: initial fragmentation of the species' distribution, and later, possibly during Pleistocene glacial oscillations. Today, the three parapatric groups exhibit mixing of genetic components, indicating ongoing inter-group gene flow (Fig. 1 and 2). The Yarlung Tsagpo drainage system presumably promotes the gene flow from W to C, and it occurs along a narrow path at the intersection of Yadong-Gulu rift and Yarlung Tsagpo (Fig. 4). This pattern was detected previously for the frog *Nanorana parkeri* (Liu et al., 2015; Wang et al., 2018), where intraspecific divergence also occurred between two main groups (Wang et al., 2018). Associated with the population founder events in group E, our simulations indicate limited gene flow occurs between group E and the other two groups. Accordingly, the Nyenchen Tanglha Mountains may restrict gene flow between E and both other groups. Notwithstanding, conclusions about when and how gene flow might have occurred during the groups' divergence were still uncertain. The specific geographical and environmental conditions on the QTP likely also played important roles in divergence.

Living at extremely high-elevations, *T. baileyi* possesses many adaptive traits, including those related to extreme UV radiation and hypoxia (Li et al., 2018). Survival of these snakes requires thermophilic habitats. Since the Cenozoic Indo-Asian collision, dramatic geomorphological changes and five major specific geothermal activities were identified on the QTP (Fig. 4): the Yarlung Tsagpo (YT) geothermal fields, Yarlung Tsagpo Valley (YTV) geothermal fields, Nyenchen Tanglha Mountains Southeast Land (NTMSL) geothermal fields, North Tibetan (NT) geothermal activities, and geothermal activities in Yarlu Tsagpo Great Bend and Shiquan River (Lu, 1989). We hypothesize that geothermal activities promoted dispersal and gene flow during warmer interglacial periods, while driving intraspecific divergence and speciation during glacial times by serving as refugia. This hypothesis best explains a novel evolutionary pattern for species on the QTP.

Positive Selection

Annotated genes under positive selection and showing high intergroup divergence (Tables S12 & S13) relate functionally to different categories (e.g., reproduction, immunity, and development). This may underlie ecological divergence between the groups. Importantly, many genes relate to reproduction: *Herc4* participates in spermatogenesis as its protein can remove the cytoplasmic droplet of the spermatozoon to ensure the spermatozoon maturation and fertility (Rodriguez & Stewart, 2007); *TAF4B* , encodes a protein involved in both spermatogenesis and oogenesis, and it can regulate ovarian granulosa cell lines (Wu, Lu, Hu, & Li, 2005) and its truncating mutations drive recessive azoospermia (Ayhan et al., 2014); and the protein of *Spag6* is important for structural integrity of the central apparatus in the sperm tail and for flagellar motility (Sapiro et al., 2002; Sapiro et al., 2000) (Tables S18 & S19). These proteins, which appear to have diverged rapidly across groups, are candidates involved in postzygotic isolation between groups and the drivers of differentiation and speciation. This is similar to male reproductive genes reported in primates (Wyckoff, Wang, & Wu, 2000), birds (Ellegren et al., 2012), arthropods (Ting, Tsaur, Wu, & Wu, 1998), and amphibians (Malone & Michalak, 2008).

Genes *FBLN2* , *IMMT* , *ITB2* and *VETZ* exhibit rapid evolution and positive selection either inter- or intraspecifically. *VETZ* functionally relates with reproduction because its protein is required for morphogenesis of the preimplantation embryo, and for the implantation (Hyenne et al., 2005; Hyenne et al., 2007). It may play an essential role during speciation by breaking embryo implantation and building postzygotic isolation. Meanwhile, other selected genes also appear to involve speciation because selection and fixation of alleles in a group or species renders them incompatible or inducing other linked alleles incompatible with associated proteins in other groups or species (Johnson, 2010; Wu & Ting, 2004). Thus, the genomic effects of these genes are more important than their functional effects. Notwithstanding, numerous single nucleotide polymorphisms outside of HDRs also contribute to differentiation between groups.

Conclusions

Our results reveal divergence within *T. baileyi* , some gene flow among groups W, C and E. Pleistocene geological changes and climate oscillations is the driver of the species' evolution and demography. HDRs that distinguish the three groups appear to result from ancient isolation and selection. This novel evolutionary pattern not only broadens knowledge about species' evolution on the QTP, but also enriches theories on ecological adaptations. Our results reveal that integrating information on demographic history, gene flow, and selection genes will greatly contribute to distinguish the essence of evolutionary processes potentially involved in speciation. Such analyses are also essential for informed conservation planning.

