Genomic signatures of mitonuclear coevolution across populations of *Tigriopus californicus*

Felipe S. Barreto1,2*, Eric T. Watson3, Thiago G. Lima2,4, Christopher S. Willett4, Suzanne Edmands3, Weizhong Li5 and Ronald S. Burton2

The copepod *Tigriopus californicus* shows extensive population divergence and is becoming a model for understanding allopatric differentiation and the early stages of speciation. Here, we report a high-quality reference genome for one population (~190 megabases across 12 scaffolds, and ~15,500 protein-coding genes). Comparison with other arthropods reveals 2,526 genes presumed to be specific to *T. californicus*, with an apparent proliferation of genes involved in ion transport and receptor activity. Beyond the reference population, we report re-sequenced genomes of seven additional populations, spanning the continuum of reproductive isolation. Populations show extreme mitochondrial DNA divergence, with higher levels of amino acid differentiation than observed in other taxa. Across the nuclear genome, we find elevated protein evolutionary rates and positive selection in genes predicted to interact with mitochondrial DNA and the proteins and RNA it encodes in multiple pathways. Together, these results support the hypothesis that rapid mitochondrial evolution drives compensatory nuclear evolution within isolated populations, thereby providing a potentially important mechanism for causing intrinsic reproductive isolation.

Intrinsic reproductive isolation occurs when gene interactions that are neutral or beneficial in their original genetic background become detrimental in a hybrid genetic background. In recent years, there has been considerable focus on the origin and identity of these genetic incompatibilities. One category of hybrid conflict found across a diversity of eukaryotes is mitonuclear incompatibility1. In the shift from free-living microbe to organelle, mitochondria transferred most of their genome to the nucleus, leaving their function dependent on a large number (~1,500) of nuclear-encoded gene products4,5 and the handful of genes retained in the mitochondrial genome. Mitochondrial DNA (mtDNA) is particularly prone to accumulating deleterious mutations due to elevated mutation rates, limited recombination and the absence of sexual reproduction. This favours a pattern of compensatory coevolution where nuclear genes must repeatedly evolve to rescue mitochondrial function. The recent proliferation of next-generation sequencing methods now allows genome-wide tests for signatures of mitonuclear coevolution across a diversity of taxa.

Here, we present a high-quality draft genome of the intertidal copepod *Tigriopus californicus* (Fig. 1A), with genomic comparisons across eight populations. Copepod genomes are important in their own right because copepods are among the most abundant organisms on the planet and play critical roles in both marine and freshwater environments. Despite their diversity and abundance, only a few copepod draft genomes have been assembled and published1-9, two of which are in the genus *Tigriopus* (*T. japonicus*8 and *T. kingsejongensis*11). The *T. californicus* genome is especially important because the species is an emerging model for understanding hybrid breakdown12,13, with particular focus on nuclear-mitochondrial coevolution14-16. Mitonuclear conflicts may be a particularly important driver of postzygotic isolation in this system due to elevated mitochondrial substitution rates (synonymous site substitution rate ~55-fold higher for mtDNA relative to nuclear DNA17) and the absence of heteromorphic sex chromosomes21, meaning the species lacks the rapidly accumulating X-autosome interactions thought to be the first incompatibilities to arise in many other systems12-14.

Results and discussion

The *T. californicus* genome is one of the most compact among copepods. A single cytometric study suggests a haploid genome of 244 megabases (Mb) (or 0.25 pg). However, we argue that this measurement may not be particularly accurate because the calibration standards used ranged from 2.5 to 6.3 pg; hence, extrapolation down to 0.5 pg (that is, the total DNA in a *T. californicus* nucleus) is probably not appropriate. Our high-coverage sequencing effort with long and short reads, as well as Hi-C technology, has consistently resulted in a total length estimate of 181–191.15 Mb (Supplementary Table 1). Moreover, based on the Benchmarking Universal Single-Copy Orthologs approach, the assembly captured 94.5% complete transcripts of the benchmarking arthropod gene set (1,007 out of 1,066) and 92.9% of the metazoan set (909 out of 978) (Supplementary Fig. 1), suggesting that the actual genome size is probably closer to 200 Mb. The missing sequence is probably enriched in repetitive DNA, which is known to hinder efforts to assemble eukaryotic genomes.

Although the assembly is fragmented into 459 scaffolds after removing those of bacterial origin (Supplementary Tables 1 and 2, and Supplementary Note), 99.3% of the sequence is contained in only 12 scaffolds that range in size from 13.38 to 18.07 Mb, with only 3.77 Mb (1.97%) of gaps within scaffolds. Markers from a linkage map17 readily identified these 12 scaffolds as the 12 autosomes, placing over 99% of the genome assembly in chromosomes. Repetitive elements occupy nearly 29% of the copepod genome, with long interspersed nuclear elements and long terminal repeats composing as much as 39% of all repeats (Supplementary Table 3). The MAKER2 pipeline18 predicted 15,646 gene models. Of these,

1Department of Integrative Biology, Oregon State University, Corvallis, OR, USA. 2Marine Biology Research Division, Scripps Institution of Oceanography, University of California, San Diego, La Jolla, CA, USA. 3Department of Biological Sciences, University of Southern California, Los Angeles, CA, USA. 4Center for Research in Biological Systems, University of California, San Diego, La Jolla, CA, USA. *e-mail: felipe.barreto@oregonstate.edu

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.
NATURE ECOLOGY & EVOLUTION

14,233 (91%) were supported by external transcriptomic data, such as BLASTN alignments of a high-coverage transcriptome assembly and mapping of RNA sequencing (RNA-Seq) reads, and a high proportion (69%) showed significant homology to annotated proteins in Swiss-Prot (Supplementary Note).

