The complex effects of demographic history on the estimation of substitution rate: concatenated gene analysis results in no more than twofold overestimation

Christopher H. Martin¹⁺, Sebastian Höhna²³⁴⁺, Jacob E. Crawford², Bruce J. Turner³, Emilie J. Richards¹ and Lee H. Simons⁶

¹Department of Biology, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
²Department of Integrative Biology, and ³Department of Statistics, University of California, Berkeley, CA, USA
⁴Division of Evolutionary Biology, Ludwig-Maximilian-Universität, München, Germany
⁵Department of Biology, Virginia Tech, VA, USA
⁶Retired. U.S. Fish and Wildlife Service, 7123 Grousel Street, Las Vegas, Nevada 89131, USA

Our recent estimation of the divergence time and isolation of Death Valley pupfishes, including the iconic Devil’s Hole pupfish (DHP), rewrote widespread assumptions about this group. These species were previously assumed to be relic populations isolated over millions of years; our genomic analyses indicated recent colonization of Devil’s Hole within the past 105–830 years and frequent gene flow among Death Valley populations [1]. These results understandably attracted substantial attention given the iconic battle for conservation and intense management of DHP [2]; nonetheless, a young age for this species should not diminish its conservation value. Indeed, we argue that the unique natural history of this species makes it a prime candidate for exhibiting one of the fastest mutation rates observed in any vertebrate [3].

Sağlam et al. [4] argue that we overestimated the substitution rate, the rate at which mutations occur and fix between lineages over time, in pupfishes due to our analysis of a concatenated dataset of RADseq loci and therefore underestimated the age of DHP. Specifically, Sağlam et al. argue that a multi-species coalescent analysis would remove our bias and provide strikingly different results. Here, we test this assumption by reanalysing our original RADseq dataset under a multi-species coalescent model and comparing the estimated substitution rates to an analysis of concatenated RADseq loci. It is well known that divergence times estimated from concatenated, multi-gene analyses can overestimate species divergence times due to older gene-tree divergence compared with the true species-tree divergence time, a fact that we both cited [5] and demonstrated in our study: our concatenated gene analysis estimated DHP divergence at 2500–6500 years (fig. 2 of [1]) versus our coalescent analysis using dadi [6] at 105–830 years (fig. 3 and table S2 of [1]). However, the effect of concatenation on substitution rate estimation, as opposed to divergence time, is a new area of inquiry (contra Sağlam et al.’s claims). This is supported by the fact that two recent papers introducing this topic and coining a new term for a similar phenomenon (SPILS: substitutions produced by incomplete lineage sorting) [7,8] were published after our study was published in 2016. To the best of our knowledge, no study has yet estimated the bias due to concatenation of loci into a supermatrix on the estimation of substitution rates.

Here, we estimate that gene concatenation results in no more than twofold overestimation of substitution rates in our dataset using 80 subsets of 100 loci each from our original dataset (figure 1r; see data and scripts: doi:10.5061/dryad.9727v [9]). For each dataset, we compared substitution rates under concatenation and a multi-species coalescent approach using the flexible REVBayes software [10].
We also estimated substitution rates in 10 simulated datasets using either concatenation or a multi-species coalescent approach (figure 1 b,d; see the electronic supplementary material). Each simulated dataset contained 100 loci of 100 bp each simulated under a multi-species coalescent model with an effective population size ($N_e$) of 1000 constant across all branches of the tree and a substitution rate of $2.5 \times 10^{-7}$ (i.e. close to our estimate for the Laguna Chichancanab species). We observed the same minimal overestimation of substitution rate that never exceeded twofold in each simulated dataset and the expected increase in substitution rate variance under the multi-species coalescent (figure 1 b,d). We conclude that our original estimates of DHP divergence time may be revised to no more than 210–1660 years (95% confidence interval across a range of models). These results still rewrite the history of this species as extremely recent and do not challenge any of the conclusions in our original study. Nonetheless, we caution that these inferences are based on a biased reduced-representation sample of the Cyprinodon genome [12] and limited population sampling. For example, excluding loci with missing data disproportionately removes loci with higher mutation rates, resulting in a bias towards underestimation of substitution rate [13].

