UC San Diego

UC San Diego Previously Published Works

Title

Transcriptome-wide patterns of divergence during allopatric evolution.

Permalink

<https://escholarship.org/uc/item/15j8f15v>

Journal

Molecular ecology, 25(7)

ISSN

0962-1083

Authors

Pereira, Ricardo J
Barreto, Felipe S
Pierce, N Tessa
et al.

Publication Date

2016-04-01

DOI

10.1111/mec.13579

Peer reviewed

1Title:

2Transcriptome-Wide Patterns of Divergence During Allopatric Evolution

3

4**Authors:** Ricardo J. Pereira^{1,2}, Felipe S. Barreto^{1,3}, N Tessa Pierce¹,

5Miguel Carneiro⁴, Ronald S. Burton¹

6

7Affiliations:

81- Marine Biology Research Division, Scripps Institution of

9Oceanography, University of California, San Diego, La Jolla, CA, 92093-

100202, USA

112- Centre for GeoGenetics, Natural History Museum of Denmark,

12University of Copenhagen, Øster Voldgade 5-7, 1350, Copenhagen,

13Denmark

143- Department of Integrative Biology, Oregon State University,

15Corvallis, OR, 97331,USA

164- CIBIO, Centro de Investigação em Biodiversidade e Recursos

17Genéticos, Campus Agrário de Vairão, 4485-661, Vairão, Portugal.

18Abstract

19Recent genomic studies have revealed repeated patterns of genomic
20divergence associated with species formation. Such patterns suggest
21that natural selection tends to target a set of available genes, but is
22also indicative that closely related taxa share evolutionary constraints
23that limit genetic variability. Studying patterns of genomic divergence
24among populations within the same species may shed light on the
25underlying evolutionary processes. Here, we examine transcriptome-
26wide divergence and polymorphism in the marine copepod *Tigriopus*
27*californicus*, a species where allopatric evolution has led to replicate
28sets of populations with varying degrees of divergence and hybrid
29incompatibility. Our analyses suggest that relatively small effective
30population sizes have resulted in an exponential decline of shared
31polymorphisms during population divergence and also facilitated the
32fixation of slightly deleterious mutations within allopatric populations.
33Five interpopulation comparisons at three different stages of
34divergence show that non-synonymous mutations tend to accumulate
35in a specific set of proteins. These include proteins with central roles in
36cellular metabolism, such as those encoded in mtDNA, but also include
37an additional set of proteins that repeatedly show signatures of
38positive selection during allopatric divergence. Although our results are
39consistent with a pervasive role of genetic drift generating repeatable
40patterns of genomic divergence in closely related taxa, they also
41indicate that adaptive evolution targeting a specific set of genes
42contributes to this pattern, yielding insights into the predictability of
43evolution at the gene level.

44**Introduction**

45

46Repeated evolution, or the independent evolution of similar
47phenotypes in different taxa or populations, provides strong evidence
48for natural selection (Harvey & Pagel 1991) and for evolutionary
49constraints in limiting the available variation upon which natural
50selection can act (Wake 1991; Gould 2002). Tracing the evolution of
51those phenotypes to repeated changes in homologous genes is the
52most compelling evidence that gene-level evolution is both repeatable
53and predictable (Gould 2002). Yet, the evolutionary processes under
54which genetic evolution is predictable remain unclear (Stern &
55Orgogozo 2009).

56 Several studies applying new genomic methods have uncovered
57repeated patterns of genomic divergence associated with species
58formation. Genome wide studies of species that evolved via parallel
59evolution (Gagnaire *et al.* 2013; Renaut *et al.* 2014; Soria-Carrasco *et*
60*al.* 2014; Arnegard *et al.* 2014) show repeated patterns of genomic
61divergence, suggesting that evolution often results from changes at a
62relatively small subset of genes. Indeed, the probability of gene reuse
63in parallel phenotypic evolution is particularly high between closely
64related species (between 30 to 50% of the time), and declines with
65increasing age of the common ancestor of the compared taxa (Conte
66*et al.* 2012). This pattern may arise by two different processes. On one
67hand, natural selection acting in independent populations may target a
68limited number of genes with functions related to the shared selective
69pressure driving divergence (Jones *et al.* 2012; Renaut *et al.* 2014). On
70the other hand, genetic drift may also contribute for repeated
71evolution at the gene level because of gene-specific constraints (e.g.
72standing genetic variation, mutation rate, variation in recombination
73rate, linkage relationships or pleiotropic effects; Stern & Orgogozo

742009; Christin *et al.* 2015). These processes are not exclusive because
75genetic **drift** will condition the supply and fixation of beneficial
76mutations where natural selection can act, or may generate selective
77pressures that elicit adaptive evolution. For example, during
78“compensatory coadaptation”, fixation of a deleterious mutation at one
79neutrally evolving locus establishes a positive selection pressure for a
80compensatory mutation in a functional interacting locus to regain some
81component of fitness (Rand *et al.* 2004). The relative contribution of
82genetic drift and positive selection for repeated patterns of genomic
83divergence is difficult to evaluate, but they result in distinct genetic
84signatures (Tsagkogeorga *et al.* 2012; Gayral *et al.* 2013). Scanning for
85such signatures in genomes of taxa at early stages of divergence
86might provide insight into the timing and relative contributions of these
87two evolutionary processes to predictable patterns of gene evolution.