Orthology analyses identified 4,803 orthologous gene groups that are shared among T. californicus and three arthropods with sequenced genomes (Fig. 1b and Supplementary Table 12). Between the two crustaceans (that is, T. californicus and Daphnia pulex), 5,767 gene groups are shared in total, and 433 groups are shared uniquely. Moreover, 661 gene groups, containing 2,526 genes, were shared among T. californicus that are shared among D. melanogaster, A. mellifera and D. melanogaster, 500 gene groups are shared in total, and 37 gene groups are shared uniquely. A. mellifera and Daphnia pulex (Fig. 1b). Among these, 77 Gene Ontology functional terms were found to be over-represented relative to all other T. californicus genes (Supplementary Table 4). We observed a single copy in most arthropod genomes characterized, including Daphnia (Crustacea: Cladocera), but was detected as 14 copies in T. kingsejongensis and as at least 60 copies in T. californicus (Supplementary Note, Supplementary Table 2, and Supplementary Tables 12, 13, 15 and 17).

The T. californicus system provides an outstanding opportunity to investigate mechanisms of evolutionary divergence along the full continuum of speciation, as populations along its distribution exhibit varying levels of reproductive incompatibility when hybridized. The genome sequences of seven additional T. californicus populations (Fig. 1c) were assembled by mapping Illumina reads from each population to the San Diego reference and reconstructing consensus sequences. Therefore, these assemblies do not capture structural differences among populations, but provide abundant sequence information for protein-coding regions (Supplementary Table 5).

We found extreme levels of mitochondrial divergence between natural populations of T. californicus, including one (San Roque) within a clade being described as the new species T. bajaensis. Nucleotide divergence across the mitochondrial genome ranged from 9.5 to 26.5% relative to the reference San Diego mitochondrial genome, with an average of 19.6% across all populations (Supplementary Table 6). Among the 37 mitochondrial genes (Fig. 2a), the 13 protein-coding genes have the highest average population divergence. The 22 transfer RNA (tRNA) genes were also found to be highly variable across populations, ranging from 0 to 30% nucleotide divergence, with an average of 9%.

To investigate the evolutionary forces shaping differentiation, we examined deviations from neutrality using two statistical tests (the ratio of the rate of non-synonymous to the rate of synonymous nucleotide substitutions (dN/dS) and the direction of selection (DoS)). The mitochondrial genome-wide dN/dS ratio is well below 1 at 0.326, suggesting purifying selection as the principal force in mitochondrial evolution. Despite this, mitochondrial dN/dS in T. californicus is among the highest seen in animal mtDNA (1,855 species). Analysis of nucleotide polymorphism confirms the broad pattern of purifying selection, revealing an average excess of amino acid polymorphism relative to divergence (DoS = −0.34).

A population comparison of individual protein-coding genes revealed that despite the overall pattern of purifying selection across the mitochondrial genome, some genes contain an excess of amino acid divergence (dN/dS > 1; DoS > 0; Fig. 2b). This suggests that despite the net evolutionary effects of linkage in a non-recombining genome and the cumulative action of purifying selection against amino acid polymorphisms, a handful of genes show evidence for diversifying selection (atp6, nad3, nad5 and nad6). Protein-coding genes for the different respiratory chain complexes experience different selection pressures within and between populations (dN/dS; Kruskal–Wallis rank sum test, P < 10^{-6}; DoS: P = 0.001; Supplementary Tables 7 and 8). We find that dN/dS and DoS values are generally higher across complex I genes compared with other respiratory complexes (dN/dS; Mann-Whitney U-test, PUW < 10^{-6}; DoS: PD < 0.001; Fig. 2c), and we detected a strong signature of positive selection (ω = 4.39) on ~2% of the codons of one of the complex I genes (nd4l; Table 1). Despite showing low overall levels of dN/dS (Fig. 2c), an mtDNA-encoded component of complex IV (mt-co3) also exhibited
Fig. 2 | Mitochondrial divergence and evolution. a, Pairwise nucleotide divergence ($D_{st}$) for re-sequenced populations relative to the San Diego reference across all mtDNA-encoded genes. b, Estimates of $d_{st}/d_{is}$ (top) and DoS (bottom) relative to the San Diego reference across the 13 protein-coding genes. Missing data are represented as grey squares. AB, Abalone Cove; BB, Bodega Bay; BR, Bird Rock; CAT, Catalina Island; FHL, Friday Harbor; SC, Santa Cruz; SRQ, San Roque. c, $d_{st}/d_{is}$ (top) and DoS (bottom) for protein subunits of the OXPHOS complexes. For each complex, values were pooled across the seven interpopulation alignments. Superscript, lower-case letters above the bars denote the results of Mann-Whitney U-tests, followed by correction for multiple testing; complexes not sharing a letter are significantly different (FDR < 5%). Boxplots show the median, lower and upper quartiles, with whiskers extending to 1.5 times the interquartile range; single points are data beyond this range.

