BMC Ecology and Evolution

RESEARCH

Open Access

Base-substitution mutation rate across the nuclear genome of *Alpheus* snapping shrimp and the timing of isolation by the Isthmus of Panama

Katherine Silliman[1,2*], Jane L. Indorf[3], Nancy Knowlton[4], William E. Browne[3] and Carla Hurt[3,5]

Abstract

Background: The formation of the Isthmus of Panama and final closure of the Central American Seaway (CAS) provides an independent calibration point for examining the rate of DNA substitutions. This vicariant event has been widely used to estimate the substitution rate across mitochondrial genomes and to date evolutionary events in other taxonomic groups. Nuclear sequence data is increasingly being used to complement mitochondrial datasets for phylogenetic and evolutionary investigations; these studies would benefit from information regarding the rate and pattern of DNA substitutions derived from the nuclear genome.

Results: To estimate the genome-wide neutral mutation rate (μ), genotype-by-sequencing (GBS) datasets were generated for three transisthmian species pairs in *Alpheus* snapping shrimp. A range of bioinformatic filtering parameters were evaluated in order to minimize potential bias in mutation rate estimates that may result from SNP filtering. Using a Bayesian coalescent approach (G-PhoCS) applied to 44,960 GBS loci, we estimated μ to be 2.64E−9 substitutions/site/year, when calibrated with the closure of the CAS at 3 Ma. Post-divergence gene flow was detected in one species pair. Failure to account for this post-split migration inflates our substitution rate estimates, emphasizing the importance of demographic methods that can accommodate gene flow.

Conclusions: Results from our study, both parameter estimates and bioinformatic explorations, have broad-ranging implications for phylogeographic studies in other non-model taxa using reduced representation datasets. Our best estimate of μ that accounts for coalescent and demographic processes is remarkably similar to experimentally derived mutation rates in model arthropod systems. These results contradicted recent suggestions that the closure of the Isthmus was completed much earlier (around 10 Ma), as mutation rates based on an early calibration resulted in uncharacteristically low genomic mutation rates. Also, stricter filtering parameters resulted in biased datasets that generated lower mutation rate estimates and influenced demographic parameters, serving as a cautionary tale for the adherence to conservative bioinformatic strategies when generating reduced-representation datasets at the species level. To our knowledge this is the first use of transisthmian species pairs to calibrate the rate of molecular evolution from GBS data.

Keywords: *Alpheus*, Mutation rate, Isthmus of Panama, Genotype-by-sequencing, Molecular evolution

*Correspondence: kes0132@auburn.edu
[1] School of Fisheries, Aquaculture, and Aquatic Sciences, Auburn University, Auburn, AL 36849, USA
Full list of author information is available at the end of the article

Introduction

The rate of DNA substitution is an essential parameter in evolutionary biology because it is used to establish a timeline for the history of life. In the field of phylogeography, molecular clocks have been applied extensively, as they enable investigators to put absolute values on measures of interest such as timing of speciation, patterns of historical migration, and estimates of effective population sizes [3, 8, 12, 55, 70]. Estimates of DNA substitution rates can be calibrated using experimental approaches [6], or by associating molecular phylogenies with independent information regarding the timing of species divergence [32]. The fossil record has been the most widely used source for calibrating rates of molecular evolution, however, in groups that lack a good fossil record, well-dated biogeographic barriers can be used for establishing the timing of species divergence [33].

The final closure of the Central American Seaway (CAS) and formation of the Isthmus of Panama provides a useful calibration point for examining the rates and patterns of molecular evolution, as the completion of the Isthmus of Panama created a nearly impenetrable barrier to gene flow for thousands of marine taxa. The splitting of multiple independent populations is particularly useful for molecular clock calibrations because it provides both absolute rates of divergence, and critical information regarding the constancy of molecular evolution rates across independent evolutionary lineages [29]. By far the most cited transisthmian-based molecular clock calibrations come from the snapping shrimp genus *Alpheus* [35, 43, 47]. *Alpheus* contains more transisthmian species pairs than any other genus studied to date, providing a naturally replicated data set ideal for testing evolutionary hypotheses. However, comparisons of genetic distance estimates at mitochondrial genes across this genus have been shown to vary more than fourfold [43], which could be due to irregularity of the molecular clock or non-simultaneous divergence of transisthmian sister species. The latter explanation has been supported by multiple lines of evidence, including concordance of mitochondrial divergences with patterns of mating incompatibility and estimates of divergence from protein electrophoresis [44]. Previously, Hurt et al. [35] used a Bayesian coalescent approach to test simultaneous divergence of eight alpheid sister species using population-level sampling of multi-locus nuclear and mitochondrial genes. This work identified five transisthmian species pairs for which molecular data was consistent with recent and simultaneous divergence as a result of the closure of the CAS; these taxa are thus particularly well-suited for examining patterns of molecular divergence across the Isthmus of Panama.

The formation of the Isthmus is one of the most well-studied biogeographic vicariant events. Studies based on Foraminifera, isotope ratios, molecular phylogeography, and fossils suggest completion of the Isthmus had occurred by 3.5–2.7 Ma [11, 37, 40, 47, 52, 56, 65], although there has been some recent debate about this conclusion. For example, some geological work has suggested an earlier seaway blockage, where the primary closure of the CAS occurred before 10 Ma, with only minor connections after that time via narrow, transient, shallow channels [38, 53, 54, 67]. Bacon et al. [4, 5] supported this earlier formation date using molecular and fossil data to determine that an initial land bridge was present 23–25 Ma and formation of the Isthmus occurred between 10 and 6 Ma. However, the assumptions and methods underlying these studies have been challenged, in particular, the inappropriate application of a universal rate of mitochondrial DNA divergence across clades and failure to account for ancestral lineage sorting [46, 50, 56] Phylogeographic analyses that incorporate multiple loci can provide more robust estimates of divergence times and inform understanding of the timing of the closure of the CAS [5].

The vast majority of transisthmian molecular clock calibrations have been applied to nucleotide sequence data from mitochondrial genes [45, 47]. However, improvements in DNA sequencing technology and increased awareness of the limitations of single locus mitochondrial data sets have transformed the fields of population genetics and phylogeography [24]. Reduced representation sequencing techniques, such as genotyping-by-sequencing (GBS), systematically target a subset of the genome by relying on restriction enzymes and shared cut sites. These techniques have proven to be useful and cost-efficient methods for screening polymorphisms at thousands of loci without the need for a reference genome [2, 21]. However, little is known about the rate and variance of nuclear DNA substitutions across GBS loci. Estimates of demographic parameters (e.g., effective population size, migration rates) from GBS data would benefit from a GBS-derived mutation rate (μ), as μ is often required to calculate absolute parameter estimates [22].

Inherent characteristics of GBS methodologies, including bioinformatic processing and the often large, non-random proportion of missing data, have the potential to bias demographic analyses [19, 58]. Because these methods utilize restriction enzymes to reduce the genome, they require conservation of enzyme recognition cut sites to recover shared data among individuals. This typically results in a non-uniform distribution of reads across loci and individuals, and thus a reduced set of loci shared across all individuals [19]. The proportion of shared restriction sites (and sequenced loci) across sampled

individuals is expected to decline as divergence times increase. Conserved regions of the genome, possibly regions under purifying selection, will be disproportionately represented when filtering criteria are strict, while faster evolving, neutral regions may be filtered out. This pattern has important implications for optimizing filtering parameters in bioinformatics pipelines. Many GBS/RAD papers have taken a conservative approach and employed strict filters for missing data [10]. However, in silico and empirical work have begun to show that a "total evidence approach" including loci with missing data is acceptable and may even be preferable in phylogenetic and population genetic studies [19, 34, 68, 72]. Empirical investigations examining the influence of filtering criteria on estimates of demographic parameters would be useful for optimizing bioinformatic pipelines.

Here we report results from a comparative genomic study utilizing *Alpheus* species pairs to examine patterns of molecular divergence across the nuclear genome. Reduced representation GBS datasets were generated for three transisthmian species pairs: *A. malleator*/*A. wonkimi*, *A. formosus*/*A. panamensis*, and *A. colombiensis*/*A. estuariensis.* First, we investigated the phylogenetic signal of shared GBS-derived sequence tags in order to identify potential sequence bias due to divergence of restriction sites. The optimized GBS dataset was then used to estimate the timing of divergence (τ) for the selected species pairs using a Bayesian coalescent modelling approach while correcting for variance in divergence times. We then estimated the rate of base-substitutions (μ) using the final closure of the CAS as a calibration point. In order to evaluate claims of an earlier closure of the CAS, both 3 Ma and 10 Ma were used as calibration points for calculating μ, and the results were then compared to estimates of substitution rates in other multicellular eukaryotes. To our knowledge, this is the first use of transisthmian species pairs to calibrate the rate of molecular evolution across the nuclear genome.