Acknowledgments

We thank Bao-Lin Zhang for technical assistance about the genomic data analysis; Prof. Robert W. Murphy for improving this manuscript; and Yong Zhang, Xingru Xiang for sample collection. This study was funded by the Strategic Priority Research Program of Chinese Academy of Sciences (XDB31000000); the National Natural Science Foundation of China (31722049, 31772434, 31911530101); the Key Research Program of Frontier Sciences, CAS (QYZDB-SSW-SMC058); the Youth Innovation Promotion Association of CAS (2014338); Southeast Asia Biodiversity Research Institute (Y4ZK111B01); the CAS "Light of West China" Program (2018XBZG_JCTD_001); the Second Tibetan Plateau Scientific Expedition and Research Program (STEP) (2019QZKK0501); the Biological Resources Program, Chinese Academy of Sciences (KFJ-BRP-017-14); and the International Partnership Program of Chinese Academy of Sciences (151751KYSB20190024).

References

Alexander, D., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research, 19* (9), 1655-1664.

Alföldi, J., Di Palma, F., Grabherr, M., Williams, C., Kong, L., Mauceli, E., . . . Lindblad-Toh, K. (2011). The genome of the green anole lizard and a comparative analysis with birds and mammals.*Nature* , 477(7366), 587-591.

Avise, J. C. (2004). *Molecular markers, natural history and evolution* . Sunderland, MA: Sinauer Associates

Ayhan, Ö., Balkan, M., Guven, A., Hazan, R., Atar, M., Tok, A., & Tolun, A. (2014). Truncating mutations in *TAF4B* and*ZMYND15* causing recessive azoospermia. *Journal of Medical Genetics, 51* (4), 239-244.

Brawand, D., Wagner, C. E., Li, Y. I., Malinsky, M., Keller, I., Fan, S., . . . Di Palma, F. (2014). The genomic substrate for adaptive radiation in African cichlid fish. *Nature, 513* (7518), 375-381.

Brouard, J., Schenkel, F. S., Marete, A., & Bissonnette, N. (2019). The GATK joint genotyping workflow is appropriate for calling variants in RNA-seq experiments. *Journal of animal science and biotechnology, 10* (1), 1-6.

Campbellstaton, S. C., Cheviron, Z. A., Rochette, N., Catchen, J. M., Losos, J. B., & Edwards, S. V. (2017). Winter storms drive rapid phenotypic, regulatory, and genomic shifts in the green anole lizard.*Science, 357* (6350), 495-498.

Capella-Gutiérrez, S., Silla-Martínez, J. M., & Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics, 25* (15), 1972-1973.

Castoe, T. A., de Koning, J. A., Hall, K. T., Yokoyama, K. D., Gu, W., Smith, E. N., . . . Pollock, D. D. (2011) Sequencing the genome of the Burmese python (*Python molurus bivittatus*) as a model for studying extreme adaptations in snakes. *Genome Biology* , 12(7), 406.

Chang, C. C., Chow, C. C., Tellier, L. C. A. M., Vattikuti, S., Purcell, S., & Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience, 4* (1), 7-7.

Cheviron, Z. A., Natarajan, C., Projectogarcia, J., Eddy, D. K., Jones, J. M., Carling, M. D., . . . Storz, J. F. (2014). Integrating evolutionary and functional tests of adaptive hypotheses: a case study of altitudinal

differentiation in hemoglobin function in an Andean Sparrow, *Zonotrichia capensis* . *Molecular Biology and Evolution, 31* (11), 2948-2962.

Danecek, P., Auton, A., Abecasis, G. R., Albers, C. A., Banks, E., Depristo, M. A., . . . 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics, 27* (15), 2156-2158.

Card, D. C., Adams, R. H., Schield, D. R., Perry, B. W., Corbin, A. B., Pasquesi, G. I. M., . . . Castoe, T.A. (2019). Genomic Basis of Convergent Island Phenotypes in Boa Constrictors. *Genome Biology and Evolution* 11(11), 3123-3143.

Dorge, T., Hofmann, S., Wangdwei, M., Duoje, L., Solhoy, T., & Miehe, G. (2007). The ecological specialist, *Thermophis baileyi* (Wall, 1907) - new records, distribution and biogeographic conclusions. *The Herpetological Bulletin* , (101) ,8-12.

Ellegren, H., Smeds, L., Burri, R., Olason, P., Backstrom, N., Kawakami, T., . . . Qvarnstrom, A. (2012). The genomic landscape of species divergence in *Ficedula flycatchers* . *Nature, 491* (7426), 756-760.

Emms, D., & Kelly, S. L. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology, 16* (1), 157-157.

Emms, D., & Kelly, S. L. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology, 20* (1), 238.

Evanno, G., Regnaut, S., & Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molcular Ecology, 14* (8), 2611-2620.

Excoffier, L., Dupanloup, I., Huertasanchez, E., Sousa, V., & Foll, M. (2013). Robust demographic inference from genomic and SNP data.*PLOS Genetics, 9* (10).

Fan, Z., Liu, S., Liu, Y., Liao, L., Zhang, X., & Yue, B. (2012). Phylogeography of the South China field mouse (*Apodemus draco*) on the southeastern Tibetan Plateau reveals high genetic diversity and glacial refugia. *PLOS ONE, 7* (5).