Table 1 | Evidence of positive selection in mitochondrial proteins involved in translation and oxidative phosphorylation

| Predicted protein                        | Gene abbreviation | $d_{st}/d_{is}$ | Likelihood ratio | Parameter estimates | P value |
|-----------------------------------------|-------------------|-----------------|-----------------|---------------------|---------|
| NADH dehydrogenase subunit 4L           | nd4l              | 0.049           | 7.442           | $p_s = 0.019, \omega_s = 4.39$ | 0.0242  |
| Cytochrome c oxidase 3                   | mt-co3            | 0.028           | 7178            | $p_s = 0.005, \omega_s = 1.36$ | 0.0276  |
| 28S ribosomal protein S31               | mrps31            | 0.275           | 13.631          | $p_s = 0.0025, \omega_s = 2.77$ | 0.00110 |
| 28S ribosomal protein S2                | mrps2             | 0.237           | 12.688          | $p_s = 0.0032, \omega_s = 3.66$ | 0.00175 |
| 28S ribosomal protein S35               | mrps35            | 0.106           | 9.892           | $p_s = 0.0163, \omega_s = 3.02$ | 0.00711 |
| 39S ribosomal protein L46               | mrpl46            | 0.272           | 9.828           | $p_s = 0.0129, \omega_s = 4.58$ | 0.00734 |
| NADH dehydrogenase [ubiquinone] 1 beta subcomplex subunit 8 | ndufb8 | 0.173           | 10.88           | $p_s = 0.0157, \omega_s = 4.73$ | 0.00434 |
| Cytochrome c oxidase subunit 4 isofrom 1 | cox4i1            | 0.155           | 10.561          | $p_s = 0.0054, \omega_s = 3.9$ | 0.00510 |
| ATP synthase subunit epsilon             | atpSe             | 0.497           | 12.201          | $p_s = 0.094, \omega_s = 7.38$ | 0.00224 |
| ATP synthase subunit beta                | atpsynbeta        | 0.019           | 9.628           | $p_s = 0.012, \omega_s = 2.23$ | 0.00812 |
| ATP synthase subunit f                   | atpSj2            | 0.151           | 9.360           | $p_s = 0.027, \omega_s = 6.98$ | 0.00928 |

*Overall ratio across sites and branches (from model M0). *Comparison of models M7 and M8. *From model M8; $p_s$ is the proportion of amino acid sites for which $d_{st}/d_{is}$ is estimated to be >1 and which achieved a $d_{st}/d_{is}$ value of $\omega_s$. *From a comparison of the likelihood ratio with a $\chi^2$ distribution. After correction for multiple testing, all genes have an FDR of <10%.

statistical evidence for positive selection, albeit at a moderate level ($\omega = 1.36$; Table 1).

We examined rates of coding sequence evolution via $d_{st}/d_{is}$ levels in 13,538 predicted nuclear genes across the eight T. californicus population genomes. We used these estimates to test the hypothesis of compensatory evolution, which predicts that nuclear-encoded proteins that interact with mtDNA-encoded elements will show elevated rates of amino acid changes. Specifically, we performed
four sets of comparisons: (1) mitochondrial versus cytosolic counterparts within ribosomal protein (RP) genes and (2) within aminoacyl tRNA synthetases (ARSs); (3) oxidative phosphorylation (OXPHOS) complexes I, II, IV and V versus complex II; and (4) within mitochondrially targeted proteins (MTPs), all genes in pathways predicted to interact with mtDNA (OXPHOS, mitochondrial RP (mt-RP), mitochondrial ARS (mt-ARS), mtDNA replication, transcription and elongation/initiation factors) versus all genes not in these groups (that is, those that probably form no interaction with any mtDNA-encoded product).

Our analyses revealed elevated rates of amino acid substitution in proteins that putatively interact with mtDNA-encoded products across multiple mitochondrial pathways. Consistent with a previous study of 2 T. californicus populations32, mt-RPs exhibited $d_\text{s}/d_\text{d}$ levels that were 6.5 times higher than those of cytosolic RPs (cyt-RPs; Mann-Whitney U-test, $P = 5.87 \times 10^{-6}$; Fig. 3a). In this case, after accounting for the markedly different levels of gene expression ($t = 6.8$, d.f. = 16, $P = 1.73 \times 10^{-5}$; Supplementary Fig. 3), $d_\text{s}/d_\text{d}$ among the two groups did not differ significantly (ANOVA, $P = 0.136$). Therefore, we cannot exclude the hypothesis that elevated $d_\text{s}/d_\text{d}$ in mt-ARS is due to relaxed functional constraints. Similar results for ARS have been observed in flies and chickens36.

Within the OXPHOS pathway, a meaningful test of our hypothesis is the comparison of evolutionary rates between complex II and each of the other four complexes, since complex II is the only complex that does not contain an mtDNA-encoded subunit, and expression levels of complex II genes were not significantly different from those of the other complexes ($t$-tests, all $P > 0.183$; Supplementary Fig. 3). Our results are consistent with a pattern of compensatory evolution of nuclear-encoded OXPHOS components: complexes I, III and IV showed elevated $d_\text{s}/d_\text{d}$ compared with complex II ($P = 0.018$, 0.040 and 0.013, respectively), although the difference between complexes II and V was not significant ($P = 0.214$) (Fig. 3c).