Methods

Sample collections

Three transisthmian *Alpheus* sister species pairs (six species total) were selected for GBS sequencing and analysis: the eastern Pacific/western Atlantic pairs *A. malleator*/*A. wonkimi*, *A. colombiensis*/*A. estuariensis*, and *A. panamensis*/ *A. formosus* (Table 1). Previous work suggested that divergence times for these taxa were contemporaneous and likely to have resulted from the final closure of the Isthmus [35]. All shrimp were collected from the Caribbean and Pacific coasts of Panama. *Alpheus panamensis* and *A. formosus* were collected from intertidal or subtidal habitats, *A. malleator* and *A. wonkimi* from exposed shores, burrowed inside crevices in hard substrate, and *A. colombiensis* and *A. estuariensis* were collected from mudflats near mangroves. All samples were frozen in liquid nitrogen and stored at − 80 °C. DNA sequences from the mitochondrial gene cytochrome oxidase I (COI) were generated for all included individuals and compared to previously recorded COI sequences from the corresponding species. Primers and PCR conditions for amplification of COI followed Hurt et al. [36].

Molecular methods

Genomic DNA was extracted from 24 individuals using the DNeasy tissue kit (Qiagen Inc., Valencia, California), and samples were treated with RNase following the manufacturer's protocol. Three to four replicate GBS libraries per individual were optimized, generated, and sequenced at the Cornell University Biotechnology Resource Center Genomic Diversity Facility following the protocol of

Table 1 Collection information for *Alpheus* samples used for GBS sequencing including sample size (N), distribution (EP = Eastern Pacific, WA = Western Atlantic), known habitat, and GenBank Accession numbers

Species	N	Distribution	Habitat	COI Genbank Accessions
A. panamensis	5	EP	Intertidal under rocks and coral rubble	MZ229847, MZ229848, MZ229849, MZ229850, MZ229851
A. formosus	5	WA	Intertidal under rocks and coral rubble	EF532605, MZ229852, MZ229853, MZ229854, MZ229855
A. colombiensis	4	EP	Burrows in mangroves	FJ013882, FJ013883, MZ229841, MZ229842
A. estuariensis	4	WA	Burrows in mangroves	MZ229843, MZ229844, MZ229845, MZ229846
A. wonkimi[a]	4	EP	Endolithic in rock crevices	MZ229837, MZ229838, MZ229839, MZ229840
A. malleator	2	WA	Endolithic in rock crevices	AF309912, MZ229836

[a] *Alpheus wonkimi* was referred to as *A. cf malleator* or *A. isthmalleator* in Hurt et al. [35]

[21], resulting in a total of 96 samples. Briefly, genomic DNA was digested with EcoT22I (A|TGCAT) and barcoded adapters were ligated onto resulting restriction fragments. Pooled libraries were sequenced on a single Illumina HiSeq 2000/2500 lane, obtaining 100 base pair, single-end sequencing reads. Sequence reads from replicate libraries were combined for downstream analyses.

Quality filtering, locus assembly, and genotyping

Raw sequencing reads were de-multiplexed, quality filtered, and de novo clustered using pyRAD v.3.0.2 [18], a pipeline optimized to produce aligned orthologous loci across distantly related taxa using restriction-site associated DNA. Demultiplexing used sample-specific barcode sequences, allowing one mismatch in the barcode sequence. Base calls with a Phred quality score under 20 were converted to Ns, and reads containing more than 4 Ns were discarded. Adapter sequences, barcodes, and the cut site sequences were trimmed from reads passing filter, with only reads greater than 50 bp retained. For within-sample clustering, a minimum coverage cutoff of 5 × was employed. Consensus sequences with more than eight heterozygous sites were discarded as potential paralogs. Clustered orthologs containing heterozygous sites that were shared by more than two samples were also discarded as putative paralogs. The same clustering threshold of 85% was used for both within- and across-sample clustering [18].

We generated 11 datasets that varied in included samples (10–24), the minimum number of samples (*m*) that had to be shared by each locus (3–6), and the minimum number of species (*s*) shared by each locus (1–3). In particular, one *A. malleator* individual had very few sequencing reads and therefore was excluded from some datasets. These additional datasets were used to investigate the impact of filtering by sample coverage and missing data on estimates of demographic parameters (Additional file 1: Table S1). The primary dataset, *Am4s2*, used in the following analyses includes all samples (*A*), with at least four individuals (*m4*) and two species (*s2*) genotyped at each locus. Transition/transversion (Ts/Tv) ratios were calculated for all datasets using VCFtools [13].

Estimating homology to coding regions

In order to identify putative protein-coding loci that may be subject to selection, a comprehensive dataset including all individuals and loci genotyped in at least 4 individuals (*Am4*) was blasted against Metazoa sequences in both the NCBI remote BLAST nucleotide database (*nt*) and protein database (*nr*) in April 2016, using the programs blastx and blastn in BLAST + 2.3.0 (ftp://ftp.ncbi.nih.gov/blast/executables/LATEST/). Analyses used default settings and an "*E*-value" significance threshold of 1. These results were imported into Blast2Go, where InterProScan was used to identify putative protein coding sequences. Loci with significant matches to mRNA sequences, proteins, or matches to an InterPro database were classified conservatively as putative nonneutral loci. A local BLAST database was created with these loci and used to identify nonneutral loci in the other datasets, including *Am4s2*. Loci with InterPro matches or BLAST results with an "*E*-value" less than 1E-3 were blasted again in August 2020 to be annotated with updated Gene Ontology (GO) terms in Blast2Go [26] (Additional file 1: Figure S1).

Phylogenetic analyses

For phylogenetic reconstructions, a concatenated matrix was produced for the *Am4s2* dataset and partitioned into putative neutral and coding sites, with model parameters estimated independently for each partition. Maximum Likelihood inferences were conducted using RAxML v8.1 [69] under the General Time-Reversible nucleotide model with gamma-distributed rate heterogeneity (GTRGAMMA) and 1000 bootstrap replicates. Bayesian inferences were performed using Exabayes v1.4 (Exelixis Lab, http://sco.hits.org/exelixis/web/software/exabayes/). Four independent MCMC runs were run for 1,000,000 generations, sampling every 500 generations. Runs were initiated from a random order addition parsimony tree. All other settings were default. Pairwise branch length distance between each species was calculated using the *ape* v3.5 package in R [61].

Locus bias

To assess patterns of locus-sharing among individuals, the R package RADami [30] was used to construct a locus presence-absence (LPA) matrix and display the proportion of shared loci between pairs of individuals (Fig. 3a). Pairwise Jaccard's distances calculated from this matrix were visualized using nonmetric multidimensional scaling in the R package vegan [57]. The dimensionality of the ordination was determined by performing 50 replicate runs at random starting configurations for K = 1 to 10 axes, with K = 1 to 3 showing the largest decreases in final stress. Both the K = 2 and K = 3 ordinations were rerun with 2000 replicates each, and converged on best solutions. Only the K = 3 ordinations are reported in this study, as they provided the clearest visualization of sample clustering. The poorly sequenced *A. malleator* individual was excluded from ordinations because low overlap in locus coverage between this individual and all others dominated the ordinations in preliminary analyses (not shown).

To understand the influence of phylogenetic distance and sequencing depth on locus bias, we built linear

models of pairwise shared loci across all samples using the *lm* function in base R. First, we constructed a 24×24 matrix of the number of pairwise shared loci between all samples using the pyRAD .loci file, based on Python and R code from [75]. We then used a linear model to predict the number of shared loci between two samples based only on the combined number of reads after quality filtering. This model was compared to a linear model where filtered reads and phylogenetic distance between samples were the independent variables. Phylogenetic distances were determined from the RAxML maximum likelihood tree using the R package *ape* v3.5 [63].

Demography, gene flow, and mutation rates

Estimates of demographic parameters, post-divergence gene flow, and genome-wide mutation rate were performed using the Generalized Phylogenetic Coalescent Sampler (G-PhoCS) version 1.2.3 [27], which infers divergence times (τ), ancestral effective population sizes (Θ), and migration rates. A Markov Chain Monte Carlo (MCMC) sampling strategy was used to sample parameters from a full coalescent isolation-with-migration model, where post-divergence migration bands are optional and specified by the user. This model assumes a separate constant population size for each branch of the phylogeny, and a separate constant migration rate for any migration bands specified. All demographic parameters are scaled by the mutation rate (μ), which can either be held constant or allowed to vary across loci. G-PhoCS takes as input a given phylogeny, specified directional migration bands, and a collection of aligned neutrally evolving loci, where heterozygous genotypes are unphased and the likelihood computation analytically sums over all possible phasings.

