MOLECULAR ECOLOGY

Molecular Ecology (2014) **23**, 2228–2241 doi: 10.1111/mec.12736

The advantages of going large: genome-wide SNPs clarify the complex population history and systematics of the threatened western pond turtle

PHILLIP Q. SPINKS,*† ROBERT C. THOMSON‡ and H. BRADLEY SHAFFER*†

**Department of Ecology and Evolutionary Biology, University of California, 621 Charles E. Young Dr. South, Los Angeles, CA 90095-1606, USA, †La Kretz Center for California Conservation Science, Institute of the Environment and Sustainability, University of California, La Kretz Hall, Suite 300, 619 Charles E. Young Dr. South, Los Angeles, CA 90095-14966, USA, ‡Department of Biology, University of Hawaii at Manoa, Honolulu, HI 96822, USA*

Abstract

As the field of phylogeography has matured, it has become clear that analyses of one or a few genes may reveal more about the history of those genes than the populations and species that are the targets of study. To alleviate these concerns, the discipline has moved towards larger analyses of more individuals and more genes, although little attention has been paid to the qualitative or quantitative gains that such increases in scale and scope may yield. Here, we increase the number of individuals and markers by an order of magnitude over previously published work to comprehensively assess the phylogeographical history of a well-studied declining species, the western pond turtle (*Emys marmorata*). We present a new analysis of 89 independent nuclear SNP markers and one mitochondrial gene sequence scored for rangewide sampling of >900 individuals, and compare these to smaller-scale, rangewide genetic and morphological analyses. Our enlarged SNP data fundamentally revise our understanding of evolutionary history for this lineage. Our results indicate that the gains from greatly increasing both the number of markers and individuals are substantial and worth the effort, particularly for species of high conservation concern such as the pond turtle, where accurate assessments of population history are a prerequisite for effective management.

Keywords: conservation genetics, ecological genetics, phylogeography, population genetics – empirical, reptiles

Received 4 October 2013; revision received 20 March 2014; accepted 24 March 2014

Introduction

Phylogeography has become a cornerstone of basic population biology, species delimitation and applied conservation genetics. Theory and empirical work show that traditional phylogenetic analyses based on single or a few genes may suffer from a number of biases and shortfalls related to the stochasticity of evolutionary processes operating at the population scale, and therefore that increased sampling is fundamental to improve phylogenetic accuracy (Zwickl & Hillis 2002; Heath *et al.* 2008). The impact of these increases for phylogeographical analyses is not well characterized, but recent work indicates that increasing taxon and data sampling both can improve phylogeographical resolution (McCormack *et al.* 2012; Merz *et al.* 2013). Here, we provide a comprehensive case study that examines the importance of scaling up the number of individuals and independent nuclear markers in phylogeographical research. Although recent analyses of full mitochondrial genomes have emphasized the increase in phylogeographical resolution compared to analyses of single mitochondrial genes (Morin *et al.* 2004; Chan *et al.* 2010; Knaus *et al.* 2011; Shamblin *et al.* 2012), the mitochondrion still represents a single genetic locus that may or may not provide a synthetic view of the history of populations and lineages. We therefore focus on multiple,

Correspondence: Phillip Spinks, Fax: 310-206-3987;
E-mail: pqspinks@ucla.edu

independent nuclear markers in this work. Our empirical analysis of the western pond turtle phylogeographical history based on an expanded nuclear data set indicates that a 10-fold increase in both genes and individuals fundamentally changes our understanding of the evolutionary history of this taxon.

Phylogeography seeks to discover both the patterns of genetic divergence across landscapes and the processes that give rise to those patterns. Since its inception as a discipline, analyses of mitochondrial DNA (mtDNA) have been the dominant tool for phylogeographical inquiry (Avise 2000). More recently, the field has moved towards increasingly complex, multilocus analyses. Although the tools and markers have changed, the main objectives of most phylogeographical analyses have remained quite stable: assign individuals into more inclusive lineages (Dimmick *et al.* 1999; Mayden & Wood 1995) and infer the evolutionary processes leading to the observed distribution of those lineages on landscapes (Avise 2000; Avise *et al.* 1987; Pease *et al.* 2009; Walstrom *et al.* 2012).

A key, but virtually unexplored question is the extent to which phylogeographical inferences based on limited taxon sampling and small portions of the genome accurately portray those patterns generated from deeper taxon and data sampling. The answer to this question is critical to the discipline, because the vast majority of phylogeographical knowledge accumulated over the last two decades is based on one or a few gene trees and relatively sparse taxon sampling. Although often limited in scope, the phylogeographical literature is vast: an ISI Web Of Science search (conducted 13 March 2014) on the term 'Phylogeography' yielded 8925 titles that have been cited over 195 000 times since 1988. The importance of this body of knowledge goes beyond purely academic science given the importance of phylogeographical inference in conservation and management, where funding challenges and the frequent impossibility of collecting large numbers of individuals make smaller-scale studies attractive (Fraser and Bernatchez 2001; Moritz 1994; Ryder 1986). However it is accomplished, identifying genetic diversity, including the units of biological conservation, is an essential prerequisite for managers, and getting it right matters.

Here, we compare and contrast the inferences drawn from traditional mtDNA, nuclear DNA sequence data and a large, comprehensive analysis of single-nucleotide polymorphism (SNP) data to explore how analyses of these diverse data sets can refine our insights into the biogeographical history of organisms. We focus on the western pond turtle, *Emys* [*Actinemys*] *marmorata*, an important case study for phylogeographical analysis. Early mtDNA-based analyses presented a strikingly different pattern of genetic divergence compared to that obtained from subsequent analyses of a small number of nuclear loci, leading to conflicting conservation and management interpretations for this threatened species (Spinks & Shaffer 2005; Spinks *et al.* 2010). The new SNP data presented here take the analysis of this species complex to the next standard of analysis: comprehensive population, geographical and genome-wide coverage to infer the history and management units contained within species.

Emys marmorata is the sole freshwater aquatic turtle across the west coast of North America (Stebbins 2003) (Fig. S1, Supporting Information) and is declining over most of its range. A rangewide morphological analysis identified northern and southern groups and suggested that an area of extensive intergradation was restricted to the San Joaquin Valley (Seeliger 1945). Later mtDNA phylogeographical analyses of 135 individuals from 73 localities covering the range of the species recovered four well-supported, geographically cohesive, mtDNA clades with evidence of admixture between northern and San Joaquin Valley clades in the central coast range of California (Fig. S2, Supporting Information, Spinks & Shaffer 2005). Finally, an analysis of five nuclear DNA (nuDNA) sequences with rangewide sampling of 90 turtles revealed two primary groups from northern and southern California, with the central coast range showing some admixture (Fig. S1, Supporting Information). Thus, these earlier studies disagreed on both the number of evolutionary lineages (four vs. two) and the hypothesized regions of intergradation between identified lineages.

A fundamental question is whether these initial studies failed to converge on a single resolution because of a lack of sufficient marker coverage, insufficient sampling within populations or the stochastic nature of a few markers that may be strongly affected by selection and/or lineage sorting. To resolve these questions, and to provide an empirical test case for the assumption that modest phylogeographical analyses capture the most important elements of species' history, we replicated this previous work with a much larger, comprehensive genetic analysis. We collected and analysed mtDNA and nuclear SNP data from geographically comprehensive population sampling comprising 923 individuals and assessed changes in phylogeographical patterns derived from this increased sampling programme. Analyses of this magnitude are probably what can be reasonably expected for the immediate future in nonmodel vertebrate species, and we present our case study as one that answers a simple question: can we feel confident in the primary results derived from sparse sampling that are currently available for most taxa, or should we be much more cautious in our interpretation of them as reflections of population and species history? The results presented here suggest that we should be cautious in interpreting mtDNA/limited nuclear results as accurate reflections of population and species history. Analyses of our large-scale nuclear data reveal a much more biogeographically informative and detailed depiction of the evolutionary history of the western pond turtle, and one that is incongruent with earlier results over a critical part of the species' range. These new results have novel conservation/management implications for this California Species of Special Concern.