Finally, our annotation of nuclear-encoded MTPs allowed us to separate 2 broad groups of proteins within the organelle: nuclear-encoded proteins in pathways that involve mtDNA-encoded products ($n = 147$ and those in pathways that do not interact with such products ($n = 458$). This comparison includes proteins across multiple pathways and functions, considering only whether they putatively interact with mtDNA-encoded elements. We found that mtDNA-interacting proteins showed, on average, significantly higher molecular evolutionary rates than those that do not interact with mtDNA-encoded elements ($P = 8.37 \times 10^{-2}$; Fig. 4), despite possible constraints as a result of them having higher transcription levels ($t = 5.2$, d.f. = 261, $P < 4.1 \times 10^{-6}$; Supplementary Fig. 3).

To allow $d_\text{s}/d_\text{d}$ to vary among codons and scan the genomes for patterns of amino acid changes consistent with positive selection, we applied the CODEML ‘sites’ model15 across the 13,538 genes.
This analysis revealed 775 genes (530 with BLAST annotation) with significant evidence for positive selection (false discovery rate (FDR) < 10%; Supplementary Table 9). These genes were distributed among many functional categories, with no over-representation of any Gene Ontology term. Nevertheless, we detected strong evidence for compensatory evolution to mtDNA divergence in adenosine triphosphate (ATP) synthesis and translation machineries, with four mt-RPs and five nuclear-encoded OXPHOS genes showing signatures of positive selection for at least one amino acid site (Table 1). Three of the OXPHOS genes belong to complex V, and one to each of complexes I and IV, consistent with elevated evolutionary rates of mtDNA-encoded proteins within these complexes. Coadaptation of these genes to interacting mtDNA OXPHOS proteins within populations is consistent with the highly reduced enzymatic activity of these complexes when divergent genomes are recombined. The five OXPHOS genes highlighted here are therefore strong candidates for future functional studies.

As mt-RPs associate closely with mtDNA-encoded ribosomal RNA (rRNA) in the assembly of mitochondrial ribosomes, adaptive evolution of mt-RP in response to rapid evolution in the rRNA may result in functional ‘mismatches’ between these interacting components in recombinant hybrids. We predict that low-fitness inter-population hybrids will have mismatched mitonuclear genotypes at one or more of these mt-RPs, and at the cellular level, mtDNA gene translation may be strongly affected. Such a disruption of mitochondrial translation machinery was detected in Drosophila from an incompatibility between an mt-ARS and its mtDNA-encoded tRNA in T. californicus, we found no evidence of adaptive differentiation in mt-ARS (all P > 0.61), and we hypothesize that mt-RP coadaptation may play a larger role among hybrid incompatibilities in this species.

With the recent availability of genomic resources for two other Tigriopus species, we assessed patterns of coding sequence evolution on each lineage. While the ‘sites’ model examined above detects diversifying selection across all taxa in the phylogeny, ‘branch-sites’ models permit the detection of adaptive evolution that may have occurred episodically on specific branches. A total of 4,863 orthologs were analysed using ‘branch-sites’ in CODEML, with each of the six branches serving as a foreground branch (Supplementary Table 10). Branches leading to T. japonicus and T. kingsejongensis exhibited totals of 52 and 34 positively selected genes, respectively, with only one in each species interacting with mtDNA (Fig. 5). Notably, both of these mtDNA-interacting genes were associated with OXPHOS complex I, but they were different (ndufb9 in T. japonicus, and ndufa3 in T. kingsejongensis). No Gene Ontology terms were over-represented on branches leading to either of these two species. The two T. californicus populations examined exhibited only one and five positively selected genes on the branches separating these lineages. However, the branch leading to T. californicus showed evidence of positive selection on 591 genes (Fig. 5), and these were enriched with DNA metabolic processes such as replication (GO:00062621), recombination (GO:0006310), DNA repair (GO:0006281), and cellular response to DNA damage (GO:0006974) (Supplementary Table 11). Moreover, 28 nuclear-encoded MTPs (GO:0006281), and cellular response to DNA damage (GO:0006974) exhibited positive selection, 7 of which are mtDNA-interacting tRNA. In T. californicus, we found no evidence of adaptive differentiation in mt-ARS (all P > 0.61), and we hypothesize that mt-RP coadaptation may play a larger role among hybrid incompatibilities in this species.

We detected strong signatures of adaptive molecular evolution in mt-RPs and OXPHOS proteins. Although positive selection in these gene groups has been observed during the evolution of other arthropod species, our study shows that these nuclear-encoded mitochondrial proteins are evolving rapidly among conspecific populations. We argue that adaptive evolution of these proteins represents compensatory changes within each population in response to its rapidly changing mtDNA, and that this divergence can contribute to nascent reproductive isolation among populations.

**Methods**

**Tissue sampling and DNA isolation.** T. californicus of mixed sex and life stages were collected from high intertidal rocky pools in San Diego (32°44′44″N, 117°15′18″W) and kept en masse in 11 beakers with filtered seawater (0.2 μm pore filter) at 20°C. To reduce the contribution of gut contents and bacteria to the DNA extract, the seawater was treated with an antibiotic mixture (ampicillin, 200 mg l⁻¹; streptomycin, 50 mg l⁻¹; spectinomycin, 50 mg l⁻¹; kanamycin, 50 mg l⁻¹; penicillin, 200 units ml⁻¹; neomycin, 50 mg l⁻¹; and chloramphenicol, 20 mg l⁻¹) and the animals were moved to clean filtered seawater daily for 3 days, without food, before extraction. DNA was isolated from groups of hundreds of copepods using a Qiagen DNeasy Blood and Tissue Kit, and samples were pooled to ~100 μg of DNA.