Due to the computationally intensive nature of this analysis, we were unable to analyze the full *Am4s2* neutral dataset (44,960 total loci). Thus the full dataset was randomly sampled to generate three reduced datasets with 14,986 randomly sampled loci in each. All analyses involving the *Am4s2* dataset were replicated on each of these three datasets, with the replicate runs combined to calculate the mean and confidence intervals of parameter estimates. Analyses were initially performed under the assumption of no gene flow after divergence, estimating 16 parameters (11 population sizes and 5 divergence times). Replicate analyses were conducted with mutation rate held constant, as well as an additional analysis with random locus-specific mutation rates estimated by G-PhoCS. As mutation rates are known to vary across genomes, we expected the latter analyses to provide a more accurate estimation of overall mutation rates. All MCMC runs were executed using the same settings, unless otherwise indicated. Each Markov chain included 100,000 burn-in iterations, after which parameter values were sampled every 10 iterations for 200,000 iterations. The prior distributions over model parameters were defined by a product of Gamma distributions (Additional file 1: Table S2). The fine-tune parameters of the MCMC procedure were set automatically during the first 10,000 burn-in iterations (using the 'find-finetunes TRUE' option in the G-PhoCS control file). We conditioned on the phylogenetic relationships of taxa based on the ML tree and phylogenetic inference in [74]. Convergence for each run was inspected manually in Tracer [64].

We conducted multiple G-PhoCS runs on ten datasets that varied in taxa composition and the minimum number of individuals/species recovered at a locus to explore the influence of filtering by sample coverage on parameter estimates. The purpose of this approach was to 1) determine how different stringency filters for missing data affected demographic estimates, and 2) ensure that demographic estimates were robust to the selection of species included in the analysis. Three of the datasets included representatives from all six species, and one dataset only included two species pairs (*PFECm3s2*). The other six datasets contained three species each—one transisthmian species pair and one outgroup species. These triplet datasets were classified into an *a* group and a *b* group for visualization in Fig. 4. In total, τ and Θ were estimated across 23 runs for *A. estuariensis*/*A. colombiensis* (τ_{EC}, Θ_{EC}) and *A. panamensis*/*A. formosus* (τ_{PF}, Θ_{PF}) and 21 runs for *A. wonkimi*/*A. malleator* (τ_{WM}, Θ_{WM}) (Additional file 1: Table S1).

We also conducted G-PhoCS analyses that allowed for migration between sister species in order to explicitly test for post-divergence gene flow and determine the effect of gene flow on estimates of population divergence times and effective population sizes. Three replicate runs on the *Am4s2* dataset included 6 directional 'migration bands' representing gene flow between each sister species pair, with the mutation rate (μ) allowed to vary across loci. Following [23], a migration band was inferred to have significant gene flow if the 95% Bayesian credible interval of the migration rate (*M*) did not include 1E-5 in any of the replicate runs. We then conducted three replicate G-PhoCS analyses incorporating the migration bands between the sister species pair that showed significant gene flow. The effective number of migrants per generation was calculated as $M_{A \rightarrow B} \times \Theta_B$.

We used the outputs of the three replicate G-PhoCS analyses with random locus-specific mutation rates that incorporated the migration bands between the one sister species pair that showed significant gene flow to obtain a best estimate of τ and Θ for each species pair. MCMC samples were combined from the posterior distributions for the replicate runs to determine the mean and

95% confidence interval estimates for the demographic parameters. The timing of species divergences (τ) estimated by G-PhoCS were calibrated with estimates of the final closure of the CAS to estimate the absolute rate of μ and its variation among sister species pairs. Previous work has shown that these three species pairs have contemporaneous divergence times (within 5 my); therefore, we divided the τ estimated by G-PhoCS for each species pair by a similar absolute divergence time to obtain estimates of μ. We estimated μ using both 3 Ma and 10 Ma as calibration points, as the final closure of the Isthmus has been proposed to occur within this time interval.

Results

Sequence assembly and gene ontology

Sequencing of 24 individuals across 96 libraries yielded 189,408,593 total raw sequencing reads (average of 7,350,671 ± 3,318,208 reads per sample). After quality filtering, replicate samples were combined to assemble consensus sequences for each individual, with a mean read depth of 14.33 (± 63.42). These were further filtered to 43,831 ± 25,944 consensus sequences per individual. Datasets differing by sample coverage or included taxa varied in the total number of loci (3,481–48,062), Ts/Tv ratios (1.359–1.594), and the amount of missing data (Additional file 1: Table S1). In neutral loci from the *Am4s2* dataset, the frequency of C/G substitutions was less relative to other substitutions (Fig. 1).

Of the 56,838 loci in the *Am4* dataset, 3925 loci (6.9%) were identified as non-neutral based on inferred homology to mRNA or protein coding sequences using BLAST tools and InterProScan. Whiteleg shrimp (*Litopenaeus vannamei*) had the most top hits in the *nr* database; 707 loci had a hit to InterProScan and 647 loci had a Blast E value of 0.001 to the *nr* or *nt* databases, of which 261 were annotated with GO terms (Additional file 1: Figure S1).

Locus bias and phylogenetics

Phylogenetic reconstructions resulted in monophyletic species and sister-species pairs with 100% bootstrap support, and *A. estuariensis* and *A. colombiensis* clustered separately from the other four species (Fig. 2). These topologies are consistent with previous phylogenetic work based on the mitochondrial gene cytochrome oxidase I (COI) and two nuclear genes [74]. Ordination of the pairwise shared-locus matrix showed a strong phylogenetic signal, with individuals from the same sister species pair clustering together (Fig. 3). This result is consistent across datasets varying in sample coverage (Additional file 1: Figure S2). Of the two linear models tested, our model that included phylogenetic distance and the log-root product of total number of reads passing filter performed the best [$r^2 = 0.872$, $p = 0$] (Additional file 1: Figure S3).

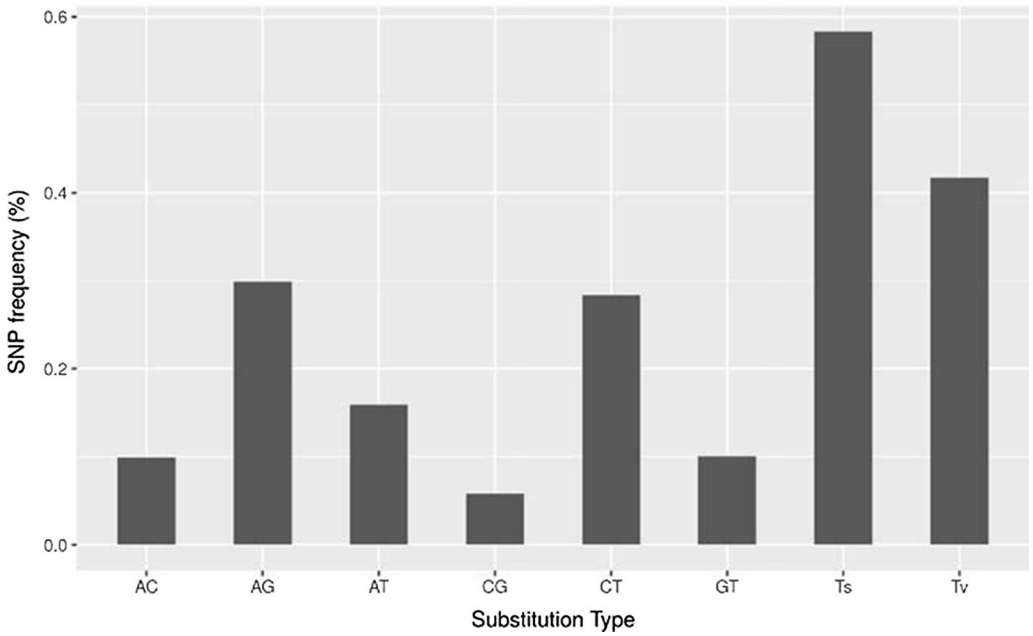

Fig. 1 Distribution of six nonpolarized substitution types based on 222,174 SNPs across 44,960 loci in the *Am4s2* dataset, after filtering for putative neutral loci

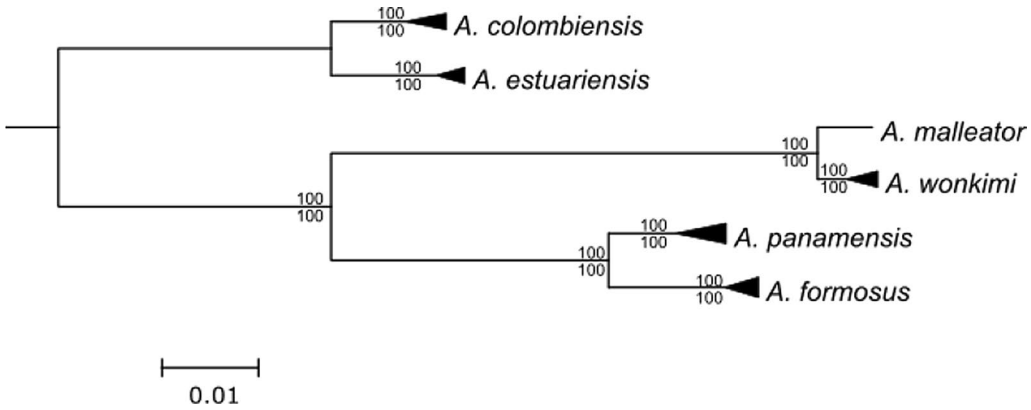

Fig. 2 Results from Bayesian phylogenetic reconstructions based on concatenated matrices from the full dataset (*23m6s2*). Nodes were collapsed by species. All species and transisthmian sister-species pairs were monophyletic with 100% posterior probability support (values above node). Maximum likelihood reconstructions using the same dataset resulted in an identical topology with 100% bootstrap support for every node (values below nodes)

Demographic inference and post-divergence gene flow

We conducted multiple G-PhoCS runs on datasets that varied in species composition and the minimum number of individuals/species recovered at a locus, as the number of loci shared between samples was shown to be influenced by phylogenetic distance. Replicate runs on the same dataset were nearly identical, indicating there were a sufficient number of simulations to achieve convergence. We found the greatest consistency in parameter estimation across runs on the *Am4s2*, *23m6s2*, and species triplet datasets (Fig. 4). Dataset *Am3s3*, which required at least three species to be sequenced at every locus and thus had the least missing data, produced significantly lower estimates for τ_{root}, τ_{PF}, and τ_{EC}. The results from *Am3s3* were similar to those from *23m6s2* for τ_{WM}, but both were significantly lower than the results from *Am4s2* and the species triplet datasets (Fig. 4).

