MOLECULAR ECOLOGY

Molecular Ecology (2016) **25**, 1478–1493 doi: 10.1111/mec.13579

Transcriptome-wide patterns of divergence during allopatric evolution

RICARDO J. PEREIRA,*† FELIPE S. BARRETO,*‡ N. TESSA PIERCE,* MIGUEL CARNEIRO§ and RONALD S. BURTON*

**Marine Biology Research Division, Scripps Institution of Oceanography, University of California, San Diego, La Jolla, CA 92093-0202, USA, †Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Øster Voldgade 5-7, 1350 Copenhagen, Denmark, ‡Department of Integrative Biology, Oregon State University, Corvallis, OR 97331, USA, §CIBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, Campus Agrário de Vairão, 4485-661 Vairão, Portugal*

Abstract

Recent studies have revealed repeated patterns of genomic divergence associated with species formation. Such patterns suggest that natural selection tends to target a set of available genes, but is also indicative that closely related taxa share evolutionary constraints that limit genetic variability. Studying patterns of genomic divergence among populations within the same species may shed light on the underlying evolutionary processes. Here, we examine transcriptome-wide divergence and polymorphism in the marine copepod *Tigriopus californicus*, a species where allopatric evolution has led to replicate sets of populations with varying degrees of divergence and hybrid incompatibility. Our analyses suggest that relatively small effective population sizes have resulted in an exponential decline of shared polymorphisms during population divergence and also facilitated the fixation of slightly deleterious mutations within allopatric populations. Five interpopulation comparisons at three different stages of divergence show that nonsynonymous mutations tend to accumulate in a specific set of proteins. These include proteins with central roles in cellular metabolism, such as those encoded in mtDNA, but also include an additional set of proteins that repeatedly show signatures of positive selection during allopatric divergence. Although our results are consistent with a contribution of nonadaptive processes, such as genetic drift and gene expression levels, generating repeatable patterns of genomic divergence in closely related taxa, they also indicate that adaptive evolution targeting a specific set of genes contributes to this pattern. Our results yield insights into the predictability of evolution at the gene level.

Keywords: comparative transcriptomics, copepod, genetic drift, positive selection, protein evolution, speciation

Received 16 September 2015; revision received 3 December 2015; accepted 6 January 2016

Introduction

Repeated evolution, or the independent evolution of similar phenotypes in different taxa or populations, provides strong evidence for natural selection (Harvey & Pagel 1991) and for evolutionary constraints in limiting the available variation upon which natural selection can act (Wake 1991; Gould 2002). Tracing the evolution of those phenotypes to repeated changes in homologous genes is the most compelling evidence that gene-level evolution is both repeatable and predictable (Gould 2002). Yet, the evolutionary processes under which genetic evolution is predictable remain unclear (Stern & Orgogozo 2009).

Several studies applying new genomic methods have uncovered repeated patterns of genomic divergence associated with species formation. Genomewide studies of species that evolved via parallel evolution (Gagnaire *et al.* 2013; Arnegard *et al.* 2014; Renaut *et al.* 2014;

Correspondence: Ricardo J. Pereira, E-mail: ricardojn.pereira@gmail.com

Soria-Carrasco *et al.* 2014) show repeated patterns of genomic divergence, suggesting that evolution often results from changes at a relatively small subset of genes. Indeed, the probability of gene reuse in parallel phenotypic evolution is particularly high between closely related species (between 30% and 50% of the time) and declines with increasing age of the common ancestor of the compared taxa (Conte *et al.* 2012). This pattern may arise by two different processes. On one hand, natural selection acting in independent populations may target a limited number of genes with functions related to the shared selective pressure driving divergence (Jones *et al.* 2012; Renaut *et al.* 2014). On the other hand, genetic drift may also contribute to repeated evolution at the gene level because of gene-specific constraints (e.g. standing genetic variation, mutation rate, variation in recombination rate, linkage relationships or pleiotropic effects; Stern & Orgogozo 2009; Christin *et al.* 2015). These processes are not exclusive because genetic drift will condition the supply and fixation of beneficial mutations where natural selection can act or may generate selective pressures that elicit adaptive evolution. For example, during 'compensatory coadaptation', fixation of a deleterious mutation at one neutrally evolving locus establishes a positive selection pressure for a compensatory mutation in a functional interacting locus to regain some component of fitness (Rand *et al.* 2004). The relative contribution of genetic drift and positive selection for repeated patterns of genomic divergence is difficult to evaluate, but they result in distinct genetic signatures (Tsagkogeorga *et al.* 2012; Gayral *et al.* 2013). Scanning for such genetic signatures in genomes of taxa at early stages of divergence might provide insight into the timing and relative contributions of these two evolutionary processes to generate predictable patterns of gene evolution.

Here, we focus on diverging populations of the marine copepod *Tigriopus californicus*. This species ranges along the west coast of North America, from central Baja California to southern Alaska. Restriction to pools in the upper intertidal of isolated rock outcrops has resulted in allopatric differentiation without gene flow (Burton 1998). Parallel adaptation in *T. californicus* might occur in response either to an extrinsic ecological environment or an intrinsic genomic environment. High tidal pools are characterized by rapid changes in abiotic parameters and are highly affected by desiccation during the summer and rainfall during the winter, resulting in high variability in salinity, oxygen and pH (Vittor 1971; Altermatt *et al.* 2012). Although some of these stressors are expected to affect allopatric populations across the entire species range in a parallel fashion, other stressors, such as temperature, vary with latitude and favour population-specific adaptation to local environmental conditions (Willett 2010). Parallel adaptation in *T. californicus* is also known to occur in response to intrinsic selective pressures caused by gene coadaptation (Burton *et al.* 2006). For example, allopatric populations experience similarly high mutation rates in mitochondrial genes (on average 55 times faster than nuclear genes; Willett 2012). Nuclear-encoded proteins that interact with mitochondrial proteins are thus likely to be under positive selection to maintain proper cellular function (Willett & Burton 2004; Barreto & Burton 2013a), resulting in an open-ended molecular evolutionary arms race. Experimental hybridization studies in *T. californicus* seem to support this gene coadaptation hypothesis: hybrid breakdown among multiple populations is attributable to co-evolution between mitochondrial and nuclear genes (Ellison & Burton 2008b), as well as among nuclear genes (Pritchard *et al.* 2011). Observation of hybrid breakdown in crosses between multiple populations (Edmands 1999; Ellison & Burton 2008b) suggests that gene coadaptation has evolved repeatedly in multiple allopatric populations.

The availability of many populations with varying degrees of genetic divergence and reproductive isolation makes *T. californicus* an attractive model for analysis of genomic evolution during the continuum of allopatric divergence. Here, using transcriptome sequence data from six geographically isolated populations, we test: (i) whether geographic isolation leads to rapid differentiation among population, (ii) whether population differentiation is associated with repeatable patterns of protein divergence; (iii) whether molecular signatures of accelerated protein evolution are predictable at several stages of allopatric evolution; and (iv) the relative roles of genetic drift and positive selection in contributing to the observed pattern of repeated genomic divergence.

Materials and methods

Population collection and sequencing

We collected copepods from high intertidal rocky pools in three northern and three southern populations of California (Fig. 1): San Diego (SD: 32°44′N, 117°15′W), Bird Rock (BR: 32°48′N, 117°16′W), Abalone Cove (AB: 33°44′N, 118°22′W), Santa Cruz (SCN: 36°57′N, 122°03′W), Pescadero (PES: 37°15′N, 122°24′) and Bodega Bay (BB: 38°19′N, 123°4′W). All cultures were maintained at common garden; multiple cultures from each site were periodically mixed to limit inbreeding. We extracted RNA from pools of 300–400 individuals from all developmental stages, using the standard Tri-Reagent (Sigma) protocol. Resuspended RNA pellets were further purified with RNeasy Mini columns

(Qiagen), and final sample integrity and quantity were assessed with an Agilent 2100 BioAnalyzer. This protocol was repeated twice for each population, resulting in 12 RNA samples. Libraries were constructed by the Beijing Genomics Institute (BGI) and by Cofactor Genomics, using 100-bp paired-end sequencing on an Illumina HiSeq II.

Population de novo assemblies

We constructed de novo assemblies for each population separately, to avoid mapping biases caused by high interpopulation divergence (20% across mitochondrial genes; Burton *et al.* 2007). FASTQ reads were trimmed via PRINSEQ (prinseq-lite-0.19.5; Schmieder & Edwards 2011) to remove base pairs with a Phred score < 20 and trimming of poly-A tails >8 bp in length. Reads that were trimmed below 55 bp in length or read pairs that were no longer complete after trimming were removed prior to assembly. TRINITY de novo assembly was performed using default parameters (Grabherr *et al.* 2011). To reduce gene redundancy within each assembly, we performed two additional rounds of assembly using the CAP3 assembler (Huang & Madan 1999) with default parameters. We evaluated the quality of each complete assembly by calculating the percentage of annotated core genes from eukaryotic genomes using CEGMA (Parra *et al.* 2007). We identified sets of orthologous nuclear genes across the six population assemblies using a reciprocal best BLASTN hit strategy between the SD assembly and each of the other assemblies (Camacho *et al.* 2008). We retained a total of 12 573 nuclear orthologous genes across all six populations. We assembled the 13 mitochondrial genes separately for each population. Using BOWTIE2 (Langmead & Salzberg 2012), we mapped our libraries to the three complete mitochondrial genomes published for this species (SD, AB and SCN, from Burton *et al.* 2007): SD and BR mapped to SD; AB mapped to AB; and SCN, PES and BB mapped to SCN. We extracted uniquely mapped reads covering the mitochondrial genes, identified single nucleotide polymorphisms (SNPs) relative to the reference using SAMTOOLS (Li *et al.* 2009) and exported consensus mitochondrial sequences using the majority rule for each base, substituting *N*'s for any base with <8 reads coverage. We used the reciprocal best BLAST hit strategy to check for redundancies between mitochondrial and nuclear genes within each population assembly.

