Supplemental Material:

Diabolical survival in Death Valley: recent pupfish colonization, gene flow, and genetic assimilation in the smallest species range on earth

CHRISTOPHER H. MARTIN[1], JACOB E. CRAWFORD[2,3,4], BRUCE J. TURNER[5], LEE H. SIMONS[6]

[1]Department of Biology, University of North Carolina at Chapel Hill, NC, USA

[2]Department of Integrative Biology, University of California, Berkeley, CA, USA

[3]Center for Theoretical Evolutionary Genomics, University of California, Berkeley, CA, USA

[4]Google, Inc., 1600 Amphitheatre Parkway, Mountain View, CA, USA

[5]Department of Biological Sciences, Virginia Tech, VA, USA

[6]Southern Nevada Fish and Wildlife Office, Las Vegas, NV, USA

Supplemental Methods

Sample collection

C. diabolis is one of the most endangered fish on earth and thus collecting tissue from live animals was impossible at the time of this study. From 2007 – 2012, all dead fish encountered in Devils Hole (*n* = 20) were collected by National Park Service staff after ~12 – 48 hours of putrefication in the 32º C water (Appendix S1). Specimens were sometimes fixed in formalin (Davidson's solution) and stored in 70% ethanol at room temperature. Highly-degraded DNA showing a large fragment size distribution was successfully extracted from 13 samples with Qiagen blood and tissue kits. Additional samples from the School Spring refuge population collected in 1989 (*n* = 3) were also used. All other Death Valley samples came from archived specimens used for previous studies [1,2]. Outgroup *Cyprinodon* samples were previously collected in the wild [3] or, if extinct in the wild (*n* = 6), provided by the American Killifish Association *Cyprinodon* species maintenance group from existing captive populations (Appendix S1). *Cyprinodon* species were sampled from all major extant lineages, including the earliest split within the clade between the *artifrons*+Chichancanab endemic species flock and all other extant species [3,4].

Genomic library preparation and bioinformatics

Double-digest RADseq libraries were prepared following Peterson et al. [5] with minor modifications as described in Martin et al. [6]. SbfI and NlaIII restriction enzymes were used for digestion. The *Cyprinodon variegatus* genome assembly (v. 1.0, 1035 Mb, 81x coverage) used for aligning reads is relatively high-quality, containing 9,258 scaffolds with an N50 scaffold size of 835 kb (NCBI: Wesley Warren, "Whole genome assembly resources for aquatic models of human disease", Grant ID 8 R24 OD011198-02, National Center for Research Resources). Empirical fragment size selection windows ranged from 300-400 bp using a Blue Pippin Prep (Sage Science). Twelve cycles were used for amplification across two independent reactions per library to limit PCR error. 145 individuals with 4-8 bp molecular barcodes (described in [7]) were sequenced on one and a half Illumina 2000 HiSeq lanes at the Vincent J. Coates Genomic Sequencing Center at UC Berkeley (one lane was pooled with 47 individuals from another study). Respectively, 43.6 and 154.7 million 95-bp and 120-bp single-end raw reads were sequenced with 67% and 76% recovery of high-quality, barcoded reads with an intact restriction site using default settings in sort_reads (Stacks v. 1.20; [8]). Read quality did not substantially decline along each read, ranging from a median Phred quality score of 42 (0.99994% accuracy) to 34 (0.9996% accuracy) from read positions 15 to 100 in both Illumina lanes, starting around position 55.

Raw reads were de-multiplexed and sorted for quality using default settings in *process_radtags* in the Stacks pipeline [9] and aligned to the *Cyprinodon variegatus* draft genome (v. 1.0) using bowtie 2 (v. 2.2.3; [10]) with very high sensitivity settings and end-to-end alignment. Aligned reads were merged into homologous loci by their genomic position, not sequence identity (*cstacks -g*). SNPs were called using a likelihood model across individuals. We then used rxstacks to exclude problematic loci with a log-likelihood less than -100 or if more than 25% of individuals contained multiple loci matching a single catalog locus (*conf_limit* = 0.25) or any non-biological haplotypes (*--prune_haplo*). Loci with a minimum of 8 sequenced reads were exported from the Stacks pipeline in .plink format (*-m* 8 *--plink*). We used PLINK [11] to exclude low-coverage individuals genotyped at less than 5% of total loci over all populations/species and retained only those loci present in >50% of all high-coverage individuals (n = 56) for downstream analyses.

Population genetic structure and introgression analyses

Principal components of genetic variance were calculated using probabilistic PCA in the pcaMethods package in R [12]. Bayesian clustering analyses with STRUCTURE sampled one SNP per locus (4,679 SNPs) and were aggregated using CLUMPP [13] and STRUCTURE Harvester [14] from 10 independent runs of 50,000 generations each after discarding the first 50,000 generations as burn-in (Table S4). Confidence in estimates of ancestry proportions was assessed by comparing estimates across independent runs of STRUCTURE.

Inference of introgression was made using three complementary approaches. First, formal tests of introgression used D-statistics, also known as ABBA/BABA tests [15–17], to determine if any populations shared more residual alleles than expected under a tree-like model of branching. D-statistics were calculated with a custom script after thinning to one informative site (i.e. ABBA or BABA) per locus. Z-scores were calculated based on 500 bootstrap datasets sampled from the thinned dataset. Second, estimated ancestry proportions of each individual in STRUCTURE were used to complement these formal tests. Third, Treemix (v. 1.12; [18]) was used to visualize variance-covariance relationships in allele frequencies among Death Valley populations. Four migration events were fit to a maximum likelihood population tree to estimate which populations showed the strongest evidence for introgression.

Phylogenetic analyses and time-calibration

We constructed a new catalogue of homologous loci for taxa used in phylogenetic analyses by merging loci by genomic position and extracting loci present in at least 4 taxa following recommendations for clustering thresholds in phylogenetic analyses of RADseq data [19,20]. A fasta file was exported from Stacks and sorted by locus with a custom perl script (provided as a supplemental file in the supplemental material) and then concatenated into a nexus file using

Geneious (v. 7.1.7; [21]). A single haplotype was sampled from one high-coverage individual per population. We used a coalescent process with constant population size for our tree prior. Nucleotide substitution rates were modeled by the general time-reversible model (GTR) plus gamma-distributed rate variation across loci. We used an uncorrelated lognormal model or a random local model for the molecular clock. Four independent MCMC chains were run on the CIPRES cluster [22] using BEAST (v. 1.8.1; [23]), totaling 186 million generations after discarding burn-in. We confirmed the convergence of all four runs in ≤ 4 million generations using Tracer (v. 1.6) and all parameters exceeded an effective sample size of 153. We also explored the effects of additional phylogenetic models on parameter estimation (discussed below).