88 Here, we focus on diverging populations of the marine copepod
89*Tigriopus californicus*. This species ranges along the west coast of
90North America, from central Baja California to southern Alaska.
91Restriction to pools in the upper intertidal of isolated rock outcrops has
92resulted in allopatric differentiation without gene flow (Burton 1998).
93Parallel adaptation in *T. californicus* might occur in response either to
94an extrinsic ecological environment or an intrinsic genomic
95environment. High tidal pools are strongly affected by desiccation
96during the summer and rainfall during the winter, resulting in strong
97osmotic, pH, and oxygen stress. While some stressors, such as
98temperature, seem to influence populations differently along a
99latitudinal range (Willet 2010), others are expected to affect allopatric
100populations in a parallel fashion.

101 Parallel adaptation in *T. californicus* is also known to occur in
102response to intrinsic selective pressures caused by gene coadaptation
103(Burton *et al.* 2006). For example, allopatric populations experience

104similarly high mutation rates in mitochondrial genes (on average 55
105times faster than nuclear genes; Willet 2012). Nuclear-encoded
106proteins that interact with mitochondrial proteins are thus likely to be
107under positive selection to maintain proper cellular function (Willett &
108Burton 2004; Barreto *et al.* 2011), resulting in an open-ended
109molecular evolutionary arms race. Experimental hybridization studies
110in *T. californicus* seem to support this gene coadaptation hypothesis:
111hybrid breakdown among multiple populations is attributable to
112coevolution between mitochondrial and nuclear genes (Ellison & Burton
1132008b), as well as among nuclear genes (Pritchard *et al.* 2011).
114Observation of hybrid breakdown in crosses between multiple
115populations (Edmands 1999, Ellison & Burton 2008b) suggests that
116gene coadaptation evolved repeatedly in multiple allopatric
117populations.

118 The availability of many populations with varying degrees of
119genetic divergence and reproductive isolation makes *T. californicus* an
120attractive model for analysis of genomic evolution during the
121continuum of allopatric divergence. Here, using transcriptome
122sequence data from six geographically isolated populations, we test: i)
123if geographic isolation leads to rapid differentiation among populations,
124ii) if population differentiation is associated with repeatable patterns of
125protein divergence; iii) if molecular signatures of accelerated protein
126evolution are predictable at several stages of allopatric evolution; and
127iv) the relative roles of genetic drift and positive selection in
128contributing to the observed pattern of repeated genomic divergence.

129

130

131**Materials and Methods**

132

133**Population collection and sequencing**

134We collected copepods from high intertidal rocky pools in three
135northern and three southern populations of California (Fig. 1): San
136Diego (SD: 32°44'N, 117°15'W), Bird Rock (BR: 32°48'N, 117°16'W),
137Abalone Cove (AB: 33°44'N, 118°22'W), Santa Cruz (SCN: 36°57'N,
138122°03'W), Pescadero (PES: 37°15'N, 122°24'), and Bodega Bay (BB:
13938°19'N, 123° 4'W). All cultures were maintained at common garden
140and outbreeding conditions. We extracted RNA from pools of 300-400
141individuals from all developmental stages, using the standard Tri-
142Reagent (Sigma) protocol. Resuspended RNA pellets were further
143purified with RNeasy Mini columns (Qiagen), and final sample integrity
144and quantity were assessed with an Agilent 2100 BioAnalyzer. This
145protocol was repeated twice for each population, resulting in 12 RNA
146samples. Libraries were constructed by the Beijing Genomics Institute
147(BGI) and by Cofactor Genomics, using 100bp paired-end sequencing
148on an Illumina HiSeq II.

149

150**Population de novo assemblies**

151We constructed de novo assemblies for each population separately, to
152avoid mapping biases caused by high interpopulation divergence (20%
153across mitochondrial genes; (Burton et al. 2007). FASTQ reads were
154trimmed via Prinseq (prinseq-lite-0.19.5; (Schmieder & Edwards 2011)
155to remove base pairs with a Phred score <20 and trim poly-A tails
156>8bp in length. Reads that were trimmed below 55 bp in length or
157read pairs that were no longer complete after trimming were removed
158prior to assembly. Trinity *de novo* assembly was performed using
159default parameters (Grabherr et al. 2011). To reduce gene redundancy

160within each assembly we performed two additional rounds of assembly
161using the CAP3 assembler (Huang & Madan 1999) with default
162parameters. We evaluated the quality of each complete assembly by
163calculating the percentage of annotated core genes from eukaryotic
164genomes using CEGMA (Parra et al. 2007). We identified sets of
165orthologous nuclear genes across the 6 population assemblies using a
166reciprocal best BLASTN hit strategy between the SD-assembly and
167each of the other assemblies (Camacho et al. 2008). We retained a
168total of 12,573 nuclear orthologous genes across all 6 populations. We
169assembled the 13 mitochondrial genes separately for each population.
170Using Bowtie2 (Langmead & Salzberg 2012) we mapped our libraries to
171the three complete mitochondrial genomes published for this species
172(SD, AB and SCN, from Burton et al. 2007): SD and BR mapped to SD;
173AB mapped to AB; and SCN, PES and BB mapped to SCN. We extracted
174uniquely mapped reads covering the mitochondrial genes, identified
175SNPs relative to the reference using SAMtools (Li et al. 2009), and
176exported consensus mitochondrial sequences using the majority rule
177for each base, substituting N's for any base with < 8 reads coverage.
178We used the reciprocal best BLAST hit strategy to check for
179redundancies between mitochondrial and nuclear genes within each
180population assembly.