Population divergence

We estimated a 'species tree' for the six populations from independent gene trees, using the NJ_{ST} approach (Liu & Yu 2011). We used our 12 573 orthologous nuclear genes to estimate a six-population alignment for each gene, using MUSCLE (Edgar 2004). To avoid inflating interpopulation divergence due to assembly artefacts and misalignments, we filtered each alignment using a sliding window approach and rejecting windows of 21 bp that contained more than seven SNPs (the maximum divergence between these populations estimated at mitochondrial genes; Burton *et al.* 2007). To incorporate uncertainty in gene tree estimation from our alignments, we use RAXML to generate 100 bootstrapped alignments for each locus (resample sites with replacement) and estimate respective maximum-likelihood

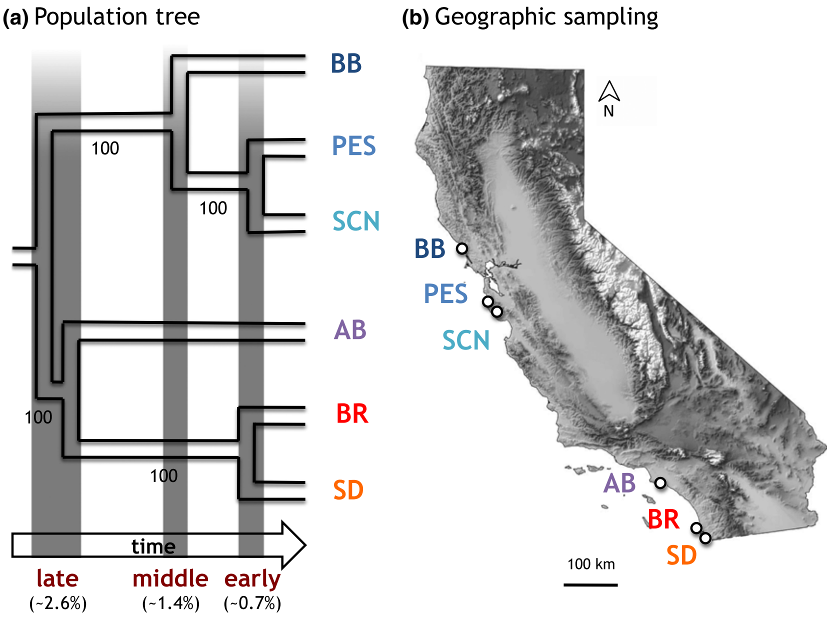

Fig. 1 Geographic isolation in *Tigriopus californicus* results in a continuum of population divergence. (A) Population tree (NJ_{st}; Liu & Yu 2011). Branch lengths reflect average genetic divergence (Table S1, Supporting information), and numbers represent bootstrap support for topology. 'Early', 'middle' and 'late' refer to the three stages of population divergence tested for repeatability of genomic divergence. (B) Sampling of allopatric populations of copepods.

(ML) trees. This resulted in 100 sets of 12 014 ML trees. We used STRAW (Species TRee Analysis Web server; Shaw *et al.* 2013) to estimate a consensus species tree (majority rule) with bootstrap values. For visualization purposes, we rooted the tree based on a species-scale phylogeny of *Tigriopus californicus* (Peterson *et al.* 2013). We estimated interpopulation divergence by calculating uncorrected *P*-distances for each locus with trimmed alignments >100 bp (11 560 nuclear loci) and averages across loci. We replicated the same procedure for the complete 13 mitochondrial genes.

Genetic diversity

We identified SNPs in each population by mapping libraries to their respective assembly using BOWTIE2 (seed size 22 bp; Langmead & Salzberg 2012).

First, to assess how quickly genetic differentiation accumulates during allopatric divergence, we focus on SNPs detected throughout the genome. We used SAMTOOLS (Li *et al.* 2009) to export mapped reads and identify SNPs segregating within each population according to the following criteria: mapping quality > 20, SNP quality > 20 and coverage > 8×. We produced a trimmed six-population alignment, using MUSCLE (Edgar 2004), and projected the identified SNPs using custom scripts. This analysis was restricted to positions where all six populations had callable genotypes according to the criteria described above. This allowed us to compute shared and fixed mutations for every interpopulation comparison, across 12 044 nuclear loci, encompassing a total of 16 415 579 bp. We tested whether fixed and shared mutations changed linearly or exponentially with genetic divergence (as a proxy for time since divergence) by fitting a linear ($y = b + x$) and an exponential ($y = ax^b$) model. The linear model for fixed mutations was forced through 0, assuming that at time of population split, there are no fixed differences between populations. Models were fitted and compared using AIC in R (available from http://www.R-project.org).

Second, to assess the fraction of polymorphisms that are slightly deleterious and neutral, we focus on SNPs that occur in synonymous (π_S) and nonsynonymous sites (π_N), following the approach of Gayral *et al.* (2013). This approach is developed for transcriptome data from nonmodel organisms where a reference genome is not available. In short, de novo assemblies are used both to predict open reading frames (ORFs) for homologous genes sequences and to map reads. Mapped reads are then filtered for paralogues, and SNPs are called with a conservative coverage threshold to account for variation in gene expression among individuals and loci. Interpopulation comparisons of the same genomic regions allow the estimation of summary statistics for population genetics analysis. Using the program READS2SNP (Tsagkogeorga *et al.* 2012), we filtered high-coverage SNPs while controlling for possible paralogous mapping (spa option), according to the following criteria: mapping quality > 20, SNP quality > 20 and coverage > 30×. We repeated filtering using 20× and 40× coverages to assess a potential effect of library size in SNP detection. We identified ORFs using the program TRANSDECODER (Haas *et al.* 2013) and computed π_N, π_S and π_N/π_S for every gene in each population. Due to differences in coverage, this approach retrieved results from a varying number of nuclear loci across populations: 3620 in SD; 2904 in BR; 3845 in AB; 3005 in SCN; 3574 in PES; and 2754 in BB. To avoid biases created by fewer polymorphic genes, average population indexes of diversity were computed as the sum of π_N and the sum of π_S across all loci.

Protein divergence and adaptive evolution

We focused on five population comparisons representative of three stages of divergence: early a—SCN-PES, early b—SD-BR, middle—SCN-BB, late a—SD-AB, late b—BR-BB. We annotated our sets of orthologous genes relative to the SD assembly using BLAST2GO (Conesa *et al.* 2005), retaining the highest hit with *E*-value ≤ 10^{-3} and its predicted gene name. We used a custom Perl script to extract the most likely ORF for SD contigs according to their BLASTX best hits. The transcriptome assemblies from the other five populations were then each blasted (BLASTN) against the SD ORFs, extracting the best ORF for every ortholog. For each focal population comparison listed above, we produced pairwise alignments using MACSE (Ranwez *et al.* 2011), which accommodates possible frameshifts introduced by sequencing and hence retains the appropriate reading frame. A custom script was employed to automate the ~7000 alignments in each two-population comparison. For each protein alignment, we calculated average d_N and d_S for the entire protein using the ML method of Yang & Nielsen (2000), implemented in the script YN00 within the PAML package (version 4.4; Yang 2007). We visually inspected all alignments with $d_S > 2$ and removed those that showed evidence of possible assembly or alignment errors, resulting in a total of 7005 nuclear genes plus 13 mitochondrial genes. Pairwise ω (ω = d_N/d_S) was estimated for 5677 genes with $d_S > 0$ at any interpopulation comparison. For each focal population comparison, we computed which loci were among the top 5% and 10% for d_N, d_S, and ω and calculated overlaps across the five comparisons. We assessed whether the number of overlapping genes could occur by chance by comparing the observed value to a null distribution of overlap between three independent population comparisons, a conservative scenario relative to our five population comparisons. A null distribution of number of genes with overlap was obtained by: (i) simulating three random draws of 5% and 10% genes from a total of 7018 genes, (ii) calculating overlaps among the three lists and (iii) repeating this process 1000 times.

In a second analysis, for the 36 target genes (i.e. top 10% of ω across the five population comparison), we tested whether a subset of codons exhibited $\omega > 1$ across all populations. We first aligned all six orthologs for each gene using MACSE, followed by removal of poorly aligned regions using the 'codon' option in Gblocks (Castresana 2000; parameters: -b1=4 -b2=5 -b3=6 -b4=10 -b5=h). We then used CODEML in PAML to test the fit of a null model that does not allow $\omega > 1$ (M7; neutral evolution) and a more general model that does (M8; positive selection). The model fitting was compared statistically using a likelihood ratio test; the negative of twice the log-likelihood difference between models was compared with the chi-square distribution with d.f. = 2.

