MOLECULAR ECOLOGY

Molecular Ecology (2012) doi: 10.1111/mec.12049

Parallel tagged amplicon sequencing reveals major lineages and phylogenetic structure in the North American tiger salamander (*Ambystoma tigrinum*) species complex

ERIC M. O'NEILL,* RACHEL SCHWARTZ,†‡ C. THOMAS BULLOCK,§ JOSHUA S. WILLIAMS,* H. BRADLEY SHAFFER,¶** X. AGUILAR-MIGUEL,†† GABRIELA PARRA-OLEA‡‡ and DAVID W. WEISROCK*

**Department of Biology, University of Kentucky, Lexington, KY 40506, USA, †Department of Biology, Colorado State University, Fort Collins, CO 80523, USA, ‡The Biodesign Institute, Arizona State University, Tempe, AZ 85287, USA, §Department of Computer Science, University of Kentucky, Lexington, KY 40506, USA, ¶Department of Ecology and Evolutionary Biology, University of California, Los Angeles, CA 90095, 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, ††CIRB, Facultad de Ciencias, Universidad Autónoma del Estado de México, Toluca, Edo. de México, México, ‡‡Instituto de Biología, Universidad Nacional Autónoma de México, Distrito Federal, México*

Abstract

Modern analytical methods for population genetics and phylogenetics are expected to provide more accurate results when data from multiple genome-wide loci are analysed. We present the results of an initial application of parallel tagged sequencing (PTS) on a next-generation platform to sequence thousands of barcoded PCR amplicons generated from 95 nuclear loci and 93 individuals sampled across the range of the tiger salamander (*Ambystoma tigrinum*) species complex. To manage the bioinformatic processing of this large data set (344 330 reads), we developed a pipeline that sorts PTS data by barcode and locus, identifies high-quality variable nucleotides and yields phased haplotype sequences for each individual at each locus. Our sequencing and bioinformatic strategy resulted in a genome-wide data set with relatively low levels of missing data and a wide range of nucleotide variation. STRUCTURE analyses of these data in a genotypic format resulted in strongly supported assignments for the majority of individuals into nine geographically defined genetic clusters. Species tree analyses of the most variable loci using a multi-species coalescent model resulted in strong support for most branches in the species tree; however, analyses including more than 50 loci produced parameter sampling trends that indicated a lack of convergence on the posterior distribution. Overall, these results demonstrate the potential for amplicon-based PTS to rapidly generate large-scale data for population genetic and phylogenetic-based research.

Keywords: barcode, bioinformatic, gene tree, next-generation sequencing, nuclear DNA

Received 21 May 2012; revision received 10 August 2012; accepted 21 August 2012

Introduction

In a relatively short period of time, the population genetic, phylogeographic and phylogenetic communities have begun to embrace newly developed massively parallel sequencing methods (Kircher & Kelso 2010; Glenn 2011; McCormack *et al.* 2012a)—widely dubbed next-generation sequencing (NGS)—as effective ways to generate the types of data sets that are needed for robustly resolving evolutionary history. This has been motivated in part by a growing understanding that: (i) accurate

Correspondence: Eric M. O'Neill, Fax: 859-257-1717; E-mail: eric.oneill@uky.edu

estimates of population-level parameters require data from multiple loci (e.g. Carling & Brumfield 2007) and (ii) discordance among gene trees is often expected (Maddison 1997). Recognition of the latter point has resulted in a shift away from traditional approaches that utilize concatenated data sets or focus on consensus among gene trees and towards approaches that seek reconciliation among discordant gene trees within a single species phylogeny (Brito & Edwards 2009; Edwards 2009). In addition, advances in algorithms and software have made both population genetic and phylogenetic analyses more tractable for large multi-locus data sets (Hey & Nielsen 2007; Liu & Pearl 2007; Heled & Drummond 2010; O'Meara 2010; Yang & Rannala 2010), providing further encouragement to the empirical geneticist that robust reconstructions of population and evolutionary history are within their grasp.

To date, the use of NGS methods in population genetics and phylogenetics has largely relied on random 'shotgun' sequencing of genomic fragments generated from some form of a reduced representation technique [e.g. CRoPS (van Orsouw *et al.* 2007), modified AFLP protocols (Gompert *et al.* 2010) and RAD tag sequencing (Baird *et al.* 2008)]. While these methods generally result in very large data sets, random sequencing of DNA fragments and mutations in restriction enzyme recognition sites often result in large numbers of missing orthologs among samples, particularly for species with large and complex genomes (e.g. salamanders). In addition, most NGS efforts have thus far focused on generating data in the form of short sequence reads (e.g. 100 bp) and the identification of one or a few single-nucleotide polymorphisms (SNPs) per read. Although these data have proved useful for population genetic analyses (e.g. Hohenlohe *et al.* 2010; Forister *et al.* 2011; Zellmer *et al.* 2012), longer sequences, which allow for easier phasing of multiple linked SNPs, are important for accurately reconstructing gene trees and adequately accounting for gene tree discordance when reconstructing species trees (Huang *et al.* 2010).

Targeted sequencing of specific loci using NGS platforms may provide a more efficient means of generating data sets that overcome the shortcomings of less targeted approaches. Loci of known genomic location, orthology, size and expected level of variation can be enriched prior to NGS using well-established PCR techniques or newly developed hybridization techniques (e.g. Briggs *et al.* 2009; Gnirke *et al.* 2009; Maricic *et al.* 2010) and then pooled for high-throughput sequencing. These targeted approaches increase the sequence coverage for any individual locus and reduce the probability of missing data, which may have negative impacts on data analyses (Lemmon *et al.* 2009). The potential for a targeted NGS approach to be applied to population genetic and phylogenetic studies has been further advanced by the development of barcoding (i.e. indexing) strategies that permit pooling and subsequent parallel tagged sequencing (PTS) of multiple individual samples within a single NGS run (Binladen *et al.* 2007; Meyer *et al.* 2007; Bybee *et al.* 2011a). However, targeted PTS methods have primarily been applied to the recovery of organellar genome data (Parks *et al.* 2009; Morin *et al.* 2010; Lerner *et al.* 2011) or one to a few nuclear loci (Babik *et al.* 2009; Bybee *et al.* 2011b; Griffin *et al.* 2011; Puritz *et al.* 2012), and no study has fully explored the potential for a PTS strategy to be applied to a large number of nuclear loci sequenced from a large number of individuals.

In this study, we present results from the first large-scale population genetic and phylogenetic studies using a targeted PTS strategy. This work represents the exploratory phase of a larger project aimed at using a large set of nuclear DNA sequence markers to delimit population-level lineages and resolve phylogenetic history among salamanders of the *Ambystoma tigrinum* complex. The *A. tigrinum* complex (Shaffer & McKnight 1996) is one of the most widely distributed amphibian species complexes in North America (Fig. 1), and this complex is exceptional among amphibians in its diversity of adaptive life-history phenotypes, from populations expressing a permanently paedomorphic and aquatic phenotype to populations that obligately express a metamorphic terrestrial adult phenotype (Shaffer 1984; Collins *et al.* 1980; Shaffer & Voss 1996). Considerable attention has been paid to this species complex in studies of development (the complex contains the well-known axolotl, *Ambystoma mexicanum*), evolution and ecology; yet, phylogenetic relationships between species have remained a challenge to resolve (Shaffer 1984; Shaffer & McKnight 1996; Weisrock *et al.* 2006), limiting a comparative framework for such studies. This result may reflect a recent history of rapid bursts of speciation within the complex (Shaffer & McKnight 1996; Shaffer & Thomson 2007), which is expected to generate substantial gene tree discordance and may require phylogenetic information from many independent gene trees to reconstruct an accurate species tree (Edwards *et al.* 2007).

Of perhaps greater importance for research in these areas is the fact that species-level boundaries within the complex have been poorly explored using genetic data, with a taxonomy that primarily reflects early twentieth-century descriptive field and museum studies of morphological character variation (Frost 2008). Population-level genetic analyses are expected to be crucial for establishing the geographic boundaries of species-level genetic lineages within this complex.

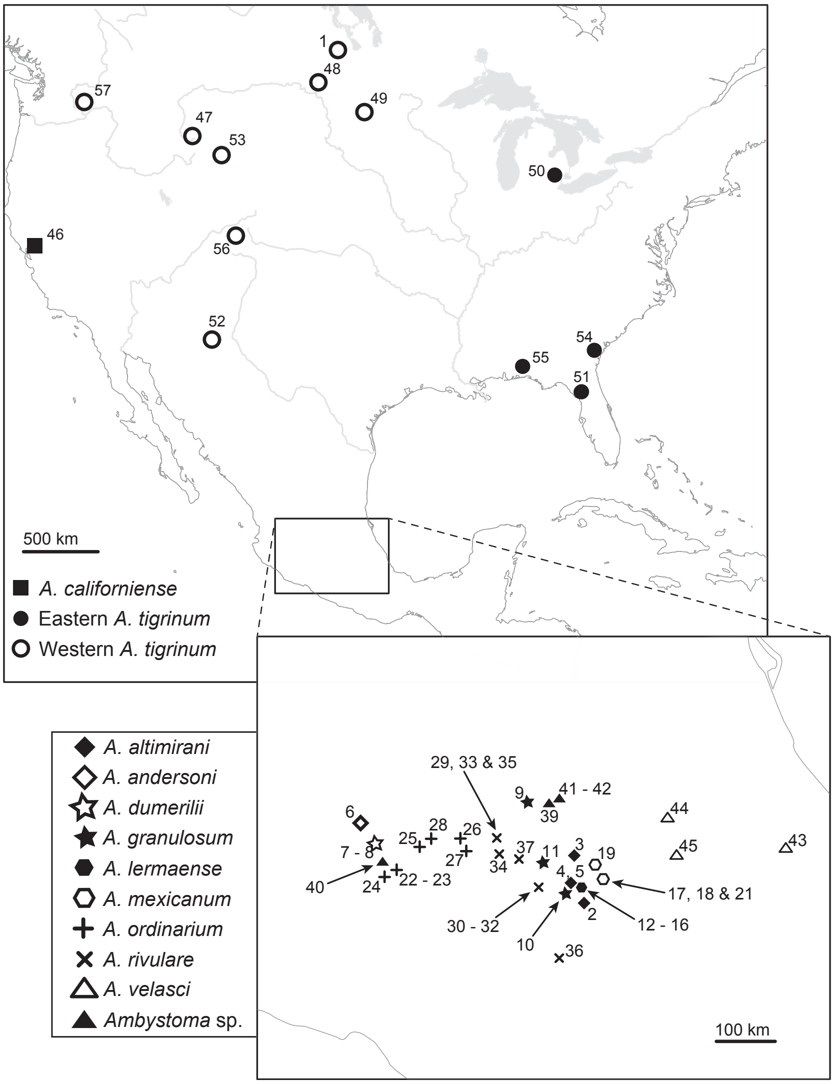

Fig. 1 North American map detailing the geographic positions for sampled localities presented in this study. The inset highlights a dense sampling within the taxonomically diverse region of central Mexico. Detailed sampling information for the numbered localities is presented in Table S1.

