Morphological species and discordant mtDNA: A genomic analysis of *Phrynocephalus* lizard lineages on the Qinghai-Tibetan Plateau (pre-print version)

Running title: Morphological species and genomic divergence on the QTP

Yuanting Jin[1] and *Richard P. Brown[1,2]

[1]College of Life Sciences, China Jiliang University, Hangzhou, 730000, P. R. China

[2]School of Natural Sciences and Psychology, Liverpool John Moores University, Liverpool, L3 3AF, UK.

*Please list both author for correspondence R.P.Brown@ljmu.ac.uk orjinyuanting@126.com

Declarations of interest: none

Abstract

Many species have been established on the basis of morphology, with markers such as mtDNA used to confirm the existence of independent historical lineages. Discordance between morphology and gene trees makes this less straightforward. Genotyping by sequencing (GBS) was used to analyse general genomic divergence across two recognized high altitude lizard species found in the eastern Qinghai-Tibetan Plateau. One of the species (*Phyrnocephalus guinanensis*) is found on a large area of sand dune habitat and distinguished from the other (*P. putjatia*) by morphology. We found that the primary pattern of genomic divergence is discordant with these morphological species: northern *P. putjatia* populations from around the large saline Qinghai lake are genomically distinct from *P. putjatia* and *P. guinanensis* populations located south of the Qinghai South and Riyue Mountains. Two competing historical scenarios were assessed using approximate Bayesian computation which unequivocally favoured a split between populations separated by the Qinghai South and Riyue mountains over a split between morphological species. The findings indicate that historical vicariance due to geographical features underpins the phylogenetic split rather than ecology-mediated divergence between sand dune and non-sand areas which i) is predicted by the mtDNA tree (showing the utility of this marker in species delimitation) and ii) demonstrates the unsuitability of the morphology-based taxonomy (indicating that large morphological differences do not always reflect historical lineages). In addition, we found a clear signal of isolation-by-distance around the periphery of Qinghai lake which suggests: i) a high level of resolution by GBS for detecting local divergence and ii) restricted gene flow over relatively short geographic distances. Overall, we show how morphological variation can mislead taxonomic conclusions and the utility of GBS for resolving these issues.

Keywords

Approximate Bayesian computation; genomics; Qinghai-Tibetan Plateau; morphological species; speciation; taxonomy

Introduction

Characterisation of general genomic divergence should provide significant advantages over analyses of a small number of loci for studies of recent lineages. This may be most important when the primary aim is to reveal population splits. First, recent splits generally lead to low levels of divergence in nuclear markers and so well-supported statistical inferences may require the large amounts of sequence that are provided by genomic approaches. Second, although all loci have been subject to the same splitting process, conflicting patterns between loci are likely due to differential introgression, local selection and incomplete lineage sorting (Pamilo and Nei, 1988; Moore, 1995; Takahashi *et al.*, 2001; McCracken and Sorenson, 2005). Misinterpretations caused by this discordance are less likely if large numbers of loci are used.

There have been relatively few examples of multispecies coalescent analyses using SNP data from reduced representation genomic libraries, possibly due to a lack of corresponding statistical approaches. Restriction Associated DNA sequencing (RADseq: Baird *et al.*, 2008; Hohenlohe *et al.*, 2010) and Genotyping By Sequencing (GBS: Elshire *et al.*, 2011) offer genome-wide SNPs from non-model species and have already been employed in species/population-level studies (e.g., Brown et al., 2016; Eaton and Ree, 2013; Combosh and Vollmer 2015). Analyses of these data within a coalescent framework will make major contributions to populations genetics. Bryant et al. (2012) devised a suitable method for estimating species trees from SNPs, that has been implemented in the Bayesian package BEAST (Drummond et al., 2014) and subsequently used in a multispecies coalescent analysis of species delimitation (e.g., Leaché *et al.*, 2014). Coalescent-based analyses can also be undertaken using the approximate Bayesian computation (ABC) approach implemented within the program DIYABC (Cornuet et al., 2014). This now implements the simulation algorithm of Hudson (2002) enabling analyses of SNPs, although to date this does not appear to be very widely applied (but see Schafer et al., 2015). Unlike other related methods used in evolution, such as

Bayesian Markov chain Monte Carlo approaches, the likelihood function is not calculated in ABC but rather approximated by means of simulations. Hence ABC is most useful when the likelihood is too difficult or costly to compute.

Applications of these statistical approaches to RADseq/GBS data have great potential for assessing species-level taxonomy. Morphology frequently forms the basis for many taxonomic designations and it is not unusual for morphological characters to show discordance with regard to some loci, particularly mtDNA (Leaché *et al.*, 2009; Bauer *et al.*, 2010; Wiens and Penkrot, 2002; Barley *et al.*, 2013). Appropriate analyses of large numbers of cross-genomic SNPs could provide definitive assessments of whether or not previously-described morphological species represent groups that are following independent evolutionary trajectories.

In this study we use a genomic approach to assess the interesting patterns of morphological and genetic divergence found in toad-headed lizards (*Phrynocephalus*) on the Qinghai-Tibetan Plateau (QTP), China. Three species are known from our study area: *P. guinanensis*, *P. putjatia* and *P. vlangalii* (Ji *et al.*, 2009; Jin *et al.*, 2014; Jin *et al.*, 2018). It should be noted that several synonyms of *P. putjatia* have been used. The original species description was *P. putjatai* Bedriaga 1909, but several versions of this species name have been used, including "*P. putjatae*" in the study that described the new species *P. guinanensis* (Ji et al., 2009). However, *P. putjatia* is currently the most widely-used spelling and will be used here. Here, we aimed to focus on *P. guinanensis* and *P. putjatia* which reach altitudes of around 3500m within the Qinghai and neighbouring Gansu provinces (Fig. 1). The region includes the large saline Qinghai lake (4317 km^2) and comprises potential physical barriers which divide northern and southern areas, namely the South Qinghai Lake and Riyue Mountains and the Yellow river (including the Longyangxia gorge/reservoir).