Relative contribution of adaptive and nonadaptive evolution

The strength of genetic drift experienced by individual genes will depend on their genomic location and effective rates of recombination. For example, genes experiencing higher drift will accumulate more slightly deleterious polymorphisms (π_N) relative to neutral ones (π_S) (Charlesworth & Charlesworth 2010). To assess the relative contribution of genetic drift generating signatures of adaptive evolution in our data set, we focused on population comparisons representative of later stages of divergence (late a—SD-AB, late b—BR-BB). We estimate genetic drift experienced by each gene by computing gene-specific π_N/π_S. We computed the direction of selection (negative or positive) using the DoS statistic (Stoletzki & Eyre-Walker 2011): $\mathrm{DoS} = d_N/(d_N + d_S) - \pi_N/(\pi_N + \pi_S)$. We assessed differences between candidate and remaining genes using an ANOVA.

Gene-specific evolutionary constraints will affect the variability upon which genetic drift and positive selection can act, impacting rates of protein evolution (Subramanian 2004; Drummond *et al.* 2005). Although the mechanisms behind this observation are not clear, gene expression is a commonly used proxy for such constraints. To estimate expression, we measured the number of Fragments Per Kilobase of transcript per Million mapped reads (FPKM) in two biological replicates for each population and computed averages per gene. We tested for correlations between FPKM and π_N/π_S relative to ω (d_N/d_S) in all sampled genes, using a Mann–Whitney *U*-test. We then tested for differences between candidate genes within the top 10% of ω and the remaining proteins, using a Wilcoxon test.

Results and discussion

To investigate patterns of transcriptome-wide divergence during allopatric evolution of *Tigriopus californicus*, we generated high-quality de novo transcriptome assemblies for each of the six populations, using between 22 and 104 million paired-end 100 bp reads. More than 98% of the core genes of eukaryotic genomes (Parra *et al.* 2007) are represented in every assembly (Table S1, Supporting information). Reciprocal BLAST searches resulted in the identification of 12 573 orthologous loci from the nuclear genome plus all 13 mitochondrial genes across the populations. The orthologous genes have similar sizes across populations and coverage ranges from 16 to more than 65 000 reads per contig (Fig. S1, Supporting information).

Geographic isolation results in three stages of population divergence

In assessing the phylogenetic relationships among the populations, closely related populations may present discordance among independent gene trees due to the influence of differential gene sorting. Thus, we estimated a single 'species tree' ($\mathrm{NJ_{ST}}$; Liu & Yu 2011) using gene trees estimated from each of the 12 014 nuclear loci with reliable alignments. Our results show that the tree topology recapitulates geography (Fig. 1), in agreement with the restricted dispersal ability known for this species. Northern and southern populations form two reciprocally monophyletic groups, with geographically close populations sharing a common ancestor more recently than with the third more distant population. Our results are in agreement with previous phylogenetic trees estimated from mitochondrial genes alone (Willett & Ladner 2009; Peterson *et al.* 2013), except that we now find strong support for deeper nodes that were previously unresolved.

As a proxy for relative divergence time between each population pair, we calculated uncorrected average pairwise divergence for the 11 560 nuclear loci larger than 100 bp (Table S1, Supporting information). Our results indicate that interpopulation comparisons reflect three different stages of divergence (Fig. 1A). Neighbouring populations from the northern and southern clades (SCN-PES and SD-BR, respectively) are 0.7% divergent, reflecting two, phylogenetically independent, vicariant events at an early stage of divergence. BB is 1.3% and 1.4% divergent from the other northern populations, describing an intermediate stage of divergence. Finally, all the remaining population comparisons, including AB relative to the other southern populations, are 2.3–2.8% divergent, revealing later stages of population divergence. The same relative magnitude of divergence is mirrored by the mitochondria (Table S2, Supporting information) with the notable exception of the populations at the early stage of divergence, where the southern population pair shows a level of mitochondrial divergence almost eight times larger than the northern population pair.

Although the high levels of genetic divergence reported here are typically found between reproductively isolated species, experimental crosses in *T. californicus* show that, in this species, hybrid inviability evolves at much later stages of divergence, as observed in other species lacking sex chromosomes (Lima 2014). Hybrids between SD and BR show some reduced fitness relative to parentals, suggesting that such early stages of allopatric evolution have already resulted in some genetic incompatibilities (Pereira *et al.* 2014). The magnitude of hybrid breakdown increases with genetic divergence among parental populations, being most pronounced at the later stages of divergence reported here (e.g. SD-SCN, BR-SCN; Edmands 1999; Pereira *et al.* 2014). In this context, comparisons among the six populations of *T. californicus* studied here reflect a continuum of allopatric evolution, presenting a rare opportunity to study genomic divergence along the continuum between population and species level divergence.

Low N_e leads to an exponential decrease of shared polymorphisms

Patterns of genomic divergence during allopatric evolution are not only conditioned by time since geographic isolation and the accumulation of new mutations (divergence), but also by demographic effects that condition how quickly pre-existing or new variation becomes fixed in descendent populations (differentiation). The neutral theory of evolution (Kimura 1968; Ohta 1992) posits that the strong effect of genetic drift in small populations leads to the potential fixation of slightly deleterious mutations and loss of slightly advantageous mutations at a higher rate than in larger populations. Here, we use SNPs to assess how quickly genetic differentiation accumulates during the continuum of allopatric evolution reflected by *T. californicus* populations and assess how such differentiation might be affected by genetic drift.

Based on 12 044 nuclear loci, our results show that even at early stages of divergence (0.7% nuclear divergence between SCN-PES and SD-BR), most variable nucleotide sites are fixed (62.7% and 72.2% of SNPs, between SCN-PES and SD-BR, respectively) while only 0.6% or less are shared (Table S3, Supporting information, Fig. 2); the remaining SNPs are polymorphic within each population. These population pairs are, respectively, 46 and 8 km apart, which indicates that geographic isolation in *T. californicus* can occur at fine spatial scales. By inspecting levels of differentiation across all 15 interpopulation comparisons, we observe that genomic differentiation based on polymorphisms does not appear to be linearly related to divergence time and is more consistent with an exponential change (Fig. 2; $\mathrm{AIC}_{\mathrm{exponential}} \ll \mathrm{AIC}_{\mathrm{linear}}$ for fixed and shared mutations). More population comparisons with intermediate stages of divergence are needed to confirm this exponential change in polymorphisms. Although not all the 15 pairwise population comparisons are statistically independent (because multiple comparisons use the same population), changes remain exponential when this analysis is restricted to the three independent pairwise comparisons (Fig. S2, Supporting information). This result shows that, in *T. californicus*, allopatric evolution leads to a rapid decrease of shared polymorphisms between diverging populations.

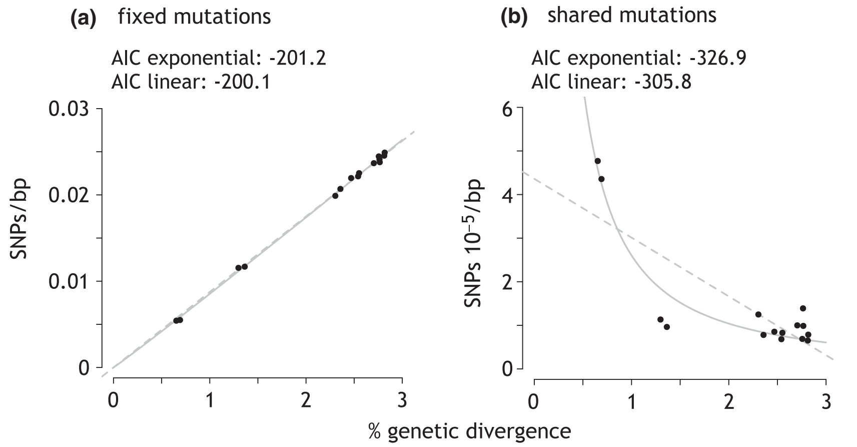

Fig. 2 Interpopulation variability changes exponentially with time since divergence. Points represent all pairwise comparisons between the six populations for fixed (A) and shared (B) mutations. Solid lines represent the exponential model of single nucleotide polymorphism (SNP) variation relative to the percentage of genetic divergence (proxy for time); dashed lines represent the linear model.

Neutral genetic polymorphisms, or genetic diversity (π), are expected to increase with effective population size ($\pi = 4N_e\mu$; where N_e is the effective population size and μ is the mutation rate). Thus, the high levels of population differentiation observed in *T. californicus* could potentially be caused by a relatively small N_e within each population. Based on a subset of 2754 high-coverage nuclear loci, we estimated the fraction of genetic polymorphisms in nonsynonymous (π_N) and synonymous (π_S) sites for each population. Because these loci are highly expressed across populations, library size had a negligible impact in SNP detection with 30× coverage (Fig. S3, Supporting information). Assuming that π_S reflects neutral polymorphisms, species with small N_e are expected to show lower π_S than species with large N_e. Accordingly, our results show that populations of *T. californicus* have relatively low average π_S compared to most taxa (Romiguier *et al.* 2014), including other Arthropoda (Fig. 3A). Despite the high reproductive rate and census size of this organism, our results indicate that the average number of individuals contributing to reproduction over large evolutionary timescales is relatively small, similar to other crustaceans such as crabs and shrimps (Romiguier *et al.* 2014). Such small N_e values are likely caused by demographic bottlenecks that frequently affect populations of *T. californicus*, such as high mortality caused by rainfall and wave scouring of pools during winter and desiccation in the summer, followed by recolonization from interconnected pools and rapid population expansion that favour the random fixation of low-frequency variants in allopatric populations (Excoffier & Ray 2008).