Frichot, E., & Francois, O. (2015). LEA: An R package for landscape and ecological association studies. *Methods in Ecology and Evolution, 6* (8), 925-929.

Georges, A., Li, Q., Lian, J., O'Meally, D., Deakin, J., Wang, Z., . . . Zhang, G. (2015). High-coverage sequencing and annotated assembly of the genome of the Australian dragon lizard Pogona vitticeps. *GigaScience* , 4, 45.

Ge, R. L., Cai, Q., Shen, Y. Y., San, A., Ma, L., Zhang, Y., . . . Wang, J. (2013). Draft genome sequence of the Tibetan antelope. *Nature Communications* , *4* , 1858

Gronau, I., Hubisz, M. J., Gulko, B., Danko, C. G., & Siepel, A. (2011). Bayesian inference of ancient human demography from individual genome sequences. *Nature Genetics, 43* (10), 1031-1034.

Han, F., Lamichhaney, S., Grant, B. R., Grant, P. R., Andersson, L., & Webster, M. T. (2017). Gene flow, ancient polymorphism, and ecological adaptation shape the genomic landscape of divergence among Darwin's finches. *Genome Research, 27* (6), 1004-1015.

Hofmann, S. (2012). Population genetic structure and geographic differentiation in the hot spring snake *Thermophis baileyi*(Serpentes, Colubridae): Indications for glacial refuges in southern-central Tibet. *Molecular Phylogenetics and Evolution, 63* (2), 396-406.

Hofmann, S., Kraus, S., Dorge, T., Nothnagel, M., Fritzsche, P., & Miehe, G. (2014). Effects of Pleistocene climatic fluctuations on the phylogeography, demography and population structure of a high-elevation snake species, *Thermophis baileyi* , on the Tibetan Plateau.*Journal of Biogeography, 41* (11).

Hu, J., Fan, J., Sun, Z., & Liu, S. (2019). NextPolish: a fast and efficient genome polishing tool for long read assembly.*Bioinformatics* .

Huson, D. H., & Bryant, D. (2005). Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution, 23* (2), 254-267.

Hyenne, V., Louvet-Vallee, S., El-Amraoui, A., Petit, C., Maro, B., & Simmler, M.-C. (2005). Vezatin, a protein associated to adherens junctions, is required for mouse blastocyst morphogenesis.*Developmental biology, 287* (1), 180-191.

Hyenne, V., Souilhol, C., Cohen-Tannoudji, M., Cereghini, S., Petit, C., Langa, F., . . . Simmler, M.-C. (2007). Conditional knock-out reveals that zygotic vezatin-null mouse embryos die at implantation.*Mechanisms of development, 124* (6), 449-462.

Jiang, W., Lv, Y., Cheng, L., Yang, K., Bian, C., Wang, X., . . . Shi, Q. (2019). Whole-genome sequencing of the giant devil catfish,*Bagarius yarrelli* . *Genome Biology and Evolution, 11* (8), 2071-2077.

Jin, Y.-T., & Liu, N. (2010). Phylogeography of *Phrynocephalus erythrurus* from the Qiangtang Plateau of the Tibetan Plateau.*Molecular Phylogenetics and Evolution, 54* (3), 933-940.

Che, J., Jiang, K., Yan, F., & Zhang, Y.-P. (2020). *Amphibians and Reptiles in Tibet — Diversity and Evolution* . Beijing: Science Press.

Johnson, N. A. (2010). Hybrid incompatibility genes: remnants of a genomic battlefield? *Trends in Genetics, 26* (7), 317-325.

Kajitani, R., Toshimoto, K., Noguchi, H., Toyoda, A., Ogura, Y., Okuno, M., . . . Itoh, T. (2014). Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Research, 24* (8), 1384-1395.

Korneliussen, T. S., Albrechtsen, A., & Nielsen, R. (2014). ANGSD: Analysis of next generation sequencing data. *BMC Bioinformatics, 15* (1), 356-356.

Kumar, S., Stecher, G., Suleski, M., & Hedges, S. B. (2017). TimeTree: A resource for timelines, timetrees, and divergence times.*Molecular Biology and Evolution, 34* (7), 1812-1819.

Lamichhaney, S., Berglund, J., Almen, M. S., Maqbool, K., Grabherr, M., Martinezbarrio, A., . . . Andersson, L. (2015). Evolution of Darwin's finches and their beaks revealed by genome sequencing. *Nature, 518* (7539), 371-375.

Lawniczak, M. K., Emrich, S. J., Holloway, A. K., Regier, A. P., Olson, M., White, B., . . . Besansky, N. J. (2010). Widespread divergence between incipient *Anopheles gambiae* species revealed by whole genome sequences. *Science, 330* (6003), 512-514.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics, 27* (21), 2987-2993.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics, 25* (14), 1754-1760.

Li, H., & Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature, 475* (7357), 493-496.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., . . . 1000 Genome Project Data Processing Subgroup. (2009). The sequence Alignment/Map format and SAMtools. *Bioinformatics, 25* (16), 2078-2079.