The species *P. guinanensis* was described from sand dune habitat in Guinan County and the designation supported by considerable divergence in morphology, coloration and ecological

preferences (Ji *et al.*, 2009). Previous analyses have not supported the hypothesis that this morphological split reflected distinct evolutionary lineages (Jin *et al.*, 2014), although whether or not these represent distinct species requires more extensive genomic analysis. Jin *et al.* (2014) showed that mtDNA haplotypes were clustered in two phylogeographical groups, one corresponding to individuals from the Qinghai lake basin (which in turn are substructured between the east and the west of the lake) and the other to more southern areas. This did not correspond to the geographical variation in morphology. A nuclear locus (*RAG-1*) showed slightly discordant but similar phylogeographical patterns, but with relatively low resolution (Jin *et al.*, 2014). Nevertheless, use of only two markers meant that relationships between historical lineages and morphology were uncertain and so the finding was tentative.

The primary aim of the present study was to use genome-wide SNP data to analyze the genomic diversity within *P. putjatia*/*P. guinanensis*, and definitively examine the divergence within/between populations ascribed to *P. putjatia*/*P. guinanensis*. Use of ABC will allow phylogeographic and taxonomic insights from the genomic divergence of these species on the QTP.

Materials and Methods

Sampling

We obtained specimens from the major areas of the distribution of *P. putjatia* and the morphological species described by Ji et al., (2009) *P. guinanensis*. Specimens from around the Qinghai Lake (sites 9-16) were all identified as *P. putjatia* and will be referred to as 'North *P. putjatia*', while 'South *P. putjatia*' will be used for specimens from sites 8 and 17, and *P. guinanensis* as those from sites 1-7 (Fig. 1; Supporting Information S1). One individual from site 17 showed some morphological similarity with a related species, *P. vlangalii*, while one individual within the *P. guinanensis* area (from site 7) appeared to have a *P. putjatia* morphology. Nevertheless, all individuals were included in the initial analysis. All specimens were captured in August, 2015. Individuals were humanely euthanized

using a barbiturate and tissue samples were frozen in liquid nitrogen. DNA was extracted from tissues using a Qiagen DNeasy Blood and Tissue Kit.

GBS-seq library preparation

GBS is a suitable approach because ascertainment bias is expected to be low (Cornuet et al., 2014). A previously published GBS protocol was used (Davey *et al.*, 2011; Elshire *et al.*, 2011), with some modifications. Approximately 50-100ng DNA was digested using 1U of the restriction enzymes ApeK and PstI, after which individually barcoded/indexed adapters were ligated to the sticky ends. Following purification, the size-selected DNA was then amplified using standard PCR with Phusion high fidelity DNA polymerase (Thermo Scientific, Pittsburgh, PA), and the products were purified using magnetic AMPure XP beads (Beckman Coulter, Inc., Indianapolis, IN). The ligated samples were run on a 2% low-melt agarose gel, and DNA in the size range of 400 to 500 bp was excised from the gel and purified using a MinElute Gel Extraction Kit (Qiagen, Valencia, CA).

Sequencing and *SNP-calling*

Paired-end Illumina HiSeq 2000 sequence data from 43 individuals from all sites were analysed using the denovo_map.pl pipeline for the STACKS program (Catchen *et al.,* 2011) to identify single nucleotide polymorphisms (SNPs) across all sequences and summarize them in a single VCF file. During this procedure, the minimum number of raw reads required to form an initial stack was set as m=6 and the maximum distance allowed for merging of stacks was set as M=3 (following several tests). The distance between catalogue loci parameter was set as n=5. SNPS that were missing in more than 8 individuals were removed.

Genetic structuring

The genetic structure across all individuals was investigated using two alternative approaches with very different assumptions. Discriminant Analysis of Principal Components (DAPC) was applied using

the R package *adegenet* (Jombart, 2008). It uses well-known general statistical procedures but no explicit evolutionary model. In the first step of this analysis, a Principal Components Analysis (PCA) was applied to all biallelic genome sites coded as 0, 1 or 2, corresponding to homozygosity of the reference allele, heterozygosity or homozygosity of the alternative alleles, respectively. All Principal Components (PCs) were retained from the PCA, and the *k*-means clustering algorithm used to compare different numbers of groups (*k*) under the Bayesian Information Criterion (BIC) (results for *k*=1 to *k*=17 clusters were compared). In the second step, a Discriminant Function Analysis (DFA) was applied to the first 10 PC scores with grouping determined from the *k*-means clustering analysis. The number of PCs was determined from the successful assignment rates and Root Mean Square Errors (RMSE) obtained using the cross-validation procedure in *adegenet* (Jombart, 2008).

The Bayesian clustering algorithm incorporated within the program TESS was also used to determine genetic structuring. Geographical coordinates were entered for each individual and the spatial interaction parameter specified as 0.6. A no-admixture model was used which allowed rapid MCMC convergence for these large datasets (5000 steps burnin, 20000 subsequent steps). Ten replicates were performed for each specified number of clusters (progressively from k_{max}=2 to k_{max}=10). The Deviance Information Criterion (DIC) was obtained for each run and the runs with the lowest DICs selected for analyses. The program Clumpp (ver 1.1.2) was used to obtain general membership coefficient matrices from different runs for a given k_{max}, contingent on them providing a DIC that was within the lowest 20% of DIC values.

The degree to which isolation-by-distance determined structuring was examined within identified genetic groups using Mantel's test in *adegenet* (Jombart, 2008). Chord distances (excluding loci under selection: see later) were used to describe genomic divergence between sites. Genomic distances were compared with measures of geographical distance between sites. The latter were obtained both as straight line distances calculated from latitude/longitude coordinates and also

around-geographical-feature (lake) distances that were recorded using Google Earth. The correlation test statistic was compared with that obtained for 99999 random permutations of the genetic distance matrix.

Tests of selection

Identification of loci that were candidates for selection was performed using Bayescan (ver 2.1)(Foll and Gaggioti 2008) which employs an F_{ST} outlier approach. One aim of these analyses was to eliminate loci under selection from subsequent coalescent analyses. Bayescan implements a reversible jump MCMC approach that can move between a selection model, containing a population-specific component and a locus-specific component, and a model with just a population-specific component (i.e., no locus-specific component and therefore no selection). The posterior probability for selection at a locus is determined by the proportion of MCMC samples that include the model with the locus-specific component. Only individuals corresponding to *P. guinanensis* and *P.putjatia* (see Results) were used. The prior on the proportion of sites under selection was set at 1:100. This prior can have considerable influence on the number of sites detected, so runs were also carried out with 1:10 and 1:1000 proportions. Several runs were made and results compared, but the MCMC characteristics of the definitive analysis were: 25 pilot runs of 5000 steps, 130000 iterations with 50000 discarded as burnin, and a sampling interval of 10. Outliers were identified from the results using the R code supplied with Bayescan. A false discovery rate of 5% was used.