The *Am4s2* dataset was chosen for downstream analysis as it represented all taxa and gave consistent results. We conducted three G-PhoCS run replicates on *Am4s2*, allowing the rate of mutation to vary randomly across loci. This represents a more realistic scenario than simply estimating a single mutation rate for all loci; however, it increases the computational time (by at least 10%). In the original G-PhoCS paper, the authors determine through simulations and empirical analyses that a "random rates" model can possibly influence the estimation of root population divergence and ancestral effective population size, but it likely has only a minor effect on divergence times of more recent branches in the phylogeny [27]. We found that allowing for mutation rate to vary across loci also widened the confidence intervals around our estimates of τ_{root}, as well as increasing estimates of τ at all population splits and decreasing estimates of Θ (Fig. 4).

Tests for post-divergence gene flow between transisthmian sister species suggested that, of the six migration bands tested, gene flow was only significant for the *A. panamensis* / *A. formosus* species pair. Migration rate estimates from G-PhoCS were significant across all replicate runs for this species pair. When G-PhoCS was run with variable mutation rates across loci and migration only allowed between *A. panamensis* and *A. formosus*, we found an expected number of 0.130 migrants per generation from *A. panamensis*→ *A. formosus* and an expected number of 0.044 migrants per generation from *A. formosus*→ *A. panamensis*. Tests for post-divergence migration were not consistently significant for the other two species pairs.

To obtain our best estimate of τ in order to calculate μ, we performed three replicate runs on the *Am4s2* dataset allowing for migration only between *A. formosus* and *A. panamensis* and variable rates of mutation across loci. *A. malleator*/*A. wonkimi* had the smallest estimated divergence time (τ_{WM} ~ 5.95E-3) and therefore likely diverged most recently, followed by *A. estuariensis* /*A. colombiensis* (τ_{EC} ~ 7.93E-3), and *A. panamensis*/*A. formosus* (τ_{PF} ~ 12.5E-3) (Table 2). As *A. malleator* only had sequence data for one individual at the majority of loci, we calculated μ from τ_{EC}. If we assume annual generation times and divergence at the proposed final closing of the Isthmus (3 Ma) we get a substitution rate of 2.64E-9 (2.53E-9—2.75E-9). If we use the more controversial closure calibration of 10 Ma, as is determined in [67], supported by [54], and cited in [4], we get a substitution rate of 7.93E-10 (7.6E-10—8.24E-10).

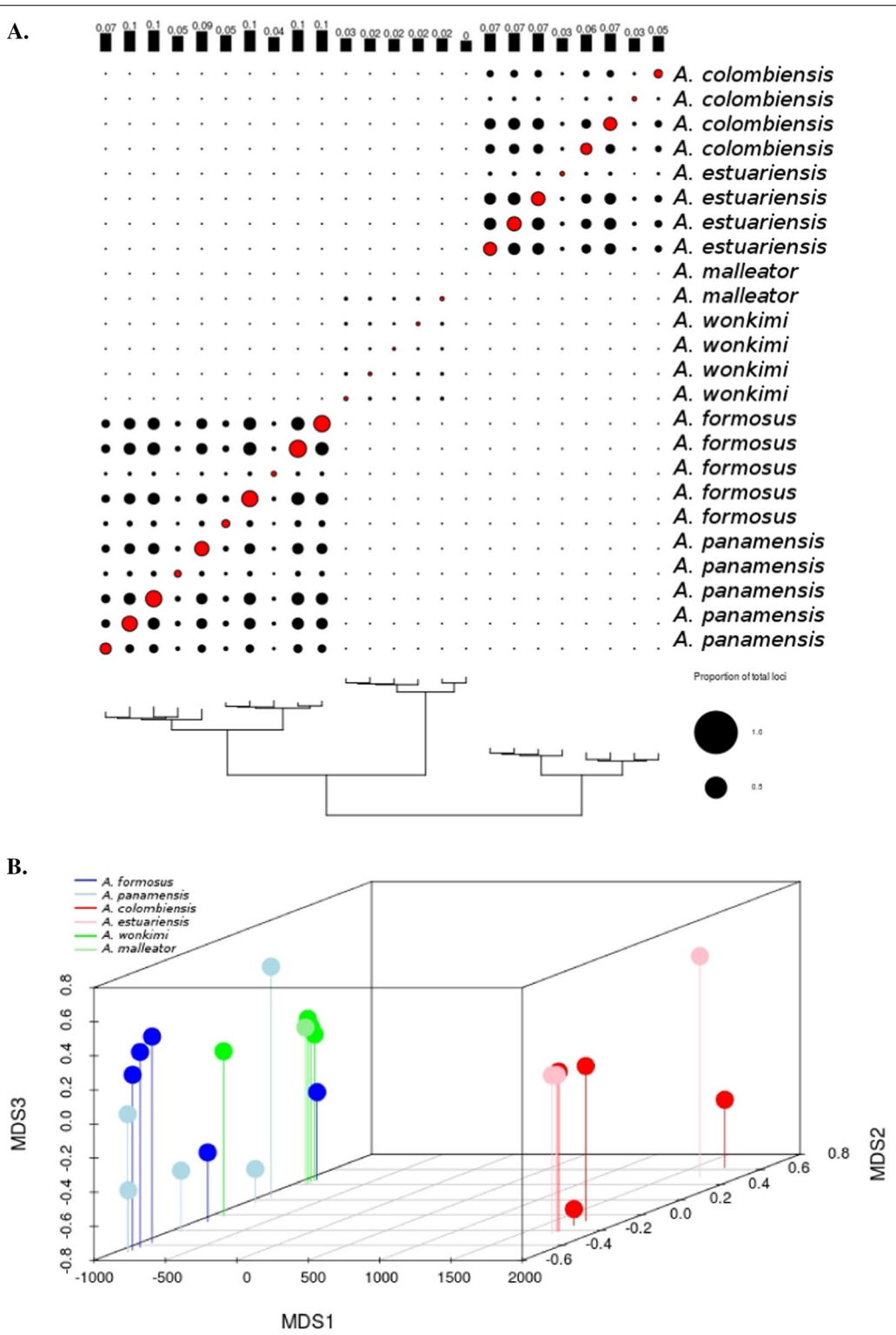

Fig. 3 **A** Proportion of loci shared among individuals in the *Am4s2* dataset. Loci shared between individuals (black circles) or successfully amplified within an individual (red circles) expressed as the proportion from 0 to 1 of all 48,062 loci scored in the *Am4s2* dataset. Plotted with the phylogenetic tree based on maximum likelihood inference, with branch lengths scaled in substitutions per nucleotide. **B** Ordination of *Alpheus* samples based on nonmetric multidimensional scaling of the locus presence-absence matrix, K = 3, colored by species