Li, J.-T., Gao, Y.-D., Xie, L., Deng, C., Shi, P., Guan, M.-L., . . . Zhang, Y.-P. (2018). Comparative genomic investigation of high-elevation adaptation in ectothermic snakes. *Proceedings of the National Academy of Sciences, 115* (33), 8406-8411.

Li, M., Tian, S., Jin, L., Zhou, G., Li, Y., Zhang, Y., . . . Li, R. (2013). Genomic analyses identify distinct patterns of selection in domesticated pigs and Tibetan wild boars. *Nature Genetics, 45* (12), 1431-1438.

Liu, J., Wang, C., Fu, D., Hu, X., Xie, X., Liu, P., . . . Li, M. (2015). Phylogeography of *Nanorana parkeri* (Anura: Ranidae) and multiple refugia on the Tibetan Plateau revealed by mitochondrial and nuclear DNA. *Scientific Reports, 5* (1), 9857-9857.

Lu, L. Z. (1989). Analysis on the geological background of geothermal activities in Tibet. *Earth Science-Journal of China Universtiy of Geoscience* , *4* , 53-59 (in Chinese with English abstract).

Malone, J. H., & Michalak, P. (2008). Physiological sex predicts hybrid sterility regardless of genotype. *Science, 319* (5859), 59-59.

Martin, S. H., Dasmahapatra, K. K., Nadeau, N. J., Salazar, C., Walters, J. R., Simpson, F., . . . Jiggins, C. D. (2013). Genome-wide evidence for speciation with gene flow in *Heliconius* butterflies.*Genome Research, 23* (11), 1817-1828.

Martin, S. H., Davey, J. W., & Jiggins, C. D. (2014). Evaluating the Use of ABBA–BABA statistics to locate introgressed loci.*Molecular Biology and Evolution, 32* (1), 244-257.

McCracken, K. G., Bulgarella, M., Johnson, K. P., Kuhner, M. K., Trucco, J., Valqui, T. H., . . . Peters, J. L. (2009). Gene flow in the face of countervailing selection: adaptation to high-altitude hypoxia in the βA hemoglobin subunit of yellow-billed pintails in the Andes.*Molecular Biology and Evolution, 26* (4), 815-827.

Mittermeier, R., Gil, P., Hoffman, M., Pilgrim, J., Brooks, T., Mittermeier, C., . . . Ford, H. (2004). Hotspots revisited: earth's biologically richest and most endangered terrestrial ecoregions Cemex.*Mexico City* .

Myers, N., Mittermeier, R. A., Mittermeier, C. G., da Fonseca, G. A., & Kent, J. (2000). Biodiversity hotspots for conservation priorities.*Nature, 403* (6772), 853-858.

Ortíz-Barrientos, D., Reiland, J., Hey, J., & Noor, M. A. F. (2002). Recombination and the divergence of hybridizing species. *Genetica, 116* (2-3), 167-178.

Peng, Y., Yang, Z., Zhang, H., Cui, C., Qi, X., Luo, X., . . . Su, B. (2011). Genetic variations in Tibetan populations and high-altitude adaptation at the Himalayas. *Molecular Biology and Evolution, 28* (2), 1075-1081.

Pickrell, J. K., & Pritchard, J. K. (2012). Inference of population splits and mixtures from genome-wide allele frequency data. *PLOS Genetics, 8* (11).

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics, 155* (2), 945-959.

Purcell, S., Neale, B. M., Toddbrown, K., Thomas, L., Ferreira, M. A. R., Bender, D., . . . Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses.*American Journal of Human Genetics, 81* (3), 559-575.

Qiu, Q., Zhang, G., Ma, T., Qian, W., Wang, J., Ye, Z., . . . Liu, J. (2012). The yak genome and adaptation to life at high altitude.*Nature Genetics, 44* (8), 946-949.

Qu, Y., Chen, C., Xiong, Y., She, H., Zhang, Y., Cheng, Y., . . . Lei, F. (2020). Rapid phenotypic evolution with shallow genomic differentiation during early stages of high elevation adaptation in Eurasian Tree Sparrows. *National Science Review, 7* (1), 113-127.

Rambaut, A., Drummond, A. J., Xie, D., Baele, G., & Suchard, M. A. (2018). Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Systematic Biology, 67* (5), 901-904.

Riesch, R., Muschick, M., Lindtke, D., Villoutreix, R., Comeault, A. A., Farkas, T. E., . . . Nosil, P. (2017). Transitions between phases of genomic differentiation during stick-insect speciation. *Nature Ecology & Evolution, 1* (4), 0082.

Rodriguez, C. I., & Stewart, C. L. (2007). Disruption of the ubiquitin ligase *HERC4* causes defects in spermatozoon maturation and impaired fertility. *Developmental Biology, 312* (2), 501-508.

Sapiro, R., Kostetskii, I., Olds-Clarke, P., Gerton, G. L., Radice, G. L., & Strauss, I. J. (2002). Male infertility, impaired sperm motility, and hydrocephalus in mice deficient in sperm-associated antigen 6.*Molecular and Cellular Biology, 22* (17), 6298-6305.