Historical population splitting

A coalescent-based analysis of the pattern of divergence between genetic groups was carried out on the SNPs using the Approximate Bayesian Computation (ABC) approach implemented in DIYABC ver. 2.1.0 (Cornuet *et al.*, 2014). Candidate sites for selection were removed, together with sites that were rendered monomorphic after the removal of individuals identified as different species. Also, only one SNP was used per sequence to reduce linkage disequilibrium between SNPs. The analysis

tested a simple morphology-based scenario in which historical population splitting events are concordant with current taxonomy, i.e., (*P. guinanensis*, (North *P. putjatia*, South *P. putjatia*)) against a mtDNA-based scenario that the major split within the group divided the South *P. putjatia*/*P. guinanensis* lineage from the North *P. putjatia* lineage, i.e., (North *P. putjatia*, (*P. guinanensis*, South *P. putjatia*)), as suggested by phylogeographical analysis (Jin *et al.*, 2014). Each scenario contains only five parameters (two divergence times and three effective population sizes). Two divergent individuals (from sites 17 and 5) were excluded from the analysis (see later).

Three effective population sizes were estimated, one for each lineage. Uniform priors of [10, 100000] were placed on the main lineages, while a U[1,40000) prior was placed on the South *P. putjatia* lineage. Time priors of U[100, 100000] and U[1, 50000] were placed on the basal and most recent splits, respectively. One million datasets were simulated, i.e., $5x10^5$ for each of the two scenarios, and the proportion of simulated data with values below that of the observed dataset were assessed, with a PCA being used to compare priors under the two scenarios with the observed data (see Cornuet *et al.*, 2010). We also examined 10000 data sets under the mtDNA scenario-posterior combination to test their fit to the observed data set. A PCA was performed in the space of summary statistics, with PCs being computed from the simulated data sets with parameter values drawn from the prior. The observed data set as well as the data sets simulated from the posterior parameter distributions were added to each PCA dimension, allowing visual assessment.

Results

SNP identification

After filtering, a total of 20365 SNPs from 1269 loci were detected from the 43 individuals.

Genetic structuring

DAPC analysis.- The *k*-means clustering algorithm revealed lowest BIC values for two (269.6) and three clusters (269.8); values for one (283.1) and four (271.1) or more clusters were all higher. To ensure incorporation of all major genetic groupings, three groups were used in the subsequent DAPC. Cross-validation showed that use of the first 10 PCs (60.3% of variance) provided higher assignment rates (90.4%) and a lower RMSE (0.179) than for more than 10 PCs, justifying use of this subset of PCs in the analysis. Patterns of group differentiation were similar, independently of whether all or just this subset of PCs were used, indicating that the results were robust.

All three genetic clusters were clearly differentiated, with no overlap (Fig. 2). Cluster 1 contained all northern *P. putjatia* from around the Qinghai lake (19 individuals). Cluster 2 contained 22 individuals from the Guinan sand-dune region approximately 30 km south of the Longyangxia reservoir (corresponding to *P. guinanensis*), two individuals from site 17 north of the reservoir and 3 southern *P. putjatia* from site S8. Cluster 3 comprised just two individuals from different southern areas: one (of three) individuals from site 5 (Guinan sand-dune region) and one (of three) individuals from site 17. The first two clusters were the largest, least divergent, and corresponded to two isolated regions. One of the two individuals that made up cluster 3 had a distinct morphology and resembled *P. vlangalii*.

A separate DAPC was used to further-explore divergence within the subset of 22 individuals that formed the second of these clusters to further-assess genomic divergence between (southern) *P. putjatia* and *P. guinanensis* morphologies. The same approach described for the DAPC across all individuals was used. Previous SNP sites that were non-polymorphic within this subset group and those with missing data were removed, leaving 4553 SNPs for analysis. Three clusters appeared to best-describe the structure in the data. Cross-validation showed that use of the first three PCs in the DAPC provided 100% assignment rate success. A plot of the two Discriminant Functions (73.4% and 26.6% of variation) showed similar separation between the cluster containing *P. guinanensis* and

clusters representing each of the two other southern *P. putjatia* sites (Fig. 3). The genomic divergence between the latter two southern *P. putjatia* clusters is greater than that between these respective groups and the *P. guinanensis* cluster.

Bayesian clustering.- TESS identified the same number of groups (three) with identical compositions as the DAPC analysis on all specimens. The lowest 20% of DIC values across all runs were in the interval (531496, 531505), and gave average log-likelihoods in the range (-256230,-256243) compared with average log-likelihoods of (-256235, -258963) for runs with lower DIC values. The runs with lowest DIC values were obtained for different values of k_{max}, specifically k_{max} = 3 (7 runs), k_{max} = 4 (4 runs), k_{max} = 5 (6 runs) and k_{max} = 10 (1 run). Crucially, when individual cluster assignments were examined for these values of k_{max} three identical genetic clusters were identified in all cases. These clusters corresponded exactly in composition with those identified by DAPC.

Following the DAPC and TESS analyses, the two most divergent individuals (from sites 5 and 17: cluster 3 in the full analysis) were eliminated because they appeared to represent *P. vlangalii* or individuals derived from recent hybridization with *P. vlangalli*.

Tests of selection

A total of 391 candidate sites for selection (within 337 loci) were determined by Bayescan under a prior ratio of 1:100 for selected: neutral sites. This increased to 705 sites when a ratio of 1:10 was used, and decreased to 206 sites when the prior ratio was 1:1000 (based on the number of SNPs analysed, this suggested a true ratio of between 1:10 and 1:100 and so our use of a 1:100 prior provides quite conservative results). Nevertheless, patterns of divergence were relatively even across all SNPs.

