Title
The Added Value of Large-Scale Genomic Data in Conservation

Author
McCartney-Melstad, Evan

Publication Date
2016

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

The Added Value of Large-Scale Genomic Data in Conservation

A dissertation submitted in partial satisfaction of the requirements

for the degree Doctor of Philosophy in Biology

by

Evan McCartney-Melstad

2016

ABSTRACT OF THE DISSERTATION

The Added Value of Large-Scale Genomic Data in Conservation

by

Evan McCartney-Melstad

Doctor of Philosophy in Biology

University of California, Los Angeles, 2016

Professor Howard Bradley Shaffer, Chair

Large-scale genomic datasets have transformed the way we approach problems in biology. While genomic data are now common in studies of human disease and variation, they are rarely used in molecular ecology and conservation genetics. In this dissertation I investigate the added value of genomic-scale data to answer conservation-relevant questions for threatened and endangered amphibians and reptiles. The first chapter is a synthetic literature review of recent studies that used genetic data to learn about the ecology of amphibian species to aid in conservation. This review reveals that genomic studies in amphibians, likely due to their extremely large and complex genomes, have lagged behind other taxa, and includes future directions aimed squarely at large-genome analyses. Chapter two addresses this deficiency empirically by optimizing a genomic data collection strategy for amphibians with large genomes. This study shows that target capture is feasible in large-genome amphibians and experimentally

examines two specific modifications that greatly improve efficiency and substantially reduce cost when gathering population-scale genomic data. Chapter three makes use of advancements from chapter two, and presents empirical results from a genomic exon capture dataset of endangered tiger salamanders (*Ambystoma tigrinum*) on Long Island, NY. This chapter demonstrates the substantial benefits and insights derived from our genomic dataset for answering conservation-relevant questions compared to a previous study of the same system using traditional microsatellite analyses. The final chapter attempts to elevate conservation genomics to the next level by using whole genomes to quantify the impacts of alternative solar energy development scenarios on gene flow of the Mojave desert tortoise (*Gopherus agassizii*). We generated 270 low-coverage whole genome tortoise sequences to build a historical map of gene flow for the species across the Mojave Desert, modeled the effects on gene flow of a set of proposed renewable energy development alternatives, and supplied management agencies with a ranked list of impacts of the green energy alternatives. Overall, this dissertation provides context for the current use of genomic tools in conservation biology, methodological advances for their application to large-genome amphibians, and two examples of large-scale studies that showcase the added value of genomic datasets in conservation.

The dissertation of Evan McCartney-Melstad is approved.

James O Lloyd-Smith

Kirk Edward Lohmueller

Beth Shapiro

Howard Bradley Shaffer, Committee Chair

University of California, Los Angeles

2016

Dedication

This dissertation is dedicated to my mother, Kimberley McCartney, who selflessly forwent a PhD to give her children the best opportunity for a good education and a full life, and to my wife Catherine McCartney Reibel, my beloved partner in the appreciation of the natural world.

Table of Contents

List of Figures ..... **xi**

List of Tables ..... **xvi**

Acknowledgments ..... **xviii**

Vita ..... **xxi**

1 Amphibian molecular ecology and how it has informed conservation **1**

Introduction **1**

Within-population processes **2**

1.1.1 Methods and their applications to amphibians 2

1.1.2 Box 1: California tiger salamander (*Ambystoma californiense*) 3

1.1.3 Box 2: Great crested newt (*Triturus cristatus*) 5

1.1.4 Box 3: Mountain yellow-legged frog (*Rana muscosa*) 6

1.1.5 Amphibian conservation 7

1.1.6 Box 4: Chinese giant salamander (*Andrias davidanus*) 10

1.1.7 Box 5: Cane toad (*Rhinella marina*) 11

Genes on landscapes **11**

1.1.8 Methods and their applications to amphibians 11

1.1.9 Amphibian conservation 13

Genes that matter **15**

1.1.10 Methods and their applications to amphibians 15

1.1.11 Amphibian conservation 16

Future directions and critical knowledge gaps **18**

Conclusions **19**

1.1.12 For molecular ecologists 19

1.1.13 For conservation ecologists 19

1.1.14 For managers 19

Acknowledgements 19

References 19

2 Exon capture optimization in amphibians with large genomes 27

Abstract 27

Introduction 27

Materials and Methods 28

2.1.1 Array design and laboratory methods 28

2.1.2 Genetic data analysis 30

2.1.3 Assessing the importance of c0t-1 and individual input DNA amounts 31

Results 31

2.1.4 Presequencing library quantitation 31

2.1.5 Sequence data 31

2.1.6 Reference assembly and read mapping 32

2.1.7 Effects of c0t-1 and input DNA amount in capture reactions 32

Discussion 33

Acknowledgements 35

References 36

3 Population genomics of endangered tiger salamanders (*Ambystoma tigrinum*) on Long Island, NY reveals a highly structured species impacted by major roads 41

Abstract 42

Introduction **42**

Methods **46**

3.1.1 Sampling and laboratory 46

3.1.2 Reference assembly 47

3.1.3 SNP calling and genotyping 48

3.1.4 Population genetic analysis 49

3.1.5 Impacts of roads 52

Results **52**

3.1.6 Sampling 52

3.1.7 Reference assembly 53

3.1.8 SNP calling and genotyping 53

3.1.9 Genetic variation within cohorts 54

3.1.10 Isolation by distance (IBD) 54

3.1.11 Principal component analysis 55

3.1.12 Population clustering 55

3.1.13 Effective population size 56

3.1.14 Roads as barriers to dispersal 57

Discussion **57**

Conclusion **63**

Acknowledgements **64**

Figures **65**

Tables **71**

References **74**

4 Desert tortoises in the genomic age: population genetics and the landscape 79

Preface.....79

4.1.1 Context.....79

4.1.2 Future plans.....79

4.1.2.1 Current state of landscape-level planning for the Mojave desert tortoise.....79

Title page.....81

Abstract.....82

Executive summary: objectives and deliverables.....83

Introduction.....85

Current state of knowledge of desert tortoise landscape genetics.....86

Deciding among genomic approaches.....87

Methods.....88

4.1.3 Sampling.....88

4.1.4 Laboratory procedures.....89

4.1.5 Genomic data.....89

4.1.6 Modeling whole-genome sequencing.....91

4.1.7 Inference of genetic relationships.....91

4.1.8 Isolation by environment.....93

4.1.9 Landscape resistance models.....95

4.1.10 Evaluation of alternatives.....96

4.1.11 Reference locations.....96

4.1.12 Measure of isolation.....97

Results **98**

4.1.13 Overview 98

4.1.14 Sampling 98

4.1.15 Genetic data 99

4.1.16 Mapping statistics 99

4.1.17 Population structure 99

4.1.18 Isolation by distance 100

4.1.19 Model fit 101

4.1.20 Effect on gene flow of removing habitat 102

4.1.21 Effects on gene flow to single locations: examples 102

4.1.22 Combined effects on gene flow across the range 103

4.1.23 Comparing total effects of each alternative 104

4.1.24 Effects of removing each chunk 105

4.1.25 Comparison of all chunks across all alternatives 108

Discussion **109**

Acknowledgements **111**

Appendices **112**

Supplemental figures **125**

References **135**

List of Figures

Figure 1: Coverage across a sample target. The black bar on the bottom corresponds to the target region from which probes were synthesized. Each grey line represents a single library. There are two peaks of coverage, one centred on the target region, and a much higher spike of coverage at the left edge of the contig, likely corresponding to a repetitive region in the genome. The latter type of spike is reduced through the chimera-filtering steps described in the text. 30

Figure 2: Relationship between post-enrichment DNA concentration and percentage of raw reads mapping to targets. Each dot is an individual library: square = CTS, triangle = F1, circle = BTS*. For the full data set, adjusted $R^2 = 0.224$, P = 0.000204. After removing the single F1 outlier, adjusted $R^2 = 0.1732$, P = 0.00126. 31

Figure 3: Average sequencing depths across targets. The average sequencing depth across all targets regions averaged between all samples, calculated using samtools depth. The highest 31 values, which had depths higher than 30, are not shown here. 32

Figure 4: Relationship between individual input DNA and c_0t-1 amounts and PCR duplication rates and enrichment efficiency. Each dot is an individual library: square = CTS, triangle = F1, circle = BTS*. P-values for slope coefficients in the four panels are as follows: top left P = 1.399×10^{-7}, top right P=9.289×10^{-6}, bottom left P = 0.000672, bottom right P = 0.00896. 33

Figure 5: Predicted vs. actual unique reads on target using two-variable model. The model contains both c_0t-1 and individual input DNA, and the diagonal line shows the 1:1 relationship between predicted and actual unique reads on target. Points close to the line mean their unique reads on target are well predicted by the two variables, and points farther away from the line are not as well predicted. Each dot is an individual library: square = CTS, triangle = F1, circle =

BTS*. 34

Figure S1: The change in raw mapping rate as a function of post-enrichment qPCR cycle number. 38

Figure 6: Map of sampling localities. 65

Figure 7: Relationship between genetic similarity and geographic distance between individuals. The plot on the left uses raw Euclidean distance between individuals, while the plot on the left uses log-transformed Euclidean distances. 66

Figure 8: Relationship between genetic distance and geographic distance between ponds. The plot on the left uses raw Euclidean distance between ponds, while the plot on the left uses log-transformed Euclidean distances.....66

Figure 9: First eight principal components of the data. Letters on the graph correspond to samples from the same pond. Colors are used only to aid in distinguishing between letters. 67

Figure 10: Cross-validation error mean and standard deviations from 10 ADMIXTURE runs using different seeds. The red line is drawn at the mean+SD of the best-performing K value (K=12). The standard deviations for K=9 through K=13 overlap this line. 68

Figure 11: Admixture results from all 282 samples. Letters correspond to ponds from the sample map (Figures 1 and 4). White vertical lines separate sampling years within ponds, and black vertical lines separate ponds from one another. 69

Figure 12: Relationship between pond area and effective population size estimate. Ne estimates represent multiple-cohort calculations if multiple cohorts were samples, otherwise adjusted

single-year estimates were used. Ponds i, L, and Q were omitted because they did not contain enough samples to generate an estimate of Ne. 70

Figure 13: Visualizing the impacts of major roads on genetic differentiation between ponds. For the same geographic distance, ponds separated by major roads (indicated by red dots) tend to have higher levels of genetic differentiation. 70

Figure 14: Sample map. 89

Figure 15. Comparison of two different sequencing approaches in their ability to differentiate very slightly differentiated populations (Fst=0.001). 91

Figure 16. Reference points used to compute changes in gene flow across desert tortoise habitat. 97

Figure 17. Tortoise sample map with samples colored continuously by their score on PC1. Additionally, samples on the left side of PC1 (mostly the red samples) were divided again into two sets based on PC2, with samples in the top half plotted with triangles and the other samples plotted in circles. The Ivanpah Sample Map is an expansion of this area that shows how genes move around the mountains in more detail. Background colors show elevation. 100

Figure 18. Showing the slope of isolation by distance both within each group and between the two groups. Groups are defined by their scores on PC1 and are shown in the left panel. 101

Figure 19. Pattern of isolation by distance solely within a small geographic range in the Ivanpah Valley. The map on the right shows the location of each tortoise sample. 101

Figure 20. Showing the effects on hitting times to a single spot in the western Mojave as a result of removing the land in the preferred alternative. The dark green areas in the left panel were deemed inaccessible under this model. 103

Figure 21. Showing the effects of hitting times to a single spot in the central Mojave as a result of removing the land in the preferred alternative. 103

Figure 22. Mean and relative difference in commute times across the range as a result of removing the land in the preferred alternative. 104

Figure 23. Spatial configuration of the proposed development chunks (see Appendix 4). 108

Figure S2. Visualization of genetic structure. Dots represent individual tortoises, and they are colored by their scores on PC1. Background color corresponds to elevation. 125

Figure S3. Visualization of genetic structure. Dots represent individual tortoises, and they are colored by their scores on PC2. Background color corresponds to elevation. 125

Figure S4. Effects of removing the development areas in Alternative 1. 126

Figure S5. Effects of removing the development areas in Alternative 2. 126

Figure S6. Effects of removing the development areas in Alternative 3. 127

Figure S7. Effects of removing the development areas in Alternative 4. 127

Figure S8. List of 83 landscape layers. 128

Figure S9. Matrix of correlations between spatial data layers. 129

Figure S10. List of environmental data layers evaluated in 6, 12, and 24-layer models. 130

Figure S11. Spatial configuration of proposed development chunks in Alternative 1 that we analyzed. 131

Figure S12. Spatial configuration of proposed development chunks in Alternative 2 that we analyzed. 132

Figure S13. Spatial configuration of proposed development chunks in Alternative 3 that we analyzed. 133

Figure S14. Spatial configuration of proposed development chunks in Alternative 4 that we analyzed. 134

List of Tables

Table 1. Key recent amphibian studies exploring within-population processes.....8

Table 2. Key recent amphibian studies exploring genes on landscapes. 14

Table 3. Key recent amphibian studies exploring genes that matter.....17

Table 1. Model comparison predicting percentage of unique reads on target, sorted by AIC values. 33

Table 2. Model comparison predicting average depth across tar- get region, sorted by AIC values. 33

Table S1. Individual libraries (1-53), their treatment levels, and description of yields and sequencing statistics. Number in parenthesis in library (first column) is the enrichment (1-24), and shows how libraries were pooled. For example, 22(18) and 25(18) indicates libraries 22 and 25 were pooled into a single tube (number 18) prior to enrichment.....39

Table S2. Table S2—Post-enrichment concentrations and sequencing efficiency results. Number in parenthesis is library name as in Table S1. 40

Table 4. Pond localities, areas, Watterson's θ estimates, sampling, and effective population size estimates. Pond areas were estimated from Google Earth satellite images taken in March 2007. Single-year estimates were corrected for iteroparity-induced downward bias as explained in Methods, and both single-year and pooled-year estimates were corrected for dense locus sampling on chromosomes. Infinite values indicate that sample sizes were likely too small to estimate Ne. N=number of samples included in analyses. Ne=Effective population size estimates

using LD method.....71

Table 5. Mantel test results: P-values calculated using 999,999 permutations. R_M is the Mantel R statistic, and R_M^2 is the square of the Mantel R statistic.....72

Table 6. Pairwise Fst values between ponds. Cells are colored by the magnitude of difference between ponds, with red being relatively low differentiation and green being relatively high differentiation. Bolded cells/values are not significantly different from 0 (p > Benjamini-Yekutieli-corrected 0.05). 73

Table 7: Effects of development alternatives on tortoise gene flow. 105

Table 8: Effects of removing each development "chunk" within the preferred alternative. 106

Table S3: Effects of removing individual chunks from all evaluated alternatives. 119

Acknowledgements

I am extremely grateful to Brad Shaffer for his mentorship over the years. Working together with him has been an absolute joy, and I owe much of who I am as a scientist to him. Brad consistently improves my ideas and writing, and is a perfect role model for how to make science and conservation work together. I thank the other members of my dissertation committee, Jamie Lloyd-Smith, Kirk Lohmueller, and Beth Shapiro, for the influence they've had on me, both through their direct mentorship and as positive examples through the quality of their own work. I thank Thomas J. Near and Martin Mendez for their dedicated and excellent mentorship throughout my undergraduate and master's work, respectively.

My wife Catherine has supported me in every way imaginable, from talking through challenges and giving me perspective to labeling tubes on the weekend; I thank her for her patience, encouragement, and heart. I thank my parents Mike Melstad and Kimberley McCartney for instilling in me respect for the natural world and showing me the power of working hard to solve important problems. I thank my siblings Anna, Austin, and Chris McCartney-Melstad for challenging their little brother to be better and for serving as role models for making world a healthier, more equitable, and safer place for everyone.

Mark Phuong, Richard Hedley, and Anna Taylor have all been supportive and inspiring friends at UCLA. Past and present members of the Shaffer Lab, including Phil Spinks, Gary Bucciarelli, Gideon Bradburd, Muge Gidis, Jannet Vu, Erin Toffelmeier, Kevin Neal, Robert Cooper, Ben Wielstra, Jaime Ashander, Genevieve Mount, Sarah Wenner, Tara Luckau, Mario Colon, and Sid Shah have all been exemplary colleagues. I thank Peter Ralph for tremendously helpful and enjoyable collaborations. The staff in the Department of Ecology and Evolutionary

Biology at UCLA have been extremely supportive, particularly Jocelyn Yamadera, who was very much appreciated as a persistent and effective advocate.

This work was supported by Graduate Assistance in Areas of National Need Fellowships from the US Department of Education and research support from the National Science Foundation, California Department of Fish and Wildlife, US Fish and Wildlife Service, La Kretz Center for California Conservation Science, and the Andrew Sabin Family Foundation.

Chapter 1 is a reprint of McCartney-Melstad E, Shaffer HB (2015) Amphibian molecular ecology and how it has informed conservation. *Molecular Ecology*, **24**, 5084–5109. A license for reuse of the publication in this dissertation was provided by John Wiley & Sons, Inc. H. Bradley Shaffer was the Principal Investigator in support of this paper.

Chaper 2 is a reprint of McCartney-Melstad E, Mount GG, Shaffer HB (2016) Exon capture optimization in amphibians with large genomes. *Molecular Ecology Resources*, **16**, 1084–1094. A license for reuse of the publication in this dissertation was provided by John Wiley & Sons, Inc. I contributed to the design of the study, performed some of the molecular work, analyzed the results and wrote the manuscript. Genevieve G. Mount contributed to the design of the study, performed most of the laboratory work and revised the manuscript. H. Bradley Shaffer contributed to the design of the study and interpretation of results and revised the manuscript. All authors read and approved the final manuscript.

For Chapter 3, I performed some of the laboratory work, conducted the analyses, and wrote the manuscript. Jannet Vu performed most of the laboratory work, assisted with writing the introduction, and edited the manuscript. H. Bradley Shaffer obtained funding for the project, conducted the field work, and edited the manuscript

Chapter 4 is a version of a report submitted to the California Department of Fish and Wildlife. This project represents a truly collaborative effort across several individuals and institutions. I played four main roles in this work. First, I performed the simulations that helped to convince us that a low-coverage full genome approach would be preferable to other approaches for this study. Second, I conducted the laboratory work to generate all of the genomic data. Third, I performed all of the bioinformatic analyses to bring the raw sequence data to the various stages required for different analyses, wrote the software to quickly estimate pairwise genetic relationships between individuals using read count data in low coverage sequence data (www.github.com/atcg/cPWP), and performed some of the population genetic analyses. Finally, I wrote and edited several sections of the report. Peter Ralph (in collaboration with Gideon Bradburd and Erik Lundgren) invented and implemented the random walk-based gene flow model that we used to estimate reductions in gene flow due to development, and also developed the theory behind the read-based pairwise pi and genetic covariance estimation used here, in addition to writing and editing several sections of the report. Gideon Bradburd also performed some of the population genetic analyses and wrote and edited several sections of the report. Jannet Vu collected and curated the spatial environmental data and generated the maps that are included in the report, and also wrote two appendices. Bridgette Hagerty, Fran Sandmeier, Chava Weitzman, and C. Richard Tracy contributed approximately 1,000 desert tortoise blood samples that they collected (at great effort), in addition to knowledge of tortoise ecology and conservation, as well as the results of previous microsatellite-based genetic analyses and editing of the report. H. Bradley Shaffer wrote and edited several sections of the report, and is listed as the lead author for his role in conceiving of and obtaining funding support for the project.

VITA

EDUCATION

Columbia University
MA, Conservation Biology (May 2012)
New York, NY

Yale University
B.S., Biology (May 2008)
New Haven, CT

PUBLICATIONS

Spinks PQ, Thomson RC, McCartney-Melstad E, Shaffer HB. 2016. Phylogeny and temporal diversification of the New World pond turtles (Emydidae). *Molecular Phylogenetics and Evolution*, 103:85-97.

McCartney-Melstad E, Shaffer HB. 2015. Amphibian molecular ecology and how it has impacted conservation. *Molecular Ecology*, 24:5084-5109.

McCartney-Melstad E, Mount G, Shaffer HG. 2015. Exon capture optimization in large-genome amphibians. *Molecular Ecology Resources,* 16:1084-1094.

Shaffer HB, Gidis M, McCartney-Melstad E, Neal K, Oyamaguchi H, Tellez M, Toffelmier E. 2015. Conservation genetics and genomics of reptiles and amphibians. *Annual Reviews of Animal Biosciences,* 3:113-138.

McCartney-Melstad E, Waller T, Micucci PA, Barros M, Draque J, Amato G, and Mendez M. 2012. Population structure and gene flow of the yellow anaconda (*Eunectes notaeus*) in Northern Argentina. *PLoS ONE*. 7(5):e37473

PRESENTATIONS

Using Exon Capture Data to Measure the Tempo and Extent of Hybridization Between an Endangered Amphibian and an Introduced Congener. 2016. Eighth World Congress of Herpetology. Hangzhou, China. *Oral presentation.*

Complete genome sequences predict the impacts of alternative energy development scenarios on the endangered Mojave desert tortoise (Gopherus agassizii). 2016. Eighth World Congress of Herpetology. Hangzhou, China. *Oral presentation.*

Exon Capture Reveals Insights into the Conservation Status and Genome Organization of the Endangered California Tiger Salamander. 2016. Annual meeting of the American Society of Icthyologists and Herpetologists. New Orleans, LA. *Oral presentation.*

Whole Genome Resequencing Provides Novel Landscape Genomic Insights for Desert Tortoise Conservation. 2015. Desert Tortoise Council 40th Annual Symposium. Las Vegas, NV. *Oral presentation.*

Conservation Genomics of the California Tiger Salamander (*Ambystoma californiense*): Past, Present, and Future. 2015. California/Nevada Amphibian Population Task Force. Malibu, CA. *Oral presentation.*

Exon Capture Optimization for a Large-Genome Amphibian. 2014. Annual meeting of the American Society of Ichthyologists and Herpetologists. Chattanooga, TN. *Oral presentation.*

A Landscape Genetic Approach to Predicting Pesiticide Impacts on the California Tiger Salamander. 2013. UCLA EcoEvoPub. *Oral presentation.*

Fossil-calibrated Molecular Phylogenies and Divergence Time Estimates of the Two Extant Coelacanth Species (*Latimeria*). 2007. Annual meeting of the American Society of Ichthyologists and Herpetologists. St. Louis, MO. *Oral presentation.*

Freshwater fish field work and coelacanth evolution. 2008. Mellon Forum, Yale University. *Oral presentation*.

MOLECULAR ECOLOGY

Molecular Ecology (2015) **24**, 5084–5109 doi: 10.1111/mec.13391

INVITED REVIEWS AND SYNTHESES

Amphibian molecular ecology and how it has informed conservation

EVAN MCCARTNEY-MELSTAD and H. BRADLEY SHAFFER

Department of Ecology and Evolutionary Biology, La Kretz Center for California Conservation Science, and Institute of the Environment and Sustainability, University of California, Los Angeles, 610 Charles E Young Drive South, Los Angeles, CA, USA

Abstract

Molecular ecology has become one of the key tools in the modern conservationist's kit. Here we review three areas where molecular ecology has been applied to amphibian conservation: genes on landscapes, within-population processes, and genes that matter. We summarize relevant analytical methods, recent important studies from the amphibian literature, and conservation implications for each section. Finally, we include five in-depth examples of how molecular ecology has been successfully applied to specific amphibian systems.

Keywords: amphibians, conservation genetics, landscape genetics, natural selection and contemporary evolution, population genetics—empirical, wildlife management

Received 20 June 2015; revision received 15 September 2015; accepted 16 September 2015

Introduction

Amphibians are facing many challenges leading to global extinctions and population declines. Of the 7450 currently recognized species of amphibians, about one-third are classified as vulnerable, endangered or critically endangered according to the IUCN Red List (Stuart *et al*. 2004; Wake & Vredenburg 2008). A great deal of attention has been focused on this issue since 1989, when scientists at the first World Congress of Herpetology in Canterbury realized that diverse groups of frogs were unexpectedly declining (Wake 1998). In the following two decades, researchers have proposed a large set of factors that may be contributing to this decline, including overharvesting, land use change, pesticides and toxicants, disease, global climate change, invasive species and hybridization (Collins & Storfer 2003).

Amphibian conservation is particularly problematic because so many amphibian declines have occurred within protected areas, suggesting that habitat loss is not solely responsible for these declines (Bosch *et al*. 2001; Knapp & Matthews 2001; Vredenburg 2004; Wake & Vredenburg 2008). This trend is deeply disturbing, as it indicates that our best traditional tool for conservation, habitat protection, is not sufficient to protect amphibians in many parts of the world. It also forces us to critically evaluate additional factors that may be causing declines.

Correspondence: Evan McCartney-Melstad, Fax: +1 (310) 206-3987; E-mail: evanmelstad@ucla.edu

Molecular genetic analyses are exceptionally well suited for this task. Many amphibians are difficult or impossible to observe directly in the field: most species are nocturnal, and the vast majority spend significant portions of their lives underwater, underground, or in otherwise cryptic habitats. Information regarding population size and landscape use by animals has traditionally been gathered through direct observation (Arnason 1973; Peterson & Cederholm 1984; Johnson *et al*. 2001), but these methods are often not feasible for amphibian population biologists. Genetic data are relatively easy to collect, especially for amphibian species that gather in breeding pools where entire cohorts can often be sampled nondestructively with a relatively minimal effort (Polich *et al*. 2013). As genomic tools become increasingly available for nonmodel amphibian systems, molecular ecological approaches can and should be a primary tool for amphibian conservation biologists.

Analysis of genetic information can be useful for conservation in several broad areas. One is the estimation of demographic parameters and trends, a subject of tremendous importance as we strive to understand the patterns and time-course of decline and historical

population bottleneck events. Another is population and phylogeographic structure, two key elements in delineating management units (Waples 1991), determining critical habitat areas and deducing likely movement corridors (Storfer *et al.* 2006). Finally, genomic studies of wild populations potentially allow researchers to identify 'genes that matter', for local adaptation or that play a critical role in mediating key life history traits that are important for conservation and management.

Amphibians can be categorized in at least two fundamental ways: phylogenetically and by their life history. The three major clades of amphibians (Anura—frogs and toads; Caudata—salamanders and newts; and Gymnophiona—caecilians) are taxonomically and morphologically diverse, as might be expected given their crown age of *c.* 360 million years (Roelants *et al.* 2007). However, even more striking is their diversity of life history strategies, which rival those of vertebrates as a whole (Duellman & Trueb 1985). At the most fundamental level, amphibians can be split into two major groups with respect to life history: those with a biphasic life cycle with aquatic larval and terrestrial postmetamorphic phases, and direct-developing species that lack the aquatic larval phase entirely. Direct-developing amphibians are often exceedingly difficult to sample and observe in nature, because they are not constrained to spend weeks or months at a breeding site as larvae and adults. However, they can reach truly astonishing population densities, and in some cases, population-level samples are relatively easy to acquire (Semlitsch *et al.* 2014). Biphasic amphibians, on the other hand, often lend themselves quite well to population genetic studies, given the ease of studying postmetamorphic juveniles and adults as they leave and enter breeding ponds (Pechmann *et al.* 1989; Trenham *et al.* 2001), as well as entire cohorts of larvae at a breeding site. The insights that can be derived from such sampling snapshots are often invaluable for conservation planning efforts.

Several reviews of genetic studies in amphibians have been published in the last 15 years (Jehle & Arntzen 2002; Beebee 2005; Storfer *et al.* 2009; Emel & Storfer 2012). We therefore focus primarily on the most recent literature, highlighting studies that make use of the most modern data acquisition and analysis techniques wherever possible. Our review of amphibian molecular ecology is divided into three areas—within-population processes, genes on landscapes, and genes that matter. In each section, we outline some of the most commonly used methods, review recent studies of amphibian molecular ecology and discuss conservation relevance and management outcomes. Finally, we explore in depth how studies in molecular ecology have informed conservation efforts in five different amphibian case studies (Boxes 1–5).

Within-population processes

Methods and their applications to amphibians

Estimation of demographic parameters, including population size and trajectory, is an area where genetic analysis has proven extremely useful and can provide critical information for amphibian conservation efforts (Beebee & Griffiths 2005). Genetic drift acts to reduce within-population variability. The speed at which genetic drift acts is proportional to the effective population size, which is often much smaller than the census population size (Funk *et al.* 1999; Waples 2002). Very low effective population sizes lead to increased inbreeding and a continual loss of genetic variation within populations, which may in turn lead to many problems including inbreeding depression, reduced disease resistance and loss of adaptive capacity (Frankham 1995a; Waples 2002). In 2010, Allentoft & O'Brien reviewed a collection of studies that examined the relationship between reduced genetic diversity and fitness in amphibians. Of the 19 studies they analysed that directly quantified these genetic diversity–fitness relationships, 15 of them provided evidence that levels of genetic diversity impacted important traits such as growth or survival (Allentoft & O'Brien 2010).

Several methods are currently in use that exploit either temporally spaced genetic sampling across multiple generations, or single-sample snapshots of genetic variation to estimate effective population size (N_e). As single-generation cohort sampling of larvae is often possible with pond-breeding amphibian species, it is possible to infer effective population sizes using the temporal sampling strategy for many amphibian studies, although overlapping generations of breeding adults that produce those cohorts can still influence estimates (Serbezov *et al.* 2012). These methods make use of the variance in the change of allele frequencies between sampling events (that is, between larval cohorts) to infer the size of a Hardy–Weinberg-ideal population that is losing heterozygosity at the observed rate (Krimbas & Tsakas 1971; Nei & Tajima 1981). A key element of this method is that the true N_e need not remain stable across sampling periods, although the harmonic mean N_e will be returned in that case (Waples 2005). The effective population size for a population is usually lower than the census population size, with estimates of <100 (and often far less) for many amphibian populations (Phillipsen *et al.* 2011).

One of the main reasons that inference of effective population size is important is to estimate and forecast

Box 1. System 1: California tiger salamander (*Ambystoma californiense*)

The California tiger salamander (*Ambystoma californiense*, CTS) is a state- and federally protected species endemic to California (US Fish and Wildlife Service 2000; Bolster 2010). It breeds primarily in vernal pools and faces threats from extensive habitat conversion to agricultural lands (Davidson *et al.* 2002), pesticide usage (Ryan *et al.* 2012) and hybridization with the invasive barred tiger salamander [*Ambystoma* (*tigrinum*) *mavortium*, BTS], which was introduced to California in the 1950s (Riley *et al.* 2003). Our laboratory has been working on this system for the past 20 years, and we have, along with others, learned a great deal about tiger salamanders in California by studying their molecular ecology.

Prior to 2000, it was clear that most populations of CTS were in danger of extinction (Barry & Shaffer 1994; Fisher & Shaffer 1996). It is, however, extremely difficult to characterize the extent to which the species has declined through direct observation given their cryptic nature and the difficulties in finding and accessing many of their breeding sites. To better understand the demography of CTS populations, Wang *et al.* (2011) used microsatellites to explore several aspects of within-population processes in 10 ponds in Merced County, California. They found that ponds generally had small effective population sizes (an average of 30) and that effective population sizes were highly correlated with the size of the breeding ponds where animals were sampled (particularly true for natural breeding pools, less so for human-constructed ones). They also found genetic evidence of population bottlenecks within all sampled populations. While microsatellites have been effective in investigating recent within-population processes of different ponds, phylogeographic studies using mitochondrial DNA have helped to reinforce the validity of the Santa Barbara and Sonoma County populations as distinct population segments worthy of protection under state and federal law (Shaffer *et al.* 2004).

Understanding habitat use and movement patterns is a critical component of threatened species conservation. In the case of CTS, two questions are particularly important: (i) How far do salamanders generally disperse from their ponds after breeding? and (ii) Which parts of the landscape do salamanders use when moving between ponds? The first question has benefitted greatly from long-term ecological (drift fence) studies (Trenham *et al.* 2001; Trenham & Shaffer 2005; Searcy *et al.* 2013, 2014). The second is more difficult to address with observational data, particularly in large landscapes. Preliminary landscape genetic work has helped to answer this question using least-cost path analysis to determine vegetation types that are conducive to salamander movement (Wang *et al.* 2009a). The results of this study were surprising, as chaparral was found to be less resistant to salamander movement than grassland habitat, even though the species is typically considered to be a grassland specialist (Trenham *et al.* 2000). When visualized in the context of existing breeding ponds, this information can help prioritize upland habitat for conservation that facilitates salamander movement and metapopulation dynamics between breeding ponds. Further work is currently underway to understand these patterns at multiple spatial scales with greater genetic resolution and larger genomic data sets.

Hybridization with the invasive BTS is among the greatest threats facing CTS. Hybrid animals are more robust and have increased survival rates compared to pure native CTS, leading to concerns that hybrids will be strongly favoured by natural selection and replace pure CTS (Fitzpatrick & Shaffer 2007b; Ryan *et al.* 2012). Additionally, the presence of CTS/BTS hybrids has been found to negatively impact other amphibians breeding in the same ponds (Ryan *et al.* 2009; Searcy *et al.* 2015).

Hybrids were first confirmed in 2003 from samples collected at the Johnson Canyon Landfill in Salinas, CA by analysing a combination of targeted mitochondrial and nuclear restriction fragment length polymorphisms (Riley *et al.* 2003). After increasing the sampling to a set of 12 diverse ponds in the Salinas Valley, Fitzpatrick & Shaffer (2004) discovered that artificial perennial ponds were especially favourable to hybrids, which is particularly concerning given the large number of perennial ponds in agricultural landscapes.

Further expanding their pond sampling to 85 ponds across the range of CTS, Fitzpatrick & Shaffer (2007a) used eight ancestry-informative genetic markers (seven nuclear, one mitochondrial) to characterize the extent to which non-native alleles had spread 50 years after their introduction. They found the same pattern in all eight markers—that non-native alleles were mostly restricted to the Salinas Valley, largely within 12 km of a diverse set of known introduction sites. Scaling up to 68 EST-derived ancestry-informative markers, however, revealed a novel pattern of genetic introgression: while 65 of the markers showed a similar pattern to the eight-marker study (introgression limited mostly to the Salinas Valley), three of the non-native alleles had synchronously moved ~94 km further north and had almost completely replaced their native allele counterparts (Fitzpatrick *et al.* 2009, 2010). This

Box 1. *Continued*

implied that those three markers were close to genomic regions experiencing strong natural selection and that they epistatically interact in unison.

Current work in the laboratory is using new genomic resources to score thousands of markers throughout the genome at hundreds of ponds for samples collected over the past 30 years to track the movement of non-native alleles in real time, so that we may better understand how natural selection is acting in this system.

A California tiger salamander (on right) and a hybrid salamander (on left). Photo: Jarrett Johnson

the effects of inbreeding depression, which can have significant impacts on fitness in wild populations (Crnokrak & Roff 1999). This is a particularly difficult quantity to investigate explicitly, as it requires estimates of the extent of inbreeding in different individuals as well as measures of fitness. While some studies in amphibians have found correlations between degree of inbreeding and fitness (Halverson *et al.* 2006), others have not found such evidence (Williams *et al.* 2008). One interesting corollary to this in r-selected amphibians is that natural selection can act strongly on eggs or larvae to reduce the observed amount of inbreeding in the metamorphosing population, a pattern that could be missed if sampling only eggs or larvae. Such a pattern was seen in the Italian agile frog (*Rana latastei*), with tadpoles showing a positive inbreeding coefficient and metamorphosing froglets exhibiting a negative inbreeding coefficient (Ficetola *et al.* 2011; but see Phillipsen *et al.* 2010).

Conservation plans for species also benefit from information regarding recent population increases or decreases, and understanding how effective population sizes have changed through time provides context to current patterns of amphibian decline. Reconstructing historical fluctuations of effective population size typically relies on coalescent approaches (Kingman 1982a, b). Using the coalescent, a sample of alleles taken from a population may be stochastically modelled backwards through time, with the effective population size (usually multiplied by the mutation rate) as an estimated parameter (Kuhner *et al.* 1995). This effective population size parameter need not be a static number—it can itself be a value that increases or decreases through time (Griffiths & Tavare 1994). By discretizing historical periods and allowing different periods to exhibit patterns of population increase or decrease, one can model complex histories of effective population size increase or decrease that can be compared using likelihood and model comparison methods (Griffiths & Tavare 1994; Kuhner *et al.* 1998; Pybus *et al.* 2000; Drummond *et al.* 2005; Minin *et al.* 2008). Blair *et al.* (2013) used such an approach to determine that population sizes in lineages of Asian tree frogs (*Polypedates leucomystax*) had recently increased, helping them to understand how climate has affected these frogs through time.

Although coalescent modelling is very useful for understanding historical fluctuations of population size, additional methods can also be used to detect patterns of recent population expansion or decline. Several of these methods test for deviations from mutation–drift equilibrium using DNA sequence data or microsatellites (Tajima 1989; Cornuet & Luikart 1996; Garza & Williamson 2001). These tests have become widespread in molecular ecology studies and species management recommendations. For instance, Richter *et al.* (2009) analysed microsatellite data in the endangered dusky gopher frog (*Rana sevosa*), showing that these frogs had markedly lower genetic variation than closely related species, and also exhibited excess heterozygosity, indicating a genetic bottleneck. Richter & Nunziata (2014) followed up on this work and found strong evi-

Box 2. System 2: Great crested newt (*Triturus cristatus*)

The *Triturus cristatus* species complex is comprised of at least seven species, although our understanding of the relationships between these species and their distributions has changed considerably over the past several years and is still not fully resolved (Wielstra *et al.* 2014). The great crested newt, *Triturus cristatus*, is itself found throughout northern Europe, and although it has a large geographic range, many populations have been found to be decreasing over the past several decades (Edgar & Bird 2006). Some of the main threats facing the species include habitat loss and fragmentation, agricultural pollutants, and introduced predatory fish (Edgar & Bird 2006). Great crested newts are protected under Appendix 2 of the Bern Convention, Annexes II and IV of Europe's Habitats Directive, and separately in many countries in Europe.

The natural history and molecular ecology of great crested newts has been relatively well studied compared to most amphibians. Of particular importance for conservation has been accurately estimating population size and population structure across multiple scales. To get a sense of effective and census population sizes, Jehle *et al.* (2001) analysed eight microsatellite loci in *Triturus cristatus* from three ponds. They found that effective population sizes were very small, on the order of 10 individuals. They also performed a mark–recapture study to estimate the adult census size of the populations and found that census population sizes were approximately one order of magnitude larger. This is in general agreement with the ratios of effective to census population sizes in many other species (Frankham 1995b; Brede & Beebee 2006). Finally, the authors did not find strong statistical support for recent population bottlenecks, which was unexpected because the region had been colonized by *Triturus cristatus* only decades ago, presumably from a small number of founders.

Molecular analyses of population structure and migration have also been important in improving our understanding of how great crested newts interact with the landscape. Jehle *et al.* (2005a) assayed seven microsatellites from *Triturus cristatus* populations spread across 15 ponds in a small area (~25 km^2). They used *Structure* (Pritchard *et al.* 2000) and *BAPS* (Corander *et al.* 2003) analyses to compare the configuration of genetically identifiable populations with the sampled ponds. They found evidence for 7 or 11 clusters (depending on the methodology), indicating that while some of the ponds did not segregate into unique population units, many of them are recoverable as genetically distinct breeding groups. This information is quite useful in a conservation framework, as it argues for conserving as many ponds as possible, even within a relatively small landscape, to maintain the highest levels of genetic diversity (Jehle *et al.* 2005a).

A later analysis of samples from the same ponds used a novel method to infer rates of recent migration (Jehle *et al.* 2005b). This algorithm is an extension of BayesAss (a Bayesian method that estimates rates of recent migration) that makes use of the fact that the sampled aquatic larvae are not able to themselves be migrants among populations (Wilson & Rannala 2003; Jehle *et al.* 2005b). Jehle *et al.* (2005b) analysed eight microsatellites and determined that migration between populations was usually asymmetric and that no pond had received more than seven reproductively successful migrants in their sample year. A similar BayesAss analysis in California tiger salamanders also found extremely limited movement between ponds (Wang *et al.* 2009a). These results reinforce the conservation-relevant conclusion that many ponds represent distinct populations, as several of the ponds in both studies received no reproducing migrants at all.

Triturus cristatus. Photo: Michael Fahrbach

Box 3. System 3: Mountain yellow-legged frog (*Rana muscosa*)

Rana muscosa is endemic to southern and central California, living above 1074 m elevation (Stebbins 2003; Vredenburg *et al.* 2007). It is classified as an endangered species under the Endangered Species Act by both the US Federal Government (US Fish and Wildlife Service 2002, 2014a) and the IUCN Red List (Hammerson 2008). The main threats to this species are predation by introduced fish (Knapp & Matthews 2001), disease caused by *Bd* (Fellers *et al.* 2001; Vredenburg *et al.* 2010), and airborne pesticides (Davidson *et al.* 2002; Fellers *et al.* 2004). Reintroductions of captive-bred mountain yellow-legged frogs have been undertaken with mostly disappointing results (Fellers *et al.* 2007; US Fish and Wildlife Service 2012, but see Santana *et al.* 2015), underscoring the need to understand the ecology of the species and the causes of its decline before conservation actions will be effective. Genetic studies have been a key to ongoing conservation efforts in *Rana muscosa*, spanning several of the categories covered in this review.

Perhaps most importantly, genetic analyses played a crucial role in establishing the distinctness of *Rana muscosa* to its sister species, the Sierra Nevada yellow-legged frog (*Rana sierrae*). Vredenburg *et al.* (2007) analysed 1901 bp of mitochondrial DNA from frogs across the range of the species and uncovered a deep phylogenetic divide between samples in the northern/central Sierra Mountains and other parts of the range (Kimura-corrected average sequence divergence of 4.6%). They also showed that several morphological characters (including advertisement calls) covaried with this split. Consistent with the two-species interpretation, the authors found evidence of isolation by distance within each putative species, but not in models that included both species. Based on this evidence, they elevated the populations in the northern and central portions of the Sierra Nevada mountains to species status, and all populations of both of these species are now protected by US federal law (US Fish and Wildlife Service 2002, 2014a). Given the complex taxonomic history of the group and that the genetic evidence for this species split is purely mitochondrial, we hope that additional nuclear data can test this critical systematic decision, although the corroborating morphological and call data suggest that it probably does identify two non-interbreeding lineages.

Following the *R. muscosa/sierrae* split, the known range of *Rana muscosa* approximately halved in size to the southern Sierra Nevada (Tulare and Fresno County south), plus the San Jacinto, San Bernardino, San Gabriel and Palomar mountain elements of the Peninsular ranges in Southern California (Vredenburg *et al.* 2007). To better set management priorities within *Rana muscosa*, and also to aid in the selection of sources for frogs to establish captive breeding colonies, Schoville *et al.* (2011) quantified genetic variation within and between populations of frogs in these southerly ranges. They analysed data from nine microsatellite and one mitochondrial locus and found low genetic variation within the species generally, and strong population structure with little migration between the nine remaining populations that in total have an estimated adult census size of 166 individuals (Backlin *et al.* 2013). Their microsatellite data also showed evidence of recent population bottlenecks. Using biogeographic analyses of the mitochondrial sequence data, Schoville *et al.* determined that the populations in the different mountain ranges were likely previously connected by more continuous habitat and that they have been separated for over 40 000 years. Taken together, this evidence argues strongly for managing populations within the separate southern California mountain ranges as independent, nontransferable conservation units. It also suggests that frogs from additional locations should be integrated into captive breeding programmes (but not mixed with other distinct lineages) to maintain the rangewide genetic diversity of the species.

Extensive knowledge of the ecology and history of *Rana muscosa* is imperative for captive breeding and translocation strategy, but releasing animals into hostile locations can lead to low survival rates (Fellers *et al.* 2007). One major worry for reintroductions of mountain yellow-legged frogs into the wild is that they are apparently susceptible to chytridiomycosis. To better understand their infection response to *Bd*, Rosenblum *et al.* (2012) quantified expression patterns for several tissues in infected *Rana muscosa*. They did not observe a strong immune

dence that frogs with higher heterozygosity survived better than more inbred individuals. These studies have in turn informed captive breeding and recovery planning (US Fish and Wildlife Service 2014b).

Phylogeographic analyses can also yield valuable information for making conservation decisions. Since the field began in the 1980s, phylogeographic studies have used molecular data to connect the fields of phylogenetics and population genetics to ask questions about how geology and geography have structured species and populations through time (Avise *et al.* 1987). The first method that attempted to evaluate alternative models of the mechanisms underlying phylogeographic patters was nested clade phylogeographic analysis (NCPA), and its first application was in salamanders (Templeton *et al.* 1995). NCPA uses haplotype networks and an a posteriori inference key to draw conclusions about past population processes (Templeton *et al.* 1995;

Box 3. *Continued*

response, consistent with studies indicating that *Bd* has played a central role in species declines (Vredenburg *et al.* 2010). Further studies investigating the role genetic (or perhaps bacterial, see Woodhams *et al.* 2007) variation plays in conferring resistance or susceptibility could be instrumental in establishing targeted captive breeding programmes designed to increase the numbers of *Bd*-resistant populations in the wild.

Rana muscosa. Photo: C. Brown, USGS

Templeton 1998). Researchers have vigorously debated the merits of NCPA for over a decade (for a summary, see Hickerson *et al.* 2010; Bloomquist *et al.* 2010). More recent methods in phylogeography have focused on modelling the stochasticity of demographic processes and explicitly estimating statistical error (Knowles & Maddison 2002; Hickerson *et al.* 2010). Current phylogeographic analyses often utilize coalescent simulations of many unlinked nuclear gene genealogies (Kingman 1982a; Beerli & Felsenstein 2001), which allows researchers to estimate demographic parameters and compare different demographic and geological explanations for an observed pattern of population subdivision (Knowles & Maddison 2002; Hickerson *et al.* 2010).

The empirical literature using molecular tools to explore within-population processes in amphibians has focused on a variety of systems, most prominently Nearctic anurans and salamanders (Table 1). As might be expected, most of these studies utilize traditional mitochondrial, microsatellite and (more rarely) allozyme markers to make demographic and phylogeographic inferences, although we expect that the next decade will see a radical expansion into next-generation genomic studies (Shaffer *et al.* 2015). In Table 1, we summarize some of the key studies that highlight different demographic outcomes in amphibian molecular ecological work of the past decade.

Amphibian conservation

An accurate understanding of effective and census population size is critical for conservation planning (Frankham 1995a; Luikart *et al.* 2001; Palstra & Ruzzante 2008). However, we lack even basic demographic knowledge for the vast majority of amphibian species, making this a critical area of need for future research. Molecular ecological analyses offer an important pathway forward in the absence of long-term field ecological studies, and many of the existing amphibian molecular ecology studies have characterized demographic parameters to better understand threatened species (several such recent studies are summarized in Table 1). Estimates of modern-day effective population sizes as well as historical fluctuations in N_e are regularly used by management agencies in species listings and recovery plans (US Fish and Wildlife Service 2005; Hallock

Table 1 Key recent amphibian studies exploring within-population processes

Citations	Species	Data	Conclusions
Dufresnes *et al*. (2013)	European tree frog (*Hyla arborea*)	2 mt regions, 1 nuclear gene and 30 microsats	Historical processes that decreased genetic variation are interacting with modern conservation challenges to magnify the effects of human disturbance
Lawson (2013)	African spotted reed frog (*Hyperolius substriatus*)	1 mt gene and 2 nuclear genes	The Malawian Highlands contain the most genetic diversity for this species, and hydrological features rather than mountain systems structure lineages
McMenamin & Hadly (2012)	Blotched tiger salamander (*Ambystoma tigrinum melanostictum*)	1 mt locus (>700 bp) for 100- to 3300-year-old samples	Mitochondrial diversity has remained constant over the past 3300 years, indicating that recent fish stocking in Yellowstone National Park has not led to the modern-day patterns of low sequence diversity in mitochondria in the area
Wang (2012)	Yosemite toad (*Bufo canorus*)	10 microsats	Effective population size is correlated with precipitation
Wang *et al*. (2011)	California tiger salamander (*Ambystoma californiense*)	11 microsats	Effective population size is correlated with breeding pond size
Schoville *et al*. (2011)	Mountain yellow-legged frog (*Rana muscosa*)	9 microsats and 2 mt genes	Populations are highly isolated and underwent recent bottlenecks
Nuñez *et al*. (2011)	*Eupsophus calcaratus*	3 mt genes	Some populations underwent demographic, but not range, expansions after the last glaciation. This suggests that some populations were able to persist in refugia within glaciated areas
Phillipsen *et al*. (2011)	*Rana pretiosa*, *Rana luteiventris*, *Rana cascadae* and *Lithobates pipiens*	Between 6 and 11 microsats for the different species	The *Rana* species investigated all have effective population sizes under 50, while *Lithobates pipiens* has an effective population size in the hundreds or thousands. *Lithobates pipiens* effective population size estimates were smaller in the westernmost part of its range
Lind *et al*. (2011)	Foothill yellow-legged frog (*Rana boylii*)	2 mt genes and 1 nuclear gene	Populations are structured by hydrological regions, suggesting that different river basins represent distinct lineages
Phillipsen *et al*. (2010)	Oregon spotted frog (*R. pretiosa*)	7 microsats	Sampling eggs vs. larvae made no difference in measuring effective breeding size of the population, even though egg mass mortality is expected to be high
Canestrelli *et al*. (2008)	Italian stream frog (*Rana italica*)	21 allozymes and 1 mt gene	Populations survived the last glacial cycle in two refugia, one of which underwent subsequent population expansion to recolonize the area postglaciation. Suggests that multiple refugia existed for multiple species in Italian peninsula
Halverson *et al*. (2006)	Wood frog (*Rana sylvatica*)	10 microsats	Inbreeding correlated with lower survival in the wild, but not the

Table 1 *Continued*

Citations	Species	Data	Conclusions
			laboratory. Growth and development were not correlated with inbreeding in the wild or the laboratory
Martínez-Solano *et al.* (2006)	Bosca's newt (*Lissotriton boscai*)	2 mt genes	Populations of Bosca's newts were likely restricted to a collection of refugia during the Plio-Pleistocene, which contributes to modern-day patterns of population structure. Bayesian skyline plots suggest that *L boscai* populations underwent population bottlenecks *circa* 40 000–75 000 years ago
Funk *et al.* (2005)	Columbia spotted frog (*R. luteiventris*)	6 microsats	High-elevation populations have less genetic variation and lower effective population sizes than low-elevation populations

2013), and many of these studies would benefit from paired field-based censuses to better understand the relationship between N_e and census size. However, care must also be taken in the interpretation of molecular population size estimation, particularly when using small numbers of markers or individuals to estimate large populations (Nunziata *et al.* 2015). While it is clear that no two amphibian systems are identical, several conservation-relevant conclusions have emerged in the amphibian literature.

First, effective population sizes are often quite low in amphibians—usually under 100 and frequently closer to 10—and this pattern is heavily influenced by the natural history of many amphibian populations (Schmeller & Merilä 2007; Phillipsen *et al.* 2011; Richmond *et al.* 2014). Because this is a 'natural' trait of some amphibian populations, it is important to understand, on a case-by-case basis, whether effective population size has remained constant through time or if it has recently crashed (Beebee & Rowe 2001; Phillipsen *et al.* 2011). Amphibian populations have a varied history of bottlenecking, with strong evidence of recent population bottlenecks for some populations (Schoville *et al.* 2011). Bayesian skyline estimates of effective population size fluctuations through time have shown the full range of scenarios, with reports of stable or increasing population size for some populations (Nuñez *et al.* 2011) and historical bottlenecks for others (Martínez-Solano *et al.* 2006).

Second, the factors that influence amphibian effective population size are diverse and not well understood. Several studies have attempted to correlate effective population size with factors such as pond size (Wang *et al.* 2011), precipitation (Wang 2012), elevation (Funk *et al.* 2005), fire history (Richmond *et al.* 2013) and glacial movements (Canestrelli *et al.* 2008; Nuñez *et al.* 2011). Of particular interest here is the comparison of population sizes among populations of *Rana pretiosa*, *Rana luteiventris*, *Rana cascadae* and *Rana* (*Lithobates*) *pipiens* performed by Phillipsen *et al.* (2011). While populations of the *Rana* species had similar (small) effective population sizes across their range, *R.* (*L.*) *pipiens* populations exhibited a much higher variance and a longitudinal gradient in N_e. Additionally, both *R. luteiventris* and *R.* (*L.*) *pipiens* (but not *R. pretiosa* or *R. cascadae*) showed a pattern where higher elevation populations had lower effective population sizes (also found in Funk *et al.* 2005). These results indicate that factors influencing effective population size are not broadly generalizable across amphibian taxa, even for the same or related genera.

Finally, phylogeographic studies can be tremendously helpful for putting current patterns of genetic diversity into both historical and conservation contexts. For example, phylogeography can point to historical population size fluctuations and barriers to dispersal that may re-emerge in future climate scenarios. The locations of past refugia as inferred through phylogeographic studies can also be used to verify that a species distribution model performs well in hindcasting (Scoble & Lowe 2010), making it more believable for forecasting. Phylogeographic studies can also be used to understand how diversity is partitioned in modern times (Canestrelli *et al.* 2008; Lind *et al.* 2011; Lawson 2013) and how modern conservation challenges may be a result of historical landscape processes (Dufresnes *et al.* 2013). In a similar vein, studies of ancient DNA allow us to more explicitly measure how genetic diversity has changed through time (McMenamin & Hadly 2012).