We begin to address the issues of species delimitation and phylogeny reconstruction by using a PTS strategy to generate sequence data for 95 unlinked nuclear DNA sequence markers sampled from 93 individuals distributed across the geographic range and described taxa of the *A. tigrinum* complex. We highlight three different components of our study that we believe will be most relevant to researchers interested in using targeted PTS in genetic studies at the population–species interface. First, we describe our use of a barcoding approach developed by Meyer *et al.* (2007, 2008) for labelling the many thousands of unique PCR amplicons generated through our study design. Second, we present the bioinformatic developments that arose from a need to process our PTS data in an automated fashion. Finally, to demonstrate the utility of our PTS data, we present the results of population genetic and phylogenetic analyses of the assembled nuclear data sets for the *A. tigrinum* complex.

Methods

Geographic sampling

To explore geographic patterns of genetic variation within the *Ambystoma tigrinum* complex, we assembled tissue samples from 93 individuals collected from 58 localities (Fig. 1; Table S1, Supporting information). Our sampling included 12 of the 18 currently recognized taxa, including the federally endangered California tiger salamander, *A. californiense*, the sister group to the remainder of the complex (Shaffer & McKnight 1996). For this study, we aimed to include 2–3 individuals per locality; however, some localities had as many as five, while others had only one. We are currently collecting data from additional samples and populations to further address fine-scaled structure within this species complex.

Locus development

We began this study with the goal of sequencing 95 nuclear loci from all sampled individuals. To develop this set of markers, we used PCR to screen a larger set of 228 loci available from an expressed sequence tag (EST) data set developed for the eastern tiger salamander (*A. t. tigrinum*) and the Mexican axolotl (*A. mexicanum*) (Putta *et al.* 2004; Smith *et al.* 2005a,b; http://www.ambystoma.org). To maximize coverage of the genome and independence of loci, we chose loci that ranged from approximately 200–650 bp in length, were widely distributed across all 14 linkage groups and were on average about 50 cM from other included loci on the *Ambystoma* linkage map (Smith *et al.* 2005a; http://www.ambystoma.org). This pool of potential loci included 26 of the loci used to study hybridization between native and introduced tiger salamanders in California (Fitzpatrick *et al.* 2009, 2010).

Using PCR, we screened each locus against a test panel of 16 representative individuals from across the range of the *A. tigrinum* complex (see below for DNA extraction and PCR conditions). These 16 individuals represent major geographic linages identified in previous mtDNA-based studies (Shaffer & McKnight 1996; Weisrock *et al.* 2006) and serve as our best initial hypotheses for genetically diverged lineages. Positive amplification across all lineages in this test panel was expected to identify nuclear markers, and their corresponding primers, that could be used successfully in all populations of the *A. tigrinum* complex. In addition, to facilitate amplification of all loci for a single individual in a single 96-well plate, we screened for loci that produced strong amplification across test individuals at a single annealing temperature of 55 °C.

Of the 228 screened loci, 125 amplified successfully in most of the 16 representative individuals. From this group, we chose a subset of 95 loci with the highest rates of amplification and the most even spacing across linkage groups with an average distance of approximately 50 cM, to be used in a larger round of data collection (Table S2, Supporting information). To confirm that these loci contain genetic variation across populations of the *A. tigrinum* complex, we randomly chose a subset of 24 loci and sequenced all 16 representative individuals using standard Sanger sequencing (Big Dye Terminator v3.1 Cycling Sequencing Kit). All 24 loci contained at least a single variable site. To generate our full data set, we PCR amplified all 95 nuclear loci from all 93 individuals (a total of 8835 individual PCRs).

We also note that many of the original EST sequences used to develop markers were located in untranslated gene regions (Putta *et al.* 2004). Our use of BLAST against the *A. mexicanum* deep-sequence transcriptome assembly (version 4.0; www.ambystoma.org) produced many significant hits to updated contig assemblies, but only rarely identified annotated coding regions. A similar application of BLAST to the NCBI nucleotide collection did not yield significant hits to known and annotated loci. As a result, for this study, we do not discriminate between patterns of sequence variation in coding vs. noncoding regions of our data.

DNA extraction and PCR amplification

We extracted genomic DNA from tissues using either a phenol–chloroform method or a DNeasy Blood & Tissue Kit (Qiagen, Inc.). We found that DNA quality was an important factor in producing successful PCR products across all targeted loci. Consequently, we ran 4.0 μL of each DNA extraction on a 1.0% agarose gel to confirm the presence of a high molecular weight band and minimal signs of degradation. When individual samples did not exhibit these properties, we excluded them from further use and extracted samples from new individuals from the same or a nearby locality. We determined the concentration and purity of high-quality DNA extractions using a NanoDrop 2000 spectrophotometer (Thermo Scientific). To improve consistency across PCRs, we diluted DNA for each individual to equal concentrations (50 ng/μL) prior to PCR. To minimize the potential for cross-contamination and facilitate pooling across loci in later steps, we performed PCRs in 96-well plates with one individual per plate and 96 PCRs (95 nuclear loci and a single negative control using dH_20 as the template and primers for the *NAD2* mtDNA gene). We performed PCR in 20-μL total volumes: 2.0 μL of 10× NEB Standard *Taq* Buffer, 0.7 μL of 10 μM each primer, 0.4 μL of 10 mM dNTPs, 0.1 μL of 5 U/μL NEB *Taq* Polymerase, 14.1 μL of dH_20 and 2.0 μL of 50 ng/μL whole genomic DNA. The thermocycler program included an initial 3-min denaturation step at 95 °C, 40 cycles of denaturation at 95 °C for 45 s, annealing at 55 °C for 45 s and elongation at 72 °C for 30 s, followed by a final extension step at 72 °C for 5 min. To confirm amplification, we visualized all PCR products on 1.3% agarose gels. For failed reactions, or those yielding relatively little product, we performed a second reaction under identical conditions.

Pooling, barcoding and 454 sequencing

To prepare PCR products for individual-specific barcoding reactions, we combined equal volumes (2.0 μL) of each of the 95 PCRs for each individual into two amplicon pools, one for smaller amplicons (177–299 bp: 49 loci) and one for larger amplicons (300–657 bp: 46 loci). This pooling strategy was based on recommendations from Roche representatives, who suggested that this might help to reduce biases that can occur during emulsion PCR (emPCR), which is more efficient for smaller fragments. To pool at this step, we used a liquid-handling robot (TECAN Genesis 200 Robotic Workstation), which we expected to provide consistency in pipetting across thousands of samples and minimize the potential for error. This pooling step resulted in two amplicon pools for each of the 93 individuals.

For each individual, we tagged both pools of amplicons for multiplexing on Roche's 454 sequencing platform using an individual-specific 8-nucleotide barcode following the techniques described in Meyer *et al.* (2007, 2008). All barcodes used in this study were at least two substitutions apart from each other, minimizing the potential for false assignment due to sequencing error. Barcodes were made up of self-hybridizing oligonucleotides that contain the barcode sequence on each end and an *Srf*I restriction cut site in the middle. This is a rare cutting site that occurs approximately every 150 kb in the human genome (Simcox *et al.* 1991).

We began the barcoding protocol with a blunt-end repair reaction on the pooled PCR products for each individual. We then ligated barcoded adaptors to both ends of each PCR product in a pool using a unique adaptor for each individual-specific pool. After ligation, we filled in single-stranded nicks using a strand-displacing polymerase. To assess the efficiency of barcode tagging, we used one PCR product from a single locus (approximately 350 bp) in this adapter ligation process and verified successful ligation by comparing the lengths of the untagged and tagged amplicons on a 2% agarose gel. Between each enzymatic step, we used Ampure SPRI beads (Agencourt) to clean sample pools following the protocols described in Meyer *et al.* (2008). We quantified each of the barcoded amplicon pools (93 × 2) using QuantIt Pico Green (Invitrogen). We then made two final pools of barcoded amplicons (representing small and large amplicon sizes) by pooling barcoded products across all 93 individuals in approximately equimolar concentrations. We dephosphorylated the two final amplicon pools and cut off half of each barcode adaptor using the *Srf*I restriction enzyme, which leaves 5′ phosphates for the ligation of universal 454 adapters during sequencing library preparation. We performed the dephosphorylation step to exclude amplicons without adaptors from 454 sequencing. After dephosphorylation reactions and restriction enzyme digests, we cleaned individual pools using the MinElute PCR purification kit (Qiagen) following the protocols described in Meyer *et al.* (2008).

We chose Roche's 454 sequencing platform over alternative platforms (e.g. Illumina) because 454 sequencing has the potential to generate longer reads (~500 bp at the time), which would allow us to more easily phase all the variable positions in an amplicon. To perform 454 sequencing, we constructed low molecular weight GS FLX Titanium General Libraries—typically used for shotgun sequencing—for each of the two amplicon pools following the manufacturer's instructions (Roche). We performed an initial round of small-volume emPCRs to assess the best DNA-copy-per-amplification-bead (cpb) conditions for clonal amplification and subsequent pyrosequencing. For the small-locus library, we performed a 0.5-cpb and 1.0-cpb emPCR, resulting in a 5.5% and 7.8% enrichment of successfully amplified beads, respectively. For the large-locus library, we also performed 0.5 cpb and 1.0 cpb emPCRs, resulting in a 3.3% and 5.8% enrichment of successfully amplified beads, respectively. We pooled the two emPCRs for the small-locus library. We also pooled the two emPCRs for the large-locus library, and each pool was sequenced separately on ⅛ of a 454 FLX PicoTiterPlate™ using titanium sequencing reagents. In general, optimal 454 sequencing results are obtained with bead enrichments of 5–20% (GS FLX Titanium Series emPCR Method Manual — Lib-L SV). Our initial enrichment percentages were on the low end of this range; therefore, we performed a second round of small-volume emPCR on the small-locus and large-locus libraries using ratios of 2 cpb (resulting in 6.1% enrichment) and 3 cpb (resulting in 7.2% enrichment), respectively. Each emPCR was sequenced separately on one-eighth of a 454 FLX PicoTiterPlate™ using titanium sequencing reagents. Finally, we performed a round of medium-volume emPCR on each sample pool using a ratio of 5 cpb for the small-locus library and 10 cpb for the large-locus library. Both emPCRs in this round resulted in successfully amplified bead enrichments of 25%. These two emPCRs were each sequenced separately on ⅛ of 454 FLX PicoTiterPlate™ using Titanium series sequencing reagents. In total, our 454 sequencing runs added up to three-fourth of a PicoTiterPlate™.

Bioinformatics of initial sequence data

We developed a bioinformatic pipeline that uses custom Perl and Python scripts, in conjunction with existing software, to sort PTS data, detect SNPs and phase haplotypes. Sequence reads (in FASTA format) were input into this pipeline and initially sorted by barcode using exact string matching. Reads without barcodes or with errors in the barcode regions were removed from further analyses. Sequence reads for each individual were then sorted into the 95 loci using BLAST and a database of reference sequences for each locus. We initially used sequences that were available from the *Ambystoma* EST database (www.ambystoma.org) as our reference sequences; however, because these reference sequences were based on transcriptome sequence data, matches were not ideal for several loci. Therefore, we subsequently generated our final reference sequences from 454 reads (excluding barcode and primer regions) that best matched each EST. Barcode and primer sequences were trimmed from the individual reads by removing all sequence data that extended beyond the reference sequences. This obviated the need to create separate files with primer sequences and efficiently removed primer sequences even if errors existed. For each locus, sequence reads were aligned for each individual separately (hereafter referred to as the primary alignment) using MAFFT under the default parameters (L-INS-i algorithm with default settings). By default, MAFFT orders alignments based on their similarity to the consensus sequence, which facilitates the evaluation of data quality, sequence variation and possible errors in sequences or in alignments.