**Genome sequencing and assembly.** Cofactor Genomics generated 3 short-insert libraries (200, 500 and 800 bp) and 2 long-insert libraries (2 and 5 kb), and then barcoded and sequenced the libraries in 1 lane using the Illumina HiSeq 2000 platform to obtain 2x100 bp paired-end (short inserts) or mate-paired sequences (Supplementary Table 1). Illumina sequences from the five libraries were first filtered using an internal quality control script to remove the low-quality reads. For the 200 and 500 bp insert libraries, which have high coverage, the quality...
control parameters were more stringent. All raw reads with unknown base ‘N’ were removed. In addition, raw reads were removed if the sum of per-base error probability across all bases was greater than two. For other libraries, whose coverage of repetitive reads with ‘N’ were not filtered, if the sum of per-base error probability was greater than three. We only kept read pairs for which both read ends passed the quality filter. Duplicate read pairs were identified by CD-HIT-DUP in the CD-HIT package with the parameters ‘-u 50 –e’. Here, the read pairs having near-identical sequences with up to 2 mismatches within the first 50 bases on both ends were considered duplicated. The duplicated read pairs with lower-quality scores were removed. The processed reads from these five libraries were assembled using AllPaths-LG (release 46212) following AllPaths’ standard workflow, including initial input file preparation and final assembly runs. The assembled scaffold sequences were further extended using the GapCloser programme in the SOAPdenovo package.

To improve contiguity of the *T. californicus* assembly, proximity ligation libraries based on the Chicago and Hi-C methods were prepared by Dovetail Genomics as described previously. Briefly, the Chicago method started with high-molecular-weight DNA (~75kb) while the Dovetail Hi-C method started with intact chromosomes in fresh copepod tissue, and chromatin was reconstituted in vitro and cross-linked with formaldehyde. While cross-linked, chromatin then underwent several steps of endonuclease digestion, biotinylation and ligation to preserve information on proximity and the association of fragments. DNA was then purified and fragmented for traditional Illumina library preparation and HiSeq10 sequencing. Finally, the Chicago sequence data were used to directly prepare the AllPaths assembly via the HiRise computational pipeline, and the Hi-C data were then used to increase contiguity of the AllPaths plus Chicago assembly.

In an effort to fill assembly gaps, 1.68 million PacBio reads were obtained for the San Diego population and used to improve the final HiRise assembly using PBjelly. Pacbio sequencing was performed by the University of North Carolina High Throughput Genomic Sequencing Facility at first using size selection to isolate larger fragments of DNA template. The reads were first filtered by mapping them to the putative bacteria scaffolds from our assembly (see Supplementary Note) and removing those that appeared to be bacterial. Over 1.5 million PacBio reads with a mean read length of 5,470 bp remained. Next, these filtered reads were mapped to the HiRise assembly using BLASR via PBjelly, and gaps were filled or ends extended for regions for which there was a depth of at least ten PacBio reads. Overfilled gaps (gaps that were larger than the estimated size in the HiRise assembly) were set to an arbitrary size of 51 bp when they could not be completely filled with this set of PacBio reads at this minimum coverage threshold of 10. The resulting assembly then went through two more rounds of gap filling and end extension using this procedure, yielding our current *T. californicus* assembly, SDv2.1. These three rounds of PBjelly dropped the percentage of Ns from 4.67 to 1.97%, decreased the number of contigs in scaffolds from 7,105 to 4,789, and increased the overall size of the assembly by 4.7 megabase pairs (Mb) while increasing the sequence in contigs by 9.7 Mb.

**Genome annotation.** We used the pipeline in MAKER2 (ref. 31) to annotate putative protein-coding regions of the *T. californicus* assembly. First, we generated a de novo repeat library for *T. californicus* using the programme RepeatModeler44 and its integrated tools (RECON, TRF and RepeatScout). We also trained two ab initio gene predictors: AUGUSTUS50 and SNAP51, using a high-quality subset of protein-coding regions of *Drosophila melanogaster* and *Apis mellifera*, and a branchiopod crustacean (*D. pulex*) (Supplementary Table 12). To assess orthology among many taxa (Supplementary Table 13), we used OrthoMCL to assign *T. californicus* proteins to pre-clustered orthologous groups curated in OrthoMCL DB version 5 (ref. 56), which contains proteomes of 150 taxa. For both analyses, BLASTP e-value threshold and inflation values were left at default (e=0 and 1.5, respectively).

**Sequencing and assembly of additional *T. californicus* populations.** Copepods were collected from intertidal pools in San Roque (25° 13' 12" N, 114° 23' 52" W), Bird Rock (32° 48' 54" N, 117° 16' 23" W), Abalone Cove (33° 44' 16" N, 118° 22' 31" W), Catalina Island (33° 26.8' N, 118° 28.6' W), Santa Cruz (36° 56' 58" N, 122° 02' 49" W), Bodega Bay (38° 19' 4" N, 123° 4' 23" W) and Friday Harbor (48° 32' 47" N, 123° 0' 35" W). Animals were maintained as described previously. For each population, 250–300 adult individuals were pooled and DNA was isolated using the Qagen DNeasy Blood and Tissue Kit or a phenol–chloroform procedure.