Materials and methods

Genetic data

Our initial nuDNA taxon sampling included 946 samples of *Emys marmorata* collected from 103 sites throughout its range including Baja California (33 individuals), California (736), Nevada (10), Oregon (140) and Washington (27). We sampled ~10 individuals/site and included two European pond turtles (*Emys orbicularis*) as the outgroup (Appendix S1, Supporting Information). Previous mtDNA analyses utilized a fragment of the nicotinamide adenine dehydrogenase subunit four gene and flanking tRNA^{His} and tRNA^{Ser} (hereafter referred to as *ND4*). We sequenced *ND4* for all but 24 of the *E. marmorata* that were genotyped for nuDNA, and also included all available *ND4* sequences from GenBank (Appendix S1, Supporting Information). DNA extraction methods as well as PCR conditions for *ND4* follow Spinks & Shaffer (2005).

SNP loci were discovered using a targeted sequencing approach. We identified 84 nuclear markers from the literature and developed an additional 20 markers for this analysis (Appendix S4, Supporting Information). We used BLAST (Zhang *et al.* 2000) to compare genomic scaffolds of the painted turtle (*Chrysemys picta*; Shaffer *et al.* 2013) on GenBank to the chicken genome (International Chicken Genome Sequencing Consortium 2004) and identified markers from alignments of highly conserved regions between *C. picta* and chicken. In addition, we redesigned some markers from the literature because they did not amplify or sequence well for our study organism. The redesigned primer sequences and those generated for this analysis are provided in Appendix S4 (Supporting Information).

We sequenced all 104 markers for a discovery panel of eight geographically dispersed western pond turtles and identified SNPs manually (GenBank Accession nos in Appendix S4, Supporting Information). Individual SNPs were then evaluated by Illumina Inc. (San Diego, CA), and the 96 highest scoring SNPs (one each from 96 loci) were used to design a custom GoldenGate oligo pool assay (OPA). Some of our nuclear alignments contained multiple SNPs, but we sampled only one SNP/locus because these tightly linked SNPs provide less information than multiple unlinked loci for population-level analyses (Morin *et al.* 2004). Samples were genotyped at the University of California Davis Genome Center using the GoldenGate BeadExpress platform (http://dnatech.genomecenter.ucdavis.edu/). A potential confounding issue for SNP analyses is ascertainment bias. However, because our samples were drawn from a rangewide sampling of individuals, ascertainment bias should not be a major issue for our analyses (Morin *et al.* 2004). Our initial taxon sample consisted of two *Emys orbicularis* and 946 *E. marmorata*. However, 23 *E. marmorata* contained more than 10% missing SNP data and were excluded from the analysis. In addition, we excluded five nuclear loci that appeared to be variable in the discovery panel but were invariant and two loci that failed to genotype for more than 10% of the samples (Appendix S1, Supporting Information). Our final SNP data set consisted of 925 individuals genotyped at 89 loci. All SNP genotypes are provided in Appendix S1 (Supporting Information) and are available from Dryad (doi:10.5061/dryad.pr907).

Mitochondrial phylogenetic analyses

We translated the *ND4* sequences into protein sequences using Geneious (Drummond *et al.* 2010) to check for nuclear-mitochondrial pseudogenes and nonsense/frameshift mutations; none were found. Partitioned-model Bayesian analyses used MrBayes version 3.1.2 (Huelsenbeck & Ronquist 2001; Ronquist and Huelsenbeck 2003) on unique sequences only. We identified identical sequences using the ALTER webserver (Glez-Peña *et al.* 2010, http://sing.ei.uvigo.es/ALTER/) and removed them from our alignment prior to analyses. We partitioned the *ND4* data by codon and selected models of molecular evolution for parameter estimation using MrModeltest (Nylander 2002), executed in paup* 4.0b10 (Swofford 2002), under the Akaike information criterion (AIC) (Guindon and Gascuel, 2003). Mixed-model analyses were performed in MrBayes (Ronquist and Huelsenbeck 2003). We ran two replicates each with four incrementally heated chains for ten million generations, sampling from the cold chains every 1000 generations. Stationarity was determined as the point when the potential scale reduction factor (PSRF) reached one and the average standard deviation of split frequencies between independent runs approached 0. We also visually examined the Markov chain Monte Carlo (MCMC) output using TRACER (Rambaut and Drummond, 2009) and AWTY (Nylander *et al.* 2008) to ensure that all chains were sampling from the same target distribution for both the continuous and tree parameters in the model. The first 25% of samples were discarded as burn-in.

We generated nuclear SNP data from 923 individuals, but our mtDNA data matrix consisted of a 725-bp segment of *ND4* from a total of 983 *E. marmorata* complex turtles including 215 GenBank sequences and 768 generated for this analysis. However, 903 of the 983 sequences were redundant and removed from the phylogenetic analysis. Thus, our final mtDNA data matrix was composed of 81 sequences including 80 *E. marmorata* sequences and one *E. orbicularis* outgroup sequence. The mitochondrial clade membership of each *E. marmorata* sample is provided in Appendix S1 (Supporting Information). All of the newly generated *ND4* sequences translated to amino acid sequences (excluding the tRNA region), and this data set was almost complete with 0.6% missing data. The *ND4* data matrix is available from Dryad (doi:10.5061/dryad.pr907), and GenBank numbers for all *ND4* sequences used are provided in Appendix S1 (Supporting Information).

Population assignment analyses

We used Structure (Pritchard *et al.* 2000) and our multilocus SNP data to assign individuals to population clusters. We employed the correlated allele frequency and the admixture ancestry models, assessed values of *K* from 1–10 and determined a preferred value of *K* with the Δ*K* method outlined in Evanno *et al.* (2005) using the Structure Harvester webserver (Earl & vonHoldt 2012). We ran five independent analyses for 1 million generations for each value of *K*, each with a burn-in of 100 000 MCMC generations, and classified individuals with admixture proportions greater than 0.10 to be admixed. For secondary Structure analyses, we assigned the admixed individuals to a population based on their admixture proportions. Two admixed individuals were assigned to the northern group and the remaining 14 to the southern group (Appendix S1, Supporting Information). For the secondary Structure analyses, we assessed *K* from 1–10 and determined a preferred value of *K* using the ΔK method outlined in Evanno *et al.* (2005) separately for each group identified in the primary Structure analysis.

Species delimitation

We used the BPP 2.1 software (Rannala and Yang, 2003; Yang & Rannala 2010) to determine whether the genetic groups revealed by the Structure analyses of our SNP data might comprise species-level divergences. We used the multilocus sequence data from Spinks *et al.* (2010), consisting of 5 nuclear loci sequenced from 90 *Emys marmorata* collected from throughout the range of the species for this analysis. We scored 77 of these 90 individuals for our SNP panel (the remaining 13 failed to genotype). We then assigned these 90 individuals into one of four geographically cohesive groups that were identified in the two sequential Structure analyses of the multilocus SNP data (see Results section below) including a 'central coast/Southern California' (CCSC) group, plus Cascade, foothill and Baja groups. The 13 individuals lacking SNP genotype data were assigned to groups based on their locality data (Appendix S1, Supporting Information). BPP evaluates the posterior probability of species divergences under a user-specified bifurcating guide tree. We used the guide tree treating all four groups as clades, and geographically proximate clades as sister groups [i.e. ((CCSC, Baja), (Cascade, foothill))]. In addition, the authors of BPP suggest that the number of alleles sampled for each putative species included in the BPP analyses should be approximately equal (Yang & Rannala 2010). Spinks *et al.* (2010) generated 100 alleles/locus for the CCSC group, which greatly exceeded that of the Cascade (32), foothill (28) and Baja (20) populations. To address this issue, we generated five subsample data sets each consisting of 20 alleles sampled from each putative species. We sampled alleles at random without replacement from the original pool of alleles for each group except for Baja where all 20 alleles were included in each data set. Data set randomization and assembly was carried out in R (http://www.r-project.org). We also assessed the potential impact of our choice of priors on the final BPP outputs by running analyses under two alternative settings: a Γ(2, 200) prior on the population coalescent parameters (θ) and a Γ(2, 200) prior on the age of the root in the species tree (τ_0), and θ = Γ(2, 2000), τ_0 = Γ(2, 2000). The latter settings place more prior probability on models containing fewer lineages (Yang & Rannala 2010) and should therefore be relatively conservative with respect to the recognition of species. The remaining divergence time parameters were estimated under a uniform Dirichlet prior (Yang & Rannala 2010). For each of the five pseudoreplicate data sets, we ran four individual analyses using: (i) algorithm 0, (θ) 200, τ_0 = 2000; (ii) algorithm 0, (θ) 2000, τ_0 = 2000; (iii) algorithm 1, (θ) 200, τ_0 = 2000; and (iv) algorithm 1, (θ) 2000, τ_0 = 2000. All analyses were run for 100 000 generations with a burn-in of 5000 generations and sampled every 10 generations, using different starting seeds for each analysis.