An essential step in processing DNA sequence data is error detection and removal. Per base error rates for 454 pyrosequencing are somewhat higher than those of Sanger sequencing (Huse *et al.* 2007). With a large number of reads generated for each sample, the absolute number of errors across an entire data set of hundreds of thousands of reads becomes quite large. About 96% of 454 pyrosequencing errors are the result of insertion and deletion events, especially in homopolymer regions, and substitution errors are relatively rare (Quinlan *et al.* 2008). Both of these error types should generally have lower quality scores than true base calls; however, even true base calls, especially in and near homopolymer regions, are frequently assigned low quality scores (Quinlan *et al.* 2008). Alternatively, errors that occur during PCR (e.g. single-base substitutions and chimaeras) are not expected to influence quality scores that are calculated during pyrosequencing. Both pyrosequencing and PCR errors are expected to be less common in a data set than the true base calls (Quince *et al.* 2011), and as a result, multiple methods have been developed to statistically identify and remove errors, improve base calling and estimate true genotypes for high-coverage genomic data sets (Quinlan *et al.* 2008; Lynch 2009; Hohenlohe *et al.* 2010; Quince *et al.* 2011).

We explored two methods to identify and remove errors: (i) excluding low-quality nucleotides and (ii) using a likelihood ratio test (see *Haplotype phasing of NGS data* below) to statistically identify erroneous nucleotides. When excluding low-quality (<Q20) nucleotides, we found that inferred genotypes from different individuals were often difficult to align because of large amounts of missing data. Alternatively, including all statistically identified nucleotides resulted in the inference of haplotypes that were largely consistent in their variable sites, easily aligned across multiple individuals and generally free of rare nucleotides that were most likely the result of error. Therefore, we chose to include all nucleotide positions regardless of quality score and rely on statistical methods to resolve errors found within our reads.

Haplotype phasing of NGS data

An individual sequence read generated using NGS technologies is produced from a single strand of DNA and thus provides sequence data from a single haplotype. While a number of bioinformatic methods have been developed for genotyping a single variable nucleotide from an assembly of NGS reads (Emerson *et al.* 2010; Hohenlohe *et al.* 2010; Hird *et al.* 2011), no methods are currently available for the direct extraction of phased sequence information from assembled primary alignments with multiple variable nucleotides. To meet this need, we developed a bioinformatic pipeline that uses as input an individual and locus-specific primary alignment of multiple sequence reads and identifies the one (homozygous) or two (heterozygous) unique haplotypes present in an individual. We provide a general description of this process here.

For each site in an alignment, we counted each base or gap (A, G, C, T, -) and conducted a likelihood ratio test (LRT) as described in Hohenlohe *et al.* (2010). If the site was identified as homozygous, the same base was appended to both developing haplotype sequences. If the site was identified as heterozygous, we first determined whether a previous site within the read had been identified as heterozygous. If not, then one of the two bases was appended to each developing haplotype. However, if previous sites were heterozygous, we constructed lists of all sequences in the alignment containing each base for both the current SNP of interest (two lists) and the previous SNP (two additional lists). Bases for the current SNP of interest were then appended to the developing haplotypes according to the intersection of the lists of sequences for each base of each SNP. If there was only one possible way to pair the SNPs (i.e. if list one overlapped with list three, but not list four, and list two overlapped with list four, but not list three or vice versa), the bases were appended to each haplotype appropriately. However, if more than one possible pairing of SNPs could be made (e.g. if chimaeras were formed during PCR), then the frequencies of each SNP combination were used to determine the correct phase. Specifically, if the most common SNP combination occurred at least three times more frequently than the next most common SNP combination, then the more common one was considered correct. This frequency cut-off is based on the fact that any chimaera would have experienced at least one less PCR cycle than either parent sequence; therefore, both parent sequences should have frequencies at least equal to the chimaera. If the base at any position could not be determined, either because of low sequence coverage or because of an unresolvable mismatch between the intersections of lists, then the position was coded as 'N' in both haplotypes. This process, which was carried out for all positions in each primary alignment, served to both remove substitution and indel errors and infer phased haplotypes (Fig. S2, Supporting information). We recorded the number of cases where there was more than one way to pair the bases in the previous and current SNP, and the frequency cut-off was used to assign the appropriate pairing (Figs S3 and S4, Supporting information).

After the haplotypes were inferred for all individuals at a given locus, they were aligned in a single data matrix (hereafter referred to as the secondary alignment) using MAFFT under the default parameters. Secondary alignments were inspected manually. In addition to the correction of minor alignment errors at this stage, manual inspection was also useful for identifying aberrations in the haplotype reconstruction process for some individuals, usually resulting from suboptimal primary alignments. Common errors identified in primary alignments (e.g. misalignments) were manually corrected in secondary alignments. After manual inspection of secondary alignments and the correction of errors, final alignments including two inferred haplotypes for each individual for each locus were output as full DNA sequences in NEXUS file format.

The bioinformatic steps described previously, including processing of the data and phasing of haplotypes, are implemented in a scripted pipeline, 'NextAllele'. For an overview of the pipeline, see Fig. 2. For an overview of the haplotype inference steps, see Fig. S2 (Supporting information). All scripts for NextAllele are available for download: DRYAD entry doi: 10.5061/dryad.03s86. We have also recently developed a java version of NextAllele: http://wars.ca.uky.edu/NextAllele/.

Generating genotype data sets

Conversion of DNA sequence data to allelic data for population genetic analyses can be managed in several ways. One approach is to treat each full haplotype sequence as a single allele, where the multiple linked and variable sites in a locus are reduced to a single column in the genotype matrix (e.g. Weisrock *et al.* 2010). This method can be effective when all nucleotide sites are unambiguous and phased. However, even a small amount of missing data or ambiguous character states in an individual's sequences will make it impossible to assign them to a single genotype and will result in missing alleles. Characterizing an entire sequence as missing data when only a small number of sites are missing or ambiguous seems overly conservative.

A second approach is to sample a single SNP from each independent locus. If some positions are better represented in the data set than others, then these could be chosen. However, if many SNPs are well represented in the data matrix, the choice of SNPs could influence the results of some population genetic and phylogenetic analyses (Brumfield *et al.* 2003).

A third approach is to treat all SNPs as independent loci regardless of their physical linkage within each locus (e.g. Falush *et al.* 2003b; Conrad *et al.* 2006). The main disadvantage of nonindependence among SNPs within a nonrecombining locus is that population genetic analyses may overestimate the certainty of particular parameter estimates. For example, Bayesian assignment analyses in the program STRUCTURE may be misled when the majority of scored genotypes are in strong linkage disequilibrium, as would be the case if a single gene, or just a few genes, were sequenced. However, STRUCTURE is expected to perform well when there is sufficient independence across regions such that linkage disequilibrium within regions does not dominate the data (see pages 17–18 of the STRUCTURE manual). This genotyping approach has been used in population structure studies of bacteria (Falush *et al.* 2003b) and humans (Conrad *et al.* 2006). A benefit of treating each SNP as a separate genotype is that it allows for the retention of data from many sequences even when ambiguities or missing data exist. Furthermore, this approach does not require phasing of SNPs, which may be difficult with longer loci or with NGS platforms that generate shorter reads.

We chose to adopt the third genotyping option and scored each variable position across all loci as a genotyped SNP. While many loci contain multiple variable sites, most of which are expected to be in strong linkage disequilibrium, our use of such a large number of loci suggests that population genetic signal from multiple loci separated by large map distances and substantial recombination should outweigh the background linkage disequilibrium within loci (as suggested in the STRUCTURE manual). Within NextAllele, secondary alignments were used to assign SNP genotypes to individuals across all variable sites and produce a genotype matrix for use in subsequent population genetic analysis. Indels and 'N's were ignored for the purpose of SNP identification.

Population structure analyses

To assess population structure across the range of the *A. tigrinum* complex and begin to better understand the

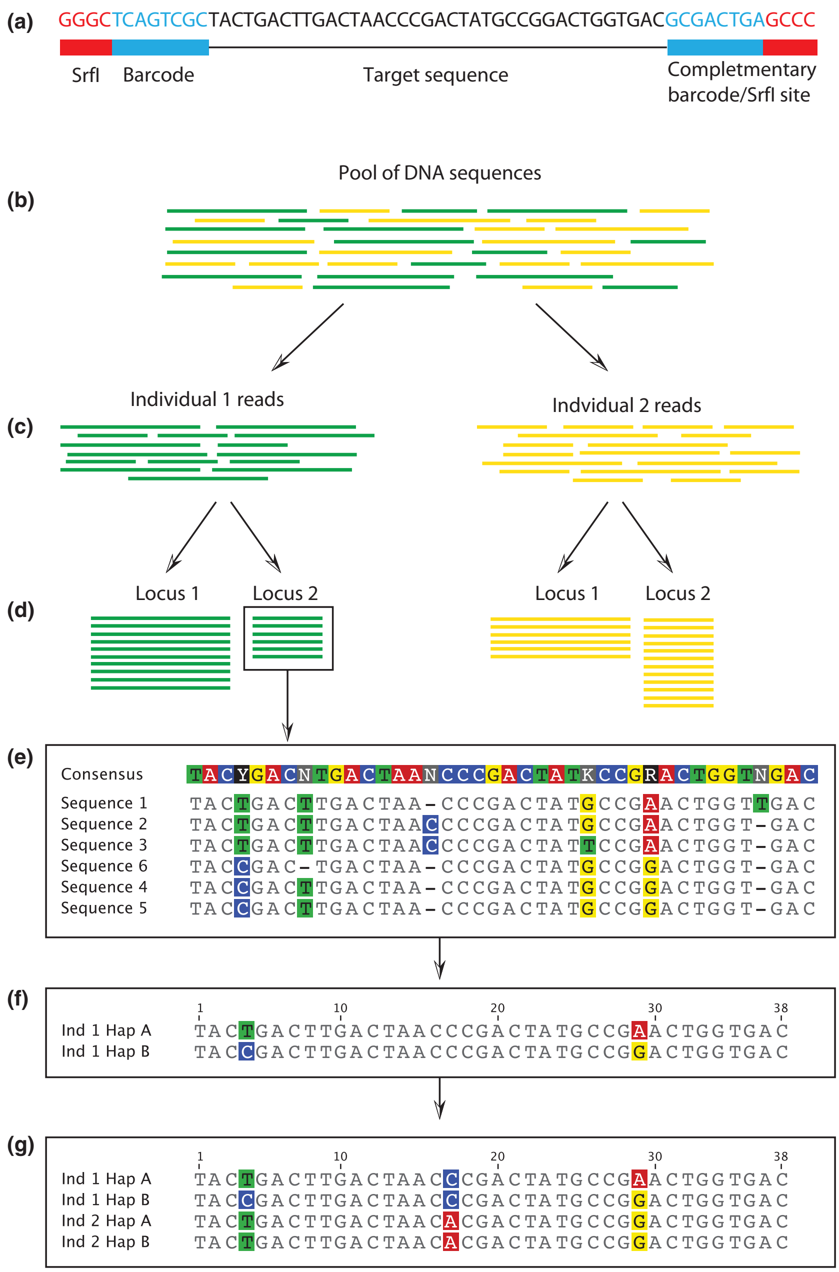

Fig. 2 A diagram of the bioinformatic pipeline developed in this study. Barcoded PCR products (a) are pooled and sequenced on a NGS platform (b). Pooled data are sorted by individual using barcode sequences (c) and by locus using a reference sequences (d). Primary alignments of multiple overlapping reads are made for each locus sequenced for each individual (e). Haplotype sequences are inferred for each primary alignment using a likelihood ratio test (f). Secondary alignments of inferred haplotypes for all individuals are made for each locus (g). These secondary alignments can then be used to generate SNP-based allelic matrices for population genetic analyses or can be used directly for sequence-based analyses.