Species with small N_e are more prone to genetic drift, which in turn will increase the probability of slightly deleterious mutations (presumably reflected by π_N) relative to neutral mutations (π_S). As such, taxa with small N_e are expected to show lower π_N, lower π_S, and higher π_N/π_S ratio than taxa with large N_e. Similar to what was found in other taxa (Gayral *et al.* 2013; Romiguier *et al.* 2014), populations of *T. californicus* show that π_N/π_S is inversely proportional to π_S ($r^2 = 0.9$, Mann–Whitney *U*-test: $P = 0.002$), supporting a role of genetic drift driving divergence. Because π_S variation between populations of the same species is not confounded, in principle, by variation of mutation rate (μ) or of life-history traits, the observed negative correlation directly reflects variation of N_e. Within *T. californicus*, diversity within populations ranges from 0.006 to 0.017, suggesting that northern populations (mean π_S: 0.015) are on average two times larger than southern populations (mean π_S: 0.007; Fig. 3B). Because the efficiency of selection relative to genetic drift is higher in larger populations, northern populations are expected to purge deleterious mutations by purifying selection more efficiently, while southern populations are expected to accumulate more slightly deleterious alleles. Mitochondrial genes are particularly prone to the effect of genetic drift because the N_e of the mitochondria is one-fourth as large as N_e for autosomal genes (Pool & Nielsen 2007). Such differences in N_e among genes and populations likely contribute to our finding that smaller populations in the south (SD and BR) show eight times higher mitochondrial divergence relative to the larger populations in the north (SCN and PES) with comparable genomewide divergence (Table S2, Supporting information).

Together, our results show that allopatric evolution in *T. californicus* is strongly affected by genetic drift. Despite population-specific variation of N_e, allopatric populations are relatively small, leading to an exponential decay of shared polymorphisms and the fixation of slightly deleterious mutations in descendent populations.

Nonsynonymous mutations accumulate more often in a subset of proteins

Recent studies comparing genomes of multiple related species have found that certain genes or regions repeatedly show higher divergence relative to the mean of the

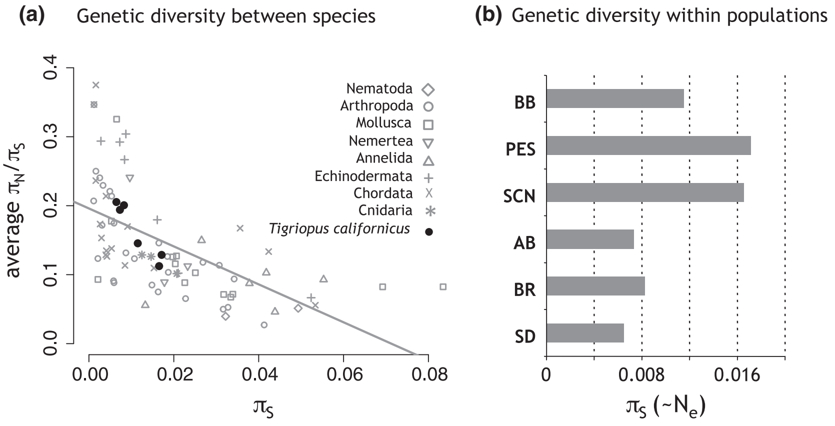

Fig. 3 Population differentiation is largely affected by genetic drift. (A) Relationship between π_N/π_S and π_S. Grey symbols refer to published data in nonmodel species (Romiguier *et al.* 2014), while black symbols refer to the six allopatric populations of *Tigriopus californicus*. Line reflects the linear model for all data ($r^2 = 0.35$, $P < 10^{-8}$). (B) Genetic diversity within populations of *T. californicus* shown as the fraction of polymorphic single nucleotide polymorphisms (SNPs) in neutral sites (π_S).

genome (e.g. Gagnaire *et al.* 2013; Renaut *et al.* 2014) that are consistent with shared selective pressures and with shared evolutionary constraints. Even in cases where high divergence is not necessarily adaptive, faster evolution of nearly neutral genes can secondarily elicit adaptive compensatory changes in interacting loci ('compensatory coadaptation'; Rand *et al.* 2004; Presgraves 2010). To test for repeatable patterns of genomic divergence in *T. californicus* and identify candidate proteins driving divergence, we focus on five population comparisons at the three different stages of divergence: early (a: SCN-PES; b: SD-BR), middle (SCN-BB) and late (a: SD-AB; b: BR-BB).

Functional divergence of the 7018 annotated protein-coding genes used in the five comparisons follows the average pairwise divergence described by all nuclear loci, with the rate of nonsynonymous substitutions (d_N) increasing with time since divergence (Fig. S4, Supporting information). With respect to d_N, population comparisons at consecutive stages of divergence significantly differ from each other (test of equality on the distributions of d_N; all P-values ≤ 0.01), confirming that these represent distinct stages of genomic divergence. In addition, the two replicated population comparisons at early (SCN-PES; SD-BR) and at late (SD-AB; BR-BB) stages of divergence do not differ from each other (test of equality; both P-values ≥ 0.44), confirming that these represent natural replicates of vicariant events with similar amounts of genomic divergence.

For each population comparison, we identified the proteins with the highest protein divergence (top 5% and 10% with respect to d_N) and assessed overlaps across the five comparisons. Our results show a large overlap across the five population comparisons (Table S4, Supporting information); 39 proteins are among the 5% most divergent in all five interpopulation comparisons, and when we consider the top 10%, the number of overlapping proteins increases to 103. It is important to note that only three of the five population comparisons are statistically independent (SCN-PES, SD-AB, and BR-BB). Nevertheless, the observed number of overlapping genes remains highly significant (P-values < 0.0001) considering a very conservative null expectation based on the three statistically independent comparisons (5% overlap: range—0 to 5 genes, median—1 gene; 10% overlap: range—0 to 19 genes, median—7 genes; Fig. S5, Supporting information). This large number of overlapping proteins indicates that nonsynonymous mutations often accumulate in a specific subset of loci.

Our results show that seven of the 13 mtDNA-encoded proteins are among the 103 fast-evolving genes. Mitochondrial genes are expected to contribute disproportionately to genetic incompatibilities between populations because they (i) tend to accumulate deleterious mutations faster than nuclear genes due to a Muller's ratchet effect (lack of recombination and 1/4 N_e relative to nuclear genes), (ii) have faster mutation rates relative to nuclear genes and (iii) constitute mitonuclear enzyme complexes with key roles in metabolism in all eukaryotic cells (Rand *et al.* 2004; Burton & Barreto 2012). Experimental hybridizations between multiple populations of *T. californicus* (Ellison & Burton 2008b) have shown that divergence in mitochondrial genes results in Dobzhansky–Muller incompatibilities with interacting nuclear loci as a result of intergenomic coadaptation (Rand *et al.* 2004; Burton *et al.* 2013). Yet, identifying which mitochondrial and nuclear genes are involved in such compensatory coadaptation is challenging. In *T. californicus*, mismatch between nuclear and mitochondrial components of the oxidative phosphorylation (OXPHOS) system has been shown to reduce enzymatic activity of complex IV (Edmands & Burton 1999; Rawson & Burton 2002) and ATP production (Ellison & Burton 2008a). Our results show that COX1, one of the three mitochondrial proteins that compose complex IV, is among our most-divergent genes. Interestingly, the remaining mtDNA-encoded proteins with accelerated evolution (NAD1, 2, 4, 5, 6 and 4L) constitute six of the seven mtDNA-encoded components of complex I. It is unknown whether mitonuclear interactions in complex I cause fitness breakdown in *T. californicus*, but studies in humans have shown that about 50% of all mitochondrial disorders affecting the energy metabolism can be traced to mutations in one of the subunits of complex I (Smeitink *et al.* 2001; Brandt 2006).

Our finding that mutations tend to accumulate in a particular set of genes shows that allopatric evolution leads to repeatable patterns of genomic divergence. The evolutionary processes underlying these patterns can be either neutral, for example if these genes are more affected by genetic drift than the remaining genome, or adaptive, if these genes are directly targeted by selection. Distinguishing between these alternatives is difficult, but scanning for genomic signatures of neutral and adaptive evolution can provide valuable insights on the relative contribution of the two evolutionary processes.

Patterns of accelerated protein evolution are predictable

Genomic methods allow investigation of how frequently the same genes are targeted by natural selection in different populations, yielding insights into the predictability of evolution at the genetic level. A meta-analysis of taxa undergoing parallel speciation found that the probability of gene reuse is surprisingly high (30–50% of the time), particularly in closely related taxa (Conte *et al.* 2012). This suggests that strong selective biases and constraints affect adaptive evolution, resulting in change at a relatively small subset of available genes. We scanned the transcriptome for signatures of rapid protein evolution and tested whether those signatures are predictable during the continuum of allopatric evolution reflected by our study populations of *T. californicus*.