We calibrated our phylogeny (16,567 concatenated loci, 38,069 informative sites) with the only well-defined recent geological event known for *Cyprinodon*: the 8,000 ± 200 year age of Laguna Chichancanab [24,25], an endorheic basin which contains an endemic species flock of *Cyprinodon* pupfishes (Fig. 2d; Humphries & Miller 1981). It is unlikely that the Chichancanab species flock diverged before the basin formed because these species cannot tolerate fish predators found in all neighboring surface waters (at least 3 Chichancanab pupfish species are now extinct due to invasive fishes [3,27]); therefore, our calibration places a lower bound on the spontaneous mutation rate [28]. We placed a normal prior on the divergence time between *C. artifrons* (the most closely related species from the Yucatan coast) and the stem age of the Chichancanab lineage with a mean of 8,000 years and standard deviation of 100 years. This age and associated error (95% confidence interval: ± 200 years) were based on multiple core samples and multiple lines of evidence, including stable isotope data and shifts from terrestrial to aquatic invertebrate communities [24,25]. No other accurate fossil or geological age estimates for *Cyprinodon* exist (reviewed in Martin & Wainwright 2011: supplement). There is a single posterior half of one fossil assigned to *Cyprinodon* which was collected in Death Valley; however, no synapomorphies were used for this designation and the rock was ascribed to Late Pliocene strata based only on "the presence of a *Cyprinodon*" (p. 316, Miller 1945). Furthermore, the vertebral count of this fossil lies outside the extant range of Cyprinodontinae (T. Echelle, pers. comm.).

Estimation of the mutation rate in pupfishes

Estimating mutation rates across animal taxa, and even within humans, remains a difficult and controversial problem [29,30]. For example, phylogenetic estimates of substitution rates calibrated with ancient fossil or geographic vicariance events appear to be at least an order of magnitude slower than mutation rates observed at more recent timescales (<100,000 years) based on high-coverage sequencing of pedigrees, comparisons between ancient and modern DNA samples, and mutation-accumulation lines [31–35]. Estimates of mutation rates in fishes are sparse, particularly for nuclear DNA. One study found that substitution rates at four-fold degenerate sites were twice as high between two pufferfish species ($1.46e^{-8}$ per site per year) as between humans and mouse for unknown reasons [36]. One of the key studies documenting that substitution rates are dependent on the time-scale of priors used for calibration found that mtDNA substitution rates are an order of magnitude faster in the past 200 kya for riverine fishes using internal calibrations based on the age of different river basins [34]. Overall, one recommendation emerging from this controversy is to calibrate recent phylogenies with internal calibrations on a similar timescale to the focal group, rather than distantly related outgroups with a better fossil record [34,37]. We have followed this approach here. However, additional uncertainty is introduced by the largely unknown variation in mutation rates across taxa and the biased genomic sampling provided by double-digest RADseq library preparation.

We explored several strategies to determine whether our methods or dataset may have biased our mutation rate estimate. First, we explored additional phylogenetic models (random local clock), more stringent filtering of RAD loci (*m* = 20 reads instead of 8 to reduce sequencing error), and taxon subsets (only the Chichancanab species and closest outgroup) to determine how these variables affected our estimate of the mutation rate (Table S2). We discarded burn-in and checked for stationarity in our BEAST analyses as described previously.

Second, we also completely reran our pipeline from raw reads trimmed to 53 bp to remove later positions with decreased read qualities, which declined from median Phred quality scores of 42 (0.99994% accuracy) to 34 (0.9996% accuracy) from read positions 15 to 100, starting around position 55. We used this empirical evaluation of declining read qualities in FastQC (Babraham Bioinformatics) to guide our trimming strategy. We re-aligned trimmed reads and used the latest version of Stacks (v. 1.34: [9]) to assemble mapped reads into homologous loci and call SNPs as described previously. We then estimated a new time-calibrated phylogeny from a concatenated set of 4,159 53-bp loci genotyped in more than 50% of individuals to explore how this trimming procedure and new pipeline affected our estimate of the mutation rate (Table S2, Fig. S4) and a new principal component analysis of genetic variance to explore how trimming affected population structure (Fig. S5). We attempted to redo our *dadi* analysis; however, trimming removed nearly 50% of our data (including all true positive SNP calls in this region) and our *dadi* model did not converge due to insufficient data to constrain the prior.

There are many reasons to expect RADseq data to be a biased under- or over-representation of genomic diversity due to selective targeting of GC-rich loci, PCR amplification bias, allele dropout at polymorphic sites [40], and other unknown biases [41,42]. For example, our infrequent-cutting restriction enzyme SbfI targets extremely GC-rich sites (6 out of 8 sites in the recognition sequence are GC). Although restriction sites are removed for downstream analyses, this means that GC-rich genomic regions are targeted (such as protein-coding regions) which may result in the overestimation of the genome-wide mutation rate due to mutation rates at CpG sites [43,44]. Second, PCR amplification during library preparation may preferentially amplify GC-rich fragments and any errors introduced will be amplified in each cycle, resulting in genotyping errors despite seemingly sufficient read depths [42]. Third, filtering for loci shared across taxa biases the mutation rate due to allelic dropout: homologous loci shared by more taxa are more likely to be evolving more slowly and retain a shared restriction site needed for detection. Thus, more stringent filtering for shared loci will bias estimated mutation rates downward while more lenient filtering will bias mutation rates upward and increase the amount of sequencing error and spurious loci. This has now been demonstrated in simulation studies [41], empirically [45], and we observed this pattern in our own dataset (unpublished data). Finally, allelic dropout results in the underestimation of genetic diversity due to incorrectly calling all polymorphic restriction sites as homozygous [40]. Genetic diversity estimates in Table S1 may be underestimated, but this bias is not expected to affect estimates of genetic differentiation or introgression among species [40]. We pooled two independent PCR reactions for each library and compared different levels of read depths and taxon filtering in our analyses to examine the effects of these biases. However, the biased genomic sampling of RADseq is inescapable.