geographic distribution of genetically distinct groups of populations, we analysed our genotypic data using the program STRUCTURE version 2.3 (Pritchard *et al.* 2000). We first analysed our full genotype matrix including all sampled individuals. We ran a series of STRUCTURE analyses under models assuming 1–15 genotypic clusters (*K*). In each iteration, individuals were assigned probabilistically to a genetic cluster based on their multi-locus genotype. No prior information about an individual's sampling locality was used in the analyses. Ten replicate analyses were performed for each *K*, each using 100 000 MCMC generations for the burn-in period and an additional 1 000 000 generations to estimate the posterior distribution. Our analyses used an admixture model that incorporated the possibility for some individuals to have mixed cluster ancestry and an 'F model' to account for correlated allele frequencies among populations that result from migration or shared ancestry (Falush *et al.* 2003a). A default setting was used for the F_k prior, which has a density proportional to a gamma distribution, and a uniform prior was used for the α parameter. Convergence was assessed by monitoring plots of the log probability of the data [$\ln\Pr(X|K)$] and α across the course of individual runs for stable patterns and by comparing mean values of $\ln\Pr(X|K)$ and α across replicate runs of the same *K* value. Generally, similar mean estimates of these parameters were interpreted as a sign of convergence on the posterior distribution.

We parsed the STRUCTURE results using the Structure-Harvester Python script version 0.6.8 (Earl & vonHoldt 2012) to calculate the mean lnPr(X|K) across replicates for each *K* and to calculate Δ*K* (Evanno *et al.* 2005), which is based on the rate of change in the lnPr(X|K) between successive *K* values. Δ*K* tends to favour smaller values of *K* that represent more highly differentiated sets of populations in systems that deviate from an island model (Evanno *et al.* 2005).

After a first round of STRUCTURE analyses on the full genotypic data set, we created three separate data partitions to explore the potential for our data to reveal finer levels of population structure. These partitions included the following: (i) all individuals sampled from U.S. populations, which were identified as three distinct clusters in the full data analysis; (ii) all individuals sampled from Mexican populations, which were identified as four distinct clusters in the full data analysis; and (iii) all sampled individuals of *A. ordinarium*, which were placed in a single cluster in the full data analysis and were previously diagnosed as a phylogenetic species using multi-locus nuclear sequence data (Weisrock *et al.* 2006). These three data sets were each analysed with STRUCTURE as described previously, using a range of *K* values from 1 to 10 for the first two data sets and a *K* range of 1–7 for the third data set.

Species tree reconstruction

We performed phylogeny reconstruction in a multispecies coalescent framework using the program *BEAST v.1.6.1 (Drummond & Rambaut 2007; Heled & Drummond 2010). We used genetic clusters identified in our STRUCTURE analyses to place individuals into nine OTUs that were used as the tips in the species tree analyses. The putative species included four clusters identified across U.S. populations and five clusters identified across the Mexican populations. We treated *A. ordinarium* as a single genetic cluster. Individuals with posterior probability (PP) assignment values >0.05 for two or more groups were considered to be admixed and were excluded from analysis.

To infer species trees, we used *BEAST analyses on 94 of our recovered nuclear loci (see Results for explanation). In addition, we ranked our loci in the order of decreasing phylogenetic information [as determined by parsimony-informative sites (Table S2, Supporting information)] and performed analyses on sets of the first 10, 20, 30, 40, 50, 60, 70, 80 and 90 loci. Phased haplotypes were critical for all species tree analyses, which benefit from longer reads with multiple variable positions and because we treated each phased haplotype sampled from each individual as a separate sample in the analyses. To determine the most appropriate substitution model for each locus, we used jModelTest 0.1.1 (Guindon & Gascuel 2003; Posada 2008) and the Akaike Information Criterion. In all analyses, we used a relaxed uncorrelated lognormal clock for gene tree estimation at each locus and a Yule prior for the species tree. We performed five replicate analyses for each set of genes and ran each MCMC analysis for one billion generations. In the analysis of the 94-locus data set, we sampled every 100 000 generations. In the analysis of the reduced-locus data sets, we sampled every 50 000 generations. We assessed convergence by comparing -lnL and ESS values of the sampled distributions across replicates using the program TRACER v1.5 (Rambaut & Drummond 2007) and by comparing the maximum clade credibility (MCC) tree generated from each individual replicate to identify concordance in tree topologies and PPs. After discarding the burn-in for each replicate (identified in Fig. S1, Supporting information), we combined their sampled distributions using LogCombiner and reconstructed a MCC tree.

Results

454 targeted resequencing

Our combined 454 runs (¾ of a PicoTiterPlate™) resulted in 344 330 sequence reads. The number of sequence reads for each of the ⅛ plate runs ranged from 29 225 to 81 407. The average sequence read length was 273 bp, and the average quality score was 33.6 (Table 1).

Bioinformatics of the sequence data

Using our bioinformatic pipeline, 317 624 sequence reads (92.2%) were sorted by barcode. 26 704 sequence reads (7.8%) did not match one of the 93 barcode sequences used, either because there was no barcode attached to the target sequence or because of sequence error in the barcode. Only two sequence reads matched multiple barcodes, and these were not included in further analyses. Sequence reads were generated for all 93 barcoded individuals with a mean number of reads per individual of 2992.60 ± 129.84 (SE) (Fig. 3a). The mean percentage of the total number of reads represented by an individual was 1.075%.

Of the 317 624 sequences that were sorted by barcode, 278 312 (87.6%) sorted by locus to one of our reference sequences. No reads sorted to multiple reference sequences. The 39 312 sequences that did not sort

Table 1 Emulsion PCR and 454 pyrosequencing results

emPCR round	Sample pool	cpb ratio (% enrichment)	454 pyrosequencing	No. of reads	Read length mean ± 1 SD (range)	Mean quality score
1	1	0.5 (5.5) 1.0 (7.8)	Combined for 1/8 run (#1)	65 388	250 ± 61.6 (40–753)	33.1
	2	0.5 (3.3) 1.0 (5.8)	Combined for 1/8 run (#2)	29 225	294 ± 103.9 (40–790)	34.6
2	1	2.0 (6.1)	1/8 run (#3)	81 407	246 ± 60.6 (40–744)	32.4
	2	3.0 (7.2)	1/8 run (#4)	60 428	275 ± 93.5 (40–528)	33.1
3	1	5.0 (25)	1/8 run (#3)	51 507	267 ± 64.6 (41–844)	34.3
	2	10.0 (25)	1/8 run (#4)	56 735	328 ± 104.1 (41–891)	34.8

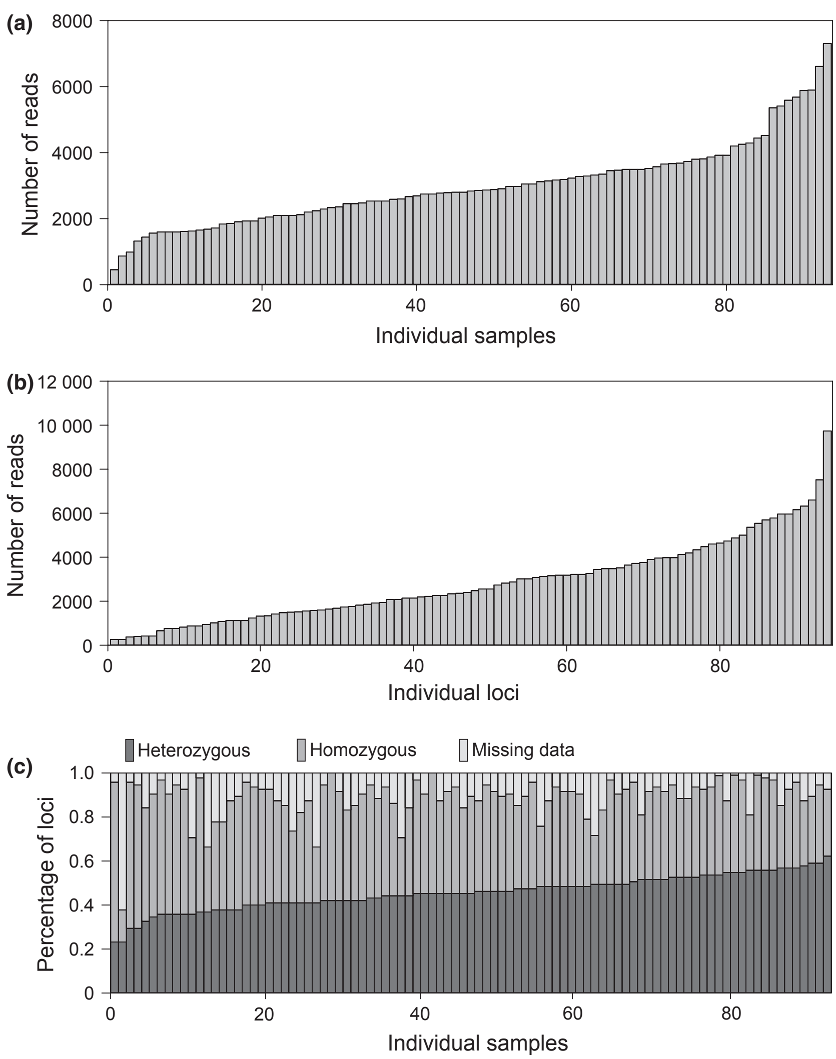

Fig. 3 Results of parallel tagged sequencing for the 93 individuals and 95 nuclear loci used in this study after automated bioinformatic processing. (a) A plot of the total number of sequence reads recovered for all individuals based on barcode matching. (b) A plot of the mean number of sequence reads across individuals for each locus. In both (a) and (b), individuals and loci were ordered according to the number of sequence reads, from smallest to largest. (c) A plot detailing the proportion of loci categorized as heterozygous, homozygous or containing missing data for all individuals. Individuals were ordered according to the proportion of heterozygous loci, from smallest to largest.

to a reference sequence were either too short or too poor of a match to be sorted to any particular reference sequence. These sequence reads were not included in further analyses. The mean number of reads per locus was 2006.06 ± 205.82 (SE) (Fig. 3b). All 93 individuals and all 95 loci were represented in the output after sorting by locus; however, not all individuals were represented at every locus. The mean coverage across all 93 individuals for all 95 loci was 31.5 ± 2.3 (SE). The number of reads for each locus generally decreased with the length of the locus ($y = -11.675x + 6094.6$; $R^2 = 0.36$), which is expected given the mechanics of emPCR.