Relationships among individuals based on these 391 SNPs were investigated using PCA to examine whether these sites might be associated with morphology and explain the *P. guinanensis*/*P. putjatia* division. The analysis did not reveal this pattern, although intriguingly, one southern *P.*

putjatia from site 17 was closely associated with northern *P. putjatia* suggesting a difference from patterns of generalized genomic divergence in these SNPs.

Patterns of population divergence

After removal of SNPs that were monomorphic, under selection, or from the same sequence, 4148 SNPs remained for analysis. Prior checking revealed suitability of the models in the ABC analysis ($P>0.01$ for all statistics, when observed are compared to simulated data) while the PCA on the model-posterior combination showed a tight cluster of points from the simulated data around the value for the observed data (mtDNA scenario) (Supporting information S2).

The ABC analysis unequivocally favoured the previously-determined mitochondrial tree, i.e., (North *P. putjatia*, (*P. guinanensis*, South *P. putjatia*)), relative to the morphology-based current taxonomy tree (Fig. 4). This was independent of whether posterior probabilities were obtained directly (posterior probabilities of 0.000 and 1.000 for the morphological and mtDNA scenarios, respectively), or using the logistic regression approach (posterior probabilities of 0.000 and 1.000, respectively)(Cornuet *et al.*, 2008). ABC analyses are generally thought to be more robust when a small number of relatively simple models are compared (Cabrera and Palsbøll, 2017) and posterior probabilities are decisive, as was the case here.

The Mantel test detected a highly significant relationship between genetic and geographical distance for the Qinghai Lake Basin samples despite relatively few sample sites (8). This suggested a strong genomic signal of isolation-by-distance. The association was highly significant when round-lake distances were used ($r=0.700$, $P<0.001$). These distances reflect the effective current geographical proximity between populations. Linear cross-lake distances represented historical geographical proximity at the time that no lake was present and were still significant, but the correlation was much weaker ($r=0.413$, $P<0.02$).

Discussion

We have used analyses of genomic data to resolve uncertainty over historical relationships that emanated from discordance between morphological variation and single loci. This has important biogeographical and taxonomic implications. First, we show that the morphologically-described species known as *P. guinanensis* (Ji *et al.*, 2009) from a sandy desert region in Guinan County on the QTP is most similar to neighbouring southern *P. putjatia* populations, which together are divergent from northern *P. putjatia* populations. This demonstrates discordance between morphological divergence (on which the taxonomy is based) and overall genomic divergence. Second, we find support for a historical population split between northern *P. putjatia* and southern *P. putjatia/P. guinanensis* populations. These two groups are separated by the Qinghai South and Riyue mountains which appear high enough to form a dispersal barrier (see later). Finally, we provide a clear demonstration of how isolation-by-distance (likely associated with the dispersal barrier created by a large saline lake) has shaped genomic structuring in northern *P. putjatia*. This provides further insights into the importance of spatial isolation in generating lizard diversity on the QTP.

The generalized genomic patterns of divergence largely supported the mtDNA relationships described by Jin *et al.* (2014), that is, mtDNA provided an accurate, well-resolved picture of historical population splitting. Prior to current genomic approaches becoming available, some authors had observed that the information content provided by mtDNA for species delimitation was potentially underestimated, e.g., Wiens and Penkrot (2002). We find support for this assertion here: our cross-genomic divergence is concordant with mtDNA phylogeography but not morphology. Nonetheless, an important general caveat is that mtDNA does not always reliably diagnose evolutionary trajectories of different populations. Geographical discordance between mtDNA and single nuclear markers is well-known at the intraspecific level (see review in Toews and Brelsford, 2012).

The historical divergence between populations definitively shows the unsuitability of the morphology-based taxonomy. The genome-wide divergence between northern *P. putjatia* and

297 southern *P. putjatia*/*P. guinanensis* populations, and previous dating of mtDNA divergence to around
298 2 mya (Jin *et al.*, 2014) suggest these two groups could be candidate species. Nevertheless, the
299 current lack of clear differentiating diagnostic characters means that there is little additional support
300 for this designation (see Bauer *et al.*, 2010). The divergence between *P. guinanensis* and southern
301 populations of *P. putjatia* is also important, but the latter populations i) do not group together, ii) do
302 not appear to show much more divergence than expected from their relative geographical isolation,
303 iii) show relatively little divergence relative to the divergence between northern *P. putjatia* and *P.*
304 *guinanensis*/southern *P. putjatia*. Hence we do not consider this to be particularly relevant for the
305 current taxonomy. It may be most appropriate to simply synonymize *P. guinanensis* with *P. putjatia*.

306 More anecdotal evidence of the discordance between genomic and morphological
307 divergence was provided by the identification of two divergent genomes in morphologically quite
308 similar specimens (that we subsequently assigned to *P. vlangalii*). Despite considerable experience
309 identifying these species in the wild only one of these specimens appeared to display *P. vlangalii*
310 characteristics. One of the *P. guinanensis* specimens from the desert area (site 7) also resembled *P.*
311 *putjatia*, morphologically, but was genomically similar to other specimens from neighbouring sites.

312 The primary effect that seems to have driven genomic divergence is reduced gene flow. The
313 Qinghai South and Riyue Mountains are extremely high with average elevations around 4000m and
314 peaks reaching 5139 m. They divide the two main genomic groups and this clearly implies
315 interruption of dispersal. There is a general consensus that mountains on the QTP have been
316 important regional drivers of divergence and speciation, largely based on observations of
317 phylogeographic divisions across mountain ranges (e.g., Jin *et al.*, 2008; Wang *et al.*, 2013; Zhou *et*
318 *al.*, 2013). Here we strengthen this evidence by demonstrating genome-wide as opposed to single
319 marker effects (previous mtDNA studies also suggested a phylogeographic break in this area) . The
320 finding that one *P. putjatia* individual from site 17 (south of these mountains) is very similar to all

northern *P. putjatia* when only outlying SNPs (i.e., candidates for selection) were analysed is quite surprising as the same individual was more similar to *P. guinanensis* in terms of generalized genomic divergence. This suggests that loci under strong selection might show slightly different spatial patterns compared with other parts of the genome.