Nonetheless, although our dataset may be biased, Bayesian posterior estimates of divergence time are extremely sensitive to calibration priors, rather than the observed heterozygosity within a dataset [46]. Thus, our estimate of the age of *diabolis* depends mainly on the accuracy of our calibration choice, not the underlying bias in our dataset, because any mutational bias present is rescaled to an external timescale and we used this same dataset for later demographic analysis. For example, if we time-calibrate our phylogeny using a fixed molecular clock with the human mutation rate of $0.5e^{-9}$ mutations/site/year, this places the age of the Laguna Chichancanab species flock at 4.9 million years, vastly greater than the 8,000-year geological age of this basin [24,25]. This strongly suggests that either pupfish mutation rates greatly exceed human rates or our RADseq dataset is a biased sample of heterozygosity.

Demographic modeling with dadi

We used *dadi* to fit a simple demographic model including divergence time, migration between populations, and effective population sizes before and after the split to the observed two-dimensional site frequency spectrum between these species (Fig. 3, Table S2). We used a generation time of 9 months for *diabolis* based on the observed peak reproductive periods in March and October and annual lifecycle of 1 year [47,48], which captures the age at which these fish are likely to contribute most to the next generation. To increase our sample sizes, we pooled all *mionectes*, *amargosae/shoshone/nevadensis*, and *salinus/milleri* populations into three groups based on their genetic clustering (Fig. 2a-b). We polarized (unfolded) the allele frequency spectrum using *salinus/milleri*. We then collapsed the site frequency spectrum to eight chromosomes to maximize the number of sites and sampled one SNP per locus to reduce the effects of linkage disequilibrium in our dataset. We bootstrapped 500 samples from this dataset to obtain empirical 95% confidence intervals for demographic parameters in our model.

References

1. Duvernell, D. D. & Turner, B. J. 1998 Variation and Divergence of Death Valley Pupfish Populations at Retrotransposon-Defined Loci. , 363–371.

2. Echelle, A. & Dowling, T. 1992 Mitochondrial DNA variation and evolution of the Death Valley pupfishes (Cyprinodon, Cyprinodontidae). *Evolution (N. Y).* **46**, 193–206.

3. Martin, C. H. & Wainwright, P. C. 2011 Trophic novelty is linked to exceptional rates of morphological diversification in two adaptive radiations of *Cyprinodon* pupfish. *Evolution* **65**, 2197–212. (doi:10.1111/j.1558-5646.2011.01294.x)

4. Echelle, A. a., Carson, E. W., Echelle, A. F., Van Den Bussche, R. a., Dowling, T. E. & Meyer, A. 2005 Historical Biogeography of the New-World Pupfish Genus Cyprinodon (Teleostei: Cyprinodontidae). *Copeia* **2005**, 320–339. (doi:10.1643/CG-03-093R3)

5. Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S. & Hoekstra, H. E. 2012 Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS One* **7**, e37135. (doi:10.1371/journal.pone.0037135)

6. Martin, C. H., Cutler, J. S., Friel, J. P., Dening, T., Coop, G. & Wainwright, P. C. 2015 Complex histories of repeated colonization and hybridization cast doubt on the clearest examples of sympatric speciation in the wild. *Evolution (N. Y).*

7. Martin, C. H. & Feinstein, L. C. 2014 Novel trophic niches drive variable progress towards ecological speciation within an adaptive radiation of pupfishes. *Mol. Ecol.* **23**, 1846–62. (doi:10.1111/mec.12658)

8. Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A. & Cresko, W. A. 2013 Stacks: an analysis tool set for population genomics. *Mol. Ecol.* **22**, 3124–40. (doi:10.1111/mec.12354)

9. Catchen, J., Hohenlohe, P. a, Bassham, S., Amores, A. & Cresko, W. a 2013 Stacks: an analysis tool set for population genomics. *Mol. Ecol.* **22**, 3124–40. (doi:10.1111/mec.12354)

10. Langmead, B. & Salzberg, S. 2012 Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359.

11. Purcell, S., Neale, B. & Todd-Brown, K. 2007 PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575.

12. Stacklies, W., Redestig, H., Scholz, M., Walther, D. & Selbig, J. 2007 pcaMethods--a bioconductor package providing PCA methods for incomplete data. *Bioinformatics* **23**, 1164–7. (doi:10.1093/bioinformatics/btm069)

13. Jakobsson, M. & Rosenberg, N. a 2007 CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* **23**, 1801–6. (doi:10.1093/bioinformatics/btm233)

14. Earl, D. A. 2012 STRUCTURE HARVESTER: a webite and program for visualizing STRUCTURE output and implementing the Evanno method. *Conserv. Genet. Resour.* **4**, 359–361.

15. Heliconius, T. & Consortium, G. 2012 Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature* **487**, 94–8. (doi:10.1038/nature11041)

16. Green, R. E. et al. 2010 A draft sequence of the Neandertal genome. *Science* **328**, 710–22. (doi:10.1126/science.1188021)

17. Durand, E. Y., Patterson, N., Reich, D. & Slatkin, M. 2011 Testing for ancient admixture between closely related populations. *Mol. Biol. Evol.* **28**, 2239–52. (doi:10.1093/molbev/msr048)

18. Pickrell, J. K. & Pritchard, J. K. 2012 Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. *PLoS Genet.* **8**, e1002967. (doi:10.1371/journal.pgen.1002967)

19. Eaton, D. a. R. 2013 PyRAD: assembly of de novo RADseq loci for phylogenetic analyses. (doi:10.1101/001081)

20. Rubin, B. E. R., Ree, R. H. & Moreau, C. S. 2012 Inferring Phylogenies from RAD Sequence Data. *PLoS One* **7**, e33394. (doi:10.1371/journal.pone.0033394)

21. Kearse, M. et al. 2012 Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647–9. (doi:10.1093/bioinformatics/bts199)

22. Miller, M. A., Pfeiffer, W. & Schwartz, T. 2010 Creating the CIPRES Science Gateway for inference of large phylogenetic trees. *2010 Gatew. Comput. Environ. Work.* , 1–8. (doi:10.1109/GCE.2010.5676129)

23. Drummond, A. J. & Rambaut, A. 2007 BEAST : Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **8**, 1–8. (doi:10.1186/1471-2148-7-214)

24. Covich, A. & Stuiver, M. 1974 Changes in the oxygen 18 as a measure of long-term fluctuations in tropical lake levels and molluscan populations. *Limnol. Oceanogr.* **19**, 682–691.