Box 4. System 4: Chinese giant salamander (*Andrias davidianus*)

The Chinese giant salamander (*Andrias davidianus*) is the largest amphibian in the world and is listed by the IUCN as critically endangered (Gang *et al.* 2004), by CITES under Appendix I, and by the Chinese government as a Class II protected species under the Law of the People's Republic of China on the Protection of Wildlife. Like all members of the family Cryptobranchidae, *Andrias* are obligatorily aquatic and are restricted to large, permanent riverine habitats. The species faces a host of threats, ranging from habitat loss and poaching (Wang *et al.* 2004) to disease (Geng *et al.* 2011). Wang *et al.* (2004) evaluated *A. davidianus* habitat, conducted population surveys and polled wildlife managers and villagers in 13 sites throughout five different provinces and the city of Chongqing. They found that while salamanders were common several decades ago, they are now rare and usually smaller when captured, and that the price of Chinese giant salamander meat has risen considerably since 1980. Further, while more than 350 000 hectares have been set aside for nature reserves to conserve *A. davidianus*, the authors could find no salamanders in the six reserves they visited (despite spending 4 months looking), and documented a lack of funding and poor protection for salamanders within reserves.

Despite the significant threats facing the species and recent dramatic declines, the molecular ecology of the Chinese giant salamander is still relatively poorly understood. Microsatellite loci (Yoshikawa *et al.* 2011; Meng *et al.* 2012) and a full mitochondrial genome (Zhang *et al.* 2003) have been developed for the species, but thus far population genetic work on Chinese giant salamanders has mainly focused on geographic patterns of variation in proteins and single mitochondrial genes. One study examined 40 allozymes and two mitochondrial genes in 19 individuals to assess variation and population structure among nine locations across much of the current range of the species (Murphy *et al.* 2000). The authors found evidence for population substructuring among their sampling locations, but after building phylogenetic trees with their data, they did not uncover monophyletic groupings that reflected major river systems (Pearl, Yellow and Yangtze). Murphy *et al.* (2000) asserted this was likely due to an extensive history of human transplantations dating back over 2300 years.

To further explore these rangewide patterns of genetic structuring, Tao *et al.* (2005) sequenced 771 bp from the mitochondrial D-loop of 28 Chinese giant salamanders from four different sites (again associated with either the Pearl River, Yellow River or Yangtze River). They found significant (but very low—$F_{ST} < 0.01$) levels of population differentiation between the Pearl River and the Yellow/Yangtze Rivers, but not between the Yellow and Yangtze Rivers. An AMOVA further revealed that among-group comparisons were responsible for <1% of the genetic variation, supporting the notion that modern-day populations are not strongly structured with respect to river systems.

Chinese giant salamanders are kept in captivity in China for both conservation and exploitative purposes (Wang *et al.* 2004). Recently several captive salamander colonies have experienced significant die offs due to disease (Dong *et al.* 2010). After ruling out bacterial infections as the cause of the deaths, researchers observed high numbers of viral particles in different cell types using electron microscopy, which, combined with the symptoms they observed in the infected salamanders, led them to believe it was a ranavirus (Dong *et al.* 2010). After sequencing several genes from the virus, researchers determined that it was closely related to several other known ranaviruses (Dong *et al.* 2010; Geng *et al.* 2011; Zhou *et al.* 2013). Given the high mortality associated with this virus in *A. davidianus* (and other ranaviruses in amphibian species across the world), the spread of this disease to wild populations presents a major risk to the persistence of the world's largest amphibian, and immediate action to prevent the movement of ranavirus is critical.

Andrias davidianus. Photo: Theodore Papenfuss

Box 5. System 5: Cane toad (*Rhinella marina*)

While the cane toad itself is not a threatened species, it is an ecologically devastating invasive in many regions, particularly in Australia where it was first introduced in 1935 (Shine 2010). Because cane toads are highly toxic, they have led to extensive declines in the native amphibian predator community, including native amphibians that die after naively eating them or their eggs (Crossland *et al.* 2008; Shine 2010). Considerable genetic research has been conducted on *Rhinella marina* populations in many parts of the world to better understand their ecology, predict their future impacts, and devise strategies to control their spread.

Understanding the history of cane toads in their native habitats and their natural biological control should help decision-makers limit their impacts across their invasive range. Slade & Moritz (1998) investigated the rangewide phylogeography (from both native and introduced locations) of cane toads using 468 bp of mitochondrial DNA, confirming that introduced populations in Australia and Hawaii are most closely related to the toads from the eastern parts of Venezuela and French Guiana. They also discovered a high degree of divergence between populations on the east and west sides of the Andes. This led them to suggest that research into pathogens for controlling cane toads in Australia should include the western Venezuelan lineage, as those pathogens would likely be novel, and therefore potentially more damaging, to Australian populations.

Cane toads have also provided an interesting system to study the genetics of a rapidly expanding invasive amphibian species. Estoup *et al.* (2001) compared the patterns of population bottlenecks in Australian cane toad populations derived from allozyme and microsatellite markers, with microsatellites showing greater evidence of population bottlenecks associated with the introduction of founder individuals. Later, they analysed 10 microsatellites and found that established populations exchange migrants each generation and that migration between populations is higher in the north than in the east (Estoup *et al.* 2004). The model-based approach in that study supported a general pattern of isolation by distance. This is in contrast to earlier work by Leblois *et al.* (2000), which found no evidence of isolation by distance among 120 continuously distributed cane toads in Byron Bay, Australia. Fully understanding the spatial scale at which isolation by distance operates will provide important information regarding the distance and frequency of migration events in the system. More recently, Estoup *et al.* (2010) combined historical occurrence reports with microsatellite data to infer founder effect-induced bottlenecks in cane toads in Australia using an approximate Bayesian computation technique. There they also determined that the first introduction of cane toads into northern Australia likely occurred at Lagoon Creek in 1979.

Management of cane toads in Australia would benefit from understanding whether particular alleles are enabling population expansion and persistence, and whether there are genetically detectable vulnerabilities in some populations. If certain populations are found to be more genetically robust than others, then eradication efforts could benefit from focusing on those populations first. One important such effort by Abramyan *et al.* (2009) found that cane toads possessed Z and W chromosomes responsible for sex determination—opening the door to using genetic modifications that shift the sex balance of wild populations (Gutierrez & Teem 2006). While the functional genomics of *Rhinella marina* has not been extensively studied, Lillie *et al.* (2014) sequenced genes from the Major Histocompatibility Complex in cane toads and found low levels of variation in MHC class I genes. The authors suggested that the toads may be susceptible to a biological control strategy (although this is highly controversial and has thus far been unsuccessful, see Shanmuganathan *et al.* 2010).

Additionally, an assembled transcriptome of *Rhinella marina* has recently become available, which will provide a framework on which researchers can design studies of functional differentiation between cane toad populations (Genomic Resources Development Consortium *et al.* 2015). One important recent advance found several differences

Genes on landscapes

Methods and their application to amphibians

Molecular genetic analyses are extremely useful for determining the limits of populations and how members are exchanged between these populations. Early approaches relied on methods that required minimal genetic information and computing power. However, recent advances in both of these areas have allowed researchers to develop new model-based approaches that are computationally intensive and data-hungry, but that avoid some of the problematic assumptions that plagued previous methods while also allowing one to estimate additional parameters. Broadly, this area can be divided into two subcategories: population structure and migration analysis.

The traditional, and still most common, approach to measuring population structure is with genetic fixation indices such as F_{ST} (Wright 1951). F_{ST} uses allele fre-

Box 5. *Continued*

in gene expression between toads at the expanding range edge in Australia compared to those at the range core (Rollins *et al.* 2015). Cane toads at the range edge have been shown to disperse farther (Lindström *et al.* 2013) and more linearly (Brown *et al.* 2014) than those in the core of the range, and although these traits were determined to be heritable, the genetic loci responsible are not well understood. Rollins *et al.* (2015) sequenced transcriptomes from toads at the range edge and range core and found hundred of genes differentially expressed between the two groups. Many of these differentially expressed genes were involved with metabolism and cell repair, with the largest differences in energy-producing pathways, suggesting that these genes may play an important role in dispersal ability.

Rhinella marina. Photo: Arthur Georges

quencies (which were readily obtainable from early allozyme data) to measure the reduction of heterozygosity due to genetic drift operating on subdivided populations, and is often used as a measure of genetic differentiation between populations of a species (Wright 1951). Additionally, analogues have been proposed which extend this approach to apply to mitochondrial sequences (Takahata & Palumbi 1985) and also to incorporate the number of mutations between haplotypes (Excoffier *et al.* 1992).

Fixation indices require that researchers make a priori hypotheses of which individuals should be grouped together as (sub)populations. For a given group of populations, there may be several arrangements of individuals that yield significantly nonzero population differentiation, and it can be difficult to determine the optimal grouping. A different approach to determining population structuring that addresses this shortcoming has gained popularity as genetic data have become more available and computing power has increased. These methods are collectively referred to as Bayesian clustering or individual-based methods (Excoffier & Heckel 2006). In general, these methods model a finite number of source populations with their own respective allele frequencies (Pritchard *et al.* 2000), assigning individuals proportionally to these source populations. While these methods usually do not require researchers to make a priori assignments of individuals to populations, they typically require a priori establishment of the number of source populations.

While information regarding population structure is useful for conservation, it is often even more informative to evaluate these hypothesized populations in the context of realized movements through the landscape. Most populations (and individuals) in nature exhibit a spatial genetic pattern known as isolation by distance, whereby populations further from one another in physical distance are more genetically different than closer populations (Wright 1943). Because terrestrial landscapes are often heterogeneous, animals rarely migrate in straight lines between populations. Rather, movement between populations is often circuitous and shaped by underlying features of the landscape (Coulon *et al.* 2004; Krivoruchko & Gribov 2004), and this may be particularly important for small animals subject to desiccation like amphibians.

Determination of realistic distances between populations is a key step in studies relying on the assumptions of isolation by distance. The simplest measurement between two populations is straight-line Euclidean distance (or great circle distance on the earth), and this can serve as a spatially explicit null hypothesis. Alternative approaches should consider the makeup of the landscape to better approximate physical distances between

populations. For instance, in a riparian or riverine species, it could make sense to measure the physical distance along rivers to other populations (McCartney-Melstad *et al.* 2012; Unger *et al.* 2013). More generally, areas in the landscape that are relatively more or less hospitable to a particular species can be categorized as minimally or highly resistive to movement, respectively (Zeller *et al.* 2012). This categorization can be used to build resistance maps for a landscape and species of interest.

Resistance surface maps can be used to infer how different amphibian species move through the landscape. Traditionally, this has been done through a least-cost path analysis using GIS (Adriaensen *et al.* 2003). This framework designates a high cost to habitat patches (pixels) deemed highly resistive to movement, and a relatively lower cost for habitat that is less resistive. Under least-cost path analysis, the most likely route of movement between populations is the route that minimizes the additive costs of traversing particular pathways between populations.

Although least-cost path modelling is still widely used, a new approach that borrows from electrical circuit theory has recently gained in popularity due to its relaxed assumptions regarding consistency of migration routes (McRae & Beier 2007; McRae *et al.* 2008). The circuit theory approach considers landscape resistance as analogous to electrical resistance, and it thus assesses the contribution of multiple routes of dispersal between different areas. Effective distances between populations here are represented as pairwise resistances. After building resistance landscapes and estimating effective distances between populations, the next logical step is to determine which physical distance best correlates with the observed genetic distances. Given the principle of isolation by distance, the physical distance that best correlates with genetic distance represents the best available hypothesis for how a species moves through the landscape. Mantel tests are often used to investigate correlations between genetic and geographic distance matrices (Mantel 1967) and to determine correlations between genetic and environmental variation; recently, the Mantel approach has been criticized as inappropriate (Guillot & Rousset 2013), and newer methods are currently being developed to address this deficiency (Bradburd *et al.* 2013).

Amphibian conservation

Population structure is essential for conservation planning. Management decisions under the US Endangered Species Act (ESA) can hinge on these analyses, as the ESA contains protections for vertebrate 'distinct population segments' (Waples 1991). Analyses of population structure may be used to determine those groups of individuals that represent distinct population segments, and therefore can be instrumental in setting protection priorities. Additionally, knowledge of how a species traverses its landscape can and should be a critical component of conservation planning (Segelbacher *et al.* 2010). The studies in Table 2 all use molecular data to draw conclusions about how amphibian populations are related to one another and how they interact with their surrounding landscape, both of which can be immediately useful in conservation planning. These studies test a wide range of taxon-specific hypotheses, and some general themes appear to be emerging.

Physical and ecological characteristics of amphibians contribute to the scales at which geneticists can ask questions regarding their population structure. Amphibians often exhibit breeding site philopatry and limited dispersal capabilities (Smith & Green 2005). This means that there are often low levels of migration between populations, and a correspondingly high degree of population structuring at fine scales (Beebee 2005). For instance, Jehle *et al.* (2005a) studied great crested newts and found detectable population structure among ponds, many of which were separated by <1 km (see Box 2). Similarly, Savage *et al.* (2010) found an extremely high amount of spatial genetic structure between nearby ponds in the southern long-toed salamander in California, as did Wang *et al.* (2009b) for endangered California tiger salamanders (CTSs); these latter two analyses may suggest that relatively arid habitats promote limited dispersal among breeding sites.

One recent trend in the landscape genetics literature is to test the generality of patterns by conducting experiments on multiple species or spatial scales. A key finding of several of the studies in Table 2 is the different patterns seen over diverse scales and sampling regimes. For instance, researchers have found differing effects of landscapes on amphibians by looking at different regions inhabited by a single species (Johansson *et al.* 2005; Wang *et al.* 2009a, 2011; Moore *et al.* 2011; Trumbo *et al.* 2013), different species within the same region (Goldberg & Waits 2010; Richardson 2012), different time periods for the same metapopulation (Savage *et al.* 2010) and different spatial scales for the same analysis (Angelone *et al.* 2011). These results collectively challenge the generality of results that stem from landscape genetic studies, and suggest that results from molecular studies of amphibian–landscape interactions should generally be interpreted within the scope of a specific study or region, but not beyond.

Another important contribution of this field is the investigation of anthropogenic effects on amphibians. While the impacts of activities such as hunting and road mortality may sometimes be directly measurable,

Table 2 Key recent amphibian studies exploring genes on landscapes

Citations	Species	Data	Conclusions
Trumbo *et al.* (2013)	Cope's giant salamander (*Dicamptodon copei*)	11 microsats	Landscape variables have different effects on the species in different parts of its range
Peterman *et al.* (2013)	Wood frog (*Rana sylvatica*)	11 microsats	Genetic diversity decreases when approaching the edge of a the species range
Igawa *et al.* (2013)	*Odorrana ishikawae* and *Odorrana splendida*	12 microsats	Topography is a key driver of population structure in these species
Aguilar *et al.* (2013)	Coastal tailed frog (*Ascaphus truei*)	9 microsats	Gene flow is positively correlated with moisture-related variables. Populations do not show signatures of bottlenecks despite heavy logging in the region
Richardson (2012)	Spotted salamander (*Ambystoma maculatum*) and wood frog (*Rana sylvatica*)	14 microsats	Landscape variables have drastically different effects on spotted salamanders than on wood frogs
Moore *et al.* (2011)	Boreal toad (*Bufo boreas*)	11 microsats	Saltwater and landscape cover influence gene flow for boreal toads in one region of southeast Alaska, but not in another
Angelone *et al.* (2011)	European tree frog (*Hyla arborea*)	11 microsats	Short-range dispersal is only negatively impacted by the presence of rivers, while longer-range dispersal is mediated by geographic distance and a more diverse set of landscape variables
Savage *et al.* (2010)	Southern long-toed salamander (*Ambystoma macrodactylum sigillatum*)	18 microsats	Individuals rarely traverse the harsh landscape between breeding ponds, and cohorts from the same ponds in different years show a surprisingly high degree of differentiation between each other
Murphy *et al.* (2010a)	Columbia spotted frogs (*Rana luteiventris*)	8 microsats	Gene flow between populations is influenced by topography, predation by fish and length of growing season
Murphy *et al.* (2010b)	*Bufo boreas*	15 microsats	Random forests analyses uncovered several factors that impacted population connectivity for boreal toads in Yellowstone National Park, including precipitation, large fires and roads. Additionally, different ecological processes were found to influence connectivity at different scales
Goldberg & Waits (2010)	*R. luteiventris* and *Ambystoma macrodactylum*	8 microsats	These two species from the same region in northern Idaho differ in the land cover type that is least resistive to movement
Zellmer & Knowles (2009)	Wood frog (*Rana sylvatica*)	9 microsats	The observed patterns of population differentiation among populations of wood frogs (*Rana sylvatica*) are mainly the result of habitat structure post-1850s
Wang *et al.* (2009a)	California tiger salamander (CTS) (*Ambystoma californiense*)	13 microsats	Chaparral is more conducive to movement through the landscape for CTSs (*Ambystoma californiense*) than are grasslands or oak forest
Funk *et al.* (2009)	*Physalaemus petersi* and *Physalaemus freibergi*	9 microsats	Genetic distance between populations is not correlated with major landscape variables, but was positively correlated with the dominant whine frequency of male advertisement calls
Telles *et al.* (2007)	*Physalaemus cuvieri*	9 RAPD markers	Presence of humans hinders movement between populations
Giordano *et al.* (2007)	Long-toed salamander (*Ambystoma macrodactylum*)	7 microsats	Gene flow is higher among low-elevation populations than among high-elevation

Table 2 *Continued*

Citations	Species	Data	Conclusions
			populations, and little migration occurs between low- and high-elevation populations
Spear *et al.* (2005)	Blotched tiger salamander (*Ambystoma tigrinum melanostictum*)	8 microsats	Open shrub habitat and river crossings facilitate migration between populations, while elevation differences decrease gene flow
Johansson *et al.* (2005)	Common frog (*Rana temporaria*)	7 microsats	Agricultural fragmentation has markedly different impacts on the species in different parts of its range
Funk *et al.* (2005)	Columbia spotted frog (*R. luteiventris*)	6 microsats	There is a high degree of gene flow among low-elevation populations, and elevation changes serve as barriers to gene flow

human effects on genetic variation and connectivity between populations due to habitat destruction and fragmentation may be more difficult to assess. Studies in molecular ecology can evaluate the impacts of human landscape modification on amphibian population structuring (Johansson *et al.* 2005; Aguilar *et al.* 2013) or even treat the presence of humans themselves as a landscape variable that might explain patterns of genetic variation. For instance, Telles *et al.* (2007) analysed nine RAPD markers in 18 Brazilian populations of *Physalaemus cuvieri* to test whether the presence of humans between populations correlated with reduced gene flow and higher divergence. They found that fixation indices were not significantly correlated with geographic distance, but that the date of establishment of municipalities throughout the region helped to explain the broad-scale patterns of genetic differentiation better than distance alone.

Genes that matter

Methods and their application to amphibians

Biologists have been interested in the function of specific genes in the fate of individuals and species since the discovery of discrete units of inheritance. Bridging the gap in understanding how genes and the environment interact to form organisms is a major goal of modern biology, and a great deal of work has been done using model organisms (including the amphibians *Xenopus laevis*, *Xenopus tropicalis* and *Ambystoma mexicanum*) towards this end. As conservationists, we are particularly interested in how specific genes and proteins may help populations and species adapt to their current and future environments, allowing us to make more informed decisions regarding translocations and better predict future population trajectories. With this in mind, biologists have increasingly been studying the importance of specific genes in wild amphibian populations using both candidate gene and genomewide strategies.

Candidate gene approaches involve using a priori knowledge of gene functions (often deduced from a model organism) to study gene function (Tabor *et al.* 2002). Candidate genes may show extreme population substructure indicative of local selection, contain specific protein-changing mutations, or exhibit exceptional rates of molecular evolution. These lines of evidence, along with a priori information regarding the function of the candidate gene in another species, can help researchers untangle the potential role of a gene in local adaptation.

In contrast to candidate gene approaches are whole-genome-scale studies. Full (or nearly full) genome data are generally gathered for an individual by extracting DNA from a tissue sample, sequencing it on high-throughput sequencers, and aligning the resulting DNA sequence data to a reference genome. This is currently untenable for most amphibians because of their large genome sizes and the lack of suitable assembled amphibian reference genomes. Several intermediate technologies are being used in amphibians to obtain information for hundreds or thousands of distinct loci across the genome, including restriction-site-associated DNA (RAD) sequencing (Baird *et al.* 2008; Streicher *et al.* 2014) and target-capture sequencing (Mamanova *et al.* 2010; Hedtke *et al.* 2013). RAD sequencing generates data for anonymous sites in the genome that flank restriction enzyme recognition sites, and as such generally does not fall within protein-coding regions. Target-capture sequencing, on the other hand, selectively enriches pools of genomic DNA for specific targets prior to sequencing. One source of targets for such amphibian studies is exons derived from existing EST or transcriptome resources (Bi *et al.* 2012; McCartney-Melstad *et al.* 2015). Several amphibian transcriptomes

have recently been sequenced and assembled, including members of three salamander families [*Hynobius chinensis* (Che *et al.* 2014), *A. mexicanum* (Smith *et al.* 2005), *Notophthalmus viridescens* (Looso *et al.* 2013)] and three anuran families [*X. tropicalis* (Hellsten *et al.* 2010; Tan *et al.* 2013), *X. laevis* (Morin *et al.* 2006), *Pseudacris regilla*/*Rana* (*Lithobates*) *clamitans* (Robertson & Cornman 2014), *Rana chensinensis*/*Rana kukunoris* (Yang *et al.* 2012), *Rana muscosa*/*Rana sierrae* (Rosenblum *et al.* 2012), *Odorrana margaretae* (Qiao *et al.* 2013), *Espadarana prosoblepon*/*Rana* (*Lithobates*) *yavapaiensis* (Savage *et al.* 2014)]. Several of these taxa are endangered in the wild, and the availability of these transcriptome resources will allow researchers to design target-capture arrays using DNA sequence data for a large variety of amphibian species.

Another way to analyse variation in protein-coding genes is by sequencing the gene products themselves. Analyses of the full complement of transcribed gene products (the transcriptome) generally rely on extracting mRNA from tissue samples, converting it to more stable cDNA, then sequencing that cDNA on high-throughput sequencers (Bainbridge *et al.* 2006; Wang *et al.* 2009b). The resulting sequence reads are then assembled to reconstruct full-length transcripts or aligned to pre-existing transcriptome assemblies. An added benefit of this sequencing approach is that the relative production of different gene products can be computed based on the relative abundance of sequence reads aligning to a transcript (Anders & Huber 2010). Such analyses can yield information regarding both the level of variation within transcribed gene products and the up- or downregulation of particular genes between different groups of individuals (for instance, one can test for differences in cellular responses for animals exposed to different environmental challenges).

Amphibian conservation

Because amphibians exhibit such a wide array of responses to the challenges they face, it is important to understand how genetic factors may be directly contributing to the way amphibians interact with the environment. Evidence of local adaptation and rapid responses to human habitat modification may be the key for successful relocation plans and can also help the conservation community predict how amphibian populations and species may react to climate change. Some studies of local adaptation have explicitly tried to uncover the genes responsible (Bonin *et al.* 2006; Richter-Boix *et al.* 2011, 2013), while others have used principles of quantitative genetics to identify a genetic effect without identifying the actual genes (Merilä *et al.* 2004; Ficetola & Bernardi 2005; O'Neill & Beard 2010; Brady 2012, See Table 3).

Candidate gene approaches have been used in several amphibian systems to explore local adaptation and genotype–phenotype associations. For instance, Voss *et al.* (2000) investigated the thyroid hormone receptor (TR) locus and found that it was not responsible for determining life cycle strategies (to metamorphose or not to metamorphose) in *A. mexicanum*/*Ambystoma tigrinum* crossing families. A later study, however, found that the TR locus may play a role in the timing of metamorphosis for ambystomatid salamanders that do metamorphose (Voss *et al.* 2003). Understanding the basis for differences in timing of metamorphosis may play a critical role in the conservation of the CTS, especially given California's recent long-term droughts which influence vernal pool hydroperiods (see Box 1). Other candidate gene studies have found evidence that a TR gene is contributing to developmental timing in the moor frog *Rana arvalis* and may be acting to structure populations in the landscape (Richter-Boix *et al.* 2011, 2013).

Genomic-scale studies targeting genes that may be important for conservation in amphibians are less common at this time, perhaps because of the challenges posed by large genome amphibians. However, one study by Yang *et al.* (2012) sequenced the transcriptomes of two closely related frogs: *R. kukunoris*, from the high-elevation Tibetan Plateau; and *R. chensinensis*, a wide ranging low-elevation species. They compared ratios of nonsynonymous to synonymous mutations across the transcriptomes and found evidence that 14 genes may be responsible for adaptation to high elevations in *R. kukunoris*. This information helps us both predict how certain populations might adapt as climate change forces them to move upslope and identify populations that could be best suited for translocations to different elevations.

One area that has been particularly active in the investigation of amphibian conservation-relevant gene function has been characterizing cellular responses to infection with the pathogenic chytrid fungus *Batrachochytrium dendrobatidis* (*Bd*). *Bd* causes the deadly disease chytridiomycosis in many amphibians, especially high-elevation frogs (Berger *et al.* 1998; Daszak *et al.* 1999). Chytridiomycosis modifies the skin of frogs, affecting their ability to transport electrolytes across the skin and leading to depletion of key elements such as potassium, which can lead to death (Voyles *et al.* 2009). This disease has been implicated in the declines of hundreds of species around the world, and its mechanism of attacking an organ that is uniquely important to amphibians may explain why it is so deadly to a diverse set of amphibians (Skerratt *et al.* 2007; Voyles *et al.* 2009).

Table 3 Key recent amphibian studies exploring genes that matter

Citations	Species	Data	Conclusions
Savage *et al.* (2014)	*Espadarana prosoblepon*, *Rana* (*Lithobates*) *yavapaiensis*	De novo assembled transcriptomes from both species (14 309 total gene products)	Acquired immunity- and inflammation-associated genes showed greater divergence between the two species than innate immunity genes. Immune-related genes did not generally show an increased rate of evolution, except for glycosyl proteases, which may be involved with bacterial and fungal defences
Richter-Boix *et al.* (2013)	*Rana arvalis*	6 microsatellites and sequence data from a TR gene	Variation in the TR gene is associated with developmental timing and growth rate, and gene flow appears to be structured by environmental factors and breeding time (not distance)
Yang *et al.* (2012)	*Rana chensinensis*, *Rana kukunoris*	De novo assembled transcriptomes from each species (81 151 total putative transcripts)	Identified 14 genes that were likely involved with the adaptation of *R. kukunoris* to high elevation
Rosenblum *et al.* (2012)	*Rana muscosa* and *Rana sierrae*	De novo assembled transcriptomes were used to design a microarray from 28 995 putative transcripts, which was used to quantify expression in infected vs. control frogs	Both *R. muscosa* and *R. sierrae* showed very limited changes in immune-related gene transcription in response to infection with *Bd*, but did find a collection of genes that responded after infection (especially in the skin)
Savage & Zamudio (2011)	*R.* (*L.*) *yavapaiensis*	14 microsats (to establish population structure) and a 246-bp region of an MHC class IIB exon	Variability in the MHC gene exon was significantly associated with lower death rates upon being exposed to *Bd*
Richter-Boix *et al.* (2011)	*R. arvalis*	15 microsats	Found evidence of directional selection in a microsatellite marker located within a thyroid hormone receptor. This receptor is involved with metamorphosis and its pattern of variation is correlated with habitat characteristics, suggesting it is involved with adaptation of populations to their local environment
Bonin *et al.* (2006)	*Rana temporaria*	392 AFLPs	Used outlier analyses to identify a list of eight AFLPs possibly involved with (or linked to other loci involved with) adaptation to high altitudes

TR, thyroid hormone receptor.

Over the past several decades, researchers have identified a collection of genes that are involved in the immune response in model organisms, some of which are particularly useful in immune responses to fungi (Hoffmann 2003). Building off of this background, Rosenblum *et al.* characterized differences in gene expression levels for these genes (and many others) in healthy and *Bd*-infected *X. tropicalis* (Rosenblum *et al.*

2009) and ranid frogs (Rosenblum *et al.* 2012). They found surprisingly little evidence of an immune response in infected frogs—a shared pattern among susceptible species. Savage *et al.* (2014) later characterized the transcriptomes of two frogs to compare rates of evolution across different immune-associated genes. They found that these genes did not generally have a higher rate of functional evolution, with the exception of lysozyme-encoding glycosyl proteases, a key component of an organism's first response against fungal and bacterial pathogens (Savage *et al.* 2014). In addition to studying the cellular responses of amphibians to *Bd* infection, gene activity of *Bd* itself has also been characterized in an attempt to better understand how the fungus behaves in different stages of its life cycle (Rosenblum *et al.* 2008).

Although still in its infancy, genomics is beginning to inform amphibian conservation science, particularly in meeting the growing challenges of disease and climate change. Given that human interventions will almost certainly be necessary to avoid the global extinction of many species, we view the characterization of genes that matter as a high priority for the field.

Future directions and critical knowledge gaps

One clear conclusion from this review is that even recent amphibian molecular ecology studies still rely on a small number of mitochondrial and nuclear loci. Understanding the current and past effective population sizes for different populations on the landscape is a key part of effective conservation, but the error around N_e estimates depends on the number of independent loci being analysed (Heled & Drummond 2008). Likewise, the detectability of genetic differentiation between populations scales on the order of one divided by the square root of the product of the number of loci and the number of individuals genotyped (Patterson *et al.* 2006), suggesting that low values of F_{ST} are only detectable with very large numbers of individuals and/or markers. Additionally, scans for genes that are important for local adaptation benefit from information on most, if not all, genes in the genome.

Lack of existing genomic resources and large genome sizes have thus far limited genome-scale approaches to conservation genetics in most amphibians. Some amphibians, particularly salamanders, have incredibly large genomes with many over 30 gigabases and some ranging up to 120 gigabases—nearly 40 times larger than the human genome (Olmo 1973). Largely as a consequence of their genome size, only two amphibian genomes are currently sequenced and assembled—the aquatic frog *Xenopus tropicalis* (Hellsten *et al.* 2010) and the Tibetan frog *Nanorana parkeri* (Sun *et al.* 2015). This presents challenges for incorporating many modern DNA sequencing and analysis strategies that require a well-assembled reference genome for read mapping or some knowledge of where loci are located in the genome to quantify linkage among loci. This situation is changing, and chromosome-specific genomic builds should soon be available for the axolotl *Ambystoma mexicanum* (R. Voss, personal communication).

Although full-genome assemblies are still essentially nonexistent for amphibians, several amphibian transcriptomes have recently been published, and more will certainly be forthcoming. This recent influx of genome-wide information gives us the beginnings of a comparative framework that is desperately needed to model how different amphibian species interact with their environments at the molecular level. Research is already underway using these transcriptomic resources to build exon capture arrays to assay variation in coding regions across the genome for large numbers of individuals (Hedtke *et al.* 2013; McCartney-Melstad *et al.* 2015). These resources will help plug the amphibian genomics gap, but fully assembled genomes from representative lineages remain essential.

Aside from the general lack of genomic-scale studies in amphibian molecular ecology, several other areas also are underexplored. Caecilians (Gymnophiona), an evolutionarily and morphologically distinct lineage of amphibians, are particularly understudied (but see Zardoya & Meyer 2001; San Mauro *et al.* 2004; Zhang & Wake 2009; Nishikawa *et al.* 2012; Stoelting *et al.* 2014); given their cryptic subterranean lifestyle, there is much to be learned from molecular ecological studies of these elusive tropical amphibians. Likewise, comparatively little molecular ecology work has been conducted on direct-developing amphibians. Of the 552 studies analysed by Emel & Storfer (2012) between 2001 and 2010, only 18 involved direct-developing amphibian species. Given their independence from free-standing water (direct-developing species lack a free-living larval stage), these animals have fundamentally different breeding and dispersal requirements than pond-breeding amphibians (Duellman & Trueb 1985), and a deeper understanding of their molecular ecology is required before we can comfortably apply what we learn from aquatic-breeding species to conservation plans for direct-developing amphibians.

Another critical future direction is building our understanding of how chytrid fungi are affecting amphibian populations, including the mechanisms of resistance in wild populations. Particularly troubling is the recent discovery of *Batrachochytrium salamandrivorans* (*Bs*), a salamander-specific analogue to *Bd*, which is now established in Europe (Martel *et al.* 2013, 2014). The historical and functional relationships between *Bd* and *Bs*, the gene variants and regulatory responses to

these pathogens in susceptible and resistant populations, and the microbial communities on the skin of amphibians involved in response to chytridiomycosis are all critical lines of future research.

Conclusions

For molecular ecologists

These are exciting times for research in molecular ecology. High-throughput sequencing has become more accessible and allows for studies that investigate orders of magnitude more individuals and loci than was previously possible for the same cost and effort. Genomic approaches and resources for amphibians are beginning to become available, including RAD-tag sequencing (Streicher *et al.* 2014) and transcriptome-based exon capture arrays (Hedtke *et al.* 2013; McCartney-Melstad *et al.* 2015), with additional full-genome assemblies likely coming soon. New analytical methods are also being designed to take advantage of the large amounts of data these genomic studies produce to provide deeper insights into important questions in molecular ecology (Alexander *et al.* 2009; Fumagalli *et al.* 2014; Korneliussen *et al.* 2014). The challenge to amphibian conservationists is to make the best of these emerging resources to generate deep and meaningful understandings of amphibian populations in nature. Forming collaborations with relevant management agencies is a key area for molecular ecology researchers and will help identify which questions are most pressing and which solutions are most needed. A further challenge is to clearly articulate how genomic technologies can and should benefit studies in molecular ecology as the technology develops.

For conservation ecologists

Genetic and genomic approaches can complement many ecological studies, especially in amphibians. Particularly interesting to ecological insights for conservation biology may be transcriptome-enabled, gene-specific analyses that help us understand, at the most basic level, why certain populations are thriving while others decline. Further work that reconciles and utilizes field observations and data with analyses of genetic variation (as in Estoup *et al.* 2010) will be especially useful in moving both fields forward.

For managers

Wildlife managers benefit from incorporating genetic research into amphibian management plans. The state of what constitutes the best possible, as opposed to the best available, scientific technologies and methods changes constantly, and will continue to for the foreseeable future. Academic scientists and wildlife managers both benefit from collaborations that foster a clear understanding of the scope of inference possible with different genetic methods, and how they might be applied to the most important questions in conservation.

Acknowledgements

The authors would like to thank our colleagues at the US Fish and Wildlife Service and California Department of Fish and Wildlife for long-standing partnerships in genetics-based amphibian conservation. The authors would also like to thank Louis Bernatchez for helping to conceive of this review and for gentle encouragement, and three anonymous reviewers for helpful insights and suggestions. The authors were supported with funding provided by NSF-DEB 0817042, DEB 1239961 and DEB 1257648.

References

Abramyan J, Ezaz T, Graves JAM, Koopman P (2009) Z and W sex chromosomes in the cane toad (*Bufo marinus*). *Chromosome Research*, **17**, 1015–1024.

Adriaensen F, Chardon JP, De Blust G *et al.* (2003) The application of "least-cost" modelling as a functional landscape model. *Landscape and Urban Planning*, **64**, 233–247.

Aguilar A, Douglas RB, Gordon E, Baumsteiger J, Goldsworthy MO (2013) Elevated genetic structure in the coastal tailed frog (*Ascaphus truei*) in managed redwood forests. *Journal of Heredity*, **104**, 202–216.

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

Allentoft ME, O'Brien J (2010) Global amphibian declines, loss of genetic diversity and fitness: a review. *Diversity*, **2**, 47–71.

Anders S, Huber W (2010) Differential expression analysis for sequence count data. *Genome Biology*, **11**, R106.

Angelone S, Kienast F, Holderegger R (2011) Where movement happens: scale-dependent landscape effects on genetic differentiation in the European tree frog. *Ecography*, **34**, 714–722.

Arnason AN (1973) The estimation of population size, migration rates and survival in a stratified population. *Researches on Population Ecology*, **15**, 1–8.

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.

Backlin AR, Hitchcock CJ, Gallegos EA, Yee JL, Fisher RN (2013) The precarious persistence of the endangered Sierra Madre yellow-legged frog *Rana muscosa* in southern California, USA. *Oryx*, **49**, 157–164.

Bainbridge MN, Warren RL, Hirst M *et al.* (2006) Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. *BMC Genomics*, **7**, 246.

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

Barry SJ, Shaffer HB (1994) The status of the California tiger salamander (*Ambystoma californiense*) at Lagunita: a 50-year update. *Journal of Herpetology*, **28**, 159–164.

Beebee TJC (2005) Conservation genetics of amphibians. *Heredity*, **95**, 423–427.

Beebee TJ, Griffiths RA (2005) The amphibian decline crisis: a watershed for conservation biology? *Biological Conservation*, **125**, 271–285.

Beebee T, Rowe G (2001) Application of genetic bottleneck testing to the investigation of amphibian declines: a case study with natterjack toads. *Conservation Biology*, **15**, 266–270.

Beerli P, Felsenstein J (2001) Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *Proceedings of the National Academy of Sciences of the USA*, **98**, 4563–4568.

Berger L, Speare R, Daszak P *et al.* (1998) Chytridiomycosis causes amphibian mortality associated with population declines in the rain forests of Australia and Central America. *Proceedings of the National Academy of Sciences of the USA*, **95**, 9031–9036.

Bi K, Vanderpool D, Singhal S *et al.* (2012) Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. *BMC Genomics*, **13**, 403.

Blair C, Davy CM, Ngo A *et al.* (2013) Genealogy and demographic history of a widespread amphibian throughout Indochina. *Journal of Heredity*, **104**, 72–85.

Bloomquist EW, Lemey P, Suchard MA (2010) Three roads diverged? Routes to phylogeographic inference. *Trends in Ecology & Evolution*, **25**, 626–632.

Bolster BC (2010) A status review of the California tiger salamander (*Ambystoma californiense*). *State of California Natural Resources Agency Department of Fish and Game. Report to the Fish and Game Commission*.

Bonin A, Taberlet P, Miaud C, Pompanon F (2006) Explorative genome scan to detect candidate loci for adaptation along a gradient of altitude in the common frog (*Rana temporaria*). *Molecular Biology and Evolution*, **23**, 773–783.

Bosch J, Martínez-Solano I, García-París M (2001) Evidence of a chytrid fungus infection involved in the decline of the common midwife toad (*Alytes obstetricans*) in protected areas of central Spain. *Biological Conservation*, **97**, 331–337.

Bradburd GS, Ralph PL, Coop GM (2013) Disentangling the effects of geographic and ecological isolation on genetic differentiation. *Evolution*, **67**, 3258–3273.

Brady SP (2012) Road to evolution? Local adaptation to road adjacency in an amphibian (*Ambystoma maculatum*). *Scientific Reports*, **2**, 235.

Brede EG, Beebee TJC (2006) Large variations in the ratio of effective breeding and census population sizes between two species of pond-breeding anurans. *Biological Journal of the Linnean Society*, **89**, 365–372.

Brown GP, Phillips BL, Shine R (2014) The straight and narrow path: the evolution of straight-line dispersal at a cane toad invasion front. *Proceedings of the Royal Society of London B: Biological Sciences*, **281**, 20141385.

Canestrelli D, Cimmaruta R, Nascetti G (2008) Population genetic structure and diversity of the Apennine endemic stream frog, *Rana italica*– insights on the Pleistocene evolutionary history of the Italian peninsular biota. *Molecular Ecology*, **17**, 3856–3872.

Che R, Sun Y, Wang R, Xu T (2014) Transcriptomic analysis of endangered Chinese salamander: identification of immune, sex and reproduction-related genes and genetic markers. *PLoS ONE*, **9**, e87940.

Collins JP, Storfer A (2003) Global amphibian declines: sorting the hypotheses. *Diversity and Distributions*, **9**, 89–98.

Corander J, Waldmann P, Sillanpää MJ (2003) Bayesian analysis of genetic differentiation between populations. *Genetics*, **163**, 367–374.

Cornuet JM, Luikart G (1996) Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. *Genetics*, **144**, 2001–2014.

Coulon A, Cosson JF, Angibault JM *et al.* (2004) Landscape connectivity influences gene flow in a roe deer population inhabiting a fragmented landscape: an individual–based approach. *Molecular Ecology*, **13**, 2841–2850.

Crnokrak P, Roff DA (1999) Inbreeding depression in the wild. *Heredity*, **83**, 260–270.

Crossland MR, Brown GP, Anstis M, Shilton CM, Shine R (2008) Mass mortality of native anuran tadpoles in tropical Australia due to the invasive cane toad (*Bufo marinus*). *Biological Conservation*, **141**, 2387–2394.

Daszak P, Berger L, Cunningham AA *et al.* (1999) Emerging infectious diseases and amphibian population declines. *Emerging Infectious Diseases*, **5**, 735.

Davidson C, Shaffer HB, Jennings MR (2002) Spatial tests of the pesticide drift, habitat destruction, UV-B, and climate-change hypotheses for California amphibian declines. *Conservation Biology*, **16**, 1588–1601.

Dong WZ, Zhang XM, Yang CM *et al.* (2010) Iridovirus infection in Chinese giant salamanders. *Emerging Infectious Diseases*, **17**, 2388–2389.

Drummond AJ, Rambaut A, Shapiro B, Pybus OG (2005) Bayesian coalescent inference of past population dynamics from molecular sequences. *Molecular Biology and Evolution*, **22**, 1185–1192.

Duellman WE, Trueb L (1985) *Biology of Amphibians*. McGraw Hill Higher Education, New York, New York.

Dufresnes C, Wassef J, Ghali K *et al.* (2013) Conservation phylogeography: does historical diversity contribute to regional vulnerability in European tree frogs (*Hyla arborea*)? *Molecular Ecology*, **22**, 5669–5684.

Edgar P, Bird DR (2006) Action Plan for the Conservation of the Crested Newt *Triturus cristatus* species complex in Europe [Internet]. Convention on the Conservation of European Wildlife and Natural Habitats. Strasbourg, 2006 Nov 26-30." Cited 2007 Apr 2. Available at http://www.coe.int/t/e/cultural_cooperation/environment/nature_and_biological_diversity/nature_protection/sc26_inf17_en.pdf

Emel SL, Storfer A (2012) A decade of amphibian population genetic studies: synthesis and recommendations. *Conservation Genetics*, **13**, 1685–1689.

Estoup A, Wilson IJ, Sullivan C, Cornuet J-M, Moritz C (2001) Inferring population history from microsatellite and enzyme data in serially introduced cane toads, *Bufo marinus*. *Genetics*, **159**, 1671–1687.

Estoup A, Beaumont M, Sennedot F, Moritz C, Cornuet J-M (2004) Genetic analysis of complex demographic scenarios: spatially expanding populations of the cane toad, *Bufo marinus*. *Evolution*, **58**, 2021–2036.

Estoup A, Baird SJ, Ray N *et al.* (2010) Combining genetic, historical and geographical data to reconstruct the dynamics of bioinvasions: application to the cane toad *Bufo marinus*. *Molecular Ecology Resources*, **10**, 886–901.

Excoffier L, Heckel G (2006) Computer programs for population genetics data analysis: a survival guide. *Nature Reviews Genetics*, **7**, 745–758.

Excoffier L, Smouse PE, Quattro JM (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics*, **131**, 479–491.

Fellers GM, Green DE, Longcore JE, Gatten RE Jr (2001) Oral chytridiomycosis in the mountain yellow-legged frog (*Rana muscosa*). *Copeia*, **2001**, 945–953.

Fellers GM, McConnell LL, Pratt D, Datta S (2004) Pesticides in mountain yellow-legged frogs (*Rana muscosa*) from the Sierra Nevada Mountains of California, USA. *Environmental Toxicology and Chemistry*, **23**, 2170–2177.

Fellers GM, Bradford DF, Pratt D, Wood LL (2007) Demise of repatriated populations of mountain yellow-legged frogs (*Rana muscosa*) in the Sierra Nevada of California. *Herpetological Conservation and Biology*, **2**, 5–21.

Ficetola GF, Bernardi F (2005) Supplementation or in situ conservation? Evidence of local adaptation in the Italian agile frog *Rana latastei* and consequences for the management of populations. *Animal Conservation*, **8**, 33–40.

Ficetola GF, Garner TWJ, Wang J, De Bernardi F (2011) Rapid selection against inbreeding in a wild population of a rare frog. *Evolutionary Applications*, **4**, 30–38.

Fisher RN, Shaffer HB (1996) The decline of amphibians in California's Great Central Valley. *Conservation Biology*, **10**, 1387–1397.

Fitzpatrick BM, Shaffer HB (2004) Environment-dependent admixture dynamics in a tiger salamander hybrid zone. *Evolution*, **58**, 1282–1293.

Fitzpatrick BM, Shaffer HB (2007a) Introduction history and habitat variation explain the landscape genetics of hybrid tiger salamanders. *Ecological Applications*, **17**, 598–608.

Fitzpatrick BM, Shaffer HB (2007b) Hybrid vigor between native and introduced salamanders raises new challenges for conservation. *Proceedings of the National Academy of Sciences of the USA*, **104**, 15793–15798.

Fitzpatrick BM, Johnson JR, Kump DK *et al.* (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 *et al.* (2010) Rapid spread of invasive genes into a threatened native species. *Proceedings of the National Academy of Sciences of the USA*, **107**, 3606–3610.

Frankham R (1995a) Conservation genetics. *Annual Review of Genetics*, **29**, 305–327.

Frankham R (1995b) Effective population size/adult population size ratios in wildlife: a review. *Genetics Research*, **66**, 95–107.

Fumagalli M, Vieira FG, Linderoth T, Nielsen R (2014) ngsTools: methods for population genetics analyses from next-generation sequencing data. *Bioinformatics*, **30**, 1486–1487.

Funk WC, Tallmon DA, Allendorf FW (1999) Small effective population size in the long-toed salamander. *Molecular Ecology*, **8**, 1633–1640.

Funk WC, Blouin MS, Corn PS *et al.* (2005) Population structure of Columbia spotted frogs (*Rana luteiventris*) is strongly affected by the landscape. *Molecular Ecology*, **14**, 483–496.

Funk WC, Cannatella DC, Ryan MJ (2009) Genetic divergence is more tightly related to call variation than landscape features in the Amazonian frogs *Physalaemus petersi* and *P. freibergi*. *Journal of Evolutionary Biology*, **22**, 1839–1853.

Gang L, Baorong G, Ermi Z (2004) *Andrias davidianus*. The IUCN red list of threatened species. Version 2014.3.

Garza JC, Williamson EG (2001) Detection of reduction in population size using data from microsatellite loci. *Molecular Ecology*, **10**, 305–318.

Geng Y, Wang KY, Zhou ZY *et al.* (2011) First report of a ranavirus associated with morbidity and mortality in farmed Chinese giant salamanders (*Andrias davidianus*). *Journal of Comparative Pathology*, **145**, 95–102.

Genomic Resources Development Consortium, Arthofer W, Banbury BL *et al.* (2015) Genomic resources notes accepted 1 August 2014–30 September 2014. *Molecular Ecology Resources*, **15**, 228–229.

Giordano AR, Ridenhour BJ, Storfer A (2007) The influence of altitude and topography on genetic structure in the long-toed salamander (*Ambystoma macrodactulym*). *Molecular Ecology*, **16**, 1625–1637.

Goldberg CS, Waits LP (2010) Comparative landscape genetics of two pond-breeding amphibian species in a highly modified agricultural landscape. *Molecular Ecology*, **19**, 3650–3663.

Griffiths RC, Tavare S (1994) Sampling theory for neutral alleles in a varying environment. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, **344**, 403–410.

Guillot G, Rousset F (2013) Dismantling the Mantel tests. *Methods in Ecology and Evolution*, **4**, 336–344.

Gutierrez JB, Teem JL (2006) A model describing the effect of sex-reversed YY fish in an established wild population: the use of a Trojan Y chromosome to cause extinction of an introduced exotic species. *Journal of Theoretical Biology*, **241**, 333–341.

Hallock L (2013) *Draft State of Washington Oregon Spotted Frog Recovery Plan*. Washington Department of Fish and Wildlife, Olympia, Washington. 93 +v pp.

Halverson MA, Skelly DK, Caccone A (2006) Inbreeding linked to amphibian survival in the wild but not in the laboratory. *Journal of Heredity*, **97**, 499–507.

Hammerson G (2008) *Rana muscosa*. The IUCN red list of threatened species. Version 2014.3.

Hedtke SM, Morgan MJ, Cannatella DC, Hillis DM (2013) Targeted enrichment: maximizing orthologous gene comparisons across deep evolutionary time. *PLoS ONE*, **8**, e67908.

Heled J, Drummond AJ (2008) Bayesian inference of population size history from multiple loci. *BMC Evolutionary Biology*, **8**, 289.

Hellsten U, Harland RM, Gilchrist MJ *et al.* (2010) The genome of the western clawed frog *Xenopus tropicalis*. *Science*, **328**, 633–636.

Hickerson MJ, Carstens BC, Cavender-Bares J *et al.* (2010) Phylogeography's past, present, and future: 10 years after. *Molecular Phylogenetics and Evolution*, **54**, 291–301.

Hoffmann JA (2003) The immune response of *Drosophila*. *Nature*, **426**, 33–38.

Igawa T, Oumi S, Katsuren S, Sumida M (2013) Population structure and landscape genetics of two endangered frog

species of genus *Odorrana*: different scenarios on two islands. *Heredity*, **110**, 46–56.

Jehle R, Arntzen JW (2002) Review: microsatellite markers in amphibian conservation genetics. *The Herpetological Journal*, **12**, 1–9.

Jehle R, Arntzen JW, Burke T, Krupa AP, Hödl W (2001) The annual number of breeding adults and the effective population size of syntopic newts (*Triturus cristatus*, *T. marmoratus*). *Molecular Ecology*, **10**, 839–850.

Jehle R, Burke T, Arntzen JW (2005a) Delineating fine-scale genetic units in amphibians: probing the primacy of ponds. *Conservation Genetics*, **6**, 227–234.

Jehle R, Wilson GA, Arntzen JW, Burke T (2005b) Contemporary gene flow and the spatio-temporal genetic structure of subdivided newt populations (*Triturus cristatus*, *T. marmoratus*). *Journal of Evolutionary Biology*, **18**, 619–628.

Johansson M, Primmer CR, Sahlsten J, Merilä J (2005) The influence of landscape structure on occurrence, abundance and genetic diversity of the common frog, *Rana temporaria*. *Global Change Biology*, **11**, 1664–1679.

Johnson CJ, Parker KL, Heard DC (2001) Foraging across a variable landscape: behavioral decisions made by woodland caribou at multiple spatial scales. *Oecologia*, **127**, 590–602.

Kingman JFC (1982a) On the genealogy of large populations. *Journal of Applied Probability*, **19**, 27–43.

Kingman JFC (1982b) The coalescent. *Stochastic Processes and their Applications*, **13**, 235–248.

Knapp RA, Matthews KR (2001) Non-native fish introductions and the decline of the mountain yellow-legged frog from within protected areas. *Conservation Biology*, **14**, 428–438.

Knowles LL, Maddison WP (2002) Statistical phylogeography. *Molecular Ecology*, **11**, 2623–2635.

Korneliussen TS, Albrechtsen A, Nielsen R (2014) ANGSD: analysis of next generation sequencing data. *BMC Bioinformatics*, **15**, 356.

Krimbas CB, Tsakas S (1971) The genetics of *Dacus oleae*. V. changes of esterase polymorphism in a natural population following insecticide control-selection or drift? *Evolution*, **25**, 454–460.

Krivoruchko K, Gribov A (2004) Geostatistical interpolation and simulation in the presence of barriers. In: *geoENV IV — Geostatistics for Environmental Applications Quantitative Geology and Geostatistics* (eds Sanchez-Vila X, Carrera J, Gómez-Hernández J), pp. 331–342. Springer, Dordrecht, The Netherlands.

Kuhner MK, Yamato J, Felsenstein J (1995) Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. *Genetics*, **140**, 1421–1430.

Kuhner MK, Yamato J, Felsenstein J (1998) Maximum likelihood estimation of population growth rates based on the coalescent. *Genetics*, **149**, 429–434.

Lawson LP (2013) Diversification in a biodiversity hot spot: landscape correlates of phylogeographic patterns in the African spotted reed frog. *Molecular Ecology*, **22**, 1947–1960.

Leblois R, Rousset F, Tikel D, Moritz C, Estoup A (2000) Absence of evidence for isolation by distance in an expanding cane toad (*Bufo marinus*) population: an individual-based analysis of microsatellite genotypes. *Molecular Ecology*, **9**, 1905–1909.

Lillie M, Shine R, Belov K (2014) Characterisation of major histocompatibility complex class I in the Australian cane toad, *Rhinella marina*. *PLoS ONE*, **9**, e102824.

Lind AJ, Spinks PQ, Fellers GM, Shaffer HB (2011) Rangewide phylogeography and landscape genetics of the Western U.S. endemic frog *Rana boylii* (Ranidae): implications for the conservation of frogs and rivers. *Conservation Genetics*, **12**, 269–284.

Lindström T, Brown GP, Sisson SA, Phillips BL, Shine R (2013) Rapid shifts in dispersal behavior on an expanding range edge. *Proceedings of the National Academy of Sciences of the USA*, **110**, 13452–13456.

Looso M, Preussner J, Sousounis K *et al.* (2013) A de novo assembly of the newt transcriptome combined with proteomic validation identifies new protein families expressed during tissue regeneration. *Genome Biology*, **14**, R16.

Luikart G, Cornuet J-M, Allendorf FW (2001) Temporal changes in allele frequencies provide estimates of population bottleneck size. *Conservation Biology*, **13**, 523–530.

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

Mantel N (1967) The detection of disease clustering and a generalized regression approach. *Cancer Research*, **27**, 209–220.

Martel A, der Sluijs AS, Blooi M *et al.* (2013) *Batrachochytrium salamandrivorans* sp. nov. causes lethal chytridiomycosis in amphibians. *Proceedings of the National Academy of Sciences of the USA*, **110**, 15325–15329.