A second important gene-flow associated element of divergence is isolation-by-distance in populations around Lake Qinghai. The spatial pattern is highly significant across a relatively small number of populations (8) indicating a strong signal. The total lake circumference is generally reported as around 360km (e.g., Virkutyte and Sillanpää, 2006) so our results suggest gene flow is sufficiently restricted to allow local divergence between populations separated by (on average) around 45 km. The closely-related *P. vlangalii* on the QTP also appears to show low dispersal, leading to detectable genetic structure (in microsatellites) over even shorter distances (20 km)(Qi *et al.*, 2013). Other similar microsatellite studies of ectothermic vertebrates have detected isolation-by-distance over similar spatial scales (e.g., Arizona treefrogs; Mims *et al.*, 2016). The very strong pattern detected here, despite fewer individuals per site, shows the power of the GBS approach to detect this effect. It contrasts with a previous mtDNA study which did not reveal isolation-by-distance (Jin *et al.*, 2014). The possibility that the pattern was due to differing environments rather than geographical distance per se appears unlikely as all samples were obtained at similar elevations in similar habitats around the lake (approximately 3300 m above sea level). Also, testing of this group of individuals alone (data not shown) did not reveal SNPs with signatures of selection when just this group was tested, supporting neutral evolution.

Overall, our analyses show how isolation-by-distance and, to a greater extent, physical dispersal barriers (lakes, mountains) lead to genomic divergence among *Phrynocephalus* lizards on the QTP, possibly because of relatively low dispersal. This appears to differ from the factors that have determined the pattern of geographical variation in morphology. Ecological factors appear to have

helped drive genetic divergence in some groups (e.g., Præbel *et al.*, 2013; Wang *et al.*, 2013) but here the general genomic divergence can be largely explained by isolation due to dispersal barriers or simply distance alone. The fact that this genomic divergence can arise over short geographical distances suggests that this may have played a major role in generating diversity on the QTP.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant numbers: 31772447, 31372183). China Jiliang University and LJMU funded exchange visits that led to the formulation of this project. We wish to thank Dr. Guimin Liu for help in the laboratory.

References

Baird, N.A., Etter, P.D., Atwood, T.S., Currey, M.C., Shiver, A.L., Lewis, Z.A., Selker, E.U., Cresko, W.A., Johnson, E.A., 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PloS One,* 3, e3376.

Barley, A.J., White, J., Diesmos, A.C., Brown, R.M., 2013. The challenge of species delimitation at the extremes: diversification without morphological change in Philippine sun skinks. *Evolution*, 67, 3556-72.

Bauer, A.M., Parham, J.F., Brown, R.M., *et al.,* 2010 Availability of new Bayesian-delimited gecko names and the importance of character-based species descriptions. *Proceedings of the Royal Society of London B: Biological Sciences*, 10.1098/rspb.2010.1330.

Brown, R.P., Paterson, S., Risse, J., 2016. Genomic signatures of historical allopatry and ecological divergence in an island lizard. *Genome Biology and Evolution*, 8, 3618-26.

Bryant, D., Bouckaert, R., Felsenstein, J., Rosenberg, N. A., RoyChoudhury, A., 2012. Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. *Molecular Biology and Evolution*, 29, 1917-1932.

368 Cabrera, A.A., Palsbøll, P.J., 2017. Inferring past demographic changes from contemporary genetic
369 data: A simulation-based evaluation of the ABC methods implemented in diyabc. *Molecular*
370 *Ecology Resources*, 17, e94-110.
371 Catchen, J.M., Amores, A., Hohenlohe, P., Cresko, W., Postlethwait, J.H., 2011. Stacks: building and
372 genotyping loci de novo from short-read sequences. *G3: Genes, Genomes, Genetics*, 1, 171-
373 182.
374 Combosch, D.J., Vollmer, S.V., 2015. Trans-Pacific RAD-Seq population genomics confirms
375 introgressive hybridization in Eastern Pacific Pocillopora corals. *Molecular Phylogenetics and*
376 *Evolution*, 88, 154-162.
377 Cornuet, J.M., Santos, F., Beaumont, M.A. *et al.,* 2008. Inferring population history with DIY ABC: a
378 user-friendly approach to approximate Bayesian computation. *Bioinformatics*, 24, 2713-9.
379 Cornuet, J.M., Ravigné, V. and Estoup, A., 2010. Inference on population history and model checking
380 using DNA sequence and microsatellite data with the software DIYABC (v1. 0). *BMC*
381 *Bioinformatics*, 11, 401.
382 Cornuet, J.M., Pudlo, P., Veyssier, J. *et al.* 2014. DIYABC v2. 0: a software to make approximate
383 Bayesian computation inferences about population history using single nucleotide
384 polymorphism, DNA sequence and microsatellite data. *Bioinformatics*, 30, 1187-9.
385 Davey, J.W., Hohenlohe, P.A., Etter, P.D., Boone, J.Q., Catchen, J.M., Blaxter, M.L., 2011. Genome-
386 wide genetic marker discovery and genotyping using next-generation sequencing. *Nature*
387 *Reviews Genetics*, 12, 499-510.
388 Bouckaert, R., Heled, J., Kühnert, D. *et al.,* 2014. BEAST 2: a software platform for Bayesian
389 evolutionary analysis. *PLoS Computational Biology*, 10, e1003537.
390 Eaton, D.A, Ree, R.H., 2013. Inferring phylogeny and introgression using RADseq data: an example
391 from flowering plants (Pedicularis: Orobanchaceae). *Systematic Biology* 62, 689-706.