181

182**Population divergence**

183We estimated a “species-tree” for the six populations from
184independent gene trees, using the NJ_{st} approach (Liu & Yu 2011). We
185used our 12,573 orthologous nuclear genes to estimate a six-
186population alignment for each gene, using MUSCLE (Edgar 2004). To
187avoid inflating interpopulation divergence due to assembly artifacts
188and misalignments, we filtered each alignment using a sliding window

189 approach and rejecting windows of 21bp that contained more than 7
190 SNPs (the maximum divergence between these populations estimated
191 at mitochondrial genes; Burton *et al.* 2007). To incorporate uncertainty
192 in gene tree estimation from our alignments, we use RAxML to
193 generate 100 bootstrapped alignments for each locus (resample sites
194 with replacement), and estimate respective maximum likelihood (ML)
195 trees. This resulted in 100 sets of 12,014 ML trees. We used STRAW
196 (Species TRee Analysis Web server; Shaw *et al.* 2013) to estimate a
197 consensus species tree (majority rule) with bootstrap values. For
198 visualization purposes we rooted the tree based in a species-scale
199 phylogeny of *T. californicus* (Peterson *et al.* 2013). We estimated
200 interpopulation divergence by calculating uncorrected P-distances for
201 each locus with trimmed alignments > 100bp (11,560 nuclear loci) and
202 averages across loci. We replicated the same procedure for the
203 complete 13 mitochondrial genes.

204

205 **Genetic diversity**

206 We identified SNPs at each population by mapping libraries to their
207 respective assembly using Bowtie2 (seed size 22 bp; Langmead &
208 Salzberg 2012).

209 First, to assess how quickly genetic differentiation accumulates
210 during allopatric divergence, we focus on SNPs detected throughout
211 the genome. We used SAMtools (Li *et al.* 2009) to export mapped
212 reads and identify SNPs segregating within each population according
213 to the following criteria: mapping quality > 20, SNP quality > 20, and
214 coverage > 8x. We produced a trimmed 6-population alignment, using
215 MUSCLE (Edgar 2004), and projected the identified SNPs using custom
216 scripts. This analysis was restricted to positions where all six
217 populations had callable genotypes according to the criteria described

218above. This allowed us to compute shared and fixed mutations for
 219every interpopulation comparison, across 12,044 nuclear loci,
 220encompassing a total of 16,415,579 bp. We tested whether fixed and
 221shared mutations changed linearly or exponentially with genetic
 222divergence (as a proxy for time since divergence) by fitting a linear
 223($y = b + x$) and an exponential ($y = ax^b$) model. The linear model for
 224fixed mutations was forced through 0, assuming that at time of
 225population split there are no fixed differences between populations.
 226Models were fitted and compared using AIC in R-studio (available from
 227<http://www.rstudio.org/>).

228 Second, to assess the fraction of polymorphisms that are slightly
 229deleterious and neutral we focus on SNPs that occur in synonymous
 230(π_S) and non-synonymous sites (π_N), following the approach of Gayral
 231*et al.* (2013). This approach is developed for transcriptome data from
 232non-model organisms where a reference genome is not available. In
 233short, de novo assemblies are used both to predict open reading
 234frames (ORFs) for homologous genes sequences and to map reads
 235from a pool of individuals. Mapped reads are then filtered for
 236paralogues and SNPs are called with a conservative coverage threshold
 237to account for pool-sequencing. Interpopulation comparisons of the
 238same genomic regions allow the estimation of summary statistics for
 239population genetics analysis. Using the program reads2snps
 240(Tsagkogeorga et al. 2012), we filtered high coverage SNPs while
 241controlling for possible paralogous mapping (spa option), according to
 242the following criteria: mapping quality > 20, SNP quality > 20, and
 243coverage > 30x. We repeated filtering using 20x and 40x coverages to
 244assess a potential effect of library size in SNP detection. We identified
 245ORFs using the program TransDecoder (Haas et al. 2013), and
 246computed π_N , π_S and π_N/π_S for every gene in each population. Due to
 247differences in coverage, this approach retrieved results from a varying

248number of nuclear loci across populations; 3,620 in SD, 2,904 in BR,
2493,845 in AB, 3,005 in SCN, 3,574 in PES, and 2,754 in BB. To avoid
250biases created by less polymorphic genes, average population indexes
251of diversity were computed as the sum of π_N and the sum of π_S across
252all loci.