To improve contiguity of the *T. californicus* assembly, we prepared additional Mate Pair libraries by Dovetail Genomics. We used the pipeline in MAKER2 (ref. 28) to annotate genomes, and to detect possible gene family expansions in the copepod. In OrthoVenn55, orthology was assessed among the copepod, two insects (48° 32′′ N, 114° 23′′ W) and a branchiopod crustacean (48° 32′′ N, 114° 23′′ W). Animals were maintained as described previously. For each population, 250–300 adult individuals were pooled and DNA was isolated using the Qagen DNeasy Blood and Tissue Kit or a phenol–chloroform procedure.

For each of the eight *T. californicus* populations, we used the ‘extractfeat’ script within GenomeTools to extract the coding sequence from all genes using genomic coordinates from the gff files. We filtered out genes with too many missing data (‘N’/s) by excluding those that had less than 150 bp of sequence, and retained only the longest isoform of each. We checked that the extracted sequences corresponded to the correct coding sequence using, with frame +1, by aligning them with BLASTX to the San Diego protein models.

We used the programme PRANK-codon57 to align sequences for each of the 14,000 orthologous coding sequences recovered (Supplementary Table 5). Poorly aligned regions were then removed using Gblocks version 0.91b (ref. 52) with the following parameters: -t e -b 5 -b2 5 -b3 0 -v 0 -b4 9 -b5 0 -h. Alignments shorter than 150 bp were removed, leaving a total of 15,610 coding sequence alignments for analysis. For each alignment, we estimated overall d/ds using the model M0 in the programme CODEML within PAML version 4.7 (ref. 52), with a well-supported ‘species’ tree estimated in RAXML. We used these gene-wide d/ds estimates to test for general patterns consistent with the hypothesis of compensatory evolution. Using DNA sequence data as above, we quantified transcription levels across the San Diego genes, and used ANCOVAs to account for expression in the different environments.

We screened the genome for signals of positive selection on amino acid changes by employing codon-level evolutionary models of nucleotide substitution implemented in CODEML. We tested the ‘sites’ hypothesis by applying models M7 and M8 on each coding sequence alignment, and compared the resulting log-likelihood ratios with a chi-squared distribution with two degrees of freedom. After the analysis, we applied an FDR correction for multiple testing. Further details and parameters are included in the Supplementary Note.

**Molecular evolution and population genetic analysis of mtDNA.** Multiple sequence alignment of each mtDNA gene was done in MEGA6 (ref. 71) by first aligning translated coding sequences according to the invertebrate mitochondrial code (transl_table = 5) and then replacing the amino acids with the original codons. Distance matrices of non-synonymous and synonymous substitutions were estimated in MEGA6 using the Nei–Gojobori model. Substitution distance
matrices for non-protein-coding genes were calculated from CLUSTAL multiple alignments using the Analyses of Phylogenetics and Evolution R package 11. To estimate nucleotide diversity, mitochondrial reads from the re-sequenced populations were mapped back to their mitochondrial genome assembly using BReadMap. Sequence pileups were created using SAMtools14. 4 bp regions surrounding indels were masked to avoid false positive SNPs and the pileups were re-sampled to equal depth of coverage (200x). High-quality SNPs (Phred > 20) were used to estimate mean Tajima’s τ, weighted for gene length, for each protein-coding gene using PoPoolation17. DoS was measured for each protein-coding gene according to Stoletzki and Eyre-Walker18. Finally, CODEML models M0, M7 and M8, as above, were also applied to the 13 protein-coding mtDNA genes.

Patterns of positive selection across the genus Tigriopus. We downloaded the annotated gene set of T. kingsejongensis29 and RNA-Seq raw data for T. japonicus (accession numbers SAMN05933003 and SAMN05933004) generated previously 1. We assembled a de novo transcriptome for T. japonicus using Trinity29, and retrieved the correct reading frames using custom scripts. To identify orthologues, we performed reciprocal BLASTX searches among the three species. We identified 5,351 orthologues. For analyses of positive selection, we were interested in testing patterns of adaptive evolution across different lineages of Tigriopus species, including T. japonicus, T. kingsejongensis, our reference T. californicus from San Diego and the reproductively isolated T. californicus from San Roque. Therefore, the orthologues identified above were aligned across the four taxa, then quality-trimmed with TRANKS26 and GBlocks, as above. We removed any alignments that were shorter than 150 bp. Using the maximum likelihood method of Yang and Nielsen78, we computed pairwise log-likelihood ratio between a model allowing for positive selection of the form

\[
\frac{N_d}{N_s} = \theta
\]

in the absence of heteromorphic sex chromosomes? BioEssays 34, 166–169 (2012).

22. Rieseberg, L. H. Chromosomal rearrangements and speciation. Trends Ecol. Evol. 16, 351–358 (2001).

23. Phillips, B. C. & Edmaiers, S. Does the speciation clock tick more slowly in the absence of heteromorphic sex chromosomes? BioEssays 34, 166–169 (2012).

24. Lima, T. G. Higher levels of sex chromosome heteromorphism are associated with markedly stronger reproductive isolation. Nat. Comm. 5, 4743 (2014).

25. Gregory, T. R. Animal Genome Size Database (2015): http://www.genomesize.com

26. Wyngaard, G. A. & Rasch, E. M. Patterns of genome size in the Copepoda. Hydrobiologia 417, 43–56 (2000).

27. Pál, C., Papp, B. & Hurst, L. Highly expressed genes in yeast evolve slowly. Genetics 158, 927–931 (2001).

28. Holt, C. & Yangell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics 12, 491 (2011).

29. Barreto, F. S., Pereira, R. J. & Burton, R. S. Hybrid dysfunction and physiological compensation in gene expression. Mol. Biol. Evol. 32, 613–622 (2015).