392 Elshire R.J., Glaubitz, J.C., Sun, Q. *et al.,* 2011. A robust, simple genotyping-by-sequencing (GBS)
393 approach for high diversity species. *PloS One*, *6*, e19379.
394 Foll, M., Gaggiotti, O., 2008. A genome-scan method to identify selected loci appropriate for both
395 dominant and codominant markers: a Bayesian perspective. *Genetics*, 180, 977-993.
396 Hohenlohe, P.A., Bassham, S., Etter, P.D., Stiffler, N., Johnson, E.A., Cresko, W.A., 2010. Population
397 genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS*
398 *Genetics*, 6, e1000862.
399 Hudson, R., 2002. Generating samples under a Wright-Fisher neutral model of genetic variation.
400 *Bioinformatics*, 18, 337-338.
401 Ji, X., Wang, Y.Z., Wang, Z., 2009. New species of *Phrynocephalus* (Squamata, Agamidae) from
402 Qinghai, Northwest China. *Zootaxa*, *1988*, 61-68.
403 Jin, Y.T., Brown, R.P., Liu, N.F., 2008. Cladogenesis and phylogeography of the lizard *Phrynocephalus*
404 *vlangalii* (Agamidae) on the Tibetan plateau. *Molecular Ecology*, *17*, 1971-1982.
405 Jin, Y., Yang, Z., Brown, R.P., Liao, P., Liu, N., 2014. Intraspecific lineages of the lizard *Phrynocephalus*
406 *putjatia* from the Qinghai-Tibetan Plateau: Impact of physical events on divergence and
407 discordance between morphology and molecular markers. *Molecular Phylogenetics and*
408 *Evolution*, *71*, 288-297.
409 Jin, Y., Wo, Y., Tong, H., Song, S., Zhang, L., Brown, R.P., 2018. Evolutionary analysis of
410 mitochondrially encoded proteins of toad-headed lizards, *Phrynocephalus*, along an
411 altitudinal gradient. *BMC Genomics*, 19, 185.
412 Jombart, T., 2008. adegenet: a R package for the multivariate analysis of genetic
413 markers. *Bioinformatics*, *24*, 1403-1405.
414 Leaché, A.D., 2009. Species tree discordance traces to phylogeographic clade boundaries in North
415 American fence lizards (*Sceloporus*). *Systematic Biology*, *58*, 547-559.

Leaché, A.D., Fujita, M.K., Minin, V.N., Bouckaert, R.R., 2014. Species delimitation using genome-wide SNP data. *Systematic Biology*, 63, 534-542.

McCracken, K.G., Sorenson, M.D., 2005. Is homoplasy or lineage sorting the source of incongruent mtDNA and nuclear gene trees in the stiff-tailed ducks (*Nomonyx-Oxyura*)? *Systematic Biology*, *54*, pp.35-55.

Mims, M.C., Hauser, L., Goldberg, C.S., Olden, J.D., 2016. Genetic differentiation, isolation-by-distance, and metapopulation dynamics of the Arizona treefrog (*Hyla wrightorum*) in an isolated portion of its range. *PloS One*, *11*, e0160655.

Moore, W.S., 1995. Inferring phylogenies from mtDNA variation: mitochondrial-gene trees versus nuclear-gene trees. *Evolution*, *49*, 718-726.

Pamilo, P., Nei, M., 1988. Relationships between gene trees and species trees. *Molecular Biology and Evolution*, *5*, 568-583.

Præbel, K., Knudsen, R., Siwertsson, A. *et al.,* 2013. Ecological speciation in postglacial European whitefish: rapid adaptive radiations into the littoral, pelagic, and profundal lake habitats. *Ecology and Evolution*, *3*, 4970-4986.

Qi, Y., Yang, W., Lu, B., Fu, J., 2013. Genetic evidence for male-biased dispersal in the Qinghai toad-headed agamid *Phrynocephalus vlangalii* and its potential link to individual social interactions. *Ecology and Evolution*, *3*, 1219-1230.

Shafer, A.B., Gattepaille, L.M., Stewart, R.E. and Wolf, J.B., 2015. Demographic inferences using short‐read genomic data in an approximate Bayesian computation framework: in silico evaluation of power, biases and proof of concept in Atlantic walrus. *Molecular Ecology*, 24, 328-345.

Takahashi, K., Terai, Y., Nishida, M., Okada, N., 2001. Phylogenetic relationships and ancient incomplete lineage sorting among cichlid fishes in Lake Tanganyika as revealed by analysis of the insertion of retroposons. *Molecular Biology and Evolution*, *18*, 2057-2066.

441 Toews, D.P., Brelsford, A., 2012. The biogeography of mitochondrial and nuclear discordance in
442 animals. *Molecular Ecology*, *21*, 3907-3930.
443 Virkutyte, J., Sillanpää, M., 2006. Chemical evaluation of potable water in Eastern Qinghai Province,
444 China: Human health aspects. *Environment International*, 32, 80-86.
445 Wang, Y., Zhao, L.M., Fang, F.J., Liao J.C., Liu, N.F., 2013. Intraspecific molecular phylogeny and
446 phylogeography of the *Meriones meridianus* (Rodentia: Cricetidae) complex in northern
447 China reflect the processes of desertification and the Tianshan Mountains uplift. *Biological*
448 *Journal of the Linnean Society*, *110*, 362-383.
449 Wiens, J.J., Penkrot, T.A., 2002. Delimiting species using DNA and morphological variation and
450 discordant species limits in spiny lizards (*Sceloporus*). *Systematic Biology*, *51*, pp.69-91.
451 Zhou, W., Yan, F., Fu, J. *et al.*, 2013. River islands, refugia and genetic structuring in the endemic
452 brown frog *Rana kukunoris* (Anura, Ranidae) of the Qinghai-Tibetan Plateau. *Molecular*
453 *Ecology*, *22*, 130-142.

Figure Legends

Figure 1. Map of sample locations. Sites 1-7 are marked with circles, located on the sand-dune areas of Guinan and correspond to *P. guinanensis*. Sites 9-17 (triangles) correspond to the species *P. putjatia*. Sites 5, 8 and 17 are marked with an asterisk because slightly unusual individuals (in terms of morphology or genome) were found there (described in text).