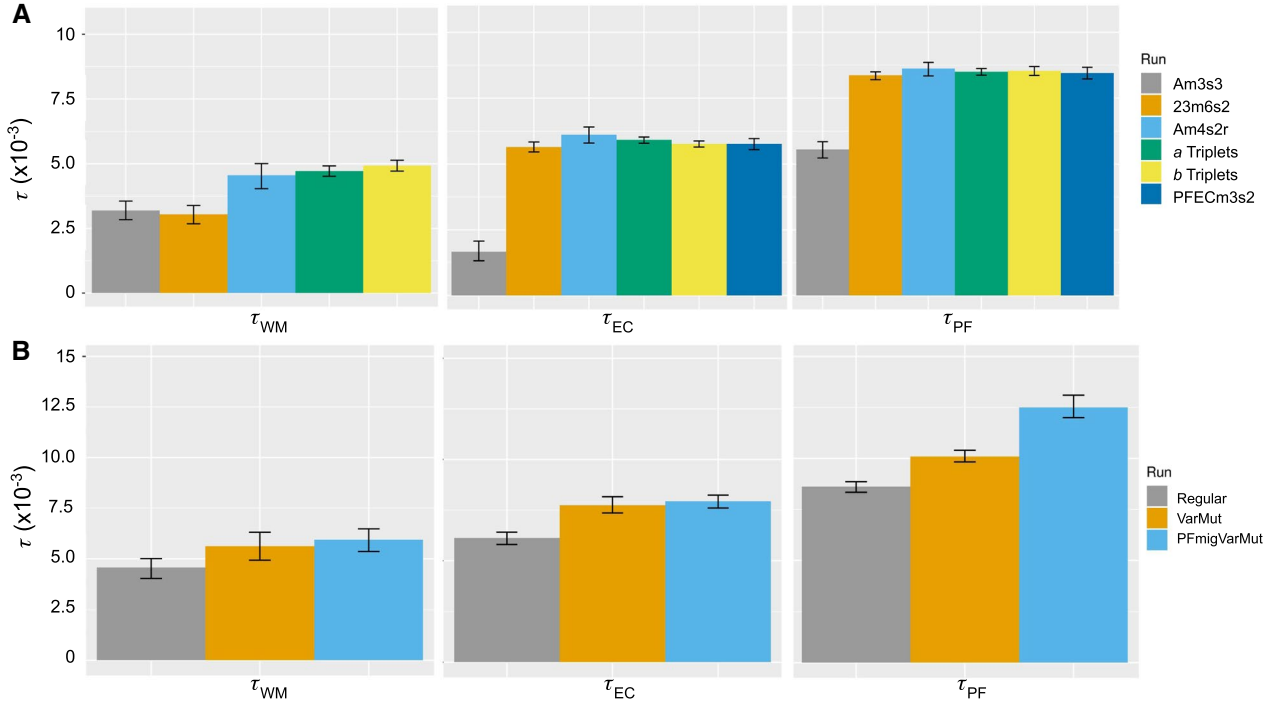

Fig. 4 Estimates of divergence times (τ) for three transisthmian *Alpheus* species pairs (*A. wonkimi/ A. malleator* (WM), *A. estuariensis/ A. colombiensis* (EC), *A. panamensis/ A. formosus* (PF)) across datasets, scaled by the substitution rate (μ). Error bars indicate 95% Bayesian credible intervals across all replicate runs. **A** Datasets that varied in species composition and the minimum number of individuals (*m*) or species (*s*) recovered at a locus, modeled with a constant mutation rate across loci and no migration. Parameter values for each dataset are listed in Additional file 1: Table 1. *a* and *b* triplet datasets include the focal species pair plus one outgroup species. **B** The *Am4s2r* dataset, modeled with either constant μ across loci and no migration (Regular), variable μ across loci (VarMut), or variable μ and migration allowed between *A. panemensis/A. formosus* (PFmigVarMut)

Table 2 Demographic inferences based on G-PhoCS including effective ancestral population size (Θ), absolute effective ancestral population size in number of individuals (N_e), divergence time (τ), absolute divergence time in millions of years (T), and calculated mutation rate (μ)

Species Pair	Θ	N_e (× 10^4) (μ = 2.64)	τ (E-3)	T (μ = 2.64)	μ using 3 Ma (E-9)	μ using 10 Ma (E-9)
A. wonkimi/A. malleator	0.018 (0.016,0.019)	167 (153,182)	5.95 (5.36,6.48)	2.25 (2.03,2.45)	1.98 (1.79,2.16)	0.59 (0.54,0.65)
A. colombiensis/A. estuariensis	0.027 (0.025,0.028)	252 (240, 264)	7.93 (7.60,8.24)	3 (2.88,3.12)	2.64 (2.53,2.75)	0.79 (0.76,0.82)
A. panamensis/A. formosus	0.0205 (0.019,0.022)	194 (184, 205)	12.5 (12.0,13.1)	4.73 (4.55,4.96)	4.17 (4.00, 4.37)	1.25 (1.20,1.31)

N_e and T for all pairs are calculated using μ = 2.64. Numbers in parentheses indicate 95% confidence intervals

Discussion

The field of evolutionary biology is rapidly transitioning from its reliance on a handful of mitochondrial loci to the incorporation of genome-wide sequence data for reconstructing evolutionary histories. Reduced representation methods for genome sampling, such as GBS, have seen widespread applications for genomic investigations involving non-model organisms [71, 73]. Interpretation of these genomic datasets will require an understanding of the rate of DNA substitution across the nuclear genome. Molecular clock calibrations utilizing the well-examined closure of the Isthmus of Panama are among the most widely used parameters for dating cladogenic events [4, 12, 43, 47]. The genus *Alpheus* includes more transisthmian sister species pairs than any other taxonomic group, facilitating the development of replicated datasets needed for robustly estimating mutation rate. Our GBS dataset included 2,844,991 bp from 44,960 neutral loci and represented three alpheid transisthmian species-pairs known to have diverged comparatively recently [35].

Collectively, results from our study can inform other studies utilizing reduced representation sequencing for evolutionary investigations. Below we outline the

implications of our results for 1) using such datasets to estimate mutation rates and reducing the role of locus bias in these analyses; and 2) understanding the evolutionary processes associated with the rise of the Isthmus of Panama, both the timing of divergence events and implications for the debate on when final closure occurred.

Mutation rate estimates

Accounting for coalescence within ancestral populations and post-split migration, our best estimate for the per site mutation rate (μ), was 2.64E-9 (2.53E-9–2.75E-9) substitutions/site/year using the more widely accepted estimated time of 3 Ma for closure of the Isthmus. A previous analysis of the nuclear mutation rate in transisthmian *Alpheus* pairs [35] yielded estimates an order of magnitude lower than most experimentally derived rates and those reported here. This earlier study used Sanger sequencing data from eight nuclear genes (4457 bp) and employed a coalescent-based method (MCMCcoal) to estimate an average per site mutation rate across loci of 2.3E-10 substitutions/year. As this estimate was based on sequence data exclusively from protein coding regions, the reduced substitution rate was likely the result of purifying selection. By excluding putative protein coding genes from our data set, our new estimate based on 2,844,991 bp, is more similar to experimental mutation rate estimates in other arthropods (e.g., 3.46E-9 in *Drosophila* [39], 3.8E-9 in *Daphnia* [41]. Our value for μ may still represent a slight underestimate, as some non-coding regions can also be under evolutionary constraints [77] . Furthermore, while four-fold degenerate sites are often used for estimating neutral mutation rates, without a reference genome we are unable to confidently identify these sites. Both MCMCcoal and the method used in our study, G-PhoCS, estimate ancestral population sizes and population divergence times by comparing and integrating across genealogies at multiple neutrally evolving loci. Importantly, G-PhoCS extends the MCMCcoal model by allowing for gene flow between diverged populations and facilitating the use of unphased genotype data. The latter is often necessary for GBS data when phasing is not possible.

The popularity of mitochondrial markers for phylogenetic and evolutionary studies is largely due to their higher mutation rate compared to nuclear markers. Our understanding of the ratio of substitution rates in the mitochondrial genome relative to the nuclear genome (μ_{mit}/μ_{nuc}) has largely been based on observations in vertebrate taxa [9]. Allio et al. [1] analyzed μ_{mit}/μ_{nuc} in 121 multilocus datasets covering 4,676 animal species and found that μ_{mit}/μ_{nuc} varies widely across taxonomic groups. In vertebrates, μ_{mit}/μ_{nuc} is typically above 10 and averages around 20. Invertebrates tend to have much lower μ_{mit}/μ_{nuc} values, ranging from 2 to 6. Across crustaceans, μ_{mit}/μ_{nuc} estimates range from 2.0 to 10.4 with an average of 5.9. We estimated μ_{mit}/μ_{nuc} using our genome-wide μ estimate and a μ_{mi} estimate of 1.1E-8 substitutions/site/year for the mitochondrial COI gene, derived from the most closely related transisthmian species pair (*A. colombiensis*/*A. estuariensis)*. Our estimate of μ_{mit}/μ_{nuc} was 4.3, a value very similar to the average value observed across other crustacean groups. Several factors have been used to explain the lower μ_{mit}/μ_{nuc} in arthropods including a lower mass-specific metabolic rate [49], taxonomic differences in the ratio of mtDNA/nuDNA replication cycles per generation [51], and a negative correlation between per generation mutation rates and effective population size [48].

SNPs identified in our GBS data can provide a rough approximation of DNA substitution types and the ratio of transitions (Ts) to transversions (Tv) in *Alpheus* (Fig. 1). However, unlike experimental assays of mutation rate, our analysis cannot determine the directionality of substitution types. The stringency for missing data and taxa inclusion influences our estimates of Ts/Tv slightly (Additional file 1: Table S1), an effect that has been observed in other reduced representation sequencing datasets [68]. Our primary dataset, *Am4s2*, had a Ts/Tv ratio of 1.399, while the more conservative *Am3s3* dataset had a Ts/Tv ratio of 1.489. These Ts/Tv ratios are comparable to the Ts/Tv ratio of 1.22 in *Drosophila* [62] but lower than the accepted Ts/Tv ratio of ~ 2.1 for humans [16, 20]. Variation in Ts/Tv ratios may be due to fundamental differences in point substitution processes [42] or an artifact of the sequencing approach [7, 14].