253

254**Protein divergence and adaptive evolution**

255We focused on five population comparisons representative of three
256stages of divergence: early a – SCN-PES, early b – SD-BR, middle –
257SCN-BB, late a – SD-AB, late b – BR-BB. We annotated our sets of
258orthologous genes relative to the SD assembly using Blast2GO (Conesa
259et al. 2005), retaining the highest hit with E-value $\leq 10^{-3}$ and its
260predicted gene name. We used a custom Perl script to extract the most
261likely open reading frame (ORF) for SD contigs according to their
262BLASTX best hits. The transcriptome assemblies from the other five
263populations were then each blasted (BLASTN) against the SD ORFs,
264extracting the best ORF for every ortholog. For each focal population
265comparison listed above, we produced pairwise alignments using
266MACSE (Ranwez et al. 2011), which accommodates possible
267frameshifts introduced by sequencing and hence retains the
268appropriate reading frame. A custom script was employed to automate
269the $\sim 7,000$ alignments in each two-population comparison. We
270calculated average d_N and d_S for the entire protein alignment using the
271maximum likelihood method of Yang and Nielsen (Yang & Nielsen
2722000), implemented in the script YN00 within the PAML package
273(version 4.4; Yang 2007). We visually inspected all alignments with d_S
274 > 2 , and removed those that showed evidence of possible assembly or
275alignment errors, resulting in a total of 7,005 nuclear genes plus 13
276mitochondrial genes. Pairwise ω ($\omega = d_N/d_S$) was estimated for 5,677

277 genes with $d_s > 0$ at any interpopulation comparison. For each focal
 278 population comparison, we computed which loci were among the top 5
 279 and 10% for d_N , d_S , and ω , and calculated overlaps across the five
 280 comparisons. We assessed whether the number of overlapping genes
 281 could occur by chance by comparing the observed value to a null
 282 distribution of overlap between 3 independent population comparisons,
 283 a conservative scenario relative to our 5 population comparisons. A
 284 null distribution of number of genes with overlap was obtained by: 1)
 285 simulating three random draws of 5 and 10% genes from a total of
 286 7,018 genes, 2) calculating overlaps among the three lists, and 3)
 287 repeating this process 1,000 times.

288 In a second analysis, for the 36 target genes (i.e. top 10% of ω
 289 across the five population comparison), we tested whether a subset of
 290 codons exhibited $\omega > 1$ across all populations. We first aligned all 6
 291 orthologs for each gene using MACSE, followed by removal of poorly
 292 aligned regions using the 'codon' option in Gblocks (Castresana 2000;
 293 parameters: -b1=4 -b2=5 -b3=6 -b4=10 -b5=h). We then used
 294 CODEML in PAML to test the fit of a null model that does not allow $\omega >$
 295 1 (M7), and a more general model that does (M8; positive selection).
 296 The model fitting was compared statistically using a likelihood ratio
 297 test; the negative of twice the log-likelihood difference between
 298 models was compared with the χ^2 distribution with d.f. = 2.

299

300 **Tests for genetic drift, evolutionary constraints and negative** 301 **selection**

302 We focused on population comparisons representative of latter stages
 303 of divergence (late a - SD-AB, late b - BR-BB), to avoid higher variance
 304 of called polymorphisms typical of earlier stage of divergence. To
 305 estimate gene effect size, we computed π_N/π_S for each gene. To

306 estimate expression, we measured the number of fragments per
 307 kilobase per million mapped reads (FPKM) in two biological replicates
 308 for each population and computed averages per gene. We tested for
 309 correlations between FPKM and π_N/π_S relative to ω (d_N/d_S) in all sampled
 310 genes, using a Mann-Whitney U test. We then tested for differences
 311 between candidate genes within the top 10% of ω and the remaining
 312 proteins, using a Wilcoxon test. We tested for genetic linkage among
 313 target genes within the same scaffold and the same chromosome by
 314 aligning the respective sequences from the SD-assembly against the
 315 draft genome of *Tigriopus californicus*
 316 (https://i5k.nal.usda.gov/Tigriopus_californicus) using BLASTN. We
 317 computed the direction of selection (negative or positive) using the
 318 DoS statistic (Stoletzki & Eyre-Walker 2011): $\text{DoS} = d_N/(d_N+d_S) -$
 319 $\pi_N/(\pi_N+\pi_S)$. We assessed differences between candidate and remaining
 320 genes using an ANOVA.

321

322

323 **Results and Discussion**

324

325 To investigate patterns of transcriptome-wide divergence during
 326 allopatric evolution of *Tigriopus californicus* we generated high-quality
 327 de novo transcriptome assemblies for each of the six populations,
 328 using between 22 and 104 million paired-end 100 bp reads. More than
 329 98% of the core genes of eukaryotic genomes (Parra et al. 2007) are
 330 represented in every assembly (Table S1). Reciprocal BLAST searches
 331 resulted in the identification of 12,573 orthologous loci from the
 332 nuclear genome plus all 13 mitochondrial genes across the
 333 populations. The orthologous genes have similar sizes across

334populations and coverage ranges from 16 to more than 65,000 reads
335per contig (Fig. S1).

336

337**Geographic isolation results in three stages of population** 338**divergence**

339In assessing the phylogenetic relationships among the populations,
340closely related populations may present discordance among
341independent gene trees due to the influence of differential gene
342sorting. Thus, we estimated a single “species-tree” (NJst; Liu & Yu
3432011) using gene trees estimated from each of the 12,014 nuclear loci
344with reliable alignments. Our results show that the tree topology
345recapitulates geography (Fig. 1), in agreement with the restricted
346dispersal ability known for this species. Northern and southern
347populations form two reciprocally monophyletic groups, with
348geographically close populations sharing a common ancestor more
349recently than with the third more distant population. Our results are in
350agreement with previous phylogenetic trees estimated from
351mitochondrial genes alone (Willett & Ladner 2009; Peterson et al.
3522013), except that we now find strong support for deeper nodes that
353were previously unresolved.