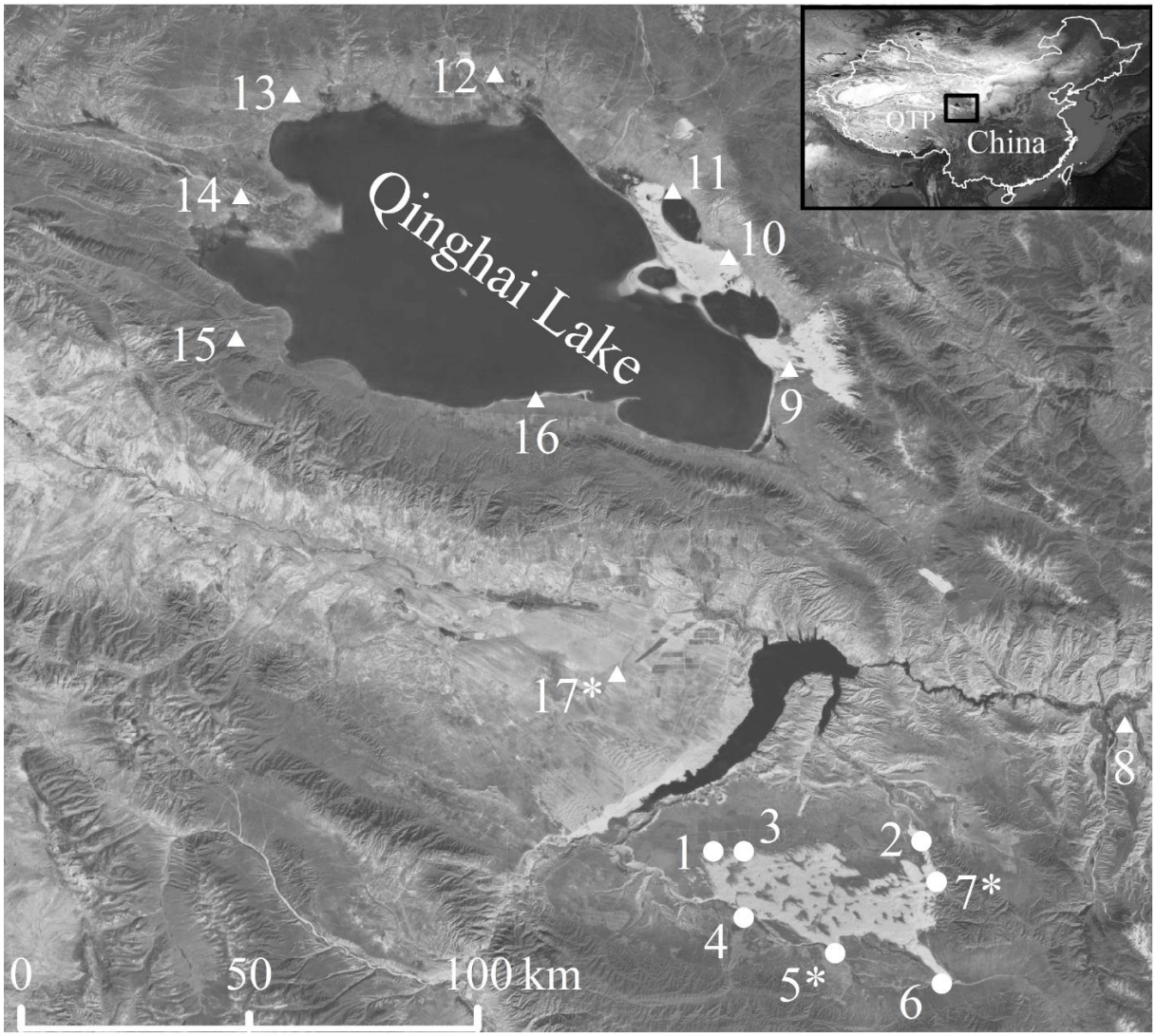

Figure 2. Scatterplot of scores from the two discriminant functions from the DAPC analysis of all SNPs across all individuals. The first discriminant function, representing 58.2% of variation, is on the horizontal axis, the second discriminant function (vertical axis) represents the remaining variation; the inset bar chart displays the difference in magnitude of the two corresponding eigenvalues. Cluster 1 contains only northern *P. putjatia* from around Qinghai Lake, Cluster 2 contains *P. guinanensis* and southern *P. putjatia* from sites 1-8 and 17, Cluster 3 contains two individuals (sites 7 and 17), one of which showed superficial resemblance to *P. vlangalii*.