In contrast to the simplified framing of Sağlam et al. [4], the effect of concatenation on substitution rates is complex and depends on the demographic history of the group under consideration. Both concatenated gene and multi-species coalescent analyses assume a strict bifurcating model of population divergence with no gene flow [5,10,11]. By contrast, we found evidence for substantial secondary gene flow among species in Death Valley, including DHP [1], and have found

![Figure 1.](http://rsb.royalsocietypublishing.org/)

Figure 1. (a) Empirical clock rate estimates compared between concatenated supermatrix and multi-species coalescent approaches from 80 subsets of 100 ddRADseq loci each from [1], pruned to the Laguna Chichancanab species and outgroup coastal *C. artifrons*. (b,d) Substitution rate simulations compared between concatenated supermatrix and multi-species coalescent approaches in 10 datasets simulated on the fixed species tree with $N_e = 1000$, estimated from the empirical dataset, and a Jukes–Cantor model of nucleotide substitution. (c) Distribution of root ages in the simulated gene trees ($n = 1000$) across 10 independent runs.

---

**Figure 1. (a)** Empirical clock rate estimates compared between concatenated supermatrix and multi-species coalescent approaches from 80 subsets of 100 ddRADseq loci each from [1], pruned to the Laguna Chichancanab species and outgroup coastal *C. artifrons*. **(b,d)** Substitution rate simulations compared between concatenated supermatrix and multi-species coalescent approaches in 10 datasets simulated on the fixed species tree with $N_e = 1000$, estimated from the empirical dataset, and a Jukes–Cantor model of nucleotide substitution. **(c)** Distribution of root ages in the simulated gene trees ($n = 1000$) across 10 independent runs.
similar patterns in Caribbean pupfishes [14] and other fish groups [15]. Colonization of Laguna Chichancanab after its initial formation or secondary gene flow from coastal populations would result in underestimation, not overestimation, of substitution rates in concatenated gene and multi-species coalescent phylogenetic analyses [16].

Sagal et al. [4] use a simplified point calculation to quantify the expected discrepancy in divergence times between concatenated gene and multi-species coalescent analyses. However, their point calculation ignores the enormous amount of variance present in this estimate which has subsequently been addressed by 17 years of investigation in this field [5, 6, 10, 11]. Although we do not advocate for this approach, phylogenetic analyses of concatenated supertrees are still a common method for estimation of substitution rates in new taxa [7, 8].

Furthermore, Sagal et al. [4] assume an unrealistic N_{e} of 50,000 in C. artifrons based on a single study of microsatellite variation within a much more widely distributed species; reported N_{e} ranged from 3500 to 30,000 based on an assumed microsatellite mutation rate which is unknown for this species [17]. Using our multi-species coalescent analysis, we estimated ancestral N_{e} in the coastal C. artifrons population to be 2000 individuals, similar to estimates for other Caribbean pupfishes which underwent a recent population bottleneck [18, 19]. If we use Sagal et al.’s point estimate, this would result in a calibration of 12,000 years and consequently a 0.5 slower substitution rate and median divergence time of 383 years for DHP. This result does not challenge the major conclusions in our original study.

Finally, Sagal et al. [4] argue that ‘the typical vertebrate mutation rate’ of 1 x 10^{-8} mutations/site/generation is appropriate for DHP. This is the rate for humans [20, 21], not teleosts [22], and ignores the negative scaling of mutation rate with N_{e} over more than an order of magnitude demonstrated in the reference Sagal et al. cite for their claim [20]. Population genetic principles predict that smaller populations are expected to exhibit higher mutation rates due to the weaker strength of purifying selection relative to genetic drift in removing alleles that increase the mutation rate [20]. Indeed, DHP inhabits the smallest species range known, with the smallest N_{e} ever observed in a natural population [1], and may reasonably exhibit one of the fastest mutation rates observed in a vertebrate. Moreover, the unique life-history characteristics of the DHP—small size, short generation time, high metabolism, high temperatures and severe environmental stressors, such as starvation—are each associated with increased mutation rates [23–25]. These characteristics, which are often shared across Cyprinodon species, suggest that we should expect higher mutation rates in pupfishes, and DHP in particular, than the typical teleost [3].

In conclusion, we have demonstrated that our original concatenated gene analysis may have overestimated substitution rate, and subsequently underestimated the age of DHP, by a maximum of no more than twofold. Substitution rate may also have been underestimated due to widespread gene flow and stringent filtering. These results do not in any way challenge our original major conclusions of a rapid time scale for speciation, genetic assimilation and the evolution of intrinsic reproductive incompatibilities in pupfishes [1]. Furthermore, nearly all unique life-history characteristics of DHP suggest that its mutation rate may indeed be unprecedented among vertebrates. We consider this line of inquiry to deserve greater attention, rather than dismissal.

Data accessibility. All data and scripts are available from Dryad Digital Repository (doi:10.5061/dryad.97z7v).

Competing interests. We declare we have no competing interests.

Funding. We received no funding for this study.

References


cast doubt on one of the clearest examples of sympatric speciation. Evolution 69, 1406 – 1422. (doi:10.1111/evo.12674)