354 As a proxy for relative divergence time between each population
355pair, we calculated uncorrected average pairwise divergence for the
35611,560 nuclear loci larger than 100 bp (Table S1). Our results indicate
357that interpopulation comparisons reflect three different stages of
358divergence (Fig. 1A). Neighboring populations from the northern and
359southern clades (SCN-PES and SD-BR, respectively) are 0.7%
360divergent, reflecting two, phylogenetically independent, vicariant
361events at an early stage of divergence. BB is 1.3 and 1.4% divergent
362from the other northern populations, describing an intermediate stage

363of divergence. Finally, all the remaining population comparisons,
364including AB relative to the other southern populations, are 2.3 to 2.8%
365divergent, revealing later stages of population divergence. The same
366relative magnitude of divergence is mirrored by the mitochondria
367(Table S2) with the notable exception of the populations at the early
368stage of divergence, where the southern population pair shows a level
369of mitochondrial divergence almost 8 times larger than the northern
370population pair.

371 Although the high levels of genetic divergence reported here are
372typically found between reproductively isolated species, experimental
373crosses in *Tigriopus californicus* show that in this species, hybrid
374inviability evolves at much later stages of divergence, as observed in
375other species lacking sexual chromosomes (Lima 2014). Hybrids
376between SD and BR show some reduced fitness relative to parentals,
377suggesting that such early stages of allopatric evolution have already
378resulted in some genetic incompatibilities (Pereira *et al.* 2014). The
379magnitude of hybrid breakdown increases with genetic divergence
380among parental populations, being most pronounced at the latter
381stages of divergence reported here (e.g. SD-SCN, BR-SCN; Edmands
3821999; Pereira *et al.* 2014). In this context, comparisons among the six
383populations of *T. californicus* studied here reflect a continuum of
384allopatric evolution, presenting a rare opportunity to study genomic
385divergence along the continuum between population and species level
386divergence.

387

388**Low N_e leads to an exponential decrease of shared** 389**polymorphisms**

390Patterns of genomic divergence during allopatric evolution are not only
391conditioned by time since geographic isolation and the accumulation of

392new mutations (divergence), but also by demographic effects that
393condition how quickly pre-existing or new variation becomes fixed in
394descendent populations (differentiation). Neutral theory of evolution
395(Kimura 1968; Ohta 1992) posits that the strong effect of genetic drift
396in small populations leads to the potential fixation of slightly
397deleterious mutations and loss of slightly advantageous mutations at a
398higher rate than in larger populations. Here, we use single nucleotide
399polymorphisms (SNPs) to assess how quickly genetic differentiation
400accumulates during the continuum of allopatric evolution reflected by
401*T. californicus* populations, and assess how such differentiation might
402be affected by genetic drift.

403 Based on 12,044 nuclear loci, our results show that at early
404stages of divergence (0.7% nuclear divergence between SCN-PES and
405SD-BR), most variable nucleotide sites are fixed (62.7% and 72.2% of
406SNPs, between SCN-PES and SD-BR respectively) while only 0.6% or
407less are shared (Table S3, Fig. 2); the remaining SNPs are polymorphic
408within each population. These population pairs are respectively 46 and
4098 km apart from each other, which indicates that geographic isolation
410in *T. californicus* can occur at fine spatial scales. By inspecting levels
411of differentiation across all 15 interpopulation comparisons, we
412observe that genomic differentiation based on polymorphisms is not
413linearly related to divergence time and is more consistent with an
414exponential change (Fig. 2; $AIC_{\text{exponential}} \ll AIC_{\text{linear}}$ for fixed and shared
415mutations). Although not all pairwise comparisons are statistically
416independent because of multiple comparisons using the same
417population, changes remain exponential when this analysis is
418restricted to the 3 independent pairwise comparisons (Fig. S2). This
419result shows that, in *T. californicus*, allopatric evolution leads to a rapid
420decrease of shared polymorphisms between diverging populations.

Neutral genetic polymorphisms, or genetic diversity (π), is 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* can potentially be caused by a relatively small N_e within each population. Based on a subset of 2,754 high-coverage nuclear loci we estimated the fraction of genetic polymorphisms in non-synonymous (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 30x coverage (Fig. SX). 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 time scales is relatively small, similar to other crustaceans such as crabs and shrimps (Romiguier et al. 2014). Such small N_e s are likely caused by demographic bottlenecks that frequently affect populations of *T. californicus*, such as high mortality caused by rainfall during winter and desiccation in the summer, followed by recolonization from interconnected pools and rapid population expansion that favors the 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

451 taxa (Gayral et al. 2013; Romiguier et al. 2014), populations of *T.*
 452 *californicus* show that π_N/π_S is inversely proportional to π_S ($r^2 = 0.9$,
 453 Mann-Whitney U test: $p = 0.002$), supporting a role of genetic drift
 454 driving divergence. Because π_S variation between populations of the
 455 same species is not confounded, in principle, by variation of mutation
 456 rate (μ) or of life history traits, the observed negative correlation
 457 directly reflects variation of N_e . Within *T. californicus*, diversity within
 458 populations ranges from 0.006 to 0.017, suggesting that northern
 459 populations (mean π_S : 0.015) are on average two times larger than
 460 southern populations (mean π_S : 0.007; Fig. 3B). Mitochondrial genes
 461 are particularly prone to the effect of genetic drift because the N_e of
 462 the mitochondria is one-fourth as large as the N_e for autosomal genes
 463 (Pool & Nielsen 2007). Such differences in N_e among genes and
 464 populations likely contribute to our finding that smaller populations in
 465 the south (SD and BR) show 8 times higher mitochondrial divergence
 466 relative to the larger populations in the north (SCN and PES) with
 467 comparable genome-wide divergence (Table S2).