We estimated pairwise ω in our five focal population comparisons (ω = d_N/d_S, where d_N is the rate of nonsynonymous substitutions per nonsynonymous site and d_S is similarly the rate of synonymous substitutions). Codons with values of ω > 1 are suggestive of adaptive evolution promoting accelerated divergence between taxa (Nei & Kumar 2000; Charlesworth & Charlesworth 2010). However, when ω is calculated across an entire protein sequence, a criterion of ω > 1 as evidence for adaptive evolution is extremely stringent (Swanson *et al.* 2001a). A sequence-wide ω threshold of 0.5 has been shown to consistently identify genes subjected to adaptive evolution (Swanson *et al.* 2001b, 2004). Our pairwise estimations of ω for 5677 protein-coding genes show high values across the five population comparisons (up to 4.7 in early a, to 3.9 in early b, to 4.4 in middle, to 1.9 in late a, and to 1.4 in late b; Fig. S6, Supporting information). Yet, whether we consider a threshold of ω > 1 or >0.5, the number of genes putatively under adaptive evolution decreases with population divergence (Fig. S7A, Supporting information). Comparative genomic studies in various taxa, such as bacteria (Rocha *et al.* 2006), mammals and birds (Wolf *et al.* 2009), have reported a similar negative correlation between estimates of ω and genetic distance in pairwise comparisons of the sequences for the same gene. This recurrent pattern has been interpreted either as (i) evidence for varying strength of selection during species divergence (Read *et al.* 2002; Baker *et al.* 2004), (ii) an effect of a decrease in nonsynonymous polymorphisms during divergence due to a larger time frame for purifying selection to operate (Rocha *et al.* 2006), or (iii) an artefact due to a larger number of polymorphisms segregating in recently diverged taxa (Peterson & Masel 2009). Our results show that, in *T. californicus*, <0.6% of variable nucleotide sites are shared among populations at early stages of divergence (Fig. 2B; Table S3, Supporting information), suggesting that retention of ancestral polymorphism is unlikely to account for the observed decrease in the number of putatively adaptive genes with time since divergence. Moreover, we observed that, although both kinds of mutations increase during population divergence, d_N accumulates at a much slower rate than d_S (Fig. S7B, Supporting information; both r^2 = 0.99; *P*-values ≪ 0.01), affecting the number of candidate genes based on a fixed threshold.

We used our genealogy of *T. californicus* (Fig. 1A) to test whether loci under accelerated evolution at an early stage of divergence can predict rates of evolution at the remaining four population comparisons, representative of all three stages of population divergence. Our results show that loci with ω > 0.5 at an early stage of divergence (early a: SCN-PES) also have higher ω than the rest of the genome in a phylogenetically independent population comparison with the same level of divergence (early b: SD-BR, Mann–Whitney *U*-test: $p < 2 \times 10^{-6}$; Fig. 4). In addition, this pattern is maintained at

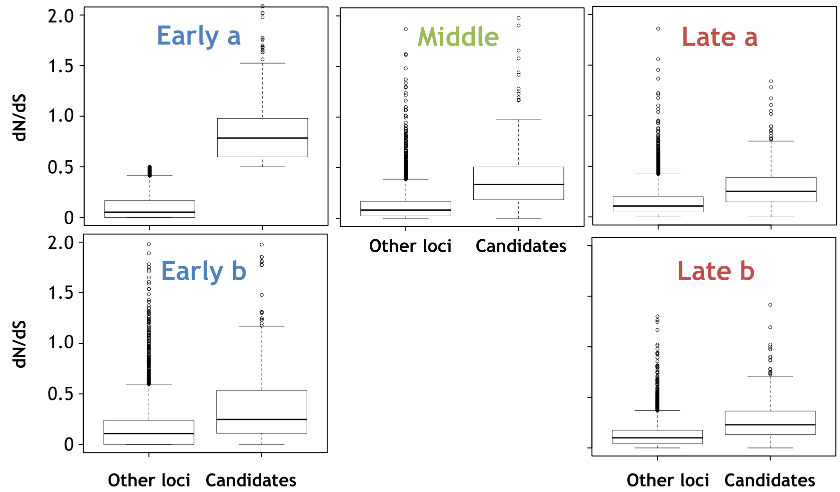

Fig. 4 Proteins with accelerated evolution are predictable at independent vicariant events and along the continuum of species formation. Based on ω estimates in the early a comparison (SCN-PES), 5677 genes were classified either as candidate genes under accelerated evolution (mean ω ≥ 0.5) or neutral (mean ω < 0.5). We tested whether these candidate genes have higher signature of rapid evolution (ω) than the rest of the genome at an independent vicariant event with similar divergence (early b: SD-BR) and at later stages of population divergence (middle: SCN-BB, late a: SD-AB and late b: BR-BB). All differences were statistically significant (Mann–Whitney *U*-test; $P < 2 \times 10^{-6}$).

later stages of population divergence (all Mann–Whitney *U*-tests: *P*-values $< 2 \times 10^{-6}$), suggesting that signatures of rapid protein evolution are predictable both in independent vicariant events and along the continuum of population divergence.

To identify proteins with accelerated evolution (relative to the genome mean) during the continuum of population divergence of *T. californicus*, we assessed which proteins are in the top 5% and 10% of ω for each comparison and calculated overlap across the five comparisons. Our results show that 12 proteins of 5677 are consistently among the top 5% of ω (values across the five population comparisons range between 0.44 and 4.70; Table S5, Supporting information), and 36 are among the top 10% (ω > 0.33). Similar to the overlaps in d_N, this number of overlapping genes remains highly significant considering a very conservative null expectation based on the three statistically independent comparisons (Fig. S5, Supporting information); *P*-values < 0.0001. Of these 36 proteins under accelerated evolution, all are nuclear-encoded. Interestingly, none of these 36 proteins are known to functionally interact with the mitochondria, as would be predicted whether mitonuclear coadaptation would result in accelerated evolution across the entire nuclear protein. Previous studies on candidate genes underlying mitonuclear incompatibilities in *T. californicus* found that hybrid breakdown in the activity of the mitochondria-encoded COX protein can be attributed to a single amino acid substitution in the interacting nuclear-encoded CYC protein (Harrison & Burton 2006). Thus, when adaptive changes in a protein result from a few amino acid changes and the rest of the protein evolves neutrally, those genes can be missed in genome scans using ω values averaged across protein sequences.

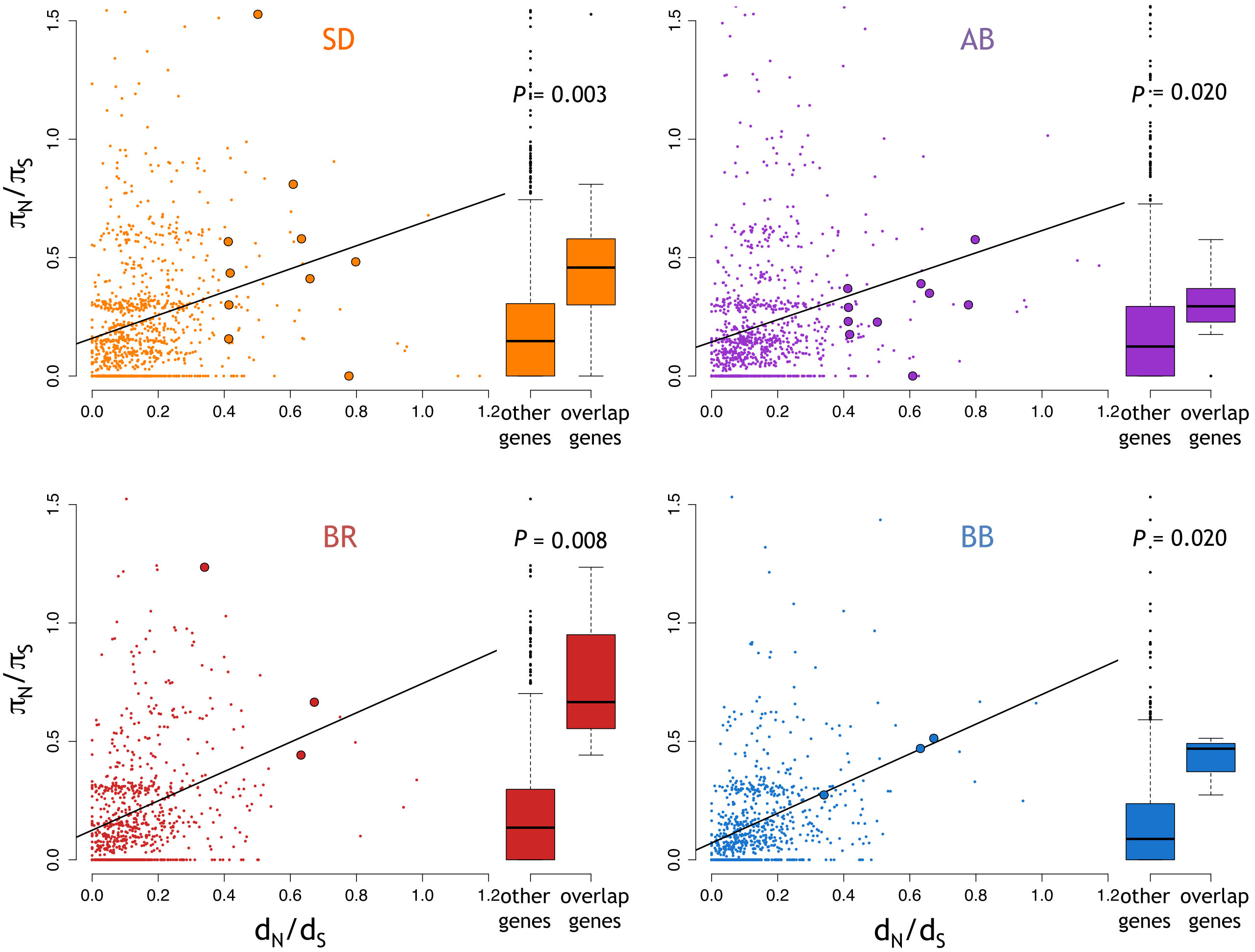

Fig. 5 Rates of protein evolution are constrained by genetic drift. Dot plots refer to values of genetic diversity (π_N/π_S) and divergence (d_N/d_S) by gene; genes without synonymous polymorphism (π_S) or mutations were excluded (d_S), as well as those with <100 sites. Larger dots refer to values in the genes with overlap of d_N/d_S across the five focal population comparisons (i.e. candidate genes under adaptive evolution); lines refer to linear models on each correlation (*P*-values $< 10^{-10}$). Box plots show differences between overlap and the remaining genes; *P*-values from Mann-Whitney U tests are denoted above.