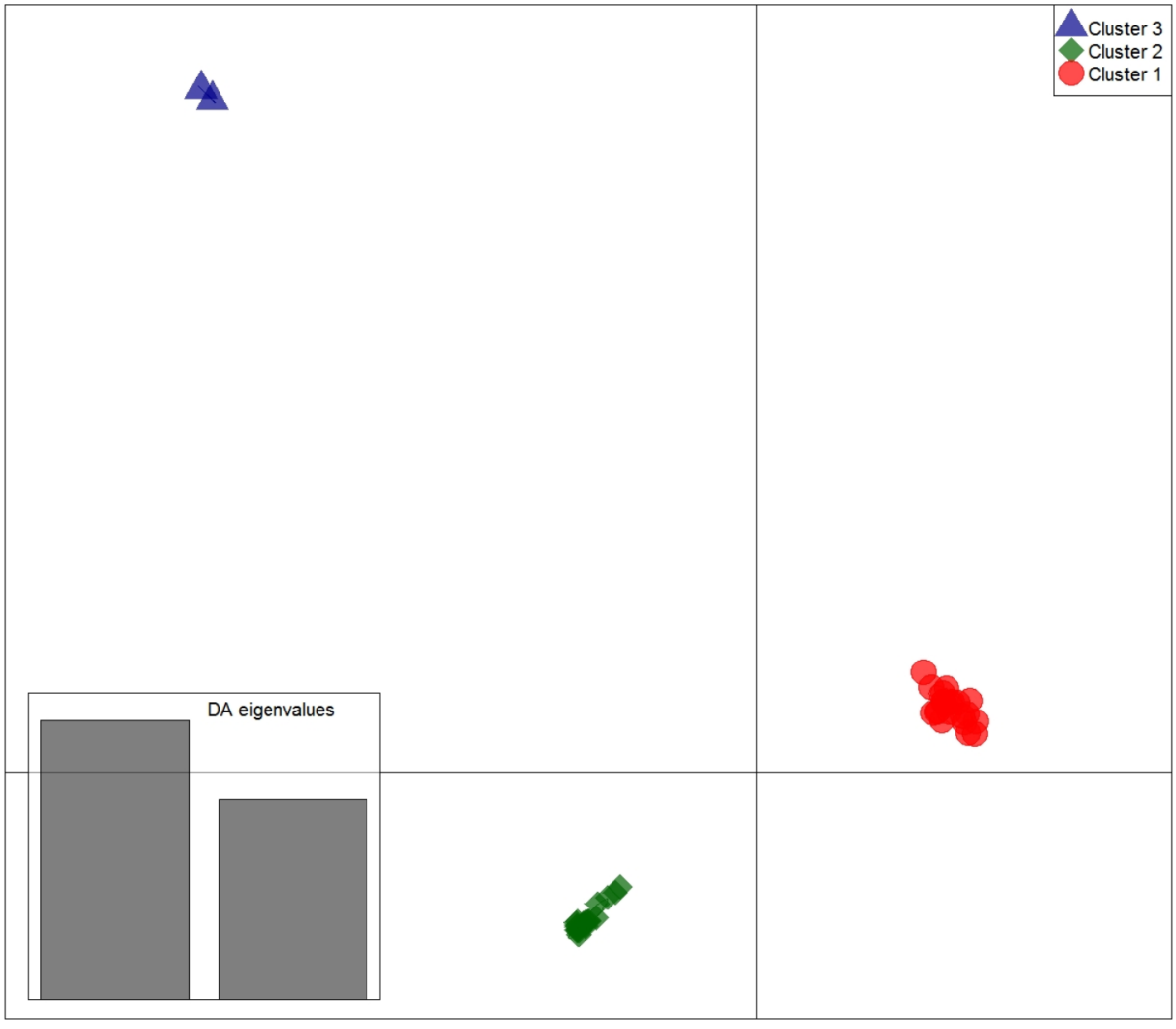

Figure 3. Scatterplot of scores from the two discriminant functions from the DAPC analysis of 4553 SNPs from southern *P. putjatia* and *P. guinanensis*. The first discriminant function, representing 73.4% of variation, is on the horizontal axis. The second discriminant function (vertical axis) represents the remaining variation, with the inset displaying the magnitude of the difference between the two corresponding eigenvalues. The cluster denoted as A contains only two *P. putjatia* individuals from site 17, while B contains three *P. putjatia* individuals from site 8. The largest cluster (C) contains *P. guinanensis* from sites 1-7. Lines represent a minimum spanning tree that connects group centres.

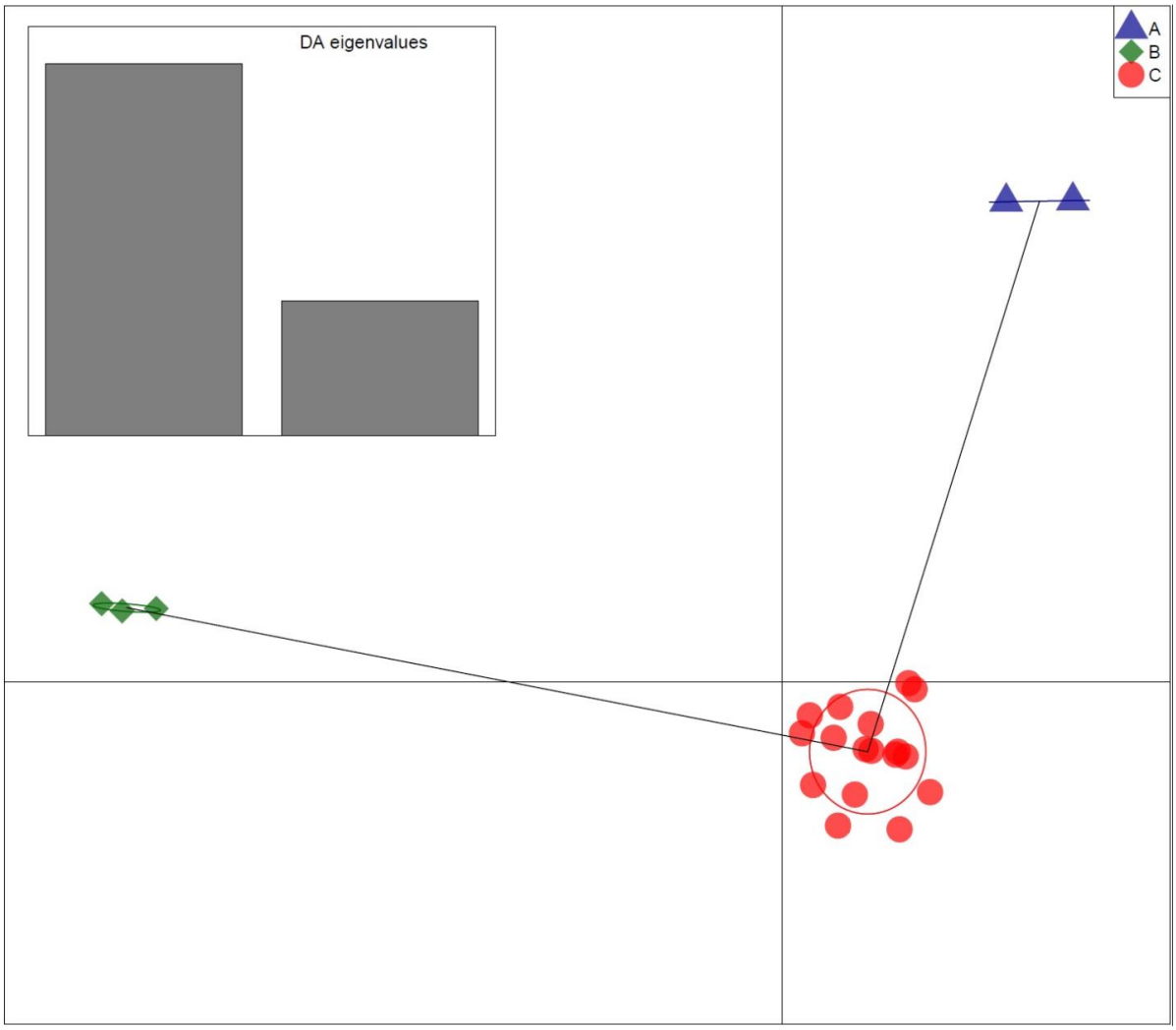

Figure 4. The two historical splitting scenarios tested using Approximate Bayesian Computation. Scenario i) is based on the mtDNA phylogeographic pattern in Jin *et al.* . 2014, and scenario ii) is the scenario based on previously-described morphological species (Ji *et al.*, 2009). Posterior probabilities for these alternatives were i) 1.0000 and ii) 0.0000 under both the direct and logistic regression approaches (see results).