Martel A, Blooi M, Adriaensen C *et al.* (2014) Recent introduction of a chytrid fungus endangers Western Palearctic salamanders. *Science*, **346**, 630–631.

Martínez-Solano I, Teixeira J, Buckley D, García-París M (2006) Mitochondrial DNA phylogeography of *Lissotriton boscai* (Caudata, Salamandridae): evidence for old, multiple refugia in an Iberian endemic. *Molecular Ecology*, **15**, 3375–3388.

McCartney-Melstad E, Waller T, Micucci PA *et al.* (2012) Population structure and gene flow of the yellow anaconda (*Eunectes notaeus*) in northern Argentina. *PLoS ONE*, **7**, e37473.

McCartney-Melstad E, Mount GG, Shaffer HB (2015) Exon capture optimization in large-genome amphibians. *bioRxiv*, 021253.

McMenamin SK, Hadly EA (2012) Ancient DNA assessment of tiger salamander population in Yellowstone National Park. *PLoS ONE*, **7**, e32763.

McRae BH, Beier P (2007) Circuit theory predicts gene flow in plant and animal populations. *Proceedings of the National Academy of Sciences of the USA*, **104**, 19885–19890.

McRae BH, Dickson BG, Keitt TH, Shah VB (2008) Using circuit theory to model connectivity in ecology, evolution, and conservation. *Ecology*, **89**, 2712–2724.

Meng Y, Zhang Y, Liang HW, Xiao HB, Xie CX (2012) Genetic diversity of Chinese giant salamander (*Andrias davidianus*) based on the novel microsatellite markers. *Russian Journal of Genetics*, **48**, 1227–1231.

Merilä J, Söderman F, O'Hara R, Räsänen K, Laurila A (2004) Local adaptation and genetics of acid-stress tolerance in the moor frog, *Rana arvalis*. *Conservation Genetics*, **5**, 513–527.

Minin VN, Bloomquist EW, Suchard MA (2008) Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Molecular Biology and Evolution*, **25**, 1459–1471.

Moore JA, Tallmon DA, Nielsen J, Pyare S (2011) Effects of the landscape on boreal toad gene flow: does the pattern–pro-

cess relationship hold true across distinct landscapes at the northern range margin? *Molecular Ecology*, **20**, 4858–4869.

Morin RD, Chang E, Petrescu A *et al*. (2006) Sequencing and analysis of 10,967 full-length cDNA clones from *Xenopus laevis* and *Xenopus tropicalis* reveals post-tetraploidization transcriptome remodeling. *Genome Research*, **16**, 796–803.

Murphy RW, Fu J, Upton DE, De Lema T, Zhao E-M (2000) Genetic variability among endangered Chinese giant salamanders, *Andrias davidianus*. *Molecular Ecology*, **9**, 1539–1547.

Murphy MA, Dezzani R, Pilliod DS, Storfer A (2010a) Landscape genetics of high mountain frog metapopulations. *Molecular Ecology*, **19**, 3634–3649.

Murphy MA, Evans JS, Storfer A (2010b) Quantifying *Bufo boreas* connectivity in Yellowstone National Park with landscape genetics. *Ecology*, **91**, 252–261.

Nei M, Tajima F (1981) Genetic drift and estimation of effective population size. *Genetics*, **98**, 625–640.

Nishikawa K, Matsui M, Yong H-S *et al*. (2012) Molecular phylogeny and biogeography of caecilians from Southeast Asia (Amphibia, Gymnophiona, Ichthyophiidae), with special reference to high cryptic species diversity in Sundaland. *Molecular Phylogenetics and Evolution*, **63**, 714–723.

Nuñez JJ, Wood NK, Rabanal FE, Fontanella FM, Sites JW Jr (2011) Amphibian phylogeography in the antipodes: refugia and postglacial colonization explain mitochondrial haplotype distribution in the Patagonian frog *Eupsophus calcaratus* (Cycloramphidae). *Molecular Phylogenetics and Evolution*, **58**, 343–352.

Nunziata SO, Scott DE, Lance SL (2015) Temporal genetic and demographic monitoring of pond-breeding amphibians in three contrasting population systems. *Conservation Genetics*, doi: 10.1007/s10592-015-0743-z.

Olmo E (1973) Quantitative variations in the nuclear DNA and phylogenesis of the amphibia. *Caryologia*, **26**, 43–68.

O'Neill EM, Beard KH (2010) Genetic basis of a color pattern polymorphism in the coqui frog *Eleutherodactylus coqui*. *Journal of Heredity*, **101**, 703–709.

Palstra FP, Ruzzante DE (2008) Genetic estimates of contemporary effective population size: what can they tell us about the importance of genetic stochasticity for wild population persistence? *Molecular Ecology*, **17**, 3428–3447.

Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. *PLoS Genetics*, **2**, e190.

Pechmann JHK, Scott DE, Gibbons JW, Semlitsch RD (1989) Influence of wetland hydroperiod on diversity and abundance of metamorphosing juvenile amphibians. *Wetlands Ecology and Management*, **1**, 3–11.

Peterman WE, Feist SM, Semlitsch RD, Eggert LS (2013) Conservation and management of peripheral populations: spatial and temporal influences on the genetic structure of wood frog (*Rana sylvatica*) populations. *Biological Conservation*, **158**, 351–358.

Peterson NP, Cederholm CJ (1984) A comparison of the removal and mark-recapture methods of population estimation for juvenile Coho salmon in a small stream. *North American Journal of Fisheries Management*, **4**, 99–102.

Phillipsen IC, Bowerman J, Blouin M (2010) Effective number of breeding adults in Oregon spotted frogs (*Rana pretiosa*): genetic estimates at two life stages. *Conservation Genetics*, **11**, 737–745.

Phillipsen IC, Funk WC, Hoffman EA, Monsen KJ, Blouin MS (2011) Comparative analyses of effective population size within and among species: ranid frogs as a case study. *Evolution*, **65**, 2927–2945.

Polich RL, Searcy CA, Shaffer HB (2013) Effects of tail-clipping on survivorship and growth of larval salamanders. *The Journal of Wildlife Management*, **77**, 1420–1425.

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

Pybus OG, Rambaut A, Harvey PH (2000) An integrated framework for the inference of viral population history from reconstructed genealogies. *Genetics*, **155**, 1429–1437.

Qiao L, Yang W, Fu J, Song Z (2013) Transcriptome profile of the green odorous frog (*Odorrana margaretae*). *PLoS ONE*, **8**, e75211.

Richardson JL (2012) Divergent landscape effects on population connectivity in two co-occurring amphibian species. *Molecular Ecology*, **21**, 4437–4451.

Richmond JQ, Barr KR, Backlin AR, Vandergast AG, Fisher RN (2013) Evolutionary dynamics of a rapidly receding southern range boundary in the threatened California Red-Legged Frog (*Rana draytonii*). *Evolutionary Applications*, **6**, 808–822.

Richmond JQ, Backlin AR, Tatarian PJ, Solvesky BG, Fisher RN (2014) Population declines lead to replicate patterns of internal range structure at the tips of the distribution of the California red-legged frog (*Rana draytonii*). *Biological Conservation*, **172**, 128–137.

Richter SC, Nunziata SO (2014) Survival to metamorphosis is positively related to genetic variability in a critically endangered amphibian species. *Animal Conservation*, **17**, 265–274.

Richter SC, Crother BI, Broughton RE (2009) Genetic consequences of population reduction and geographic isolation in the critically endangered frog, *Rana sevosa*. *Copeia*, **2009**, 799–806.

Richter-Boix A, Quintela M, Segelbacher G, Laurila A (2011) Genetic analysis of differentiation among breeding ponds reveals a candidate gene for local adaptation in *Rana arvalis*. *Molecular Ecology*, **20**, 1582–1600.

Richter-Boix A, Quintela M, Kierczak M, Franch M, Laurila A (2013) Fine-grained adaptive divergence in an amphibian: genetic basis of phenotypic divergence and the role of non-random gene flow in restricting effective migration among wetlands. *Molecular Ecology*, **22**, 1322–1340.

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

Robertson LS, Cornman RS (2014) Transcriptome resources for the frogs *Lithobates clamitans* and *Pseudacris regilla*, emphasizing antimicrobial peptides and conserved loci for phylogenetics. *Molecular Ecology Resources*, **14**, 178–183.

Roelants K, Gower DJ, Wilkinson M *et al*. (2007) Global patterns of diversification in the history of modern amphibians. *Proceedings of the National Academy of Sciences of the USA*, **104**, 887–892.

Rollins LA, Richardson MF, Shine R (2015) A genetic perspective on rapid evolution in cane toads (*Rhinella marina*). *Molecular Ecology*, **24**, 2264–2276.

Rosenblum EB, Stajich JE, Maddox N, Eisen MB (2008) Global gene expression profiles for life stages of the deadly

amphibian pathogen *Batrachochytrium dendrobatidis*. *Proceedings of the National Academy of Sciences of the USA*, **105**, 17034–17039.

Rosenblum EB, Poorten TJ, Settles M *et al*. (2009) Genome-wide transcriptional response of *Silurana* (*Xenopus*) *tropicalis* to infection with the deadly chytrid fungus. *PLoS ONE*, **4**, e6494.

Rosenblum EB, Poorten TJ, Settles M, Murdoch GK (2012) Only skin deep: shared genetic response to the deadly chytrid fungus in susceptible frog species. *Molecular Ecology*, **21**, 3110–3120.

Ryan ME, Johnson JR, Fitzpatrick BM (2009) Invasive hybrid tiger salamander genotypes impact native amphibians. *Proceedings of the National Academy of Sciences of the USA*, **106**, 11166–11171.

Ryan ME, Johnson JR, Fitzpatrick BM *et al*. (2012) Lethal effects of water quality on threatened California salamanders but not on co-occurring hybrid salamanders. *Conservation Biology*, **27**, 95–102.

San Mauro D, Gower DJ, Oommen OV, Wilkinson M, Zardoya R (2004) Phylogeny of caecilian amphibians (Gymnophiona) based on complete mitochondrial genomes and nuclear RAG1. *Molecular Phylogenetics and Evolution*, **33**, 413–427.

Santana FE, Swaisgood RR, Lemm JM, Fisher RN, Clark RW (2015) Chilled frogs are hot: hibernation and reproduction of the endangered mountain yellow-legged frog *Rana muscosa*. *Endangered Species Research*, **27**, 43–51.

Savage AE, Zamudio KR (2011) MHC genotypes associate with resistance to a frog-killing fungus. *Proceedings of the National Academy of Sciences of the USA*, **108**, 16705–16710.

Savage WK, Fremier AK, Bradley Shaffer H (2010) Landscape genetics of alpine Sierra Nevada salamanders reveal extreme population subdivision in space and time. *Molecular Ecology*, **19**, 3301–3314.

Savage AE, Kiemnec-Tyburczy KM, Ellison AR, Fleischer RC, Zamudio KR (2014) Conservation and divergence in the frog immunome: pyrosequencing and de novo assembly of immune tissue transcriptomes. *Gene*, **542**, 98–108.

Schmeller DS, Merilä J (2007) Demographic and genetic estimates of effective population and breeding size in the amphibian *Rana temporaria*. *Conservation Biology*, **21**, 142–151.

Schoville SD, Tunstall TS, Vredenburg VT *et al*. (2011) Conservation genetics of evolutionary lineages of the endangered mountain yellow-legged frog, *Rana muscosa* (Amphibia: Ranidae), in southern California. *Biological Conservation*, **144**, 2031–2040.

Scoble J, Lowe AJ (2010) A case for incorporating phylogeography and landscape genetics into species distribution modelling approaches to improve climate adaptation and conservation planning. *Diversity and Distributions*, **16**, 343–353.

Searcy CA, Gabbai-Saldate E, Bradley Shaffer H (2013) Microhabitat use and migration distance of an endangered grassland amphibian. *Biological Conservation*, **158**, 80–87.

Searcy CA, Gray LN, Trenham PC, Shaffer HB (2014) Delayed life history effects, multilevel selection, and evolutionary trade-offs in the California tiger salamander. *Ecology*, **95**, 68–77.

Searcy CA, Rollins HB, Shaffer HB (2015) Ecological equivalency as a tool for endangered species management. *Ecological Applications*, in press.

Segelbacher G, Cushman SA, Epperson BK *et al*. (2010) Applications of landscape genetics in conservation biology: concepts and challenges. *Conservation Genetics*, **11**, 375–385.

Semlitsch RD, O'Donnell KM, Thompson FR (2014) Abundance, biomass production, nutrient content, and the possible role of terrestrial salamanders in Missouri Ozark forest ecosystems. *Canadian Journal of Zoology*, **92**, 997–1004.

Serbezov D, Jorde PE, Bernatchez L, Olsen EM, Vøllestad LA (2012) Short-term genetic changes: evaluating effective population size estimates in a comprehensively described brown trout (*Salmo trutta*) population. *Genetics*, **191**, 579–592.

Shaffer HB, Pauly GB, Oliver JC, Trenham PC (2004) The molecular phylogenetics of endangerment: cryptic variation and historical phylogeography of the California tiger salamander, *Ambystoma californiense*. *Molecular Ecology*, **13**, 3033–3049.

Shaffer HB, Gidiş M, McCartney-Melstad E *et al*. (2015) Conservation genetics and genomics of amphibians and reptiles. *Annual Review of Animal Biosciences*, **3**, 113–138.

Shanmuganathan T, Pallister J, Doody S *et al*. (2010) Biological control of the cane toad in Australia: a review. *Animal Conservation*, **13**, 16–23.

Shine BR (2010) The ecological impact of invasive cane toads (*Bufo marinus*) in Australia. *The Quarterly Review of Biology*, **85**, 253–291.

Skerratt LF, Berger L, Speare R *et al*. (2007) Spread of chytridiomycosis has caused the rapid global decline and extinction of frogs. *EcoHealth*, **4**, 125–134.

Slade RW, Moritz C (1998) Phylogeography of *Bufo marinus* from its natural and introduced ranges. *Proceedings of the Royal Society of London B: Biological Sciences*, **265**, 769–777.

Smith MA, Green D (2005) Dispersal and the metapopulation paradigm in amphibian ecology and conservation: are all amphibian populations metapopulations? *Ecography*, **28**, 110–128.

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

Spear SF, Peterson CR, Matocq MD, Storfer A (2005) Landscape genetics of the blotched tiger salamander (*Ambystoma tigrinum melanostictum*). *Molecular Ecology*, **14**, 2553–2564.

Stebbins RC (2003) Mountain yellow-legged frog. In: *A Field Guide to Western Reptiles and Amphibians* (ed. Peterson RT), pp. 233–234. Houghton Mifflin Harcourt, Boston, Massachusetts.

Stoelting RE, Measey GJ, Drewes RC (2014) Population genetics of the São Tomé caecilian (Gymnophiona: Dermophiidae: *Schistometopum thomense*) reveals strong geographic structuring. *PLoS ONE*, **9**, e104628.

Storfer A, Murphy MA, Evans JS *et al*. (2006) Putting the "landscape" in landscape genetics. *Heredity*, **98**, 128–142.

Storfer A, Eastman JM, Spear SF (2009) Modern molecular methods for amphibian conservation. *BioScience*, **59**, 559–571.

Streicher JW, Devitt TJ, Goldberg CS *et al*. (2014) Diversification and asymmetrical gene flow across time and space: lineage sorting and hybridization in polytypic barking frogs. *Molecular Ecology*, **23**, 3273–3291.

Stuart SN, Chanson JS, Cox NA *et al*. (2004) Status and trends of amphibian declines and extinctions worldwide. *Science*, **306**, 1783–1786.

Sun Y-B, Xiong Z-J, Xiang X-Y *et al.* (2015) Whole-genome sequence of the Tibetan frog *Nanorana parkeri* and the comparative evolution of tetrapod genomes. *Proceedings of the National Academy of Sciences of the USA*, **112**, E1257–E1262.

Tabor HK, Risch NJ, Myers RM (2002) Candidate-gene approaches for studying complex genetic traits: practical considerations. *Nature Reviews Genetics*, **3**, 391–397.

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585–595.

Takahata N, Palumbi SR (1985) Extranuclear differentiation and gene flow in the finite island model. *Genetics*, **109**, 441–457.

Tan MH, Au KF, Yablonovitch AL *et al.* (2013) RNA sequencing reveals a diverse and dynamic repertoire of the *Xenopus tropicalis* transcriptome over development. *Genome Research*, **23**, 201–216.

Tao F, Wang X, Zheng H, Fang S (2005) Genetic structure and geographic subdivision of four populations of the Chinese giant salamander (*Andrias davidianus*). *Zoological Research*, **26**, 162–167.

Telles MPdeC, Diniz-Filho JAF, Bastos RP *et al.* (2007) Landscape genetics of *Physalaemus cuvieri* in Brazilian Cerrado: correspondence between population structure and patterns of human occupation and habitat loss. *Biological Conservation*, **139**, 37–46.

Templeton AR (1998) Nested clade analyses of phylogeographic data: testing hypotheses about gene flow and population history. *Molecular Ecology*, **7**, 381–397.

Templeton AR, Routman E, Phillips CA (1995) Separating population structure from population history: a cladistic analysis of the geographical distribution of mitochondrial DNA haplotypes in the tiger salamander, *Ambystoma tigrinum*. *Genetics*, **140**, 767–782.

Trenham PC, Shaffer HB (2005) Amphibian upland habitat use and its consequences for population viability. *Ecological Applications*, **15**, 1158–1168.

Trenham PC, Bradley Shaffer H, Koenig WD, Stromberg MR, Ross ST (2000) Life history and demographic variation in the California tiger salamander (*Ambystoma californiense*). *Copeia*, **2000**, 365–377.

Trenham PC, Koenig WD, Shaffer HB (2001) Spatially autocorrelated demography and interpond dispersal in the salamander *Ambystoma californiense*. *Ecology*, **82**, 3519–3530.

Trumbo DR, Spear SF, Baumsteiger J, Storfer A (2013) Range-wide landscape genetics of an endemic Pacific northwestern salamander. *Molecular Ecology*, **22**, 1250–1266.

Unger SD, Rhodes OE Jr, Sutton TM, Williams RN (2013) Population genetics of the Eastern Hellbender (*Cryptobranchus alleganiensis alleganiensis*) across multiple spatial scales. *PLoS ONE*, **8**, e74180.

US Fish and Wildlife Service (2000) Emergency rule to list the Santa Barbara County distinct population of the California tiger salamander as endangered; rule and proposed rule. *Federal Register*, **65**, 3096–3109.

US Fish and Wildlife Service (2002) Endangered and threatened wildlife and plants; Determination of endangered status for the Southern California distinct vertebrate population segment of the mountain yellow-legged frog (*Rana muscosa*). *Federal Register*, **67**, 44382–44392.

US Fish and Wildlife Service (2005) *Barton Springs Salamander (Eurycea sosorum) Recovery Plan*. Southwest Region, USFWS, Albuquerque, New Mexico.

US Fish and Wildlife Service (2012) Mountain yellow-legged frog (*Rana muscosa*). Southern California distinct population segment. 5-Year review: summary and evaluation.

US Fish and Wildlife Service (2014a) Endangered and threatened wildlife and plants; endangered species status for Sierra Nevada yellow-legged frog and northern distinct population segment of the mountain yellow-legged frog, and threatened species status for Yosemite toad. *Federal Register*, **79**, 24256–24310.

US Fish and Wildlife Service (2014b) Technical/agency draft recovery plan for the dusky gopher frog (*Rana sevosa*).

Voss SR, Shaffer HB, Taylor J, Safi R, Laudet V (2000) Candidate gene analysis of thyroid hormone receptors in metamorphosing vs. nonmetamorphosing salamanders. *Heredity*, **85**, 107–114.

Voss SR, Prudic KL, Oliver JC, Shaffer HB (2003) Candidate gene analysis of metamorphic timing in ambystomatid salamanders. *Molecular Ecology*, **12**, 1217–1223.

Voyles J, Young S, Berger L *et al.* (2009) Pathogenesis of chytridiomycosis, a cause of catastrophic amphibian declines. *Science*, **326**, 582–585.

Vredenburg VT (2004) Reversing introduced species effects: experimental removal of introduced fish leads to rapid recovery of a declining frog. *Proceedings of the National Academy of Sciences of the USA*, **101**, 7646–7650.

Vredenburg VT, Bingham R, Knapp R *et al.* (2007) Concordant molecular and phenotypic data delineate new taxonomy and conservation priorities for the endangered mountain yellow-legged frog. *Journal of Zoology*, **271**, 361–374.

Vredenburg VT, Knapp RA, Tunstall TS, Briggs CJ (2010) Dynamics of an emerging disease drive large-scale amphibian population extinctions. *Proceedings of the National Academy of Sciences of the USA*, **107**, 9689–9694.

Wake DB (1998) Action on amphibians. *Trends in Ecology & Evolution*, **13**, 379–380.

Wake DB, Vredenburg VT (2008) Colloquium Paper: are we in the midst of the sixth mass extinction? A view from the world of amphibians. *Proceedings of the National Academy of Sciences of the USA*, **105**, 11466–11473.

Wang IJ (2012) Environmental and topographic variables shape genetic structure and effective population sizes in the endangered Yosemite toad. *Diversity and Distributions*, **18**, 1033–1041.

Wang X, Zhang K, Wang Z *et al.* (2004) The decline of the Chinese giant salamander *Andrias davidianus* and implications for its conservation. *Oryx*, **38**, 197–202.

Wang IJ, Savage WK, Shaffer HB (2009a) Landscape genetics and least-cost path analysis reveal unexpected dispersal routes in the California tiger salamander (*Ambystoma californiense*). *Molecular Ecology*, **18**, 1365–1374.

Wang Z, Gerstein M, Snyder M (2009b) RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, **10**, 57–63.

Wang IJ, Johnson JR, Johnson BB, Shaffer HB (2011) Effective population size is strongly correlated with breeding pond size in the endangered California tiger salamander, *Ambystoma californiense*. *Conservation Genetics*, **12**, 911–920.

Waples RS (1991) Pacific salmon, *Oncorhynchus spp.*, and the definition of "species" under the Endangered Species Act. *Marine Fisheries Review*, **53**, 11–22.

Waples RS (2002) Definition and estimation of effective population size in the conservation of endangered species. In: *Population Viability Analysis* (eds Beissinger SR, McCullough DR), pp. 147–168. University of Chicago Press, Chicago, Illinois.

Waples RS (2005) Genetic estimates of contemporary effective population size: to what time periods do the estimates apply? *Molecular Ecology*, **14**, 3335–3352.

Wielstra B, Arntzen JW, van der Gaag KJ, Pabijan M, Babik W (2014) Data concatenation, Bayesian concordance and coalescent-based analyses of the species tree for the rapid radiation of *Triturus* newts. *PLoS ONE*, **9**, e111011.

Williams RN, Bos DH, Gopurenko D, DeWoody JA (2008) Amphibian malformations and inbreeding. *Biology Letters*, **4**, 549–552.

Wilson GA, Rannala B (2003) Bayesian inference of recent migration rates using multilocus genotypes. *Genetics*, **163**, 1177–1191.

Woodhams DC, Vredenburg VT, Simon M-A *et al*. (2007) Symbiotic bacteria contribute to innate immune defenses of the threatened mountain yellow-legged frog, *Rana muscosa*. *Biological Conservation*, **138**, 390–398.

Wright S (1943) Isolation by distance. *Genetics*, **28**, 114–138.

Wright S (1951) The genetical structure of populations. *Annals of Eugenics*, **15**, 323–354.

Yang W, Qi Y, Bi K, Fu J (2012) Toward understanding the genetic basis of adaptation to high-elevation life in poikilothermic species: a comparative transcriptomic analysis of two ranid frogs, *Rana chensinensis* and *R. kukunoris*. *BMC Genomics*, **13**, 588.

Yoshikawa N, Kaneko S, Kuwabara K *et al*. (2011) Development of microsatellite markers for the two giant salamander species (*Andrias japonicus* and *A. davidianus*). *Current Herpetology*, **30**, 177–180.

Zardoya R, Meyer A (2001) On the origin of and phylogenetic relationships among living amphibians. *Proceedings of the National Academy of Sciences of the USA*, **98**, 7380–7383.

Zeller KA, McGarigal K, Whiteley AR (2012) Estimating landscape resistance to movement: a review. *Landscape Ecology*, **27**, 777–797.

Zellmer AJ, Knowles LL (2009) Disentangling the effects of historic vs. contemporary landscape structure on population genetic divergence. *Molecular Ecology*, **18**, 3593–3602.

Zhang P, Wake MH (2009) A mitogenomic perspective on the phylogeny and biogeography of living caecilians (Amphibia: Gymnophiona). *Molecular Phylogenetics and Evolution*, **53**, 479–491.

Zhang P, Chen Y-Q, Liu Y-F, Zhou H, Qu L-H (2003) The complete mitochondrial genome of the Chinese giant salamander, *Andrias davidianus* (Amphibia: Caudata). *Gene*, **311**, 93–98.

Zhou ZY, Geng Y, Liu XX *et al*. (2013) Characterization of a ranavirus isolated from the Chinese giant salamander (*Andrias davidianus*, Blanchard, 1871) in China. *Aquaculture*, **384–387**, 66–73.

E.M.M. and H.B.S. conceived of and wrote the manuscript.

MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2016) **16**, 1084–1094 doi: 10.1111/1755-0998.12538

SPECIAL ISSUE: SEQUENCE CAPTURE

Exon capture optimization in amphibians with large genomes

EVAN MCCARTNEY-MELSTAD,* GENEVIEVE G. MOUNT*†‡ and H. BRADLEY SHAFFER*
Department of Ecology and Evolutionary Biology, La Kretz Center for California Conservation Science, Institute of the Environment and Sustainability, University of California, Los Angeles, CA 90095, USA, †Museum of Natural Science, Louisiana State University, Baton Rouge, LA 70803, USA, ‡Department of Biological Sciences, Louisiana State University, Baton Rouge LA 70803, USA

Abstract

Gathering genomic-scale data efficiently is challenging for nonmodel species with large, complex genomes. Transcriptome sequencing is accessible for organisms with large genomes, and sequence capture probes can be designed from such mRNA sequences to enrich and sequence exonic regions. Maximizing enrichment efficiency is important to reduce sequencing costs, but relatively few data exist for exon capture experiments in nonmodel organisms with large genomes. Here, we conducted a replicated factorial experiment to explore the effects of several modifications to standard protocols that might increase sequence capture efficiency for amphibians and other taxa with large, complex genomes. Increasing the amounts of c_0t-1 repetitive sequence blocker and individual input DNA used in target enrichment reactions reduced the rates of PCR duplication. This reduction led to an increase in the percentage of unique reads mapping to target sequences, essentially doubling overall efficiency of the target capture from 10.4% to nearly 19.9% and rendering target capture experiments more efficient and affordable. Our results indicate that target capture protocols can be modified to efficiently screen vertebrates with large genomes, including amphibians.

Keywords: amphibian, exon capture, large genome, target enrichment

Received 15 September 2015; revision received 13 April 2016; accepted 6 May 2016

Introduction

Reduced representation sequencing technologies enrich DNA libraries for selected genomic regions, allowing researchers to attain higher sequencing depth over a predetermined subset of the genome for a given cost. Several techniques are now in widespread use in population genetics and evolutionary biology. The most popular of these include RAD-seq (Miller *et al*. 2007) and target enrichment approaches such as ultra-conserved element (UCE) sequencing (Faircloth *et al*. 2012), anchored enrichment (Lemmon *et al*. 2012), and exon/exome sequencing.

These methods are all useful for different purposes. RAD-seq targets anonymous loci flanking restriction enzyme sites and is a cost-effective strategy for collecting information on thousands of anonymous loci for individuals within a population. However, it suffers from bias and large amounts of missing data, especially when divergent individuals are analysed (Arnold *et al.* 2013). Conversely, UCE and anchored enrichment sequencing target regions of the genome that are moderately or highly conserved among species and are designed to generate relatively complete data sets across distantly related species using a single set of probes (Faircloth *et al.* 2012; Lemmon *et al.* 2012). However, the biological functions of these conserved loci are mostly unknown, and by definition they enrich for regions of relatively low divergence.

Exon capture differs from RAD-seq in that it targets pre-identified sequence regions, and it is distinct from UCE sequencing in that it often targets known gene regions that are expressed as RNAs and are functionally important. As such, exon capture is a promising technology for gathering large amounts of targeted genomic data for population-level studies exploring patterns of population structure and natural selection (Hodges *et al*. 2007; Yi *et al*. 2010). It is particularly useful for collecting data from species without assembled reference genomes, as the prerequisite genomic information may be gathered from existing collections of expressed sequence tag (EST) sequences or transcriptome sequencing (Bi *et al*. 2012; Neves *et al*. 2013).

The laboratory approaches for UCE and exon capture and sequencing are the same (Gnirke *et al*. 2009; Blumenstiel *et al*. 2010). Both rely on the hybridization of synthetic biotinylated RNA or DNA probes to library

Correspondence: Evan McCartney-Melstad, Fax: 310-206-3987; E-mail: evanmelstad@ucla.edu

fragments from samples of interest. After hybridization, the biotin on these probes is bound to streptavidin molecules attached to paramagnetic beads, allowing the target sequences to be magnetically captured, and nonhybridized DNA is washed away. Unfortunately, capture of off-target DNA can happen for several reasons and can drastically reduce the efficiency of sequencing (Hodges *et al.* 2009). Because library fragments are often longer than the probe sequences, part of the hybridized library fragment is usually free to bind to other molecules in the pool. Repetitive DNA sequences are often present at high concentrations in large-genome organisms; if this exposed region is from a repetitive element, it has a high probability of binding to another such fragment, pulling the entire construct through to the final library pool. Adapter sequences are also present at very high concentrations, presenting another opportunity for molecules to bind to captured fragments, creating 'daisy chains' of random library molecules (Hodges *et al.* 2009; Nijman *et al.* 2010). To mitigate these factors, several 'blockers', designed to hybridize to high-copy-number regions early in the protocol and prevent daisy chaining, are typically added to target capture reactions. One such blocker, c_0t-1, is a solution of high-copy repetitive DNA fragments that hybridizes with repetitive library fragments and blocks them from attaching to captured fragments. Amphibian genomes contain complex patterns of repetitive elements that may be present at an even higher concentration than normal (Straus 1971; Sun & Mueller 2014; Keinath *et al.* 2015), and we hypothesize that increasing the amount of c_0t-1 in solution may improve hybridization efficiency.

Amphibians have large genomes, ranging up to 117 gigabases (Gregory 2002), currently rendering full-genome sequencing approaches untenable, but exon capture is well suited to bridge the gap between single-locus comparative studies and whole-genome analyses for these and other diploid species with large genomes. Several amphibian species have large collections of EST sequences available (Abdullayev *et al.* 2013; Robertson & Cornman 2014), and transcriptome sequencing and *de novo* assembly are becoming increasingly accessible for species that lack such resources. Despite the available resources, few exon capture studies have been performed in amphibians (but see Hedtke *et al.* 2013). This likely reflects limited success of those who have tried and reticence of others to attempt sequence capture approaches with large, highly repetitive genomes.

Beyond the initial purchase price of custom exon probe sets, laboratory costs of exon capture experiments primarily hinge on the efficiency of the enrichment process. Increasing the percentage of reads 'on target' (sequence reads that align to regions targeted in the capture array) directly reduces the amount of sequencing required to attain a desired coverage level. Off-target reads may be present for several reasons, including nonspecific hybridization of capture probes to off-target regions, hybridization of off-target DNA to the ends of captured target fragments, and failure to wash away DNA not hybridized to capture probes following enrichment (Hodges *et al.* 2009). The ratio of off-target reads may be particularly problematic in amphibians because their large genome size often reflects a massive increase in the amount of repetitive DNA (Straus 1971; Sun & Mueller 2014; Keinath *et al.* 2015), which leads to a greatly increased concentration of off-target DNA in solution relative to on-target fragments.

We conducted a series of experiments to optimize existing protocols for exon capture experiments for amphibians and other taxa with large genomes. Our focus is on three different *Ambystoma* salamanders—the California tiger salamander (*Ambystoma californiense*), the barred tiger salamander (*Ambystoma mavortium*) and an F1 hybrid between the two (*Ambystoma californiense* × *mavortium*, referred to as F1). Given the enormous size of their genomes, estimated at about 32 gigabases (Keinath *et al.* 2015), and the observation that they, like many amphibians, have genomes rich in repetitive DNA, we altered the amount of c_0t-1 blocker, under the assumption that highly repetitive genomes may benefit from an increased amount of repetitive sequence blocker. We also manipulated the amount of individual input and total DNA in sequence capture reactions to manipulate the total number of copies of the genome, estimating tradeoffs among multiplexibility and enrichment efficiency to maximize the number of individuals that can be sequenced for each sequence capture reaction.

Materials and methods

Array design and laboratory methods

We designed an array from 8706 putative exons (8706 distinct genes) using EST sequences from the closely related Mexican axolotl (*Ambystoma mexicanum*) (Smith *et al.* 2005). Mitochondrial sequence divergence between the California tiger salamander and the Mexican axolotl is approximately 6.4% (Samuels *et al.* 2005), and is approximately 1.2% between the barred tiger salamander and Mexican axolotl (Shaffer and McKight 1996), suggesting that less-diverged nuclear exons from the axolotl should serve as appropriate targets for our species. In our design, we attempted to avoid targeting regions that span exon/intron boundaries, as these targets have been found to be much less efficient (Neves *et al.* 2013). Exon boundaries can be found by mapping EST sequences to a reference genome while allowing for long gaps that represent introns. However, no salamander genome is

currently available, and the two available frog genomes [*Xenopus tropicalus* (Hellsten *et al.* 2010) and *Nanorana parkeri* (Sun *et al.* 2015)] last shared a common ancestor with salamanders approximately 290 million years ago (San Mauro 2010). To account for this, we developed a comparative method for conservatively predicting intron splice sites within EST sequences (E. McCartney-Melstad & H. B. Shaffer, unpublished data). Target sequences were an average of 290 bp in length (minimum length = 88 bp, maximum = 450 bp, standard deviation = 71 bp), for a total target region length of 2.53 megabases. A total of 39 984 100-bp probe sequences were tiled across these target regions at an average of 1.8× tiling density. These probes were synthesized as biotinylated RNA oligos in a MYBAITS kit (MYcroarray, Ann Arbor, MI).

We used a salt extraction protocol (Sambrook & Russell 2001) to extract genomic DNA from three individual salamanders: one California tiger salamander (*Ambystoma californiense* #HBS127160—CTS), one barred tiger salamander (*Ambystoma mavortium* #HBS127161—BTS) and one F1 hybrid between the two species (#HBS109668). Multiple independent extractions were performed for each individual to attain the amount needed for preparing several libraries. Extractions were then combined into pools and used for library preparations. Two of these pools consisted of pure California tiger salamander DNA or pure F1 DNA and are labelled CTS and F1, respectively. The third pool, which was intended to be pure BTS, was found to consist of approximately 70% barred tiger salamander DNA and 30% California tiger salamander DNA, apparently due to a pooling error, which was later verified through reextraction of the original tissues and Sanger sequencing. We refer to this pool as BTS* and treat it as a third sample in our experimental design. DNA was diluted to 20 ng/μL and sheared to approximately 500 bp on a BioRupter (Diagenode, Denville, NJ). For each of the 53 individual library preparations (Table S1, Supporting information), we used approximately 450 ng of DNA for library preparations. Standard Illumina library preparations (end repair, A-tailing and adapter ligation) were performed using Kapa LTP library preparation kits (Kapa Biosystems, Wilmington, MA). Samples were dual-indexed with 8-bp indices that were added via PCR (adapters from Travis Glenn, University of Georgia). Following library preparation, we performed a double-sided size selection with SPRI beads (Bronner *et al.* 2009) to attain a fragment size distribution centred around 400 bp and ranging from 200 bp to 1000 bp. Species-specific c_0t-1 was prepared using DNA extracted from a California tiger salamander and a single-strand nuclease as follows: First, extracted DNA was treated with RNase and brought to 500 μL at 1000 ng/μL in 1.2× SSC. This DNA was then sheared on a BioRuptor (Diagenode, Denville, NJ) to approximately 300 bp. Next, the solution was denatured at 95 °C for 10 min, partially renatured at 60 °C for 5 min and 45 s, placed on ice for 2 min and then put in a 42 °C incubator. A preheated 250 μL aliquot of S1 nuclease (in buffer) was then added to the partially renatured DNA and incubated for 1 h at 42 °C. The DNA was then precipitated with 75 μL of 3 M sodium acetate and 750 μL isopropanol and centrifuged for 20 min at 10 000 ***g*** at 4 °C. Isopropanol was then removed and the pellet was washed with 500 μL of cold 70% ethanol, centrifuged again at 10 000 ***g*** for 10 min (4 °C) and dried following ethanol removal. We rehydrated this pellet with 50 μL of 10 mM Tris-HCl, pH 8 and dried it down to the appropriate concentration (for 1× c_0t-1, 500 ng/μL; for 6× and 12× c_0t-1, 1000 ng/μL).

We then multiplexed prepared libraries into capture reactions (Table S1, Supporting information). Total DNA input into the sequence capture was either 500 ng or 1000 ng, and individual library input DNA for multiplexing ranged from 20 to 1000 ng (Table S1, Supporting information). The repetitive DNA blocker c_0t-1 was added to the 24 different capture reactions in one of three amounts—2500, 15 000 or 30 000 ng, corresponding to 1×, 6× and 12× protocol recommendation. Libraries were enriched using the MYBAITS protocol (version 2.3.1), hybridizing probes for 24.5 h and implementing the optional high-stringency washes. Following the three wash steps in the MYBAITS protocol, we amplified the remaining enriched DNA (with streptavidin beads still in solution) using 14 cycles of PCR. Multiple separate PCR reactions were performed for each capture reaction, which were subsequently pooled after amplification to reduce PCR amplification bias (Barnard *et al.* 1998).

Postcapture, post-PCR libraries were quantitated and characterized with qPCR using the Kapa Illumina library quantification kit (PicoGreen® Life Technologies, Grand Island, NY and Kapa Biosystems, Wilmington, MA) on a LightCycler 480 (Roche, Basel, Switzerland). We also visualized fragment size distributions using a BioAnalyzer 2100 DNA HS chip (Agilent, Santa Clara, CA). All capture reactions were tested for preliminary evidence of enrichment via qPCR. We developed five primer pairs for different test loci chosen from our targets as positive controls, and one primer pair from a mitochondrial locus we were not targeting as a negative control. We used these to measure the relative concentrations of target molecules in solution by calculating the mean number of cycles required for qPCR reactions to reach the crossing point (C_p) in libraries pre- and post-enrichment. Changes in (C_p) were measured for each test locus for all samples and averaged across all five test loci. For targeted loci, we expected that the number of cycles needed to reach this point would decrease, because target sequences

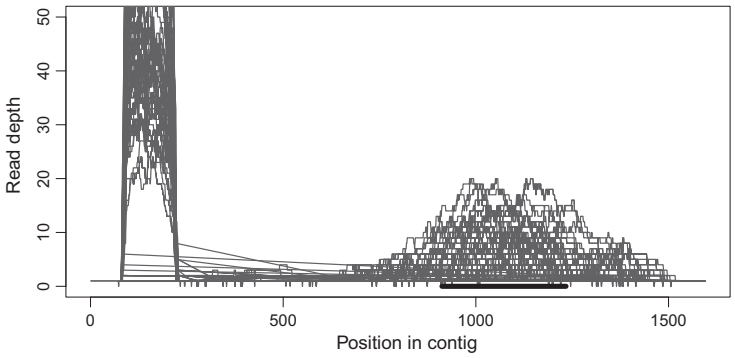

Fig. 1 Coverage across a sample target. The black bar on the bottom corresponds to the target region from which probes were synthesized. Each grey line represents a single library. There are two peaks of coverage, one centred on the target region, and a much higher spike of coverage at the left edge of the contig, likely corresponding to a repetitive region in the genome. The latter type of spike is reduced through the chimera-filtering steps described in the text.

would be present in higher concentrations. Conversely, we expected the number of cycles for the mitochondrial DNA locus to increase after enrichment, because that sequence was not targeted and we expected its concentration to decrease.

All capture reactions were combined together for sequencing on an Illumina HiSeq 2500 with 150 bp paired-end reads. Reactions were pooled such that all individual libraries would receive at least 1.5 million reads (Table S1, Supporting information). Sample pooling and sequencing was performed at the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley.

Genetic data analysis

Demultiplexed reads were checked for adapter contamination and quality trimmed using Trimmomatic 0.32 (Bolger *et al.* 2014). Quality trimming was performed using several criteria. First, leading base pairs with a phred score <5 were removed. Next, trailing (3′) base pairs with a phred score <15 were removed. Finally, we used a four base pair sliding window (5′ to 3′), trimming all trailing bases when the average phred score within that window dropped below 20. We discarded all reads under 40 bp after trimming, and overlapping reads were merged using fastq-join (Aronesty 2013).

Genetic data from all of the California tiger salamander libraries were combined for assembly to create the most complete possible single-species *de novo* assembly of our target regions. Targets were *de novo* assembled using the Assembly by Reduced Complexity (ARC) pipeline (Hunter *et al*. 2015). This assembly pipeline separates reads that align to target regions and performs small, target-specific *de novo* assemblies on these read pools. Each assembled contig then replaces its original target sequence, and the process is repeated iteratively. Within ARC, read mapping was performed using Bowtie2 (Langmead & Salzberg 2012), error correction with BayesHammer (Nikolenko *et al.* 2013), and assemblies were generated using SPAdes (Bankevich *et al*. 2012). The ARC pipeline was run for six iterations, which was enough to exhaust all of the reads assignable to most targets.

Following assembly, all contigs were compared against the original target sequences using blastn (Camacho *et al.* 2009), and reciprocal best blast hits (RBBHs) were found (Rivera *et al.* 1998). Chimeric assemblies are pervasive and problematic for studies involving *de novo* assembly of target sequences, because they can insert repetitive sequences into the contigs, making it appear that many reads are mapping to a target when those reads are actually from repetitive regions in the genome (for instance, see the coverage across the example contig in Fig. 1). To attempt to reduce the presence of chimeric assemblies and repetitive sequences in our data, the RBBHs were blasted to themselves (blastn *e*-value of 1×10^{-20}), and base pairs in sequence regions that positively matched other targets were replaced with *N*′s. These chimera-masked RBBHs served as our final assembled target set.

After assembly, reads from each individual were mapped against the chimera-masked RBBH target set using BWA-MEM (Li 2013). BAM file conversion, sorting, and merging was done using SAMtools v1.0 (Li *et al*. 2009). PCR duplicates were marked using picard tools v. 1.119 (http://broadinstittute.github.io/picard) which finds fragments that have identical 5′ and 3′ mapping coordinates, under the assumption that two different chromosomal copies are unlikely to shear in the exact same positions with random sonication. Under this assumption, fragments with identical 5′ and 3′ mapping coordinates most likely are the result of sequencing multiple amplified copies of the same original DNA molecule, which is undesirable. Finally, mapping rates and PCR duplication rates were inferred by counting the relevant SAM flags using SAMtools flagstat (Li *et al*. 2009).

In addition to measuring the total percentage of unique reads that mapped to target regions, target-level performance was also evaluated. Because most targets showed a characteristic peak of read depth centred over

the middle of the target where probes were tiled, and because a few targets maintained confounding repetitive sequences at the periphery of the assembled contigs, we characterized the read depths of targets over bases that had direct overlap with our target probes. That is, for target-level metrics, we did not consider read depth for the flanking regions that are naturally appended to the ends of each target during the assembly process. For each individual library preparation, we calculated the average unique read sequencing depth across (i) entire target regions and (ii) across the 100-bp window within each target that had the highest average coverage. For all read depth comparisons, depths were corrected for the total number of reads a library received in sequencing by multiplying by a scaling factor n_f/n_i, where n_f is the fewest number of reads received by any individual in the experiment and n_i is the number of reads received by the individual under consideration. Assembled target sequences <100 bp were not included in read depth calculations because 100 bp is significantly less than the average read length and these targets tended to recruit very few reads.

Assessing the importance of c_0t-1 and individual input DNA amounts

Linear regression was used to quantify the relationships between c_0t-1 and individual input DNA to the percentage of unique reads that mapped to targets. Because three different biological individuals were used for library preparations in this experiment, we also included the identity of the individual as a possible source of variation to explain enrichment efficiency. Models were built that included different combinations of c_0t-1, individual input DNA and the identity of the individual (CTS, BTS* or F1) as predictor variables, and unique reads mapping to targets as the response variable. A similar approach was used to model the average sequencing depths across all targets. All models were evaluated by examining the regression coefficients, adjusted R^2 and AIC values.

Results

Presequencing library quantitation

DNA concentration yields for post-enrichment, post-PCR samples were lower than anticipated. After 14 PCR cycles, amplified enrichment pools contained an average of 279.5 ng of DNA (after amplifying 15 μL of a total 33 μL in the post-enrichment pools with a 50 μL PCR reaction). One capture reaction (Library #18, see Table S2, Supporting information) had a much higher yield after post-enrichment PCR (2150 ng). Mean C_P in qPCR enrichment verification reactions decreased by an average of 9.1 cycles across the five test loci after enrichment, while the number of cycles required for amplification of a nontargeted negative control locus increased by an average of 2.17 cycles. We found a positive correlation between the mean change in C_P averaged across the five test loci and the raw percentage of reads on target after sequencing for each library (Fig. S1, Supporting information, adjusted $R^2 = 0.1136$, $P = 0.00784$), although the relationship was stronger between post-enrichment, post-PCR DNA concentration and raw mapping rate (Fig. 2, adjusted $R^2 = 0.224$, $P = 0.000204$).

Sequence data

We generated 45 641 469 300 base pairs of sequence data in the form of 150 bp paired-end reads. All libraries received at least 1 207 605 read pairs passing filter (mean = 2 766 149 read pairs, SD = 1 582 161 read pairs). Average base quality phred scores for samples ranged from 33.6 to 34.8 (mean = 34.4, SD = 0.29). An average of 93% of all read pairs both passed the Trimmomatic filter, whereas 5.2% of all read pairs had either the forward or reverse read removed, and 1.8% had both members removed. Because our insert size was mostly larger than 300 bp (which is two times the read length), fastq-join did not merge most reads—percentages of

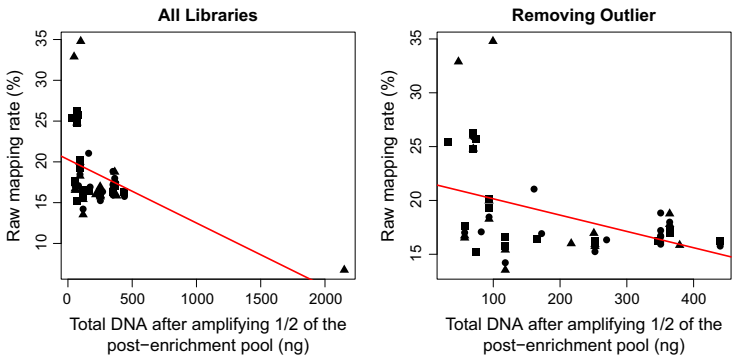

Fig. 2 Relationship between post-enrichment DNA concentration and percentage of raw reads mapping to targets. Each dot is an individual library: square = CTS, triangle = F1, circle = BTS*. For the full data set, adjusted $R^2 = 0.224$, $P = 0.000204$. After removing the single F1 outlier, adjusted $R^2 = 0.1732$, $P = 0.00126$.

joined reads ranged from 24.0% to 35.1% for the different samples. Nuclear sequence divergence between the Mexican axolotl (the species from which probes were designed) and California tiger salamander in the exon targets averaged 1.84%.

Reference assembly and read mapping

A total of 78 674 304 reads (all of the reads from the CTS individual) representing 11 960 279 114 bp were supplied to ARC for *de novo* assembly of targets. An average of 905 reads in iteration 1, 1496 reads in iteration 2, 1999 reads in iteration 3, 4485 reads in iteration 4, 8132 reads in iteration 5 and 11 199 reads in iteration 6 were assigned to each target for *de novo* assembly. The final assembly, after six iterations of the ARC assembly pipeline, contained 120 617 sequences for a total of 69 873 191 bp. After blasting the target sequences to the assembly and *vice versa*, we found a total of 8386 RBBHs, or 96.3% of all targets. These assembled target contigs were 1409 bp on average, for a total reference length of 11 813 341 bp. This average extension of 1119 to each target sequence was expected, as the insert size in our genomic library preparations ranged up to approximately 550 bp. Thus, 550 bp fragments that contained target sequence on either end could still be hybridized by the capture probes and their sequence at the other end recruited into the target assembly. Self-blasting the target RBBHs to one another resulted in 1060 targets that also had hits with other targets. A total of 361 949 bp of such overlap was found between targets, and the overlapping bases were replaced with *N*'s to reduce the effects of repetitive sequences and chimeric assemblies.

An average of 18.21% of all reads across samples mapped to the chimera-masked reciprocal blast hit target assembly. Individual sample raw read mapping rates ranged from 6.7% to 34.8% (Table S2, Supporting information). The percentage of PCR duplicates present also varied widely across samples, ranging from 8.5% to 48.6% (mean = 24.5%, SD = 11.7%). After subtracting PCR duplicates from mapped reads, the percentage of unique reads on target varied between 5.4% and 30.8%, with a mean of 14.0% and standard deviation of 4.4% (Table S2, Supporting information).

Target-level metrics indicated some targets performed significantly better than others (Fig. 3). To control for variation in the number of reads received between samples, all libraries had their depths corrected to what would have been observed if they had received the same number of reads as the least-sequenced library in this study. To give an idea of the sequencing effort required to generate the depths listed below, this corresponded to approximately 2.4 million 150 bp reads against just over 2.5 million bp of total target sequence. Among all

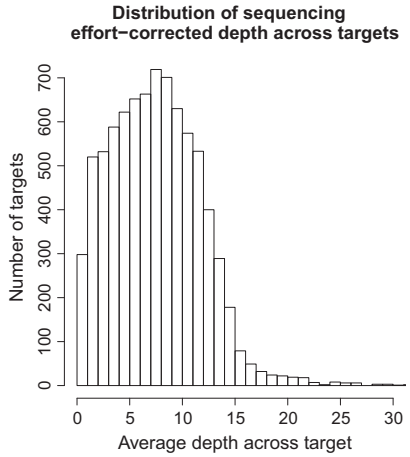

Fig. 3 Average sequencing depths across targets. The average sequencing depth across all targets regions averaged between all samples, calculated using *samtools depth*. The highest 31 values, which had depths higher than 30, are not shown here.

libraries, the average depth across target sequences was 7.99 (SD = 3.33), and the average for the highest 100 bp window within targets was 9.50 (SD = 3.89). A total of 5648 targets had a sequencing effort corrected average depth across the target region >5, and 2283 had average depths >10. For the 100 bp windows with the greatest depth for each target, 6100 had depths >5 and 3313 had depths >10.

Effects of c_0t-1 and input DNA amount in capture reactions

All models that incorporated the identity of the individual DNA pool underperformed (higher AIC value) nested models that did not incorporate information regarding the identity of the input DNA. Because of this, and because slope coefficients for the identity term in all models were never significant (P = 0.44 or greater), the identity of the individual did not significantly improve capture efficiency or read mapping, and models including this variable are not included in the summary tables.

Increasing amounts of individual input DNA and c_0t-1 blocker were both associated with higher percentages of unique reads on target and higher realized sequence depth across targets (Tables 1 and 2, Fig. 4). Linear regression recovered positive and significant slopes for both variables separately and when combined in multiple linear regression. Models predict an extra 1% of unique reads on target for every 166 ng of extra individual input DNA (P = 0.000672) or every 6750 ng of extra c_0t-1 blocker (P = 0.00896) used in enrichment reactions. Regression coefficients for models that contained both

Table 1 Model comparison predicting percentage of unique reads on target, sorted by AIC values

Model	R^2	Adj. R^2	AIC
c_0t1** + inputDNA***	0.3252	0.2982	−193.6057
inputDNA***	0.2046	0.189	−186.8963
c_0t-1**	0.1265	0.1094	−181.9297

***Signifies $P < 0.001$, **signifies $0.001 < P < 0.01$, *signifies $0.01 < P < 0.05$.

Table 2 Model comparison predicting average depth across target region, sorted by AIC values

Model	R^2	Adj. R^2	AIC
c_0t1* + inputDNA**	0.252	0.222	269.6817
inputDNA**	0.1676	0.1513	273.3437
c_0t-1*	0.08887	0.07101	278.1344

***Signifies $P < 0.001$, **signifies $0.001 < P < 0.01$, *signifies $0.01 < P < 0.05$.

individual input DNA and c_0t-1 were quite similar to the single-variable model, differing by <3%. Individual input DNA and c_0t-1 did a better job predicting the percentage of unique reads on target than the average depth across target regions (adjusted R^2 of 0.325 vs. 0.252 for the combined models). Finally, the models that contained both input DNA and c_0t-1 as variables had better (lower) AIC scores and higher R^2 values than the nested single-variable models (see Fig. 5), and within the single-variable tests, individual input DNA models outperformed c_0t-1 models for both success measures in AIC and R^2 (Tables 1 and 2).