25. Hodell, D., Curtis, J. & Brenner, M. 1995 Possible role of climate in the collapse of Classic Maya civilization. *Nature* **375**, 391–394.

26. Humphries, J. & Miller, R. R. 1981 A remarkable species flock of pupfishes, genus *Cyprinodon*, from Yucatan, Mexico. *Copeia* **1981**, 52–64.

27. Strecker, U. 2006 The impact of invasive fish on an endemic Cyprinodon species flock (Teleostei) from Laguna Chichancanab, Yucatan, Mexico. *Ecol. Freshw. Fish* **15**, 408–418. (doi:10.1111/j.1600-0633.2006.00159.x)

28. Lanfear, R., Kokko, H. & Eyre-Walker, A. 2014 Population size and the rate of evolution. *Trends Ecol. Evol.* **29**, 33–41. (doi:10.1016/j.tree.2013.09.009)

29. Scally, A. & Durbin, R. 2012 Revising the human mutation rate: implications for understanding human evolution. *Nat. Rev. Genet.* **13**, 824–824. (doi:10.1038/nrg3353)

30. Ho, S. Y. W., Phillips, M. J., Cooper, A. & Drummond, A. J. 2005 Time dependency of molecular rate estimates and systematic overestimation of recent divergence times. *Mol. Biol. Evol.* **22**, 1561–8. (doi:10.1093/molbev/msi145)

31. Santos, C., Montiel, R., Sierra, B., Bettencourt, C., Fernandez, E., Alvarez, L., Lima, M., Abade, A. & Aluja, M. P. 2005 Understanding differences between phylogenetic and pedigree-derived mtDNA mutation rate: A model using families from the Azores Islands (Portugal). *Mol. Biol. Evol.* **22**, 1490–1505. (doi:10.1093/molbev/msi141)

32. Millar, C. D., Dodd, A., Anderson, J., Gibb, G. C., Ritchie, P. a, Baroni, C., Woodhams, M. D., Hendy, M. D. & Lambert, D. M. 2008 Mutation and evolutionary rates in adélie penguins from the antarctic. *PLoS Genet.* **4**, e1000209. (doi:10.1371/journal.pgen.1000209)

33. Subramanian, S., Denver, D. R., Millar, C. D., Heupink, T., Aschrafi, A., Emslie, S. D., Baroni, C. & Lambert, D. M. 2009 High mitogenomic evolutionary rates and time dependency. *Trends Genet.* **25**, 482–6. (doi:10.1016/j.tig.2009.09.005)

34. Burridge, C. P., Craw, D., Fletcher, D. & Waters, J. M. 2008 Geological dates and molecular rates: fish DNA sheds light on time dependency. *Mol. Biol. Evol.* **25**, 624–33. (doi:10.1093/molbev/msm271)

35. Ho, S. Y. W., Saarma, U., Barnett, R., Haile, J. & Shapiro, B. 2008 The effect of inappropriate calibration: three case studies in molecular ecology. *PLoS One* **3**, e1615. (doi:10.1371/journal.pone.0001615)

36. Jaillon, O. et al. 2004 Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *Nature* **431**, 946–957. (doi:10.1038/nature03025)

37. Ho, S. Y. W. 2007 Calibrating molecular estimates of substitution rates and divergence times in birds. *J. Avian Biol.* **38**, 409–414. (doi:10.1111/j.2007.0908-8857.04168.x)

38. Brix, K. V & Grosell, M. 2012 Comparative characterization of Na+ transport in Cyprinodon variegatus variegatus and Cyprinodon variegatus hubbsi: a model species complex for studying teleost invasion of freshwater. *J. Exp. Biol.* **215**, 1199–209. (doi:10.1242/jeb.067496)

39. Leffler, E. M., Bullaughey, K., Matute, D. R., Meyer, W. K., Ségurel, L., Venkat, A., Andolfatto, P. & Przeworski, M. 2012 Revisiting an old riddle: what determines genetic diversity levels within species? *PLoS Biol.* **10**, e1001388. (doi:10.1371/journal.pbio.1001388)

40. Arnold, B., Corbett-Detig, R. B., Hartl, D. & Bomblies, K. 2013 RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. *Mol. Ecol.* **22**, 3179–3190. (doi:10.1111/mec.12276)

41. Huang, H. & Knowles, L. L. 2014 Unforeseen Consequences of Excluding Missing Data from Next-Generation Sequences: Simulation Study of RAD Sequences. *Syst. Biol.* **0**, 1–9. (doi:10.1093/sysbio/syu046)

42. Puritz, J. B., Matz, M. V, Toonen, R. J., Weber, J. N., Bolnick, D. I. & Bird, C. E. 2014 Demystifying the RAD fad. *Mol. Ecol.* **23**, 5937–42. (doi:10.1111/mec.12965)

43. Guryev, V., Koudijs, M. J., Berezikov, E., Johnson, S. L., Plasterk, R. H. a, Eeden, J. M. Van, Cuppen, E. & Eeden, F. J. M. Van 2006 Genetic variation in the zebrafish Genetic variation in the zebrafish. , 491–497. (doi:10.1101/gr.4791006)

44. Nachman, M. W. & Crowell, S. L. 2000 Estimate of the mutation rate per nucleotide in humans. *Genetics*

45. Leache, a. D., Chavez, a. S., Jones, L. N., Grummer, J. a., Gottscho, a. D. & Linkem, C. W. 2015 Phylogenomics of Phrynosomatid Lizards: Conflicting Signals from Sequence Capture Versus Restriction Site Associated DNA Sequencing. *Genome Biol. Evol.* **7**, 706–719. (doi:10.1093/gbe/evv026)

46. Warnock, R. C. M., Parham, J. F., Joyce, W. G., Lyson, T. R. & Donoghue, P. C. J. 2014 Calibration uncertainty in molecular dating analyses : there is no substitute for the prior evaluation of time priors.

47. Deacon, J. E., Taylor, F. R., Pedretti, J. W. & Pedretti, W. 1995 Egg viability and ecology of Devils Hole pupfish : Insights from captive propagation. *Southwest. Nat.* **40**, 216–223.