Sapiro, R., Tarantino, L. M., Velazquez, F., Kiriakidou, M., Hecht, N. B., Bucan, M., & Strauss^{3rd}, J. F. (2000). Sperm antigen 6 is the murine homologue of the Chlamydomonas reinhardtii central apparatus protein encoded by the *PF16* locus.*Biology of Reproduction, 62* (3), 511-518.

Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics, 31* (19), 3210-3212.

Simonson, T. S., Yang, Y., Huff, C. D., Yun, H., Qin, G., Witherspoon, D. J., . . . Ge, R. (2010). Genetic evidence for high-altitude adaptation in Tibet. *Science, 329* (5987), 72-75.

Song, B., Cheng, S., Sun, Y., Zhong, X., Jin, J., Guan, R., . . . Liu, X. (2015). A genome draft of the legless anguid lizard, *Ophisaurus gracilis* . *GigaScience* , 4, 17.

Vonk, F. J., Casewell, N. R., Henkel, C. V., Heimberg, A. M., Jansen, H. J., McCleary, R. J., . . . Richardson, M. K. (2013) The king cobra genome reveals dynamic gene evolution and adaptation in the snake venom system. *Proceedings of the National Academy of Sciences of the United States of America,* 110(51), 20651-20656.

Wu. C.-I. & Ting, C. T. (2004). Genes and speciation. *Nature Reviews Genetics, 5* (2), 114-122.

Ting, C. T., Tsaur, S. C., Wu, M. L., & Wu, C.-I. (1998). A rapidly evolving homeobox at the site of a hybrid sterility gene. *Science, 282* (5393), 1501-1504.

Turelli, M., Barton, N. H., & Coyne, J. A. (2001). Theory and speciation. *Trends in Ecology and Evolution, 16* (7), 330-343.

Wang, G. D., Zhang, B. L., Zhou, W. W., Li, Y. X., Jin, J. Q., Shao, Y., . . . Che, J. (2018). Selection and environmental adaptation along a path to speciation in the Tibetan frog *Nanorana parkeri* .*Proceedings of the National Academy of Sciences of the United States of America, 115* (22), E5056-E5065.

Waterhouse, R. M., Seppey, M., Simao, F. A., Manni, M., Ioannidis, P., Klioutchnikov, G., . . . Zdobnov, E. M. (2018). BUSCO applications from quality assessments to gene prediction and phylogenomics.*Molecular Biology and Evolution, 35* (3), 543-548.

Wigginton, J. E., Cutler, D. J., & Abecasis, G. R. (2005). A note on exact tests of hardy-weinberg equilibrium. *The American Journal of Human Genetics, 76* (5), 887-893.

Wingfield, J. C., Patrick Kelley, J., Angelier, F., Chastel, O., Lei, F., Lynn, S. E., . . . Wang, G. (2011). Organism–environment interactions in a changing world: a mechanistic approach. *Journal of Ornithology, 152* (1), 279-288.

Wu, Y., Lu, Y., Hu, Y., & Li, R. (2005). Cyclic AMP-dependent modification of gonad-selective TAF(II)105 in a human ovarian granulosa cell line. *J Cell Biochem, 96* (4), 751-759.

Wyckoff, G. J., Wang, W., & Wu, C.-I. (2000). Rapid evolution of male reproductive genes in the descent of man. *Nature, 403* (6767), 304-309.

Yang, Z. (2007). PAML 4: Phylogenetic analysis by maximum likelihood.*Molecular Biology and Evolution, 24* (8), 1586-1591.

Ye, C., Hill, C. M., Wu, S., Ruan, J., & Ma, Z. (2016). DBG2OLC: Efficient assembly of large genomes using long erroneous reads of the third generation sequencing technologies. *Scientific Reports, 6* (1), 31900-31900.

Yi, X., Liang, Y., Huertasanchez, E., Jin, X., Cuo, Z. X. P., Pool, J. E., . . . Wang, J. (2010). Sequencing of 50 human exomes reveals adaptation to high altitude. *Science, 329* (5987), 75-78.

Yu, G., Wang, L.-G., Han, Y., & He, Q.-Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics A Journal of Integrative Biology, 16* (5), 284-287.

Yu, G., Zhang, M., Rao, D., & Yang, J. (2013). Effect of Pleistocene climatic oscillations on the phylogeography and demography of red Knobby Newt (*Tylototriton shanjing*) from southwestern China. *PLOS ONE, 8* (2).

Zhang, C., Dong, S.-S., Xu, J.-X., He, W.-M., & Yang, T.-L. (2019). PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. *Bioinformatics, 35* (10), 1786-1788.

Zhao, K., Duan, Z., Peng, Z., Gan, X., Zhang, R., He, S., & Zhao, X. (2011). Phylogeography of the endemic *Gymnocypris chilianensis*(Cyprinidae): Sequential westward colonization followed by allopatric evolution in response to cyclical Pleistocene glaciations on the Tibetan Plateau. *Molecular Phylogenetics and Evolution, 59* (2), 303-310.