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

473

474 **Non-synonymous mutations accumulate more often in a subset** 475 **of proteins**

476 Recent studies comparing genomes of multiple related species have
 477 found that certain genes or regions repeatedly show higher divergence
 478 relative to the mean of the genome (e.g. Gagnaire *et al.* 2013; Renaut
 479 *et al.* 2014) that are consistent with shared selective pressures and

480with shared evolutionary constraints. Even in cases where high
481divergence is not necessarily adaptive, faster evolution of nearly
482neutral genes can secondarily elicit adaptive compensatory changes in
483interacting loci (“compensatory coadaptation”; Rand *et al.* 2004). To
484test for repeatable patterns of genomic divergence in *T. californicus*
485and identify candidate proteins driving divergence, we focus on five
486population comparisons at the three different stages of divergence:
487early (a: SCN-PES; b: SD-BR), middle (SCN-BB), and late (a: SD-AB; b:
488BR-BB).

489 Functional divergence of the 7,018 annotated protein-coding
490genes used in the five comparisons follows the average pairwise
491divergence described by all nuclear loci, with the rate of non-
492synonymous substitutions (d_N) increasing with time since divergence
493(Fig. S3). With d_N , population comparisons at consecutive stages of
494divergence significantly differ from each other (test of equality on the
495distributions of d_N ; all *p-values* ≤ 0.01), confirming that these represent
496distinct stages of genomic divergence. In addition, the two replicated
497population comparisons at early (SCN-PES; SD-BR) and at late (SD-AB;
498BR-BB) stages of divergence do not differ from each other (test of
499equality; both *p-values* ≥ 0.44), confirming that these represent
500natural replicates of vicariant events with similar amounts of genomic
501divergence.

502 For each population comparison we identified the proteins with
503the highest protein divergence (top 5 and 10% with respect to d_N) and
504assessed overlaps across the five comparisons. Our results show a
505large overlap across the five population comparisons (Table S4); 39
506proteins are among the 5% most divergent in all five interpopulation
507comparisons, and when we consider the top 10%, the number of
508overlapping proteins increases to 103. It is important to note that only
509three of the five populations comparisons are statistically independent

510(SCN-PES, SD-AB, and BR-BB). Nevertheless, the observed number of
511overlapping genes remains highly significant ($p\text{-values} < 0.0001$)
512considering a very conservative null expectation based on the three
513statistically independent comparisons (5% overlap: range- 0 to 5
514genes, median- 1 gene; 10% overlap: range- 0 to 19 genes, median- 7
515genes; Fig. S4). This large number of overlapping proteins indicates
516that functional mutations often accumulate in a subset of loci.

517 Our results show that 7 out of the 13 mtDNA-encoded proteins
518are among the 103 fast-evolving genes. Mitochondrial genes are
519expected to contribute disproportionately to genetic incompatibilities
520between populations because they i) tend to accumulate deleterious
521mutations faster than nuclear genes due to a Muller's ratchet effect
522(lack of recombination and $1/4 N_e$ relative to nuclear genes), ii) have
523faster mutation rates relative to nuclear genes, and iii) constitute
524mitonuclear enzyme complexes with key roles in metabolism in all
525eukaryotic cells (Rand *et al.* 2004; Burton & Barreto 2012).
526Experimental hybridizations between multiple populations of *T.*
527*californicus* (Ellison & Burton 2008b) have shown that divergence in
528mitochondrial genes results in Dobzhansky-Muller incompatibilities
529with interacting nuclear loci as a result of intergenomic coadaptation
530(Rand *et al.* 2004; Burton *et al.* 2013). Yet, identifying which
531mitochondrial and nuclear genes are involved in such compensatory
532coadaptation is challenging. In *T. californicus*, mismatch between
533nuclear and mitochondrial components of the oxidative
534phosphorylation (OXPHOS) system have been shown to reduce
535enzymatic activity of complex IV (Edmands & Burton 1999; Rawson &
536Burton 2002) and ATP production (Ellison & Burton 2008a). Our results
537show that COX1, one of the 3 mitochondrial proteins that compose
538complex IV, is among our hotspot genes. Interestingly, the remaining
539mtDNA encoded proteins with accelerated evolution (NAD1, 2, 4, 5, 6,

540and 4L) constitute 6 of the 7 mtDNA encoded components of complex
541I. It is unknown whether mitonuclear interactions in complex I cause
542fitness breakdown in *T. californicus*, but studies in humans have shown
543that about 50% of all mitochondrial disorders affecting the energy
544metabolism can be traced to mutations in one of the subunits of
545complex I (Smeitink et al. 2001; Brandt 2006).

546 Our finding that mutations tend to accumulate in a particular set
547of genes shows that allopatric evolution leads to repeatable patterns of
548genomic divergence. The evolutionary processes underlying this
549patterns can be either neutral, if these genes evolve under specific
550constraints, or adaptive, if these genes adapt in response to an
551ecologic or genomic environment. Distinguishing between these
552alternatives is difficult, but scanning for genomic signatures of neutral
553and adaptive evolution can provide valuable insights on the relative
554contribution of the two evolutionary processes.