Population splits and mixtures

Recently, Pickrell & Pritchard (2012) developed a method for inferring population history, including divergence and gene flow using allele frequency data under a

Gaussian approximation to genetic drift. The model of Pickrell & Pritchard (2012), implemented in the TREEMIX version 1.12 software and available at http://treemix.googlecode.com, relates a sample of populations to their common ancestor using a graph of ancestral populations (Pickrell & Pritchard 2012). The output of the TREEMIX software includes a maximum-likelihood (ML) tree of population ancestry and a graph of the ML tree showing estimated migration events (m). In addition, the direction of migration events among populations can be displayed on the graphs. For our TREEMIX analyses, SNP data were converted from diploid genotype calls for each individual into population-level allele counts using a custom Perl script (available at https://github.com/atcg/SNPs2alleles). We performed analyses on the 923 *Emys marmorata* samples grouped into the CCSC, Baja, Cascade and foothill groups revealed in the Structure analyses with the two *Emys orbicularis* individuals included as the outgroup. However, unlike the BPP analyses, we enforced no topological constraints among these four groups, as TREEMIX does not require a guide tree. We also assessed from one to four migration events (m1–m4) to provide an independent assessment of historical migration and admixture among *Emys marmorata* complex taxa. Because admixed individuals may have a disproportionately large effect on inferred migration, we ran additional analyses with admixed individuals (identified in the Structure analyses below) excluded to assess the impact of those admixed individuals on the estimated migration events. In addition, the Baja population consisted of mostly missing data at two SNP loci (*NB06374* and *NB22443*, Appendix S1, Supporting Information), and we ran additional analyses with these two loci excluded to assess the impact of these missing data on the TREEMIX analyses. We performed 100 bootstrap replicates with different starting seeds/replicate and we sampled blocks of five contiguous SNPs/replicate in order to assess how robust these results may be to stochastic sampling error.

Finally, we converted the SNP data into a matrix of pairwise distances (Nei 1972) using the R package Adegenet (Jombart 2008) and generated a distance network from the SNP data using SPLITSTREE version 4.11.3 (Huson & Bryant 2006) and the NEIGHBORNET algorithm (Bryant & Moulton 2004).

Results

Mitochondrial phylogeography

A partitioned Bayesian phylogenetic analysis of our expanded taxon sampling (923 samples vs. 147 in a previous analysis) recovered the same four well-supported mtDNA clades with the same lack of support for relationships among those four clades (Spinks & Shaffer 2005; Spinks *et al.* 2010) (Fig. 1; Fig. S3, Supporting Information). A few details emerged from our sevenfold increase in individual and population sampling: we uncovered a single northern clade mitochondrial haplotype at the extreme southern end of the San Joaquin Valley, a few San Joaquin Valley clade mitochondrial haplotypes in the southern Sacramento Valley and Nevada and a clear indication that the previously observed admixture between San Joaquin Valley and northern populations is largely restricted to the southern San Joaquin River and adjacent coastal areas. However, this massive increase in individual and population sampling revealed few novel insights for either phylogeography or conservation biology (Fig. 1, Fig. S2, Supporting Information).

SNP population assignment analyses

Bayesian population assignment analyses of the entire SNP data set revealed overwhelming support for two clusters where 893/923 individuals (97%) were assigned with posterior probabilities (PP) of ≥0.95, 14 individuals were assigned with PP between 0.90 and 0.949 and the remaining 16 individuals were admixed (i.e. assigned with PP < 0.90). The two clusters included a northern group (516 individuals) ranging from the southern San Joaquin Valley north to Washington, including the Nevada population, and a southern group (391 samples) extending from the Central Coast Range south to Baja California, including a few sites in the San Joaquin Valley and the Mojave population (Fig. 2). Contrary to earlier molecular and morphological results, only 16 individuals appeared to be admixed and these were distributed in a restricted arc extending from the northern central coast range southeastwards to the Sierra Nevada foothills. Over half of these individuals occurred in two adjacent populations a few kilometres apart and may represent human-mediated introductions. These two clusters are somewhat consistent with previous results from more limited nuclear sequence data (Spinks *et al.* 2010), although the pattern and extent of admixture is strikingly different. Our enlarged SNP data set recovered essentially pure southern group animals throughout the central coast range, in sharp contrast to earlier mitochondrial and nuclear analyses, but consistent with Seeliger's earlier morphological analyses (Fig. 2, Fig. S1, Supporting Information). However, in contrast with Seeliger's (1945) morphological analysis (Fig. S1, Supporting Information), populations from the San Joaquin Valley were recovered as primarily pure northern group animals rather than admixed intergrades.

All individuals that clustered in the northern population based on the SNP data contained either north-

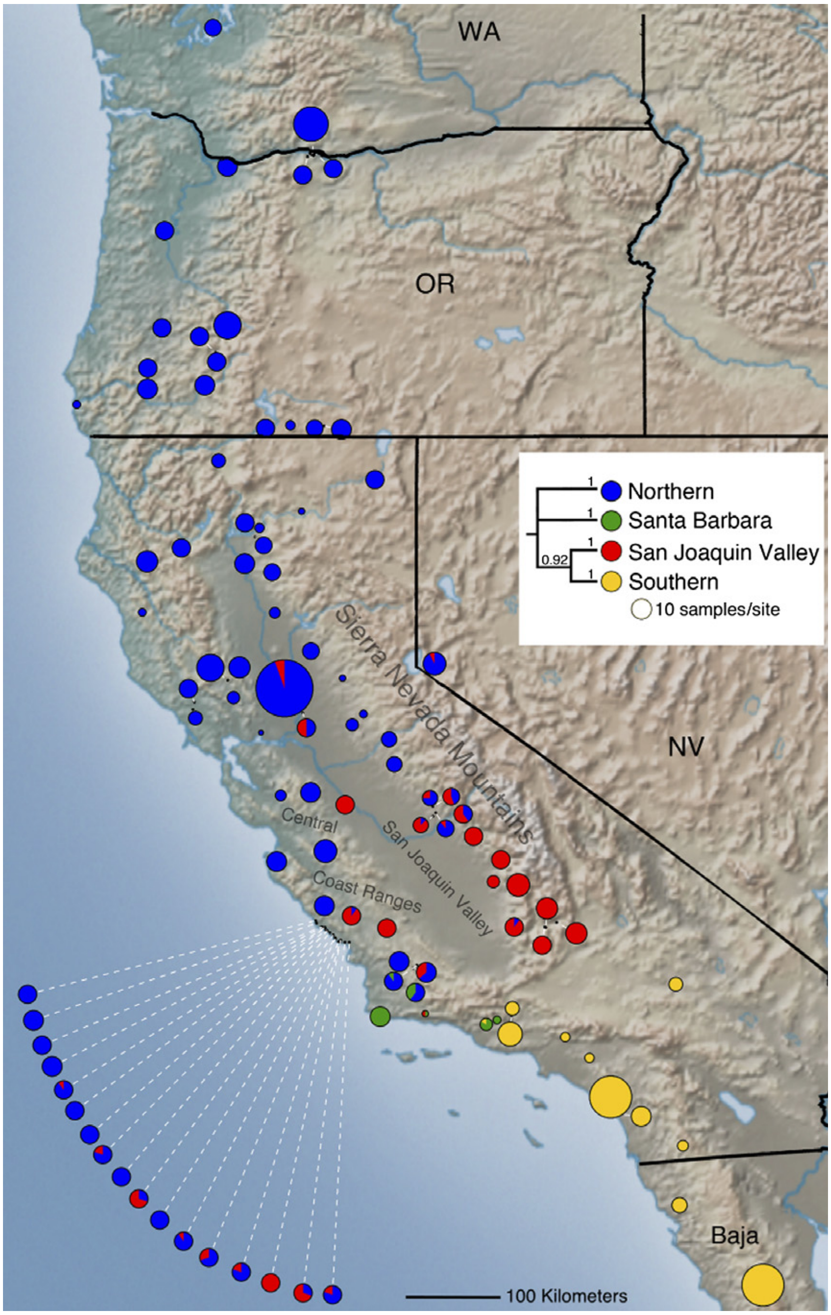

Fig. 1 Map showing mitochondrial DNA (mtDNA) clade membership based on analyses of *ND4* sequence data (Appendix S1, Supporting Information). Numbers above branches are Bayesian posterior probabilities. Circle diameter corresponds to number of samples/site. The map was generated using the GenGis software (Parks *et al.* 2009).

ern clade or San Joaquin Valley clade mitochondrial haplotypes, while the southern population south of the Tehachapi mountains that separate southern from central California exclusively carried southern clade mitochondrial haplotypes. The greatest discrepancies between data sets occurred in the central coast range, where the SNP data assigned all but a few individuals unambiguously to the southern group, yet all of those samples have northern, San Joaquin Valley or Santa Barbara clade mitochondrial haplotypes.