As found in other systems (Conte *et al.* 2012), our results reveal repeated patterns of accelerated evolution during allopatric evolution, supporting the general observation that adaptive evolution may be restricted to a limited set of genes. We also find that proteins with signature of accelerated evolution at early divergence stages maintain higher rates of evolution than the remaining genes, suggesting that patterns of genomic divergence during allopatric evolution are to some extent predictable.

Adaptive and nonadaptive processes contribute to repeatability of genomic divergence

Repeatable patterns of genomic divergence in multiple taxa are consistent both with adaptive evolution targeting a specific set of genes, and with nonadaptive processes, such as genetic drift and gene expression constraining the available variation upon which natural selection can act. Here, we use multiple allopatric populations of *T. californicus* to evaluate the relative contribution of these two evolutionary processes during allopatric evolution within the same species.

Genomic architecture may contribute to changes in protein polymorphisms among genes. For example, genomic regions with low recombination rates, as is typical around centromeres, within inversions, or around genes under selection, experience local reductions in effective population size and thus higher genetic drift. If there is a significant effect of genetic drift in certain genomic regions, levels of protein polymorphism measured as π_N/π_S are expected to correlate with protein divergence (Charlesworth & Charlesworth 2010). To test for the effect of genetic drift contributing for repeated patterns of genomic divergence as measured by d_N/d_S (ω), we focus on population comparisons reflective of later stages of divergence (late a: SD-AB; late b: BR-BB), to avoid higher variance of called polymorphisms typical of earlier stage of divergence. Our results show significant positive correlations throughout the genome between protein polymorphism and divergence (all Mann–Whitney *U*-tests: *P*-values < 10^{-10}), with rapidly evolving proteins showing significantly lower polymorphisms relative to the remaining genes (all Wilcoxon tests: *P*-values ≤ 0.02; Fig. 5). It is important to note that analyses of π_N and π_S are restricted to loci with high coverage (>60 reads per site) and exclude loci with no neutral polymorphism ($\pi_S = 0$) and short length (<100 sites). Although this bias will likely inflate average π_N/π_S, it does not affect the classification of candidate genes based on d_N/d_S. This result is in agreement with the general observation that genetic drift experienced by individual genes affect rates of protein evolution in *T. californicus*. Because the magnitude of genetic drift experienced by certain genomic regions is likely similar across populations, the efficiency of negative selection purging slightly deleterious mutations will also be lower in those genes, contributing significantly to the observed pattern of repeated and predictable evolution at the gene level. To further test whether the 36 candidate proteins showing accelerated evolution across multiple population comparisons are mostly affected by positive or negative selection, we computed the direction of selection (DoS) statistic (Stoletzki & Eyre-Walker 2011). In this modified version of the neutrality index (Rand & Kann 1996), positive values of DoS (excess of d_N/d_S relative to π_N/π_S) are indicative of positive selection, while negative values of DoS reflect the influence of purifying selection. Our results show across the entire transcriptome, most genes have DoS values around zero or slightly negative, suggestive of neutrality (medians: SD = −0.008, AB = 0.000, BR = −0.006, BB = 0.008; Fig. 6). However, the median

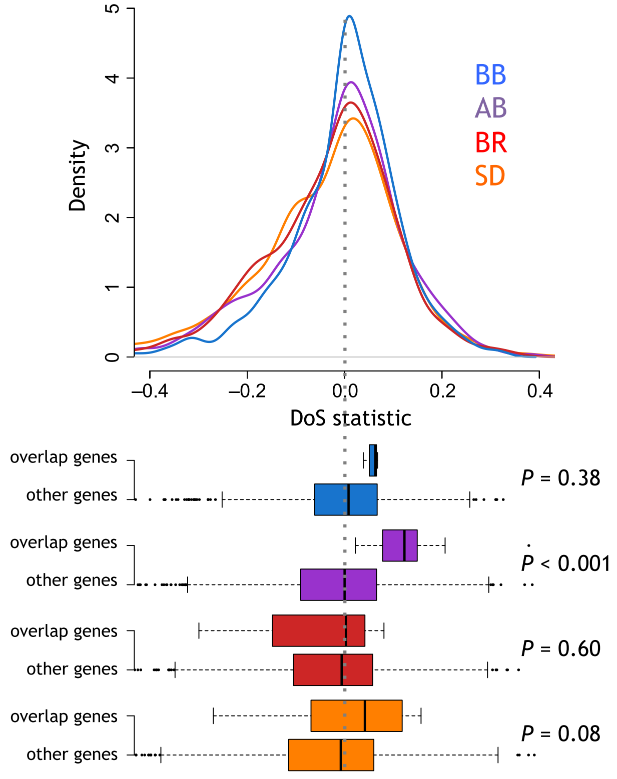

Fig. 6 Genes with repeated signature of accelerated evolution are not affected by purifying (negative) selection. Density plots show the distribution of the direction of selection (DoS; Stoletzki & Eyre-Walker 2011) throughout the genome; dotted line demarks the expectation under neutrality. Box plots show differences between candidate and the remaining genes; *P*-values are denoted.

values of DoS for candidate genes are all positive (SD = 0.042, AB = 0.123, BR = 0.003, BB = 0.063), suggestive of a stronger contribution of positive selection in the signature of adaptive divergence. It is important to note that only one of these differences is significant (ANOVA, $p < 0.001$). In addition, we found that in *T. californicus*, gene expression is inversely proportional to protein divergence (all Mann–Whitney *U*-tests: *P*-values $< 10^{-15}$; Fig. S8, Supporting information), suggesting that evolutionary constraints condition the variability upon which neutral and adaptive processes can act. Together, these results suggest that although neutral evolution is a strong contributor for protein divergence throughout the genome, proteins with accelerated evolution across multiple population comparisons are also affected by positive selection.

To further test these 36 rapidly evolving genes for signatures of positive selection, we aligned all six orthologs for each gene and applied models that allowed ω to vary among codons, instead of assuming one ω for the entire gene. We compared a null model that does not allow for any codons to have $\omega > 1$ against a more general model that does (Yang 2007). Although our power to detect positive selection is low both due to the small number of aligned taxa and the low level of sequence divergence observed between populations (relative to studies comparing different species or genera), we find that a model allowing for positive selection in some codons was a significantly better fit in 42% of our candidate genes (Table 1). For those genes, values of ω ranged from 19.0 in 2% of the protein to 1.9 in 55% of the protein suggesting that positive selection can either target key amino acids or most of the protein sequence. Together, our results suggest that repeated patterns of accelerated evolution in a subset of genes might be, at least partially, driven by positive selection.

Genome scans such as the one employed here can provide a manageable list of candidate loci underlying adaptation in independent evolutionary units. Although most of these genes remain unknown due to insufficient annotation and lack of homology to other organisms, some of our candidate genes belong to functional classes of proteins known to be involved in adaptive evolution in diverse taxa. For example, the two highest values of ω (Table 1) were found in a lectin (19.05) and a histone (15.31). Lectins are involved in binding to carbohydrates displayed in cell surfaces; lectins mediating sperm–egg incompatibility were shown to evolve under positive diversifying selection in oyster, mussel, sea urchin and fruit fly (Mah 2004; Moy *et al.* 2008; Findlay *et al.* 2009; Lima & McCartney 2013). Histones play an important role in DNA binding and gene regulation and have also been suggested to underlie genomic coadaptation in mammal species (Nowick *et al.* 2013; Carneiro *et al.* 2014). Surprisingly, other genes that are known to evolve under diversifying selection in other organisms, such as immune genes (Obbard *et al.* 2009), are not present in this list.

Table 1 Positive selection in proteins with repeated accelerated evolution

			Likelihood				
Protein name	Omega overlap (%)	Alignment size (bp)	Neutral	Selection	*P*-value	*p*	ω
Aleurain-like protease	5	1377	−3096.5	−3092.9	0.025	0.55	1.87
Protein DEP, isoform b	5	3411	−6780.6	−6734.7	0.000	0.08	8.55
Oryzain gamma chain precursor	5	1320	−2836.2	−2810.0	0.000	0.22	4.81
c-Type lectin domain family member f	5	552	−1126.0	−1116.4	0.000	0.02	19.05
Pan domain-containing	5	1719	−3118.7	−3115.5	0.040	0.04	6.13
cg2206-pb-like protein	5	2037	−4249.2	−4227.4	0.000	0.07	6.10
Cysteine protease	10	927	−2018.8	−2008.1	0.000	0.18	4.89
cg13310 cg13310-pa	10	1260	−2378.7	−2361.7	0.000	0.02	12.43
Conserved hypothetical protein	10	969	−1968.7	−1964.6	0.017	0.13	4.64
Adenosine deaminase	10	1095	−2096.8	−2079.4	0.000	0.05	12.22
Histone h1-delta	10	1377	−2638.2	−2634.5	0.025	0.01	15.31
g protein-coupled receptor mth2-like	10	2202	−4161.2	−4153.8	0.001	0.02	8.67
Intraflagellar transport protein 140 partial	10	1425	−2368.5	−2361.9	0.001	0.10	4.84
Isoform b	10	2445	−4455.0	−4451.0	0.019	0.01	10.11
PREDICTED: uncharacterized protein LOC101241641	10	3378	−6764.7	−6761.0	0.023	0.01	8.48

Omega overlap = presence among the top 5% or 10% on each interpopulation comparison according to ω values averaged for the entire protein alignment; proteins in bold are referred to in the text. *p* = proportion of sites estimated to be under positive selection with $\omega > 1$; ω = value of ω in those sites.

