Failure to differentiate between divergence of species and their genes can result in over-estimation of mutation rates in recently diverged species.

İsmail K. Sağlam[1,2,*], Jason Baumsteiger[1,3] and Michael R. Miller[1,3,*]

1. Genetic Diversity Research Group, Department of Animal Science, University of California, One Shields Avenue, Davis, CA, 95616, U.S.A.
2. Ecological Sciences Research Laboratories, Department of Biology, Hacettepe University, Beytepe, 06800, Ankara, Turkey.
3. Center for Watershed Sciences, University of California, One Shields Avenue, Davis, CA, 95616, U.S.A.

* Corresponding authors: iksaglam@ucdavis.edu and micmiller@ucdavis.edu

Devils Hole pupfish (*Cyprinodon diabolis*, DHP) is an icon of population persistence due to having survived in complete isolation in a small pool in the Mojave Desert (Devils Hole, southwestern USA) for thousands of years [1]. Although the exact time and mode of colonization is unclear, DHP are generally assumed to have been isolated for 10 – 40 ka [2]. However, in a recent paper, Martin *et al.* [3] use an analysis with over 13,000 genomic loci to seriously challenge this notion. Based on demographic modeling with a genomic substitution rate of 5.37×10^{-7} mutations per site per year (m/s/y) (obtained from a phylogenetic analysis of Cyprinodontidae), they estimated the age of DHP to be between 0.105 – 0.830 ka and argue evolutionary timescales in DHP and other pupfish species in the region have been overestimated [3].

The recent divergence of DHP in Martin *et al.* [3] can be linked to the extremely high genomic mutation rate used in their demographic analysis. Although some Teleost fish genomes are believed to

evolve slightly faster than the typical vertebrate rate of 1×10^{-8} m/s/g [4,5], there is no precedent for a rate of $1.79\times10^{-7} – 5.37\times10^{-7}$ m/s/g, depending on the assumed average generation interval. Martin *et al.* [3] defend this rate because it was estimated using the only well-defined internal calibration event known for Cyprinodon: the 8,000 ± 200 year age of the Laguna Chichancanab [6]. This date was used to put a lower bound on the divergence between species within Laguna Chichancanab and the coastal *C. artifrons,* likely conspecific with the ancestral species [3]. Although this date is well supported, Martin *et al.* [3] fail to account for the bias associated with determining mutation rates when conflating species divergence times with time since the most recent common ancestor (TMRCA) of sampled genes.

The phylogenetic tree presented in Martin *et al* [3] reconstructs relationships using a single haplotype (16,567 concatenated 100-bp RAD-loci) per population, and therefore represents a haplotype (gene) tree and not a species tree. Although Martin *et al.* [3] acknowledge the potential of their phylogeny to return biased results due to incomplete lineage sorting (which could produce topological incongruences), they nevertheless do not distinguish between species and gene divergence times when calibrating the tree, which could cause temporal incongruences and have severe consequences for mutation rate estimates.

Figure 1a illustrates the distinction between a species divergence time (t) and a gene divergence time (T: TMRCA of haplotypes sampled from two descendant species). Population genetic theory predicts the TMRCA of a random pair of haplotypes in the ancestral population will be, on average, $2N_e$ generations (where N_e is the coalescent effective size of the ancestral population). Hence, gene divergence (T) will be $2N_e$ generations greater than species divergence (t) on average [7]. Since the phylogeny in Martin *et al.* [3] represents a haplotype tree, the calibration node depicting the divergence between coastal and inland species reflects gene divergence and not species divergence. Thus, the appropriate calibration for this node is not t (i.e. 8,000) but T ($8,000 + 2N_e$*g, g=generation interval in

49 years). This difference is often ignored when species have been separated for long periods of time as
50 the difference between T and t is proportionally less [7]. For example, if species divergence occurred
51 10 million years ago and ancestral N_e was 50,000, the difference between gene and species divergence
52 would be minor (10.1 million vs. 10 million, when g = 1). However, when divergence times between
53 species are small, the difference between t and T can be substantial, especially if ancestral N_e's are
54 large.

55 The ancestral species in Martin *et al.* [3], *C. artifrons*, is broadly-distributed throughout
56 coastal/brackish areas around the Yucatan Peninsula (figure 1b). Although its exact coalescent N_e is
57 unknown, similar species have coalescent N_e on the order of tens to hundreds of thousands [8]. If we
58 conservatively assume the ancestral population had a coalescent N_e of 50,000, expected T would be
59 108,000 (assuming g = 1), even though t is only 8,000. Thus, unless coalescent N_e of ancestral *C.*
60 *artifrons* was very small, the calibration information used by Martin *et al.* [3] represents a substantial
61 underestimation of T and hence a substantial overestimation of mutation rates.