Secondary population assignment analyses

To more finely dissect the geographical distribution of genetic variation, we conducted secondary population assignment analyses of both the northern and southern populations (including the 16 admixed individuals). The northern individuals (n = 518) sorted into a clear pattern of K = 2 subpopulations, including a 'Cascade' group containing samples from the southern Sacramento Valley north to Washington and a 'foothill'

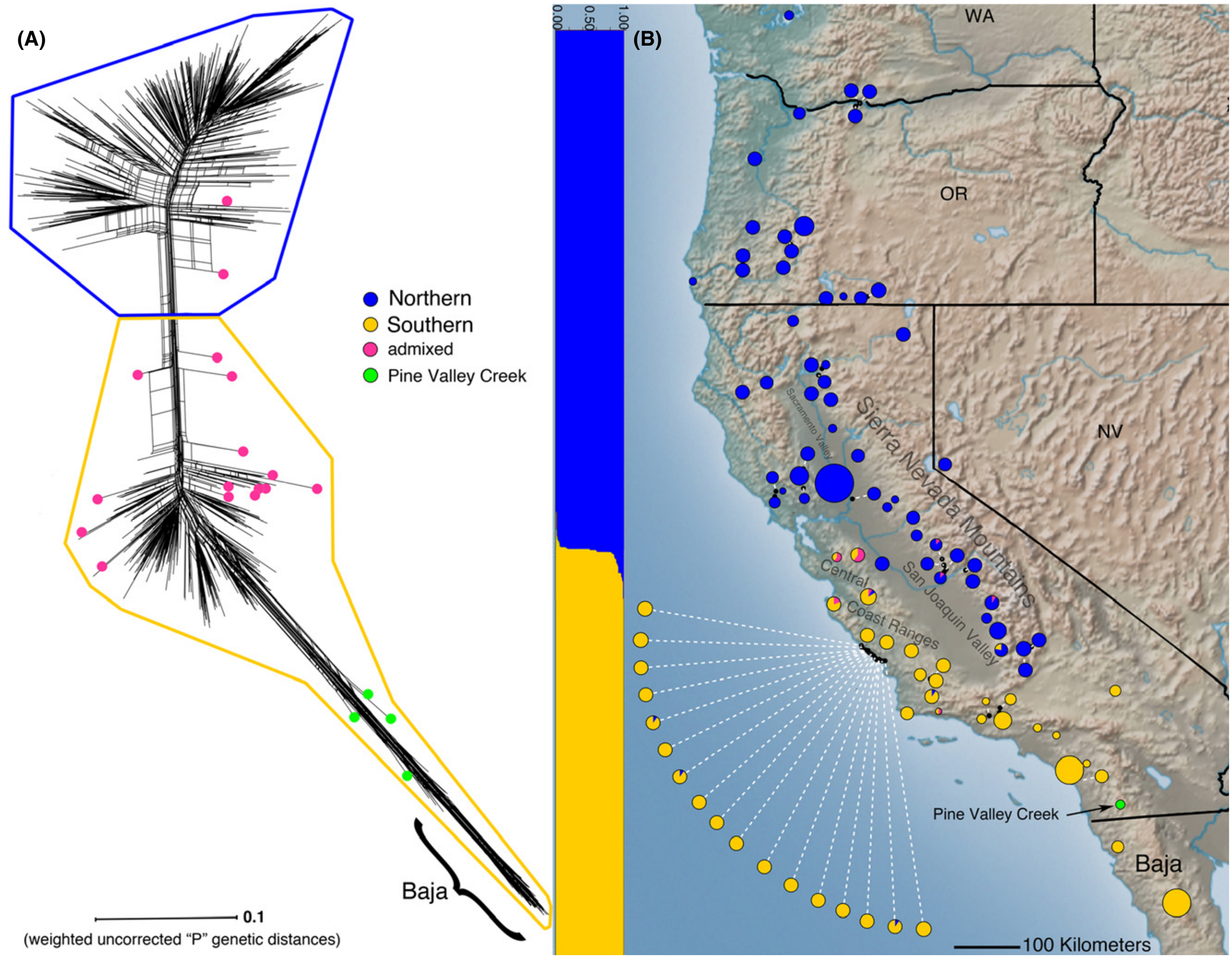

Fig. 2 Map showing results of (A) the phylogenetic network analysis and (B) the primary population assignment analyses. The phylogenetic network was generated from SNP data using all 923 individuals. This analysis recovered the northern and southern nuclear groups but also indicated a relatively deep genetic divergence of the Baja population from the southern group. Individuals that were recovered as admixed are shown with pink dots, and the individuals from Pine Valley Creek, USA, are identified with green dots. The remaining sample IDs are excluded for clarity of presentation. Results of the primary structure analyses are shown as a barplot (panel B, left-hand side) and by geographical locality (panel B, right-hand side). Circle diameters as in Figure 1. The map was generated using the GenGis software (Parks *et al.* 2009).

group comprising individuals from the San Joaquin Valley and adjacent Sierra Nevada foothills (Fig. 3). These two subpopulations intergrade along a broad contact zone from the northern central coast range across the Sacramento Valley. The Nevada population is a combination of Cascade and foothill population individuals (and admixtures between them), consistent with early records suggesting that they were introduced from California (Cary 1889).

Additional analyses of the southern population (n = 405) also recovered K = 2 subpopulations, including a CCSC group and a 'Baja' group. The CCSC group included samples from the central coast range from the San Francisco Bay Area south to the US/Mexico border including the Mojave population. Pure members of the Baja group were restricted to our limited samples from the Sierra San Pedro Martir in Baja California, while most populations from San Diego County and adjacent northern Baja California contained admixed CCSC and Baja group individuals. Pine Valley Creek, a tributary of the Tijuana River watershed in southern San Diego Co., California, contained only admixed individuals (Figs 2 and 3).

Species delimitation

Based on the SNP Structure analyses, we hypothesized that the *Emys marmorata* complex could consist of four species under a general lineage species concept (de Queiroz 1998). We tested this hypothesis under the