Zheng, B., Xu, Q., & Shen, Y. (2002). The relationship between climate change and Quaternary glacial cycles on the Qinghai–Tibetan Plateau: review and speculation. *Quaternary International, 97* , 93-101.

Zhou, S., Wang, X., Wang, J., & Xu, L. (2006). A preliminary study on timing of the oldest Pleistocene glaciation in Qinghai–Tibetan Plateau.*Quaternary International, 154* , 44-51.

Zhou, X., Meng, X., Liu, Z., Chang, J., Wang, B., Li, M., . . . Li, M. (2016). Population genomics reveals low genetic diversity and adaptation to hypoxia in Snub-Nosed Monkeys. *Molecular Biology and Evolution, 33* (10), 2670-2681.

Zhu, X., Guan, Y., Signore, A. V., Natarajan, C., Dubay, S. G., Cheng, Y., . . . Storz, J. F. (2018). Divergent and parallel routes of biochemical adaptation in high-altitude passerine birds from the Qinghai-Tibet Plateau. *Proceedings of the National Academy of Sciences of the United States of America, 115* (8), 1865-1870.

Data and materials availability

The raw genome sequencing data and raw genome resequencing data reported this study have been deposited in the China National GeneBank DataBase (CNGBdb), with accession number XXXX (raw genome sequencing data) and XXXX-XXXX (raw genome resequencing data) that are available at http://XXXX. We provide all data needed to evaluate our conclusions in this study in the paper or/and the Supplemental Information. Additional data related to this study may available from the authors.

Author Contributions

J.-T.L. conceived and supervised the project. J.-L.R. performed the sample collection. Y.L. assembled the reference genome. C.Y. and M.-H.S. designed and performed the data analyses. M.-H.S., D.J. and C.Y. wrote the initial manuscript draft. J.-T.L. and H. Z. worked on the approval of the manuscript. All authors read and approved the final version.

Declaration of Interests

The authors declare no competing interests.

Tables and Figures legends

Table 1. Quality metrics for *T. baileyi* genome compared to other published snake genomes.

Fig. 1. Population genetic structure based on whole genomes of hot-spring snake. (**A**) The geographic distribution of *T. baileyi* with different sample sites and genetic structure. Blue, Western group; Yellow, Central group; Red, Eastern group (*T. baileyi* 's photo is provided by Jin-Long Ren). The sectors of circles indicate the genotype frequency of the population at that location. (**B**) ADMIXTURE results based

on whole-genome SNPs with K = 2 to 5. (**C**) Phylogenetic network based on whole-genome SNPs. (**D**) PCA results based on whole-genome SNPs.

Fig. 2. Demographic history of *T. baileyi* species.(**A**) Changed effective population size (*Ne*) through time inferred by PSMC with a generation time (g) of 4 years and a mutation rate (μ) of 4×10^{-9} per site per generation (**B**) The divergence time, demographic history, and migration of *T. baileyi*using G-PhoCS simulation. Arrows indicate the direction of gene flow, and associated numbers indicate the estimates of total migration rates. (**C**) The maximum likelihood tree constructed by TreeMix analysis with gene flow from W to C. s.e., standard error.

Fig. 3. Collection of genes selected based on a combination analysis of inter-species and inner-species for *T. baileyi*.(**A**) Four gene sets were combined to select the potential essential genes for *T. baileyi* . Fstwe, upper 2.5% of *Fst*value (W vs E); PBSw, HDRs of PBSw; PS, positive selection genes based on comparative genomics; QE, quickly evolved genes based on comparative genomics. (**B**) Four genes were found selected either inter-species or inner-species. Arrows indicate the variation sites, red, $P < 0.05$, purple, $P < 0.01$.

Fig. 4. Correlation between distribution of *T. baileyi*and geothermal fields. Five geothermal fields were existed in the QTP, they may shape the distribution of *T. baileyi* as different roles in glaciations and interglacials during the Pleistocene.

Supplemental Information

Fig. S1. Population structure of hot-spring snake. (**A**) ADMIXTURE cluster membership for hot-spring snake. (*) Suspected introgression involved individuals in W and E groups; (**B**) ADMIXTURE cross validation errors for hot-spring snake. Red dot shows the preferable K value for hot-spring snake. **(C**) STRUCTURE cluster membership for hot-spring snake, showing the similar result with ADMIXTURE. (**D**) STRUCTURE cross validation errors for hot-spring snake. Red circle shows the preferable K value for hot-spring snake. (**E**) 3D PCA plot of hot-spring snake whole-genome SNPs data. Three groups were more clearly strong divided. (**F**) PCA plot of hot-spring snake whole-genome SNPs data using plink (v1.90b6), with PC1 and PC2 explaining 24% and 6.6% the variations, respectively. (**G**) H-W tests using all SNPs for each, pair, and all three groups of hot-spring snake. The distribution of P values indicates that for each group, most sites follows H-W balance.