555

556**Patterns of accelerated protein evolution are predictable**

557Genomic methods allow investigation of how frequently the same
558genes are targeted by natural selection in different populations,
559yielding insights into the predictability of evolution at the genetic level.
560A meta-analysis of taxa undergoing parallel speciation found that the
561probability of gene reuse is surprisingly high (30-50% of the time),
562particularly in closely related taxa (Conte et al. 2012). This suggests
563that strong selective biases and constraints affect adaptive evolution,
564resulting in change at a relatively small subset of available genes. We
565scanned the transcriptome for signatures of rapid protein evolution,
566and tested whether those signatures are predictable during the
567continuum of allopatric evolution reflected by our study populations of
568*T. californicus*.

569 We estimated pairwise ω in our five focal population comparisons
570($\omega = d_N/d_S$, where d_N is the rate of non-synonymous substitutions per
571non-synonymous site and d_S is similarly the rate of synonymous
572substitutions). Codons with values of $\omega > 1$ are suggestive of adaptive
573evolution promoting accelerated divergence between taxa (Nei &
574Kumar 2000; Charlesworth & Charlesworth 2010). However, when ω is
575calculated across an entire protein sequence, a criterion of $\omega > 1$ as
576evidence for adaptive evolution is extremely stringent (Swanson *et al.*
5772001a). A sequence-wide ω threshold of 0.5 has been shown to
578consistently identify genes subjected to adaptive evolution (Swanson
579*et al.* 2001b; 2004). Our pairwise estimations of ω for 5,677 protein-
580coding genes show high values across the five population comparisons
581(up to 4.7 in early a, to 3.9 in early b, to 4.4 in middle, to 1.9 in late a,
582and to 1.4 in late b; Fig. S5). Yet, whether we consider a threshold of ω
583 > 1 or > 0.5 , the number of genes putatively under adaptive evolution
584decreases with population divergence (Fig. S6A). Comparative genomic
585studies in various taxa such as bacteria (Rocha *et al.* 2006), mammals
586and birds (Wolf *et al.* 2009), have reported a similar negative
587correlation between estimates of ω and genetic distance in pairwise
588comparisons of the sequences for the same gene. This recurrent
589pattern has been interpreted either as 1) evidence for varying strength
590of selection during species divergence (Read *et al.* 2002; Baker *et al.*
5912004), 2) an effect of a decrease in non-synonymous polymorphisms
592during divergence due to a larger time frame for purifying selection to
593operate (Rocha *et al.* 2006), or 3) an artifact due to a larger number of
594polymorphisms segregating in recently diverged taxa (Peterson &
595Masel 2009). Our results show that, in *T. californicus*, less than 0.6% of
596variable nucleotide sites are shared among populations at earlier
597stages of divergence (Fig. 2B; Table S3), suggesting that retention of
598ancestral polymorphism is unlikely to account for the observed
599decrease in the number of putatively adaptive genes with time since

600divergence. Moreover, we observed that, although both kinds of
601mutations increase during population divergence, d_N accumulates at a
602much slower rate relative to d_S (Fig. S6B; both $r^2 = 0.99$; p -values $< <$
6030.01), affecting the number of candidate genes based on a fixed
604threshold.

605 We used our genealogy of *T. californicus* (Fig. 1A) to test whether
606loci under accelerated evolution at an early stage of divergence can
607predict rates of evolution at the remaining four population
608comparisons, representative of all three stages of population
609divergence. Our results show that loci with $\omega > 0.5$ at an early stage of
610divergence (early a: SCN-PES) also have higher ω than the rest of the
611genome at a phylogenetically independent population comparison with
612the same level of divergence (early b: SD-BR, Mann-Whitney U test: p
613 $< 2 \times 10^{-6}$; Fig. 4). In addition, this pattern is maintained at later stages
614of population divergence (all Mann-Whitney U tests: p -values $< 2 \times 10^{-6}$),
615suggesting that signatures of rapid protein evolution are predictable
616both in independent vicariant events and along the continuum of
617population divergence.

618 To identify proteins with accelerated evolution (relative to the
619genome mean) during the continuum of population divergence of *T.*
620*californicus*, we assessed which proteins are in the top 5 and 10% of ω
621for each comparison and calculated overlap across the five
622comparisons. Our results show that 12 proteins out of 5,677 are
623consistently among the top 5% of ω (values across the five population
624comparisons range between 0.44 and 4.70; Table S5), and 36 are
625among the top 10% ($\omega > 0.33$). Similar to the overlaps in d_N , this
626number of overlapping genes remains highly significant considering a
627very conservative null expectation based on the three statistically
628independent comparisons (Fig. S4); p -values < 0.0001 . Of these 36
629proteins under accelerated evolution, all are nuclear-encoded.