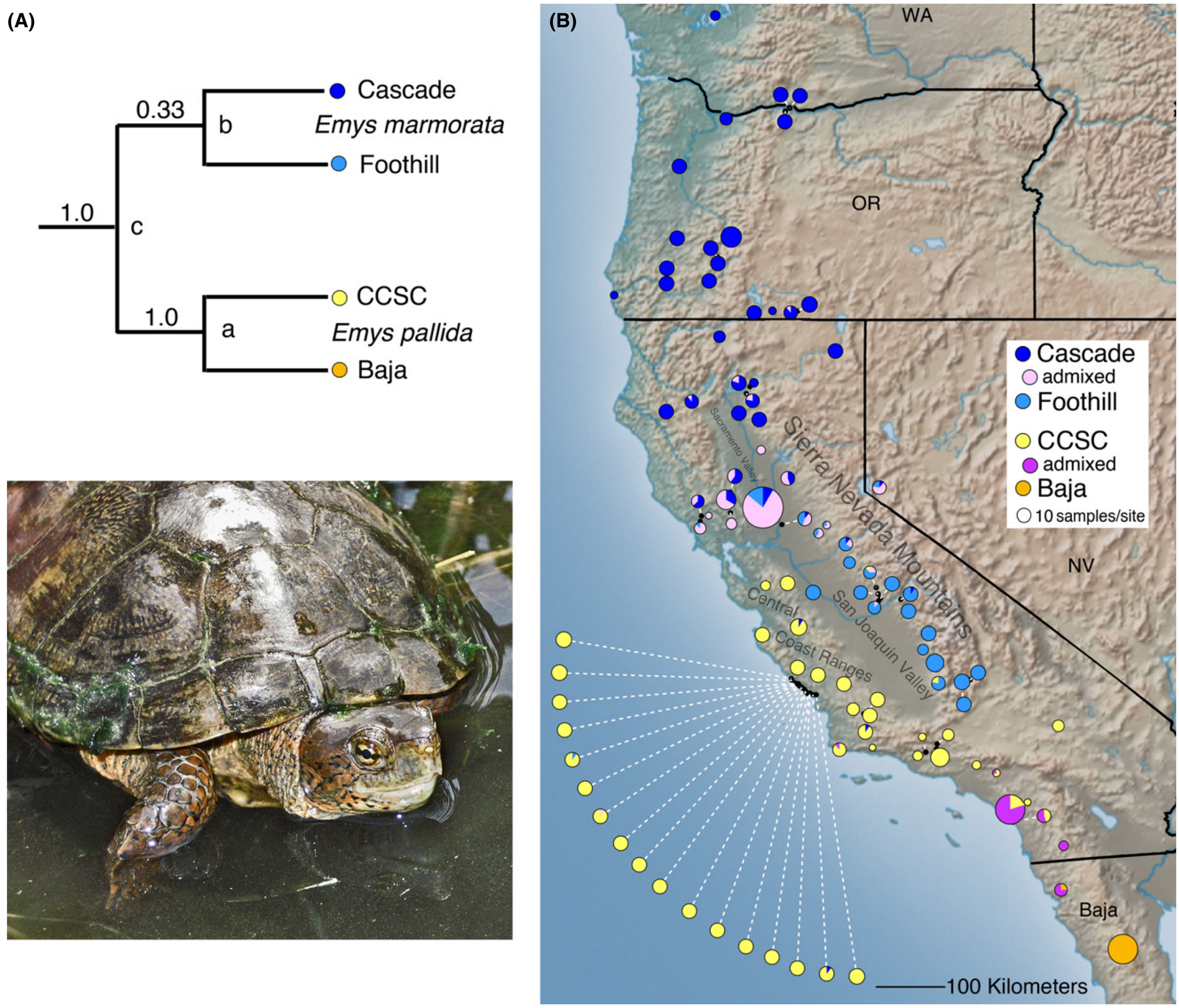

Fig. 3 Map showing results of (A) top panel: the BPP species delimitation analyses. Bottom panel: *Emys pallida*, Lusardi Creek, San Diego Co., California (photo courtesy of Robert Fisher), and (B) the secondary population assignment analyses. For species delimitation (panel A, top), we tested the hypotheses that northern (*E. marmorata*) and southern (*E. pallida*) populations, as well as their contained subpopulations, could be considered distinct species. Posterior probabilities for divergence are shown above branches. The map was generated using the GenGis software (Parks *et al.* 2009).

multispecies coalescent model using an independent nuclear sequence data set [the 90-taxon, 5-locus nuclear sequence data set from Spinks *et al.* (2010)]. By specifying the guide tree [(CCSC, Baja), (Cascade, foothill)] implied by the Structure analysis of our SNP data, we assessed the posterior probability that (i) the northern/southern population division and (ii) the two more recent divergences within each of these populations represent speciation events using BPP (Yang & Rannala 2010). These analyses returned strong support for species-level divergence between (i) northern and southern clusters (the first, $K = 2$ division identified in Structure) and (ii) between CCSC and Baja California subpopulations within the southern group. Support for these divergences was uniformly high and robust to alternative choice of priors, algorithmic settings and varying sample size experiments across populations (PP > 0.97 across all analysis settings; see Materials and methods) (Table S1, Supporting Information). The Cascade vs. foothill divergence, however, garnered little support (Fig. 3, Table S1, Supporting Information).

Population splits and mixture

Analyses of our SNP data using Treemix revealed two key results. First, the topology of the ML tree recovered from analyses of the SNP data was identical to our hypothesized guide tree. The Cascade and foothill groups are sister clades with strong support (Fig. 4), the CCSC and Baja groups are sister clades, but with weak support, and there is a deep divergence between the (Cascade + foothill) group and the (CCSC + Baja) group that is well supported (Fig. 4). These analyses also revealed up to three possible migration events, although by far the greatest weight was assigned to the migration event from the foothill group to the CCSC group (Fig. 4). Results from additional analyses with admixed individuals or the SNP loci (*NB06374*, *NB22443*) excluded (not shown) were essentially identical to analyses of the full data set, suggesting that these individuals and loci are not driving the inferred migration. A phylogenetic network-based examination of the SNP data also recovered the primary northern/southern split, and recovered the Baja California population as strongly divergent from, but nested within the southern group (Fig. 2).

Discussion

Phylogeography was built upon single gene analyses of mtDNA (Avise *et al.* 1987), although the potential problems associated with population inferences based on single loci were raised early on (Moore 1995; Tajima 1983). Only recently have larger amounts of informative nuclear data been used in phylogeographical analyses, and comparative studies that directly assess the adequacy of single- or few-locus studies have yet to be conducted. Less emphasized in the literature is the importance of deep vs. sparse taxon sampling in recovering meaningful phylogeographical patterns, even though the importance of doing so is becoming increasingly evident (e.g. Fijarczyk *et al.* 2011). Our nuclear SNP analyses revealed a strikingly different population structure across the range of the western pond turtle than what we previously inferred from single markers, while our sevenfold increase in taxon sampling for a single marker did little to change this previous understanding. Our analyses of these new data highlight the importance of deep, multilocus sampling in phylogeographical analyses and add new dimensions to our understanding of the evolutionary history of our case study.

Overall, the pattern recovered with extensive mtDNA sequence data (983 individuals) was virtually identical to that found using a much smaller data set of 147 individuals (Spinks *et al.* 2010), suggesting that increased taxon sampling for mtDNA data had little impact on phylogeographical conclusions, at least with the 725-bp section of the mitochondrial genome that we sampled (Fig. S3, Supporting Information). However, analyses of 89 independent SNP markers (see Appendix S2, S3 and Fig. S1, Supporting Information) enabled a much more comprehensive analysis of the complex historical processes leading to current biogeographical patterns, which will ultimately lead to more informed conservation and management decisions. Our overarching conclusion is that the gains from a modest (~100 marker) population genomic analysis is worth the effort, particularly for taxa of conservation importance, and that we should move away from analyses based on sparse genetic and population sampling when possible (Dupuis *et al.* 2012).

Phylogeographical insights into the biogeography of the western pond turtle complex

Analyses of our SNP data revealed two temporally distinct sets of biogeographical events in the western pond turtle: (i) a relatively ancient, deep divergence between populations from the San Joaquin Valley north to Washington and those from the central coast range south to Mexico (Figs 2–4) and (ii) more recent population subdivision across central California and northern Mexico (Fig. 3). When compared with earlier work, certain patterns have been consistent across all analyses to date (Seeliger 1945; Spinks & Shaffer 2005; Spinks *et al.* 2010). Turtles from Los Angeles and south, and those from San Francisco and north, have uniformly been identified as separate entities referred to as *E. m. pallida* and *E. m. marmorata*, respectively (although Seeliger (1945) felt that Baja turtles were sufficiently different from *pallida* that she did not include them in her morphological analyses). However, for the intervening populations from central California, multiple data sets and the power of a large SNP panel provide unique insights into the phylogeography of this system. At the SNP level, turtles inhabiting the central coast range are almost all southern *pallida*. Spinks *et al.* (2010) suggested that the central coast range populations probably were initially carrying southern mitotypes until northern and San Joaquin Valley mtDNA swept through the region following the closure of an inland seaway. The TREEMIX results reported here, based purely on the nuclear SNP data, are consistent with this hypothesis and recovered an ancestral migration event from the foothill into the CCSC clade (Fig. 4); this result is consistent regardless of whether admixed individuals are retained or excluded from the analysis. Thus, the joint analysis of a large, informative SNP data set combined with mtDNA and nuDNA sequence data suggests that extensive mitochondrial, and more limited nuclear, introgression has been prevalent in this region, with westward movement of mtDNA from the San Joaquin Valley and southward movement from northern *marmorata* into the central coast range populations of