Concluding remarks

We focused on multiple allopatric populations of the copepod *Tigriopus californicus* to understand the evolutionary processes underlying the emergent pattern of repeated genomic divergence commonly observed among different species. A phylogenetic tree based on transcriptome data confirms that the evolutionary history of these populations is largely influenced by geographic isolation and that allopatric divergence as resulted in three stages of divergence (Fig. 1). Polymorphism data show that genetic drift is an important driver of allopatric evolution, leading to rapid fixation of alleles between diverging populations (Fig. 2). Analyses of synonymous and nonsynonymous substitutions show that mutations tend to accumulate in a specific set of proteins, leading to repeated and predictable patterns of divergence at the gene-level (Fig. 4). Although this pattern can largely be explained by neutral evolutionary processes, such as drift (Fig. 5) and evolutionary constraints (Fig. S8, Supporting information), we find a strong contribution of positive selection in proteins with signatures of accelerated evolution across multiple population comparisons (Fig. 6, Table 1).

Together, these results show that, although initial stages of allopatric evolution are largely conditioned by neutral processes, adaptive divergence is an important contributor to repeated patterns of genomic evolution. Positive selection targeting specific genes during allopatric evolution can be both caused by adaptation to an extrinsic ecological environment (e.g. in response to abiotic stressors), or adaptation to an intrinsic genetic environment (e.g. in response to compensatory coadaptation among genes). In vitro experiments of isolated proteins (Rawson & Burton 2002; Harrison & Burton 2006) and recombinant inbred lines (Barreto & Burton 2013b) have been successfully used in *T. californicus* to establish links between accelerated evolution in specific genes and the phenotypes affected by them, offering unique insights into the adaptive regimes operating in this system (Burton *et al.* 2013). Follow-up studies can elucidate the function of the candidate genes identified here and provide important insights into the selective regimes causing the observed pattern of repeated accelerated evolution.

Acknowledgements

This study was supported by grants from the U.S. National Science Foundation (DEB1051057 and IOS1155030 to R.S.B). M.C. was supported by POPH-QREN funds from the European Social Fund and Portuguese MCTES (FCT Investigator grant IF/00283/2014; FCT postdoc grant SFRH/BPD/72343/2010). During the submission of this manuscript, R.P. received funding from the European Union's Horizon 2020 Research and Innovation Programme under the Marie Sklodowska-Curie grant agreement No 658706. We thank A. Geraldes, P. Ralph, S. Singhal and anonymous reviewers for their comments on the early versions of this manuscript.

References

Altermatt F, Bieger A, Morgan SG (2012) Habitat characteristics and metapopulation dynamics of the copepod *Tigriopus californicus*. *Marine Ecology Progress Series*, **468**, 85–93.

Arnegard ME, McGee MD, Matthews B *et al.* (2014) Genetics of ecological divergence during speciation. *Nature*, **511**, 307–311.

Baker L, Brown T, Maiden MC, Drobniewski F (2004) Silent nucleotide polymorphisms and a phylogeny for *Mycobacterium tuberculosis*. *Emerging Infectious Diseases*, **10**, 1568–1577.

Barreto FS, Burton RS (2013a) Evidence for compensatory evolution of ribosomal proteins in response to rapid divergence of mitochondrial rRNA. *Molecular Biology and Evolution*, **30**, 310–314.

Barreto FS, Burton RS (2013b) Elevated oxidative damage is correlated with reduced fitness in interpopulation hybrids of a marine copepod. *Proceedings of the Royal Society of London B: Biological Sciences*, **280**, 20131521.

Brandt U (2006) Energy converting NADH: quinone oxidoreductase (complex I). *Annual Review of Biochemistry*, **75**, 69–92.

Burton RS (1998) Intraspecific phylogeography across the point conception biogeographic boundary. *Evolution*, **52**, 734–745.

Burton RS, Barreto FS (2012) A disproportionate role for mtDNA in Dobzhansky-Muller incompatibilities? *Molecular Ecology*, **21**, 4942–4957.

Burton RS, Ellison CK, Harrison JS (2006) The sorry state of F2 hybrids: consequences of rapid mitochondrial DNA evolution in allopatric populations. *The American Naturalist*, **168**, S14–S24.

Burton RS, Byrne RJ, Rawson PD (2007) Three divergent mitochondrial genomes from California populations of the copepod *Tigriopus californicus*. *Gene*, **403**, 53–59.

Burton RS, Pereira RJ, Barreto FS (2013) Cytonuclear genomic interactions and hybrid breakdown. *Annual Review of Ecology Evolution and Systematics*, **44**, 281–302.

Camacho C, Coulouris G, Avagyan V *et al.* (2008) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.

Carneiro M, Albert FW, Afonso S *et al.* (2014) The genomic architecture of population divergence between subspecies of the european rabbit (JL Feder, Ed). *PLoS Genetics*, **10**, e1003519.

Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution*, **17**, 540–552.

Charlesworth B, Charlesworth D (2010) *Elements of Evolutionary Genetics*. Roberts and Company, Greenwood Village, Colorado.

Christin P-A, Weinreich DM, Besnard G (2015) Causes and evolutionary significance of genetic convergence. *Trends in Genetics*, **26**, 400–405.

Conesa A, Gotz S, Garcia-Gomez JM *et al.* (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, **21**, 3674–3676.

Conte GL, Arnegard ME, Peichel CL, Schluter D (2012) The probability of genetic parallelism and convergence in natural populations. *Proceedings of the Royal Society of London B: Biological Sciences*, **279**, 5039–5047.

Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH (2005) Why highly expressed proteins evolve slowly. *Proceedings of the National Academy of Sciences of the USA*, **102**, 14338–14343.

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, **32**, 1792–1797.

Edmands S (1999) Heterosis and outbreeding depression in interpopulation crosses spanning a wide range of divergence. *Evolution*, **53**, 1757–1768.

Edmands S, Burton R (1999) Cytochrome *C* oxidase activity in interpopulation hybrids of a marine copepod: a test for nuclear-nuclear or nuclear-cytoplasmic coadaptation. *Evolution*, **53**, 1972–1978.

Ellison CK, Burton RS (2008a) Genotype-dependent variation of mitochondrial transcriptional profiles in interpopulation hybrids. *Proceedings of the National Academy of Sciences of the USA*, **105**, 15831.

Ellison CK, Burton RS (2008b) Interpopulation hybrid breakdown maps to the mitochondrial genome. *Evolution*, **62**, 631–638.

Excoffier L, Ray N (2008) Surfing during population expansions promotes genetic revolutions and structuration. *Trends in Ecology & Evolution*, **23**, 347–351.

Findlay GD, MacCoss MJ, Swanson WJ (2009) Proteomic discovery of previously unannotated, rapidly evolving seminal fluid genes in *Drosophila*. *Genome Research*, **19**, 886–896.

Gagnaire P-A, Pavey SA, Normandeau E, Bernatchez L (2013) The genetic architecture of reproductive isolation during speciation-with-gene-flow in lake whitefish species pairs assessed by RAD sequencing. *Evolution*, **67**, 2483–2497.

Gayral P, Melo-Ferreira J, Glémin S *et al.* (2013) Reference-free population genomics from next-generation transcriptome data and the vertebrate-invertebrate gap (JJ Welch, Ed). *PLoS Genetics*, **9**, e1003457.

Gould SJ (2002) *The Structure of Evolutionary Theory*. The Belknap University Press, Cambridge, Massachusetts.

Grabherr MG, Haas BJ, Yassour M *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, **29**, 644–652.

Haas BJ, Papanicolaou A, Yassour M *et al.* (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, **8**, 1494–1512.

Harrison JS, Burton RS (2006) Tracing hybrid incompatibilities to single amino acid substitutions. *Molecular Biology and Evolution*, **23**, 559–564.

Harvey PH, Pagel MD (1991) *The Comparative Method in Evolutionary Biology*. Oxford University Press, Oxford.

Huang X, Madan A (1999) CAP3: a DNA sequence assembly program. *Genome Research*, **9**, 868–877.

Jones FC, Grabherr MG, Chan YF *et al.* (2012) The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, **484**, 55–61.

Kimura M (1968) Evolutionary rate at the molecular level. *Nature*, **217**, 624–626.

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

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

Lima TG (2014) Higher levels of sex chromosome heteromorphism are associated with markedly stronger reproductive isolation. *Nature Communications*, **5**, 4743.

Lima TG, McCartney MA (2013) Adaptive evolution of M3 lysin—a candidate gamete recognition protein in the *Mytilus edulis* species complex. *Molecular Biology and Evolution*, **30**, 2688–2698.