Discussion

Perhaps the most important conclusion from this experiment is that target capture experiments can indeed be successful in amphibians (and other taxa) with large genomes. This was not at all obvious based on prior work on these organisms, and our hope is that others will use these results to bring amphibians more fully into the

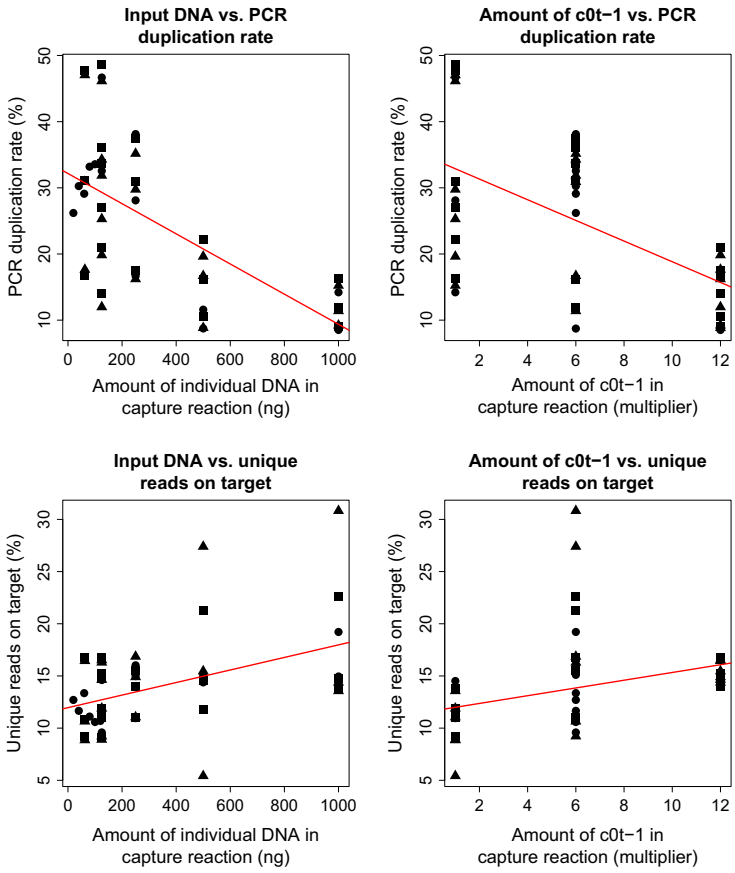

Fig. 4 Relationship between individual input DNA and c_0t-1 amounts and PCR duplication rates and enrichment efficiency. Each dot is an individual library: square = CTS, triangle = F1, circle = BTS*. *P*-values for slope coefficients in the four panels are as follows: top left $P = 1.39 \times 10^{-7}$, top right $P = 9.28 \times 10^{-6}$, bottom left $P = 0.000672$, bottom right $P = 0.00896$.

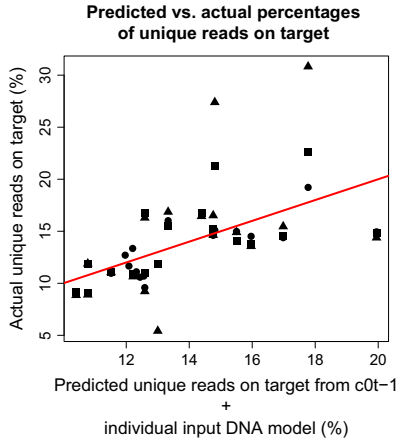

Fig. 5 Predicted vs. actual unique reads on target using two-variable model. The model contains both c_0t-1 and individual input DNA, and the diagonal line shows the 1:1 relationship between predicted and actual unique reads on target. Points close to the line mean their unique reads on target are well predicted by the two variables, and points farther away from the line are not as well predicted. Each dot is an individual library: square = CTS, triangle = F1, circle = BTS*.

realm of population and phylogenomic analyses. The percentage of unique reads on target is the most important summary metric for enrichment. Our average percentage of unique reads on target across all library treatments was 14%; only three libraries were under 9%, while our four best-performing libraries were all over 20%. These numbers suggest that it is reasonable to sequence 50–100 samples on a single HiSeq lane for a capture array size similar to ours (2.5 megabases), depending on array configuration and coverage requirements, which vary based on the particular application.

Our percentages of unique reads on target are in line with several other nonmodel exon capture studies for species with smaller genomes. For instance, Hedtke *et al.* (2013) designed Agilent probes from the *Xenopus tropicalus* genome and enriched libraries from two smaller-genome frogs, achieving rates of 7.4% unique reads on target in *Pipa pipa* and 47.8% in *Xenopus tropicalus*. Bi *et al.* (2012) recovered 25.6% to 29.1% unique reads on target for an exon capture study in chipmunks. Similarly, Cosart *et al.* (2011) designed an Agilent exon capture microarray from the bovine (*Bos taurus*) genome and attained 20–29% unique read mapping percentages in *Bos taurus*, *Bos indicus* and *Bison bison* for a similarly sized target array as this study. Finally, Neves *et al.* (2013) reached 50% raw mapping rates in multiplexed exon capture experiments in *Pinus taeda*, a pine species with a approximately 21 Gb genome (approximately 2/3 of the size of the salamander genomes in this study), although they did not report percentages of unique reads on target or levels of PCR duplication. Several factors may be important in explaining these results, including a potential negative relationship between the phylogenetic distance to the species from which the capture array was developed and the percentage of unique reads on target, and the size of the genome under investigation. As more target capture studies are reported across diverse nonmodel taxa, we will better understand the relationship between genome size and enrichment efficiency, as well as the effects of designing capture probes from divergent taxa.

Human exome capture studies, which typically use predesigned sequence capture arrays across one of several different technologies (e.g. Truseq, Nimblegen, Agilent or Nextera exome capture kits) often attain percentages of unique reads on target in the range of 40% to 70% or higher (Chilamakuri *et al.* 2014). However, the high numbers in human experiments are likely a function of many iterations of probe set optimization experiments that have been conducted, which is generally not feasible in non-human systems.

We found evidence that increasing c_0t-1 and individual input DNA into sequence capture reactions increased the percentage of unique reads mapping to targets. This effect was driven largely by the correlation of these two variables with the reduction in PCR duplication rates (Fig. 4). Because duplicate reads (reads with the same 5′ and 3′ mapping coordinates) are typically removed prior to genotyping analyses, lowering duplication rates as much as possible is critical for increasing the efficiency, and therefore reducing the sequencing costs of target enrichment studies. Researchers are generally encouraged to use paired-end sequencing whenever possible in exon capture studies, as single-end reads have a much higher false identification rate of PCR duplication (Bainbridge *et al.* 2010).

The low yields of DNA after enrichment and PCR are interesting. We speculate that they may be a consequence of libraries prepared from large genomes containing relatively low absolute numbers of on-target fragments in the pools during enrichment, so that a higher percentage of the pool is washed away. Comparing the results of qPCR from pre- and post-enrichment libraries using primers meant to amplify targeted regions is a common way to qualitatively assess enrichment efficiency, but we found post-enrichment DNA concentrations to be a better predictor of enrichment success with our protocol (Figs 2 and S1, Supporting information). Also, we note that Library #18, which had a very high post-enrichment post-PCR DNA concentration, showed correspondingly low performance in terms of percentage of raw and unique reads on target (5.4% unique read mapping rate). This suggests that off-target fragments may not have

been efficiently removed during the post-enrichment washing steps in that library.

After duplicate removal, we observed a greater than fivefold difference in unique read mapping percentages (from 5.4% to 30.8%) among the samples tested in this experiment. Given that the targeted region represents approximately 0.008% of the 32 Gb *Ambystoma* genome (Keinath *et al.* 2015), this is a striking improvement over whole-genome shotgun sequencing. While even the low end of our enrichment efficiency values are encouraging for future exon capture studies in large-genome amphibians, regularly attaining percentages of unique reads on target at the upper end of our success rate would lead to a concurrent fivefold reduction in sequencing costs for a given target coverage depth.

Based on these experiments, we recommend using at least 30 000 ng of species-specific c_0t-1 blocker, and as much input DNA as possible for each individual multiplexed into a capture reaction when working with large-genome species. One drawback of this recommendation is that producing this much c_0t-1 is challenging in species where it is difficult or impossible to extract large amounts of DNA, in which case preparing c_0t-1 from a closely related species or performing whole-genome amplification on a smaller starting quantity of species-specific c_0t-1 may suffice. The threshold at which the addition of more blocker DNA ceases to improve (and may potentially inhibit) capture efficiency is not yet known. Additional experiments should attempt to define this limit and should also seek to understand whether additional c_0t-1 blocker enhances target enrichment in more modestly sized amphibian genomes.

Amounts of individual input DNA are constrained by the total amount of DNA in the capture reaction divided by the number of samples. This means that for a set amount of total input DNA, increasing the number of individuals multiplexed into a single capture reaction will decrease the percentage of unique reads on target. Ongoing research in our laboratory to further optimize target capture for taxa with large genomes is focusing on resolving potential trade-offs resulting from different multiplexing regimes and from increasing the total amount of DNA going into capture reactions and individual library preparations. Although we can only speak directly to experiments that utilize custom MYBAITS exon enrichment reactions, we see no reason why our results should not generalize to other platforms such as UCEs (Faircloth *et al.* 2012).

There are several sources of variation in capture efficiency in addition to those explored here. One is the number of mismatches between probes and DNA, which is a function of the evolutionary distance of the species from which probes were designed to the species being enriched. For instance, Hedtke *et al.* (2013) found a negative linear relationship between age of the most recent common ancestor to the probe-design species and fold enrichment values. Those authors observed a greater than threefold decrease in capture efficiency as the enriched species approached 250 million years of divergence from *Xenopus tropicalis*, the species from which probes were designed. One reasonable approach to mitigate this effect is to first generate a reference transcriptome for the species of interest through RNA-seq, which can then be used to design probes. Another important variable is hybridization temperature. Probes with more mismatches hybridize more readily to fragments at lower temperatures, and one promising strategy is touch-down hybridization, where temperatures are sequentially lowered during hybridization (Li *et al.* 2013). Finally, the tiling density of probes across target sequences is another source of variation in target enrichment experiments. This variable is a function of the total desired target length and the total number of unique probes in the probe set, which varies according to the manufacturer and product purchased. Although information regarding the impacts of this parameter is sparse, Ávila-Arcos *et al.* (2011) found no clear differences between 5× and 10.9× tiling densities for enriching ancient plant DNA, suggesting that it may not strongly affect enrichment efficiency.

As large-scale sequencing projects become the norm for data acquisition in nonmodel systems, it is crucial to build a body of literature with standard reporting metrics for both laboratory procedures and data filtering and analysis. At a minimum, we suggest researchers report raw mapping rates to target sequences, PCR duplication rates (ideally based on paired-end reads) and average depths across the different targets, including standard deviations, for a given sequencing effort. Standardized metrics will allow researchers to evaluate whether a particular probe set may work in their study system and how much sequencing may be needed. We hope this study can help set a precedent for such reporting on successful laboratory procedures, including a thorough discussion of efficiency and success of target capture in nonmodel organisms.

Acknowledgements

We thank Randal Voss for the *Ambystoma mexicanum* sequences used to design the capture array and Brant Faircloth for input on experimental design and laboratory troubleshooting. Animal work was conducted under California Department of Fish and Wildlife permit #SC-2480 and associated MOU, USFWS permit #TE-094642-9 and UCLA IACUC protocol #2013-011. This experiment used the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley, supported by NIH S10 Instrumentation

Grants S10RR029668 and S10RR027303. EMM and HBS are supported by NSF-DEB 1257648 and NSF-DEB 1457832.

References

Abdullayev I, Kirkham M, Björklund ÅK, Simon A, Sandberg R (2013) A reference transcriptome and inferred proteome for the salamander *Notophthalmus viridescens*. *Experimental Cell Research*, **319**, 1187–1197.

Arnold B, Corbett-Detig RB, Hartl D, Bomblies K (2013) RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. *Molecular Ecology*, **22**, 3179–3190.

Aronesty E (2013) Comparison of sequencing utility programs. *Open Bioinformatics Journal*, **7**, 1–8.

Ávila-Arcos MC, Cappellini E, Romero-Navarro JA *et al.* (2011) Application and comparison of large-scale solution-based DNA capture-enrichment methods on ancient DNA. *Scientific Reports*, **1**, 74.

Bainbridge MN, Wang M, Burgess DL *et al.* (2010) Whole exome capture in solution with 3 Gbp of data. *Genome Biology*, **11**, R62.

Bankevich A, Nurk S, Antipov D *et al.* (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*, **19**, 455–477.

Barnard R, Futo V, Pecheniuk N, Slattery M, Walsh T (1998) PCR bias toward the wild-type k-ras and p53 sequences: implications for PCR detection of mutations and cancer diagnosis. *BioTechniques*, **25**, 684–691.

Bi K, Vanderpool D, Singhal S *et al.* (2012) Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. *BMC Genomics*, **13**, 403.

Blumenstiel B, Cibulskis K, Fisher S *et al.* (2010) Targeted exon sequencing by in-solution hybrid selection. *Current Protocols in Human Genetics*, **66**, 18.4.1–18.4.24.

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.

Bronner IF, Quail MA, Turner DJ, Swerdlow H (2009) Improved protocols for Illumina sequencing. *Current Protocols in Human Genetics*, doi:10.1002/0471142905.hg1802s62.

Camacho C, Coulouris G, Avagyan V *et al.* (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.

Chilamakuri CSR, Lorenz S, Madoui M-A *et al.* (2014) Performance comparison of four exome capture systems for deep sequencing. *BMC Genomics*, **15**, 449.

Cosart T, Beja-Pereira A, Chen S *et al.* (2011) Exome-wide DNA capture and next generation sequencing in domestic and wild species. *BMC Genomics*, **12**, 347.

Faircloth BC, McCormack JE, Crawford NG *et al.* (2012) Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Systematic Biology*, **61**, 717–726.

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.

Gregory TR (2002) Genome size and developmental complexity. *Genetica*, **115**, 131–146.

Hedtke SM, Morgan MJ, Cannatella DC, Hillis DM (2013) Targeted enrichment: maximizing orthologous gene comparisons across deep evolutionary time. *PLoS One*, **8**, e67908.

Hellsten U, Harland RM, Gilchrist MJ *et al.* (2010) The genome of the western clawed frog *Xenopus tropicalis*. *Science*, **328**, 633–636.

Hodges E, Xuan Z, Balija V *et al.* (2007) Genome-wide *in situ* exon capture for selective resequencing. *Nature Genetics*, **39**, 1522–1527.

Hodges E, Rooks M, Xuan Z *et al.* (2009) Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing. *Nature Protocols*, **4**, 960–974.

Hunter SS, Lyon RT, Sarver BAJ *et al.* (2015) Assembly by reduced complexity (ARC): a hybrid approach for targeted assembly of homologous sequences. *bioRxiv*, 014662. doi: http://dx.doi.org/10.1101/014662

Keinath MC, Timoshevskiy VA, Timoshevskaya NY *et al.* (2015) Initial characterization of the large genome of the salamander *Ambystoma mexicanum* using shotgun and laser capture chromosome sequencing. *Scientific Reports*, **5**, 16413.

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods*, **9**, 357–359.

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

Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv:1303.3997 [q-bio]*.

Li H, Handsaker B, Wysoker A *et al.* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Li C, Hofreiter M, Straube N, Corrigan S, Naylor GJP (2013) Capturing protein-coding genes across highly divergent species. *BioTechniques*, **54**, 321–326.

McCartney-Melstad E (2015) Scripts and data used in "Exon Capture Optimization in Large-Genome Amphibians". Zenodo.

Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA (2007) Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research*, **17**, 240–248.

Neves LG, Davis JM, Barbazuk WB, Kirst M (2013) Whole-exome targeted sequencing of the uncharacterized pine genome. *The Plant Journal*, **75**, 146–156.

Nijman IJ, Mokry M, van Boxtel R *et al.* (2010) Mutation discovery by targeted genomic enrichment of multiplexed barcoded samples. *Nature Methods*, **7**, 913–915.

Nikolenko SI, Korobeynikov AI, Alekseyev MA (2013) BayesHammer: Bayesian clustering for error correction in single-cell sequencing. *BMC Genomics*, **14**, S7.

Rivera MC, Jain R, Moore JE, Lake JA (1998) Genomic evidence for two functionally distinct gene classes. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 6239–6244.

Robertson LS, Cornman RS (2014) Transcriptome resources for the frogs *Lithobates clamitans* and *Pseudacris regilla*, emphasizing antimicrobial peptides and conserved loci for phylogenetics. *Molecular Ecology Resources*, **14**, 178–183.

Sambrook J, Russell DW (2001) *Molecular Cloning: A Laboratory Manual (3-Volume Set)*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

Samuels AK, Weisrock DW, Smith JJ *et al.* (2005) Transcriptional and phylogenetic analysis of five complete ambystomatid salamander mitochondrial genomes. *Gene*, **349**, 43–53.

San Mauro D (2010) A multilocus timescale for the origin of extant amphibians. *Molecular Phylogenetics and Evolution*, **56**, 554–561.

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.

Smith JJ, Kump DK, Walker JA, Parichy DM, Voss SR (2005) 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.

Straus NA (1971) Comparative DNA renaturation kinetics in amphibians. *Proceedings of the National Academy of Sciences of the United States of America*, **68**, 799–802.

Sun C, Mueller RL (2014) Hellbender genome sequences shed light on genomic expansion at the base of crown salamanders. *Genome Biology and Evolution*, **6**, 1818–1829.

Sun Y-B, Xiong Z-J, Xiang X-Y *et al.* (2015) Whole-genome sequence of the Tibetan frog *Nanorana parkeri* and the comparative evolution of tetrapod genomes. *Proceedings of the National Academy of Sciences of the United States of America*, **112**, E1257–E1262.

Yi X, Liang Y, Huerta-Sanchez E *et al*. (2010) Sequencing of 50 human exomes reveals adaptation to high altitude. *Science*, **329**, 75–78.

E.M.M. contributed to the design of the study, performed some of the molecular work, analysed the results and wrote the manuscript. G.G.M. contributed to the design of the study, performed most of the laboratory work and revised the manuscript. H.B.S. contributed to the design of the study and interpretation of results and revised the manuscript. All authors read and approved the final manuscript.

Data accessibility

The data set supporting the results of this article is available at [NCBI SRA:PRJNA285335, SRA:SRP058854]. Accession numbers for individual libraries are shown in Table S1 (Supporting information). The target sequences used for this study, the corresponding *Ambystoma mexicanum*-derived capture probes and the source code used to analyse the data from this experiment are available at http://dx.doi.org/10.5281/zenodo.18587 (McCartney-Melstad 2015).

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Fig. S1. The change in raw mapping rate as a function of post-enrichment qPCR cycle number.

Table S1. Individual libraries (1–53), their treatment levels, and description of yields and sequencing statistics.

Table S2. Post-enrichment concentrations and sequencing efficiency results.

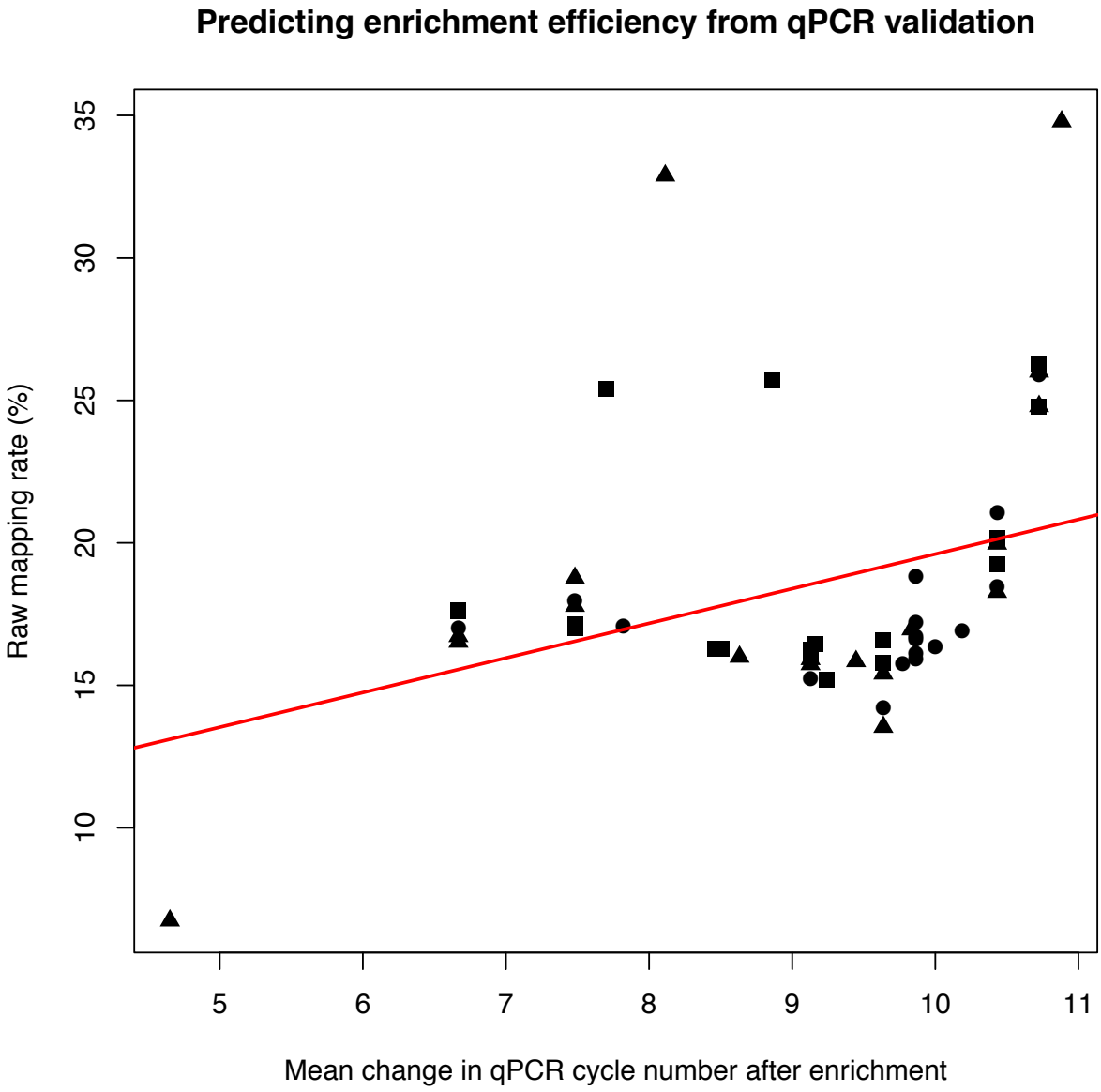

Figure S1: The change in raw mapping rate as a function of post- enrichment qPCR cycle number.

Table S1—Individual libraries (1-53), their treatment levels, and description of yields and sequencing statistics. Number in parenthesis in library (first column) is the enrichment (1-24), and shows how libraries were pooled. For example, 22(18) and 25(18) indicates libraries 22 and 25 were pooled into a single tube (number 18) prior to enrichment.

Library	X c_0t1	Total DNA in Capture (ng)	Individual DNA in Capture (ng)	Sequencing Yield (mb)	% Reads PF	# Reads	% Bases Q >= 30	Mean Quality Score	SRA Accession
1 (1)	1	500	500	527	98.36	3,574,822	88.01	34	SRX1044193
2 (5)	6	500	500	554	98.37	3,752,238	89.7	34.49	SRX1044194
3 (24)	6	500	40	994	98.29	6,739,256	89.91	34.53	SRX1044205
4 (4)	6	500	500	399	98.43	2,702,468	89.57	34.45	SRX1044216
5 (14)	6	1000	1000	426	98.48	2,880,694	88.91	34.26	SRX1044227
6 (24)	6	500	80	2,019	98.36	13,687,874	90.25	34.62	SRX1044238
7 (7)	12	500	500	499	98.45	3,382,090	90.36	34.66	SRX1044242
8 (11)	1	1000	1000	428	98.31	2,903,488	88.21	34.04	SRX1044243
9 (24)	6	500	100	2,096	98.36	14,207,310	90.18	34.6	SRX1044244
10 (10)	1	1000	1000	368	98.24	2,498,034	87.8	33.93	SRX1044245
11 (17)	12	1000	1000	520	98.38	3,523,274	90.58	34.72	SRX1044195
12 (24)	6	500	120	2,507	98.31	16,999,812	90.05	34.57	SRX1044196
13 (13)	6	1000	1000	448	98.54	3,029,216	90.01	34.57	SRX1044197
14 (3)	6	500	500	408	98.54	2,757,646	90.86	34.8	SRX1044198
15 (24)	6	500	60	954	98.52	6,456,486	90.43	34.67	SRX1044199
16 (16)	12	1000	1000	409	98.58	2,764,362	89.99	34.56	SRX1044200
17 (9)	1	1000	1000	357	98.46	2,415,210	88.81	34.21	SRX1044201
18 (2)	1	500	500	404	98.46	2,738,742	90.57	34.72	SRX1044202
19 (12)	6	1000	1000	365	98.45	2,469,008	90.33	34.66	SRX1044203
20 (8)	12	500	500	472	98.58	3,190,188	90.61	34.74	SRX1044204
21 (15)	12	1000	1000	493	98.55	3,337,582	90.94	34.82	SRX1044206
22 (18)	1	500	62.5	659	98.29	4,466,442	88.35	34.08	SRX1044207
23 (6)	12	500	500	476	98.26	3,231,866	90.61	34.74	SRX1044208
24 (19)	6	500	125	937	98.51	6,343,946	90.46	34.68	SRX1044209
25 (18)	1	500	125	1,262	98.34	8,555,454	87.09	33.72	SRX1044210
26 (19)	6	500	62.5	585	98.62	3,956,520	89.72	34.48	SRX1044211
27 (18)	1	500	62.5	527	98.46	3,571,410	86.5	33.57	SRX1044212
28 (20)	12	500	125	1,095	98.53	7,408,314	90.18	34.61	SRX1044213
29 (19)	6	500	125	1,254	98.34	8,498,554	89.44	34.41	SRX1044214
30 (20)	12	500	62.5	481	98.37	3,256,714	90.11	34.6	SRX1044215
31 (19)	6	500	62.5	597	98.33	4,044,512	89.85	34.52	SRX1044217
32 (18)	1	500	125	869	98.42	5,886,378	88.81	34.21	SRX1044218
33 (20)	12	500	125	1,067	98.45	7,228,628	89.96	34.56	SRX1044219
34 (19)	6	500	125	1,120	98.57	7,573,524	89.83	34.51	SRX1044220
35 (20)	12	500	62.5	440	98.56	2,976,292	88.85	34.25	SRX1044221
36 (20)	12	500	125	1,247	98.49	8,439,750	89.72	34.49	SRX1044222
37 (18)	1	500	125	777	98.49	5,262,556	87.91	33.96	SRX1044223
38 (21)	1	1000	250	1,056	98.32	7,162,990	88.48	34.11	SRX1044224
39 (23)	12	1000	250	1,213	98.38	8,218,358	89.81	34.51	SRX1044225
40 (22)	6	1000	250	1,222	98.46	8,271,852	89.66	34.47	SRX1044226
41 (21)	1	1000	250	1,277	98.26	8,660,718	88.3	34.07	SRX1044228
42 (23)	12	1000	250	1,281	98.45	8,673,266	90.72	34.76	SRX1044229
43 (21)	1	1000	125	451	98.36	3,059,160	88.77	34.2	SRX1044230
44 (21)	1	1000	250	1,273	98.3	8,633,572	87.49	33.84	SRX1044231
45 (22)	6	1000	250	1,041	98.42	7,048,962	89.81	34.51	SRX1044232
46 (21)	1	1000	125	501	98.32	3,400,050	88.07	34.01	SRX1044233
47 (22)	6	1000	125	481	98.38	3,256,798	89.4	34.39	SRX1044234
48 (22)	6	1000	250	1,061	98.3	7,193,568	89.62	34.46	SRX1044235
49 (23)	12	1000	250	1,228	98.39	8,319,782	90.38	34.67	SRX1044236
50 (22)	6	1000	125	636	98.62	4,296,942	89.56	34.44	SRX1044237
51 (23)	12	1000	125	483	98.61	3,263,134	90.42	34.68	SRX1044239
52 (23)	12	1000	125	568	98.28	3,851,216	89.59	34.46	SRX1044240
53 (24)	6	500	20	427	98.43	2,889,856	90.06	34.58	SRX1044241

Table S2—Post-enrichment concentrations and sequencing efficiency results. Number in parenthesis is library name as in Table S1.

Library #	Amount of DNA after post-enrichment PCR (ng)	Average change in qPCR cycle #	PCR duplication rate	Raw mapping rate	Unique read mapping rate	Average depth across target	Highest 100bp window ave. depth
1 (1)	74.17	9.2388	22.24%	15.19%	11.81%	6.27	7.68
2 (5)	47.83	8.113	16.70%	32.89%	27.40%	18.63	21.76
3 (24)	350.64	9.864	30.26%	16.72%	11.66%	6.58	7.74
4 (4)	32.16	7.6988	16.21%	25.41%	21.29%	13.22	15.85
5 (14)	99.54	10.883	11.39%	34.79%	30.83%	20.89	24.45
6 (24)	350.64	9.864	33.19%	16.63%	11.11%	6.32	7.31
7 (7)	347.14	8.505	10.58%	16.28%	14.55%	7.70	9.27
8 (11)	216.71	8.632	15.23%	16.00%	13.56%	7.74	9.42
9 (24)	350.64	9.864	33.57%	15.93%	10.59%	5.94	6.86
10 (10)	165.66	9.1638	16.33%	16.45%	13.76%	7.54	9.31
11 (17)	379.10	9.446	9.21%	15.84%	14.38%	7.78	9.30
12 (24)	350.64	9.864	33.72%	16.14%	10.69%	5.99	6.88
13 (13)	74.65	8.863	11.91%	25.69%	22.63%	13.97	16.59
14 (3)	81.97	7.82	11.59%	17.08%	15.10%	8.73	10.62
15 (24)	350.64	9.864	29.10%	18.83%	13.35%	7.86	9.20
16 (16)	439.73	8.461	9.11%	16.27%	14.79%	7.68	9.32
17 (9)	172.38	10.187	14.21%	16.92%	14.52%	8.65	10.60
18 (2)	2150.07	4.652	19.62%	6.74%	5.42%	0.27	0.47
19 (12)	161.07	10.435	8.73%	21.05%	19.22%	11.91	14.30
20 (8)	250.64	9.833	8.84%	16.96%	15.46%	8.52	10.19
21 (15)	269.70	9.999	8.51%	16.34%	14.95%	8.43	10.09
22 (18)	57.15	6.668	47.05%	16.72%	8.86%	4.69	5.72
23 (6)	439.78	9.773	8.75%	15.76%	14.38%	8.16	9.80
24 (19)	117.98	9.637	31.86%	13.54%	9.22%	4.45	5.36
25 (18)	57.15	6.668	48.59%	17.62%	9.06%	4.74	5.64
26 (19)	117.98	9.637	30.90%	15.41%	10.65%	5.46	6.63
27 (18)	57.15	6.668	47.69%	17.59%	9.20%	4.66	5.76
28 (20)	93.68	10.432	19.83%	18.28%	14.65%	8.37	9.72
29 (19)	117.98	9.637	33.69%	16.59%	11.00%	5.67	6.70
30 (20)	93.68	10.432	17.64%	19.97%	16.44%	9.67	11.49
31 (19)	117.98	9.637	31.14%	15.79%	10.87%	5.39	6.55
32 (18)	57.15	6.668	46.67%	17.01%	9.07%	5.08	6.15
33 (20)	93.68	10.432	21.03%	19.26%	15.21%	8.55	9.96
34 (19)	117.98	9.637	32.57%	14.22%	9.59%	4.96	5.90
35 (20)	93.68	10.432	16.78%	20.18%	16.79%	9.60	11.49
36 (20)	93.68	10.432	20.87%	18.46%	14.61%	8.67	10.05
37 (18)	57.15	6.668	46.14%	16.53%	8.90%	4.82	5.83
38 (21)	252.29	9.128	28.10%	15.24%	10.96%	6.09	7.24
39 (23)	364.13	7.482	17.49%	17.00%	14.03%	7.64	8.91
40 (22)	70.10	10.725	38.12%	25.89%	16.02%	10.11	11.74
41 (21)	252.29	9.128	29.75%	15.74%	11.05%	6.03	7.10
42 (23)	364.13	7.482	16.69%	17.99%	14.98%	8.82	10.22
43 (21)	252.29	9.128	25.29%	15.92%	11.89%	6.36	7.78
44 (21)	252.29	9.128	30.89%	16.01%	11.06%	5.75	6.81
45 (22)	70.10	10.725	35.17%	26.01%	16.86%	10.52	12.25
46 (21)	252.29	9.128	27.01%	16.25%	11.86%	6.15	7.53
47 (22)	70.10	10.725	34.28%	24.80%	16.30%	10.11	12.06
48 (22)	70.10	10.725	37.39%	24.78%	15.52%	9.33	10.92
49 (23)	364.13	7.482	16.22%	17.79%	14.90%	8.48	9.81
50 (22)	70.10	10.725	36.08%	26.28%	16.80%	10.22	12.10
51 (23)	364.13	7.482	11.98%	18.77%	16.52%	9.45	11.23
52 (23)	364.13	7.482	14.06%	17.15%	14.74%	7.83	9.35
53 (24)	350.64	9.864	26.20%	17.21%	12.70%	6.98	8.46

Population Genomics of Endangered Tiger Salamanders (*Ambystoma tigrinum*) on Long Island, NY Reveals a Highly Structured Species Impacted by Major Roads

Evan McCartney-Melstad[1*], Jannet Vu[1,2], H. Bradley Shaffer[1]

1: Department of Ecology and Evolutionary Biology, La Kretz Center for California Conservation Science, and Institute of Environmental Studies, University of California, Los Angeles
2: Department of Ecology and Evolutionary Biology, Stony Brook University, New York

Abstract

We used DNA sequence data from thousands of nuclear loci to characterize the population structure of endangered tiger salamanders (*Ambystoma tigrinum*) on Long Island and quantify the impacts of human development on this species. We uncovered highly genetically structured populations over an extremely small spatial scale (approximately 40 km^2) in an increasingly human-modified landscape. Geographic distance and the presence of major roads between ponds are both strong predictors of genetic divergence in this system, which suggests both natural and anthropogenic factors are responsible for the observed patterns of genetic variation. This study demonstrates the added value of genomic approaches in molecular ecology, as these patterns were not apparent in an earlier study of the same system using microsatellite loci. Ponds exhibited small effective population sizes, and there is a strong correlation between pond surface area and salamander population size. When combined with the high degree of structuring in this heavily modified landscape, our study suggests that these endangered amphibians require management at the individual pond, or pond cluster, landscape level. Particular efforts should be made to preserve large vernal pools, which harbor greater genetic diversity, and their surrounding upland habitat. Contiguous upland landscapes between ponds that encourage natural metapopulation dynamics and demographic rescue from future local extirpations should also be protected.

Introduction

Genetic, and, increasingly, genomic analyses constitute a powerful tool kit for understanding how species move through landscapes, particularly for secretive species such as reptiles and amphibians (Shaffer *et al.* 2015). When studying endangered species, we are often concerned with the degree to which human activity has impacted the size and movement of populations. This human interference often occurs at very small spatial scales compared to species range

sizes—for example, building a road between two nearby populations that exchange migrants regularly—as well as short temporal scales, given that humans often have been impacting wildlife populations for tens or hundreds of generations. As conservation and resource managers and as population biologists, we are often less interested in larger scale effects across thousands of kilometers of a species range than we are about dynamics across a few kilometers on specific landscapes. This is especially true for low-vagility species like amphibians, reptiles, small mammals, and many invertebrates that often move a kilometer or less per generation (Blaustein *et al.* 1994). For such taxa, the genetic relationships between distant populations are often a result of ancient demographic processes, but interruption of gene flow at an extremely fine spatial scale is the defining component of human impacts. For protected or endangered species, understanding the extent to which human activities at the finest spatial scales alter demographic and population processes is the key to effective management.

Discerning gene flow and differentiation at very fine spatial scales is challenging because populations located proximately to one another tend to be very closely related (Wright 1943). Furthermore, the ability to detect differentiation between genetically very closely related populations is limited by the number of samples and genetic loci assayed (Patterson *et al.* 2006). Until now, nearly all population genetic studies of amphibians have been limited to mitochondrial DNA or a small number of nuclear loci (typically microsatellites). This is due at least in part to the large, highly repetitive genomes of many amphibians that make it difficult to generate genomic resources (Licht & Lowcock 1991; Sun *et al.* 2012). While this is slowly changing as genomic technologies are beginning to be being applied to amphibians (Keinath *et al.* 2016; McCartney-Melstad *et al.* 2016; Portik *et al.* 2016; Newman & Austin 2016), most systems that could benefit from genomic scale data remain unexplored. Custom target

enrichment assays built from transcriptomic resources are promising intermediate solutions that bridge the gap between microsatellites and whole genome sequencing while allowing for flexibility in which genomic regions to study (McCartney-Melstad *et al.* 2016; Portik *et al.* 2016).

One interesting case study where the added resolution of genomic-scale datasets may make a difference is for tiger salamanders (*Ambystoma tigrinum*) on Long Island, a New York-listed endangered species (6 CRR-NY 182.5) where fine-scale population dynamics are critical for management decision making. *A. tigrinum* was historically found in scattered localities across New York at the northern limits of its range in the eastern US, including Albany County, Rockland County, and across Long Island. However, the species has experienced dramatic declines in the region, and it is currently restricted to Suffolk and Nassau Counties, primarily in central Long Island (Bishop 1941; Stewart & Rossi 1981). In recent years, surveyors have witnessed a decrease in the observed number of individuals, with approximately 90 breeding ponds remaining (New York State Department of Environmental Conservation 2015).

The species suffers a range of threats including disease, predation, pollution, invasive species, and climate change-induced sea level rise. Development is not only a source of habitat loss, but also creates direct mortality risk from road kill, degrades pond viability from pollutants, and creates barriers to migration and population fragmentation (Titus *et al.* 2014). Telemetry studies have documented individuals traveling at least 500 meters from breeding ponds, and confirmed that individuals tend to avoid paved roads, dirt roads, and grassy areas (Madison & Farrand 1998). Movements, which are often studied during the annual breeding migration, are generally oriented towards upland refugia in their preferred habitat of sandy soil, pine barren habitat (Madison & Farrand 1998; Titus *et al.* 2014).

Prior genetic work using twelve microsatellite loci recovered two distinct populations of *A. tigrinum* across 17 ponds spanning 50 km on Long Island, both of which exhibited low diversity and high relatedness among ponds (Titus *et al.* 2014). The authors attributed the low diversity and high relatedness to post-glacial colonization from North Carolina (Church *et al.* 2003) and relatively frequent migration of salamanders between ponds. Their primary conclusion was that Long Island and New Jersey tiger salamanders were genetically uniform within each state, but were differentiated between states due to geographic isolation and range fragmentation.

Most of the ponds analyzed by Titus *et al.* (2014) on Long Island were fewer than six kilometers apart, and their analyses and conclusions required genetic markers capable of discerning fine scale ecological processes. However, the microsatellite loci used showed relatively low diversity (1-13 alleles per locus across ponds and an average of 1-3 alleles per locus within ponds), and therefore were not the most imformative (Reyes-Valdés 2013). This leaves open the real possibility that these markers lacked the statistical power to detect real patterns of landscape-driven differentiation. This was not a fault of the Titus *et al.* (2014) work, but rather a reflection of the tools available when their work was undertaken.

To explore recent anthropogenic impacts on this endangered, fragmented set of populations further, we applied a genomic target capture approach with 5,237 random nuclear exons to ponds in the same system to quantify the degree to which ponds are isolated from one another and whether or not major roads act as barriers to dispersal for extant populations of *Ambystoma tigrinum* on Long Island. We sought to answer three separate questions: 1) To what degree are ponds genetically connected to or differentiated from one another?, 2) what are the effective population sizes of ponds in the system, are they related to pond area, and how do these values compare to other amphibians?, and 3) what are the effects of roads on connectivity between

ponds in the system? The increased resolution recovered from the genomic dataset collected here demonstrates the increased power and utility of genomic-scale data for population genetics of threatened species, and highlights the fundamentally different conclusions for appropriate management interventions that such data can provide.

Methods

Sampling and Data Generation

Larval tissue samples were collected in Suffolk County over three consecutive breeding seasons between 2013 and 2015 using seines and dipnets. We timed our sampling to occur in the late spring when larvae were large enough to sample non-destructively with small tail clips (Polich *et al.* 2013). Tail tips were placed in 95% ethanol within 30 seconds of clipping, larvae were immediately released at the site of capture, and tail tips were stored at -80C until use. A hand-held GPS unit was used to locate ponds in the field, and final spatial coordinates and areas of ponds were taken from tracings of Google Earth images from March 2007. We sampled larvae from multiple sites at each pond to randomly sample the genetic variation present. DNA was extracted from samples using a salt extraction protocol (Sambrook & Russell 2001), diluted to 100 ng/μL, and sheared for 28 cycles (30s on, 90s off) using the "high" setting on a Bioruptor NGS (Diagenode). After shearing, samples were dual-end size selected to approximately 300-500bp using 0.8X-1.0X SPRI beads (Rohland & Reich 2012).

Libraries were prepared with 419-2000 ng of starting input DNA using Kapa LTP library prep kit half reactions (Kapa Biosystems, Wilmington MA). Libraries were dual-indexed using the iTru system (Glenn *et al.* 2016), which adds 8bp indices to the adapters of both ends of library fragments for demultiplexing. Next, 500ng of each library were combined into pools of 8 (4,000ng total input DNA) and enriched using a MYcroarray (Ann Arbor, MI) biotinylated RNA probe set designed from 5,237 exons from unique genes from the California tiger salamander

genome (McCartney-Melstad *et al.* 2016). Given the relatively close phylogenetic relationships of all members of the tiger salamander complex (Shaffer & McKnight 1996; O'Neill *et al.* 2013), we predicted that most of the probes would also capture the eastern tiger salamander homolog. A total of 30,000 ng of c_0t-1 prepared from *Ambystoma californiense* was used for each capture reaction to block repetitive DNA from hybridizing with probes or captured fragments. Probes were hybridized for 30 hours at 60C, bound to streptavidin-coated beads, and washed four times with wash buffer 2.2 (MYcroarray). Enriched libraries were then amplified on-bead with 14 cycles of PCR, cleaned using 1.0X SPRI beads, and sequenced on three 150bp PE lanes on an Illumina HiSeq 4000.

Reference Assembly

We built a reference assembly for read mapping and SNP calling using the Assembly by Reduced Complexity (ARC) pipeline (Hunter *et al.* 2015). To do this, the reads from the 10 samples that received the greatest number of reads were pooled and mapped to the 5,237 *A. californiense* targets across which capture probes were tiled using bowtie2 v.2.2.6 (Langmead & Salzberg 2012). Pools of reads mapping to each one of these targets were independently assembled using SPAdes v.3.8.2 (Bankevich *et al.* 2012), and the contigs assembled for each target then replaced their respective targets and another round of mapping was performed to these contigs. This process was repeated for 10 iterations to extend assembled targets several hundred bp in both directions from their central probe-tiled regions. Reciprocal best blast hits (RBBHs) were then found to represent each target locus using blast+ 2.2.30 (Camacho *et al.* 2009). The set of RBBHs was then blasted against itself to find similar regions among targets, which may be indicative of chimeric assemblies. Regions within each RBBH that were found to be similar to other RBBHs were trimmed to the ends of the RBBH contigs.

SNP Calling and Genotyping

Reads for all samples were trimmed to 150bp (if the 151st base was reported by the sequencing facility) and adapters were trimmed using skewer 0.1.127 (Jiang *et al.* 2014). These trimmed reads were then mapped to the reference assembly using BWA-mem 0.7.15 (Li 2013). Read group information was added to the aligned reads and PCR duplicates were marked using picard tools v2.0.1 (https://broadinstitute.github.io/picard/).

SNP calling and genotyping was performed according to GATK best practices (DePristo *et al.* 2011; Van der Auwera *et al.* 2013). First, a set of high-quality reference SNPs was generated to assess and recalibrate base quality scores within each sample. HaplotypeCaller from GATK nightly-2016-11-21-g69e703d (McKenna *et al.* 2010) was run separately on each sample in GVCF mode followed by joint genotyping with GenotypeGVCFs. Then, any SNP that met any of the following criteria were removed from the reference set: QD < 2.0, MQ < 40.0, FS > 60.0, MQRankSum < -12.5, ReadPosRankSum < -8.0, QUAL < 100. Similarly, any indel that failed any of the following criteria were also removed from the reference set: QD < 2.0, SOR > 10.0, FS > 60.0, ReadPosRankSum < -8.0, QUAL < 100. Base quality score recalibration was then performed at the lane level (three different platform units among all of the read groups) using GATK.

HaplotypeCaller in GATK was then used with recalibrated reads to generate sample-level GVCF files that were jointly genotyped using GATK's GenotypeGVCFs function. The same hard filters outlined above were then applied to the resulting VCF files, except that all SNPs with QUAL values above 30 (instead of 100) were kept. Genotype calls with phred-scaled quality scores under 20 (1 in 100 chance of being incorrect) were set to "missing" data, and SNPs with greater than 50% missing data were removed. Samples with missing data rates greater than 30% were also removed.

Given the extremely large genomes of ambystomatid salamanders (roughly 30GB) (Licht & Lowcock 1991; Keinath *et al.* 2015), we were concerned about the possibility of including duplicated paralogous loci in our analyses. We attempted to correct for this by filtering out loci that contained excessive heterozygosity, as fixed differences between true paralogs interpreted as homologs will typically appear as variable sites that are always heterozygous. To do this, VCFtools v.0.1.15 was used to calculate p-values for heterozygote excess for every SNP (Wigginton *et al.* 2005; Danecek *et al.* 2011). Target regions that contained at least one SNP with an excess heterozygote p-value below 0.001 were removed from the analysis. A set of SNPs was then generated by randomly choosing a single SNP from each qualifying target region (those targets that did not contain any excessively heterozygous SNPs). This dataset with a single SNP taken from each target region is referred to hereafter as the "linkage-pruned" dataset.

Population Genetic Analysis

The presence of isolation by distance (IBD)—the relationship between geographic and genetic distance—was tested at both the individual and pond (population) levels. Individual genetic similarity was calculated as the percentage of SNPs that were identical-by-state using SNPRelate v1.6.4 (Zheng *et al.* 2012). These values were regressed on geographic distance and the significance of the correlation between genetic distance and geographic distance was tested using a simple Mantel test with 999,999 permutations in the R package vegan 2.4-0 (Mantel 1967; Oksanen *et al.* 2016). At the pond level, Fst/(1-Fst) (Slatkin 1995) was calculated using SNPRelate v1.6.4 and regressed on geographic distance to estimate the slope of isolation by distance. Rousset (1997) recommends regressing Fst/(1-Fst) on the logarithm of geographic distance in the case of two-dimensional habitats or non-transformed geographic distance in the case of one-dimensional habitats. Since the sampling area for this study is very narrow and is over three times longer than it is wide (approximately 15.5 km x 4.5 km), it is unclear whether it

is more appropriate to treat the study area as linear or two dimensional, and regressions and Mantel tests are reported for both raw and log-transformed geographic distances. Fst values were also calculated using Arlequin v3.5.2.2 (Excoffier & Lischer 2010) to determine significance p-values using 100,172 permutations of the data. P-values from Arlequin were adjusted for multiple testing using the Benjamini-Yekutieli correction implemented in base R (Benjamini & Yekutieli 2001). For individual-based analyses, logarithms of geographic distances were set to a minimum value of 0.

We were interested in characterizing the level of genetic diversity present in tiger salamanders on Long Island. To estimate genetic diversity we determined per-base pair Watterson's θ, an estimator that characterizes the level of genetic diversity in populations based on the number of segregating sites per base pair sequenced (Watterson 1975). We calculated θ for each pond with samples pooled across years. As a basis of comparison, a population sample of 15 California tiger salamanders (*A. californiense*) from a single pond in Great Valley Grasslands State Park, California (McCartney-Melstad and Shaffer, unpublished data) was genotyped under similar filtering parameters for the same set of loci, and θ was estimated for this group in the same way.

The linkage-pruned dataset was visualized using principal components analysis (PCA) in the R package SNPRelate v1.6.4 (Zheng *et al.* 2012). The first eight principal components were plotted with letters corresponding to the collection sites of samples. The proportion of the variance explained by each principal component was also obtained using SNPRelate v1.6.4.

To estimate the number of distinct population clusters in the data, ADMIXTURE v1.3.0 was run using the linkage-pruned dataset containing all samples from all ponds across all three years of sampling for K=1 to K=30 with ten different random number seeds (Alexander *et al.* 2009). Each replicate was subjected to 100-fold cross validation, and CV errors were used to choose a

"reasonable" set of K values. If the standard deviation of CV values for any K value overlapped with the standard deviation of the best-scoring K value, it was included as a reasonable value for K.

Effective population sizes (Ne) for each pond were estimated using the linkage disequilibrium (LD) method in NeEstimator v2.01 with a minor allele frequency cutoff of 0.05 (Hill 1981; Do *et al.* 2014). Estimates were calculated for all cohorts (a given pond in a given year), and, when more than one year of sampling was conducted for a pond, Ne was also calculated for the pooled sample of either two or three cohorts. LD-based estimates of effective population size from single cohorts represent the harmonic mean between the effective number of breeders (Nb) and the true effective population size (Ne) (Waples *et al.* 2016). Alternatively, as the number of pooled cohorts approaches the generation length (the average age of parents for a cohort), LD-based estimators should approach the true Ne (Waples & Do 2010; Waples *et al.* 2014).

Effective population size estimates using the LD method can be downwardly biased for multiple reasons. First, estimates may be biased when many loci are used due to physical linkage among loci, given that the method assumes the loci being used are unlinked (Waples *et al.* 2016). This effect is predictable, however, and can be corrected if the number of chromosomes or total linkage map length is known. Estimates of linkage map length for the closely related axolotl, *Ambystoma mexicanum*, are known, and this number (4200cm) was used to correct estimates of effective population size for dense locus sampling by dividing them by 0.9170819 (which is equal to -0.910 + 0.219 x ln(4200)) (Voss *et al.* 2011; Waples *et al.* 2016).

LD based estimates of effective population size can also be downwardly biased when analyzing mixed cohorts in iteroparous species such as *A. tigrinum*, although this bias appears to decrease as the number of sampled cohorts approaches the generation length of the species

(Waples & Do 2010; Waples *et al.* 2014). Therefore, single-cohort estimates of Ne were further corrected by dividing dense-locus adjusted estimates by 0.8781801, the product of two equations from Table 3 of Waples et al. (2014) that use the ratio of adult lifespan (estimated at 7 years for the closely related *A. californiense*) to age at maturity (4 years, also in *A. californiense*) (Trenham *et al.* 2000) to compensate for the downward bias introduced by iteroparity: (1.103-0.245 * log(7/4)) * (0.485+0.758 x log(7/4)). For ponds in which multiple years of sampling were conducted, we report both pooled-cohort estimates (corrected for dense locus sampling) and per-cohort estimates (corrected both for dense locus sampling and single-cohort sampling). We used linear regression to visualize the relationship between pond area (as traced from Google Earth images) and effective population size, using multi-year estimates of Ne when available.

Impact of Roads

We were interested in assessing to what degree human habitat modifications have restricted movement of this species, and whether or not human activity has contributed to the observed patterns of population structure. To explore this, we created a matrix that indicated whether or not pairs of ponds were separated by a major road (New York State Route 25, Suffolk CR 46, or Interstate 495, see Figure 6). This matrix was included as a predictor variable for genetic distance in linear regression and was tested for correlations to genetic distance (while controlling for geographic distance) using a partial Mantel test with vegan v2.4-0 in R (Mantel 1967; Smouse *et al.* 1986; R Core Team 2015; Oksanen *et al.* 2016).

Results

Sampling: A total of 283 salamanders were genotyped from 17 ponds spread over an approximately 40 km^2 area (Figure 6, Table 4). More than 1.9 billion 150-bp sequencing reads were generated from three Illumina HiSeq 4000 lanes across these samples (mean=6.8 million reads/sample, min=1.8 million reads, max=10.9 million reads).

Reference assembly: The ten samples that received the most sequencing reads were pooled to generate a *de novo* reference assembly, for a total of 66.9 million merged and paired-end sequencing reads (11.7 billion total bp). Assembly of target regions with the ARC assembler produced a set of 74,109 contigs (47.5 million bp) from which 5,057 reciprocal best blast hits were recovered (6.7 million bp). After blasting these contigs against themselves, trimming self-complementary regions to the ends of contigs, and re-determining reciprocal best blast hits, a 6.6 million bp assembly with 5,050 target regions (96.4% of the originally targeted regions) was recovered for mapping reads and calling SNPs.

SNP Calling and Genotyping: An average of 29.27% of raw reads mapped to the reference assembly using BWA-mem across all 283 samples (sd=2.47%, min=20.33%, max=34.30%). After removing PCR duplicates (read pairs that map to the exact same position on the reference, indicating that they may be PCR amplicons from the same molecule), an average of 17.03% unique reads mapped to the reference (sd=2.47%, min=8.51%, max=24.59%). After joint genotyping, a total of 82,005 raw SNPs were recovered across 4,400 target regions. After applying hard filters to SNP loci, setting the minimum genotype call quality to 20, discarding variants genotyped in less than 50% of all samples, and removing the one sample with a missing data rate greater than 30%, a total of 21,998 biallelic SNPs were retained across 3,631 target regions. Tests for Hardy Weinberg equilibrium revealed 533 targets contained at least one SNP with clear ($p<0.001$) heterozygote excess, which is consistent with (though not definite evidence of) the presence of an unknown paralogous copy of this gene in the genome. After removing these target regions from the analysis, a total of 12,924 biallelic SNPs remained across 3,098 target regions. The final matrix containing 282 individuals had a mean missing data rate of 7.7%

(max=27.8%, min=1.8%, sd=4.5%). The linkage-pruned dataset contained one random biallelic SNP from each final target, for a total of 3,098 variants.