We were able to reconstruct phased haplotypes for 94.5% (7850 of 8301) of the individual-specific primary alignments that were sorted by barcode and locus. Of these primary alignments, 45.8% included at least one heterozygous position. Of these heterozygotes, 48.5% had at least one case where linking the bases correctly in the haplotypes required using the 3× frequency rule (Figs S3 and S4, Supporting information). Specifically, the 3× frequency rule was required to resolve 4437 of a total of 13 806 heterozygous positions. After haplotypes were inferred, individuals were on average homozygous for 45.66% of their loci and heterozygous for 43.19% of their loci (Fig. 3c). Individuals had missing data for an average of 11.15% of their loci (Fig. 3c). Across all individuals, the average number of loci for which haplotypes could be inferred was 84.4 ± 0.94 (SE) (Table S1, Supporting information). From another perspective, on average a locus was missing data from 13% of the individuals (Table S2, Supporting information).

Of the 95 secondary alignments, 94 provided complete or nearly complete coverage for most of the inferred haplotypes. However, the largest locus (E21A3; reference sequence length 657 bp) was missing data for many individuals, contained multiple large regions of ambiguous nucleotides, many short reads and low overall coverage in the primary alignments. This locus was very difficult to align unambiguously; therefore, we excluded it from further analyses.

Properties of the data

Of the remaining 94 loci, the average alignment length was 271 bp (range: 123–477). The average number of variable sites per locus was 26.1 (range: 6–65). The average number of parsimony-informative sites was 18.5 (range: 5–56). The average of the mean pairwise differences between all haplotypes within a locus was 1.86 (range: 0.08–4.32). The average nucleotide diversity was 0.00723 (range: 0.00063–0.02403). The average number of inferred haplotypes per locus was 165.36 (range: 56–188). In general, the number of variable positions increased with the length of the alignment ($R^2 = 0.47$). We identified 2627 SNPs in the full data set of 94 loci.

Genetic structure analyses

Replicate STRUCTURE analyses of our full genotypic data set produced a ΔK that favoured a $K = 7$ (Table S3, Supporting information). Plots of assignment values revealed that most individuals had a very high PP (>0.95) for assignment to one of these seven clusters (Fig. 4a); however, a small number of individuals exhibited signs of admixture between two or more genetic clusters. Genetic clusters were geographically and taxonomically specific, a pattern also found at lower levels of genetic structure. For example, at a $K = 2$ level of structure, most individuals sampled from Mexican population were placed in a cluster separate individuals sampled from U.S. population with PPs of ~1.0 (results not shown).

STRUCTURE analysis of more exclusive sets of individuals indicated that our data were effective in resolving further levels of genetic structure. Separate analyses of individuals sampled from U.S. population produced a ΔK that favoured a $K = 4$ (Table S3, Supporting information) and revealed that our 'western' *Ambystoma tigrinum* samples were divided into two additional genetic clusters (Fig. 4b) that were not resolved in analyses of our larger data set. Similarly, separate analyses of individuals sampled from Mexican population produced a ΔK that favoured a $K = 6$ (Table S3, Supporting information). This level of K may be an overestimate of genetic structure for this group as one of the clusters was only represented at a low assignment frequency in some individuals (results not shown, see Dryad accession for full STRUCTURE results). However, an evaluation of a $K = 5$ for Mexican samples revealed strong cluster assignments for most individuals and revealed that populations of *A. andersoni*, *A. dumerilii*, *A. mexicanum* and two unidentified individuals formed a separate cluster that was highly distinct from a cluster of *A. mexicanum* and *A. velasci* individuals (Fig. 4c). Finally, analysis of *A. ordinarium* individuals produced a ΔK that favoured a $K = 2$ (Table S3, Supporting information), and all individuals were grouped geographically into western and eastern clusters (Fig. 4d).

Phylogeny reconstruction

*BEAST analysis of smaller subsets of our loci that contained the most phylogenetic information (as measured by parsimony-informative sites) produced results that were most consistent with convergence on the posterior sampling distribution, as measured by similar lnL, and parameter estimates across replicate analyses (Figs 5 and S1, Supporting information) and similar MCC trees constructed from the sampling distribution of each replicate (Fig. S1, Supporting information). Replicate analyses using the 10 and 20 most informative loci converged on similar posterior distributions with the 20-locus analyses producing the overall strongest set of relationships, with multiple well-supported clades with PPs = 1.0 (Figs 5a and S1, Supporting information): (i) a clade of all tiger salamanders, excluding *A. californiense*; (ii) a clade containing the eastern *A. tigrinum* lineage and the two western *A. tigrinum* lineages; (iii) a clade of the two western *A. tigrinum* lineages; (iv) a clade of Mexican lineages; and (v) a clade of Mexican lineages

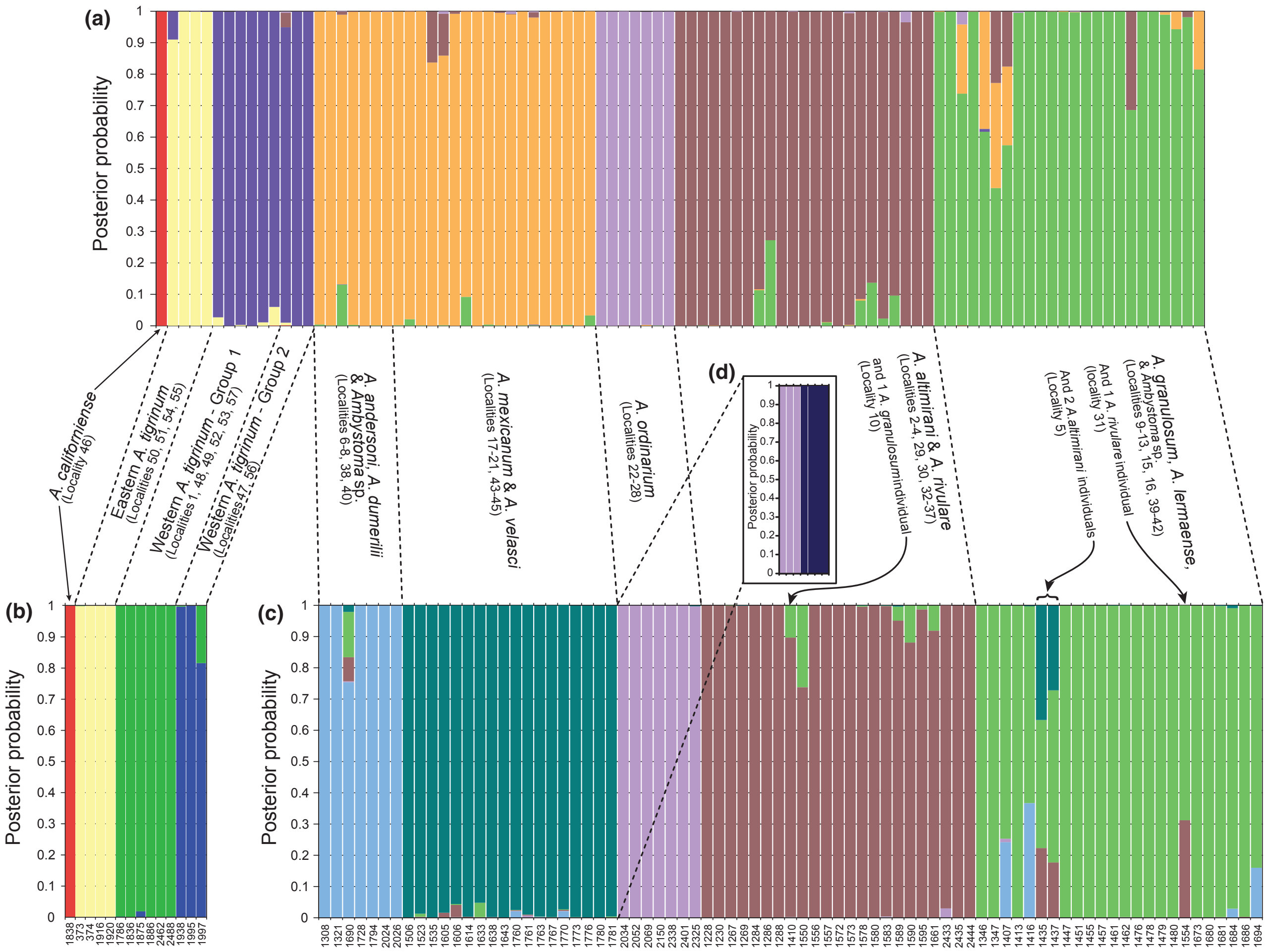

Fig. 4 Results from STRUCTURE analyses of the nuclear genotypic data. In all plots, vertical bars represent an individual's assignment to a genotypic cluster with colours designating the different clusters. (a) $K = 7$ STRUCTURE plot resulting from analysis of the full 93 individual genotypic data set. (b) $K = 4$ STRUCTURE plot resulting from analysis of genotypic data from all individuals sampled from U.S. populations. (c) $K = 5$ STRUCTURE plot resulting from analysis of genotypic data from all individuals sampled from Mexican population. (d) $K = 2$ STRUCTURE plot resulting from analysis of genotypic data from all sampled individuals of *Ambystoma ordinarium*. In all plots, the K value presented is the favoured level of structure using ΔK (Evanno *et al.* 2005). Numbers listed below the plots in (b) and (c) refer to individual identification numbers listed in Table S1.

excluding *A. ordinarium*. Analyses of the 30- and 40-locus data sets also produced lnL patterns suggestive of convergence (Figs 5b and S1, Supporting information); however, topologies from replicate analyses indicated strong support for discordant relationships between some Mexican lineages. In general, replicate analyses that included 50 or more loci failed to exhibit patters that would indicate convergence on a similar sampling distribution and showed decreased resolution and support among lineages (Figs 5c, d and S1, Supporting information). We note two important aspects of the analyses of these larger data sets. First, the increased numbers of loci (with decreased levels of variation and informativeness) greatly decreased the rate at which analyses proceeded. While most of our analyses finished, or nearly finished, the full one billion generation analysis by the time of final production of this study, the larger analyses (i.e. >50 loci) took over 3 months to complete. Second, without convergence among replicates for the larger analyses, no appropriate burn-in could be identified for the generation of a combined MCC tree. These are presented in Figs 5c, d and S1 (Supporting information) using an arbitrarily defined burn-in.