Fig. S2. Divergence situation, introgression, and demographic history of three groups in this study. (**A**) LD decay trajectories of three groups in this study. The linkage disequilibrium is significant higher in W than that of C and E groups. (**B**)*Fst* and PBS distribution of three groups in this study. The difference of *Fst* and PBS values indicated a great degree for differentiation of W. (**C**) Changed effective population size (Ne) through time inferred by PSMC for all W individuals. (**D**) Changed effective population size (Ne) through time inferred by PSMC for all C individuals. (**E**) Changed effective population size (Ne) through time inferred by PSMC for all E individuals. (**F**) Introgression matrix results for different edges between groups using TreeMix. (**G**) Distribution of the D statistics (ABBA-BABA Statistics) results. The results showed that the introgression from group W to C is great more than the values from W to E.

Table S1. Summary of sequence data of seven reptile genomes for comparative genomics retrieved in this study.

Table S2. Summary information of PacBio sequencing data of hot-spring snake in this study.

Table S3. Summary information of new reference genome of hot-spring snake in this study.

Table S4. Sampling information of hot-spring snake in this study.

Table S5. Summary of sequenced data for data quality and mapping efficient of hot-spring snake in this study.

Table S6. SNP and Indel summary of all samples of hot-spring snake in this study.

Table S7. Tarcy-Widom statistics of the top nine eigenvalues from PCA analysis of hot-spring snake.

Table S8. Summary statistics of group genomic parameters for comparisons of HDRs regions with outside HDRs regions and with whole-genome regions for all groups involved.

Table S9. Demographic parameters inferred by G-PhoCS of hot-spring snake.

Table S10. TreeMix involved assignment groups and individuals' information of hot-spring snake in this study.

Table S11. Information of different divergent model simulated by fastsimcoal2.

Other Supplementary Material for this manuscript includes the following:

Table S12 (Microsoft Excel format). List of genes found to be under positive selection in hot-spring snake (Positive selection was estimated by hot-spring snake and other seven reptile species).

Table S13 (Microsoft Excel format). List of genes found to be under quickly evolved in hot-spring snake (Quick evolution was estimated by hot-spring snake and other seven reptile species).

Table S14 (Microsoft Excel format). Ontology analysis of genes found to be under positive selection in hot-spring snake (Positive selection was estimated by hot-spring snake and other seven reptile species).

Table S15 (Microsoft Excel format). Gene Ontology analysis of genes found to be under quickly evolved in hot-spring snake (Quick evolution was estimated by hot-spring snake and other seven reptile species).

Table S16 (Microsoft Excel format). List of genes found to be rapid evolved in group W of hot-spring snake (PBSw).

Table S17 (Microsoft Excel format). List of genes found to be under directional selection in the hot-spring snake groups. (FST was estimated between Western and Eastern groups of hot-spring snake).

Table S18 (Microsoft Excel format). List of genes found to be under rapidly evolved in the hot-spring snake group W, which were associated with reproductive function.

Table S19 (Microsoft Excel format). List of genes found to be under directional selection (*Fst* W vs E) in the hot-spring snake, which were associated with reproductive function.

Table S20 (Microsoft Excel format). Ontology analysis of genes found to be rapid evolved in group W of hot-spring snake (PBSw).

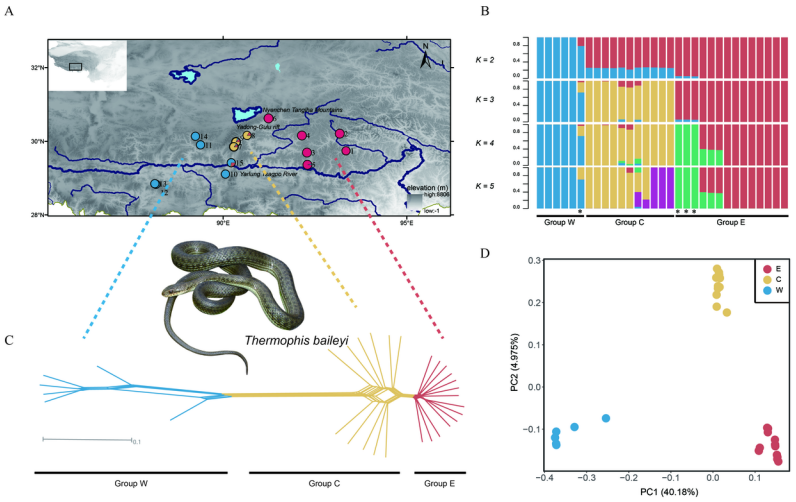
Table S21 (Microsoft Excel format). Gene Ontology analysis of genes found to be under directional selection in the hot-spring snakes' groups (*Fst* W vs E).