Genetic variation within cohorts: Values of Watterson's θ for ponds ranged from $3.26x10^{-4}$ to $5.77x10^{-4}$ (Table 4), and was $3.19\ x10^{-4}$ after pooling the 282 samples from all ponds together for a single estimate of θ. The comparative sample of 15 *A. californiense* from a pond in Merced County, CA had a θ value of $7.09x10^{-4}$, which was higher than each of the values calculated for ponds in Long Island *A. tigrinum.* This suggests that genetic diversity is lower for *A. tigrinum* in Long Island than it is for *A. californiense* in Great Valley Grasslands State Park, CA, and is in keeping with the low estimates of variation found by Titus *et al.* (2014).

Isolation by Distance (IBD): IBD was apparent at both the individual and pond level (Figures 7 and 8, Table 5). Regressions of individual identity-by-state on both raw and log-transformed geographic distances yielded negative relationships with p-values below $2x10^{-16}$ (Figure 7). Adjusted R^2 values were higher for log-transformed distances when comparing pairwise individual genetic relationships and geographic distances (0.2861 vs. 0.1764). Similarly, regression coefficients were positive and highly significant when testing for the relationship between pairwise Fst of ponds and raw and log-transformed geographic distances (Figure 8, $p < 2.6x10^{-16}$ and $p < 4.12x10^{-11}$ for raw and log-transformed distances, respectively). Unlike the individual-based measure, the pond-based model with raw geographic distances fit the data better (R^2=0.39) than log-transformed geographic distances (R^2=0.27). Testing the significance of isolation by distance using regression coefficient p-values is inappropriate because many of the pairwise observations are not independent. Therefore, simple Mantel tests were used to test the significance of correlations between pond/individual genetic and raw/log-transformed geographic distances, all of which yielded p-values lower than 0.000011 (Table 5). This indicates

that there is a significant relationship between geographic and genetic distance, even at the extremely fine scale studied here.

Pairwise Fst values between ponds ranged from 0.005 to 0.207 (136 comparisons, median=0.064, sd=0.042, Table 6). Using Benjimini-Yekutieli (BY)-corrected p-values, 118 out of 136 of these pairwise comparisons were significantly different from 0. Of the 18 non-significant pairwise comparisons, 16 were from pond L, which contained only a single sample and therefore had extremely low power. Many of the highest Fst values are from pairwise comparisons containing ponds A or Q. These ponds are both outliers separated by greater geographic distances and by major roads from all other ponds (Figure 6).

Principal Component Analysis: The first eight principal components (PCs) are shown as pairwise plots in Figure 9. In all PC graphs, samples are coded by letters representing the ponds from which they were collected (Figure 6). PC1 groups samples from pond A to the exclusion of the other samples, while PC 2 does the same for samples from ponds E, F, and G. PC 3 separates samples from ponds B, C, and D from the other ponds (especially pond N), and PC4 appears to be an axis of variation between ponds J and Q (which is also apparent in PC5). Finally, PCs 6, 7, and 8 correspond to axes that differentiate ponds N, P, and Q, along with some samples from ponds A and J. Overall, clustering of single ponds and small groups of closely adjacent ponds is quite apparent, which indicates the presence of easily detectable population structure with the genomic data that we have collected in this study.

Population Clustering: The value of K in ADMIXTURE with the lowest mean CV error was K=12. Four other K values (9, 10, 11, and 13) had CV error standard deviations that overlapped with K=12 (Figure 10). Admixture proportions for K=9 through K=13 are shown in Figure 11, and are split by both pond and sampling year (Glasbey *et al.* 2007). Results from ADMIXTURE

analyses corroborated the qualitative patterns observed in the PCA. First, pond A generally formed one to three clusters to the exclusion of all other ponds (as recapitulated in PCs 1, 6, and 8). Ponds B, C, and D form a single cluster to the exclusion of other ponds (as also seen in PC 3). Similarly, ponds E and G form a unique cluster at K=9 (corresponding to PC 2), but are separated into their own private clusters at K=10 through K=13. Pond F, geographically separated from its closest neighbors (ponds E and G) by NY State Route 25, appears strongly admixed at K=9 through K=12, and receives its own cluster at K=13. Ponds H, I, J, K, L, and M appear to be strongly associated across all K values (though ponds I, L, and M appear highly admixed at these K values), with the exception of one year of sampling in pond J (2014) that produced a group of animals that formed their own cluster. Pond N appears quite distinct across all K values (which can also be seen on PCs 3-8). Pond O appears highly admixed across all K values, but tends to share a considerable admixture component with the cluster formed by pond P (and pond Q for K=9 through K=11). At K=12 and K=13, pond Q forms its own strong cluster to the exclusion of all other ponds, a pattern that is also quite apparent in PC5.

Effective Population Size: Estimates of effective population size ranged from 10.3 for pond N to 135.0 for pond K (Table 4). For ponds with multiple years of sampling, single-cohort estimates were generally close to those for pooled-cohort, with the exception of pond O, which had a pooled-cohort estimate of 68.8 and a 2013-cohort estimate of 17,689. This single-cohort estimate was extremely sensitive to the minor allele frequency cutoff—changing the threshold to 0.10 from 0.05 lowered the estimate to less than 600. The 95% confidence interval was also extremely wide for this cohort estimate, ranging from 953.0 to Infinite/incalculable. The surface area of ponds was strongly correlated with effective population size estimates ($p=0.00122$, $R^2=0.5619$, Figure 12). The number of samples included in the calculation of Ne was not

correlated with the resulting Ne estimate (linear regression p=0.513, adj R^2= -0.0438), suggesting that sample size *per se* was not a driver of Ne estimates.

Roads as Barriers to Dispersal: Roads appear to play a strong role in structuring among-pond genetic divergence in Long Island tiger salamanders. Specifically, linear regression supports roads as an explanatory factor in pairwise Fst values between ponds, as adding this term increased the adjusted R^2 of models including only geographic distance from 0.39 to 0.68 (with both terms highly significant). This is apparent from visualizing the distances, as a distinct upwards shift in genetic distance is apparent for pairwise comparisons separated by major roads (Figure 13). Similarly, partial Mantel tests recovered strong and highly significant correlations between genetic distance and being separated (or not) by major roads after controlling for geographic distance (p=0.000608, Mantel R^2=0.48). This suggests that dispersal may be limited across major roads, and that human activity has contributed to isolation of ponds in this relatively highly developed region.

Discussion

Population structure is difficult to detect and quantify accurately in subtly differentiated populations, and populations in close geographic proximity tend to be subtly differentiated (Wright 1943). In conservation genetics, however, we are often interested in understanding limitations in gene flow at the temporal and spatial scales at which humans impact populations. Furthermore, as the number of generations over which humans have affected most populations is usually relatively small, many cases of human-induced structure will be difficult to detect with conventional genetic datasets.

Several amphibian studies have attempted to quantify spatial genetic structure of populations at very fine spatial scales. Jehle *et al.* (2005) found evidence of pond clustering in *Triturus* newts over a 26.5 km^2 landscape using a hierarchical Bayesian clustering algorithm (Corander *et al.*

2003), although ponds did not cluster cleanly in STRUCTURE analyses (Pritchard *et al.* 2000). Hitchings and Beebee (1997) used allozyme data in common frogs in the UK and found evidence for significant structuring over a few kilometers in urbanized environments, but not in rural environments, suggesting that human development was acting to isolate ponds from one another in this system. Similarly, Lampert et al. (2003) recovered significant isolation by distance over roughly 8km between ponds in Túngara frogs (*Physalaemus pustulosus*), although 51 of 64 pairwise Fst values on the same side of the 100m-wide Chagres River were non-significant, and no population clustering methods were attempted. Conversely, Newman and Squire (2001) recovered significant differentiation and isolation by distance in wood frogs (*Rana sylvatica*) ponds separated by roughly 20km but could not genetically differentiate ponds at closer distances. Lampert *et al.* (2003) attributed the differences in discriminating power between these two studies to the low levels of diversity in microsatellite loci for wood frogs. Zamudio and Wieczorek (2007) found evidence for two genetic clusters of *Ambystoma maculatum* from 29 ponds spread over 1272km^2 in upstate New York, but little support for substructuring among ponds within each cluster. A number of other studies have found strong support for population structure among breeding ponds of amphibians in small landscapes using microsatellite loci (Wang *et al*. 2009, 2011, Wang 2009b, 2012; Savage *et al*. 2010). Conversely, several amphibian studies using microsatellites have failed to find significant genetic differentiation among ponds for pond-breeding amphibians (Coster *et al.* 2015; Furman *et al.* 2016), while others have found evidence of isolation by distance and limited clustering (Sotiropoulos *et al.* 2013; Peterman *et al.* 2015).

These studies illustrate that, in amphibians, genetic differentiation is sometimes detectable at very fine spatial scales, and sometimes it is not. This may hinge largely on the variability of the

markers studied, which itself is shaped by deeper-time demographic processes such as bottlenecks and range expansions (Watterson 1984; Slatkin 1993). While microsatellite loci have been extremely valuable for conservation genetics, a panel of 20 microsatellites (which is towards the high end employed by most studies) has been shown in one instance to be approximately as effective for estimating genetic relationships as 50 SNP loci (Santure *et al.* 2010). While it is laborious to increase the number of microsatellite loci above the 20 or so that are typically used in conservation genetics, it is very straightforward to scale the number of SNPs assayed into the thousands or tens of thousands, which greatly increases our ability to distinguish barriers to gene flow that are subtle or have only been operating for a small number of generations (Patterson *et al.* 2006; Anderson *et al.* 2010). As genomic-scale datasets become comparable with microsatellites in terms of cost and feasibility, the added resolution from thousands of loci will give a particular boost to population genetic studies in systems with low genetic diversity, and will open entire new classes of analyses to both low- and high-diversity systems.

While a lack of statistical power is one reason why population structure may not be detected in pond-breeding amphibians, another possibility is that, even in low-vagility species, ponds in some systems are truly unstructured, and that failing to recover population structure reflects a biological reality of panmixia across these ponds. Differentiating between low resolving power and true panmixia is critical for conservation and management decision makers. Multiple studies of the same systems with both conventional and genomic datasets can help clarify whether the null hypothesis of population differentiation and strong isolation by distance is a general rule for pond-breeding amphibians, or whether such rules may be habitat or lineage-specific.

The current study is among the first to use thousands of nuclear loci across hundreds of individuals in a large-genome amphibian, and represents an opportunity to compare results between the two genetic approaches in the same system. While little genetic clustering was apparent in the microsatellite loci analyzed by Titus et al. (2014), our dataset of thousands of nuclear SNPs reveals clear population genetic structuring among breeding ponds of *Ambystoma tigrinum* on Long Island. The major genetic patterns in our data are readily apparent in both ADMIXTURE and PCA results. Genetic structuring of ponds generally shows consistent results across years (Figure 11), with two exceptions. First, samples from 2013 in Pond A were classified consistently as a unique population that is admixed with the Pond A lineages sampled in 2014 and 2015. Second, some of the samples from 2014 in Pond J appear to belong to a unique lineage that was not sampled in any other ponds or years. Aside from these two results, consistency between sampling years in the different ponds suggests that the observed patterns of genetic structure are likely driven by geography and not year-to-year variation.

Species with low genetic diversity require collecting data from a greater number of genetic loci to detect population structure (Patterson *et al.* 2006). One cause of low genetic diversity is a range expansion. Church *et al.* (2003) analyzed *Ambystoma tigrinum* mitochondrial DNA and determined that New York was likely recolonized by salamanders from Pleistocene refugia in North Carolina. This was corroborated by Titus et al. (2014), who found low genetic diversity in microsatellite loci in New Jersey and Long Island tiger salamanders. To try to understand whether this low genetic diversity led to the apparent differences between microsatellite and target capture datasets, we compared estimates of genetic diversity from Long Island tiger salamanders to other amphibian systems. Crawford (2003) used a single gene (*c-myc*) to estimate θ in populations of *Eleutherodactylus* frogs in Costa Rica and Panama and obtained values

ranging from 0.00080 to 0.01148 (excluding one population that was fixed for a single haplotype across eight diploid individuals). Weisrock *et al.* (2006) estimated θ at eight nuclear loci from 217 *Ambystoma ordinarium* (a member of the *Ambystoma tigrinum* complex) larvae from across the geographic range of the species (spanning roughly 200km) and obtained an average θ of 0.00208 across loci (min=0.0006, max=0.0034). Similarly, Nadachowska and Babik (2009) sequenced eight nuclear loci for 20 different populations of smooth newt subspecies in Turkey (*Lissotriton vulgaris kosswigi* and *Lissotriton vulgaris vulgaris*). They calculated θ for each population and, after averaging across loci, recovered population estimates ranging from 0.0019 to 0.0081. Finally, we calculated θ as 0.000709 in a collection of 15 *A. californiense* from Merced County, CA. This calculation was performed for a collection of individuals across the same set of nuclear loci presented here, so it is the most direct comparison available. All of these values of θ are greater than the largest value obtained in Long Island tiger salamander ponds (0.000577, mean=0.000427), which indicates that these populations likely do have lower genetic diversity than is normally seen in amphibians.

Breeding ponds that we examined generally exhibited small effective population sizes (< 100), consistent with results found for many other amphibian species (Schmeller & Merilä 2007; Phillipsen *et al.* 2011; McCartney-Melstad & Shaffer 2015). Our estimates (mean=36.9) are larger than, but of the same magnitude as microsatellite-based estimates performed by Titus *et al.* (2014) using the sibship method (Wang 2009a), which had a mean value of 20.9. We did, however, recover several ponds with effective population sizes higher than 44, which was the maximum value recovered by Titus *et al.* (2014). These included pond H (Ne=91.0), pond K (Ne=135.0), pond M (Ne=82.9), and pond O (Ne=68.8). This may indicate that the area around

these ponds, which was not directly sampled by Titus et al. (2014), may harbor greater effective population sizes than elsewhere on Long Island.

A clear relationship between pond size and effective population size was recovered ($p=0.00122$, $R^2=0.5619$, Figure 12). This relationship has been previously observed in *A. californiense* (Wang *et al.* 2011). Interestingly, the pond for which surface area did the worst job predicting Ne, Pond H, had a much higher effective population size estimate than expected by the model (that is, it had the largest residual from the regression line). Pond H is geographically closest pond to Pond K, which has the largest effective population size estimate of any pond. The landscape between Pond H and Pond K is largely forested with no major roads or other anthropogenic barriers to gene flow, the Fst value between ponds H and K is the lowest of any pairwise comparison between ponds (Fst=0.005, Table 6), and these ponds are consistently recovered in the same cluster in ADMIXTURE analyses. Taken together, this suggests that migration has been common between Pond H and Pond K, and that the effective population size of Pond H is augmented by its close relationship with the very large Pond K.

Our approach afforded us the resolution to evaluate the contributions of human disturbance on the movement of salamanders in the form of roads limiting dispersal between ponds. Based on the y-intercepts of linear regressions, the presence of a major road between ponds raised Fst values by approximately 0.04. Pond A was quite distinct from all the other ponds, as was Pond Q (Table 6). These ponds are generally separated from other ponds by greater geographic distance, but they are also separated from all other ponds by major roads. Similarly, ponds E and G tend to separate from all other ponds (PC2 in Figure 9)—these are the only ponds besides pond A that are north of New York State Route 25, a high-traffic road that constitutes a substantial barrier to salamander movement. The combination of geographic distance and roads did an excellent job of

explaining the observed genetic distances between ponds (linear regression, adj. R^2=0.6814). These results suggest that both geographic distance and the presence of roads have affected salamander dispersal for many generations, which has important implications for conservation strategies.

Conclusion

The results of this study show that *Ambystoma tigrinum* ponds on Long Island generally have relatively small effective population sizes that are correlated with the surface area of ponds, that migration is limited among most ponds in the area, and that major roads further limit dispersal. The interrelationships between these factors are important for conservation management. Small effective population sizes imply that ponds are more likely to suffer random demographic extinction, and highly structured populations indicate that locally extirpated ponds (such as those that do not fill with water for many years in a row) may not be easily recolonized by individuals from nearby ponds. Roads and other human activities add to these natural dynamics, and emphasize the critical importance of conserving blocks of contiguous habitat with a complex of ponds that can act as semi-isolated metapopulations. Within the Long Island landscape studied here, there appear to be several clusters of interconnected ponds that periodically share migrants (ponds B, C, and D; ponds H, I, J, K, L, and M; and ponds O and P). For such clusters migrants from interconnected ponds may be expected to "rescue" nearby ponds that go locally extinct, and maintaining these dynamics is probably critical to the long-term persistence of tiger salamanders locally. However, the presence of major roads appears to disrupt this pattern, as seen by the tendency of nearby ponds separated by major roads to fall out in different genetic clusters (such as Pond A vs. ponds B, C, and D and Pond F vs. ponds E and G).

A genomic approach was critical for this experiment to detect the observed population structure at such a fine spatial scale in a post-glacially recolonized area. The distinction between

inferences made from relatively few microsatellite loci from the data generated in this study have important consequences for our understanding of ecological dynamics in the system. Titus *et al.* (2014) recovered little genetic structure among endangered populations of Long Island tiger salamanders and inferred relatively high migration rates between ponds. Conversely, our genomic approach revealed the restrictions in movement between many groups of ponds, despite low overall levels of genetic differentiation.

This study suggests that monitoring of individual ponds is necessary, especially during and following droughts. Our genetic results suggest that ponds not separated by major roads may have increased resilience to local extirpation via demographic rescue from neighboring ponds, so efforts should be made to prevent activities that separate such clusters of ponds. In the event of an observed local extirpation of a pond, the genetic results herein provide information regarding the best source of animals to use for translocations to preserve the current genetic landscape, which is a result of a combination of current and historical patterns of dispersal among ponds.

Acknowledgments

Funding was provided by the Andrew Sabin Family Foundation. EMM and HBS also received support from NSF-DEB 1457832 and NSF-DEB 1257648. We thank Alvin Breisch, Andy Sabin, Pete Davis, and Jeremy Feinberg for assistance with fieldwork, and Megan Sha and Rice Zhang for laboratory support. This work used the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley, supported by NIH S10 Instrumentation Grants S10RR029668 and S10RR027303. Computing resources were provided by XSEDE and the Comet supercomputer at SDSC.

Figures

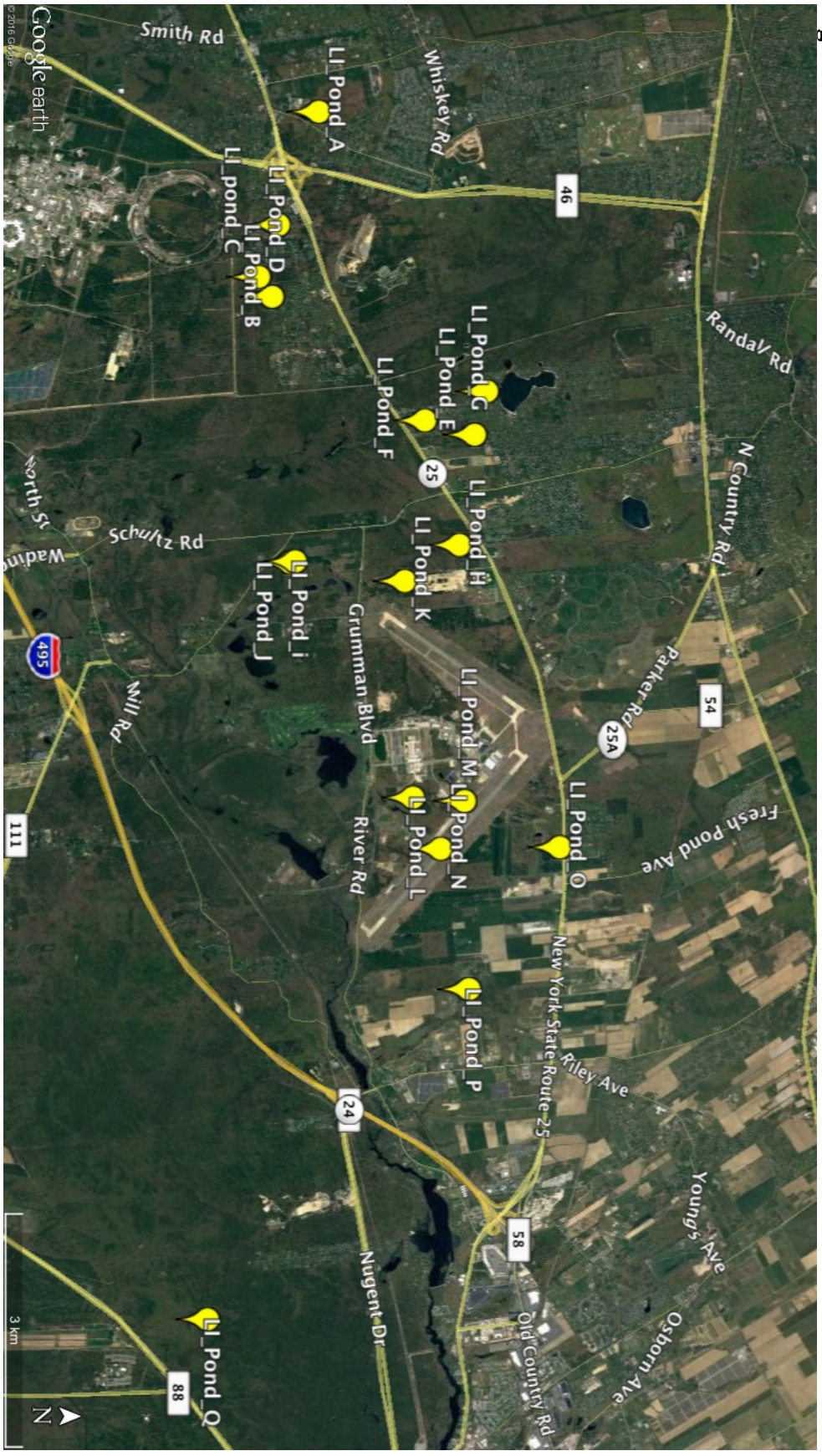

Figure 6: Map of sampling localities.

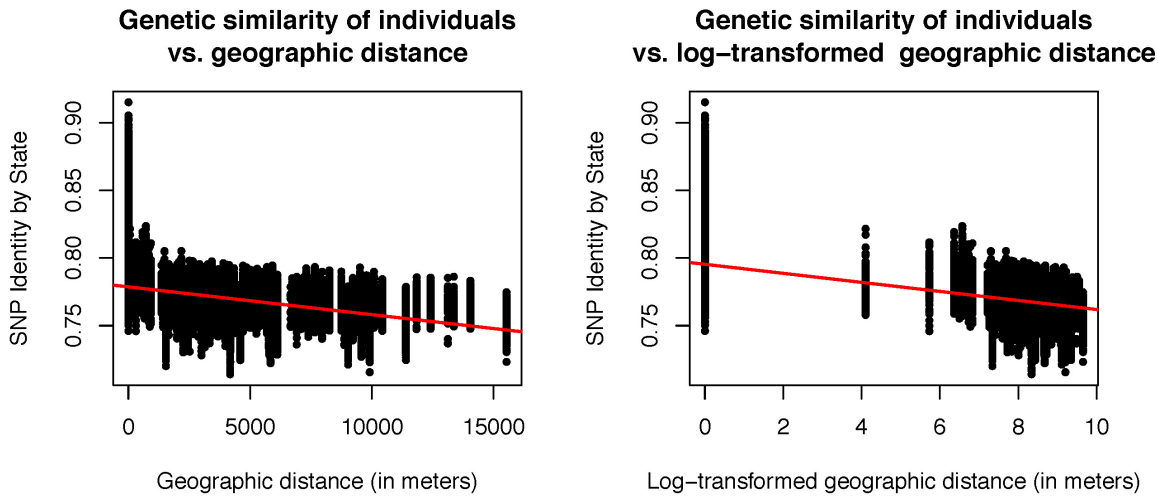

Figure 7: Relationship between genetic similarity and geographic distance between individuals. The plot on the left uses raw Euclidean distance between individuals, while the plot on the left uses log-transformed Euclidean distances.

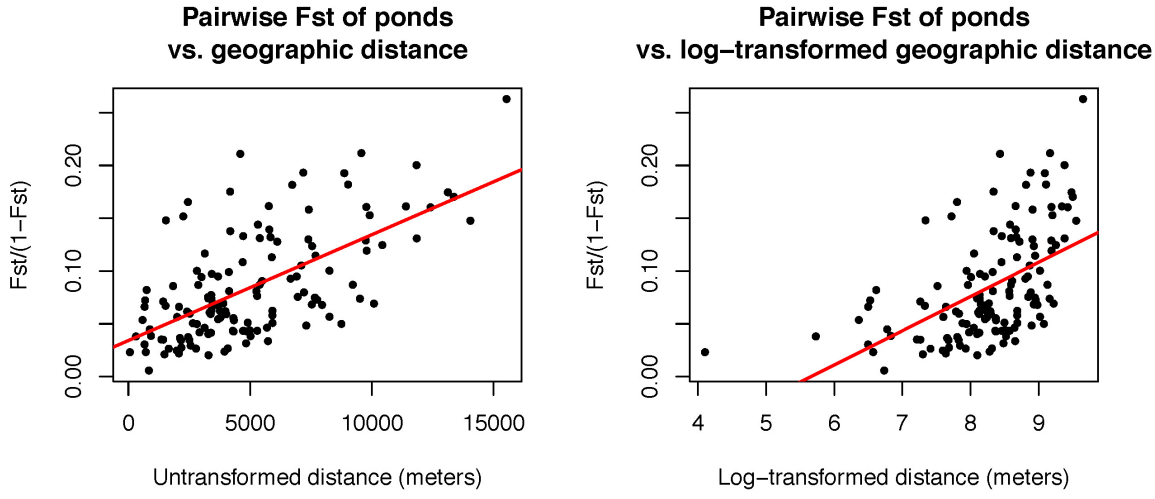

Figure 8: Relationship between genetic distance and geographic distance between ponds. The plot on the left uses raw Euclidean distance between ponds, while the plot on the left uses log-transformed Euclidean distances.

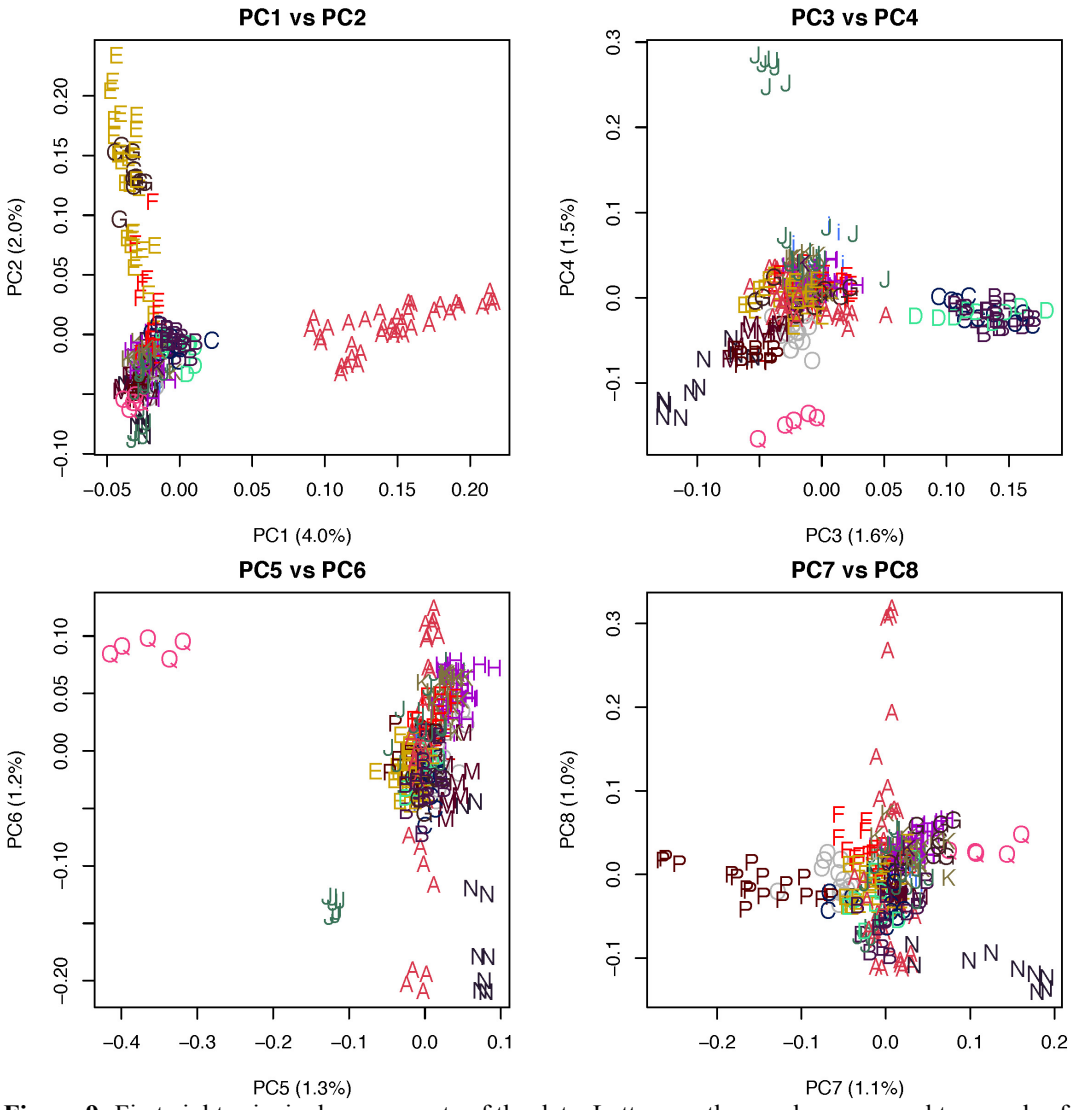

Figure 9: First eight principal components of the data. Letters on the graph correspond to samples from the same pond. Colors are used only to aid in distinguishing between letters.

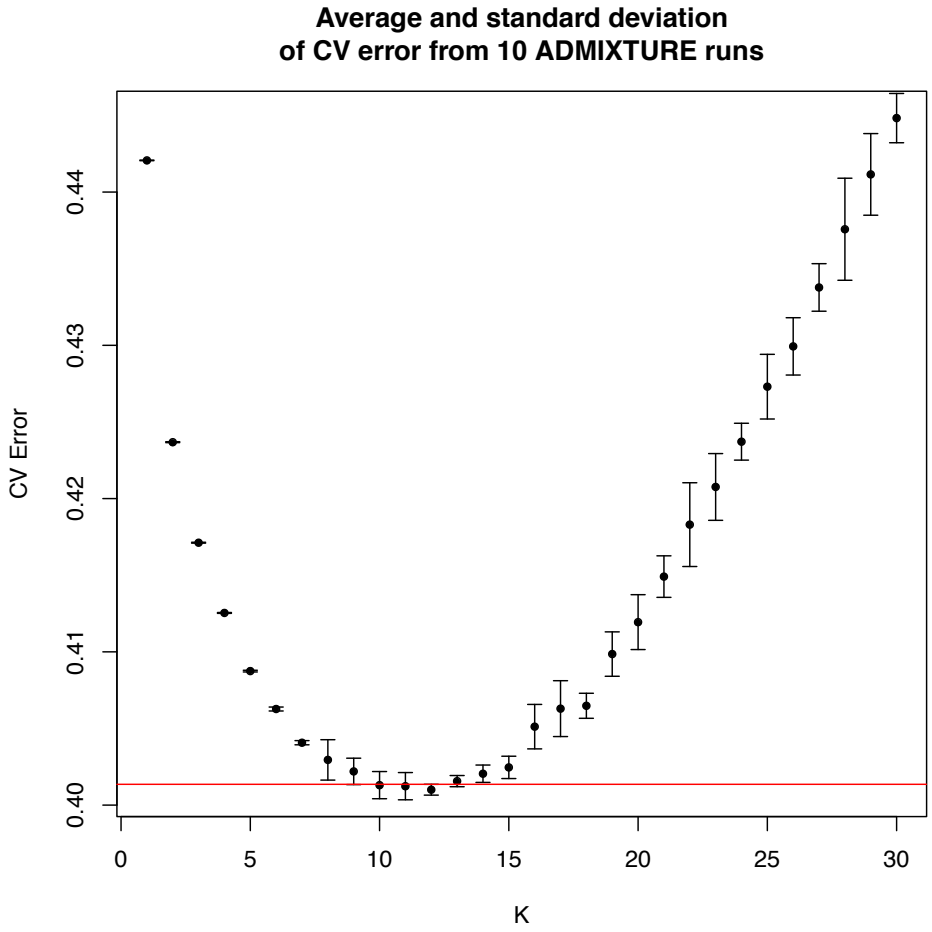

Figure 10: Cross-validation error mean and standard deviations from 10 ADMIXTURE runs using different seeds. The red line is drawn at the mean+SD of the best-performing K value (K=12). The standard deviations for K=9 through K=13 overlap this line.

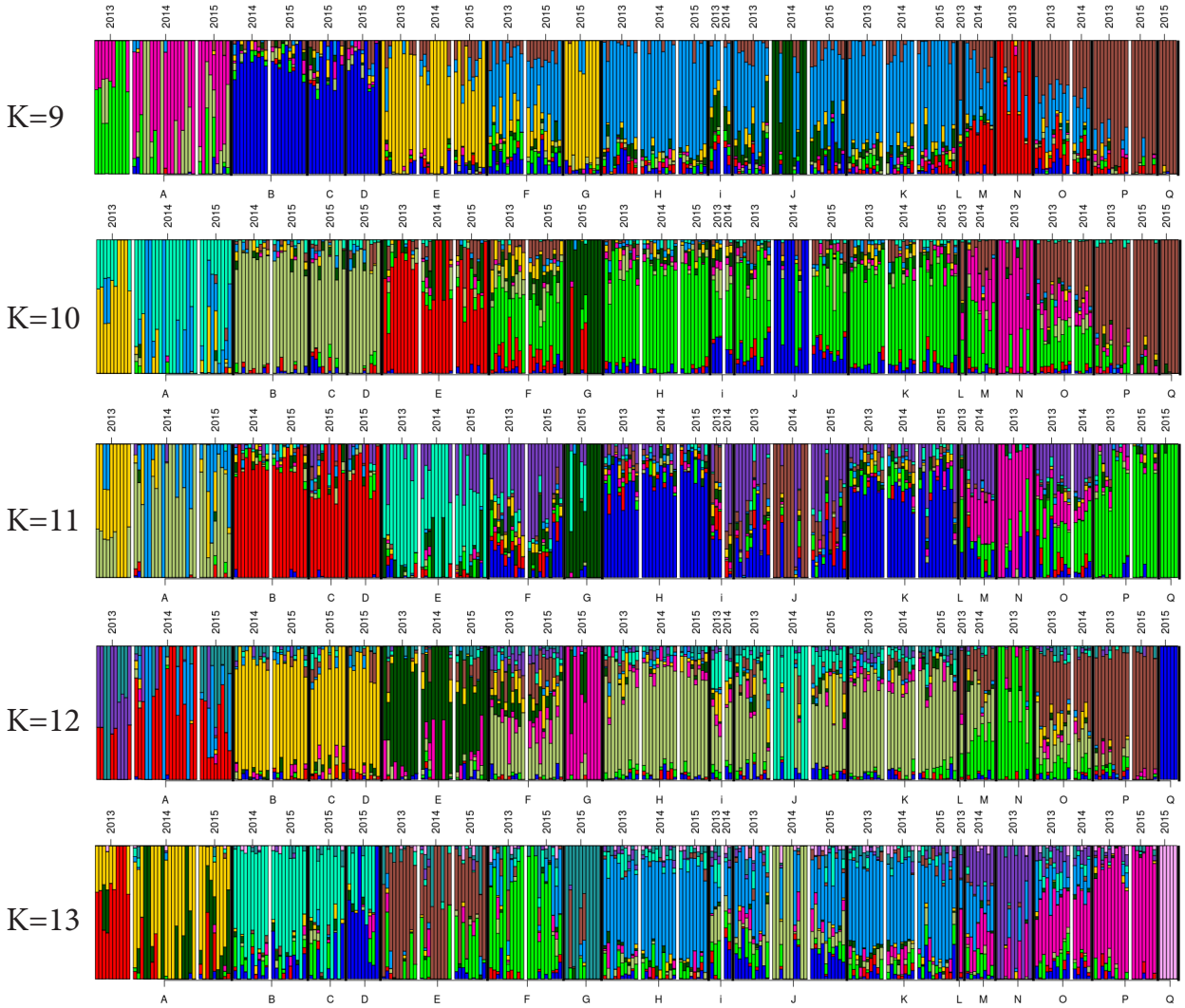

Figure 11: Admixture results from all 282 samples. Letters correspond to ponds from the sample map (Figures 6 and 9). White vertical lines separate sampling years within ponds, and black vertical lines separate ponds from one another.

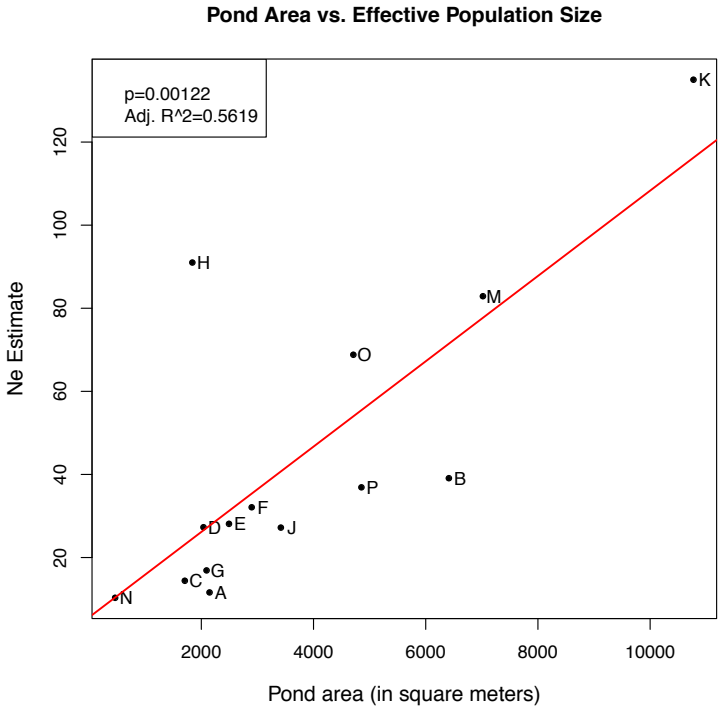

Figure 12: Relationship between pond area and effective population size estimate. Ne estimates represent multiple-cohort calculations if multiple cohorts were samples, otherwise adjusted single-year estimates were used. Ponds i, L, and Q were omitted because they did not contain enough samples to generate an estimate of Ne.

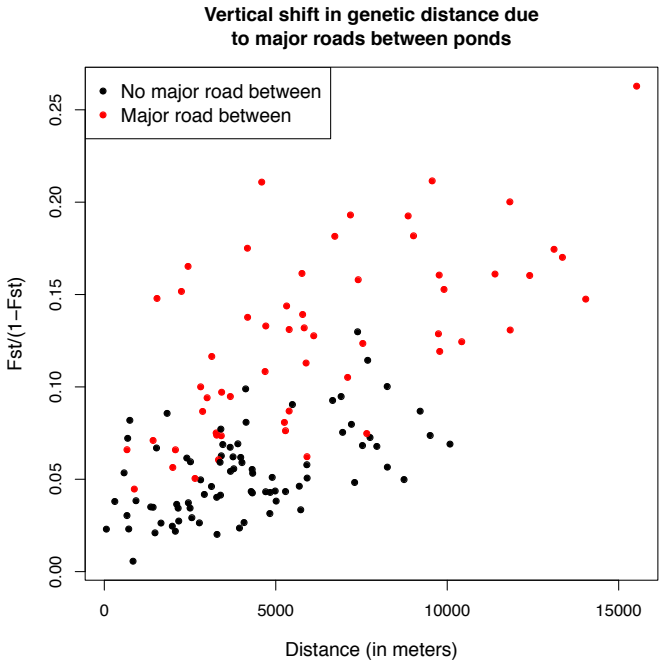

Figure 13: Visualizing the impacts of major roads on genetic differentiation between ponds. For the same geographic distance, ponds separated by major roads (indicated by red dots) tend to have higher levels of genetic differentiation.

Tables

Table 4: Pond localities, areas, Watterson's θ estimates, sampling, and effective population size estimates. Pond areas were estimated from Google Earth satellite images taken in March 2007. Single-year estimates were corrected for iteroparity-induced downward bias as explained in Methods, and both single-year and pooled-year estimates were corrected for dense locus sampling on chromosomes. Infinite values indicate that sample sizes were likely too small to estimate Ne. N=number of samples included in analyses. Ne=Effective population size estimates using LD method.

Pond	Latitude	Longitude	Pond Area (m^2)	Watterson's θ	N (2013/2014/2015)	Ne (2013/2014/2015)
A	40.896379	-72.892071	2147	$3.26x10^{-4}$	37 (10/18/9)	11.6 (6.7/7.1/20.1)
B	40.891766	-72.874854	6413	$4.05x10^{-4}$	20 (0/10/10)	39.1 (NA/37.7/47.2)
C	40.889497	-72.866932	1706	$4.45x10^{-4}$	10 (0/0/10)	14.4 (NA/NA/14.4)
D	40.891043	-72.863908	2039	$4.38x10^{-4}$	9 (0/0/9)	27.3 (NA/NA/27.3)
E	40.915705	-72.849554	2493	$3.75x10^{-4}$	28 (10/9/9)	28.1 (40.4/14.3/19.7)
F	40.908597	-72.845109	2898	$4.16x10^{-4}$	20 (10/0/10)	32.1 (30.8/NA/31.3)
G	40.914317	-72.842938	2094	$4.21x10^{-4}$	10 (0/0/10)	16.9 (NA/NA/16.9)
H	40.912580	-72.826168	1840	$4.06x10^{-4}$	28 (10/10/8)	91.0 (55.5/187.2/515.2)
I	40.893704	-72.823658	944	$4.98x10^{-4}$	5 (3/2/0)	Inf (Inf/Inf/NA)
J	40.893182	-72.823465	3418	$3.94x10^{-4}$	30 (10/10/10)	27.2 (136.2/4.0/91.3)
K	40.906296	-72.820671	10773	$4.07x10^{-4}$	29 (10/8/11)	135.0 (602.1/Inf/27.4)
L	40.907237	-72.787736	8587	$5.77x10^{-4}$	1 (1/0/0)	Inf (Inf/NA/NA)
M	40.913165	-72.787206	7020	$4.62x10^{-4}$	8 (0/8/0)	82.9 (NA/82.9/NA)
N	40.910430	-72.779946	464	$4.22x10^{-4}$	10 (10/0/0)	10.3 (10.3/NA/NA)
O	40.924112	-72.780170	4710	$4.36x10^{-4}$	15 (10/5/0)	68.8 (17689.4/Inf/NA)
P	40.913681	-72.758595	4854	$4.15x10^{-4}$	17 (10/0/7)	36.9 (Inf/NA/11.1)
Q	40.883585	-72.708374	1302	$4.08x10^{-4}$	5 (0/0/5)	Inf (NA/NA/Inf)

Test	R_M	R_M^2	p-value
Individual with log(geographic distance)	0.5349	0.2861	$1x10^{-6}$
Individual with raw geographic distance	0.4200	0.1764	$1x10^{-6}$
Ponds with log(geographic distance)	0.5276	0.2784	$1x10^{-6}$
Ponds with raw geographic distance	0.6305	0.3975	$1.1x10^{-5}$

Table 5: Mantel test results: P-values calculated using 999,999 permutations. R_M is the Mantel R statistic, and R_M^2 is the square of the Mantel R statistic.

	A	B	C	D	E	F	G	H	i	J	K	L	M	N	O	P	Q
A																	
B	0.127																
C	0.129	0.020															
D	0.129	0.030	**0.022**														
E	0.147	0.066	0.069	0.072													
F	0.119	0.043	0.044	0.048	0.040												
G	0.173	0.085	0.086	0.091	0.050	0.060											
H	0.115	0.038	0.041	0.042	0.052	0.022	0.066										
i	0.137	0.050	0.061	0.059	0.068	0.034	0.088	0.035									
J	0.120	0.048	0.050	0.045	0.056	0.031	0.080	0.026	0.022								
K	0.111	0.037	0.040	0.039	0.047	0.018	0.062	0.005	0.033	0.020							
L	**0.159**	**0.072**	**0.064**	**0.099**	**0.071**	**0.039**	**0.111**	**0.018**	**0.043**	**0.041**	**0.024**						
M	0.151	0.066	0.070	0.076	0.073	0.047	0.095	0.037	0.056	0.051	0.038	**0.028**					
N	0.173	0.090	0.100	0.102	0.101	0.082	0.123	0.064	0.088	0.074	0.064	**0.069**	0.065				
O	0.128	0.045	0.049	0.055	0.054	0.031	0.076	0.021	0.038	0.033	0.018	**0.030**	0.030	0.059			
P	0.136	0.063	0.065	0.071	0.066	0.043	0.092	0.040	0.052	0.045	0.037	**0.033**	0.056	0.077	0.031		
Q	0.207	0.131	0.142	0.144	0.139	0.118	0.164	0.111	**0.135**	0.115	0.106	**0.153**	0.135	0.151	0.110	0.116	

Table 6: Pairwise Fst values between ponds. Cells are colored by the magnitude of difference between ponds, with red being relatively low differentiation and green being relatively high differentiation. Bolded cells/values are not significantly different from 0 ($p >$ Benjamini-Yekutieli-corrected 0.05).

References

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

Anderson CD, Epperson BK, Fortin M-J *et al.* (2010) Considering spatial and temporal scale in landscape-genetic studies of gene flow. *Molecular Ecology*, **19**, 3565–3575.

Bankevich A, Nurk S, Antipov D *et al.* (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*, **19**, 455–477.

Benjamini Y, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, **29**, 1165–1188.

Bishop SC (1941) *The salamanders of New York*. University of the State of New York.

Blaustein AR, Wake DB, Sousa WP (1994) Amphibian Declines: Judging Stability, Persistence, and Susceptibility of Populations to Local and Global Extinctions. *Conservation Biology*, **8**, 60–71.

Camacho C, Coulouris G, Avagyan V *et al.* (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.

Church SA, Kraus JM, Mitchell JC, Church DR, Taylor DR (2003) Evidence for Multiple Pleistocene Refugia in the Postglacial Expansion of the Eastern Tiger Salamander, Ambystoma Tigrinum Tigrinum. *Evolution*, **57**, 372–383.

Corander J, Waldmann P, Sillanpää MJ (2003) Bayesian Analysis of Genetic Differentiation Between Populations. *Genetics*, **163**, 367–374.

Coster SS, Babbitt KJ, Cooper A, Kovach AI (2015) Limited influence of local and landscape factors on finescale gene flow in two pond-breeding amphibians. *Molecular Ecology*, **24**, 742–758.

Crawford AJ (2003) Huge populations and old species of Costa Rican and Panamanian dirt frogs inferred from mitochondrial and nuclear gene sequences. *Molecular Ecology*, **12**, 2525–2540.

Danecek P, Auton A, Abecasis G *et al.* (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.

DePristo MA, Banks E, Poplin RE *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics*, **43**, 491–498.

Do C, Waples RS, Peel D *et al.* (2014) NeEstimator v2: re-implementation of software for the estimation of contemporary effective population size (Ne) from genetic data. *Molecular Ecology Resources*, **14**, 209–214.

Excoffier L, Lischer HE (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular ecology resources*, **10**, 564–567.

Furman BLS, Scheffers BR, Taylor M, Davis C, Paszkowski CA (2016) Limited genetic structure in a wood frog (Lithobates sylvaticus) population in an urban landscape inhabiting natural and constructed wetlands. *Conservation Genetics*, **17**, 19–30.

Glasbey C, van der Heijden G, Toh VFK, Gray A (2007) Colour displays for categorical images. *Color Research & Application*, **32**, 304–309.

Glenn TC, Nilsen R, Kieran TJ *et al.* (2016) Adapterama I: Universal stubs and primers for thousands of dual-indexed Illumina libraries (iTru & iNext). *bioRxiv*, 049114.

Hill WG (1981) Estimation of effective population size from data on linkage disequilibrium. *Genetics Research*, **38**, 209–216.

Hitchings SP, Beebee TJ (1997) Genetic substructuring as a result of barriers to gene flow in urban Rana temporaria (common frog) populations: implications for biodiversity conservation. *Heredity*, **79**, 117–127.

Hunter SS, Lyon RT, Sarver BAJ *et al.* (2015) Assembly by Reduced Complexity (ARC): a hybrid approach for targeted assembly of homologous sequences. *bioRxiv*.

Jehle R, Burke T, Arntzen JW (2005) Delineating fine-scale genetic units in amphibians: Probing the primacy of ponds. *Conservation Genetics*, **6**, 227–234.

Jiang H, Lei R, Ding S-W, Zhu S (2014) Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC Bioinformatics*, **15**, 182.

Keinath MC, Timoshevskiy VA, Timoshevskaya NY *et al.* (2015) Initial characterization of the large genome of the salamander Ambystoma mexicanum using shotgun and laser capture chromosome sequencing. *Scientific Reports*, **5**, 16413.

Keinath MC, Voss SR, Tsonis PA, Smith JJ (2016) A linkage map for the Newt Notophthalmus viridescens: Insights in vertebrate genome and chromosome evolution. *Developmental Biology*.

Lampert KP, Rand AS, Mueller UG, Ryan MJ (2003) Fine-scale genetic pattern and evidence for sex-biased dispersal in the túngara frog, Physalaemus pustulosus. *Molecular Ecology*, **12**, 3325–3334.

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods*, **9**, 357–359.

Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv:1303.3997 [q-bio]*.

Licht LE, Lowcock LA (1991) Genome size and metabolic rate in salamanders. *Comparative Biochemistry and Physiology Part B: Comparative Biochemistry*, **100**, 83–92.

Madison DM, Farrand L (1998) Habitat Use during Breeding and Emigration in Radio-Implanted Tiger Salamanders, Ambystoma tigrinum. *Copeia*, **1998**, 402–410.

Mantel N (1967) The detection of disease clustering and a generalized regression approach. *Cancer Research*, **27**, 209–220.

McCartney-Melstad E, Mount GG, Shaffer HB (2016) Exon capture optimization in amphibians with large genomes. *Molecular Ecology Resources*, **16**, 1084–1094.

McCartney-Melstad E, Shaffer HB (2015) Amphibian molecular ecology and how it has informed conservation. *Molecular Ecology*, **24**, 5084–5109.

McKenna A, Hanna M, Banks E *et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, **20**, 1297–1303.

Nadachowska K, Babik W (2009) Divergence in the Face of Gene Flow: The Case of Two Newts (Amphibia: Salamandridae). *Molecular Biology and Evolution*, **26**, 829–841.

New York State Department of Environmental Conservation (2015) Eastern tiger salamander fact sheet. Available http://www.dec.ny.gov/animals/7143.html. (Accessed September, 2016).

Newman CE, Austin CC (2016) Sequence capture and next-generation sequencing of ultraconserved elements in a large-genome salamander. *Molecular Ecology*, In press.

Newman RA, Squire T (2001) Microsatellite variation and fine-scale population structure in the wood frog (Rana sylvatica). *Molecular Ecology*, **10**, 1087–1100.

Oksanen J, Blanchet FG, Friendly M *et al.* (2016) The vegan package. *https://CRAN.R-project.org/package=vegan*.

O'Neill EM, Schwartz R, Bullock CT *et al.* (2013) Parallel tagged amplicon sequencing reveals major lineages and phylogenetic structure in the North American tiger salamander (Ambystoma tigrinum) species complex. *Molecular Ecology*, **22**, 111–129.

Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. *PLoS Genet*, **2**, e190.

Peterman WE, Anderson TL, Ousterhout BH *et al.* (2015) Differential dispersal shapes population structure and patterns of genetic differentiation in two sympatric pond breeding salamanders. *Conservation Genetics*, **16**, 59–69.

Phillipsen IC, Funk WC, Hoffman EA, Monsen KJ, Blouin MS (2011) Comparative analyses of effective population size within and among species: Ranid frogs as a case study. *Evolution*, **65**, 2927–2945.

Polich RL, Searcy CA, Shaffer HB (2013) Effects of tail-clipping on survivorship and growth of larval salamanders. *The Journal of Wildlife Management*, **77**, 1420–1425.

Portik DM, Smith LL, Bi K (2016) An evaluation of transcriptome-based exon capture for frog phylogenomics across multiple scales of divergence (Class: Amphibia, Order: Anura). *Molecular Ecology Resources*, **16**, 1069–1083.

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

R Core Team (2015) R: A language and environment for statistical computing. https://www.R-project.org/.

Raj A, Stephens M, Pritchard JK (2014) fastSTRUCTURE: Variational Inference of Population Structure in Large SNP Data Sets. *Genetics*, **197**, 573–589.

Reyes-Valdés MH (2013) Informativeness of Microsatellite Markers. In: *Microsatellites* Methods in Molecular Biology Vol 1006. (ed Kantartzi SK), pp. 259–270. Humana Press.

Rohland N, Reich D (2012) Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Research*, **22**, 939–946.

Rousset F (1997) Genetic Differentiation and Estimation of Gene Flow from F-Statistics Under Isolation by Distance. *Genetics*, **145**, 1219–1228.

Sambrook J, Russell DW (2001) *Molecular cloning: a laboratory manual (3-volume set)*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

Santure AW, Stapley J, Ball AD *et al.* (2010) On the use of large marker panels to estimate inbreeding and relatedness: empirical and simulation studies of a pedigreed zebra finch population typed at 771 SNPs. *Molecular Ecology*, **19**, 1439–1451.