Discussion

In this study, we used a PTS approach to collect DNA sequence data from 93 tiger salamander individuals by using PCR to target 95 loci broadly spaced throughout the genome. We also developed a bioinformatic pipeline, NextAllele, that sorts these reads by locus and individual and

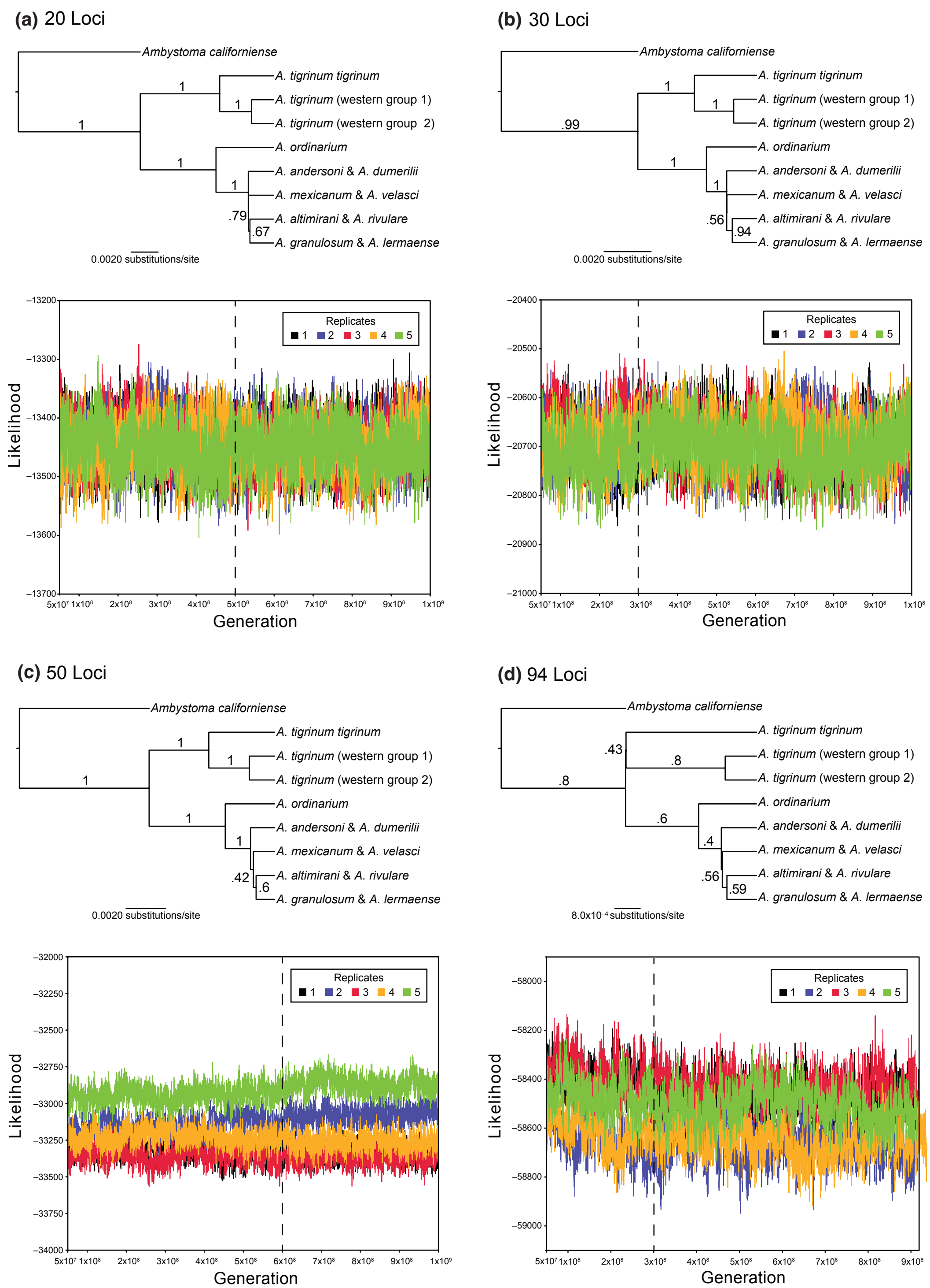

Fig. 5 Results from *BEAST species tree reconstruction analyses. (a) The maximum clade credibility (MCC) tree created from combined posterior distributions of all five replicate analyses using the 20 most informative loci (as measured by parsimony-informative sites), and the plot of log-likelihood (lnL) values from the five replicate analyses. (b) The MCC tree created from combined posterior distributions of all five replicate analyses using the 30 most informative loci, and the plot of lnL values from the five replicate analyses. (c) The MCC tree created from combined posterior distributions of all five replicate analyses using the 50 most informative loci, and the plot of lnL values from the five replicate analyses. (d) The MCC tree created from combined posterior distributions of all five replicate analyses using all 94 loci, and the plot of lnL values from the five replicate analyses. Numbers above branches in all MCC trees represent posterior probabilities. In all lnL plots, the dashed line indicates the burn-in used to generate MCC trees from the combined posterior distributions.

infers phased haplotypes from primary alignments with multiple variable positions. Our method was successful in recovering phased sequence data, allowing us to use a substantial number of variable sites that were informative at both population genetic and phylogenetic levels. As a result, we now have an initial insight into the range-wide population structure of the *A. tigrinum* complex, the potential species lineages that it contains and a first-pass estimate of species-level phylogenetic relationships.

Parallel targeted sequencing

The costs of NGS are rapidly decreasing, further enabling the use of large-scale genomic data sets for research in molecular ecology (Ekblom & Galindo 2011; Glenn 2011; McCormack *et al.* 2012a). However, for most empirical biologists, questions remain regarding the optimal methods for generating population-level genomic sequence data that provide both large amounts of sequence information identifiable to an individual and maximizes the recovery of orthologous loci across individual samples. The development of barcoding/indexing strategies for most of the major sequencing platforms has efficiently dealt with the first of these needs, providing the potential for pooled and parallel sequencing of many individuals within a single sequencing run. However, meeting this second goal in the majority of organisms is still a challenge. For organisms that lack substantial genomic resources, random sequencing of exclusive portions of the genome using RNA-seq (e.g. Jeukens *et al.* 2010) or reduced representation (e.g. McCormack *et al.* 2012b) methods may be the best sequencing strategy to use. However, these sequencing strategies often yield large numbers of sequence reads (sometimes the majority) that cannot be assembled into contigs (e.g. Gompert *et al.* 2010). Such results may be especially problematic in organisms with large genomes, including salamanders, whose genome sizes range from 15 to 120 GB (http://www.genomesize.com/). In addition, despite the overall recovery of thousands of unique loci in these types of studies, the number of loci in which the majority of individuals or taxa are represented can often be quite small (e.g. McCormack *et al.* 2012b). Increased sequencing may be one way to alleviate these problems, but for most researchers with limited budgets, this may not be a practical solution.

Our approach provides a different perspective, one in which some prior genomic resources are already available to target potentially informative markers. Using PCR in conjunction with PTS, we efficiently amplified and sequenced specific loci from across the genome for all of our samples. This sequencing strategy was highly efficient at recovering our target loci and individuals, with relatively few unsorted reads generally resulting from either missing barcodes or sequences that were too short to match to reference sequences. Only a small percentage of our reads were missing barcodes, suggesting that the tagging method we employed (Meyer *et al.* 2008) was very efficient. Short sequence reads, which mainly affected sorting by locus, were probably the result of incomplete emPCR or pyrosequencing reactions, both of which will probably improve with continued advances in NGS technology. More importantly, our targeted PTS resulted in data sets with little missing data (on average, individuals were missing data from just 11% of their loci), which could be improved with increased sequencing depth or normalization prior to pooling. This is particularly important because it demonstrates a method that minimizes missing data for studies using large numbers of individuals or species. In contrast, reduced representation techniques, which employ restriction enzyme digests, are expected to increase in missing data as more divergent lineages are included. The primary benefit of using PCR to target and enrich specific loci for NGS was that it allowed us to sequence the same orthologous regions from each individual. As a result, our NGS sequencing efforts were focused on a limited set of loci. Furthermore, verification of successful PCR by gel electrophoresis provides a guarantee that the targeted loci are included in NGS library preparation and sequencing. This is highly desirable when very specific loci are under scrutiny across individuals (Babik *et al.* 2009; Galan *et al.* 2010).

However, using PCR for enrichment is not without its drawbacks. We observed considerable variance in coverage across individuals and loci, which probably resulted from our failure to normalize the PCR products prior to the initial pooling step. This variance may explain the absence of recovered haplotypes from some individuals at particular loci, with fragments that amplified better resulting in higher sequence coverage. We made some effort to minimize this effect (i.e. choosing loci that optimally amplified at 55 °C across our test panel), but this strategy was crude at best. Several methods are currently available to normalize PCR products. These could be used to reduce this variance; however, quantification and normalization of thousands of PCR products may be less efficient and more expensive than simply generating greater overall sequence depth.

Alternatives to standard PCR amplification for targeting specific loci are now commercially available (e.g. Rainstorm) (Ekblom & Galindo 2011). In addition, sequence capture methods using single-stranded nucleotide probes provide an alternative for enriching loci and can greatly reduce the laboratory workload prior to NGS (Mamanova *et al.* 2010). This latter approach has worked well in studies where probes are targeted to known gene regions within a lineage (Nadeau *et al.* 2012), and recent studies suggest that it can be expanded to target large numbers of conserved elements across a divergent range of lineages, including those for which limited or no prior genomic sequence information is available (Crawford *et al.* 2012; Faircloth *et al.* 2012; Lemmon *et al.* 2012). These techniques show great promise as targeting or enrichment methods, but may still present some hurdles, including the initial price of commercial library and capture kits and the technical proficiency or equipment required for noncommercial approaches. Even as these hurdles are reduced, familiarity with PCR may delay some researchers from taking full advantage of these approaches. As a result, PCR will likely to remain a common technique for generating sequencing templates in the near future, and our study demonstrates the feasibility of this approach even for relatively large numbers of loci. Furthermore, our bioinformatic pipeline will be applicable to a variety of approaches including enrichment or reduced representation.

Cost in terms of time and money is also a major consideration for most molecular projects. Sanger sequencing has long been the standard for several decades; however, it is not the least expensive or fastest method for very large-scale projects. We estimated that the total cost of this PTS project to be approximately $14 000 (Table S4, Supporting information), which includes the cost of all of our exploratory sequence runs. We estimated that the comparable cost using Sanger sequencing and current rates through our sequencing facility is estimated to be at least $21 000. This latter estimate is based on two sequencing reactions per amplicon and does not factor in the cost of cloning, which would be necessary to phase haplotypes using a nonalgorithmic method. Furthermore, cloning and sequencing these 8835 amplicons using Sanger sequencing would require considerably more time than the methods we present here.

Bioinformatics of PTS data

The ease with which large data sets can now be generated is often contrasted with the difficulty of processing the data in an automated fashion. The number of commercial and noncommercial programs available for quality control, assembly and analysis of NGS data is growing rapidly (reviewed in McCormack *et al.* 2012a; Nielsen *et al.* 2011). Typically, the programs designed for genotyping for population genetic and phylogeography have scored one SNP in each short sequence read (e.g. Catchen *et al.* 2011; Hird *et al.* 2011) The goal of our data collection was not only the identification of individual SNPs, but also the reconstruction of phased haplotypes to be used in gene tree and species tree reconstruction. Our bioinformatic pipeline was therefore essential for both managing our large set of raw sequence data and identifying haplotypes from loci with multiple polymorphic sites. A benefit of using NGS for generating population-level data is that sequence reads are generated from single molecules and therefore provide readily extractable phase information. Our pipeline serves as an efficient and automated tool for producing phased data sets that not only avoids the time-consuming process of manually identifying haplotypes, but also accounts for many of the potential errors in NGS sequence data.

The use of this method does have some limitations: the recovery of phased alleles for an individual was hampered when coverage across the target locus was not complete (i.e. individual reads did not include all of the SNPs in long loci). In addition, some inferred haplotypes included ambiguities and incomplete SNP phasing, because of homopolymer regions in the sequence or primary assemblies with patterns consistent with PCR-induced recombination. To improve coverage and phase all target SNPs, shorter amplicons can be targeted; however, read lengths are increasing for most next-generation platforms, and this issue will probably improve in the near future. Putative recombinant amplicons were often resolved in cases where one combination was at least three times as common as any other; however, this was only possible in high-coverage primary alignments. Most likely, problems resulting from PCR-induced recombination can be resolved by using fewer cycles during amplification (Cronn *et al.* 2002) or by sequencing replicate PCRs for each locus.