30. Peterson, D. L. et al. Reproductive and phylogenetic divergence of tidepool copepod populations across a narrow geographical boundary in Baja California. J. Biogeogr. 40, 1664–1675 (2013).

31. Bazin, E., Glémis, S. & Galitzer, N. Population size does not influence mitochondrial genetic diversity in animals. Science 312, 570–571 (2006).

32. Barreto, F. S. & Burton, R. S. Evidence for compensatory evolution of ribosomal proteins in response to rapid divergence of mitochondrial rRNA. Mol. Biol. Evol. 30, 310–314 (2013).

33. Pál, C., Papp, B. & Hurst, L. Highly expressed genes in yeast evolve slowly. Genetics 158, 927–931 (2001).

34. Drummond, D. A., Blyth, J. D., Adams, C., Wilke, C. O. & Arnold, F. H. Why highly expressed proteins evolve slowly. Proc. Natl Acad. Sci. USA 102, 14338–14343 (2005).

35. Adron, J. R., White, P. S. & Montooth, K. L. The roles of compensatory evolution and constraint in aminocyl tRNA synthetase evolution. Mol. Biol. Evol. 33, 152–161 (2015).

36. Yang, Z., Nielsen, R., Goldman, N. & Pedersen, A. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155, 431–449 (2000).

37. Ellison, C. K. & Burton, R. S. Disruption of mitochondrial function in interpopulation hybrids of Tigriopus californicus. Evolution 60, 1382–1391 (2006).

38. Barreto, F. S., Schoville, S. D. & Burton, R. S. Reverse genetics in the tidepool: knock-down of target gene expression via RNA interference in the copepod Tigriopus californicus. Mol. Ecol. Resour. 15, 868–879 (2014).

39. Weil, J. E. & Labrador, M. C. et al. Acanthoecyly: incompatibility between a mitochondrial rRNA and its nuclear-encoded tRNA synthetase compromises development and fitness in Drosophila. PLoS Genet. 9, e1003238 (2013).

40. Zhang, J. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol. Biol. Evol. 22, 2472–2479 (2005).
42. Roux, J. et al. Patterns of positive selection in seven ant genomes. *Mol. Biol. Evol.* 31, 1661–1685 (2014).
43. Putnam, N. H. et al. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res.* 26, 342–350 (2016).
44. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326, 289–293 (2009).
45. English, A. C. et al. Mind the gap: upgrading genomes with Pacific Biosciences RS Long-Read Sequencing Technology. *PLoS ONE* 7, e47768 (2012).
46. Smit, A. & Hubley, R. *RepeatModeler Open-1.0* (2014); http://www.repeatmasker.org/RepeatModeler/
47. Bao, Z. & Eddy, S. R. Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res.* 12, 1269–1276 (2002).
48. Benson, G. *Tandem repeats finder: a program to analyze DNA sequences.* *Nucleic Acids Res.* 28, 57–60 (2000).
49. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079 (2009).
50. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* 26, 589–595 (2010).
51. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079 (2009).
52. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetic parameter estimation from sequencing data. *Bioinformatics* 27, 2987–2993 (2011).
53. Burton, R. S., Byrne, R. J. & Rawson, P. D. Three divergent mitochondrial genomes from California populations of the copepod *Tigriopus californicus*. *Gene* 403, 53–59 (2007).
54. Bushnell, B. *BBMap* (2014); http://www.repeatmasker.org/RepeatModeler/
55. Hahn, C., Bachmann, L. & Chevreux, B. Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads—a bating and iterative mapping approach. *Nucleic Acids Res.* 41, e129 (2013).
56. Bernt, M. et al. MITOS: improved de novo metazoan mitochondrial genome annotation. *Mol. Phylogenet. Evol.* 69, 313–319 (2013).
57. Greemse, G., Steinbiss, S. & Kurtz, S. *GenomeTools*: a comprehensive software library for efficient processing of structured genome annotations. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 10, 645–656 (2013).
58. Löytynoja, A. & Goldman, N. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science* 320, 1632–1635 (2008).
59. Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* 17, 540–552 (2000).
60. Yang, Z. *PAML 4: phylogenetic analysis by maximum likelihood.* *Mol. Biol. Evol.* 24, 1586–1591 (2007).
61. Stamatakis, A. *RAxML version 8*: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313 (2014).
62. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.* 57, 289–300 (1995).
63. Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. *MEGA6*: Molecular Evolutionary Genetics Analysis Version 6.0. *Mol. Biol. Evol.* 30, 2725–2729 (2013).
64. Nei, M. & Gojobori, T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* 3, 418–426 (1986).
65. Paradis, E. *Analysis of Phylogenetics and Evolution Using R* (Springer, New York, 2012).
66. Stoletzki, N. & Eyre-Walker, A. Estimation of the neutrality index. *Mol. Biol. Evol.* 28, 63–70 (2011).
67. Grabherr, M. G. et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* 29, 644–652 (2011).
68. Yang, Z. & Nielsen, R. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* 17, 32–43 (2000).

**Acknowledgements**

This work was supported by US National Science Foundation grants (IOS1154321 to S.E.; IOS1155030 to R.S.B.; and IOS1155325 to C.S.W.) and Oregon State University faculty startup funds to F.S.B. The authors thank S. Morgan and R. J. Pereira for help with sample collection.