Locus bias in GBS data

Our analyses also examined the effects of molecular divergence on locus recovery from GBS data, providing empirical evidence for a phylogenetic signal in the distribution of missing data when reduced representation methods are applied to species-level investigations. This finding has important implications for establishing unbiased filtering parameters in bioinformatic pipelines, as data from GBS and restriction-site associated DNA (RAD) sequencing are now commonly used for demographic and phylogenomic analyses at both shallow and deep time scales [2]. Because these methods utilize restriction enzymes to reduce the genome, they require conservation of enzyme recognition cut sites to recover shared data among individuals. In theory, a mutation in a cut site would result in either allelic dropout at shallow time scales and bias population genetic inferences [25], or potentially the loss of phylogenetically informative loci at deeper timescales. Missing data can also arise

from low or uneven sequencing coverage across samples [19]. Simulations of GBS datasets at phylogenetic scales found that bioinformatic filtering to reduce missing data can select for loci with lower mutation rates that are more likely to be genotyped across taxa [34]. Determining the cause of missing data in a GBS dataset can help inform bioinformatic processing decisions, which can, in turn, influence downstream phylogenetic and population genetic inferences [17, 68].

Using linear models and ordination we demonstrated that the amount of missing data between samples was strongly influenced by phylogenetic relatedness and, to a lesser extent, sequencing depth. This result suggested that a strict missing data filter may impose a phylogenetically-informed bias on the retained data. It is likely that strict filtering parameters will preferentially retain phylogenetically conserved loci that are subject to purifying selection. We examined the proportion of loci shared across three or more species for putative protein coding and non-coding loci; the proportion of loci recovered from three or more species was more than 30% higher for coding than non-coding tags (7.1% and 5.4%, respectively). We also tested the influence of filtering parameters on demographic parameter estimates obtained from G-PhoCS; a total of 10 different datasets were applied that varied in the amount of missing data allowed across individuals and/or species. Our most conservative dataset, requiring a locus to be present in at least three species, produced significantly lower divergence time estimates for all three species pairs. This result suggests that a stringent missing data filter selects for loci that are more conserved across species and therefore have lower substitution rates, supporting simulation findings [34]. Strict filtering thresholds also impacted other demographic parameters, such as estimates of current and ancestral effective population size (Θ) (Additional file 1). Other studies that have used GBS or RAD datasets for interspecific demographic modelling often only include loci found in all study species [10, 60]. Our results highlight the risk of using strong filters for missing data with G-PhoCS and other demographic methods, and instead suggest demographic models using reduced representation methods should be tested with a range of missing data.

Divergence time estimates

The sequence of divergence times for *Alpheus* transisthmian species pairs based on our GBS dataset is largely consistent with earlier, coalescent based estimates based on sequence data from nuclear protein coding genes. Model-based estimates of τ using the software IMa [28] and MCMCcoal [76] on eight nuclear genes found that *A. panamensis*/*A. formosus* was the first species pair to be separated, followed by *A. malleator*/*A. wonkimi*, with the most recent split being between *Alpheus colombiensis*/ *A. estuariensis*; the 95% confidence intervals of τ for these latter two species pairs overlapped considerably [35]. In this study of the broader nuclear genome (Table 2), we again found *A. panamensis* and *A. formosus* diverging first at an estimated T = 4.73 Ma, followed by *A. colombiensis* and *A. estuariensis* (T = 3.0 Ma), with *A. wonkimi* and *A. malleator* the most recently diverged pair (T = 2.25 Ma). It is intuitive that these often intertidal species would be clustered and recent in their divergence times as they would have had more opportunities for dispersal during the final stages of the formation of the Isthmus than species inhabiting rocky intertidal habitats [56].

Mitochondrial loci evolve separately from nuclear genes, providing an independent dataset for examining divergence. K2P pairwise sequence distances in the mitochondrial COI barcoding gene have indicated a partially different order of divergence, showing *A. malleator*/*A. wonkimi* diverging first (K2P = 11.5%), with *A. panamensis*/*A. formosus* diverging next (K2P = 9.5%), and *A. colombiensis*/*A. estuariensis* as the last-diverging pair (K2P = 6.8%) [47]. Coalescent based divergence times account for polymorphisms within ancestral taxa which can influence split estimates while K2P distances do not, thus the difference in divergence order between these two approaches may reflect ancestral lineage sorting.

Mutation rates, vicariance patterns, and the timing of final closure of the Isthmus of Panama

Comparison of our estimate of μ to other established mutation rate estimates can provide insight for the ongoing debate surrounding the timing of the closure of the Isthmus of Panama. We compared our estimate for μ when using a calibration point of divergence at the more broadly supported estimate of 3 Ma [56] vs. the suggestion of a considerably older timing of 10 Ma [4, 54, 67]. We found that the latter resulted in an almost fivefold lower estimate of mutation rate than the rates found for *Drosophila*, *Daphnia*, and other multicellular eukaryotes [15, 39, 41, 59]. The estimate of μ_{mit} is also consistent with estimates from independently calibrated arthropod taxa when calibrated with a 3 Ma Isthmus. While it is possible that *Alpheus* have a considerably lower nuclear mutation rate than other studied taxa, it is unlikely that both nuclear and mitochondrial genomes would exhibit unusually low mutation rates as these processes occur independently [43].

Not all transisthmian species pairs reflect recent, clustered vicariant events (e.g., three of eight *Alpheus* pairs studied by Hurt et al. [35] and some other taxa reviewed

in O'Dea et al. [56]). However, the contemporaneous divergence time estimates of multiple transisthmian species pairs [35, 56] supports the utility of the formation of the Isthmus as a calibration point for evolutionary histories. Overdispersion of pairwise distance estimates in mitochondrial genes has been used to refute the established timeline for completion of the Isthmus [4]. However, findings from our GBS dataset that account for polymorphisms in ancestral populations and post-divergence gene flow suggest that divergence times for multiple species pairs occurred within a narrow window at about 3 Ma. Our results support accounting for accurate taxon sampling and coalescent processes in ancestral populations when examining transisthmian speciation events [50].

Once formed, the Isthmus of Panama represented an impenetrable barrier for shallow water marine species to migrate between the eastern Pacific and western Atlantic, but opportunities for migration may have fluctuated as closure neared final completion. Results from our G-PhoCS analyses largely support a complete isolation of eastern Pacific and western Atlantic *Alpheus* populations following the closure of the CAS. Of the six migration parameters estimated, significant post-divergence gene flow was only found for *A. panamensis*/*A. formosus,* the species pair with the oldest divergence time (approx. 4.73 Ma). Best estimates of migration rates in this pair were exceedingly low; while gene flow was bidirectional, greater migration was inferred from the eastern Pacific species towards the western Atlantic species; this is consistent with models showing that strong currents passed through the straits from the Pacific into the Caribbean leading up to its final closure [56, 66]. *Alpheus panamensis*/*A. formosus* are both common, free-living species that occupy a wide range of habitats, including under rocks and in rock crevices, in dead and living coral rubble, and in sand/mud mixed substrate. Tolerance to a diversity of habitats may have facilitated trans-oceanic passage during the final stages of Isthmus formation when sub-optimal habitats may have been encountered by migrants, and even today these species are occasionally capable of producing fertile hybrid clutches [44]. We found that the inclusion of post-divergence migration parameters was important for obtaining robust estimates of mutation rates. Failure to account for post-split migration results in a negative bias in estimates of divergence times and inflates estimates of genome-wide substitution rates. For example, the estimated divergence time for *A. panamensis*/*A. formosus* (τ_{PF}) was reduced by 19% when migration was not included in the model (Fig. 3).

Conclusion

Our results add an additional layer of support for a recent closure of the Panamanian Isthmus which has broad-ranging implications across evolutionary biology. The agreement between our estimate of the genomic mutation rate and experimentally derived mutation rates in multiple model organisms makes an older Isthmus highly unlikely. Though widely criticized [46, 50, 56], the suggestion that the closure of the Isthmus may have occurred as early as 23 million years ago has cast doubt on decades of studies across multiple disciplines that have relied on the widely accepted closure date of 3 million years [31]. The multi-locus approach employed here highlights the importance of accounting for ancestral lineage sorting when using geological events to calibrate molecular processes.

The nuclear mutation rate and evidence for phylogenetic signal in loci identified here can inform studies using reduced representation methods to address phylogeographic and demographic histories in non-model taxa. Our empirical-based estimate of the nuclear mutation rate aligns well with experimentally determined mutation rates in model arthropod species suggesting that these rates may be applied more broadly to studies in other taxonomic groups. Results from our exploration of filtering parameters serve as a cautionary tale for the adherence to strict bioinformatic filtering parameters. To our knowledge, this is the first use of transisthmian species pairs to calibrate the rate of molecular evolution from reduced-representation sequencing data.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12862-021-01836-3.

Additional file 1. Table S1. Description of different datasets used in the paper. **Table S2.** Configuration parameters for G-PhoCS. **Figure S1.** Distribution of gene ontology annotations from Blast2GO for putative protein-coding loci. **Figure S2.** Ordination plots of locus-sharing matrices for additional datasets. **Figure S3.** Quantitative modelling of bias in the number of shared loci between samples.