Liu L, Yu L (2011) Estimating species trees from unrooted gene trees. *Systematic Biology*, **60**, 661–667.

Mah SA (2004) Positive selection in the carbohydrate recognition domains of sea urchin sperm receptor for egg jelly (suREJ) proteins. *Molecular Biology and Evolution*, **22**, 533–541.

Moy GW, Springer SA, Adams SL, Swanson WJ, Vacquier VD (2008) Extraordinary intraspecific diversity in oyster sperm bindin. *Proceedings of the National Academy of Sciences of the USA*, **105**, 1993–1998.

Nei M, Kumar S (2000) *Molecular Evolution and Phylogenetics*. Oxford University Press, Inc., New York, New York.

Nowick K, Carneiro M, Faria R (2013) A prominent role of KRAB-ZNF transcription factors in mammalian speciation? *Trends in Genetics*, **29**, 130–139.

Obbard DJ, Welch JJ, Kim K-W, Jiggins FM (2009) Quantifying adaptive evolution in the *Drosophila* immune system. *PLoS Genetics*, **5**, e1000698

Ohta T (1992) The nearly neutral theory of molecular evolution. *Annual Review of Ecology and Systematics*, **23**, 263–286.

Parra G, Bradnam K, Korf I (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*, **23**, 1061–1067.

Pereira RJ, Barreto FS, Burton RS (2014) Ecological novelty by hybridization: experimental evidence for increased thermal tolerance by transgressive segregation in *Tigriopus californicus*. *Evolution*, **68**, 204–215.

Peterson GI, Masel J (2009) Quantitative prediction of molecular clock and Ka/Ks at short timescales. *Molecular Biology and Evolution*, **26**, 2595–2603.

Peterson DL, Kubow KB, Connolly MJ *et al.* (2013) Reproductive and phylogenetic divergence of tidepool copepod populations across a narrow geographical boundary in Baja California. *Journal of Biogeography*, **40**, 1664–1675.

Pool JE, Nielsen R (2007) Population size changes reshape genomic patterns of diversity. *Evolution*, **61**, 3001–3006.

Presgraves DC (2010) The molecular evolutionary basis of species formation. *Nature Reviews Genetics*, **11**, 175–180.

Pritchard VL, Dimond L, Harrison JS *et al.* (2011) Interpopulation hybridization results in widespread viability selection across the genome in *Tigriopus californicus*. *BMC Genetics*, **12**, 54.

Rand DM, Kann LM (1996) Excess amino acid polymorphism in mitochondrial DNA: contrasts among genes from *Drosophila*, mice, and humans. *Molecular Biology and Evolution*, **13**, 735–748.

Rand DM, Haney RA, Fry AJ (2004) Cytonuclear coevolution: the genomics of cooperation. *Trends in Ecology & Evolution*, **19**, 645–653.

Ranwez V, Harispe S, Delsuc F, Douzery EJP (2011) MACSE: Multiple Alignment of Coding SEquences accounting for frameshifts and stop codons (WJ Murphy, Ed). *PLoS ONE*, **6**, e22594.

Rawson PD, Burton RS (2002) Functional coadaptation between cytochrome *c* and cytochrome *c* oxidase within allopatric populations of a marine copepod. *Proceedings of the National Academy of Sciences of the USA*, **99**, 12955–12958.

Read TD, Salzberg SL, Pop M *et al.* (2002) Comparative genome sequencing for discovery of novel polymorphisms in *Bacillus anthracis*. *Science*, **296**, 2028–2033.

Renaut S, Owens GL, Rieseberg LH (2014) Shared selective pressure and local genomic landscape lead to repeatable patterns of genomic divergence in sunflowers. *Molecular Ecology*, **23**, 311–324.

Rocha EPC, Smith JM, Hurst LD *et al.* (2006) Comparisons of dN/dS are time dependent for closely related bacterial genomes. *Journal of Theoretical Biology*, **239**, 226–235.

Romiguier J, Gayral P, Ballenghien M *et al.* (2014) Comparative population genomics in animals uncovers the determinants of genetic diversity. *Nature*, **515**, 261–263.

Schmieder R, Edwards R (2011) Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, **27**, 863–864.

Shaw TI, Ruan Z, Glenn TC, Liu L (2013) STRAW: Species TRee Analysis Web server. *Nucleic Acids Research*, **41**, W238–W241.

Smeitink J, van den Heuvel L, DiMauro S (2001) The genetics and pathology of oxidative phosphorylation. *Nature Reviews Genetics*, **2**, 342–352.

Soria-Carrasco V, Gompert Z, Comeault AA *et al.* (2014) Stick insect genomes reveal natural selection's role in parallel speciation. *Science*, **344**, 738–742.

Stern DL, Orgogozo V (2009) Is genetic evolution predictable? *Science*, **323**, 746–751.

Stoletzki N, Eyre-Walker A (2011) Estimation of the neutrality index. *Molecular Biology and Evolution*, **28**, 63–70.

Subramanian S (2004) Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome. *Genetics*, **168**, 373–381.

Swanson WJ, Clark AG, Waldrip-Dail HM, Wolfner MF, Aquadro CF (2001a) Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in *Drosophila*. *Proceedings of the National Academy of Sciences of the USA*, **98**, 7375–7379.

Swanson WJ, Zhang ZH, Wolfner MF, Aquadro CF (2001b) Positive Darwinian selection drives the evolution of several female reproductive proteins in mammals. *Proceedings of the National Academy of Sciences of the USA*, **98**, 2509–2514.

Swanson WJ, Wong A, Wolfner MF, Aquadro CF (2004) Evolutionary expressed sequence tag analysis of *Drosophila* female reproductive tracts identifies genes subjected to positive selection. *Genetics*, **168**, 1457–1465.

Tsagkogeorga G, Cahais V, Galtier N (2012) The population genomics of a fast evolver: high levels of diversity, functional constraint, and molecular adaptation in the tunicate *Ciona intestinalis*. *Genome Biology and Evolution*, **4**, 740–749.

Vittor BA (1971) *Effects of the environment on fitness-related life history characters in* Tigriopus californicus. PhD Thesis, University of Oregon, Eugene, Oregon.

Wake DB (1991) Homoplasy: the result of natural selection, or evidence of design limitations? *The American Naturalist*, **138**, 543–567.

Willett CS (2010) Potential fitness tradeoffs for thermal tolerance in the intertidal copepod *Tigriopus californicus*. *Evolution*, **64**, 2521–2534.

Willett CS (2012) Quantifying the elevation of mitochondrial DNA evolutionary substitution rates over nuclear rates in the intertidal copepod *Tigriopus californicus*. *Journal of Molecular Evolution*, **74**, 310–318.

Willett CS, Burton RS (2004) Evolution of interacting proteins in the mitochondrial electron transport system in a marine copepod. *Molecular Biology and Evolution*, **21**, 443–453.

Willett CS, Ladner JT (2009) Investigations of fine-scale phylogeography in *Tigriopus californicus* reveal historical patterns of population divergence. *BMC Evolutionary Biology*, **9**, 139.

Wolf JBW, Künstner A, Nam K, Jakobsson M, Ellegren H (2009) Nonlinear dynamics of nonsynonymous (dN) and synonymous (dS) substitution rates affects inference of selection. *Genome Biology and Evolution*, **1**, 308–319.

Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, **24**, 1586–1591.

Yang Z, Nielsen R (2000) Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Molecular Biology and Evolution*, **17**, 32–43.

R.J.P. conceived the project with direct input from F.S.B., M.C. and R.S.B. R.J.P. generated the data. R.J.P., F.S.B., N.T.P. and M.C. analyzed the data. R.J.P. wrote the first version of the manuscript; all authors contributed for the writing of the final version.

Data accessibility

Raw reads were deposited in the NCBI Sequence Read Archive (BioProject PRJNA304270; BioSample Acc. SAMN04311981-6). The dryad archive (doi: 10.5061/dryad.23s61) contains all de novo assemblies, list of homologous genes, complete polymorphism data set, alignments for mitochondrial and autosomal genes, bootstrapped alignments and resulting tree file, *p*-distance estimates, estimates of d_N, d_S, omega and DoS, read counts and FPKM estimates. Parsing scripts are available in a public GitHub repository (https://github.com/bluegenes/repeatability).

Supporting information

Additional supporting information may be found in the online version of this article.

Fig. S1 Quality of the de novo assemblies for the orthologous genes.

Fig. S2 Genetic variability sorts exponentially with time since population divergence in independent population comparisons.

Fig. S3 Potential bias on number of SNPs identified caused by differences in library size.

Fig. S4 Genome-wide protein divergence during the continuum of species formation in *Tigriopus californicus*.

Fig. S5 Null distributions of gene overlap.

Fig. S6 Pairwise estimation of adaptive divergence during the continuum of species formation in *Tigriopus californicus*.

Fig. S7 The number of genes with signature of accelerated evolution decreases with time since divergence.

Fig. S8 Correlation between gene expression and rates of protein divergence in populations of *Tigriopus californicus*.

Table S1 Data used to generate de novo assemblies and mapping for each population of *Tigriopus californicus*.

Table S2 Average genetic divergence between all pairwise population comparisons of *Tigriopus californicus*.

Table S3 Genetic diversity among populations of *Tigriopus californicus*.

Table S4 Overlaps in fast-evolving proteins across the five focal population comparisons.

Table S5 Overlaps in proteins with possible positive selection as informed by ω (d_N/d_S) across the five focal population comparisons.