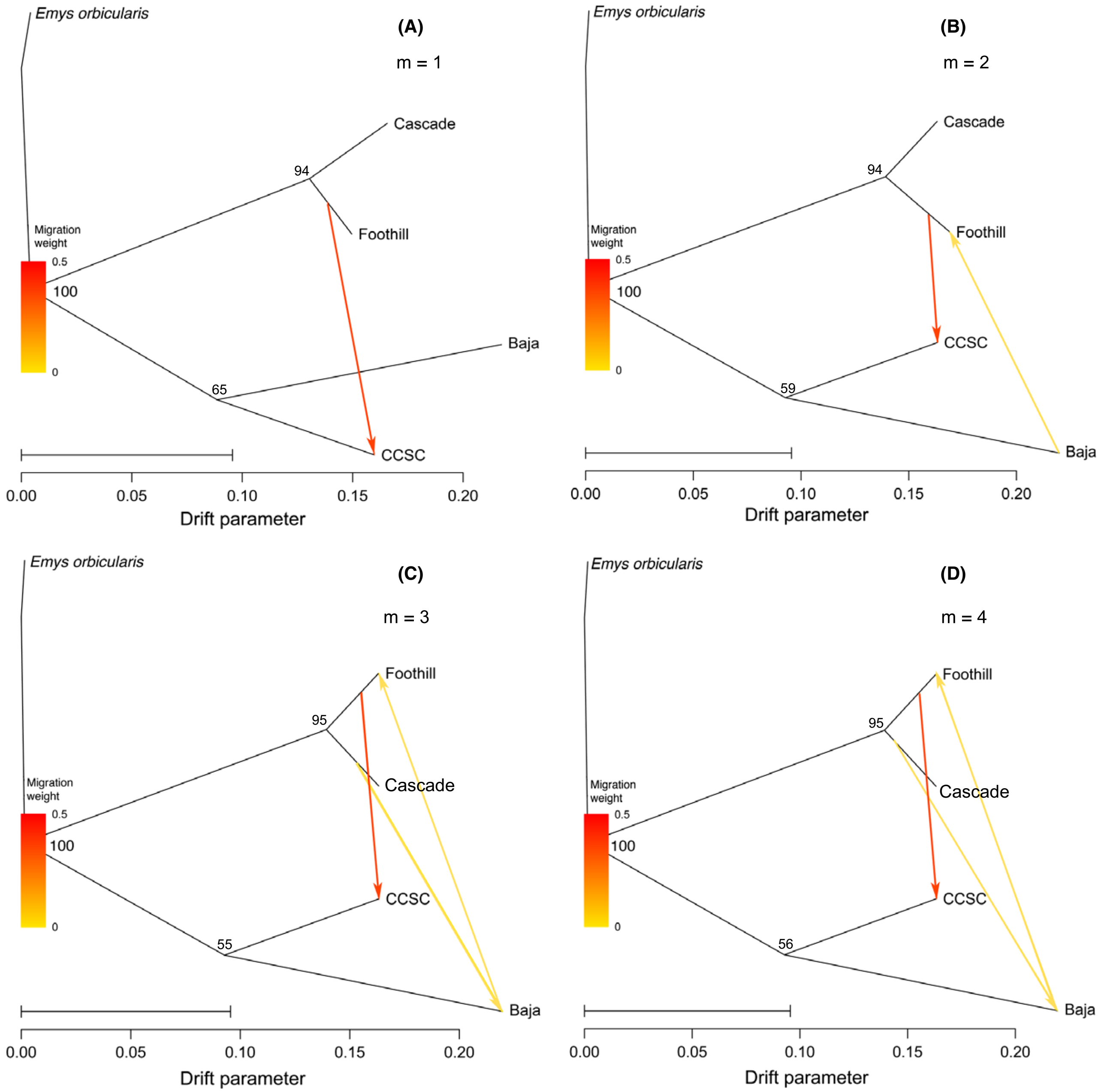

Fig. 4 Maximum-likelihood trees generated with the TREEMIX software. Graphs depict splits among Cascade, foothill, CCSC and Baja groups, allowing migration events. (A) With one migration event, migration is indicated as migrants from the foothill to the CCSC group. (B) With two migration events allowed, one is as in A, and an additional migration event is reconstructed from the Baja to the foothill population. With three (C) or four (D) migration events allowed, the first two remain as in B, and a single additional migration event from the Baja group to the foothill group is indicated. Numbers at nodes indicate bootstrap support values. The edge colour indicates the weight of the migration event measured as the fraction of alleles coming from the parental population, from red (high weight) to yellow (low weight). In addition, the direction of gene flow between populations is indicated with an arrow.

pallida (Figs 1, 2, 4). This relatively recent mtDNA pattern conceals the more ancient phylogeographical history of turtles in the central coast range and incorrectly suggests that the taxon *pallida* is restricted to southern California rather than ranging north to the San Francisco Bay region.

Taxonomy and conservation

The revised perspective provided by our SNP nuclear data brings a new level of clarity and precision to our interpretation of the evolutionary history and conservation of this species complex. Previous mitochondrial, nuclear sequence and morphological analyses (Figs S1, S2, Supporting Information) all pointed to some level of population subdivision, although the details varied leading to the subspecies names *marmorata* and *pallida* being assigned to various parts of the species' range. All previous analyses have identified a broad region of intergradation in either the central coast range (molecular analyses) or the San Joaquin Valley (morphology), leading to uncertainty over the appropriate taxonomy and conservation priorities for a large part of the range of the *Emys marmorata* complex. Given the severe declines seen in many of these regions, a clear understanding of population histories and the resulting taxonomy is critical for effective management. It now seems clear that two primary clades exist and that subpopulations within each have been incorrectly interpreted as intergrades.

The conflict between nuclear and mitochondrial patterns in the central coast range apparently represents a relatively recent mitochondrial sweep, although evidence of that sweep is generally not reflected in nearly 90 markers distributed across the nuclear genome. Our current interpretation is that the central coast range contains essentially pure *pallida* animals, albeit with *marmorata* mitochondria. The exceptions to this overall pattern include two southern clade individuals recovered in the southern Sierra Nevada foothills and five northern clade samples found along the central coast. These individuals may represent human-mediated translocations/pet releases, a phenomenon that regularly occurs in turtles (Carr 1952; Storer 1930). Whether they are native or translocated, the presence of pure northern and southern group turtles at these sites with no currently measurable admixture suggests that they can exist in microsympatry without hybridization.

Population assignment and network analysis of the SNP data, and species delimitation analyses of nuDNA sequence data, all strongly support three well-differentiated groups (Figs 2–4) with limited admixture, in contrast to four (mitochondrial DNA) or potentially three (morphology) taxa with extensive hybridization that have been previously suggested. Populations from Baja are particularly important, as these turtles have commonly been assigned to *E. m. pallida* following Carr (1952). Seeliger (1945) did not assign turtles from Baja California to either *marmorata* or *pallida* because these turtles were 'not similar to the northern or southern forms'. Thus, Seeliger (1945) implicitly identified not two, but three morphologically distinct units, and our SNP data support her interpretation. She also identified the central coast populations as pure *pallida*, a result that is again supported by our SNP data.