48. Riggs, A. & Deacon, J. 2002 Connectivity in desert aquatic ecosystems: The Devils Hole story. *Spring-fed Wetl. important Sci. Cult. Resour. Intermt. Reg.* **11**.

49. Evanno, G., Regnaut, S. & Goudet, J. 2005 Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol. Ecol.* **14**, 2611–20. (doi:10.1111/j.1365-294X.2005.02553.x)

50. Martin, A. P., Echelle, A. a., Zegers, G., Baker, S. & Keeler-Foster, C. L. 2011 Dramatic shifts in the gene pool of a managed population of an endangered species may be exacerbated by high genetic load. *Conserv. Genet.* **13**, 349–358. (doi:10.1007/s10592-011-0289-7)

Table S1. Genetic diversity (π), private alleles, total nucleotide sites examined, and percentage of polymorphic sites in Death Valley pupfishes and additional pupfish outgroups for comparison (EW = extinct in the wild based on IUCN designation or unpublished observations). Number of individuals sequenced in each population is indicated (Appendix S1).

species/subspecies	**location**	**genetic diversity**	**private alleles**	**total sites**	**%polymorphic**
Cyprinodon diabolis (n = 4)	Devils Hole	0.0009	1019	1017310	0.0938
C. nevadensis mionectes (n = 8)	Point-of-Rocks	0.0023	1572	1229973	0.2348
C. nevadensis mionectes (n = 1)	Big Spring	0.0006	272	403729	0.0577
C. nevadensis pectoralis (n = 7)	Indian Spring	0.0001	21	141445	0.0099
C. nevadensis amargosae (n = 7)	Amargosa River	0.002	1008	1261354	0.1954
C. nevadensis shoshone (n = 5)	Shoshone Spring	0.0013	539	874517	0.1337
C. nevadensis nevadensis (n = 7)	Saratoga Spring	0.0021	1360	1314878	0.2135
C. salinus salinus (n = 2)	Salt Creek	0.0009	1244	1184303	0.0888
C. salinus milleri (n = 1)	Cottonball Marsh	0.0004	264	367191	0.0400
All Death Valley species	Death Valley	0.0036	10224	1140524	0.7117
C. artifrons (n = 1)	coastal Cancun, Mexico	0.0034	3212	1185394	0.3354
C. variegatus (n = 1)	coastal San Salvador, Bahamas	0.0039	9900	1311769	0.3862
C. alvarezi (n = 1)	captive colony - EW	0.0011	2211	965072	0.1149
C. maya (n = 1)	captive colony - EW	0.0027	657	964452	0.2722
C. veronicae (n = 1)	captive colony - EW	0.0015	5160	1241190	0.1491

Table S2. Estimates of the pupfish mutation rate (median substitution rate per site per year) based on various modeling assumptions and datasets and their effect on the median divergence time for *diabolis*. The 95% credible intervals for each median substitution rate and the 95% confidence intervals for each *diabolis* divergence time estimate are indicated in brackets. Note that demographic estimates of divergence time from our *dadi* analysis scale linearly with mutation rate and we used the median substitution rate for each estimate. In our trimmed dataset, we also ran analyses after pruning two rogue taxa with minimal support in the tree (19 taxa: Fig. S4).

dataset: taxa subset	filter (min. reads)	loci	clock model	median substitution rate (mutations/site/year)	*diabolis* divergence time (years)
original (21 taxa)	8	16,567	uncorrelated lognormal	5.37e^{-7} [4.01e^{-7}-7.01e^{-7}]	255 [**105**-408]
original (21 taxa)	20	2,437	uncorrelated lognormal	2.06e^{-7} [1.64e^{-7}-2.53e^{-7}]	665 [541-**835**]
5 taxa: Chichancanab+*artifrons*	8	4,889	uncorrelated lognormal	5.69e^{-7} [5.47e^{-7}-5.91e^{-7}]	241 [232-250]
5 taxa: Chichancanab+*artifrons*	8	4,889	random local	3.14e^{-7} [2.86e^{-7}-3.68e^{-7}]	436 [372-479]
5 taxa: Chichancanab+*artifrons*	20	2,437	uncorrelated lognormal	3.17e^{-7}[2.97e^{-7}-3.37e^{-7}]	431 [406-461]
19 taxa: trimmed to first 53 bp	8	4,159	uncorrelated lognormal	2.34e^{-7}[1.85e^{-7}-2.93e^{-7}]	585 [467-740]
21 taxa: trimmed to first 53 bp	8	4,159	uncorrelated lognormal	4.42e^{-7}[3.31e^{-7}-5.98e^{-7}]	310 [229-413]

Table S3. Maximum likelihood parameter estimates in *dadi* for a simple demographic model of the split between *mionectes* and *amargosae/shoshone/nevadensis* including a symmetric migration rate, ancestral and derived effective population sizes. Based on 578,557 sites sequenced in at least 4 individuals per species. We used the median substitution rate from our original *Cyprinodon* time-calibrated phylogeny of 5.37e^{-7} per site per year and a generation time of 0.75 years.

	ML estimate	**95% confidence interval**
ancestral N_e	401.3	369.9 – 429.4
mionectes and *amargosae* divergence time (years)	209.9	59.8 – 363.4
amargosae N_e / ancestral N_e	0.28	0.22 – 0.36
mionectes N_e / ancestral N_e	0.90	0.70 – 1.15
migration rate (per generation per year)	9.28x10^{-4}	5.32 x10^{-5} – 1.46 x 10^{-3}

Table S4. D-statistics testing for introgression between *diabolis* and neighboring Death Valley and Ash Meadows pupfishes. *C. salinus* was used as an outgroup in all tests. Two-tailed *P*-values are reported for each *z*-score. Populations showing significant introgression with *diabolis* are bolded. Note that statistical tests are not independent of each other, but indicate the strength of support for introgression, or deviations from a tree-like model of population branching, across various four-taxon subsets.

four-taxon tree	**D-statistic**	**ABBA sites**	**BABA sites**	**z-score**	**P-value**
(a, **b**) , (**diabolis**, salinus) (**a**, b) , (**diabolis**, salinus)	+ **-**				
n. nevadensis, ***n. amargosae***	0.17 ± .05	148	104	3.39	0.0007
n. nevadensis, ***n. pectoralis***	0.15 ± .05	148	110	3.11	0.002
n. amargosae*, n. mionectes*	-0.11 ± .05	121	151	-2.18	0.029
n. nevadensis, n. shoshone	0.09 ± .06	105	88	1.50	0.133
n. amargosae, n. shoshone	-0.07 ± .06	70	80	-1.03	0.303
n. nevadensis, n. mionectes	0.04 ± .05	133	122	0.84	0.399
n. amargosae, n. pectoralis	-0.00 ± .05	128	129	-0.07	0.941