Savage WK, Fremier AK, Bradley Shaffer H (2010) Landscape genetics of alpine Sierra Nevada salamanders reveal extreme population subdivision in space and time. *Molecular Ecology*, **19**, 3301–3314.

Schmeller DS, Merilä J (2007) Demographic and genetic estimates of effective population and breeding size in the amphibian Rana temporaria. *Conservation Biology*, **21**, 142–151.

Shaffer HB, Gidiş M, McCartney-Melstad E *et al.* (2015) Conservation Genetics and Genomics of Amphibians and Reptiles. *Annual Review of Animal Biosciences*, **3**.

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.

Slatkin M (1993) Isolation by Distance in Equilibrium and Non-Equilibrium Populations. *Evolution*, **47**, 264–279.

Slatkin M (1995) A Measure of Population Subdivision Based on Microsatellite Allele Frequencies. *Genetics*, **139**, 457–462.

Smouse PE, Long JC, Sokal RR (1986) Multiple Regression and Correlation Extensions of the Mantel Test of Matrix Correspondence. *Systematic Zoology*, **35**, 627–632.

Sotiropoulos K, Eleftherakos K, Tsaparis D *et al.* (2013) Fine scale spatial genetic structure of two syntopic newts across a network of ponds: implications for conservation. *Conservation Genetics*, **14**, 385–400.

Stewart MM, Rossi J (1981) The Albany Pine Bush: a northern outpost for southern species of amphibians and reptiles in New York. *American Midland Naturalist*, 282–292.

Sun C, Shepard DB, Chong RA *et al.* (2012) LTR Retrotransposons Contribute to Genomic Gigantism in Plethodontid Salamanders. *Genome Biology and Evolution*, **4**, 168–183.

Titus VR, Bell RC, Becker CG, Zamudio KR (2014) Connectivity and gene flow among Eastern Tiger Salamander (Ambystoma tigrinum) populations in highly modified anthropogenic landscapes. *Conservation Genetics*, **15**, 1447–1462.

Trenham PC, Bradley Shaffer H, Koenig WD, Stromberg MR, Ross ST (2000) Life history and demographic variation in the California tiger salamander (Ambystoma californiense). *Copeia*, **2000**, 365–377.

Van der Auwera GA, Carneiro MO, Hartl C *et al.* (2013) From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al.]*, **43**, 11.10.1-33.

Voss SR, Kump DK, Putta S *et al.* (2011) Origin of amphibian and avian chromosomes by fission, fusion, and retention of ancestral chromosomes. *Genome Research*, **21**, 1306–1312.

Wang J (2009a) A new method for estimating effective population sizes from a single sample of multilocus genotypes. *Molecular Ecology*, **18**, 2148–2164.

Wang IJ (2009b) Fine-scale population structure in a desert amphibian: landscape genetics of the black toad (Bufo exsul). *Molecular Ecology*, **18**, 3847–3856.

Wang IJ (2012) Environmental and topographic variables shape genetic structure and effective population sizes in the endangered Yosemite toad. *Diversity and Distributions*, **18**, 1033–1041.

Wang IJ, Johnson JR, Johnson BB, Shaffer HB (2011) Effective population size is strongly correlated with breeding pond size in the endangered California tiger salamander, Ambystoma californiense. *Conservation Genetics*, **12**, 911–920.

Wang IJ, Savage WK, Bradley Shaffer H (2009) Landscape genetics and least-cost path analysis reveal unexpected dispersal routes in the California tiger salamander (Ambystoma californiense). *Molecular Ecology*, **18**, 1365–1374.

Waples RS, Antao T, Luikart G (2014) Effects of Overlapping Generations on Linkage Disequilibrium Estimates of Effective Population Size. *Genetics*, **197**, 769–780.

Waples RS, Do C (2010) Linkage disequilibrium estimates of contemporary Ne using highly variable genetic markers: a largely untapped resource for applied conservation and evolution. *Evolutionary Applications*, **3**, 244–262.

Waples RK, Larson WA, Waples RS (2016) Estimating contemporary effective population size in non-model species using linkage disequilibrium across thousands of loci. *Heredity*, **117**, 233–240.

Watterson GA (1975) On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, **7**, 256–276.

Watterson GA (1984) Allele frequencies after a bottleneck. *Theoretical Population Biology*, **26**, 387–407.

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.

Wigginton JE, Cutler DJ, Abecasis GR (2005) A Note on Exact Tests of Hardy-Weinberg Equilibrium. *The American Journal of Human Genetics*, **76**, 887–893.

Wright S (1943) Isolation by Distance. *Genetics*, **28**, 114–138.

Zamudio KR, Wieczorek AM (2007) Fine-scale spatial genetic structure and dispersal among spotted salamander (Ambystoma maculatum) breeding populations. *Molecular Ecology*, **16**, 257–274.

Zheng X, Levine D, Shen J *et al.* (2012) A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics (Oxford, England)*, **28**, 3326–3328.

Desert Tortoises in the Genomic Age:
Population Genetics and the Landscape

Preface

Context

The following chapter of my dissertation is in the form of a report that was submitted to the California Department of Fish and Wildlife. It details a series of analyses to help understand the impacts of several alternative spatial configurations of renewable energy development on gene flow of the federally threatened Mojave desert tortoise. These development alternatives were the centerpiece of the Desert Renewable Energy Conservation Plan (DRECP), a landscape-level land use planning initiative undertaken by the Bureau of Land Management (BLM), U.S. Fish and Wildlife Service (USFWS), California Energy Commission (CEC), and the California Department of Fish and Wildlife (CDFW). We were funded by the California Department of Fish and Wildlife to provide a detailed analysis of these alternative plans on desert tortoise gene flow, and submitted the report for the public comment period for the initial implementation of the DRECP.

Future Plans

Current state of landscape-level planning for the Mojave desert tortoise

The five proposed land use configuration alternatives analyzed in the subsequent report include public and private lands spread across several counties in California. Shortly after the end of the DRECP's public comment period, the government agencies that developed the DRECP announced that they would be splitting its implementation into two phases: one that deals with land use decisions on BLM-controlled lands and one that deals with non-BLM areas (Sahagun 2015).

Phase I of the DRECP was approved by the Bureau of Land Management on September 14, 2016 (U.S. Bureau of Land Management 2016). This phase includes land use planning decisions for BLM-administered lands. Specifically, 388,000 acres of public lands were designated as development focus areas (DFAs). In applications for leasing lands for renewable energy development, DFAs will not require the same degree of environmental evaluation prior to permitting, as they've already been evaluated in the context of the DRECP. The application process for renewable energy development within DFAs will be streamlined to encourage development in these areas. Phase I also designated a total of 6,527,000 acres for natural resource conservation. This includes California Desert National Conservation Lands, Areas of Critical Environmental Concern, and Wildlife Allocations. A further 2,691,000 acres were designated for recreation under Phase I. Phase II of the DRECP is currently under development

in conjunction with county-level governments to extend this landscape-level planning beyond BLM-administered lands.

As the DRECP has taken shape over the past year, so too have our plans for parsing, supplementing, and disseminating the research contained in the following report. We are currently planning and preparing at least four distinct publications that draw data, analyses, and inferences from the current report:

1. A paper that leverages the 270 low-coverage tortoise genome sequences from across their range to make conservation-relevant population genetic inferences about the desert tortoise. This includes visualizing the current state of genetic variation on the landscape, contextualizing levels of genetic differentiation between populations, and exploring the spatial distribution of allelic variation that is under selection in the genome. We will use these analyses to make recommendations about management units and to help managers understand the genetic consequences of different tortoise translocation schemes.
2. A study that uses the genomic data to fit a model of deeper-time demographic history of the desert tortoise. This paper will combine demographic simulations with knowledge of the geological history of the Mojave Desert to better understand the deeper observed divergences between groups of tortoises (such as those explained by PC1 in Figure 17).
3. An applied conservation paper that estimates the effects of renewable energy development on gene flow of the tortoise. This analysis will incorporate changes to the DRECP that have been implemented since the closing of the public comment period.
4. At least one paper that formally describes the methods we use to estimate genetic parameters from low-coverage genome read count data and the software we've written to implement them.

Desert Tortoises in the Genomic Age:

Population Genetics and the Landscape

Draft Final Report to the California Department of Fish and Wildlife. February 22, 2015

H. Bradley Shaffer[1], Evan McCartney-Melstad[1], Peter Ralph[2], Gideon Bradburd[3], Erik Lundgren[2], Jannet Vu[1], Bridgette Hagerty[4], Fran Sandmeier[5], Chava Weitzman[6], Richard Tracy[6]

1) University of California, Los Angeles. Los Angeles, CA 90095. 2) University of Southern California, Los Angeles, CA 90089. 3) University of California, Davis. Davis, CA 95616. 4) York College of Pennsylvania. York, PA 17403. 5) Lindenwood University – Belleville. Belleville, IL 62226. 6) University of Nevada, Reno. Reno, NV 89557.

Abstract

The California Department of Fish and Wildlife (CDFW) provided research funds to study the conservation genomics and landscape genomics of the California desert tortoise, *Gopherus agassizii*, in response to the Desert Renewable Energy Conservation Plan (DRECP). Our desert tortoise research group ("Team Tortoise") was headed by the lab of Brad Shaffer at University of California Los Angeles and included colleagues at, or previously from, the University of Southern California, University of California at Davis, and University of Nevada at Reno. Team Tortoise consolidated tissue samples of the desert tortoise from across the species range within California and southern Nevada, generated a DNA dataset consisting of full genomes of 270 tortoises, and analyzed the way in which the environment of the desert tortoise has determined modern patterns of relatedness and genetic diversity across the landscape. Here we present the implications of these results for the conservation and landscape genomics of the desert tortoise. Our work strongly indicates that several well-defined genetic groups exist within the species, including a primary north-south genetic discontinuity at the Ivanpah Valley and another separating western from eastern Mojave samples. We also incorporate existing desert tortoise habitat modeling data into a novel, spatially explicit, landscape genomic inference framework that allowed us to predict the relative impacts of five proposed development alternatives within the DRECP and rank them with respect to their likely impacts on desert tortoise gene flow and connectivity in the Mojave. Finally, we analyzed the impacts of each of the 214 distinct proposed development area "chunks," derived from the proposed development polygons, and ranked each chunk in terms of its range-wide impacts on desert tortoise gene flow. This whole-genome approach, which we have here implemented at an unprecedented scale for a non-model species, is returning spatially-explicit results at a level of detail that has not been previously possible, allowing us to evaluate alternative land use projections at a biologically meaningful level for desert tortoise movement and population connectivity.

Executive Summary: Objectives and Deliverables

Below, we summarize the project's original objectives, the results obtained, and ongoing research directions:

1. Develop new methods for generating a comprehensive genome-wide summary of genetic variability among desert tortoises in the region using high throughput, low coverage genome sequencing.

Our sequencing strategy was extremely successful, generating data for over 50 million putative polymorphic loci for 270 tortoises. Over 1.29 trillion base pairs of total genetic data were produced for this project.

2. Develop a novel, robust landscape genetic inference framework that accommodates the statistical uncertainty associated with low-coverage sequencing to accurately estimate genetic relatedness among tortoises.

Standard single nucleotide polymorphism (SNP) genotype calling is error-prone with low coverage sequencing data of the type that we generated. Consequently, we developed, tested, and employed methods to measure genetic relationships between individuals based on raw read count data averaged over tens of millions of sites in the genome, allowing us to confidently infer relatedness and other summary statistics between pairs of tortoises.

3. Assemble tissue samples from up to four target populations in the Ivanpah Valley and Pisgah, Brisbane, and Pinto Wash corridors, with a goal of one tortoise per 1-10 hectares, evenly spaced across these previously identified linkage corridors.

We established a collaboration with Professor Dick Tracy and his team to share blood samples from roughly 1,000 desert tortoises that his group has collected from throughout the Mojave Desert over the past decade. We also collaborated with Roy Averill-Murray of the US Fish and Wildlife Service and with the Desert Tortoise Conservation Center (DTCC) to increase our sampling, especially in regions where development is proposed or underway in the Ivanpah Valley. Using a subset of 270 of these samples allowed us to employ our desired spatial genetic sampling without any further disturbance to wild tortoise populations.

4. Identify and collate the finest-scale habitat models available and assemble GIS layers for vegetation, soil type, elevation, temperature, and roads at one site (Ivanpah, Pisgah, or Brisbane, depending on the identified needs of the Grantor and the United States Fish and Wildlife Service (USFWS)).

After a review of published studies on desert tortoise habitat suitability models and landscape genetics, we identified and obtained 83 landscape variables as GIS layers covering the geographical range of our tortoise samples to a resolution of 30 X 30 meters. Of these, we selected three subsets of 6, 12, and 24 layers chosen to minimize correlation among layers to use in analyses. These models were then compared to an isolation by distance model that incorporated habitat bounds specified by a desert tortoise habitat model generated by Nussear et al. (2009). We found the landscape variable models fit no better than the simpler model using only the Nussear et al. habitat projections, and so we used only that layer for the current work. However, we retained the GIS layers for future analyses.

5. Model existing landscape features and linkage corridors identified by the DRECP to determine resistance pathways for individual tortoise connectivity across the landscape.

We developed a novel statistical framework in which all pairwise genetic divergences between sampled individual tortoises were fit to a geographically explicit model of tortoise dispersal across a landscape with heterogeneous resistance to movement. We used the resulting model to evaluate the relative and absolute impacts of the five proposed development alternatives (Preferred Alternative as well as Alternatives 1, 2, 3, and 4) in the draft DRECP. In building this model, we combined development projects into a set of parcels, considered each to be inaccessible to tortoise movement, and modeled the decrease in gene flow that each alternative, and each parcel, imposes on tortoises. Alternative 1 was found to be least detrimental to tortoises in the Mojave, followed (in increasing order of impact) by the Preferred Alternative, Alternative 4, Alternative 3, and Alternative 2. The Preferred Alternative, and Alternatives 3 and 4 had roughly twice the impact of Alternative 1, while Alternative 2 had three times the impact of Alternative 1 on tortoise gene flow.

6. Using siting information from existing project applications and DRECP Development Focus Areas (DFAs) and reserve lands, identify the most likely siting of renewable energy installations - including service roads, parking structures, and buildings, and model the impacts of different configurations and placements on overall connectivity across the length of the corridor.

The inference framework we developed allows for flexibly evaluating virtually any proposed development alternative. We provide a rank order of relative and absolute impacts of 214 discrete potential areas of development ("development chunks") from all five DRECP alternative development schemes. Other potential development schemes can be similarly evaluated in the future.

Introduction

The Mojave desert tortoise (*Gopherus agassizii,* recently identified as taxonomically distinct from the Sonoran desert tortoise, *Gopherus morafkai,* see Murphy *et al.* 2011) is a widely distributed but declining resident of the Mojave Desert. Potential renewable energy development may negatively influence future population trajectories of this species, placing it into direct conflict with renewable energy projects. Anticipated increases in direct mortality, habitat loss, and especially habitat fragmentation from renewable energy development within the Desert Renewable Energy Conservation Plan (DRECP) Planning Area need to be considered in combination with other population stressors in the desert tortoise's range. Listed under both the federal and California Endangered Species Acts, the Mojave desert tortoise is and will continue to be a significant driver of reserve design under the DRECP. Gaining the most precise knowledge possible about tortoise population health, genetic diversity, population substructure, and exchange of migrants in this widespread, unevenly distributed species will contribute to a comprehensive conservation strategy for this and other species covered by the plan. It will also inform the delineation of ecologically meaningful reserves in California's deserts and facilitate siting of viable zones for renewable energy that best accomplish the goals of energy development while minimizing impacts on current and future tortoise population dynamics.

Recovery of rare and endangered species requires a series of interrelated steps and approaches. One pressing need is to apply the best available scientific techniques to understand how widespread species are genetically connected across landscapes, and therefore the extent to which seemingly discrete populations are genetically and demographically connected or isolated. Acquiring these data via direct field observations (e.g. mark-recapture, radio transmitters) is time-consuming, challenging, and expensive, especially for a cryptic species like the Mojave desert tortoise, which occurs at low density, is frequently in underground retreats, and lives for decades. Both direct and indirect (genetic) measures of gene flow and connectivity should be used to fully realize how organisms traverse landscapes, and therefore how to best preserve historical connections in the face of human modifications and the habitat fragmentation that often follows. A combination of direct and indirect genetic analyses often brings complementary information to management. For example, direct measures may indicate daily and seasonal activity and movement patterns, whereas indirect genetic measures often better indicate the extent to which dispersal results in successful reproduction and the long-distance, but often infrequent, movement of genes across landscapes.

Until recently, the principal paradigm used to study how genetic variation traverses landscapes has been one of Isolation by Distance (IBD; Wright 1943). One nearly ubiquitous field observation is that genetic differentiation between populations increases with the geographic distance between them (Jenkins *et al.* 2010) ("Everything is related to everything else, but near things are more related than distant things;" Tobler 1970). With the advent of cheaper DNA sequencing and the rise of Geographic Information Systems (GIS) data availability, the field of landscape genetics has developed to quantify other aspects of environmental heterogeneity that shape patterns of dispersal and genetic variation above and

beyond that derived from geographic distance alone. Recent innovations have brought inferences from the burgeoning fields of landscape ecology and associated ecological niche models (Elith and Leathwick 2009; Forman 1995; Peterson *et al.* 1999) into the equation, adding a much-needed GIS component to landscape genetics analyses. Fortunately, some of the groundbreaking research in this area has been carried out in desert tortoises. We briefly summarize that research below, before describing our methods and findings.

Current state of knowledge of desert tortoise landscape genetics

Andersen and colleagues (2000) modeled tortoise density, measured using field observations, as a function of 11 spatial GIS data layers, and found that "soil composition and parent materials can be important determinants of habitat suitability." In 2009, Nussear *et al.* expanded on this work, compiling a dataset of over 15,000 desert tortoise (both *G. agassizii* and *G. morafkai*) occurrence observations. They modeled desert tortoise presence and absence over its entire range as a function of 16 spatial GIS layers and built a habitat suitability model for the species.

At the same time, a parallel research program was generating DNA sequence data for desert tortoises to learn how gene flow between desert tortoise populations may be structured by landscape characteristics. Edwards *et al.* (2004) reported on genetic variation at seven microsatellite loci for 170 Arizona (*G. morafkai)* tortoises, some of which were also tracked via radiotelemetry. Their research indicated that "long-distance movements result in the exchange of genetic material among adjacent populations," but that "estimates of gene flow predate anthropogenic habitat fragmentation and should not be taken as evidence that natural immigration/emigration still occurs." That is, because of the long generation time of tortoises, it may take many years for genetic population substructure to reflect reduced patterns of migration caused by anthropogenic disturbances. They further warned that long-distance migrants may be a critical component of connectivity and metapopulation dynamics in the species.

Hagerty *et al.* (2011) used the habitat suitability model of Nussear *et al.* (2009) to parameterize a resistance surface and a least cost path map (where resistance to migration in a patch of habitat (McRae 2006) is the inverse of its suitability from the model). Using 20 variable microsatellite loci sequenced in 744 tortoises, Hagerty *et al.* (2011) found support for both an effect of geographic distance and topographical barriers in structuring patterns of gene flow over the range of the desert tortoise. Their summary figure showing resistance across the range of the Mojave desert tortoise suggested that many low-cost paths, including ones in and out of the Ivanpah Valley, existed among their 25 population samples. Latch *et al.* (2011) employed similar sampling (859 tortoises, genotyped at 16 microsatellite loci), and found support for the hypothesis that both natural landscape features (slope), and anthropogenic features (roads) were limiting gene flow.

These studies offered exciting clues into the biology of the desert tortoise and the way in which the species interacts with its landscape. They also provided the earliest foundational insights that can be gleaned from landscape genetic data within the landscape and conservation genetics communities for desert tortoises. However, these studies were also limited in their

ability to draw strong, spatially explicit conclusions by a lack of genetic resolution and the inability to precisely model anthropogenic impacts on population genetic movement probabilities. Recent advances in next generation sequencing (NGS) have enabled the relatively cheap generation of many orders of magnitude more DNA sequence data, which in turn has enabled the detection and quantification of vastly more subtle landscape effects on patterns of current and historical gene flow, and we fully embrace these advances in the current work.

In the current study, our goal is to move beyond the foundational work that has been accomplished on Mojave desert tortoise landscape genetics and beyond what is achievable with SNP-based NGS studies. By generating results based on full genome sequencing, we bring a greater level of precision to landscape genetic analyses than has previously been possible for the Mojave desert tortoise. Rather than generate genetic data for a few thousand variable single nucleotide polymorphisms (SNPs) as is now sometimes done in NGS studies, we sequenced the entire genome (an estimated two billion base pairs) of 270 geographically distributed tortoises. The resulting dataset is, to our knowledge, the most comprehensively geographically sampled dataset of whole genomes in a wild animal species. We generated this massive DNA dataset from a sample of desert tortoises that evenly covers their range, used (and invented, when necessary) cutting-edge spatial statistics tools, and applied these tools to both an expanded collection of high resolution GIS data rasters (explained in detail in Appendix I) and the habitat model of Nussear *et al.* (2009) to quantify how gene flow between tortoises has been affected by historical landscape features and how it will be affected by future anthropogenic changes. Below, we discuss these results and present specific, actionable, and data-supported recommendations for the conservation of the desert tortoise.

Deciding among genomic approaches

Researchers in population genetics now have a choice of many different types of data and strategies for genetic data collection. Traditionally, virtually all work has used single or multi-gene Sanger sequencing, and traditional data types have included microsatellites, mitochondrial DNA (mtDNA) and nuclear DNA sequence analysis. While Sanger sequencing is still considered the gold standard with regard to per-nucleotide accuracy, the amount of data generated in Sanger sequencing is limited due to the cost (it is very expensive on a per-nucleotide basis), labor, and time that such analyses often take to complete. Much more recently, several new techniques that take advantage of massively parallel next generation sequencing (NGS) platforms have begun to replace traditional Sanger sequencing approaches. These new approaches almost invariably rely on Illumina NGS technologies, and sequence much larger fractions of the genome in a single, highly parallelized sequencing experiment. Restriction site associated DNA (RAD) sequencing and targeted sequence enrichment are two recent advances that seek to isolate, sequence and analyze a consistent subset of the genome from each individual, often yielding several thousand informative SNPs. Microsatellites and RADseq are very useful for population genetic studies and can generate small (most microsatellite analyses) to larger (RADseq) amounts of informative data. However, both suffer from problems with missing data, null alleles (particularly for

microsatellites), and a general lack of information about the genomic distribution of informative markers. Targeted sequence enrichment is a technically more complex approach, and relies on having existing genomic resources that allow a researcher to identify ("target") specific genomic regions to study. Target enrichment studies can also generate thousands of SNPs whose physical location and linkage relationships within that genome are known, but at relatively high cost in comparison to RADseq.

The newest, and most radical, approach, and the one that we employed in this study, uses whole-genome sequencing to characterize the entire genome of each individual. Recently, sequencing technology has progressed to the point where entire genomes can be sequenced, at low coverage, for prices accessible to wildlife researchers. Uncertainty associated with the genotype of any SNP at any particular position in the genome in low coverage, whole genome sequencing is much greater than for other approaches, because RADseq and target capture generally sequence each site many times. However, whole genome sequencing provides information for many orders of magnitude more sites than microsatellites, RADseq or target capture, and the statistical power gained appears to far outweigh the uncertainty from low-coverage, whole genome approaches.

Methods

Sampling

Desert tortoise blood samples were obtained in the field between 2004 and 2013. Our samples come from two sources. First, we contacted all individuals who have published papers using Mojave desert tortoise DNA to determine research groups that still have tortoise samples and to explore potential collaborations. We identified Professor Richard (Dick) Tracy, including current and past members of his research team Bridgette Hagerty, Fran Sandmeier, and Chava Weitzman, as a group interested in working together in a collaborative framework. We also were given full access to all material stored at the DTCC. We used existing georeferenced data to map all available tissues, and based on a visual assessment of those samples and a day-long discussion with Dr. Kristin Berry (USGS Western Ecological Research Center) on high-priority areas in need of analysis, we chose 270 blood samples for genomic analysis (see Figure 14). Blood samples had been stored in tubes by themselves or in RNAlater (Life Technologies) and kept frozen at -80C, or as dried drops on filter paper at room temperature. All blood samples were transferred to the Shaffer lab at UCLA for genetic analysis.

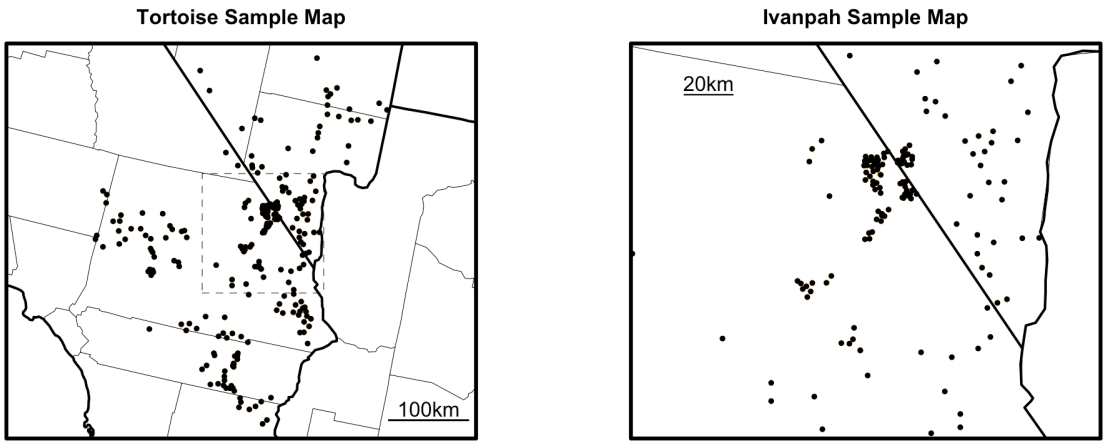

Figure 14: Sample map

Laboratory Procedures

DNA was extracted using a salt extraction protocol (Sambrook *et al.* 2001), and quantified using the Quant-It dsDNA kit (Invitrogen). After diluting to 30ng/uL, DNA extractions were physically sheared to approximately 200-600bp using a BioRuptor (Diagenode) with an average of 7 cycles on the highest setting (30 seconds on, 90 seconds off). Sheared DNA was cleaned using a SeraPure bead mixture to remove chemical contaminants and to exclude short DNA fragments, which reduce sequencing efficiency. Illumina sequencing adapters were ligated to fragmented DNA using a standard library preparation kit (Kapa BioSystems). Each sample received one of 10 adapters with distinct 10 base pair indexes that allowed for samples from different tortoises to be combined, multiplexed in a single sequencing lane, and later computationally separated back into individual-tortoise data. All indexes had an edit distance of at least three to other adapters to allow for sequencing errors in the index reads (Faircloth and Glenn 2012). Importantly, the protocol did not include any PCR amplification, which is known to introduce and amplify biases in the resulting sequence data (Aird *et al.* 2011).

A total of 270 tortoises were submitted for sequencing at the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley in three batches. Sample libraries were quantitated using qPCR and pooled, 10 samples per lane, for whole genome sequencing. All samples were sequenced in 100bp paired-end mode on an Illumina HiSeq 2000 or HiSeq 2500. Data for individual samples were de-multiplexed using Illumina's CASAVA pipeline and downloaded in FASTQ format for analysis.

Genomic Data

We developed a multi-step pipeline to remove low quality data, consistent with current best practices in processing genomic data. Particularly for low-coverage data (our data were approximately 1.5X coverage, which is very low), this is a critical step. First, reads that failed

Illumina's CASAVA filter were removed. Next, each sequencing read was checked for adapter contamination and base call quality degradation. Adapter contamination arises when the fragment being sequenced is shorter than the read length, resulting in bases being called at the 3' end of fragments that are actually part of the synthetic sequencing adapters rather than the tortoise's genome. Base call quality degradation occurs because the quality of base calls often deteriorates from 5' to 3' on Illumina sequencing reads (Fuller *et al.* 2009). To account for both of these factors, reads were processed using Trimmomatic 0.32 (Bolger e*t al.* 2014). Specifically, leading 5' base pairs with a phred quality score (a standard metric of probability of accurate base-calling) below 5 were removed, and trailing 3' base pairs with a phred quality score below 15 were removed. Then, a four base pair window was moved from 5' to 3' along each read, and the read was trimmed when the average phred base quality within the window dropped below 20. After this trimming, all reads less than 40bp in length were discarded.

Following sequence trimming, overlapping read pairs were merged using fastq-join from the ea-utils toolkit (Aronesty 2011). (Paired-end reads simply means that each 200-600 base pair DNA fragment was sequenced for 100 base pairs from both ends of the fragment, rather than only one 100 base pair read from the 5' end.) For short fragments, paired-end reads will overlap when the total length of the fragment being sequenced is less than two times the read length; merging these reads prevents artificially inflating sequencing coverage estimates where reads overlap, and results in improved mapping efficiency for these reads. Joined read pairs were then combined with singleton reads whose mate pairs were discarded in earlier quality control steps. This resulted in a set of paired reads and a set of singleton reads for each tortoise.

Paired reads and singleton reads were separately mapped to a draft of the Galapagos tortoise (*Chelonoidis nigra*) genome supplied by the laboratory of Dr. Adalgisa Caccone at Yale University. Ideally, reads would be mapped to a Mojave desert tortoise genome, since that would allow the maximum number of reads to be identified and physically arranged along the species' genome. However, a high-quality genome for the Mojave desert tortoise does not currently exist (as part of this project, we have initiated a collaboration with a group from Arizona State University that is producing this genomic resource). Mapping was done with bwa mem version 0.7.10-r998-dirty (Li 2013). Sequence alignment map (SAM) files output from bwa mem were converted to binary alignment map (BAM) and the paired and singleton alignment files for each tortoise were merged into a single alignment file for each tortoise using samtools version 1.0 (Li *et al.* 2009).

Merged BAM files were then cleaned to soft-clip alignments that extended past the end of reference contigs (CleanSam) and individual tortoise read group information was added (AddOrReplaceReadGroups) using picard 1.119 (http://broadinstittute.github.io/picard). Duplicates were then marked to identify levels of optical duplication (single molecule colonies on an Illumina flow cell that are mistakenly identified as multiple reads and can inflate coverage estimates). It was not necessary to mark or remove PCR duplicates because we utilized a PCR-free laboratory protocol, eliminating this potential source of error. Mapping rates were then calculated by counting the appropriate alignment flags using samtools flagstat (Li *et al.* 2009).

Modeling whole-genome sequencing

Given that no study has, to our knowledge, used low-coverage, whole-genome sequencing for large-scale population genomics of an endangered species, we first conducted a simulation study to compare RADseq vs. whole-genome approaches. We simulated datasets from two hypothetical populations separated by a true genetic distance of Fst=0.001. That is, for this simulation, 0.1% of the total genetic variation was between these two populations, and 99.9% was within populations. In the first simulation, we sampled two thousand polymorphic loci (SNPs), at 20X coverage (typical for RAD sequencing or target capture). In the second, we simulated one million polymorphic loci, at 1X coverage (a low, and therefore conservative number of SNPs for whole-genome sequencing). The low, but real level of genetic divergence between these simulated populations was not recoverable using the first dataset, but was confidently inferred using the second (Figure 15). This result is consistent with a known guideline for population genetic analyses (Patterson and Reich 2006): genetic differentiation (measured by Fst) between two groups or individuals becomes detectable if it exceeds 1/sqrt(n*m), where m is the number of variable markers (~1 million in our simulation) and n is the number of individuals. Based on these results, we were convinced that we should pursue a low-coverage, whole genome approach to best quantify Fst among individual tortoises.

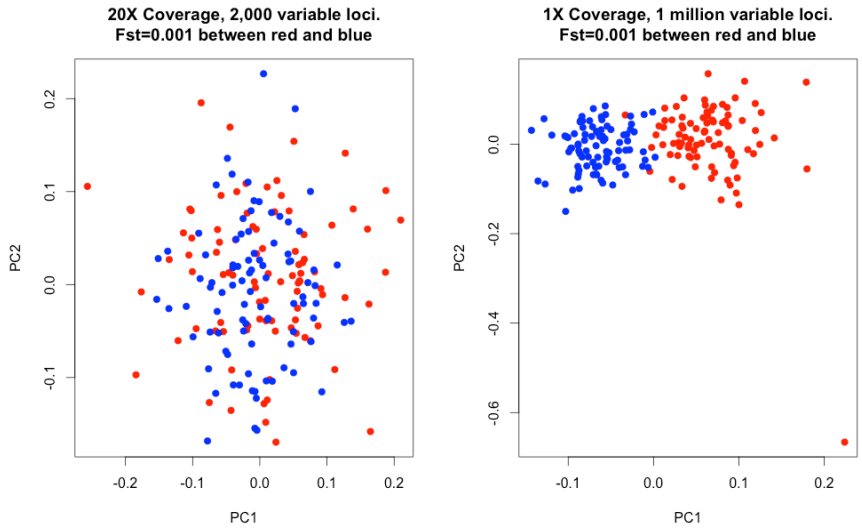

Figure 15. Comparison of two different sequencing approaches in their ability to differentiate very slightly differentiated populations (Fst=0.001)

Inference of Genetic Relationships

In this study, the individual tortoise comprised our sampling unit. To determine the relationships between the sampled tortoises, we estimated two quantities for each pair of tortoises: pairwise sequence divergence, and genotype covariance. Pairwise sequence divergence is the average density of sites at which the two sequences differ, and is hence proportional to the average time back to the most recent common ancestor of the samples, averaged across the

genome. Genotype covariance, on the other hand, decreases with average time back to the most recent common ancestor (Slatkin 1991), and is the basis for several widely used visualization methods.

To estimate these quantities, we used *angsd (http://popgen.dk/wiki/index.php/ANGSD)*, an existing set of computational tools designed to incorporate uncertainty in genotype calls deriving from low-coverage sequencing data (Kim *et al.* 2011, Korneliussen *et al.* 2014, Li *et al.* 2010). This software provided us with reliable lists of polymorphic sites (SNPs), but unfortunately, we found that genotype posterior probabilities generated by *angsd* were influenced by both the variation in sequencing depth between samples and the distance of a given sample to the rest of our tortoise samples.

Both of these influences can lead to incorrect population inferences about tortoise biology, and therefore require correction. To do so, we developed a new method that instead uses raw read counts and is robust to differences in sequencing depth between individuals. To calculate divergence between a pair of tortoises in a way that is not influenced by sequencing depth, we estimated the probability for each base pair that two homologous reads drawn from the two tortoises are different at that base and averaged this across the genome, weighting by the read depths in those tortoises at that site (Appendix 2). Using this method, pairwise genetic divergence is not correlated with sequencing depth. For the following analyses we computed these pairwise sequence divergences using the full list of 52,740,529 sites determined to be polymorphic by *angsd* with a p-value less than 1e-6 ($p<0.000001$) and for which no tortoises had a read depth greater than 10 (to avoid overweighting repetitive regions, which can also skew summary statistics). These were then corrected to the proper genomic scale by multiplying by the density of polymorphic sites (an average of 2.98% in the 1.899 billion relevant bases of the reference genome).

The mean sequence divergence between two sequences provides an estimate of the mean time since they shared a most recent common ancestor, averaged across the sequence and multiplied by twice the average substitution rate (Hudson 2007). To make our results more interpretable, for the purposes of fitting models we converted sequence divergences to years by dividing by an estimate of twice the average nucleotide substitution rate. The substitution rate was estimated by dividing the pairwise sequence divergence for a large set of genes between a tortoise (*Manouria emys*) and the painted turtle (*Chrysemys picta*) by a fossil-calibrated divergence time estimate between those two lineages. These two values were derived from a different large-scale turtle genomics project ongoing in the Shaffer lab. This estimate is probably not a completely accurate estimate of the true mean substitution rate for the desert tortoise, but is by far the best estimate currently available for turtles and tortoises from this related group of species, including the desert tortoise. It provides a reasonable, albeit approximate, idea of the time scales involved in our estimates.

We pursued multiple avenues of visualization and analysis of population structure. To investigate the geographic structure of genetic variation, we compared and plotted average pairwise sequence divergence against pairwise Euclidean distance between samples. In addition,

we performed a principal component analysis (PCA) using the sample genetic covariance matrix, and obtained simple 'geogenetic maps' (inset of Figure 17) by plotting different PC axes against each other (Menozzi *et al.* 1978, Patterson and Reich 2006, Novembre *et al.* 2008). We also plotted the first several principal component scores for each tortoise onto elevational maps to visualize how tortoise genetic differentiation was distributed across their actual range (Figures S2 and S3).

Isolation by Environment

It may well be that knowing simply where tortoises do not go (e.g. up steep mountains) suffices to describe gene flow across the species range. However, there are good reasons to suspect that other environmental factors have substantial effects. For instance, if overall habitat quality varies across the range so that there are "source" and "sink" populations, then we expect "source" populations to harbor more genetic diversity and to potentially serve as hubs connecting the "sink" populations. On the other hand, if offspring dispersal is biased such that young tortoises tend to end up in in habitats similar to their parents (beyond the correlation implied by localized dispersal), then gene flow between regions with different environmental variables will be reduced. This would imply that not only geographic distance but also ecological similarity predicts genetic differentiation, a pattern widely observed in nature (Sexton *et al.* 2013).

The current state-of-the-art method for making predictions of gene flow on continuous landscapes is to compute so-called *resistance distances* (McRae and Beier 2007). The nomenclature and formalism of this approach derives from a mathematical correspondence between electrical networks and certain quantities of reversible random walks. It turns out that if one equates the movement rates of a random walk between nodes in a network with the conductances of wires connecting those nodes, then the *effective resistance* between two points of the network (what one would measure using a volt meter) is equal to a biologically important parameter known as the *mean commute time* for the random walk, i.e. the mean time until a random walker, beginning at one of the points, first returns to its starting point, after having visited the other point (Nash-Williams 1959). The results can depend on the discretization used and the resulting random walk model is metaphorical, not predictive. We used the fact that this correspondence, usually stated for discrete networks, carries over to continuous models, where the random walk is replaced by its continuous counterpart, a diffusion process whose movement rates depend on local properties of the inhomogeneous medium (in our case, the local landscape and its quality as tortoise habitat). This, combined with robust approximation of discrete random walks with continuous diffusions (Oblój 2004), allows us to bypass both drawbacks. The resulting resistance distance is then a powerful summary of gene flow across the landscape, since it integrates movement along all possible paths between the two locations.

The resistance distance has been shown to be a useful summary, but we would like to extract concrete predictions from it for effective management decision-making. Each generation since the most recent common ancestor provides an opportunity for mutations to occur that are inherited by only one of the sequences, and so mean sequence divergence provides an estimate of

the mutation rate multiplied by the average time since the most recent common ancestor, across the genome (Hudson 2007). Focusing for a moment on a particular point on the genome, the time since the most recent common ancestor of two sequences for that part of the genome can be found by following the lineages back until they meet in their most recent common ancestor. Following a lineage backwards in this way can be seen as a random walk: the probability the lineage moves from location x to location y in a generation is the probability that a tortoise living at x has inherited the relevant bit of genome from a parent living at y. Intuitively, the motion of a lineage backwards in time looks like a random walk that is determined by the dispersal patterns of young tortoises, except that going backward in time, lineages are more likely to move towards better habitat, since more successful offspring are produced in such places. One important caveat is that it is known that under reasonable population models – in particular, those that show significant patterns of isolation by distance – the motions of two nearby lineages are not independent and therefore require a model that incorporates this non-independence (Barton, Depaulis, and Etheridge 2002). However, it is reasonable to assume that the motion of lineages is independent until the point that they are sufficiently close to each other in their path backwards to a common ancestor. Then, we can decompose the time to most recent common ancestor into two parts: the time until two lineages are close to each other, and the time from when they are close to each other until they find a common ancestor. This first part determines how sequence divergence decreases with distance, while the second part determines typical divergences between nearby individuals.

To relate this to resistance distance, we approximate the mean time until two lineages are close to each other by the average commute time. Specifically, we approximate the mean time until the lineages of tortoises at current locations x and y are within distance d of each other by one-half the sum of the mean time that a random walk begun at x takes to get within distance d of y, and the same quantity for y with respect to x. If the landscape is homogeneous then this approximation is exact, since the displacement between two independent walks is itself a walk that moves at twice the speed. On an inhomogeneous landscape it is a reasonable approximation, except in extreme circumstances (like very strong barriers to movement).

We provide the details and formal specification of this model of landscape resistance in Appendix 3. Briefly, we defined a random walk model whereby environmental rasters were each given two parameters that affect movement rates in the random walk: a stationary distribution and a relative jump rate to adjacent pixels. The application of a given set of stationary distribution and relative jump rate parameters (as well as a single overall scaling parameter) generates a resistance surface on the landscape over which commute times between tortoises can be measured. Since commute times are proportional to the coalescence times for pairs of tortoises, we can evaluate the model by testing how well random walk commute times over the generated resistance surface correlate with observed genetic distances. The optimal parameters for a set of landscape rasters are determined by minimizing the weighted mean square error for the set of tortoises used to fit the model.

Landscape Resistance Models

We used the above procedure to fit a large number of landscape models, varying which landscape layers were used, which tortoises were used to fit the model, and which habitat mask was used. In all cases, we masked (that is, eliminated) regions to the east and south of the Colorado River, since they are now considered to be in the range of *Gopherus morafkai*. For reference, the tortoise habitat model of Nussear *et al.* (2009) fit a maxent model using 16 landscape variables, of which the most important were elevation (59.7%) and annual growth potential (AGP, 19.3%).

Each model fitting procedure produced a random walk model of tortoise lineage movement, which we evaluated in a common framework, measuring model fit to all tortoises using weighted median residuals. For an exact description, see Appendix 3 under Evaluating Landscape Resistance Models. We used median, rather than mean-squared, residuals to reduce the effect of statistical (and biological) outliers, and we weighted these so that the measure of goodness of fit assigns appropriate weights to each geographic area (unweighted would significantly upweight locations with more samples). Furthermore, we only use comparisons *within* each of the two major regions that we identified (north and south of the Ivanpah), because, as argued below in the results section on population structure, the relationship between the two regions has a deep-time historical component that is not likely to be a product of temporally homogeneous tortoise movement.

We evaluated a large number of possible landscape resistance models. Below is a quick summary of the procedure that led us to the best-fitting model.

First, we found that models fit using tortoises from both regions (loosely North and South) performed poorly: none could explain the two-cloud pattern seen in Figure 18. This is not surprising, because no available landscape layer accurately differentiates between those two regions. There is a confluence of not-insubstantial physical barriers around the break between the two regions (the mountains that define the Ivanpah Valley and the Colorado River), but the constriction in tortoise passage induced by these appears to not be sufficient to cause the genetic discontinuity that we detected. Furthermore, remaining tortoise population structure is seen to be much more significant in the north than in the south, and combining them into a single analysis confounds these differences. To deal with these differences we proceeded by fitting models using only comparisons between tortoises in the same group (north-north, or south-south comparisons). This is reasonable because we expect nearby comparisons to provide more information about local movement patterns than comparisons between tortoises on opposite sides of the range.

Next, we evaluated the effects of the choice of *habitat mask*, i.e. the region where movement was allowed to occur. We compared two choices: (a) the region for which the habitat model of Nussear *et al.* (2009) had habitat score above zero; and (b) the region below 2,000m in elevation. The first mask is strictly contained within the second; in both cases we also restricted to a reasonable bounding box (see the extent of the elevation layer in Figure 17). We found that the two different choices of mask gave indistinguishable goodness-of-fit values, and so proceeded with (a), the habitat mask based on Nussear *et al.* (2009), as this represents the best available

biological prior knowledge of areas where tortoises are likely to avoid or perish, and which therefore should be excluded from our model.

Finally, we examined the impact of including different habitat layers in the model. We explored a wide variety of layers, but ultimately the best-fitting models all included only transformations of the habitat quality derived from Nussear *et al.* (2009). Therefore, we chose as our current best-fitting model the one providing the best goodness-of-fit using only transformations of the Nussear *et al.* (2009) habitat quality. (As discussed below, other models, including those with longitude, gave very similar results.) We favor this both because of its relative statistical simplicity and because it keeps our landscape model closely linked to the best available habitat model as derived by the desert tortoise biological community.

Evaluation of Alternatives

We then used the best-fitting model to evaluate how development of particular areas under the DRECP would affect gene flow between different areas of the tortoise range. To do this, we evaluated changes in gene flow between each pair of a large set of reference points spread uniformly across the range predicted by Nussear *et al*. (2009), and we then used these to quantify both the overall reduction in gene flow and the areas that would be most affected (more details below). Some analyses considered "chunks" of proposed development areas within each of the five Alternatives separately. The process by which we generated these proposed development chunks is described in Appendix 4. In modeling how development on these "chunks" would affect gene flow, we assume that they represent zones of inaccessible habitat for tortoises, in the same way that areas outside the range boundary are modeled as inaccessible. Under our modeling strategy, a tortoise that wandered into a chunk boundary would reflect off of that boundary, much as it would if the development were surrounded by an impenetrable fence. Other modeling strategies are possible, and reasonable ones might include a pure mortality scenario (where tortoises have free access to development chunks, but die upon entry) or a semi-permeable boundary (where some fraction of tortoises can cross the chunk). Given the uncertainty in exactly how development might occur in each chunk, we feel that our approach is a reasonable starting point, since it has minimal effects on demography (tortoises do not die when they reach a chunk boundary) but reasonable effects on gene flow (tortoises presumably cannot cross a large solar installation).

To quantify gene flow, we used the mean commute time to a 15-km circle (or neighborhood), since this is the same quantity used to fit the model. As discussed below, for a pair of points x and y, this is equal to one-half the sum of the mean time for a random walk from x to get within 15 km of y, and the mean time for a walk from y to get within 15 km of x. This can be concretely interpreted as the mean time since a tortoise at one location has inherited genetic material from a tortoise near the other location, along a particular lineage. Note that the neighborhood approach makes this measure independent of population density.

Reference Locations

Our samples of tortoise tissue were not distributed uniformly across the range, and uneven sampling can have profound effects on some population genetic estimates and biological interpretations. To evaluate the effects of our sampling in an integrated way across the entire range, we chose uniformly spread reference locations as follows. First, we found the area with habitat quality of at least 0.3 in the Nussear *et al.* (2009) model, since those represented relatively high-quality tortoise habitat. Then, we sampled 10,000 points uniformly from across the enclosing rectangle, and discarded all but a maximal set of points that fell within the area of high habitat quality and had no two points within 10km of each other. This resulted in 202 points uniformly spread across the area of high-quality habitat. We additionally removed those points predicted by our model to be in isolated areas, defined as the minimal set of reference points such that after removing them, all remaining mean 15 km commute times were smaller than 3×10^6 years (the maximum observed divergence between any pair of samples was slightly less than 1.5×10^6, so a distance of 3×10^6 would be equivalent to a separation of twice the width of the current range). The remaining points, shown on a map of habitat quality from Nussear *et al.* (2009), are shown in Figure 16.

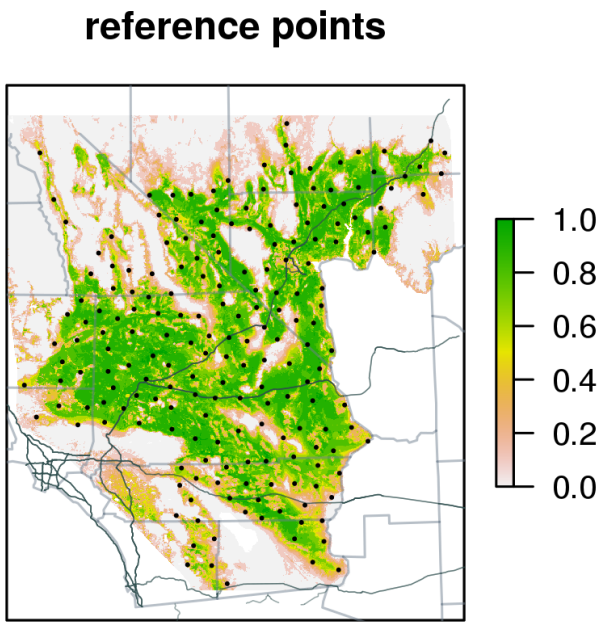

Figure 16. Reference points used to compute changes in gene flow across desert tortoise habitat.

Measure of Isolation

The mean commute times described above allow us to quantify the effect that particular development scenarios will have on gene flow between any pair of locations in the range. However, this is not yet a measure of *isolation*. To quantify isolation, we need to summarize the total effects on each location across all of our sample points shown above. Consider, for instance, what would happen if a valley were to be blocked off to tortoises from the outside: mean commute times between the valley and the outside would drastically increase, but mean

commute times within the valley would decrease, since tortoises within the valley can no longer take longer commutes outside the valley. Furthermore, and as we will see below occurs in practice, the act of removing a piece of habitat usually *reduces* commute times (or increases gene flow) between distant locations, because there are now fewer locations for transiting tortoises to visit. However, these commute times to very distant locations are not biologically relevant or important, at least on the time scale of human-mediated disturbances, simply because it takes thousands or millions of years for genes to commute to these distant locations, and that is not the scope of concern of our analyses. For these reasons, we say that a location becomes more isolated if mean commute time increases *to the bulk of the range*. We quantify this by identifying for each reference location the closest 40% of other reference locations, measured by commute time, and averaging the difference in commute time induced by removing a particular piece of habitat across those locations. This limits our summary statistics to a biologically reasonable region of space and time.

Concretely, suppose that $\overline{h}_{i,j}$ is the commute time between reference locations i and j, ordered by proximity to location i, so that $\overline{h}_{i,1} \leq \overline{h}_{i,2} \leq \cdots \leq \overline{h}_{i,202}$. Then, since $202 \times 0.4 \approx 80$, our measure of *isolation* of location i is

$$I(i) = \frac{1}{80} \sum_{j=1}^{80} \overline{h}_{i,j}.$$

To interpolate values observed only on a subset of locations (e.g., the isolation values of the reference locations), we fit a thin plate spline model using the function *fastTps* in the *fields* package in R, which uses compactly supported kernels (with range 200km).

Results

Overview

The sequencing strategy we followed produced an immense amount of data, which will serve as a tremendous resource for tortoise biologists, planners, and desert ecologists more generally. Additionally, the genetic data generated by this project will be used to improve the genome assembly of the desert tortoise currently underway by Kenro Kusumi and Dale Denardo at the University of Arizona. When complete, all of our data will be freely available both in raw form and as summary statistics on a desert tortoise genome project web page. Of course, usefulness of data is not measured solely in terabytes. As detailed below, descriptive analyses of the data show that genomic measures of relatedness can identify geographic population structure, revealing both population splits and fine-scale structure on the scale of kilometers.

Sampling

We amassed a collection of 270 tortoise tissue samples from throughout their range in the Mojave Desert (Figure 14). In our sampling, we attempted to balance an even spatial sampling of tortoises (increasing the probability of observing spatial structuring of genetic diversity partitioned by geography or environment) with a dense sampling of tortoises in regions of

conservation importance (which also allows us to observe patterns of genetic differentiation on local spatial scales). The mean distance between a pair of tortoises was ~141.8 km, with three-quarters of the distances less than 199.3 km, and ranging between 0 km and 464.8 km. The sampling was dense: the mean nearest-neighbor distance between tortoises was 6.4 km, three-quarters of the tortoises had another within 8.4 km, and only 10 did not have another tortoise within 20 km.

Genetic Data

We obtained a total of 1.29 trillion base pairs of genomic sequence data from 28 paired-end 100bp Illumina HiSeq High Output lanes. Total bases sequenced per tortoise ranged from 1.71 billion bases to 13.91 billion bases, with a mean of 4.73 billion bases and standard deviation of 1.62 billion bases. Of this raw data, an average of 86.74% of reads passed Illumina's CASAVA filter for each individual tortoise (standard deviation = 4.54%). After trimming low quality reads and merging overlapping read pairs as outlined in "Methods" above, the total number of bases going into the mapping stage ranged from 1.37 billion to 9.93 billion (average = 3.36 billion, sd = 1.12 billion).

Mapping statistics

Mapping reads to the Galapagos tortoise genome was quite successful, with an average mapping rate of 95.67% (sd=0.67%). Using the ~2.2 billion bp reference size of the Galapagos tortoise as a proxy for genome size, this yielded a mean sequencing coverage of 1.45X (sd=0.48, min=0.59X, max=4.28X).

Population Structure

A geographically explicit way of looking at the relationship between genetic and geographic distance is to use PCA to summarize the major axes of genetic variation on a landscape. We show this in several ways. The positions of the samples on the first two principal components are shown in the inset of Figure 17. The most obvious pattern is the division of the samples into two large clusters by PC1, which corresponds to a fairly sharp division between tortoises to the north and south of the New York and Providence mountains (the eastern/southern border of the Ivanpah valley), with a few intermediate tortoises (coded as purple/pink) occurring in the Kelso area and the vicinity of Searchlight, NV. As the insert map of the Ivanpah region shows, these mountains form a strong barrier to tortoise dispersal; as a consequence PC1 accounts for about 12.2% of the total genetic variance in the data set. These two groupings also explain the two clouds of points that are evident in the overall IBD plot in Figure 18; genetic comparisons of pairs of tortoises between these two groups show a significantly higher divergence than comparisons of tortoises at comparable distances within each group.