62 A scenario where the ancestral population could have a low coalescent N_e would be if *C. artifrons*
63 exhibits substantial population structure. In this case, specific locations along the coast (small brackish
64 inlets for example) would contain a relatively small number of individuals isolated from other areas.
65 However, in structured species, the TMRCA between random haplotype copies sampled from two
66 distinct populations will depend on how closely related the populations are, and as the number of
67 populations increases, so will the variance in coalescent times [7] (figure 1c). Under this scenario,
68 phylogenetic analyses must use *C. artifrons* samples that originated from the correct ancestral
69 population because the branch length describing divergence between a random *C. artifrons* haplotype
70 and a random inland haplotype will equal T ($8{,}000 + 2N_e*g$) only if haplotypes of *C. artifrons* came
71 from the true ancestral population that colonized Laguna Chichancanab (C_7, figure 1c). If the sampled
72 haplotype belonged to a distantly related population (C_9 or C_{10}, figure 1c), the branch length would be

greater than T (8,000 + $2N_e$*g), resulting in a severe and unpredictable overestimation of mutation rates. Unfortunately, since the route of colonization and original location of founding individuals is unknown, it is difficult to determine the true level of error caused by calibrating T at 8,000 in Martin *et al.* [3]. To do this would require information from multiple coastal *C. artifrons* individuals to determine overall phylogenetic structure and the true sister population of the inland species group.

The issues we raise here are hardly new, as the errors of calibrating gene trees using species level information was discussed as recently as McCormack *et al* [9]. Put simply, without adherence to proper population genetic principles (detailed above), the extremely high genomic mutation rate estimated by Martin *et al.* [3] is likely a severe overestimate of the true mutation rate. Therefore, when this mutation rate was used in demographic modeling, the resulting divergence time is likely a serious underestimate. Given this issue, a reasonable assumption is that the mutation rate of DHP is similar to typical vertebrates (1×10^{-8} m/s/g). In fact, the difference between the typical vertebrate mutation rate and the one estimated by Martin *et al.* [3] (approximately 20-50x) is consistent with the difference between species and gene divergence expected from a short divergence time and large ancestral N_e. If Martin *et al.* [3] would have used the typical vertebrate mutation rate in their demographic analysis, their divergence estimate would have been at least an order of magnitude greater than presented in their paper. Therefore, the estimate by Martin *et al.* [3] that DHP is as young as 0.105 – 0.830 ka and the conclusion of "a surprisingly rapid timescale for speciation, genetic assimilation and the evolution of intrinsic reproductive incompatibilities in this group" should be taken with utmost caution.

References

1. Reed, J. M. & Stockwell, C. A. 2014 Evaluating an icon of population persistence: the Devil's Hole pupfish. *Proc. Biol. Sci.* **281**, 20141648. (doi:10.1098/rspb.2014.1648)

2. Echelle, A. A. 2008 The western North American pupfish clade (Cyprinodontidae: *Cyprinodon*): Mitochondrial DNA divergence and drainage history. *Geol. Soc. Am. Spec. Pap.* **439**, 27–38. (doi:10.1130/2008.2439(02).)

3. Martin, C. H., Crawford, J. E., Turner, B. J. & Simons, L. H. 2016 Diabolical survival in Death Valley: recent pupfish colonization, gene flow and genetic assimilation in the smallest species range on earth. *Proc. Biol. Sci.* **283,** 20152334 (doi: 10.1098/rspb.2015.2334)

4. Lynch, M. 2010 Evolution of the mutation rate. *Trends Genet.* **26**, 345–352. (doi:10.1016/j.tig.2010.05.003)

5. 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)