Table S5. Summary of STRUCTURE runs and statistics used for calculating Evanno's Delta K [49].

k	reps	mean LnP(K)	stdev LnP(K)	Ln'(K)	|Ln''(K)|	ΔK
1	2	-71698.00	4.95	—	—	—
2	8	-64118.51	500.29	7579.49	868.16	1.74
3	8	-55670.86	171.68	8447.65	6310.13	36.75
4	8	-53533.34	221.07	2137.53	1159.91	5.25
5	8	-50235.90	331.33	3297.44	1632.25	4.93
6	6	-48570.72	1254.40	1665.18	1582.83	1.26
7	5	-48488.36	1161.24	82.36	379.41	0.33
8	8	-48785.41	1189.98	-297.05	772.27	0.65
9	2	-48310.20	74.95	475.21	1528.31	20.39
10	2	-49363.30	1627.62	-1053.10	—	—

Appendix S1. Species, location, source, collection date, and sample size per population for all individuals genotyped at more than 5% of total loci and used for analyses (out of the total number of individuals sequenced). Location numbers in parentheses from [1]. "*EW*" refers to extinct-in-the-wild species which were sequenced from captive colonies.

species	***n*** *(>5%)*	***n*** **total**	**location**	**source**	**date**
Cyprinodon diabolis	3	13	Devils Hole	Bailey Gaines	2008-2012
	1	3	School Spring Refuge	Anthony Echelle	1989
	2	3	Point-of-Rocks Refuge	LS	2013
	0	3	Mandalay Bay Refuge	LS	2013
Cyprinodon nevadensis amargosae	7	8	Amargosa R. (5)	BJT	1994
	0	3	Amargosa R., Tecopa (7)	BJT	1994
	0	3	Tecopa Spring Rd. (8)	BJT	1994
	0	3	China Ranch (6)	BJT	1994
Cyprinodon nevadensis mionectes	1	6	Big Spring	BJT	1994
	8	10	Point-of-Rocks	BJT	1994
Cyprinodon nevadensis nevadensis	7	9	Saratoga Spring	BJT	1994
Cyprinodon nevadensis pectoralis	7	9	Indian Spring	Anthony Echelle	1989
	1	5	School Spring	BJT	1994
Cyprinodon nevadensis shoshone	1	3	Amargosa R., Shoshone (11)	BJT	1994
	5	7	Shoshone Head Spring (10)	BJT	1994
Cyprinodon salinus salinus	2	4	Salt Creek	BJT	1994
Cyprinodon salinus milleri	1	3	Cottonball Marsh	BJT	1994
Cyprinodon eremus	0	2	Quitobaquito Spring, Arizona		1994
Cyprinodon fontinalis (EW)	0	1	Apache Spring	John Brill	captive
(EW)	0	2	Carbonera Spring	Al Morales	captive
Cyprinodon macularius	0	2	Coachella	BJT	1994
Cyprinodon radiosus	0	2	White Mountain	BJT	1994
Cyprinodon artifrons	1		Cancun, Mexico	Al Morales	2011
Cyprinodon labiosus	1		Laguna Chichancanab, Mexico	Al Morales	2011
Cyprinodon beltrani	1		Laguna Chichancanab, Mexico	Al Morales	2011

Cyprinodon verecundus (EW)	1	Laguna Chichancanab, Mexico	Michael Schneider	captive
Cyprinodon maya (EW)	1	Laguna Chichancanab, Mexico	Rhiannon West	captive
Cyprinodon variegatus	1	San Salvador Island, Bahamas	CHM	2008
Cyprinodon nichollsi	1	Laguna Oviedo, Dominican Republic	CHM	2011
Cyprinodon alvarezi (EW)	1	El Potosi, Mexico	Ryan Grisso	captive
Cyprinodon veronicae (EW)	1	Charco Palma, Mexico	Arcadio Valdes Gonzalez	captive
Cyprinodon albivelis	1	Rio Yaqui, Mexico	BJT	1994