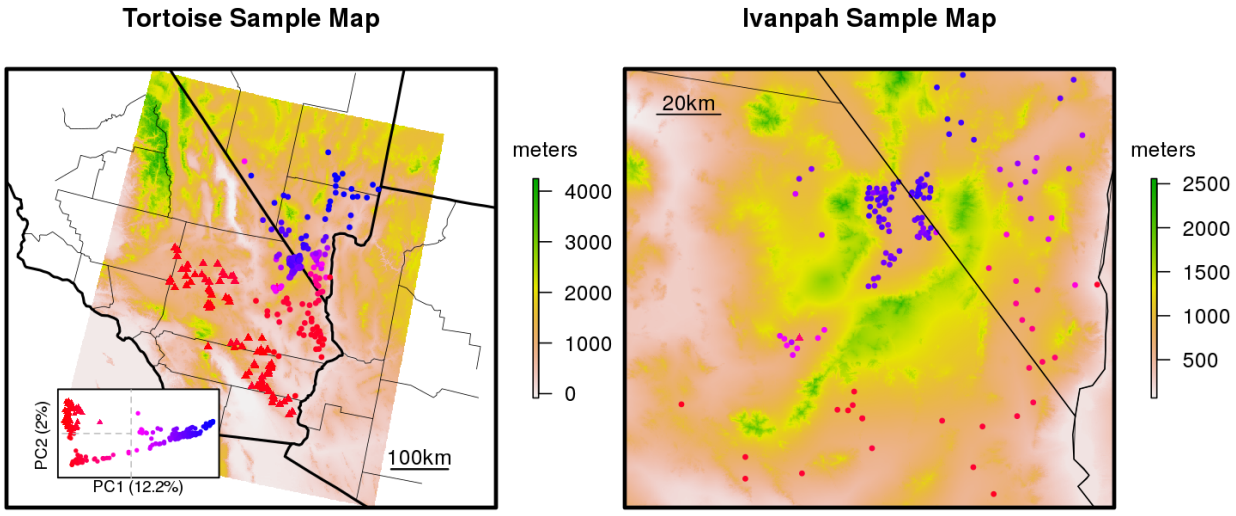

Figure 17. Tortoise sample map with samples colored continuously by their score on PC1. Additionally, samples on the left side of PC1 (mostly the red samples) were divided again into two sets based on PC2, with samples in the top half plotted with triangles and the other samples plotted in circles. The Ivanpah Sample Map is an expansion of this area that shows how genes move around the mountains in more detail. Background colors show elevation.

The first principal component is most striking, but others similarly reflect additional geographical subdivisions. PC2 further subdivides the southern tortoises roughly into eastern and western groups (red triangles vs. circles in Figure 17) on either side of the low-lying Cadiz valley lakebeds and accounts for about 2.0% of the total genetic variance. Subsequent principal components further subdivide the range, generally following geographical barriers such as mountains. These sub-groupings account for additional substructure seen in Figure 18, and represent geographic patterns of differentiation above and beyond that explained by distance alone. Overall, our emerging hypothesis is that relatedness between tortoises is well predicted by distance as traversed by tortoises on the landscape and that the genomic data contain a great deal of information on how to define distance in a biologically meaningful way.

Isolation by Distance

The mean density of nucleotide differences between tortoises ("pairwise divergence") in the sample is 5.4 differing sites per kilobase, and varies between 3.3 and 5.9 sites per kilobase. This measure of relatedness, when compared to geographic distances between tortoise pairs, demonstrates that tortoises sampled nearby each other are more closely related than ones sampled farther away (Figure 18) – the classic "isolation by distance" pattern (IBD; Wright 1943). Overall, pairwise divergence increases by about 0.0011 differences per kilobase for each additional kilometer of separation (Figure 18; $p<10^{-16}$). The "groups" referenced by Figure 18 are shown as discretely colored purple and blue dots in the left panel, and are determined by

discretizing the samples by their scores on PC1 as shown in Figure 17.

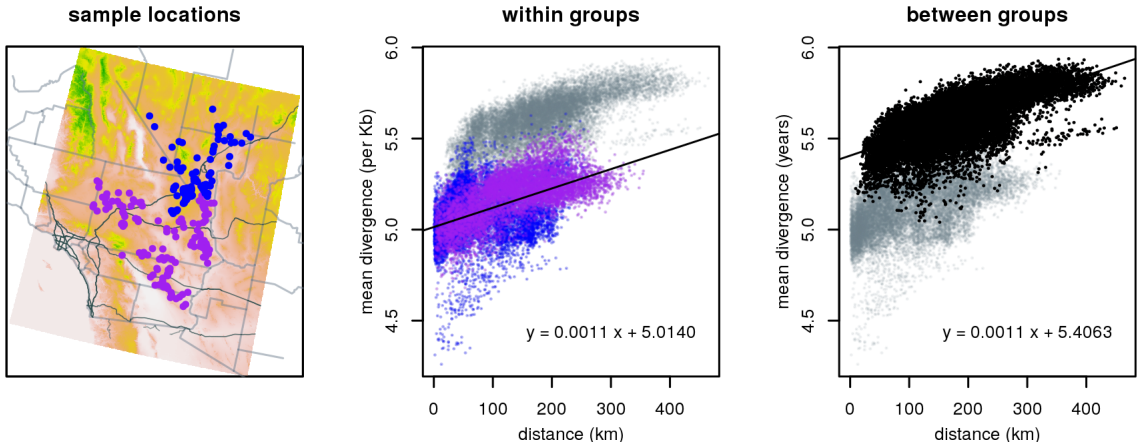

Figure 18. Showing the slope of isolation by distance both within each group and between the two groups. Groups are defined by their scores on PC1 and are shown in the left panel.

This positive correlation of genetic differentiation and geographic distance extends to the smallest spatial scales: within the Ivanpah valley, where the densest cluster of samples occurs, the relationship between pairwise divergence and geographic distance is likewise highly significant, showing an increase of 0.0024 differences per kilobase for every extra kilometer of separation (Figure 19; $p < 10^{-16}$).

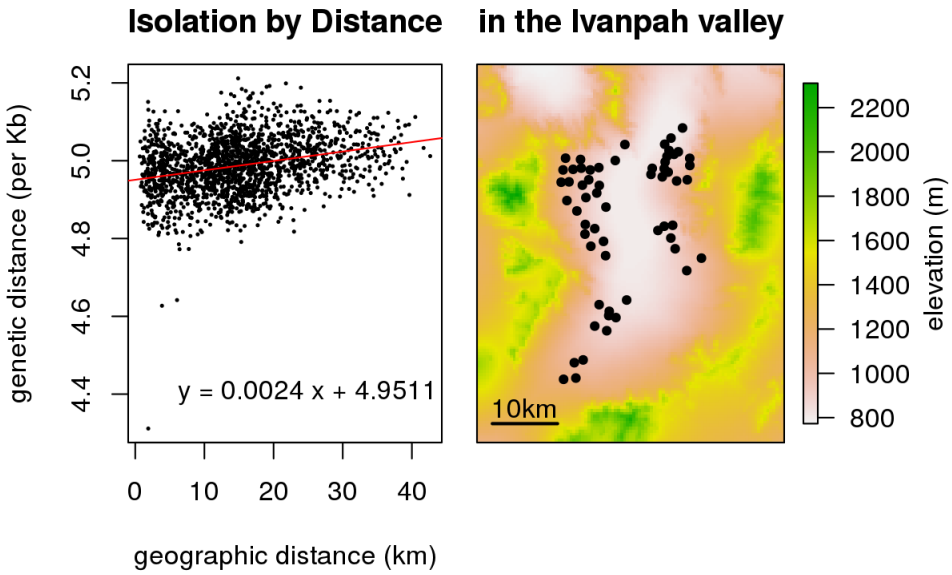

Figure 19. Pattern of isolation by distance solely within a small geographic range in the Ivanpah Valley. The map on the right shows the location of each tortoise sample.

Model fit

Of the many models that we tested, the one with the best fit included the effects of only one layer: a binary layer that takes the value 1 if habitat quality (from Nussear *et al*. 2009) is greater than 0.3, and zero otherwise. (More extensive model fitting results are given below.) This model

allows four rates of tortoise movement: on low quality habitat, on high quality habitat, from low to high quality habitat, and from high to low quality habitat. The weighted median residual value for this model was 2,133 years, while the same quantity for the linear regression of pairwise divergence against great-circle geographic distance was 18.7% greater. The difference measured by weighted mean squared error was even stronger: a 53.7% difference. This indicates that the landscape model incorporating a binary indicator of tortoise habitat quality as defined by Nussear *et al.* (2009) did a significantly better job of predicting genetic relationships between tortoises than did straight-line distance. Although additional, more complex models can and should be tested, we used this two-state model based on its simplicity, its biological realism, and its statistical performance.

Effect on gene flow of removing habitat

As discussed in the *Methods*, we quantify the effects on gene flow of removing particular pieces of habitat through the changes in mean time for the random walk that models the time, in years or generations, that it takes for tortoise lineages to travel between each pair of points, averaged across reference locations. **These analyses provide the tools by which we can directly test the effects on tortoise connectivity of alternative habitat modification plans as outlined in the DRECP.**

Effects on gene flow to single locations: examples

To show how this approach works, we consider how the removal of all of the habitat in the Preferred Alternative Plan would affect gene flow to a single location. In the left panel of Figure 20, the star located in the far western Mojave is the single location, and each map pixel is colored according to the mean time to reach the 15km circle surrounding the star in the map. Unsurprisingly, it takes longer to reach locations that are more distant from the star. On this map, the potential development areas of the DRECP Preferred Alternative are shown in grey, and the mapped mean times have been computed after blocking these areas to possible tortoise movement. The middle panel shows how this differs from the scenario where these development areas are not blocked: each area is colored according to the difference between the mean time to reach the starred area before and removing the development areas. As this map demonstrates, most parts of the range are around 40,000 years more distant (in red), although the nearby area that is also trapped between two potential development areas becomes slightly closer (in pale blue). The right panel shows the relative change: here, colors correspond to the difference (middle panel) divided by the mean time without the potential development areas removed. As might be expected, the relative effect is greatest near the star, since more distant areas are relatively less affected by a change near the star.

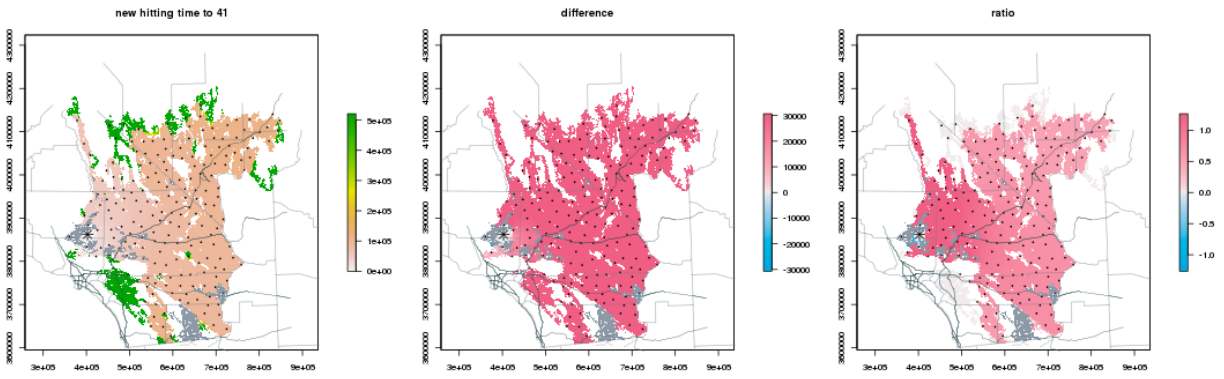

Figure 20. Showing the effects on hitting times to a single spot in the western Mojave as a result of removing the land in the preferred alternative. The dark green areas in the left panel were deemed inaccessible under this model.

Figure 21 shows the same set of analyses for a reference location near the center of the Mojave, again marked with a star. Here, we see that most of the Mojave actually becomes *closer* (blue in the center panel): this is because removing a portion of habitat means that there are fewer available locations for ancestors to live, and so all else being equal, two tortoises are expected to have ancestors living nearby to each other more recently. However, note that there is a small "shadow" of increased distance (red) just on the other side of a nearby potential development area, reflecting the reduced regional gene flow in this area that would be induced by blocking off this piece of habitat.

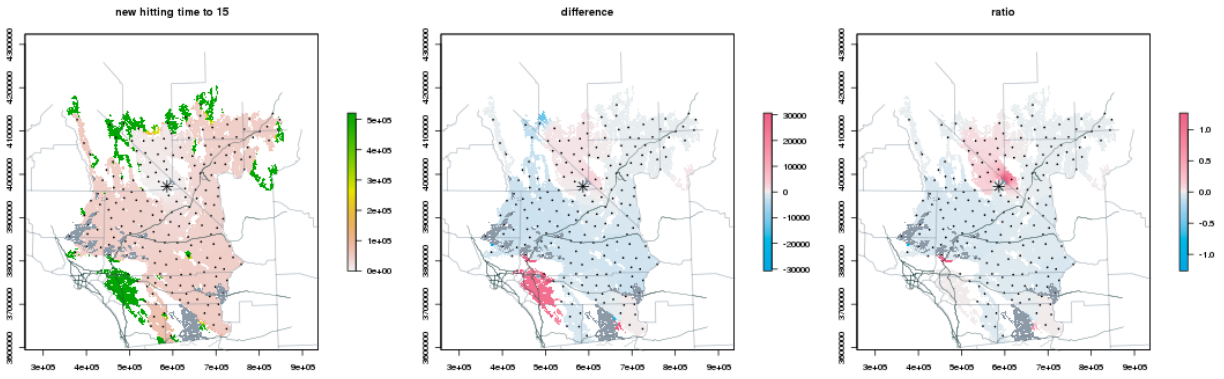

Figure 21. Showing the effects of hitting times to a single spot in the central Mojave as a result of removing the land in the preferred alternative.

Combined effects on gene flow across the range

To summarize the effects on gene flow of blocking off particular regions of the habitat, we average the difference in gene flow with and without the possible barriers, across the nearest 40% of the other reference locations. We chose the closest 40% as a reasonable compromise between the entire range of high-quality habitat of the Mojave (which is too large to reasonably affect gene flow for a tortoise) and a region immediately surrounding an animal (which does not allow for the cascading effects across generations of blocking gene flow). We computed this measure of isolation for each reference location and interpolated it to the remainder of the map;

this is shown on the left of Figure 22. We refer to this statistic as the *mean difference nearby*; it quantifies the mean difference in gene flow for the biologically relevant 40% of tortoise habitat with and without a subset of habitat removed. On the right we show the *relative difference nearby*, calculated as the mean (over the same 40% of locations) of the ratio of this difference in gene flow to the gene flow (commute time) in the original habitat without the barriers. These figures show the predicted impacts of removing the proposed development chunks of land in the DRECP Preferred Alternative.

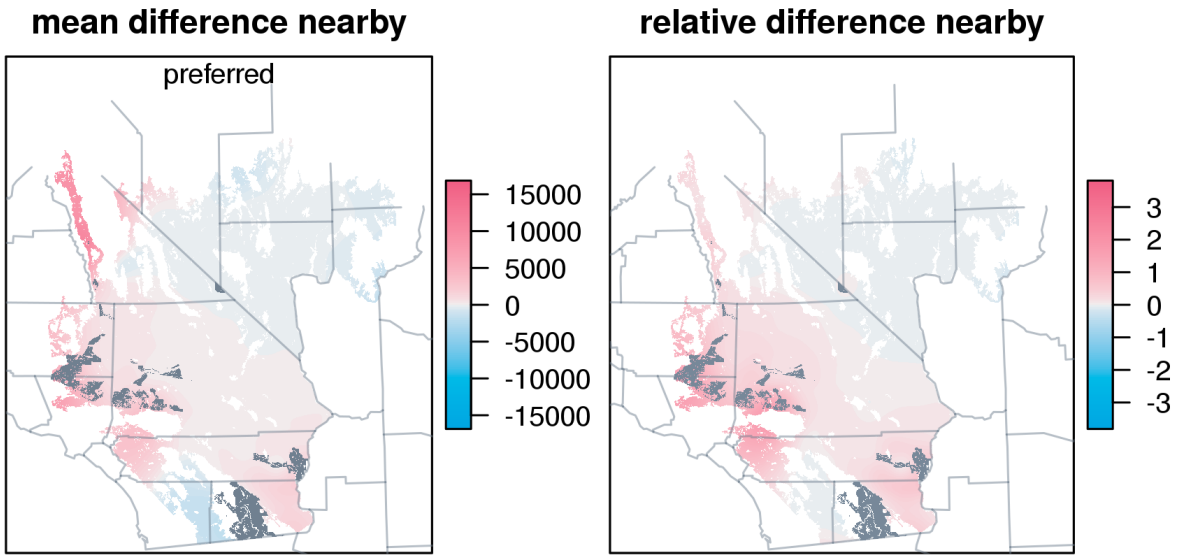

Figure 22. Mean and relative difference in commute times across the range as a result of removing the land in the preferred alternative.

In both maps, the darkest red areas are more distant from other, nearby portions of the range by 10,000-15,000 years. This is a very strong separation, because most parts of the range are separated by less than 10,000 years, as seen in the example commute time plots above.

Comparing total effects of each alternative

We can now apply this same approach to each of the four Alternative plans in the DRECP, and compare them to the Preferred Alternative. We plot these in Figures S4-S7.

In order to be able to rank these alternative development plans, we show several summary statistics for each. For each alternative, areas are given in km2 and as a percentage of the total tortoise habitat. We calculated and tabulate each of the following:

- *habitat removed* is the total amount of area either in possible development areas or completely isolated from the rest of tortoise habitat under that alternative (this occurs if, for example, land is developed in a ring, with an undeveloped hole in the center of the development)

- *isolated* is the total area for which the gene flow to nearby areas has increased (regardless of the amount by which it has increased)

- *isolation* is the mean amount by which the commute time has increased to nearby areas across this area where it has increased; **it measures the intensity of decrease in gene flow**

- *strongly isolated* is the total area to which gene flow has strongly decreased; we define "strongly" as the mean commute time to nearby areas increasing by at least 1,500 years

- ***relative isolation* is the ratio of the amount by which commute time has increased to the commute time without blocking any areas, averaged over the set of nearby locations**

Alternative	Habitat removed (km^2)	Habitat removed (%)	Isolated (km2)	Isolated (%)	Isolation (years)	Strongly isolated (km2)	Strongly isolated (%)	Relative isolation
Preferred	5061	(3.7%)	74658	(54.2%)	729	5303	(3.8%)	18%
Alternative 1	2338	(1.7%)	65677	(47.7%)	643	2864	(2.1%)	9%
Alternative 2	6772	(4.9%)	92242	(66.9%)	950	14422	(10.5%)	26%
Alternative 3	3401	(2.5%)	70976	(51.5%)	882	4416	(3.2%)	16%
Alternative 4	4458	(3.2%)	71017	(51.5%)	760	5681	(4.1%)	17%

Table 7: Effects of development alternatives on tortoise gene flow.

Several points are worth noting in Table 7. First, as a single summary statistic of the effect of a development alternative, we highlight the Isolation (in years) column of the table, which summarizes the overall increase in isolation, or the decrease in gene flow, for each alterative. Second, every alternative increases the time for genes to traverse the landscape by many hundreds of years, approaching 1000 years for Alternative 2. These are large numbers representing very significant effects across dozens of tortoise generations. Third, Alternative 1 is clearly the least harmful (which makes sense given that it is the smallest acreage), and Alternative 2 is the most harmful. And fourth, the Preferred Alternative has a very substantial effect on tortoise connectivity (729 years).

We also call attention to the Relative Isolation, particularly in comparison to the percentage of habitat removed. In most cases, the effect in terms of Relative Isolation is about five times the percentage of habitat removed, reflecting the extremely strong, cascading effects that development has on tortoise movement. For example, for the Preferred Alternative, removing 3.7% of the tortoise habitat leads to an 18% increase in relative isolation.

Effects of removing each chunk

Using the same analytical approach, we also consider the effects of removing each "chunk" of habitat for the Preferred Alternative. Figure 23 is a key showing where each chunk is found under the Preferred Alternative. The *mean isolation* is a reasonable measure of the total effect of removing a given chunk, but it must be interpreted with some caution. In particular, the absolute size of the *mean isolation* **only** measures this chunk, in isolation, without the effects of any other chunk that might be removed. There are many instances where the impacts of multiple chunks considered together have a much larger impact than the sum of the chunks individually,

reflecting the synergistic negative effects that can result when multiple chunks are removed. We provide this table to show how individually chunks may have very different impacts; to understand the impacts of removing a set of chunks, **each potential combination of removal areas must be modeled and evaluated**, and we have not done that here.

We calculated and tabulate each of the following:

- *habitat removed* is the total amount of area either in possible development areas or completely isolated from the rest of tortoise habitat for this chunk
- *isolated* is the total area over which the gene flow to nearby areas has increased
- *mean isolation* is the mean amount by which the commute time has increased to nearby areas across this area where it has increased; it reflects the decrease in gene flow for each chunk in the analysis
- *max isolation* is the maximum amount by which the commute time has increased between any two nearby reference locations given the removal of this chunk

	Habitat removed (km2)	Habitat removed (%)	Isolated (km2)	Isolated (%)	Mean Isolation (years)	Max isolation (years)
Whole preferred alternative	5061	(3.7%)	74658	(54.2%)	729	15787
chunk 31	55	(0.0%)	3129	(2.3%)	10454	11002
chunk 32	6	(0.0%)	3129	(2.3%)	549	847
chunk 25	987	(0.7%)	4406	(3.2%)	126	890
chunk 7	481	(0.3%)	41928	(30.4%)	110	1865
chunk 13	299	(0.2%)	87719	(63.6%)	65	1627
chunk 16	441	(0.3%)	34920	(25.3%)	60	1758
chunk 14	512	(0.4%)	38947	(28.3%)	57	1069
chunk 5	148	(0.1%)	64315	(46.7%)	51	994
chunk 27	336	(0.2%)	34510	(25.0%)	51	867
chunk 11	204	(0.1%)	71582	(51.9%)	40	1066
chunk 4	128	(0.1%)	54457	(39.5%)	29	726
chunk 29	95	(0.1%)	39807	(28.9%)	24	555
chunk 17	173	(0.1%)	60235	(43.7%)	20	277
chunk 28	128	(0.1%)	40702	(29.5%)	20	399
chunk 18	219	(0.2%)	95912	(69.6%)	19	371
chunk 30	121	(0.1%)	41658	(30.2%)	15	298
chunk 22	98	(0.1%)	34762	(25.2%)	10	269
chunk 8	108	(0.1%)	82666	(60.0%)	8	135
chunk 15	23	(0.0%)	21228	(15.4%)	7	243
chunk 20	65	(0.0%)	70303	(51.0%)	6	144
chunk 3	35	(0.0%)	31393	(22.8%)	5	67
chunk 9	17	(0.0%)	21228	(15.4%)	4	126
chunk 10	23	(0.0%)	35420	(25.7%)	3	62
chunk 12	28	(0.0%)	58859	(42.7%)	3	77
chunk 21	26	(0.0%)	95748	(69.5%)	2	34

chunk 23	18	(0.0%)	49594	(36.0%)	2	28
chunk 24	17	(0.0%)	33530	(24.3%)	2	75
chunk 2	120	(0.1%)	35857	(26.0%)	0	0
chunk 34	1	(0.0%)	91766	(66.6%)	0	1
chunk 35	1	(0.0%)	100854	(73.2%)	0	2
chunk 0	0	(0.0%)	0	(0.0%)	-	0
chunk 1	0	(0.0%)	0	(0.0%)	-	0
chunk 6	31	(0.0%)	0	(0.0%)	-	0
chunk 19	81	(0.1%)	0	(0.0%)	-	0
chunk 26	26	(0.0%)	0	(0.0%)	-	0
chunk 33	0	(0.0%)	0	(0.0%)	-	0

Table 8: Effects of removing each development "chunk" within the Preferred Alternative.

In Table 8 the chunks are ordered by their Mean Isolation effect, from largest (most detrimental) to smallest (least detrimental). The primary point to take from this analysis is that both the amount of habitat removed and its spatial configuration are important determinants of the effect of a chunk, or project, on tortoise gene flow. For example, chunks 31 and 32 are both quite small in terms of area, but have extremely large effects on tortoise gene flow as reflected in their *mean isolation* effect size. This presumably reflects their position near the mouth of the Owens Valley, and their effective isolation of that entire piece of tortoise habitat. In contrast, chunk 25 is relatively large, but has a much smaller effect on *mean isolation* than chunks 31 or 32.

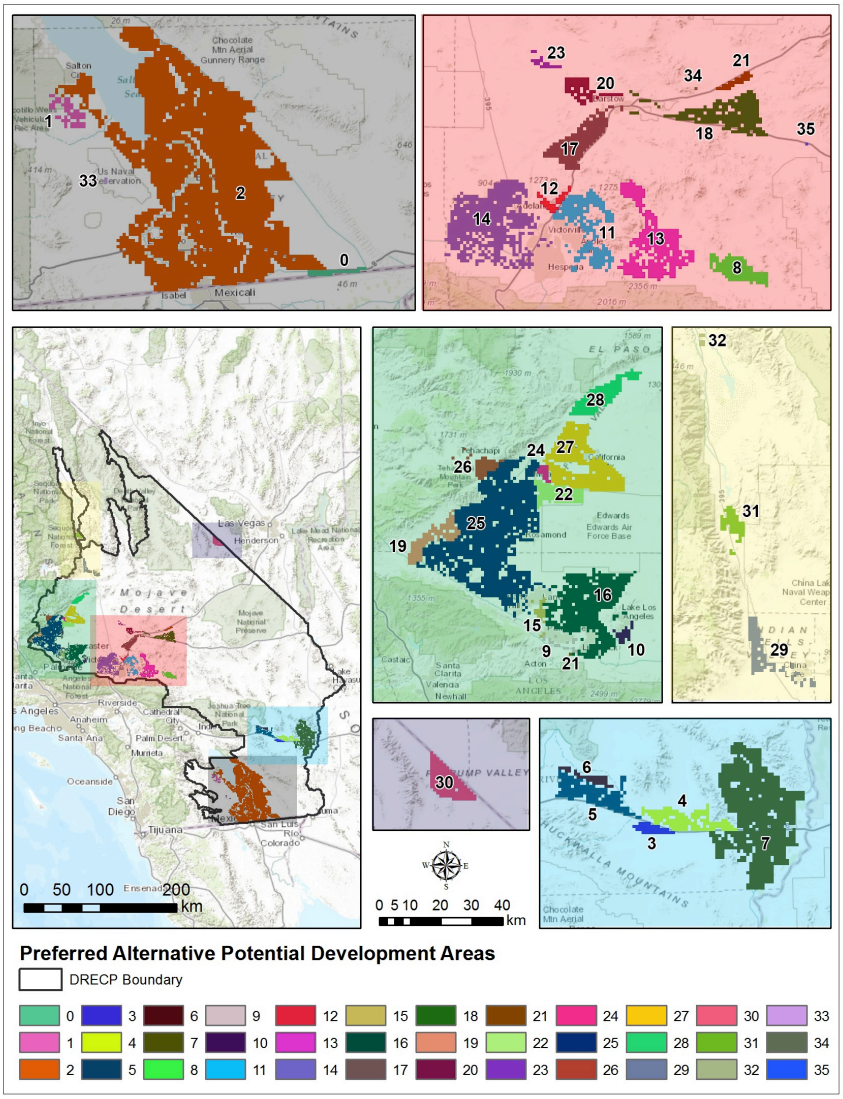

Figure 23. Spatial configuration of the proposed development chunks (see Appendix 4).

Comparison of all chunks across all alternatives

Similarly, we compiled a ranking of proposed development chunks across all five alternatives. **This information may be useful if the final development plan mixes and matches chunks from different alternatives.** However, it is critical to note that the influence of developing certain regions of the landscape is not additive. Interactions between chunks, such as several chunks directly adjacent to one another that form a long barrier to gene flow, may have a significantly higher impact on gene flow than the sum of those individual chunks alone. The best available methodology for assessing a landscape-level development plan is to evaluate the effects of removing all of the proposed development chunks simultaneously. The results showing the individual impacts of the different putative development chunks across all of the alternatives can be found in Appendix 5.

Discussion

The data we have generated for this project, including over 1.2 trillion base pairs of DNA sequence data and 80 high-resolution raster data layers covering much of the Mojave Desert will be an unparalleled resource both for informing our understanding of the natural history and ecology of the desert tortoise and for guiding conservation actions. In addition, we believe the sequencing and genetic inference methods described here will serve as a valuable template for other researchers working in conservation genomics, landscape genomics, and wildlife biology who wish to learn about the ecology of their study organism through the power of genetic data. This may be particularly true for research on long-lived organisms with cryptic life histories, for which traditional methods of assessing population density and dispersal are difficult and costly.

There are several important caveats that should be kept in mind with respect to our approach. First, we model the development chunks, as we calculated them (see Appendix 4), as project areas with impenetrable boundaries that completely repel tortoise movement, but never kill or otherwise reduce tortoise fitness. Other modeling strategies are possible, and they will have different effects on the predictions of gene flow reduction or increase after development. Second, although we include the full geographical range of *G. agassizi* in our analysis, we only model development in the California portion of its range, and only that development identified in the possible DRECP scenarios. In particular, potential development in Nevada will have consequences for gene flow in California, and ideally scenarios for both states should be evaluated simultaneously to develop a comprehensive, range-wide picture of impacts on the species. Finally, the impact of the reduction in gene flow that we model on ecological and demographic processes, and therefore on population viability, is not currently known. Modeling changes in population viability is a critical next step in our research.

Given these caveats, we feel that certain biological conclusions of key importance to Mojave desert tortoise conservation generally, and the DRECP in particular, can be made at this time.

1. By sampling the entire tortoise genome, we can detect subtle differences in population structure that have previously been impossible to detect with more conventional genetic and genomic tools. Our simulation results in Figure 15 show this result quite clearly, and provide a primary motivation for continuing to work at this genomic scale.

2. We detected a strong signal of isolation by distance among tortoises, and that signal is consistent across spatial scales and habitat regions across the range of the tortoise. Even within the relatively homogeneous Ivanpah Valley, we found a strong, statistically significant relationship between genetic and geographic distances. We conclude from this result that even tortoise populations within uninterrupted basins are not "panmictic", allowing the potential for local adaptation regionally in the desert. We also conclude that the occasional long-distance dispersal events that have been observed for the species do not seem to be leading to large areas of admixture and free interbreeding. Rather, there appears to be a general relationship between genetic isolation and geographic distance that scales across both large and small landscapes. The

extent to which locally adapted populations are, or are not, exchangeable is an important avenue for future research, given the frequent use of translocations as a management tool in tortoises.

3. Our PCA analysis identified three primary groupings of tortoises, corresponding to a north/south division, and within the southern group, and east-west division (inset of Figure 17). Certain aspects of these groupings were also suggested in previous analyses (Hagerty and Tracy 2010), although the concordance is not perfect. In particular, the split between the California Cluster and the Las Vegas Cluster of Hagerty and Tracy is almost exactly replicated on PC1, and the NC, WM, and EC splits among the California Cluster are recovered by PC2. Our data indicate that the North-South groups comprise two key tortoise management units, and the East-West division is an additional genetic grouping. For the North-South grouping, our data suggest that the mountains separating the Ivanpah Valley from the surrounding desert habitat constitute a major barrier to gene flow. Within the southern unit, the Cadiz Valley (and its extension to the north and west, the Baker Sink, see Hagerty and Tracy 2010) has similarly been a low-elevation barrier to gene flow. Both of these suggest that if alternative energy installations could be placed *within* the New York/Providence mountains or the Cadiz Valley, they would interfere relatively little with current or past tortoise metapopulation dynamics, but that installations in the corridors of tortoise habitat around these could easily isolate these areas. Our data further indicate that these three sets of tortoises should best be considered three independent management units for tortoise conservation. Other units may emerge with additional tortoise sampling or analysis, but these three are clear.

4. Our landscape genetic inference framework allowed us to estimate the relative effects that the different development alternatives put forth in the DRECP will have on desert tortoise gene flow in the Mojave. **Alternative 1 was found to have the least effect on tortoises, followed in order by the Preferred Alternative, Alternative 4, Alternative 3, and Alternative 2. However, we also note that the effects of all five of the alternative development plans have profound effects on tortoise gene flow; the additional time required for gene flow ranges from 650-950 years across alternative plans, and the relative isolation from 9% to 26%. These numbers far outstrip the actual amount of lost habitat for each plan (1.7%-4.9% of the total habitat for the tortoise), and emphasize the cascading effects that development can have on landscape connectivity.**

5. Within the Preferred Alternative, many of the individual proposed development chunks have relatively little impact on desert tortoise connectivity when considered in isolation, although some have extremely high impacts. Chunks 31, 32, 25, and 7 have (in order) the greatest impact on tortoise connectivity, and should be examined closely before they are implemented. One of these, chunk 31 is located at the entrance of the Owens Valley, and this chunk is having a disproportionate effect on gene flow across the whole range of the tortoise, because it is

effectively isolating the Owens Valley from the rest of the Mojave Desert. The general lesson from these analyses is that development that isolates regions should be avoided.

6. Across all alternatives, we identified and evaluated 214 development chunks, in terms of their individual effects on tortoise connectivity, and we encourage using this list along with other variables as a first pass for considering the order of approval of projects and habitat patches. However, we again emphasize that this list is for each chunk by itself, and it ignores the synergistic effects of developing multiple chunks.

7. Future analyses can, and should, investigate the isolating effects that multiple habitat chunks have when considered together. By sequentially adding development chunks and subdividing chunks, our work can help develop both a better final build-out and a gradual path to that build-out that minimized impacts on tortoise connectivity for as long as possible across the Mojave. Further direction from CDFW would be valuable in delineating which particular combinations of proposed development chunks would be useful to evaluate.

8. The tools we have developed can be used to predict the local effects of gene flow of specific development plans, and to recommend specific mitigation procedures, including critical issues like the location of habitat corridors. Not only selection of development areas (in the DRECP), but also situation of development within these areas in subsequent planning processes, will be key to reducing the impact of renewable energy development on the long-term viability of desert tortoise populations.

Acknowledgements

We thank Roy Averill-Murray of the US Fish and Wildlife Service, Danna Hinderle, and the staff and scientists from the DTCC, and Jim Oosterhuis from the San Diego Zoo for desert tortoise handling training and strategizing on this project, Adalgisa Caccone, Kevin White, and Zifeng Jiang for sharing an early draft assembly of the Galapagos tortoise genome, Tasha La Doux, Jim Andre, and the Sweeney Granites Mountains Desert Research Center for hospitality and field support, Ying Zhi Lim, and Sarah Wenner for sorting tortoise samples, and Mariel Villanueva for some tortoise background research. This work was made possible through financial support from the California Department of Fish and Wildlife, Agreement Number P1382008. This work used the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley, supported by NIH S10 Instrumentation Grants S10RR029668 and S10RR027303. PR is supported by NSF grant #DBI-1262645 and startup funds from USC.

Appendix 1

GIS Raster Data

We compiled and synthesized a total of 83 environmental raster data layers for this study area. These rasters fell in five principal categories: anthropogenic, biotic, climatic, topographic, and soils. We briefly describe these layers below. The resolution of the layers varied from 10m to 800m, although they were all standardized to 30m. For a list of all 83 layers, including a brief description of each, see Figure S8.

Anthropogenic layers:

Our anthropogenic GIS raster dataset included 13 layers, sourced or derived from the 2012 TIGER Census road classification (http://www.census.gov/cgi-bin/geo/shapefiles2012/main). The data describe the spatial distribution of roads in the Mojave ranging from 4WD trails and bike paths to primary roads and their on- and off-ramps. These layers were also used to calculate the Euclidean distance across the landscape to the nearest road. Also included was one layer from the National Land Cover Database 2011 (http://www.mrlc.gov/nlcd2011.php) that depicts the percent of impervious cover per cell. All anthropogenic layers have a resolution of 30m.

Biotic layers:

Our biotic GIS raster dataset included four layers describing the distribution of plant material across the Mojave. Three of these layers (shrub/scrub, grass/herb, and tree cover) were sourced from the National Land Cover Database 2011 (http://www.mrlc.gov/nlcd2011.php) and have a resolution of 30m. The fourth layer, annual growth potential, was calculated from the Moderate Resolution Imaging Spectroradiometer Enhanced Vegetation Index (MODIS-EVI) following the methods of Wallace and Thomas (2008) and Nussear *et al.* (2009). This 250m resolution layer serves as a proxy for annual plant biomass and was aggregated down to 30m resolution.

Climatic layers:

The climatic GIS raster dataset consisted of 52 layers describing the spatial distribution of climatic variables, including minimum, maximum, and mean temperature as well as mean precipitation, for each month and a yearly average, across the Mojave. These layers were taken from the PRISM Climate Group (http://www.prism.oregonstate.edu/normals/) and calculated over the last 30 years. The resolution of these layers was 800m, and were aggregated down to 30m.

Topographic layers:

The topographic GIS raster dataset consisted of 11 layers derived from the National Land Cover Database (NLCD ;http://www.mrlc.gov/nlcd2011.php) and National Elevation Database (NED) on the USGS National Map Viewer (http://viewer.nationalmap.gov/viewer/). The elevation, aspect, surface roughness, surface area, slope, and eastness and northness (the degree to which slope faces east and north, respectively) layers were all derived from the NED DEM. The land cover and barrenness layers were both extracted from the NLCD. All landscape raster layers were produced at a 30m resolution. Longitude and latitude rasters were also constructed over the study area.

Soil layers:

The soils GIS raster dataset consisted of 3 layers that describe bulk density, percent of rocks greater than 10 inches, and depth to bedrock. The data were extracted from SSURGO2 database. Data gaps in SSURGO2 were filled using STATSGO, downloaded from USDA NRCS Soil Data Mart (http://websoilsurvey.sc.egov.usda.gov/App/WebSoilSurvey.aspx). All layers were transformed to 30m rasters.

Of this complete set of rasters, three different subsets (of 6, 12, and 24 rasters) were selected to be used in inference of the correlation between ecological heterogeneity and partitioning of genetic variation in Mojave tortoises. We selected these subsets to reduce the complexity and computation time of analyses, and to produce a set of statistically more independent layers. Many of the initial 83 raster layers in the full set were highly correlated (Figure S9), and including such highly correlated layers in later analyses can both confuse the analysis and make any interpretation of their individual effects difficult or impossible. We selected rasters for inclusion to maximize overlap with layers used in previous GIS analyses of tortoises, and to minimize pairwise correlation among layers. The list of rasters included in each subset can be found in Figure S10.

Appendix 2

Divergence

Suppose that we have sequences of length L from two individuals, with $C_{i,k}$ reads that map for individual k to a position overlapping site i, for $k \in \{1,2\}$ and $1 \leq i \leq L$. Suppose that $C_{i,k}$ is marginally Poisson with mean $\lambda_k c_i$, and that each read covering site i independently draws an allele, so that given $C_{i,k}$, the allele counts $N_{i,k}(a)$ are Multinomial with probabilities $p_{i,k}(a)$, where a is the allele. Suppose also that coverages and counts are independent between samples, given p and c. The probability that a pair of reads drawn from those at site i, one drawn uniformly at random from each sample, both have allele a is $Y_i(a) = N_{i,1}(a)N_{i,2}(a)/C_{i,1}C_{i,2}$, and so that if $C_{i,1}C_{i,2} > 0$,

$$\mathbb{E}[Y_i(a) \,|\, C_{i,1}, C_{i,2}] = p_{i,1}(a)p_{i,2}(a). \tag{1}$$

We would like to estimate mean sequence divergence,

$$\pi = \frac{1}{L}\sum_{i=1}^{L}\left(1 - \sum_a p_{i,1}(a)p_{i,2}(a)\right). \tag{2}$$

Given a weighting function w with $w(0,n) = w(n,0) = 0$ for each n, an estimator of divergence is

$$D(w) = \frac{\sum_{i=1}^{L} w(C_{i,1}, C_{i,2})(1 - \sum_a Y_i(a))}{\sum_{i=1}^{L} w(C_{i,1}, C_{i,2})}. \tag{3}$$

The expection of $D(w)$ is

$$\mathbb{E}[D(w)] = \mathbb{E}\left[\frac{\sum_{i=1}^{L} w(C_{i,1}, C_{i,2})(1 - \sum_a p_{i,1}(a)p_{i,2}(a))}{\sum_{i=1}^{L} w(C_{i,1}, C_{i,2})}\right]. \qquad = \pi. \tag{4}$$

If the mean sitewise coverages c_i are independent of the $p_{i,k}$, then by exchangeability, $\mathbb{E}[D(w)] = \pi$. We take $w(x,y) = xy$:, which approximately (but not quite) does not depend on the coverages λ:

$$\mathbb{E}[D(w)] \approx \frac{\sum_{i=1}^{L} c_{i,1}c_{i,2}(1 - \sum_a p_{i,1}(a)p_{i,2}(a))}{\sum_{i=1}^{L} c_{i,1}c_{i,2}}. \tag{5}$$

Appendix 3

Model Specification

A model of landscape resistance, as discussed above, is essentially a specification of a reversible random walk on the landscape. A reversible random walk is specified by two quantities: the *stationary distribution* of each point x, denoted $\pi(x)$, and the *relative jump rates* between each pair of adjacent locations x and y, denoted $j(x,y)$; these combine to give the total rate of movement from x to y as $G(x,y) = j\,(x,y)/\pi\,(x)$. The requirement that the random walk to be reversible, i.e. $\pi(x)G(x,y) = \pi(y)G(x,y)$, means that relative jump rates must be symmetric, i.e. $j(x,y) = j(y,x)$.

We then allow these two ingredients to be determined by linear functions of the landscape layers: if we have n landscape layers whose values at location x are $L_1(x)$, ..., $L_n(x)$, then we suppose that

$$G(x,y) = \beta \times \frac{1}{1 + exp\big(-\gamma_1 L_1(x) - \cdots - \gamma_n L_n(x)\big)} \times \frac{1}{1 + exp\Big(-\delta\big(L_1(x) + L_1(y)\big) - \cdots - \delta\big(L_n(x) + L_n(y)\big)\Big)}.$$

The parameters are: β, an overall scaling factor, and for each $1 \le k \le n$, γ_k, that determines how the kth layer affects the stationary distribution, and δ_k, that determines how the kth layer affects the relative jump rates.

In practice, then, a model is determined by:

1 A *mask*, i.e. a specification of the total potential habitat area available for movement; movement rates to locations outside of this are assumed to be zero.

2 The *layers*, which provide a numerical value for each location on the landscape; we include a "constant" layer (that takes the value 1 everywhere), and normalize remaining layers to have mean zero and variance 1.

3 The *parameters* β, γ_1, ..., γ_n, and δ_1, ..., δ_n.

4 A *neighborhood size* R and a *local coalescence time* T.

These are combined to fit the data by computing for each x and y the mean time until a random walk begun at x first gets closer than R to the location y, which we denote by $h_R(x,y)$, and postulating that the observed sequence divergence between tortoises at locations x and y, denoted $d(x,y)$, is equal to T plus the mean R-commute time, i.e.

$$d(x,y) = T + \big(h_R(x,y) + h_R(y,x)\big)/2 + \varepsilon,$$

where ε is the noise due to demographic stochasticity and sequencing error.

Fitting Procedure

To fit the model above, we find parameters to minimize the weighted mean squared error

$$L = \sum_{x,y} (d(x,y) - T - (h_R(x,y) + h_R(y,x)))^2.$$

This requires computing the times $h_R(x,y)$, which can be done as follows. First, we compute the movement rates of the random walk and place them in a matrix G, with rows and columns indexed by locations, and whose (x,y)th entry is $G_{x,y}$ defined above. Fix a location y and a distance R, let $N_R(y)$ be the set of locations within distance R of location y. Then the times $h_R(x,y)$ solve the equations

$$\sum_z G_{x,z} h_R(z,y) = -1 \qquad \text{for } x \notin N_R(y),$$

and boundary conditions

$$h_R(x,y) = 0 \qquad \text{if } x \in N_R(y).$$

This forms one system of equations for each y, that we solve numerically using sparse matrix solvers in the *Matrix* package in *R* (Bates and Maechler 2014).

Analytically, the solution can be written as follows: for a given y and R let $\tilde{G}^{(y,R)}$ denote the matrix obtained by removing the rows and columns of G corresponding to $N_R(y)$. Then, seen as a vector indexed by x,

$$h_R(x,y) = (\tilde{G}^{y,R})^{-1}(-1),$$

where $(\tilde{G}^{y,R})^{-1}$ is the matrix inverce of $\tilde{G}^{y,R}$, and (-1) denotes the vector whose entries are all -1. This can be substituted into the expression for the mean squared error above, and then differentiated, to find analytic expressions for the gradient vector and Hessian matrix of L with respect to T, β, each γ, and each δ. With these in hand, we then use a "trust region" optimization routine, as coded in the package *trust* in *R* [Geyer]. This allows us to find best-fitting choices of all parameters except R; in practice, we then fix R at 15km. It would be preferrable to also optimize over R; however, R is nearly confounded with T, in that increasing R is very nearly equivalent to adding a constant to $h_R(x,y)$, and so this choice does not significantly affect results.

Landscape Resistance Models

Concretely: for the ith sampled tortoise, let n_i be the number of other sampled tortoises within 25km, and let the (i,j)th *weight* be

$$w_{i,j} = \frac{1}{n_i n_j} if i \wedge jare \in thesameregion,$$

and $w_{i,j} = 0$ otherwise. Let the (i,j)th *residual* be

$$r_{i,j} = d(x,y) - T - \big(h_R(x,y) + h_R(y,x)\big).$$

Then the *weighted median residual* is the value $\overline{r}$ such that the sum of the weights of the residuals smaller then $\overline{r}$ is equal to the sum of the weights larger than $\overline{r}$: concretely, it satisfies

$$\sum_{(i,j):r_{i,j}<\bar{r}} w_{i,j} = \sum_{(i,j):r_{i,j}>\bar{r}} w_{i,j};$$

if there is ambiguity in where $\overline{r}$ should fall, then it is specified as the weighted mean of the nearest possible samples.

Appendix 4

Aggregation of development focus area polygons into “chunks”

The potential development areas evaluated in this study were derived from the alternative shapefiles on the DRECP gateway. Polygons designated as "Development Focus Areas", which ranged between 1000-2500 polygons for the different alternatives. The polygons were grouped into “chunks” according to area and proximity amongst each other, as described below.

First, an initial set of development chunks was established by selecting all polygons with an area greater than or equal to 1000 hectares. Next, all remaining polygons within 5 km (edge to edge proximity) of an initial polygon were identified as secondary polygons and assigned to the closest initial polygon. If the secondary polygons were adjacent to multiple initial polygons, the secondary polygon took the assignment of the largest initial polygon.

The remaining polygons (polygons under 1000 hectares that are not within 5 km of an initial polygon) were grouped into “remainder clusters,” based on proximity, with a 5 km upper limit. If a remainder cluster was smaller than 5 hectares and within 10 km of other such polygon clusters, that remainder cluster was reassigned to reflect a single potential development area with the closest remainder cluster. In the preferred alternative, the polygons were grouped into 36 potential development areas, with areas ranging between 4 and 300,000 hectares. All results were rasterized to a 1 km resolution (consistent with the habitat model from Nussear *et al.* (2009)).

Appendix 5

Development chunks across all alternatives, ranked by isolation (years)

Table S3 below shows the effects of removing each chunk *across all alternatives*. Keys that show the spatial configuration all of the chunks can be found in Figures 23 and S11-S14). The *mean isolation* is a good measure of the total effect of removing a given area, but note that the absolute size of the number is not necessarily reflective of the overall effect, as it measures this piece, in isolation, without the effects of all other pieces. There are many instances where the impacts of multiple chunks considered together have a larger impact than the sum of the chunks by themselves.

- *habitat removed* is the total amount of area either in possible development areas or completely isolated from the rest of tortoise habitat under this alternative,
- *isolated* is the total area on which the gene flow to nearby areas has increased,
- *isolation* is the mean amount by which the commute time has increased to nearby areas across this area where it has increased,
- *max isolation* is the maximum amount by which the commute time has increased between any two nearby reference locations.

Table S3: Effects of removing individual chunks from all evaluated alternatives.