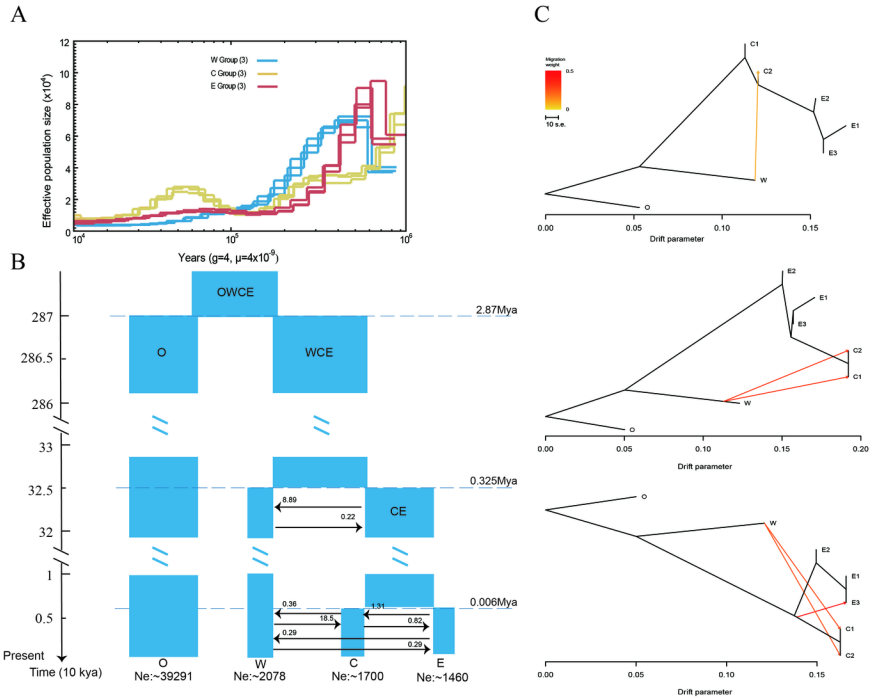
Table S22 (Microsoft Excel format). Summary of genes located in regions that strongly differentiated between groups of hot-spring snake and positively selected as well as quickly evolved in hot-spring snake.

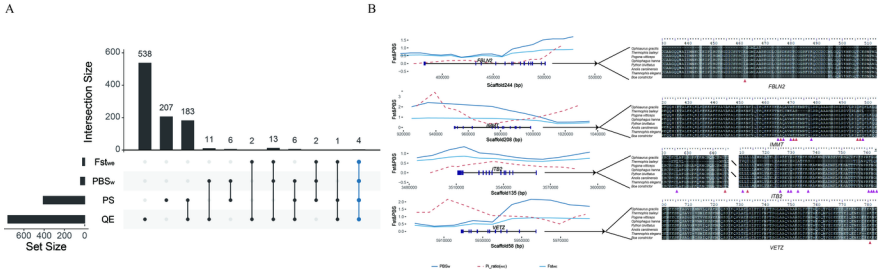
Table S23 (Microsoft Excel format). GO analysis of genes located in regions that strongly differentiated between groups of hot-spring snake and positively selected as well as quickly evolved in hot-spring snake.

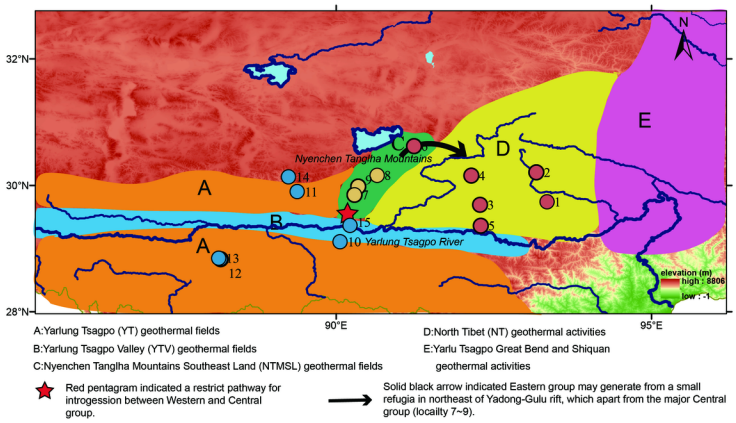
Hosted file

`Table 1.docx` available at https://authorea.com/users/468636/articles/562079-genomic-evidence-reveals-intraspecific-divergence-in-the-hot-spring-snake-thermophis-baileyi-an-endemic-reptile-of-the-qinghai-tibet-plateau









A:Yarlung Tsagpo (YT) geothermal fields

B:Yarlung Tsagpo Valley (YTV) geothermal fields

C:Nyenchen Tanglha Mountains Southeast Land (NTMSL) geothermal fields

D:North Tibet (NT) geothermal activities

E:Yarlu Tsagpo Great Bend and Shiquan geothermal activities

Red pentagram indicated a restrict pathway for introgession between Western and Central group.

Solid black arrow indicated Eastern group may generate from a small refugia in northeast of Yadong-Gulu rift, which apart from the major Central group (locailty 7~9).