Fig. S1

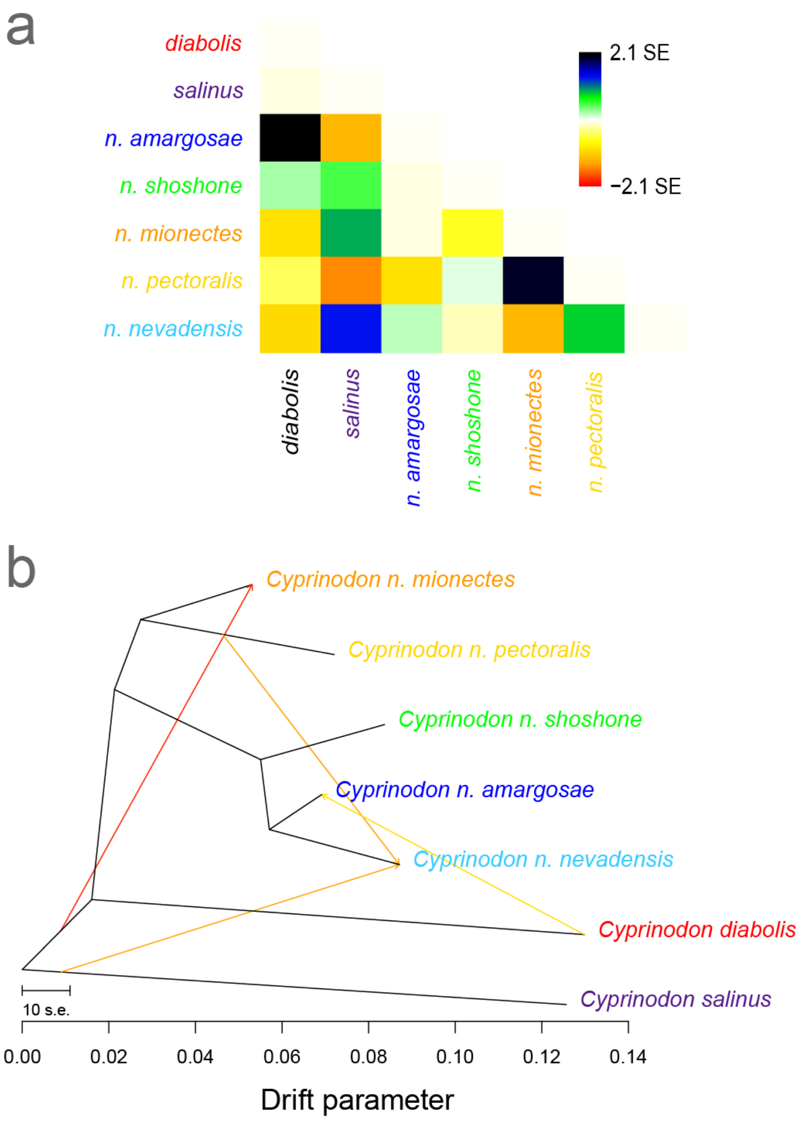

Fig. S1. *a)* Genetic variance-covariance structure of allele frequencies among Death Valley pupfish populations. *b)* Treemix graph with four migration events depicting major gene flow among Death Valley pupfish populations. Note the recent gene flow from *diabolis* into *amargosae*, consistent with Table S2 and Fig. 2c. Species colored as in Figs. 1-2. Heat color of migration lines indicate strength of admixture.

Fig. S2

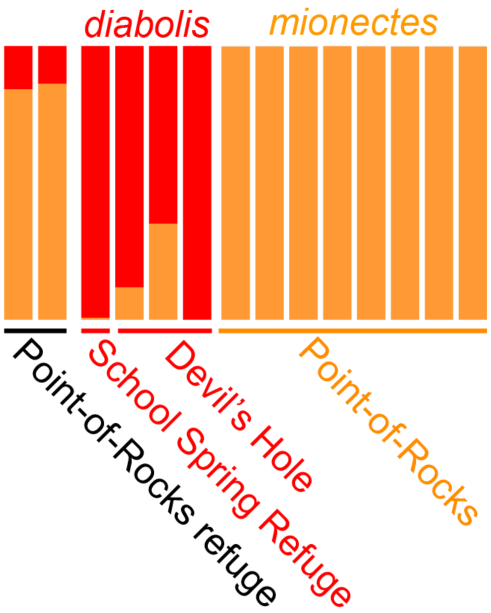

Fig. S2 Bayesian clustering analyses using STRUCTURE with *k* = 2 groups indicating the proportion of shared ancestry among *diabolis, mionectes,* and the Point-of-Rocks *diabolis* refuge population currently housed at the Ash Meadows Fish Conservation Facility. The Point-of-Rocks *diabolis* refuge population contained substantial shared ancestry with *mionectes* after less than 11 years [50].

Fig. S3

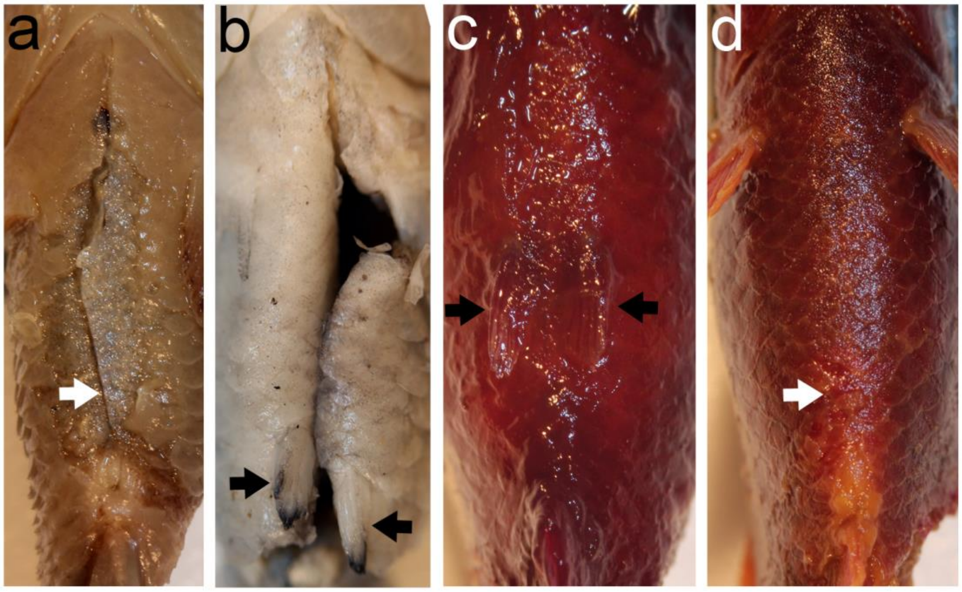

Fig. S3 Representative photographs in ventral view showing presence (black arrows) or absence (white arrow) of pelvic fins in *a)* wild *C. diabolis* (32º C), *b)* wild *C. nevadensis mionectes* (28-29ºC), *c-d) diabolis x mionectes* hybrids reared over five generations at 28-29ºC (alizarin-stained). Laboratory-rearing experiments indicate that pelvic fin loss in *diabolis* has a genetic basis. First, 100% of wild-collected *diabolis* eggs raised in the lab at 28-29ºC lacked pelvic fins (O. Feuerbacher pers. comm.), whereas 25% of *pectoralis* and 10.5% of *mionectes* found at similar temperatures in the wild lacked pelvic fins (*n* = 47; B. Turner unpublished data). Second, pelvic fin loss continues to segregate over several generations within a laboratory-reared *diabolis* x *mionectes* hybrid population (*c*-*d*).