Acknowledgements
Not applicable.

Authors' contributions
All authors contributed to the design of the study. CH and NK collected tissue samples. JI performed the molecular lab work. KS and CH analyzed the data and drafted the manuscript. All authors read and approved the final manuscript.

Funding
This work was funded by a University of Miami Scientists and Engineers Expanding Diversity and Success (SEEDS) grant (NSF #0820128) and a University of Miami General Research Support Award. KS was funded by the National Science Foundation Graduate Research Fellowship under Grant No. 1545870

and the Department of Education Graduate Assistance in Areas of National Need Fellowship Grant No. P200A150101.

Availability of data and materials
Raw demultiplexed genotype-by-sequencing DNA sequences are available on NCBI SRA PRJNA729103. COI sequences are available through GenBank, with accession numbers listed in Table 1. Output files from pyRAD and input files for G-PhoCS are available on Dryad: https://doi.org/10.5061/dryad.qnk98sfgs. Scripts used for data analysis are available on the corresponding author's Github: https://github.com/ksil91/alpheus-gbs.

Declarations

Ethics approval and consent to participate
No ethical approval was required for any of the experimental research described here.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]School of Fisheries, Aquaculture, and Aquatic Sciences, Auburn University, Auburn, AL 36849, USA. [2]Committee on Evolutionary Biology, University of Chicago, Chicago, IL 60637, USA. [3]Department of Biology, University of Miami, Coral Gables, FL 33146, USA. [4]National Museum of Natural History, Smithsonian Institution, Washington, DC, USA. [5]Department of Biology, Tennessee Tech University, Cookeville, TN 38505, USA.

Received: 18 January 2021 Accepted: 6 April 2021
Published online: 28 May 2021

References

1. Allio R, Donega S, Galtier N, Nabholz B. Large variation in the ratio of mitochondrial to nuclear mutation rate across animals: implications for genetic diversity and the use of mitochondrial DNA as a molecular marker. Mol Biol Evol. 2017;34:2762–72.
2. Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA. Harnessing the power of RADseq for ecological and evolutionary genomics. Nat Rev Genet. 2016;17:81–92.
3. Arbogast BS, Edwards SV, Wakeley J, Beerli P, Slowinski JB. Estimating divergence times from molecular data on phylogenetic and population genetic timescales. Annu Rev Ecol Syst. 2002;33:707–40.
4. Bacon CD, Silvestro D, Jaramillo C, Smith BT, Chakrabarty P, Antonelli A. Biological evidence supports an early and complex emergence of the Isthmus of Panama. Proc Natl Acad Sci USA. 2015;112:6110–5.
5. Bacon CD, Silvestro D, Jaramillo C, Smith BT, Chakrabarty P, Antonelli A. Reply to Lessios and Marko et al.: early and progressive migration across the Isthmus of Panama is robust to missing data and biases. PNAS. 2015;112(43):E5767–8.
6. Baer CF, Miyamoto MM, Denver DR. Mutation rate variation in multicellular eukaryotes: causes and consequences. Nat Rev Genet. 2007;8:619–31.
7. Ba H, Jia B, Wang G, Yang Y, Kedem G, Li C. Genome-wide SNP discovery and analysis of genetic diversity in farmed Sika deer (*Cervus nippon*) in northeast China using double-digest restriction site-associated DNA sequencing. Science. 2017;7:3169–76.
8. Bromham L, Penny D. The modern molecular clock. Nat Rev Genet. 2003;4:216–24.
9. Brown WM, George M Jr, Wilson AC. Rapid evolution of animal mitochondrial DNA. Proc Natl Acad Sci USA. 1979;76:1967–71.
10. Campagna L, Gronau I, Silveira LF, Siepel A, Lovette IJ. Distinguishing noise from signal in patterns of genomic divergence in a highly polymorphic avian radiation. Mol Ecol. 2015;24:4238–51.
11. Coates AG, Stallard RF. How old is the Isthmus of Panama? Bull Mar Sci. 2013;89:801–13.
12. Cunningham CW, Collins TM. Developing model systems for molecular biogeography: vicariance and interchange in marine invertebrates. In: Molecular ecology and evolution: approaches and applications. Basel: Birkhäuser. 1994. Pp. 405–433.
13. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R, 1000 Genomes Project Analysis Group. The variant call format and VCFtools. Bioinformatics. 2011; 27:2156–8.
14. Davey JW, Cezard T, Fuentes-Utrilla P, Eland C, Gharbi K, Blaxter ML. Special features of RAD Sequencing data: implications for genotyping. Mol Ecol. 2013;22:3151–64.
15. Denver DR, Dolan PC, Wilhelm LJ, Sung W, Lucas-Lledó JI, Howe DK, Lewis SC, Okamoto K, Thomas WK, Lynch M, Baer CF. A genome-wide view of *Caenorhabditis elegans* base-substitution mutation processes. Proc Natl Acad Sci USA. 2009;106:16310–4.
16. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43:491–8.
17. Díaz-Arce N, Rodríguez-Ezpeleta N. Selecting RAD-Seq data analysis parameters for population genetics: the more the better? Front Genet. 2019;10:533.
18. Eaton DAR. PyRAD: assembly of de novo RADseq loci for phylogenetic analyses. Bioinformatics. 2014;30:1844–9.
19. Eaton DAR, Spriggs EL, Park B, Donoghue MJ. Misconceptions on missing data in RAD-seq phylogenetics with a deep-scale example from flowering plants. Syst Biol. 2017;66:399–412.
20. Ebersberger I, Metzler D, Schwarz C, Pääbo S. Genomewide comparison of DNA sequences between humans and chimpanzees. Am J Hum Genet. 2002;70:1490–7.
21. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE. 2011;6:e19379.
22. Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M. Robust demographic inference from genomic and SNP data. PLoS Genet. 2013;9:e1003905.
23. Freedman AH, Gronau I, Schweizer RM, Ortega-Del Vecchyo D, Han E, Silva PM, Galaverni M, Fan Z, Marx P, Lorente-Galdos B, Beale H, Ramirez O, Hormozdiari F, Alkan C, Vilà C, Squire K, Geffen E, Kusak J, Boyko AR, Parker HG, Lee C, Tadigotla V, Wilton A, Siepel A, Bustamante CD, Harkins TT, Nelson SF, Ostrander EA, Marques-Bonet T, Wayne RK, Novembre J. Genome sequencing highlights the dynamic early history of dogs. PLoS Genet. 2014;10:e1004016.
24. Garrick RC, Bonatelli IAS, Hyseni C, Morales A, Pelletier TA, Perez MF, Rice E, Satler JD, Symula RE, Thomé MTC, Carstens BC. The evolution of phylogeographic data sets. News and Views. 2015. https://doi.org/10.1111/mec.13108.
25. Gautier M, Gharbi K, Cezard T, Foucaud J, Kerdelhué C, Pudlo P, Cornuet J-M, Estoup A. The effect of RAD allele dropout on the estimation of genetic variation within and between populations. Mol Ecol. 2013;22:3165–78.
26. Götz S, García-Gómez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, Robles M, Talón M, Dopazo J, Conesa A. High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res. 2008;36:3420–35.
27. Gronau I, Hubisz MJ, Gulko B, Danko CG, Siepel A. Bayesian inference of ancient human demography from individual genome sequences. Nat Genet. 2011;43:1031–4.
28. Hey J, Nielsen R. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. Genetics. 2004;167:45.
29. Hickerson MJ, Carstens BC, Cavender-Bares J, Crandall KA, Graham CH, Johnson JB, Rissler L, Victoriano PF, Yoder AD. Phylogeography's past, present, and future: 10 years after Avise, 2000. Mol Phylogenet Evol. 2010;54:291–301.
30. Hipp AL, Eaton DAR, Cavender-Bares J, Fitzek E, Nipper R, Manos PS. A framework phylogeny of the American oak clade based on sequenced RAD data. PLoS ONE. 2014;9:e93975.
31. Hoorn C, Flantua S. An early start for the Panama land bridge. Geology. 2015;14:78.
32. Ho SYW, Duchêne S. Molecular-clock methods for estimating evolutionary rates and timescales. Mol Ecol. 2014;23:5947–65.