The combined information from SNP and morphological data led us to recommend that at least two species be recognized within the *Emys marmorata* complex. The northern nuclear group closely corresponds to *E. m. marmorata* and the southern nuclear group (possibly excluding Baja California samples) closely corresponds to *E. m. pallida*. These two groups were recovered using multiple analytical methods and data sets, are statistically well supported under a model of multispecies coalescence (Fig. 3) and should be recognized as distinct species under a general lineage species concept (de Queiroz 1998). The type specimens and morphological descriptions of *E. marmorata* and *E. pallida* are presented by Seeliger (1945). We propose using the name *Emys marmorata* for all populations north of the San Francisco Bay area plus populations from the Great Central Valley north including the apparently introduced Nevada population (Fig. 2, labelled as Northern). *Emys pallida* is restricted to those populations inhabiting the central coast range south of the San Francisco Bay area to the species' southern range boundary, including the Mojave River (Fig. 2, labelled as Southern). *Emys marmorata* and *Emys pallida* show very limited intergradation in a few populations in the northern central coast range and adjacent Sierra Nevada foothills, although at all intergrade sites we also found pure individuals of the locally prevalent species. Although we tentatively include populations from Baja California in *E. pallida*, we also recognize that these animals may represent a distinct species pending results from additional analyses.

Pond turtles from southern California are in precipitous decline, with few stable, reproducing populations known between Los Angeles and the US/Mexico border. The recognition of *E. pallida* as a distinct species and the possibility that stable populations in Baja California represent a unique evolutionary lineage emphasize the critical need for immediate conservation in southern California and Baja California, Mexico.

Concluding thoughts

Few studies allow a direct comparison of sparse and dense population and gene sampling in the same species and landscapes. Our results indicate that there is little to be gained by increasing individual sampling beyond comprehensive geographical sampling for single markers, and our mitochondrial results for approximately 100 and 1000 turtles were very similar. However, deep population and individual sampling for ~90 informative SNPs allowed us to identify both recently and more anciently derived lineages, novel management and conservation units, previously obscure patterns of admixture across the landscape, and instances of microsympatry between the two species, none of which were apparent with sparse nuclear sampling. Whether increasing the size of a nuclear marker panel by another order of magnitude would provide additional increases in resolution is an open question that is probably worthy of investigation. However, it is clear that genome-enabled approaches to phylogeography have a great deal to offer, especially for systems of high conservation concern.

Acknowledgements

For tissue samples, we thank Dan Holland, Jamie Bettaso, Kat Caldera, Ed DeGrauw, Tibor Kovács, Frank Slavens, Kate Slavens and Teri Herman. We thank Evan McCartney-Melstead for contributing the Perl script and help with data formatting. We thank Joseph Pickrell for advice on using TREEMIX and Bruce Rannala for advice on using BPP. We thank Anastasia Krivoruchko for prepublication access to primer sequences and Vanessa Rashbrook for SNP genotyping. This work was supported in part by funding from NSF-DEB 0817042, DEB 1239961 and the California Department of Fish and Wildlife.

References

Avise JC (2000) *Phylogeography*. Harvard University Press, Cambridge, Mass.

Avise JC, Arnold J, Ball RM *et al.* (1987) Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. *Annual Review of Ecology and Systematics*, **18**, 489–522.

Backström N, Fagerberg S, Ellegren H (2008) Genomics of natural bird populations: a gene-based set of reference markers evenly spread across the avian genome. *Molecular Ecology*, **17**, 964–980.

Barley AJ, Spinks PQ, Thomson RC, Shaffer HB (2010) Fourteen nuclear genes provide phylogenetic resolution for difficult nodes in the turtle tree of life. *Molecular Phylogenetics and Evolution*, **55**, 1189–1194.

Berlin S, Quintela M, Höglund J (2008) A multilocus assay reveals high nucleotide diversity and limited differentiation among Scandinavian willow grouse (*Lagopus lagopus*). *BMC Genetics*, **9**, 89.

Bryant D, Moulton V (2004) Neighbor-Net: an agglomerative method for the construction of phylogenetic networks. *Molecular Biology and Evolution*, **21**, 255–265.

Carr AF (1952) *Handbook of Turtles. The Turtles of the United States, Canada, and Baja California*, p. 542. Cornell University Press, Ithaca, NY.

Cary W (1889) *Biennial report of the fish commissioner of the state of Nevada. Appendix to the Journals of the Senate and Assembly 14th session*. pp. 3–7.

Chan YC, Roos C, Inoue-Murayama M *et al.* (2010) Mitochondrial genome sequences effectively reveal the phylogeny of *Hylobates* gibbons. *PLoS ONE*, **5**, e14419.

Dimmick WW, Ghedotti MJ, Grose MJ, Maglia AM, Meinhardt DJ, Pennock DS (1999) The importance of systematic biology in defining units of conservation. *Conservation Biology*, **13**, 653–660.

Drummond AJ, Ashton B, Buxton S *et al.* (2010) *Geneious v5.3*. Available at: http://www.geneious.com.

Dupuis JR, Roe AD, Sperling FAH (2012) Multi-locus species delimitation in closely related animals and fungi: one marker is not enough. *Molecular Ecology*, **21**, 4422–4436.

Earl DA, vonHoldt BM (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources*, **4**, 359–361.

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software Structure: a simulation study. *Molecular Ecology*, **14**, 2611–2620.

Fijarczyk A, Nadachowska K, Hofman S *et al.* (2011) Nuclear and mitochondrial phylogeography of the European fire-bellied toads *Bombina bombina* and *Bombina variegata* supports their independent histories. *Molecular Ecology*, **20**, 3381–3398.

Fraser DJ, Bernatchez L (2001) Adaptive evolutionary conservation: towards a unified concept for defining conservation units. *Molecular Ecology*, **10**, 2741–2752.

Friesen VL, Congdon BC, Kidd MG, Birt TP (1999) Polymerase chain reaction (PCR) primers for the amplification of five nuclear introns in vertebrates. *Molecular Ecology Notes*, **8**, 2141–2152.

Fujita MF, Engstrom TN, Starkey DE, Shaffer HB (2004) Turtle phylogeny: insights from a novel nuclear intron. *Molecular Phylogenetics and Evolution*, **31**, 1031–1040.

Glez-Peña D, Gómez-Blanco D, Reboiro-Jato M, Fdez-Riverola F, Posada D (2010) *ALTER: program-oriented format conversion of DNA and protein alignments*. Nucleic Acids Research. Web Server issue. ISSN: 0305–1048.

Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, **52**, 696–704.

Handley LJL, Ceplitis H, Ellegren H (2004) Evolutionary strata on the chicken Z chromosome: implications for sex chromosome evolution. *Genetics*, **167**, 367–376.

Harshman J, Huddleston CJ, Bollback JP, Parsons TJ, Braun MJ (2003) True and false gharials: a nuclear gene phylogeny of Crocodylia. *Systematic Biology*, **52**, 386–402.

Heath TA, Hedtke SM, Hillis DM (2008) Taxon sampling and the accuracy of phylogenetic analyses. *Journal of Systematics and Evolution*, **46**, 239–257.

Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogeny. *Bioinformatics*, **17**, 754–755.

Huson DH, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, **23**, 254–267.

International Chicken Genome Sequencing Consortium (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*, **432**, 695–716.

Jombart T (2008) adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics*, **24**, 1403–1405.

Kent WJ (2002) BLAT—the BLAST-like alignment tool. *Genome Research*, **12**, 656–664.

Kimball RT, Braun EL, Barker FK *et al.* (2009) A well-tested set of primers to amplify regions spread across the avian genome. *Molecular Phylogenetics and Evolution*, **30**, 654–660.

Knaus BJ, Cronn R, Liston A, Pilgrim K, Schwartz MK (2011) Mitochondrial genome sequences illuminate maternal lineages of conservation concern in a rare carnivore. *BMC Ecology*, **11**, 10.

Krenz JG, Naylor GJP, Shaffer HB, Janzen FJ (2005) Molecular phylogenetics and evolution of turtles. *Molecular Phylogenetics and Evolution*, **37**, 178–191.

Krivoruchko A, Storey KB (2010) Molecular mechanisms of turtle anoxia tolerance: a role for NF-κB. *Gene*, **450**, 63–69.

Lyons LA, Laughlin TF, Copeland NG, Jenkins NA, Womack JE, O'Brien SJ (1997) Comparative anchor tagged sequences (CATS) for integrative mapping of mammalian genomes. *Nature Genetics*, **15**, 47–56.

Mayden RL, Wood RM (1995) Systematics, Species Concepts, and the Evolutionarily Significant Unit in Biodiversity and Conservation Biology. In: *Evolution and the Aquatic Ecosystem: Defining Unique Units in Population Conservation*. Symposium 17 (eds Nielsen JL), pp. 58–113. American Fisheries Society, Bethesda, MD.