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

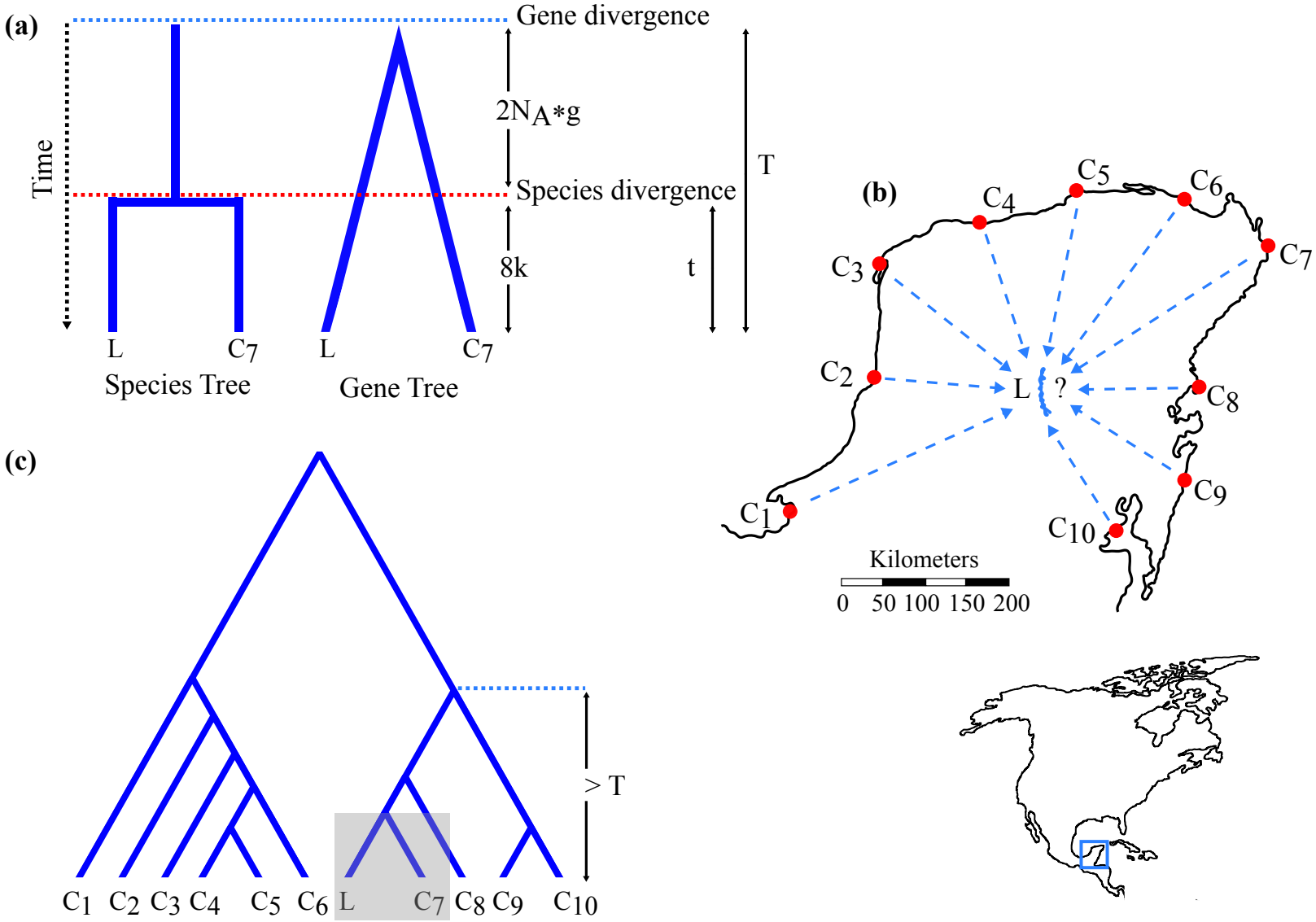
7. Edwards, S. V & Beerli, P. 2000 Perspective: gene divergence, population divergence, and the variance in coalescence time in phylogeographic studies. *Evolution* **54**, 1839–1854. (doi:10.1111/j.0014-3820.2000.tb01231.x)

8. Duvernell, D. D., Lindmeier, J. B., Faust, K. E. & Whitehead, A. 2008 Relative influences of historical and contemporary forces shaping the distribution of genetic variation in the Atlantic killifish, Fundulus heteroclitus. *Mol. Ecol.* **17**, 1344–1360. (doi:10.1111/j.1365-294X.2007.03648.x)

9. McCormack, J. E., Heled, J., Delaney, K. S., Peterson, A. T. & Knowles, L. L. 2011 Calibrating Divergence Times on Species Trees Versus Gene Trees: Implications for Speciation History of Aphelocoma Jays. *Evolution (N. Y).* 65, 184–202. (doi:10.1111/j.1558-5646.2010.01097.x)

Figure Legends

Figure 1. Differences in divergence times of Yucatan Peninsula pupfish and their genes. **(a)** Schematic diagram representing the difference between gene (triangular trees) and species/population divergence times (rectangular trees) for hypothetical coastal (C) and lake (L) populations. **(b)** Hypothetical locations of *C. artifrons* populations around the Yucatan Peninsula. The ancestral origin of Laguna Chichancanab taxa are unknown. **(c)** Hypothetical scenario where C_7 is the true ancestral population but C_9 or C_{10} were sampled for phylogenetic analysis. Laguna Chichancanab is shown for representative purposes and not drawn to scale.


(a)
Gene divergence
Time
2N_A*g
Species divergence
8k
T
t
L
C7
L
C7
Species Tree
Gene Tree
(b)
C5
C6
C4
C7
C3
C2
L
?
C8
C9
C1
C10
Kilometers
0 50 100 150 200
(c)
> T
C1 C2 C3 C4 C5 C6 L C7 C8 C9 C10