Chunk name	Habitat area removed (km2)	Habitat removed (%)	Isolated (km2)	Isolated (%)	**Isolation (years)**	Max isolation (years)
pref-31	55	0.040%	3129	2.270%	10454.96287	11002.95027
alt1-30	55	0.040%	3129	2.270%	10454.96287	11002.95027
alt2-39	55	0.040%	3129	2.270%	10454.96287	11002.95027
alt3-36	55	0.040%	3129	2.270%	10454.96287	11002.95027
alt4-40	55	0.040%	3129	2.270%	10454.96287	11002.95027
alt1-31	32	0.023%	3129	2.270%	2977.230231	5720.945775
alt3-37	32	0.023%	3129	2.270%	2977.230231	5720.945775
alt4-41	32	0.023%	3129	2.270%	2977.230231	5720.945775
alt2-41	26	0.019%	3129	2.270%	2076.042664	4267.545256
alt2-42	4	0.003%	3129	2.270%	930.5677689	1028.976902
pref-32	6	0.004%	3129	2.270%	549.4983126	847.0290059
alt2-4	441	0.320%	5962	4.326%	154.2939496	1232.719606
alt4-34	982	0.713%	4406	3.197%	126.544891	890.60505
pref-25	987	0.716%	4406	3.197%	126.4884531	890.5329366

alt2-29	1014	0.736%	4406	3.197%	125.5936494	889.4511347
alt4-15	464	0.337%	41928	30.422%	111.1056815	1866.096075
pref-7	481	0.349%	41928	30.422%	110.8387193	1865.595161
alt2-11	481	0.349%	41928	30.422%	110.8387193	1865.595161
alt1-26	229	0.166%	16041	11.639%	105.030575	1981.301324
alt3-32	232	0.168%	16041	11.639%	104.9816929	1981.223443
alt2-33	685	0.497%	32672	23.706%	102.4671119	1558.291619
alt3-26	119	0.086%	20171	14.636%	96.79866947	2509.389538
alt2-17	561	0.407%	78360	56.856%	92.70997656	2420.175936
alt2-18	658	0.477%	38341	27.819%	92.61728592	2003.122895
alt2-37	349	0.253%	39807	28.883%	83.20759122	1517.147758
alt2-13	125	0.091%	83308	60.447%	71.09490207	1043.239349
alt2-12	25	0.018%	80470	58.387%	67.64288972	970.6218473
alt4-20	279	0.202%	87719	63.647%	67.37425841	1966.676353
pref-13	299	0.217%	87719	63.647%	65.09388048	1627.105793
alt4-23	440	0.319%	34920	25.337%	60.32250515	1758.418059
pref-16	441	0.320%	34920	25.337%	60.30016361	1758.381114
alt1-18	509	0.369%	38947	28.259%	57.23993726	1069.079078
alt4-21	509	0.369%	38947	28.259%	57.23993726	1069.079078
pref-14	512	0.371%	38947	28.259%	57.19660826	1069.024854
alt3-19	512	0.371%	38947	28.259%	57.19660826	1069.024854
alt3-33	339	0.246%	34510	25.040%	56.93591435	1002.734925
alt1-20	376	0.273%	30698	22.274%	54.74882569	1668.181877
pref-27	336	0.244%	34510	25.040%	51.62989072	867.7679889
pref-5	148	0.107%	64315	46.666%	51.61584802	994.4857026
alt2-9	148	0.107%	64315	46.666%	51.61584802	994.4857026
alt2-21	372	0.270%	30698	22.274%	51.17128564	1562.243574
alt3-21	372	0.270%	30698	22.274%	51.17128564	1562.243574
alt4-13	146	0.106%	64917	47.102%	51.0937871	994.4998639
alt3-14	276	0.200%	81060	58.815%	43.42090795	1254.543749
pref-11	204	0.148%	71582	51.938%	40.38579457	1066.374699
alt4-18	199	0.144%	71582	51.938%	39.86367356	1055.615116
alt2-34	144	0.104%	129339	93.846%	39.54953954	806.3461141
alt3-15	175	0.127%	67035	48.639%	34.99917168	912.9427254
alt1-15	173	0.126%	67035	48.639%	34.73083816	906.799037
alt2-32	121	0.088%	43126	31.291%	29.82379848	373.7315794
alt1-13	212	0.154%	79400	57.611%	29.07229668	789.171887
alt4-12	129	0.094%	54457	39.513%	29.07204937	726.005056
pref-4	128	0.093%	54457	39.513%	29.07196254	726.005058
alt2-8	128	0.093%	54457	39.513%	29.07196254	726.005058
alt3-10	159	0.115%	40442	29.344%	26.64999279	472.1949398

alt3-22	129	0.094%	29122	21.130%	26.19520402	1038.755225
pref-29	95	0.069%	39807	28.883%	24.63882816	555.0311281
alt2-22	232	0.168%	100619	73.007%	24.16969985	423.3725473
alt3-34	129	0.094%	40702	29.533%	21.49991786	412.1222172
pref-28	128	0.093%	40702	29.533%	20.95496256	399.7414774
alt4-37	128	0.093%	40702	29.533%	20.95496256	399.7414774
pref-17	173	0.126%	60235	43.705%	20.31691442	277.0804954
pref-18	219	0.159%	95912	69.592%	19.78074043	371.7395754
alt3-25	216	0.157%	95912	69.592%	19.48284167	366.1448871
alt1-22	195	0.141%	97881	71.020%	18.42110525	329.3456871
alt4-26	192	0.139%	97881	71.020%	18.3888473	327.7363475
alt4-36	144	0.104%	36063	26.167%	17.34544972	391.5485222
alt2-31	83	0.060%	42545	30.870%	17.25972043	183.7982339
alt2-20	115	0.083%	104851	76.078%	15.31718378	255.6196251
pref-30	121	0.088%	41658	30.226%	15.12708329	298.3699549
alt2-38	121	0.088%	41658	30.226%	15.12708329	298.3699549
alt2-14	160	0.116%	83214	60.378%	13.92264439	235.0435836
alt2-36	51	0.037%	43732	31.731%	12.43975845	161.6208687
alt1-29	70	0.051%	42267	30.668%	12.18142083	255.2150158
alt3-35	43	0.031%	44465	32.263%	11.75473679	150.7783055
alt1-21	25	0.018%	36644	26.588%	11.65870977	557.613118
alt1-19	30	0.022%	22221	16.123%	10.68822929	311.8983999
alt4-33	66	0.048%	36063	26.167%	10.62698592	297.4178308
pref-22	98	0.071%	34762	25.223%	10.31592401	269.6499622
alt4-30	92	0.067%	34762	25.223%	10.06381054	264.2721669
alt1-11	74	0.054%	55899	40.559%	9.221810567	246.7060242
alt1-27	57	0.041%	36063	26.167%	8.527680577	266.3843285
pref-8	108	0.078%	82666	59.981%	8.144658656	135.688524
alt4-39	38	0.028%	49685	36.050%	7.942928403	87.2320839
alt2-26	65	0.047%	33010	23.951%	7.627565649	216.1114425
pref-15	23	0.017%	21228	15.403%	7.547277138	243.3179688
alt2-19	23	0.017%	21228	15.403%	7.547277138	243.3179688
alt3-20	23	0.017%	21228	15.403%	7.547277138	243.3179688
alt4-22	23	0.017%	21228	15.403%	7.547277138	243.3179688
alt3-27	74	0.054%	70303	51.010%	7.382383857	156.8901014
alt1-12	62	0.045%	34974	25.376%	7.364909383	74.60262281
alt3-11	63	0.046%	34974	25.376%	7.364909383	74.60262281
alt4-38	18	0.013%	44995	32.647%	7.306471287	107.577588
alt2-24	69	0.050%	69312	50.291%	6.88446757	145.8093543
pref-20	65	0.047%	70303	51.010%	6.569753728	144.0881695
alt4-28	63	0.046%	75593	54.849%	6.298322003	152.3734595

alt4-24	64	0.046%	54970	39.885%	5.711972865	95.90468116
pref-3	35	0.025%	31393	22.778%	5.65602406	67.75635896
alt2-7	35	0.025%	31393	22.778%	5.65602406	67.75635896
alt3-8	35	0.025%	31393	22.778%	5.65602406	67.75635896
alt3-29	31	0.022%	25251	18.322%	5.560229973	151.1221915
alt1-24	30	0.022%	25975	18.847%	5.434577802	151.1369023
alt1-28	52	0.038%	36063	26.167%	5.133053172	158.09309
alt4-11	18	0.013%	34094	24.738%	4.396711231	53.22000634
alt1-14	18	0.013%	25471	18.481%	4.393072327	178.8457954
alt2-15	18	0.013%	25471	18.481%	4.393072327	178.8457954
alt3-13	18	0.013%	25471	18.481%	4.393072327	178.8457954
alt4-16	18	0.013%	25471	18.481%	4.393072327	178.8457954
alt4-42	17	0.012%	41108	29.827%	4.303040585	52.73392572
pref-9	17	0.012%	21228	15.403%	4.072480321	126.1079045
alt4-19	30	0.022%	49241	35.728%	3.944288786	68.19111865
pref-10	23	0.017%	35420	25.700%	3.78033413	62.64543408
alt1-16	23	0.017%	35420	25.700%	3.78033413	62.64543408
alt3-16	23	0.017%	35420	25.700%	3.78033413	62.64543408
alt4-17	23	0.017%	35420	25.700%	3.78033413	62.64543408
alt2-16	28	0.020%	44698	32.432%	3.717882741	65.90024311
alt3-17	28	0.020%	44698	32.432%	3.717882741	65.90024311
pref-12	28	0.020%	58859	42.707%	3.682896042	77.54002323
alt1-32	14	0.010%	40450	29.350%	3.508712862	45.50924548
alt3-38	14	0.010%	40450	29.350%	3.508712862	45.50924548
alt2-35	16	0.012%	40450	29.350%	3.494510069	49.41277092
alt4-25	31	0.022%	69312	50.291%	3.281974681	43.96302757
alt2-27	26	0.019%	49594	35.984%	3.023469654	35.31076735
alt3-30	26	0.019%	49594	35.984%	3.023469654	35.31076735
alt1-23	26	0.019%	96483	70.006%	3.005983975	35.59140754
pref-21	26	0.019%	95748	69.473%	2.998579551	34.5690408
alt1-17	22	0.016%	44698	32.432%	2.955399174	55.77312773
alt3-12	38	0.028%	82666	59.981%	2.91868251	49.60180006
alt2-25	25	0.018%	96483	70.006%	2.912109851	34.57389117
alt3-28	25	0.018%	96483	70.006%	2.912109851	34.57389117
alt4-29	25	0.018%	96483	70.006%	2.912109851	34.57389117
alt3-24	42	0.030%	16041	11.639%	2.851573119	59.60814145
alt4-31	22	0.016%	48972	35.533%	2.808693614	32.47754302
pref-23	18	0.013%	49594	35.984%	2.458679846	28.1307359
pref-24	17	0.012%	33530	24.329%	2.096735046	75.98429227
alt1-25	17	0.012%	33530	24.329%	2.096735046	75.98429227
alt2-28	17	0.012%	33530	24.329%	2.096735046	75.98429227

alt3-31	17	0.012%	33530	24.329%	2.096735046	75.98429227
alt4-32	17	0.012%	33530	24.329%	2.096735046	75.98429227
alt3-9	21	0.015%	43789	31.772%	1.916902126	45.0516296
alt3-18	11	0.008%	76444	55.466%	0.968735427	18.69600268
alt1-35	2	0.001%	44465	32.263%	0.631432596	19.1767269
alt2-45	2	0.001%	44465	32.263%	0.631432596	19.1767269
alt3-41	2	0.001%	44465	32.263%	0.631432596	19.1767269
alt4-44	2	0.001%	44465	32.263%	0.631432596	19.1767269
alt2-46	1	0.001%	6660	4.832%	0.390993037	5.624057879
alt1-36	1	0.001%	91981	66.739%	0.131381667	2.42779131
alt2-48	1	0.001%	100854	73.178%	0.103475423	3.856345285
pref-35	1	0.001%	100854	73.178%	0.103289812	2.785222272
alt1-37	1	0.001%	100854	73.178%	0.103289812	2.785222272
alt3-43	1	0.001%	100854	73.178%	0.103289812	2.785222272
alt4-45	1	0.001%	100854	73.178%	0.103289812	2.785222272
pref-34	1	0.001%	91766	66.583%	0.090169896	1.929091762
alt2-47	1	0.001%	91766	66.583%	0.090169896	1.929091762
alt3-42	1	0.001%	91766	66.583%	0.090169896	1.929091762
alt4-46	1	0.001%	91766	66.583%	0.090169896	1.929091762
pref-2	120	0.087%	35857	26.017%	0.017469848	0.383377854
alt2-6	120	0.087%	35857	26.017%	0.017469848	0.383377854
alt3-7	66	0.048%	42948	31.162%	0.014768459	0.383490034
alt4-10	41	0.030%	47572	34.517%	0.013312576	0.383523046
alt2-43	1	0.001%	137166	99.525%	0.00058591	0.007541373
alt1-10	26	0.019%	1402	1.017%	0.000268426	0.004089324
pref-0	0	0.000%	0	0.000%	-	0
pref-1	0	0.000%	0	0.000%	-	0
pref-6	31	0.022%	0	0.000%	-	0
pref-19	81	0.059%	0	0.000%	-	0.076930398
pref-26	26	0.019%	0	0.000%	-	0
pref-33	0	0.000%	0	0.000%	-	0
alt1-0	0	0.000%	0	0.000%	-	0
alt1-1	0	0.000%	0	0.000%	-	0
alt1-2	0	0.000%	0	0.000%	-	0
alt1-3	0	0.000%	0	0.000%	-	0
alt1-4	0	0.000%	0	0.000%	-	0
alt1-5	0	0.000%	0	0.000%	-	0
alt1-6	0	0.000%	0	0.000%	-	0
alt1-7	0	0.000%	0	0.000%	-	0
alt1-8	0	0.000%	0	0.000%	-	0
alt1-9	0	0.000%	0	0.000%	-	0

alt1-33	4	0.003%	0	0.000%	-	0
alt1-34	0	0.000%	0	0.000%	-	0
alt2-0	0	0.000%	0	0.000%	-	0
alt2-1	0	0.000%	0	0.000%	-	0
alt2-2	18	0.013%	0	0.000%	-	0
alt2-3	31	0.022%	0	0.000%	-	1.037902843
alt2-5	0	0.000%	0	0.000%	-	0
alt2-10	31	0.022%	0	0.000%	-	0
alt2-23	82	0.059%	0	0.000%	-	0.003731843
alt2-30	27	0.020%	0	0.000%	-	0
alt2-40	26	0.019%	0	0.000%	-	0
alt2-44	0	0.000%	0	0.000%	-	0
alt3-0	0	0.000%	0	0.000%	-	0
alt3-1	0	0.000%	0	0.000%	-	0
alt3-2	0	0.000%	0	0.000%	-	0
alt3-3	0	0.000%	0	0.000%	-	0
alt3-4	0	0.000%	0	0.000%	-	0
alt3-5	0	0.000%	0	0.000%	-	0
alt3-6	0	0.000%	0	0.000%	-	0
alt3-23	47	0.034%	0	0.000%	-	0
alt3-39	4	0.003%	0	0.000%	-	0
alt3-40	0	0.000%	0	0.000%	-	0
alt4-0	0	0.000%	0	0.000%	-	0
alt4-1	0	0.000%	0	0.000%	-	0
alt4-2	0	0.000%	0	0.000%	-	0
alt4-3	0	0.000%	0	0.000%	-	0
alt4-4	0	0.000%	0	0.000%	-	0
alt4-5	0	0.000%	0	0.000%	-	0
alt4-6	0	0.000%	0	0.000%	-	0
alt4-7	0	0.000%	0	0.000%	-	0
alt4-8	0	0.000%	0	0.000%	-	0
alt4-9	0	0.000%	0	0.000%	-	0
alt4-14	31	0.022%	0	0.000%	-	0
alt4-27	80	0.058%	0	0.000%	-	0.09594425
alt4-35	26	0.019%	0	0.000%	-	0
alt4-43	0	0.000%	0	0.000%	-	0

Supplemental Figures

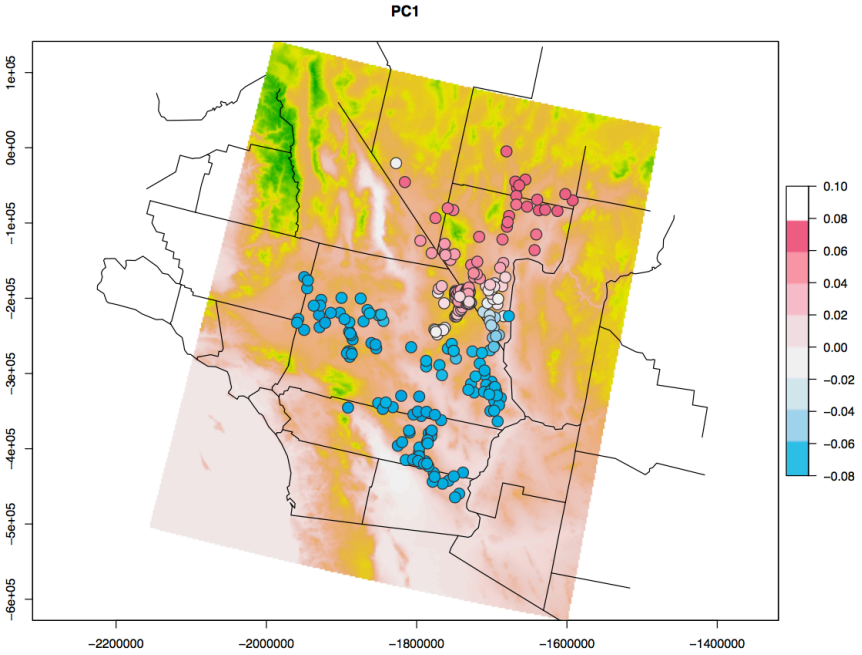

Figure S2. Visualization of genetic structure. Dots represent individual tortoises, and they are colored by their scores on PC1. Background color corresponds to elevation.

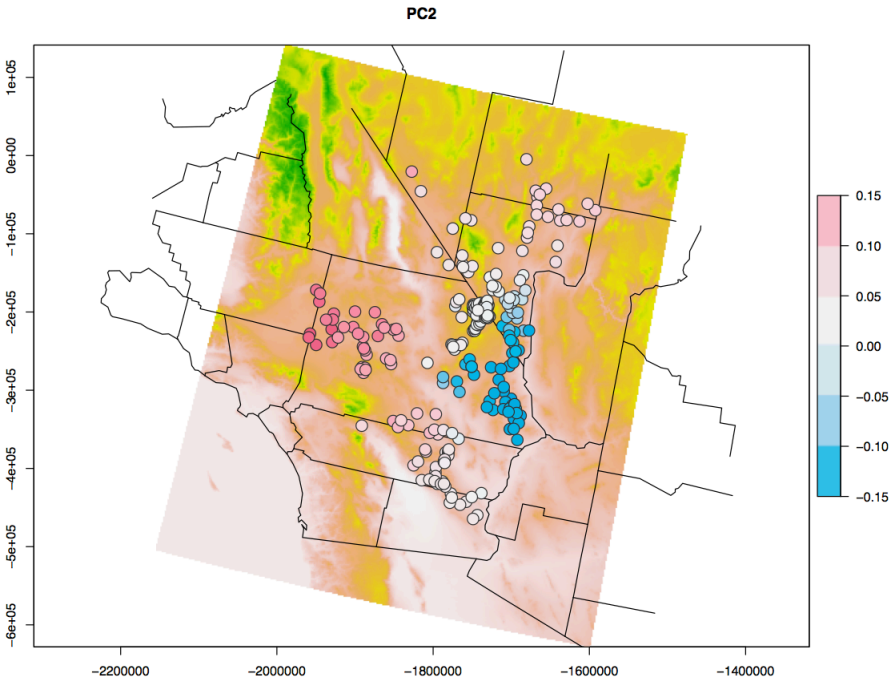

Figure S3. Visualization of genetic structure. Dots represent individual tortoises, and they are colored by their scores on PC2. Background color corresponds to elevation.

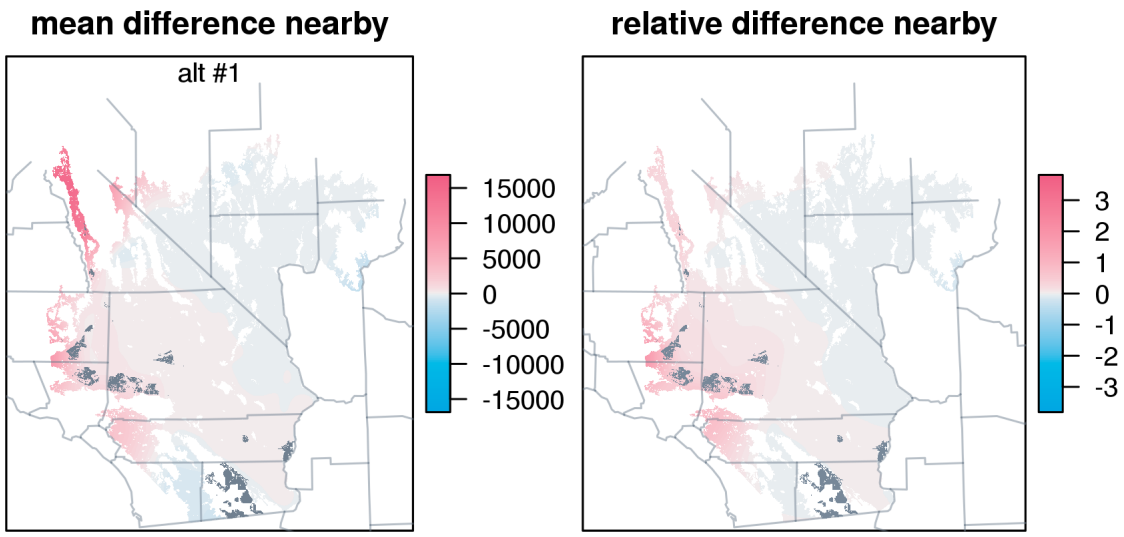

Figure S4. Effects of removing the development areas in Alternative 1.

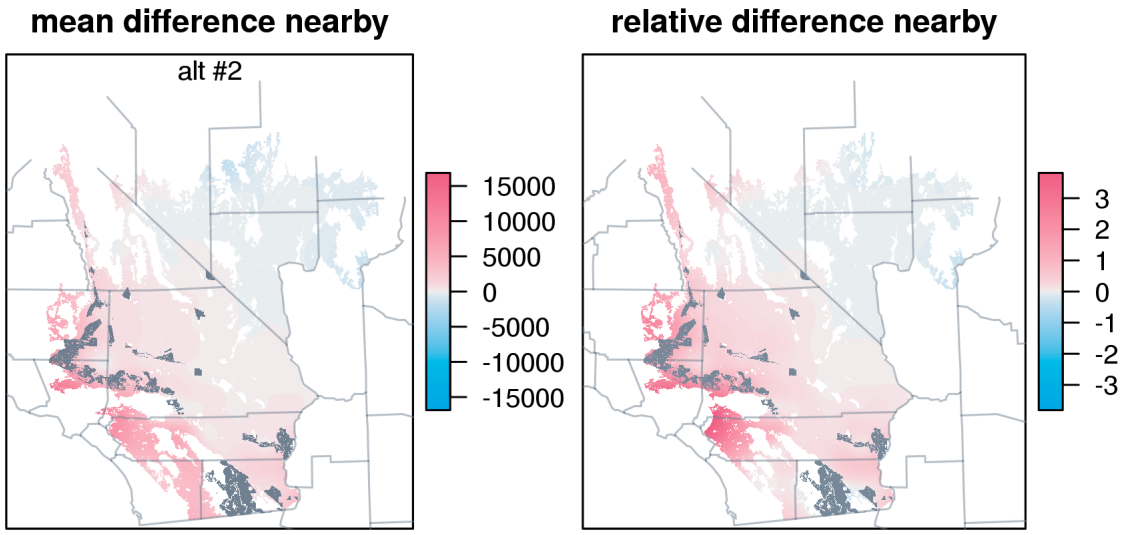

Figure S5. Effects of removing the development areas in Alternative 2.

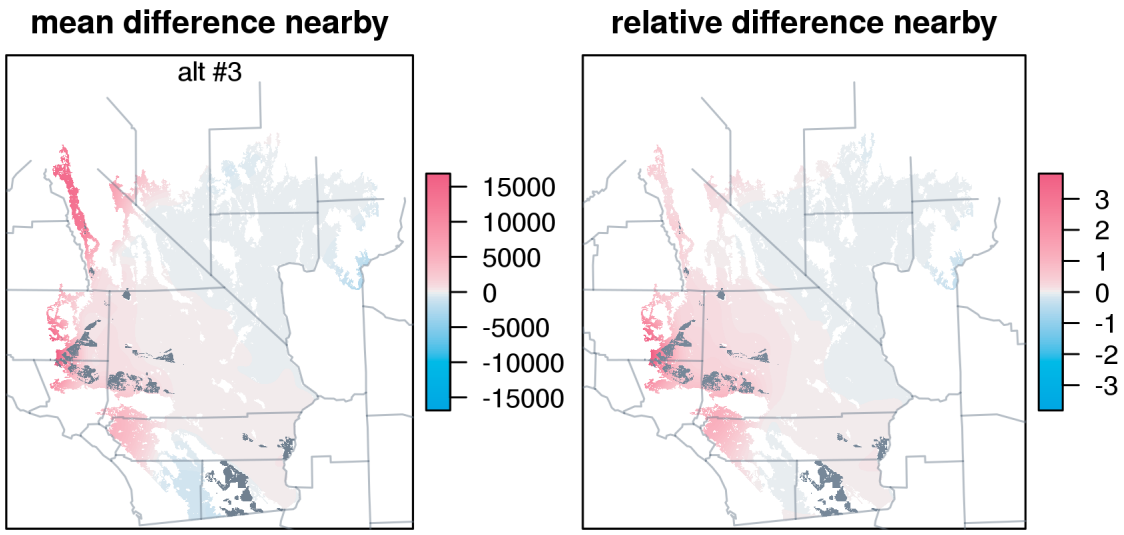

Figure S6. Effects of removing the development areas in Alternative 3.

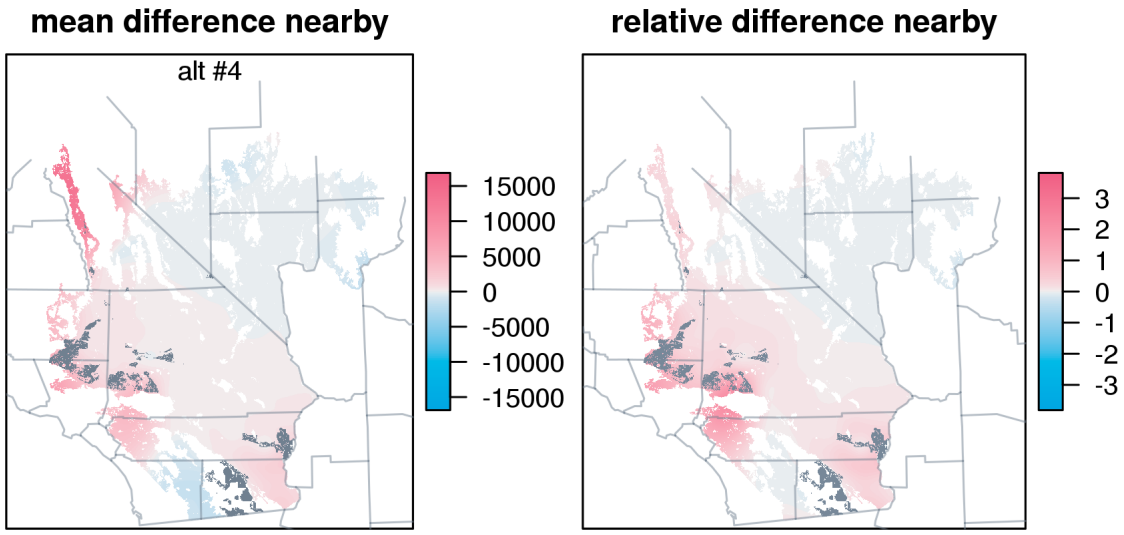

Figure S7. Effects of removing the development areas in Alternative 4.

Layer Name	Layer Description	Used in previous analyses
avg_rough_30	average surface roughness	X
agp_250	annual growth potential	X
alley_30	alley	
annual_precip	annual precipitation	
aspect_30	direction of slope face	X
barren_30	barren land	
bd_ss2_st_30	same as bulk density	X
bdrock_ss2_st	depth to bedrock	X
bike_pat_30	bike path	
dem_30	elevation	X
eastness_30	degree to which slope faces east	X
grass_herb_30	grassland/herbaceous cover	
imperv_30	percent impervious surfaces	
lat_gcs_30	latitude	
local_ro_30	ocal roads	
lon_gcs_30	longitude	
m2_01_precip	avg. precip (Jan)	
m2_01tmin	minimum temp (Jan)	
m2_02_precip	avg. precip (Feb)	
m2_02tmax	max temp (Feb)	
m2_02tmean	mean temp (Feb)	
m2_02tmin	min temp (Feb)	
m2_03_precip	avg. precip (Mar)	
m2_03tmax	max temp (Mar)	
m2_03tmean	mean temp (Mar)	
m2_03tmin	min temp (Mar)	
m2_04_precip	avg. precip (Apr)	
m2_04tmax	max temp (Apr)	
m2_04tmean	mean temp (Apr)	
m2_04tmin	min temp (Apr)	
m2_05_precip	avg. precip (May)	
m2_05tmax	max temp (May)	
m2_05tmean	mean temp (May)	
m2_05tmin	min temp (May)	
m2_06_precip	avg. precip (Jun)	
m2_06tmax	max temp (Jun)	
m2_06tmean	mean temp (Jun)	
m2_06tmin	min temp (Jun)	
m2_07_precip	avg. precip (Jul)	
m2_07tmax	max temp (Jul)	
m2_07tmean	mean temp (Jul)	
m2_07tmin	min temp (Jul)	
m2_08_precip	avg. precip (Aug)	
m2_08tmax	max temp (Aug)	
m2_08tmean	mean temp (Aug)	
m2_08tmin	min temp (Aug)	
m2_09_precip	avg. precip (Sept)	
m2_09tmax	max temp (Sept)	
m2_09tmean	mean temp (Sept)	
m2_09tmin	min temp (Sept)	
m2_10_precip	avg. precip (Oct)	
m2_10tmax	max temp (Oct)	
m2_10tmean	mean temp (Oct)	
m2_10tmin	min temp (Oct)	
m2_11_precip	avg. precip (Nov)	
m2_11tmax	max temp (Nov)	
m2_11tmean	mean temp (Nov)	
m2_11tmin	min temp (Nov)	
m2_12_precip	avg. precip (Dec)	
m2_12tmax	max temp (Dec)	
m2_12tmean	mean temp (Dec)	
m2_12tmin	min temp (Dec)	
m2_ann_precip	avg. annual precip	X
m2_anntmax	avg. annual max temp	
m2_anntmean	avg. annual mean temp	
m2_anntmin	avg. annual min temp	
nlcd_30	land cover type	
northness_30	degree to which slope faces north	X
parking_30	parking lot road	
pedestri_30	pedestrian trails	
pr_ss2_st	percent rocks	X
primary_30	primary roads	
private_30	private roads	
ramps_30	highway ramps	
road_30	euclidean distance to nearest road	
secondar_30	secondary roads	
service_30	service road	
shrub_30	shrub cover	
slope_30	inclination of landscape in degrees	X
surfarea_30	surface area of a grid cell	
tree_30	tree cover	
vehicula_30	4WD dirt trail	
win_precip	avg. winter precip	X

Figure S8. List of 83 landscape layers.

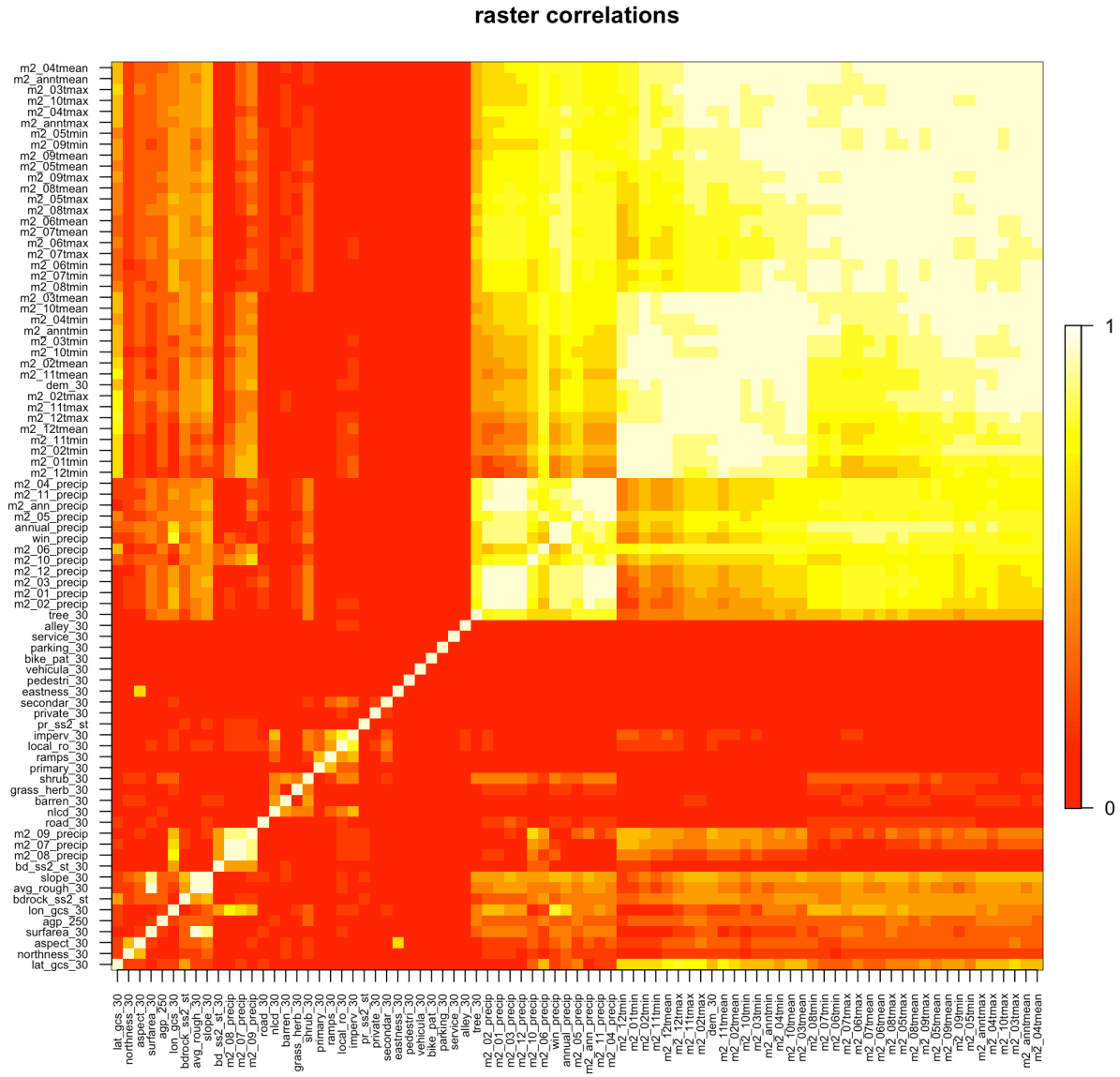

Figure S9. Matrix of correlations between spatial data layers.

layer name	Layer Category	Layer Description	Previous Use	6 rasters	12 rasters	24 rasters
imperv_30	anthropogenic	percent impervious surfaces		x	x	x
road_30	anthropogenic	euclidean distance to nearest road			x	x
agp_250	biotic	annual growth potential	X	x	x	x
grass_herb_30	biotic	grassland/herbaceous cover				x
shrub_30	biotic	shrub cover			x	x
m2_08_precip	climate	avg. precip (Aug)			x	x
m2_ann_precip	climate	avg. annual precip	X	x	x	x
m2_anntmax	climate	avg. annual max temp				x
m2_anntmean	climate	avg. annual mean temp			x	x
m2_anntmin	climate	avg. annual min temp				x
win_precip	climate	avg. winter precip	X			x
avg_rough_30	landscape	average surface roughness	X	x	x	x
aspect_30	landscape	direction of slope face	X			x
barren_30	landscape	percent barren land				x
dem_30	landscape	elevation	X	x	x	x
eastness_30	landscape	degree to which slope faces east	X			x
lat_gcs_30	landscape	latitude				x
lon_gcs_30	landscape	longitude				x
northness_30	landscape	degree to which slope faces north				x
slope_30	landscape	inclination of landscape in degrees	X			x
surfarea_30	landscape	surface area of a grid cell				x
bd_ss2_st_30	soils	bulk soil density	X		x	x
bdrock_ss2_st	soils	depth to bedrock	X	x	x	x
pr_ss2_st	soils	percent rocks	X		x	x
		TOTAL		6	12	24

Figure S10. List of environmental data layers evaluated in 6, 12, and 24-layer models.

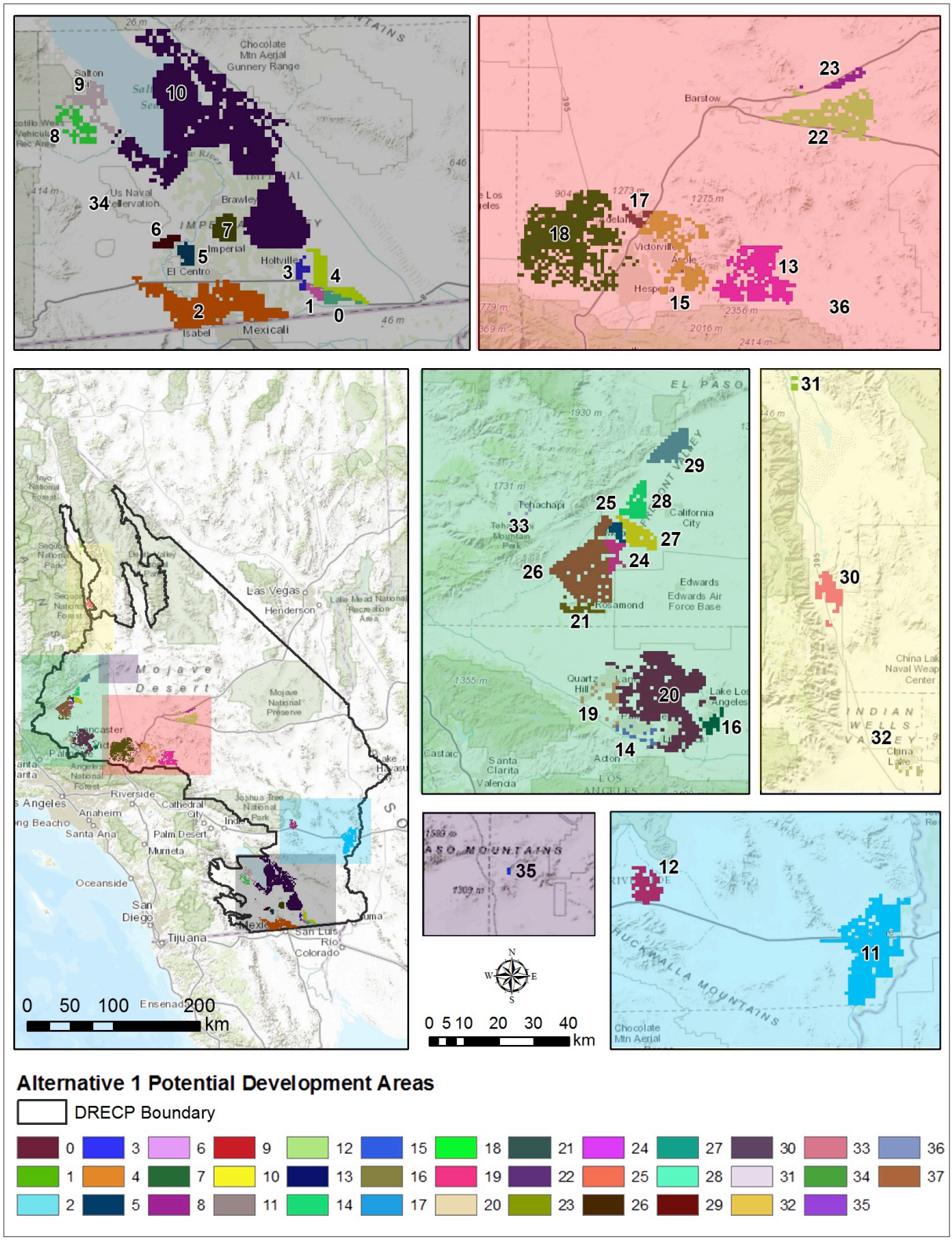

Figure S11. Spatial configuration of proposed development chunks in Alternative 1 that we analyzed.

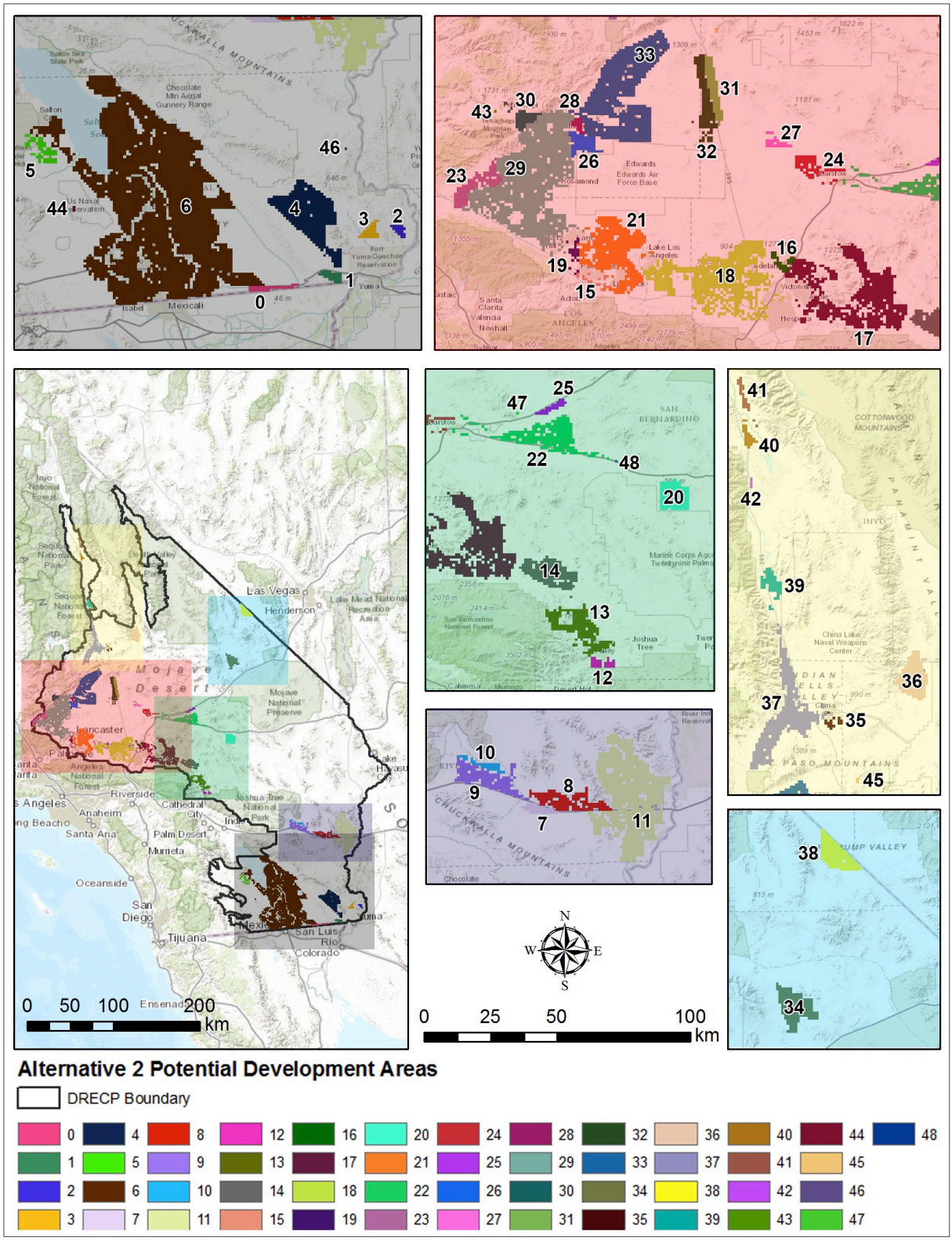

Figure S12. Spatial configuration of proposed development chunks in Alternative 2 that we analyzed.

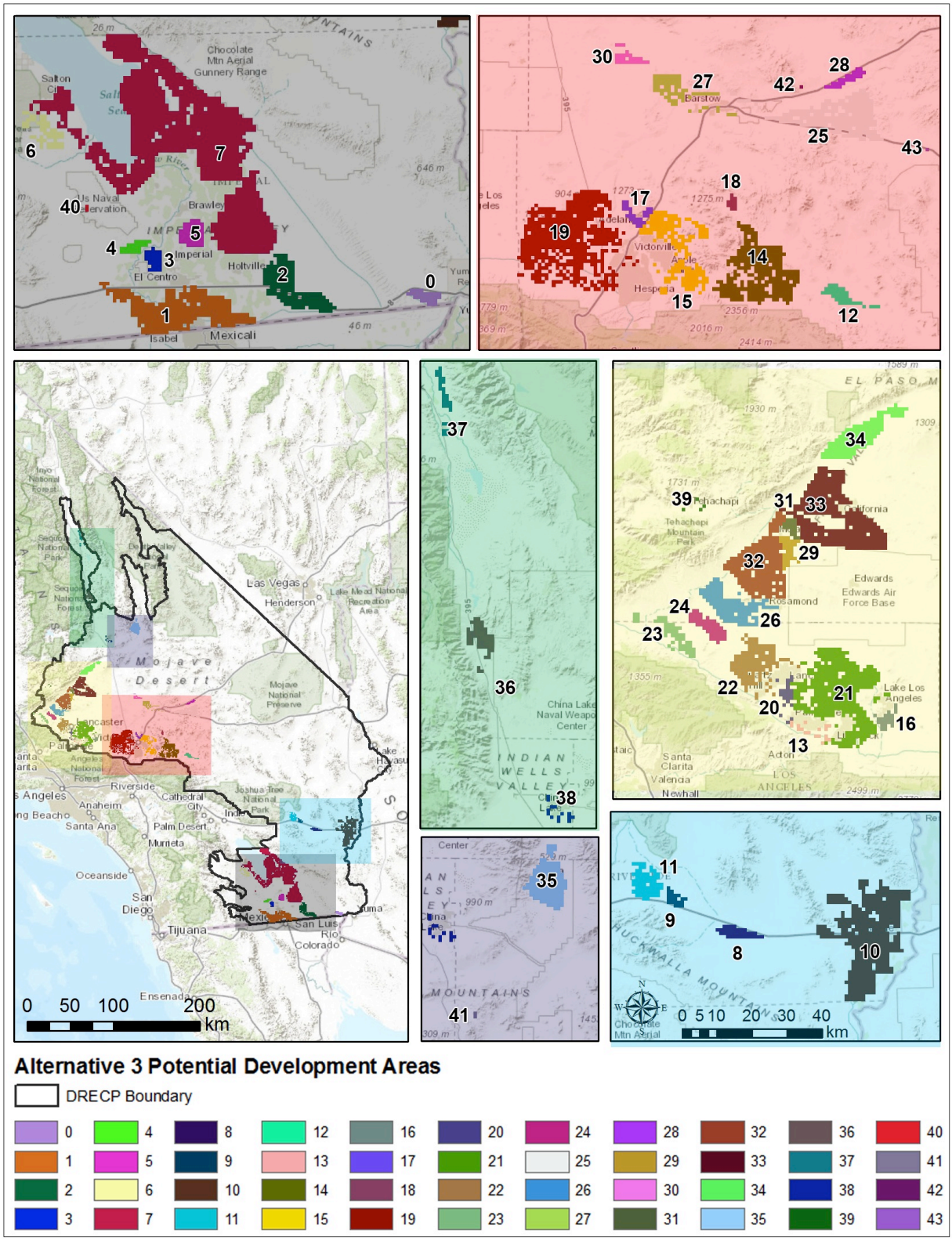

Figure S13. Spatial configuration of proposed development chunks in Alternative 3 that we analyzed.

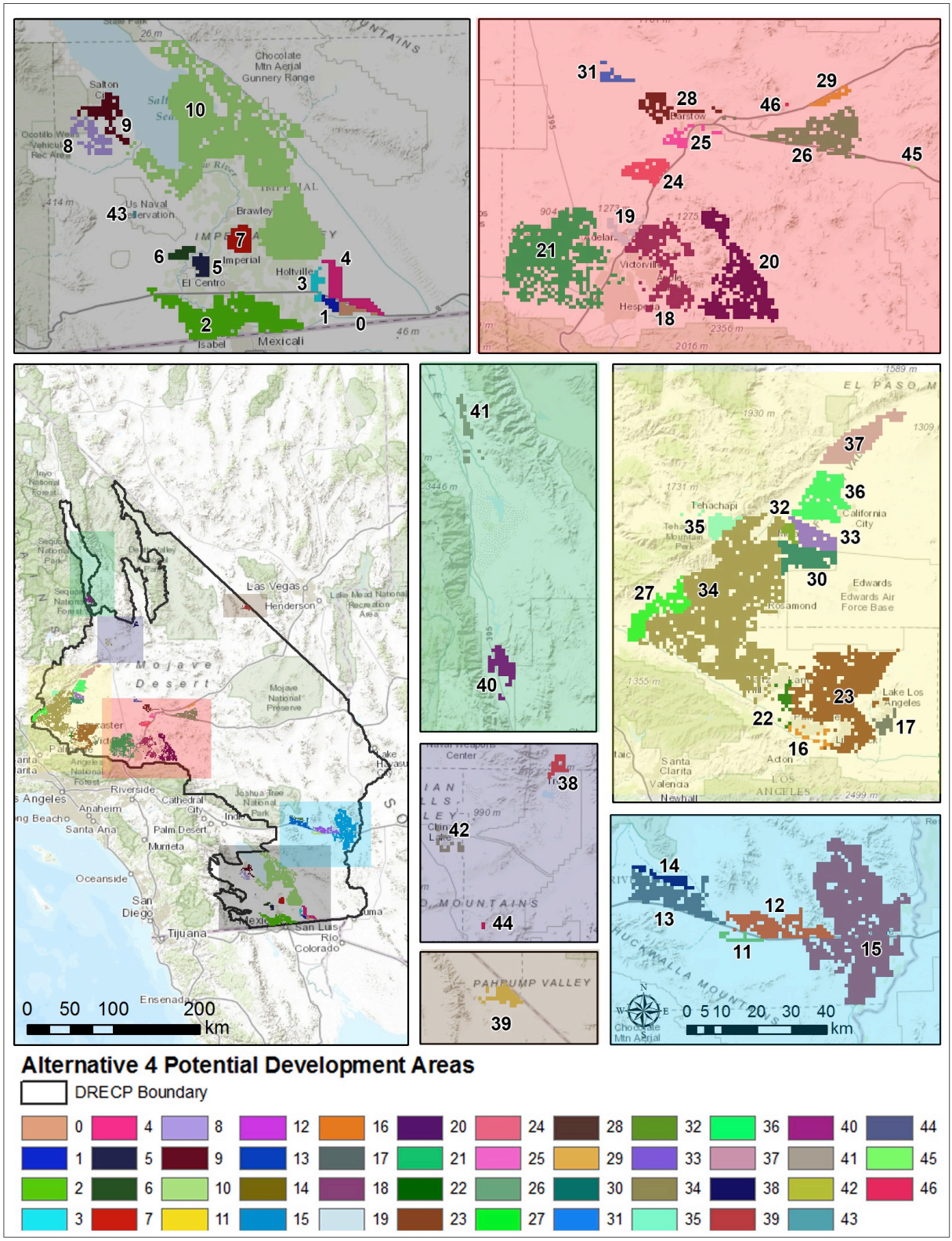

Figure S14. Spatial configuration of proposed development chunks in Alternative 4 that we analyzed.

Literature Cited

Aird D, Ross MG, Chen W-S, Danielsson M, Fennell T, Russ C, Jaffe DB, Nusbaum C, Gnirke A (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol* 12:R18.

Andersen, M.C., Joseph M. Watts, Jerome E. Freilich, Stephen R. Yool, Gery I. Wakefield, John F. McCauley, and Peter B. Fahnestock (2000). Regression-Tree modeling of desert tortoise habitat in the central Mojave desert. Ecological Applications 10:890–900.

Aronesty, Erik (2011). *ea-utils* : "Command-line tools for processing biological sequencing data"; http://code.google.com/p/ea-utils

Barton NH, Depaulis F, Etheridge AM (2002). Neutral evolution in spatially continuous populations. *Theoretical Population Biology*, **61**, 31–48.

Bates, Douglas and Maechler, Martin (2014). Matrix: Sparse and Dense Matrix Classes and Methods. R package version 1.1-4. http://CRAN.R-project.org/package=Matrix

Bolger AM, Lohse M, Usadel B (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics*, btu170.

Bradburd, G. S., Ralph, P. L. and Coop, G. M. (2013). Disentangling the effects of geographic and ecological isolation on genetic differentiation. Evolution, 67: 3258–3273.

Edwards, T, Cecil R. Schwalbe, Don E. Swann, Caren S. Goldberg. (2004). Implications of Anthropogenic Landscape Change on Inter-Population Movements of the Desert Tortoise (*Gopherus agassizii*). Conservation Genetics. Volume 5, Issue 4, pp. 485-499

Elith, J. and J. Leathwick. (2009). Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. Annual Review of Ecology, Evolution, and Systematics. Vol. 40: 677-697

Faircloth BC, Glenn TC (2012). Not All Sequence Tags Are Created Equal: Designing and Validating Sequence Identification Tags Robust to Indels. *PLoS ONE*, **7**, e42543.

Fuller CW, Middendorf LR, Benner SA *et al.* (2009). The challenges of sequencing by synthesis. *Nature Biotechnology*, **27**, 1013–1023.

Hagerty, B. E. , Kenneth E. Nussear, Todd C. Esque, C. Richard Tracy. (2011). Making molehills out of mountains: landscape genetics of the Mojave desert tortoise. Landscape Ecology. Volume 26, Issue 2, pp. 267-280

Hagerty, B. E. and C. R. Tracy. (2010). Defining population structure for the Mojave desert tortoise. Conservation Genetics 11:1795-1807.

Hudson, R. R. (2007). The Variance of Coalescent Time Estimates from DNA Sequences. Journal of Molecular Evolution 64(6):702-705.

Jenkins DG, Carey M, Czerniewska J *et al.* (2010). A meta-analysis of isolation by distance: relic or reference standard for landscape genetics? *Ecography*, **33**, 315–320.

Kim SY, Lohmueller KE, Albrechtsen A, Li Y, Korneliussen T, Tian G, Grarup N, Jiang T, Andersen G, Witte D, *et al*., (2011). Estimation of allele frequency and association mapping using next-generation sequencing data. BMC Bioinformatics 35:231.

Korneliussen TS, Albrechtsen A, Nielsen R. (2014). ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics*, **15**, 356.

Latch EK, Boarman WI, Walde A, Fleischer RC. (2011). Fine-Scale Analysis Reveals Cryptic Landscape Genetic Structure in Desert Tortoises. PLoS ONE 6(11): e27794.

Li H (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv:1303.3997 [q-bio]*.

Li H, Handsaker B, Wysoker A *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Li Y, Vinckenbosch N, Tian G, Huerta-Sanchez E, Jiang T, Jiang H, Albrechtsen A, Andersen G, Cao H, Korneliussen T, *et al*., (2010). Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. Nat Genet 42:969–972.

McRae, B. (2006). Isolation by Resistance. Evolution, 60(8): 1551-1561

McRae BH, Beier P (2007). Circuit theory predicts gene flow in plant and animal populations. *Proceedings of the National Academy of Sciences*, **104**, 19885–19890.

Menozzi P, Piazza A, and Cavalli-Sforza L (1978). Synthetic maps of human gene frequencies in Europeans. Science 201(4358):786-792.

Murphy R, Berry K, Edwards T, Leviton A, Lathrop A, Riedle J (2011). The dazed and confused identity of Agassiz's land tortoise, *Gopherus agassizii* (Testudines: Testudinidae) with the description of a new species and its consequences for conservation. ZooKeys 113: 39-71.

Nash-Williams C. (1959). Random walk and electric currents in networks. In: *Mathematical Proceedings of the Cambridge Philosophical Society*, pp. 181–194. Cambridge Univ Press.

National Land Cover Database (NLCD). http://www.mrlc.gov/nlcd2011.php

Nussear KE, Esque TC, Inman RD, Gass L, Thomas KA, Wallace CSA, Blainey JB, Miller DM, Webb RH. (2009). Modeling habitat of the desert tortoise (*Gopherus agassizii*) in the Mojave and parts of the Sonoran Deserts of California, Nevada, Utah, and Arizona. U.S. Geological Survey open-file report 2009-1102.

Novembre, J., Toby Johnson, Katarzyna Bryc, Zolta´n Kutalik, Adam R. Boyko, Adam Auton, Amit Indap, Karen S. King, Sven Bergmann, Matthew R. Nelson, Matthew Stephens, Carlos D. Bustamante. (2008). Genes mirror geography within Europe. Nature 456, 98–101.

Oblój J. (2004) The Skorokhod embedding problem and its offspring. *Probability Surveys*, **1**, 321–392.

Patterson N, Price AL, Reich D (2006). Population Structure and Eigenanalysis. PLoS Genet 2(12): e190

PRISM Climate Group. http://www.prism.oregonstate.edu/normals/

Sahagun L (2015) Desert renewable energy plan is altered to win counties' support. *Los Angeles Times*. <http://www.latimes.com/local/california/la-me-0311-desert-20150311-story.html>.

Sambrook *et al*. (2001). Molecular Cloning: A Laboratory Manual

Sexton JP, Hangartner SB, Hoffmann AA (2014). Genetic Isolation by Environment or Distance: Which Pattern of Gene Flow Is Most Common? *Evolution*, **68**, 1–15.

Slatkin M (1991). Inbreeding coefficients and coalescence times. Genet Res 58: 167–175.

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. Economic geography 46:234-240.

TIGER Census road classification, (2012). http://www.census.gov/cgi-bin/geo/shapefiles2012/main

U.S. Bureau of Land Management (2016) Desert Renewable Energy Conservation Plan Record of Decision for the Land Use Plan Amendment to the California Desert Conservation Area Plan, Bishop Resource Management Plan, and Bakersfield Resource Management Plan. BLM/CA/PL-2016/03+1793+8321.

USDA NRCS Soil Data Mart http://websoilsurvey.sc.egov.usda.gov/App/WebSoilSurvey.aspx.

USGS National Map Viewer. http://viewer.nationalmap.gov/viewer/

Wallace, C.S.A., Thomas, K.A. (2008). An annual plant growth proxy in the Mojave desert using MODIS-EVI data. Sensors, 6, 7792-7808.

Wang, I.J., Bradburd, G.S. (2014). Isolation by Environment. Molecular Ecology 23:5649-5662.

Wright, S. (1943). Isolation by distance. Genetics 28:114-138.