McCormack JE, Maley JM, Hird SM, Derryberry EP, Graves GR, Brumfield RT (2012) Next-generation sequencing reveals phylogeographic structure and a species tree for recent bird divergences. *Molecular Phylogenetics and Evolution*, **62**, 397–406.

Merz C, Catchen JM, Hanson-Smith V, Emerson KJ, Bradshaw WE, Holzapfel CM (2013) Replicate phylogenies and post-glacial range expansion of the pitcher-plant mosquito, *Wyeomyia smithii*, North America. *PLoS ONE*, **8**, e72262.

Morin PA, Luikart G, Wayne RK, the SNP workshop Group (2004) SNPs in ecology, evolution and conservation. *Trends in Ecology & Evolution*, **19**, 208–216.

Moore WS (1995) Inferring phylogenies from mtDNA variation: mitochondrial-gene trees versus nuclear-gene trees. *Evolution*, **49**, 718–726.

Moritz C (1994) Defining 'evolutionarily significant units' for conservation. *Trends in Ecology & Evolution*, **9**, 373–375.

Nei M (1972) Genetic distances between populations. *The American Naturalist*, **106**, 283–292.

Noonan BP, Chippindale PT (2006) Vicariant origin of Malagasy reptiles supports late cretaceous Antarctic land bridge. *The American Naturalist*, **168**, 730–741.

Nylander JAA (2002) *MRMODELTEST v2.1. Department of Systematic Zoology*. Uppsala University, Available from author.

Nylander JA, Wilgenbusch JC, Warren DL, Swofford DL (2008) AWTY (are we there yet?): a system for graphical exploration of MCMC convergence in Bayesian phylogenetics. *Bioinformatics*, **24**, 581–583.

Parks DH, Porter M, Churcher S *et al.* (2009) GenGIS: a geospatial information system for genomic data. *Genome Research*, **19**, 1896–1904.

Pease KM, Freedman AH, Pollinger JP *et al.* (2009) Landscape genetics of California mule deer (*Odocoileus hemionus*): the roles of ecological and historical factors in generating differentiation. *Molecular Ecology*, **18**, 1848–1862.

Pickrell JK, Pritchard JK (2012) Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genetics*, **8**, e1002967.

Primmer CR, Borge T, Lindell J, Sætre GP (2002) Single nucleotide polymorphism characterization in species with limited available sequence information: high nucleotide diversity revealed in the avian genome. *Molecular Ecology*, **11**, 603–612.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–949.

de Queiroz K (1998) The general lineage concept of species, species criteria, and the process of speciation. In: *Endless Forms. Species and Speciation* (eds Howard DJ & Berlocher SH), pp. 57–75. Oxford University Press, Oxford.

Rambaut A, Drummond AJ (2009) Tracer v15. Available at: http://beastbioedacuk/Tracer.

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

Ronquist F, Huelsenbeck JP (2003) MRBAYES 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, **19**, 1572–1574.

Ryder OA (1986) Species conservation and systematics: the dilemma of subspecies. *Trends in Ecology & Evolution*, **1**, 9–10.

Schulte JA II, Cartwright EM (2009) Phylogenetic relationships among iguanian lizards using alternative partitioning methods and *TSHZ1*: a new phylogenetic marker for reptiles. *Molecular Phylogenetics and Evolution*, **50**, 391–396.

Seeliger LM (1945) Variation in the Pacific mud turtle. *Copeia*, **1945**, 150–159.

Shaffer HB, Minx P, Warren DE *et al.* (2013) The western painted turtle genome, a model for the evolution of extreme physiological adaptations in a slowly evolving lineage. *Genome Biology*, **14**, R28.

Shamblin BM, Bjorndal KA, Bolten AB *et al.* (2012) Mitogenomic sequences better resolve stock structure of southern Greater Caribbean green turtle rookeries. *Molecular Ecology*, **21**, 2330–2340.

Spinks PQ, Shaffer HB (2005) Range-wide molecular analysis of the western pond turtle (*Emys marmorata*): cryptic variation, isolation by distance, and their conservation status. *Molecular Ecology*, **14**, 2047–2064.

Spinks PQ, Shaffer HB (2007) Conservation phylogenetics of the Asian box turtles (Geoemydidae, *Cuora*): mitochondrial introgression, numts, and inferences from multiple nuclear loci. *Conservation Genetics*, **8**, 641–657.

Spinks PQ, Thomson RC, Shaffer HB (2010) Nuclear gene phylogeography reveals the historical legacy of an ancient inland sea on lineages of the western pond turtle, *Emys marmorata* in California. *Molecular Ecology*, **19**, 542–556.

Stebbins RC (2003) *A Field Guide to Western Reptiles and Amphibians*, 3rd edn. Houghton Mifflin, New York.

Storer TI (1930) Notes on the range and life history of the Pacific fresh-water turtle, *Clemmys marmorata*. *University of California Publications in Zoology*, **32**, 429–441.

Swofford DL (2002) *PAUP* Phylogenetic Analysis Using Parsimony (*and Other Methods), Version 4*. Sinauer Associates, Sunderland, MA.

Tajima F (1983) Evolutionary relationship of DNA sequences in finite populations. *Genetics*, **105**, 437–460.

Thomson RC, Shedlock AM, Edwards SV, Shaffer HB (2008) Developing markers for multilocus phylogenetics in non-model organisms: a test case with turtles. *Molecular Phylogenetics and Evolution*, **49**, 514–525.

Townsend TM, Alegre RE, Kelley ST, Weins JJ, Reeder TW (2008) Rapid development of multiple nuclear loci for phylogenetic analysis using genomic resources: an example from squamate reptiles. *Molecular Phylogenetics and Evolution*, **47**, 129–142.

Walstrom V, Klicka J, Spellman GM (2012) Speciation in the White-breasted Nuthatch (Sitta carolinensis): a multilocus perspective. *Molecular Ecology*, **21**, 907–920.

Yang Z, Rannala B (2010) Bayesian species delimitation using multilocus sequence data. *PNAS*, **107**, 9264–9269.

Zhang Z, Schwartz S, Wagner L, Miller W (2000) A greedy algorithm for aligning DNA sequences. *Journal of Computational Biology*, **7**, 203–214.

Zwickl DJ, Hillis DM (2002) Increased taxon sampling greatly reduces phylogenetic error. *Systematic Biology*, **51**, 588–598.

P.Q.S. and H.B.S. conceived the study; P.Q.S. and R.C.T. collected some samples; P.Q.S. and R.C.T. performed the phylogenetic and population genetics analyses; all authors contributed to writing the paper.

Data accessibility

Data and associated data files generated for this analysis are available in the Molecular Ecology online supplementary Appendices and are deposited in the Dryad digital repository (http://datadryad.org) under accession number doi:10.5061/dryad.pr907. All sequences generated here are available from GenBank: Accession nos KJ580956–KJ582521.

Supporting information

Additional supporting information may be found in the online version of this article.

Fig. S1 Map showing collection localities of samples included in this analysis labelled according to Seeliger's (1945) subspecies demarcations.

Fig. S2 Map showing results from previous phylogeographical analyses of *Emys marmorata* (after Spinks *et al.* (2010)).

Fig. S3 Majority-rule consensus of the posterior distribution of trees from the Bayesian analysis of the 81-taxon *ND4* data set.

Table S1 Means for population size (θ) and divergence time estimates (τ), and posterior probabilities (PP), from the BPP analyses of the five subsampled data sets (see text).

Appendix S1 Table showing Catalogue number (HBS no), GenBank ND4 Accession no, locality information, mitochondrial clade membership, species designation, population assignment scores (posterior probabilities) from the Structure analyses and genotypes for the 925 turtles genotyped for this analysis.

Appendix S2 Table showing results of Blat (Kent 2002) searches of *Emys marmorata* sequences to *Chrysemys picta* genomic data using the UCSC Genome Browser (Kent *et al.* 2002). Available at http://genome.ucsc.edu/index.html.

Appendix S3 Table showing Blast (Zhang *et al.* 2000) search results of *Emys marmorata* sequences to *Chrysemys picta* genomic data.

Appendix S4 Table showing PCR annealing temperature, primer source and forward and reverse primer sequences for 104 nuclear markers assessed for this analysis.