**Author contributions**

F.S.B., E.T.W., T.G.L., C.S.W., S.E. and R.S.B. contributed to data interpretation and writing of the manuscript. F.S.B., E.T.W., T.G.L., C.S.W., S.E. and R.S.B. contributed to design of the project, collection of biological samples, and sequence data acquisition. W.L. contributed to initial genome sequence assembly. F.S.B., E.T.W. and C.S.W. contributed to genome annotation. F.S.B., E.T.W., T.G.L. and C.S.W. contributed to computational and statistical analyses. F.S.B., E.T.W., T.G.L., C.S.W., S.E. and R.S.B. contributed to data interpretation and writing of the manuscript.

**Competing interests**

The authors declare no competing interests.

**Additional information**

Supplementary information is available for this paper at https://doi.org/10.1038/s41559-018-0588-1.

Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to F.S.B.

**Publisher’s note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.
Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a | Confirmed
---|---
☐ | The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
☐ | An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐ | The statistical test(s) used AND whether they are one- or two-sided
☐ | Only common tests should be described solely by name; describe more complex techniques in the Methods section.
☐ | A description of all covariates tested
☐ | A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐ | A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐ | For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted
☐ | Give P values as exact values whenever suitable.
☐ | For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐ | For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐ | Estimates of effect sizes (e.g. Cohen’s d, Pearson’s r), indicating how they were calculated
☐ | Clearly defined error bars
☐ | State explicitly what error bars represent (e.g. SD, SE, CI)

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection

All data in our manuscript are of high-throughput sequencing methods, and were collected using the standard industry software.

Data analysis

For all genetic analyses, including genome assembly and annotation, we used published and well-established software packages, for which details are provided in the main methods and/or in the supplementary note.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Raw Illumina and PacBio DNA sequencing reads have been deposited in the NCBI Sequence Read Archive database (Illumina reads for the SD reference: SRX469409 - 469413; for the additional populations: SRX2746698 - 2746704; PacBio reads for SD: SX3778522- SX3778523). The T. californicus annotated genome assembly...
was deposited in the i5k Workspace of the National Agricultural Library (US Department of Agriculture), where it can be browsed, searched, and downloaded: 
https://i5k.nal.usda.gov/Tigriopus_californicus

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

| Study description | This study is aimed at describing the genome of a species that is a model for allopatric speciation and mitochondrial-nuclear genetic interactions, and then to use the genome to test hypotheses regarding the molecular evolutionary outcomes of mito-nuclear divergence across the species and the genus. |
| Research sample | We performed high-throughput sequencing and assembly of multiple divergent populations of our study copepod species *Tigriopus californicus*. These populations are geographically isolated and distributed along the Pacific coast of Northern America, and each population represents a distinct gene pool with unique genetic variants. |
| Sampling strategy | No experiments were performed in this study. Our sampling was simply aimed at querying genetic diversity along most of the geographic range of this species. Sample sizes in each of our statistical analyses is always determined intrinsically by the number of genes in each tested group. Whenever possible, all genes in each group were used. |
| Data collection | No experiments were performed in this study. |
| Timing and spatial scale | No experiments were performed in this study. We sampled tissue from multiple populations across the species range over the course of 2 years, and specimens used in DNA sequencing were acclimated to lab culture conditions. |
| Data exclusions | In all of our analyses, we aimed to include all genes available. In each analysis, we had a criterion of minimum alignment length of 150 bp. Therefore, genes that were excluded in each analyses were so because they were too short in one or more of the compared populations. |
| Reproducibility | No experiments were performed in this study. |
| Randomization | No experiments were performed in this study, therefore no randomization was applied. |
| Blinding | No experiments were performed in this study. All hypotheses tested relied on first categorizing genes according to their protein function, and hence blinding is not possible. |

Did the study involve field work? ☒ Yes ☐ No

Field work, collection and transport

| Field conditions | Field work for this study was performed solely for rapid collection of live copepods for lab culturing and eventual sequencing. No field data are used in our study, and none were recorded. |
| Location | All collection locations are given by name and coordinates in the main manuscript (line 254 and 344-347). |
| Access and import/export | Collection of this copepods species is well-established. They inhabit high rocky pools, and are away from the main tide pools. Collections were done by simply dipping a small aquarium net (3in x 3in) into the pools and transferring live copepods into collection bottles that are easily transported to lab. |
| Disturbance | As described above, collection of these copepods imposes no disturbance to the habitat. |

Reporting for specific materials, systems and methods
Materials & experimental systems

Methods

Involved in the study

- Unique biological materials
- Antibodies
- Eukaryotic cell lines
- Palaeontology
- Animals and other organisms
- Human research participants

ChIP-seq
Flow cytometry
MRI-based neuroimaging

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals

For laboratory animals, report species, strain, sex and age OR state that the study did not involve laboratory animals.

Wild animals

Provide details on animals observed in or captured in the field; report species, sex and age where possible. Describe how animals were caught and transported and what happened to captive animals after the study (if killed, explain why and describe method; if released, say where and when) OR state that the study did not involve wild animals.

Field-collected samples

Animals used in this study were minute invertebrates that are NOT under any regulations regarding animal care and maintenance. Individuals used in this study were field-collected (as described above) and maintained in lab for multiple generations, but were not experimentally treated or hybridized in any way, and hence likely represent wild types. Maintenance was standard for this system, which involves keeping high-density cultures in beakers with seawater and food ad libitum.