While automated processing of our data was essential, we also found that quality control measures were still important. It is not feasible to examine every raw sequence read; however, the examination of major steps in the bioinformatic process can improve the quality of our downstream data sets. We have included multiple output files in our pipeline to help users access both their data and the performance of the pipeline at relevant steps. We found this output to be critical, especially for the identification of SNPs in secondary alignments. MAFFT is a very effective alignment method, but errors were common enough to warrant some manual editing. Many of the errors in secondary alignments were the result of errors in primary alignments, especially in and around homopolymer regions. The primary alignment errors can often be resolved by first identifying unusual patterns in secondary alignments and then backtracking to the individual-specific primary alignment to make a manual adjustment. Furthermore, our pipeline is equipped to accommodate these subsequent changes without requiring a full re-run of the pipeline. We believe that the approach we took was conservative in that we allowed for a considerable number of ambiguities because of low coverage for some amplicons. However, once many haplotypes have been inferred, it should be possible to use 'known' SNPs to help infer genotypes for new individuals.

Utility of large data sets

Perhaps the greatest significance of this project is that our PTS data allowed us to begin to approach population genetic and phylogenetic analyses with multi-locus genome-wide sequence data. A major objective of our overall project is to characterize species boundaries of the *A. tigrinum* complex and place these lineages within a phylogenetic framework, objectives that have been challenging given the very recent origin and radiation of this group (Shaffer & McKnight 1996), the lack of reproductive isolation between populations (as evidenced by the interbreeding of introduced populations; Riley *et al.* 2003) and signatures of mitochondrial introgression across putative species boundaries (Weisrock *et al.* 2006). The collection of DNA sequence data from a large set of independently evolving nuclear loci was seen as necessary for providing a comprehensive survey of geographic genetic variation across the entire range of the *A. tigrinum* species complex and placing it in a phylogenetic context.

Our observation that individuals were assigned, with high probabilities, to genetic clusters that correlated with geography and taxonomy is consistent with the results of the small number of published studies of population structure within the complex using microsatellites (Parra-Olea *et al.* 2011) and mtDNA (Routman 1993; Shaffer & McKnight 1996), suggesting that our nuclear sequence data are at least as informative in detecting fine-scale patterns of genetic differentiation. We did not thoroughly explore the limits of fine-scale population structure in this study, mainly due to limited geographic sampling across the complex range. It is notable, however, that further exploration of population structure within one particular cluster, *A. ordinarium*, revealed additional discrete geographic clusters (Fig. 4d) not revealed in previous phylogenetic analyses of a smaller subset of nuclear loci (Weisrock *et al.* 2006). We expect that these data, once expanded to include more thorough geographic sampling, will be extremely informative in identifying lineage boundaries across the *A. tigrinum* species complex. Genetic clusters identified using STRUCTURE have previously served as hypotheses in species delimitation studies (Shaffer & Thomson 2007; Leaché & Fujita 2010; Weisrock *et al.* 2010) that can be subsequently tested using multi-locus coalescent (Yang & Rannala 2010; Ence & Carstens 2011) or nonparametric (Cummings *et al.* 2008; O'Meara 2010) tests of lineage divergence, or coalescent-based estimates of gene flow (Hey & Nielsen 2007). A major benefit of the data set we generated is that it is amenable to analysis at both population genetic and phylogenetic levels.

Our PTS data also indicate utility at a deeper phylogenetic level; however, not all loci were equally informative. *BEAST analyses of relatively variable loci produced replicate analyses that converged on similar posterior parameter and topology estimates (Figs 5a, b and S1, Supporting information) and produced strong posterior support for many branches within the consensus species tree. In contrast, analyses with larger numbers of loci that individually—and increasingly—contained less variation resulted in overall lower support for individual branches that were strongly supported in analyses with fewer loci (Figs 5c, d and S1, Supporting information). More importantly, replicate analyses of larger data sets produced parameter and topology estimates that did not converge on the same posterior distribution (e.g. Fig. 5d). The poor performance of these analyses with larger data sets is probably the result of increasing the number of parameters to be estimated in an analysis (e.g. more gene trees), but doing so with the addition of loci that have decreasing levels of information; for example, a number of loci have fewer than 10 parsimony-informative sites, yet their gene tree may have as many as 180 tips to be resolved. The posterior distribution of a gene tree with these properties is likely to be very broad with few branches estimated with certainty. As a result, these uncertain gene tree estimates are not likely to be informative in the resolution of the species tree, despite creating additional computational burden to the overall analysis. Future species tree methods that can accommodate loci with low levels of variation (i.e. at the level of a single SNP) would be beneficial in more fully utilizing the full scope of data sets that we have collected here. However, overall, we view these reconstruction patterns as encouraging for our data collection strategy because it provides a mechanism for collecting large amounts of sequence data that can be parsed to identify loci that are the most informative for phylogenetic reconstruction.

Conclusions

It is now possible to generate large data sets of multi-locus DNA sequences for addressing evolutionary questions in population genetics, phylogeography and phylogeny. The major challenges have been (i) how to generate data that are useful at multiple scales and (ii) how to analyse the data efficiently and effectively. The approach that we have demonstrated here efficiently provides large quantities of one of the most generally useful types of data, phased DNA sequence data from orthologous loci from multiple individuals. We have also demonstrated the utility of these data for understanding population genetic and phylogenetic relationships that could not be resolved effectively with smaller data sets. Our approach holds great promise for a variety of studies addressing the population and evolutionary histories of a broad range of taxa.

Acknowledgements

We thank Stephanie Mitchell, Alex Noble and Ben Tuttle for assistance in the laboratory; Jim Bogart, Ken Carbale, Sheri Church, Gerardo Herrera, Paul Moler, Robert Seib, Andrew Storfer and Luis Zambrano for providing valuable tiger salamander samples; Randal Voss and John Walker for assistance with robotics; and Yukie Kajita for providing the map. Abbe Kesterson and Jenifer Webb at the UK AGTC provided useful advice in preparation for our 454 sequencing runs. Jerzy Jaromczyk and Chris Schardl provided valuable student mentorship and guidance. Discussions with the NIMBioS Working Group on Species Delimitation were especially helpful with planning analyses. Paul Hime, Scott Hotaling, John McCormack, Brant Faircloth and an anonymous reviewer provided helpful comments that improved the quality of the manuscript. Finally, we thank the University of Kentucky Information Technology department and Center for Computational Sciences for computing time on the Lipscomb High Performance Computing Cluster and for access to other supercomputing resources. This work was supported by the Commonwealth of Kentucky and by the National Science Foundation through awards DEB-0949532 (to DWW and EMO) and KY EPSCoR grant number 0814194.

References

Babik W, Taberlet P, Ejsmond MJ, Radwan J (2009) New generation sequencers as a tool for genotyping of highly polymorphic multilocus MHC system. *Molecular Ecology Resources*, **9**, 713–719.

Baird NA, Etter PD, Atwood TS, *et al.* (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One*, **3**, e3376.

Binladen J, Gilbert MTP, Bollback JP *et al.* (2007) The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. *PLoS One*, **2**, e197.

Briggs AW, Good JM, Green RE *et al.* (2009) Targeted retrieval and analysis of five Neandertal mtDNA genomes. *Science*, **325**, 318–321.

Brito P, Edwards S (2009) Multilocus phylogeography and phylogenetics using sequence-based markers. *Genetica*, **135**, 439–455.

Brumfield RT, Beerli P, Nickerson DA, Edwards SV (2003) The utility of single nucleotide polymorphisms in inferences of population history. *Trends in Ecology & Evolution*, **18**, 249–256.

Bybee SM, Bracken-Grissom H, Haynes BD *et al.* (2011a) Targeted amplicon sequencing (TAS): a scalable next-gen approach to multilocus, multitaxa phylogenetics. *Genome Biology and Evolution*, **3**, 1312–1323.

Bybee SM, Bracken-Grissom HD, Hermansen RA *et al.* (2011b) Directed next generation sequencing for phylogenetics: an example using Decapoda (Crustacea). *Zoologischer Anzeiger – A Journal of Comparative Zoology*, **250**, 497–506.

Carling MD, Brumfield RT (2007) Gene sampling strategies for multi-locus population estimates of genetic diversity (theta). *PLoS One*, **2**, e160.

Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlewait JH (2011) *Stacks*: building and genotyping loci *de novo* from short-read sequences. *G3: Genes, Genomes, Genetics*, **1**, 171–182.

Collins JP, Mitton JB, Pierce BA (1980) *Ambystoma tigrinum*: a multispecies conglomerate?. *Copeia*, **1980**, 938–941.

Conrad DF, Jakobsson M, Coop G *et al.* (2006) A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nature Genetics*, **38**, 1251–1260.

Crawford NG, Faircloth BC, McCormack JE, Brumfield RT, Winker K, Glenn TC (2012) More than 1,000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs. *Biology Letters*, doi: 10.1098/rsbl.2012.0331

Cronn R, Cedroni M, Haselkorn T, Grover C, Wendel JF (2002) PCR-mediated recombination in amplification products derived from polyploid cotton. *TAG. Theoretical and Applied Genetics. Theoretische und angewandte Genetik*, **104**, 482–489.

Cummings MP, Neel MC, Shaw KL (2008) A genealogical approach to quantifying lineage divergence. *Evolution*, **62**, 2411–2422.

Drummond AJ, Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, **7**, 214.

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.

Edwards S (2009) Is a new and general theory of molecular systematics emerging? *Evolution*, **63**, 1–19.

Edwards S, Liu L, Pearl D (2007) High-resolution species trees without concatenation. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 5936–5941.

Ekblom R, Galindo J (2011) Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity*, **107**, 1–15.

Emerson KJ, Merz CR, Catchen JM *et al.* (2010) Resolving postglacial phylogeography using high-throughput sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 16196–16200.

Ence DD, Carstens BC (2011) SpedeSTEM: a rapid and accurate method for species delimitation. *Molecular Ecology Resources*, **11**, 473–480.

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.

Faircloth BC, McCormack JE, Crawford NG, Harvey MG, Brumfield RT, Glenn TC (2012) Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Systematic Biology*, **61**, 717–726.

Falush D, Stephens M, Pritchard JK (2003a) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, **164**, 1567–1587.

Falush D, Wirth T, Linz B *et al.* (2003b) Traces of human migrations in *Helicobacter pylori* populations. *Science*, **299**, 1582–1585.

Fitzpatrick BM, Johnson JR, Kump DK, Shaffer HB, Smith JJ, Voss SR (2009) Rapid fixation of non-native alleles revealed by genome-wide SNP analysis of hybrid tiger salamanders. *BMC Evolutionary Biology*, **9**, 176.

Fitzpatrick BM, Johnson JR, Kump DK, Smith JJ, Voss SR, Shaffer HB (2010) Rapid spread of invasive genes into a threatened native species. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 3606–3610.