33. Ho SYW, Tong KJ, Foster CSP, Ritchie AM, Lo N, Crisp MD. Biogeographic calibrations for the molecular clock. Biol Lett. 2015;11:20150194.
34. Huang H, Knowles LL. Unforeseen consequences of excluding missing data from next-generation sequences: simulation study of RAD sequences. Syst Biol. 2016;65:357–65.
35. Hurt C, Anker A, Knowlton N. A multilocus test of simultaneous divergence across the Isthmus of Panama using snapping shrimp in the genus *Alpheus*. Evolution. 2009;63:514–30.
36. Hurt C, Silliman K, Anker A, Knowlton N. Ecological speciation in anemone-associated snapping shrimps (*Alpheus armatus* species complex). Mol Ecol. 2013;22:4532–48.
37. Jackson JBC, O'Dea A. Timing of the oceanographic and biological isolation of the Caribbean Sea from the tropical eastern Pacific Ocean. Bull Mar Sci. 2013;89:779–800.
38. Jaramillo C, Montes C, Cardona A, Silvestro D, Antonelli A, Bacon CD. Comment (1) on "Formation of the Isthmus of Panama" by O'Dea et al. 2017.
39. Keightley PD, Trivedi U, Thomson M, Oliver F, Kumar S, Blaxter ML. Analysis of the genome sequences of three *Drosophila melanogaster* spontaneous mutation accumulation lines. Genome Res. 2009;19:1195–201.
40. Keigwin L. Isotopic paleoceanography of the Caribbean and East pacific: Role of Panama uplift in late Neogene time. Science. 1982;217:350–3.
41. Keith N, Tucker AE, Jackson CE, Sung W, Lucas Lledó JI, Schrider DR, Schaack S, Dudycha JL, Ackerman M, Younge AJ, Shaw JR, Lynch M. High mutational rates of large-scale duplication and deletion in *Daphnia pulex*. Genome Res. 2016;26:60–9.
42. Keller I, Bensasson D, Nichols RA. Transition-transversion bias is not universal: a counter example from grasshopper pseudogenes. PLoS Genet. 2007;3:e22.
43. Knowlton N, Weigt LA. New dates and new rates for divergence across the Isthmus of Panama. Proc R Soc Lond. 1998;265:2257–63.
44. Knowlton N, Weigt LA, Solorzano LA, Mills DK, Bermingham E. Divergence in proteins, mitochondrial DNA, and reproductive compatibility across the Isthmus of Panama. Science. 1993;260:1629.
45. Lavinia PD, Kerr KCR, Tubaro PL, Hebert PDN, Lijtmaer DA. Calibrating the molecular clock beyond cytochrome b: assessing the evolutionary rate of COI in birds. J Avian Biol. 2016;47:84–91.
46. Lessios HA. Appearance of an early closure of the Isthmus of Panama is the product of biased inclusion of data in the metaanalysis. Proc Natl Acad Sci U S A. 2015;112:E5765.
47. Lessios HA. The Great American Schism: Divergence of Marine Organisms After the Rise of the Central American Isthmus. Annu Rev Ecol Evol Syst. 2008;39:63–91.
48. Lynch M. Evolution of the mutation rate. Trends Genet. 2010;26:345–52.
49. Makarieva AM, Gorshkov VG, Li B-L, Chown SL, Reich PB, Gavrilov VM. Mean mass-specific metabolic rates are strikingly similar across life's major domains: evidence for life's metabolic optimum. Proc Natl Acad Sci USA. 2008;105:16994–9.
50. Marko PB, Eytan RI, Knowlton N. Do large molecular sequence divergences imply an early closure of the Isthmus of Panama? Proc Natl Acad Sci. 2015;112:43.
51. Mishra P, Chan DC. Mitochondrial dynamics and inheritance during cell division, development and disease. Nat Rev Mol Cell Biol. 2014;15:634–46.
52. Molnar P. Closing of the Central American Seaway and the ice age: a critical review. Paleoceanography. 2008;23:PA2201.
53. Montes C, Bayona G, Cardona A, Buchs DM, Silva CA, Morón S, Hoyos N, Ramírez DA, Jaramillo CA, Valencia V. Arc-continent collision and orocline formation: closing of the Central American seaway. J Geophys Res. 2012;117:B04105.
54. Montes C, Cardona A, Jaramillo C, Pardo A, Silva JC, Valencia V, Ayala C, Pérez-Angel LC, Rodriguez-Parra LA, Ramirez V, Niño H. Middle Miocene closure of the Central American Seaway. Science. 2015;348:226–9.
55. Obbard DJ, Maclennan J, Kim K-W, Rambaut A, O'Grady PM, Jiggins FM. Estimating divergence dates and substitution rates in the *Drosophila* phylogeny. Mol Biol Evol. 2012;29:3459–73.
56. O'Dea A, Lessios HA, Coates AG, Eytan RI, Restrepo-Moreno SA, Cione AL, Collins LS, de Queiroz A, Farris DW, Norris RD, Stallard RF, Woodburne MO, Aguilera O, Aubry M-P, Berggren WA, Budd AF, Cozzuol MA, Coppard SE, Duque-Caro H, Finnegan S, Gasparini GM, Grossman EL, Johnson KG, Keigwin LD, Knowlton N, Leigh EG, Leonard-Pingel JS, Marko PB, Pyenson ND, Rachello-Dolmen PG, Soibelzon E, Soibelzon L, Todd JA, Vermeij GJ, Jackson JBC. Formation of the isthmus of panama. Sci Adv. 2016;2:e1600883.
57. Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin PR, O'hara RB, Simpson GL, Solymos P, et al.. Vegan: Community Ecology Package. R package version 2.4–3. Vienna: R Foundation for Statistical Computing. 2016.
58. O'Leary SJ, Puritz JB, Willis SC, Hollenbeck CM, Portnoy DS. These aren't the loci you'e looking for: Principles of effective SNP filtering for molecular ecologists. Mol Ecol. 2018. https://doi.org/10.1111/mec.14792.
59. Ossowski S, Schneeberger K, Lucas-Lledó JI, Warthmann N, Clark RM, Shaw RG, Weigel D, Lynch M. The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. Science. 2009;327:984.
60. Oswald JA, Overcast I, Mauck WM 3rd, Andersen MJ, Smith BT. Isolation with asymmetric gene flow during the nonsynchronous divergence of dry forest birds. Mol Ecol. 2017;26:1386–400.
61. Paradis E, Claude J, Strimmer K. APE: analyses of phylogenetics and evolution in R language. Bioinformatics. 2004;20:289–90.
62. Petrov DA, Hartl DL. Patterns of nucleotide substitution in Drosophila and mammalian genomes. Proc Natl Acad Sci USA. 1999;96:1475–9.
63. Popescu A-A, Huber KT, Paradis E. ape 3.0: New tools for distance-based phylogenetics and evolutionary analysis in R. Bioinformatics. 2012;28:1536–7.
64. Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. Syst Biol. 2018;67:901–4.
65. Schmidt DN, Williams M, Haywood AM, Gregory FJ, Others. . The closure history of the Central American seaway: evidence from isotopes and fossils to models and molecules. Deep time perspectives on climate change marrying the signal from computer models and biological proxies. London: Geological Society of London; 2007. p. 427–42.
66. Schneider B, Schmittner A. Simulating the impact of the Panamanian seaway closure on ocean circulation, marine productivity and nutrient cycling. Earth Planet Sci Lett. 2006;246:367–80.
67. Sepulchre P, Arsouze T, Donnadieu Y, Dutay J-C, Jaramillo C, Le Bras J, Martin E, Montes C, Waite AJ. Consequences of shoaling of the Central American Seaway determined from modeling Nd isotopes. Paleoceanography. 2014;29:2013.
68. Shafer ABA, Peart CR, Tusso S, Maayan I, Brelsford A, Wheat CW, Wolf JBW. Bioinformatic processing of RAD-seq data dramatically impacts downstream population genetic inference. Methods Ecol Evol. 2017;8:907–17.
69. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30:1312–3.
70. Takahata N, Satta Y, Klein J. Divergence time and population size in the lineage leading to modern humans. Theor Popul Biol. 1995;48:198–221.
71. Titus BM, Blischak PD, Daly M. Genomic signatures of sympatric speciation with historical and contemporary gene flow in a tropical anthozoan (Hexacorallia: Actiniaria). Mol Ecol. 2019;28:3572–86.
72. Tripp EA, Tsai Y-HE, Zhuang Y, Dexter KG. RADseq dataset with 90% missing data fully resolves recent radiation of Petalidium (Acanthaceae) in the ultra-arid deserts of Namibia. Ecol Evol. 2017;7:7920–36.
73. Wells SJ, Dale J. Contrasting gene flow at different spatial scales revealed by genotyping-by-sequencing in *Isocladus armatus*, a massively colour polymorphic New Zealand marine isopod. PeerJ. 2018;6:e5462.
74. Williams ST, Knowlton N, Weigt LA, Jara JA. Evidence for three major clades within the snapping shrimp genus *Alpheus* inferred from nuclear and mitochondrial gene sequence data. Mol Phylogenet Evol. 2001;20:375–89.
75. Winston ME, Kronauer DJ, Moreau CS. Early and dynamic colonization of Central America drives speciation in Neotropical army ants. Mol Ecol. 2016. https://doi.org/10.1111/mec.13846.
76. Yang Z. Likelihood and Bayes estimation of ancestral population sizes in hominoids using data from multiple loci. Genetics. 2002;162:1811–23.
77. Halligan DL, Eyre-Walker A, Andolfatto P, Keightley PD. Patterns of evolutionary constraints in intronic and intergenic DNA of *Drosophila*. Genome Res. 2004;14:273–9.

Publisher's Note