Fig. S4

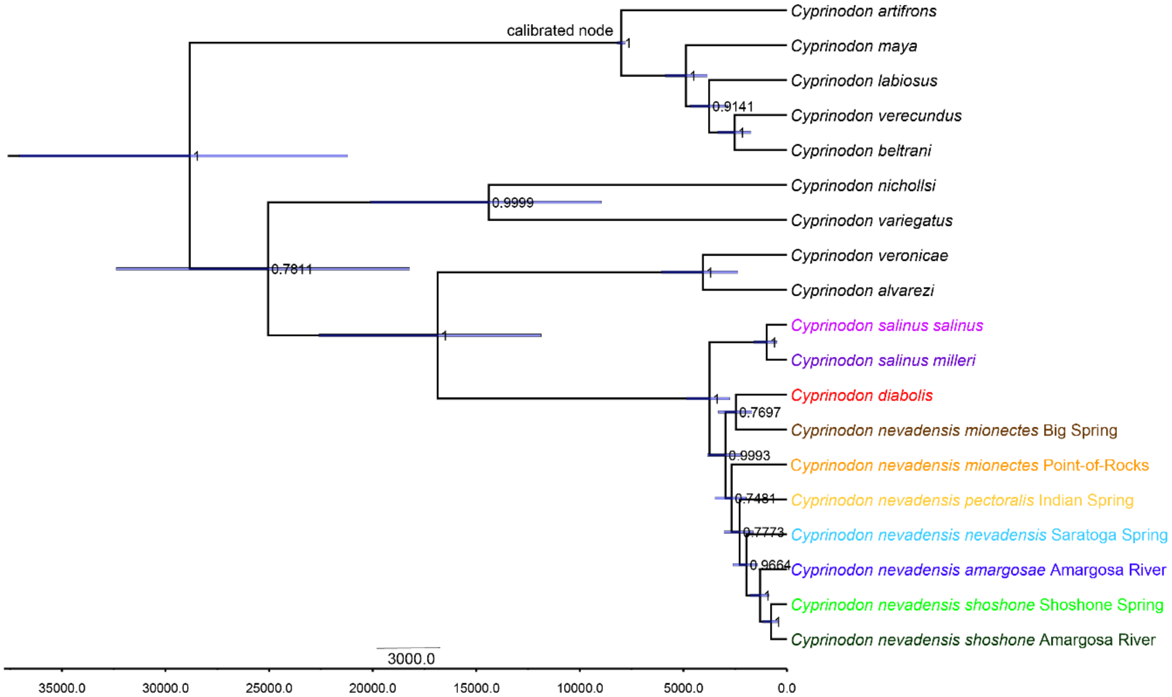
calibrated node
Cyprinodon artifrons
Cyprinodon maya
Cyprinodon labiosus
Cyprinodon verecundus
Cyprinodon beltrani
Cyprinodon nichollsi
Cyprinodon variegatus
Cyprinodon veronicae
Cyprinodon alvarezi
Cyprinodon salinus salinus
Cyprinodon salinus milleri
Cyprinodon diabolis
Cyprinodon nevadensis mionectes Big Spring
Cyprinodon nevadensis mionectes Point-of-Rocks
Cyprinodon nevadensis pectoralis Indian Spring
Cyprinodon nevadensis nevadensis Saratoga Spring
Cyprinodon nevadensis amargosae Amargosa River
Cyprinodon nevadensis shoshone Shoshone Spring
Cyprinodon nevadensis shoshone Amargosa River
1
0.9141
0.9999
0.7811
0.7697
0.9993
0.7481
0.7773
0.9664
3000.0
35000.0
30000.0
25000.0
20000.0
15000.0
10000.0
5000.0
0.0

Fig. S4 Time-calibrated maximum clade credibility tree for the Death Valley populations plus outgroup taxa across *Cyprinodon* estimated from the trimmed dataset of 4,159 concatenated 53-bp loci present in at least half of all taxa. Trees were estimated using BEAST under a coalescent model with GTR + Γ nucleotide substitution rates as described for Fig. 2. Two ‘rogue’ taxa with minimal support were trimmed for this analysis (*C. albivelis* and *C. nevadensis pectoralis* School Spring). Posterior probability of each node is indicated. Blue bars indicate 95% credible intervals for the estimated age of each node.

Fig. S5

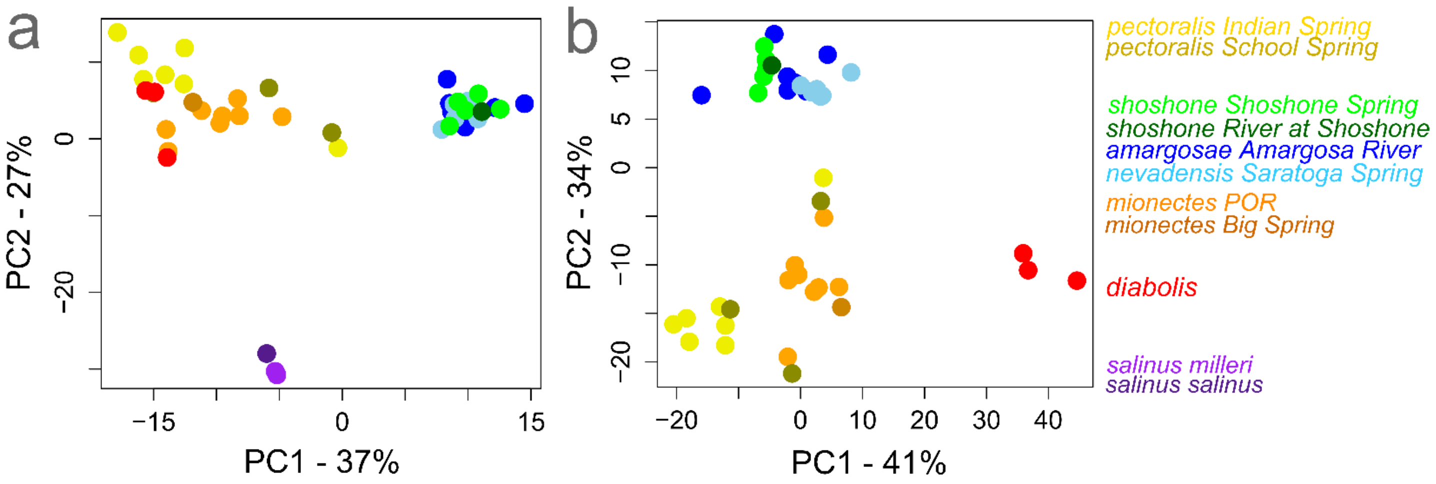

Fig. S5 First two principal components of genetic variance for 1,051 SNPs on 3,484 loci from the trimmed 53-bp dataset showing *a)* three main clusters of Death Valley populations as in Fig. 2. *b)* Excluding the distant *salinus salinus* and *salinus milleri* populations reveals four distinct genetic clusters. SNPs were filtered to one per locus to reduce the effects of linkage disequilibrium.