Forister ML, Gompert Z, Fordyce JA, Nice CC (2011) After 60 years, an answer to the question: what is the Karner blue butterfly? *Biology Letters*, **7**, 399–402.

Frost DR (2008) Amphibian Species of the World: an Online Reference. Version 5.2 (15 July, 2008). Electronic Database accessible at http://research.amnh.org/herpetology/amphibia/index.php. American Museum of Natural History, New York, USA.

Galan M, Guivier E, Caraux G, Charbonnel N, Cosson JF (2010) A 454 multiplex sequencing method for rapid and reliable genotyping of highly polymorphic genes in large-scale studies. *BMC Genomics*, **11**, 296.

Glenn TC (2011) Field guide to next-generation DNA sequencers. *Molecular Ecology Resources*, **11**, 759–769.

Gnirke A, Melnikov A, Maguire J *et al.* (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nature Biotechnology*, **27**, 182–189.

Gompert Z, Forister ML, Fordyce JA, Nice C, Wiiliamson RJ, Buerkle CA (2010) Bayesian analysis of molecular variance in pyrosequences quantifies population genetic structure across the genome of Lycaeides butterflies. *Molecular Ecology*, **19**, 2455–2473.

Griffin PC, Robin C, Hoffmann AA (2011) A next-generation sequencing method for overcoming the multiple gene copy problem in polyploid phylogenetics, applied to Poa grasses. *BMC Biology*, **9**, 19.

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

Heled J, Drummond AJ (2010) Bayesian inference of species trees from multilocus data. *Molecular Biology and Evolution*, **27**, 570–580.

Hey J, Nielsen R (2007) Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 2785–2790.

Hird S, Brumfield RT, Carstens BC (2011) PRGmatic: an efficient pipeline for collating genome-enriched second-generation sequencing data using a provisional reference genome. *Molecular Ecology Resources*, **11**, 743–748.

Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA (2010) Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genetics*, **6**, e1000862.

Huang H, He Q, Kubatko LS, Knowles LL (2010) Sources of error inherent in species-tree estimation: impact of mutational and coalescent effects on accuracy and implications for choosing among different methods. *Systematic Biology*, **59**, 573–583.

Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM (2007) Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biology*, **8**, R143.

Jeukens J, Renaut S, St-Cyr J, Nolte AW, Bernatchez L (2010) The transcriptomics of sympatric dwarf and normal lake whitefish (*Coregonus clupeaformis* spp., Salmonidae) divergence as revealed by next-generation sequencing. *Molecular Ecology*, **19**, 5389–5403.

Kircher M, Kelso J (2010) High-throughput DNA sequencing – concepts and limitations. *BioEssays*, **32**, 524–536.

Leaché AD, Fujita MK (2010) Bayesian species delimitation in West African forest geckos (*Hemidactylus fasciatus*). *Proceedings of the Royal Society B*, **277**, 3071–3077.

Lemmon AR, Brown JM, Stanger-Hall K, Lemmon EM (2009) The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and Bayesian inference. *Systematic Biology*, **58**, 130–145.

Lemmon AR, Emme SA, Lemmon EM (2012) Anchored hybrid enrichment for massively high-throughput phylogenomics. *Systematic Biology*, **61**, 727–744.

Lerner HR, Meyer M, James HF, Hofreiter M, Fleischer RC (2011) Multilocus resolution of phylogeny and timescale in the extant adaptive radiation of Hawaiian honeycreepers. *Current Biology*, **21**, 1838–1844.

Liu L, Pearl D (2007) Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. *Systematic Biology*, **56**, 504–514.

Lynch M (2009) Estimation of allele frequencies from high-coverage genome-sequencing projects. *Genetics*, **182**, 295–301.

Maddison WP (1997) Gene trees in species trees. *Systematic Biology*, **46**, 523–536.

Mamanova L, Coffey AJ, Scott CE *et al.* (2010) Target-enrichment strategies for next-generation sequencing. *Nature Methods*, **7**, 111–118.

Maricic T, Whitten M, Pääbo S (2010) Multiplexed DNA capture of mitochondrial genomes using PCR products. *PLoS One*, **5**, e14004.

McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT (2012a) Applications of next-generation sequencing to phylogeography and phylogenetics. *Molecular Phylogenetics and Evolution*, doi: 10.1016/j.ympev.2011.12.007.

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

Meyer M, Stenzel U, Myles S, Prufer K, Hofreiter M (2007) Targeted high-throughput sequencing of tagged nucleic acid samples. *Nucleic Acids Research*, **35**, e97.

Meyer M, Stenzel U, Hofreiter M (2008) Parallel tagged sequencing on the 454 platform. *Nature Protocols*, **3**, 267–278.

Morin PA, Archer FI, Foote AD *et al.* (2010) Complete mitochondrial genome phylogeographic analysis of killer whales (*Orcinus orca*) indicates multiple species. *Genome Research*, **20**, 908–916.

Nadeau NJ, Whibley A, Jones RT *et al.* (2012) Genomic islands of divergence in hybridizing Heliconius butterflies identified by large-scale targeted sequencing. *Philosophical Transactions of the Royal Society of London. Series B, Biological sciences*, **367**, 343–353.

Nielsen R, Paul JS, Albrechtsen A, Song YS (2011) Genotype and SNP calling from next-generation sequencing data. *Nature Reviews. Genetics*, **12**, 443–451.

O'Meara BC (2010) New heuristic methods for joint species delimitation and species tree inference. *Systematic Biology*, **59**, 59–73.

van Orsouw NJ, Hogers RC, Janssen A *et al.* (2007) Complexity reduction of polymorphic sequences (CRoPS): a novel approach for large-scale polymorphism discovery in complex genomes. *PLoS One*, **2**, e1172.

Parks M, Cronn R, Liston A (2009) Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. *BMC Biology*, **7**, 84.

Parra-Olea G, Zamudio KR, Recuero E, Aguilar-Miguel X, Huacuz D, Zambrano L (2011) Conservation genetics of threatened Mexican axolotls (*Ambystoma*). *Animal Conservation*, **15**, 61–72.

Posada D (2008) jModelTest: phylogenetic model averaging. *Molecular Biology and Evolution*, **25**, 1253–1256.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

Puritz JB, Addison JA, Toonen RJ (2012) Next-generation phylogeography: a targeted approach for multilocus sequencing of non-model organisms. *PLoS One*, **7**, e34241.

Putta S, Smith JJ, Walker JA *et al.* (2004) From biomedicine to natural history research: EST resources for ambystomatid salamanders. *BMC Genomics*, **5**, 54.

Quince C, Lanzen A, Davenport RJ, Turnbaugh PJ (2011) Removing noise from pyrosequenced amplicons. *BMC Bioinformatics*, **12**, 38.

Quinlan AR, Stewart DA, Stromberg MP, Marth GT (2008) Pyrobayes: an improved base caller for SNP discovery in pyrosequences. *Nature Methods*, **5**, 179–181.

Rambaut A, Drummond AJ (2007) Tracer v1.5, Available from http://beast.bio.ed.ac.uk/Tracer

Riley SPD, Shaffer HB, Voss SR, Fitzpatrick BM (2003) Hybridization between a rare, native tiger salamander (*Ambystoma californiense*) and its introduced congener. *Ecological Applications*, **13**, 1263–1275.

Routman E (1993) Population structure and genetic diversity of metamorphic and paedomorphic populations of the tiger salamander, *Ambystoma tigrinum*. *Journal of Evolutionary Biology*, **6**, 329–357.

Shaffer HB (1984) Evolution in a paedomorphic lineage. I. An electrophoretic analysis of the Mexican ambystomatid salamanders. *Evolution*, **38**, 1194–1206.

Shaffer HB, McKnight ML (1996) The polytypic species revisited: genetic differentiation and molecular phylogenetics of the tiger salamander *Ambystoma tigrinum* (Amphibia: Caudata) complex. *Evolution*, **50**, 417–433.

Shaffer HB, Thomson RC (2007) Delimiting species in recent radiations. *Systematic Biology*, **56**, 896–906.

Shaffer HB, Voss SR (1996) Phylogenetic and mechanistic analysis of a developmentally integrated character complex: alternate life history modes in ambystomatid salamanders. *American Zoologist*, **36**, 24–35.

Simcox TG, Marsh SJ, Gross EA, Lernhardt W, Davis S, Simcox MEC (1991) *SrfI*, a new type-II restriction endonuclease that recognizes the octanucleotide sequence, 5′-GCCC↓GGGC-3′ CGGG↑CCCG. *Gene*, **109**, 121–123.

Smith JJ, Kump DK, Walker JA, Parichy DM, Voss SR (2005a) A comprehensive expressed sequence tag linkage map for tiger salamander and Mexican axolotl: enabling gene mapping and comparative genomics in *Ambystoma*. *Genetics*, **171**, 1161–1171.

Smith JJ, Putta S, Walker JA *et al.* (2005b) Sal-Site: Integrating new and existing ambystomatid salamander research and informational resources. *BMC Genomics*, **6**, 181.

Weisrock DW, Shaffer HB, Storz BL, Storz SR, Voss SR (2006) Multiple nuclear gene sequences identify phylogenetic species boundaries in the rapidly radiating clade of Mexican ambystomatid salamanders. *Molecular Ecology*, **15**, 2489–2503.

Weisrock DW, Rasoloarison RM, Fiorentino I *et al.* (2010) Delimiting species without nuclear monophyly in Madagascar's mouse lemurs. *PLoS One*, **5**, e9883.

Yang Z, Rannala B (2010) Bayesian species delimitation using multilocus sequence data. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 9264–9269.

Zellmer AJ, Hanes MM, Hird SM, Carstens BC (2012) Deep phylogeographic structure and environmental differentiation in the carnivorous plant *Sarracenia alata*. *Systematic Biology*, **61**, 763–777.

E.M.O. and D.W.W. designed the study. H.B.S., G.P-O. and X.A-M. provided samples. E.M.O. performed the laboratory work. E.M.O., D.W.W., R.S. and C.T.B. designed the bioinformatic pipeline. E.M.O. performed the bioinformatics analyses. E.M.O., D.W.W. and J.S.W. designed and performed the population genetic and phylogenetic analyses. E.M.O., D.W.W., J.S.W., R.S., C.T.B. and H.B.S. wrote the manuscript. All authors read and edited the manuscript.

Data accessibility

All DNA sequences generated by the 454 runs in SFF format, NextAllele scripts, secondary alignments, STRUCTURE analyses, and *BEAST analyses: DRYAD entry doi:10.5061/dryad.03s86. A java version of NextAllele is available here: http://wars.ca.uky.edu/NextAllele/

Supporting information

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

Fig. S1 Species trees.

Fig. S2 Flowchart showing the steps involved in identifying SNPs (heterozygous positions) within each primary alignment.

Fig. S3 The frequency of the number of imperfect linkages (SNPs that could be linked in more than one way) within inferred genotypes (individual-locus combinations).

Fig. S4 Enlargement of S3 showing the bottom portion of the histogram more clearly.

Fig. S5 The frequencies of the proportions of imperfect linkages (SNPs that could be linked in more than one way) within inferred genotypes (individual-locus combination).

Table S1 Locality and sampling information for all individuals used in this study.

Table S2 Information for loci used in this study.

Table S3 Structure statistics resulting from analysis of the full data set.

Table S4 Approximate costs of this study compared with Sanger sequencing.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.