630 Interestingly, none of these 36 proteins are known to functionally
631 interact with the mitochondria, as would be predicted if mitonuclear
632 coadaptation would result in accelerated evolution across the entire
633 nuclear protein. Previous studies on candidate genes underlying
634 mitonuclear incompatibilities in *T. californicus* found that hybrid
635 breakdown in the activity of the mitochondria-encoded COX protein
636 can be attributed to a single amino acid substitution in the interacting
637 nuclear-encoded CYC protein (Harrison & Burton 2006). Thus, when
638 adaptive changes in a protein result from a few amino acid changes
639 and rest of the protein evolves neutrally, those genes can be missed in
640 genome scans using ω values averaged across protein sequences.

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

649

650 **Evolutionary constraints and positive selection contribute to** 651 **repeatability of genomic divergence**

652 Two non-mutually exclusive processes can contribute to
653 repeated patterns of genomic divergence commonly observed in
654 comparisons between closely related species: evolutionary constraints
655 and positive selection. Here, we use multiple allopatric populations of
656 *T. californicus* to evaluate the contribution of these two evolutionary
657 processes during allopatric evolution within the same species.

658 To test for the effect of evolutionary constraints on accelerated
 659 protein evolution as measured by d_N/d_S (ω), we focus on population
 660 comparisons reflective of later stages of divergence (late a: SD-AB; late
 661 b: BR-BB). Genomic architecture may contribute to changes in protein
 662 polymorphisms among genes. For example, genomic regions with low
 663 recombination rates, as is typical around centromeres, within
 664 inversions, or around genes under selection, experience local
 665 reductions in effective population size and thus higher genetic drift. If
 666 there is a significant effect of genetic drift, levels of protein
 667 polymorphism measured as π_N/π_S are expected to correlate with protein
 668 divergence (Charlesworth & Charlesworth 2010). Accordingly, our
 669 results show significant positive correlations throughout the genome
 670 (all Mann-Whitney U tests: $p\text{-values} < 10^{-10}$), and significantly lower
 671 polymorphisms in rapidly evolving proteins, relative to the remaining
 672 genes (all Wilcoxon tests: $p\text{-values} \leq 0.02$; Fig. 5). It is important to
 673 note that analyses of π_N and π_S are restricted to loci with high coverage
 674 (> 60 reads per site) and exclude loci with no neutral polymorphism
 675 ($\pi_S = 0$) and small length (< 100 sites). Although this bias will inflate
 676 average π_N/π_S it does not affect the classification of candidate genes
 677 based on d_N/d_S . This result is in agreement with the general
 678 observation that gene-specific drift conditions the supply and fixation
 679 of mutations, thus affecting rates of protein evolution in *Tigriopus*
 680 *californicus*. Because the genetic drift experienced by the same gene is
 681 similar across populations, this constraint likely leads to high ω at the
 682 same genes across population comparisons, largely contributing to the
 683 observed pattern of repeated and predictable evolution at the gene
 684 level.

685 As discussed above, genetic drift may lead to a signature of
 686 adaptive divergence by facilitating the fixation of slightly deleterious
 687 mutations via a lower efficiency of purifying (negative) selection. To

688test whether the 36 proteins showing consistent rapid evolution are
 689mostly affected by negative or positive selection, we computed the
 690direction of selection (DoS) statistic (Stoletzki & Eyre-Walker 2011), a
 691modified version of the neutrality index (Rand & Kann 1996). Positive
 692values of DoS (excess of d_N/d_S relative to π_N/π_S) are indicative of
 693positive selection, while negative values of DoS reflect the influence of
 694purifying selection. Our results show that most genes have DoS values
 695around zero or slightly negative, suggestive of neutrality (medians:
 696SD= -0.008, AB= 0.000, BR= -0.006, BB= 0.008; Fig. 6). Yet, the
 697median values of DoS for candidate genes are all positive (SD= 0.042,
 698AB= 0.123, BR= 0.003, BB= 0.063), suggestive of a stronger
 699contribution of positive selection in the signature of adaptive
 700divergence. It is important to note that only one of these differences is
 701significant (ANOVA, $p < 0.001$). To further test these 36 rapidly
 702evolving genes for signatures of positive selection, we aligned all 6
 703orthologs for each gene and applied models that allowed ω to vary
 704among codons, instead of assuming one ω for the entire gene. We
 705compared a null model that does not allow for any codons to have $\omega >$
 7061 against a more general model that does (Yang 2007). Although our
 707power to detect positive selection is low both due to the small number
 708of aligned taxa and the low level of sequence divergence observed
 709between populations (relative to studies comparing different species or
 710genera), we find that a model allowing for positive selection in some
 711codons was a significantly better fit in 42% of our candidate genes
 712(Table1). For those genes, values of ω ranged from 19.0 in 2% of the
 713protein to 1.9 in 55% of the protein, suggesting that positive selection
 714can either target key amino acids or most of the protein sequence.
 715Together, our results suggest that repeated patterns of accelerated
 716evolution in a subset of genes might be, at least partially, driven by
 717positive selection.

718 Genome scans such as the one employed here can provide a
719manageable list of candidate loci underlying adaptation in independent
720evolutionary units. Although most of these genes remain unknown due
721to insufficient annotation and lack of homology to other organisms,
722some of our candidate genes belong to functional classes of proteins
723known to be involved in adaptive evolution in taxa beyond these
724populations of *T. californicus*. For example, the two highest values of ω
725(Table 1) were found in a lectin (19.05) and a histone (15.31). Lectins
726are involved in binding to carbohydrates displayed in cell surfaces;
727lectins mediating sperm-egg incompatibility were shown to evolve
728under positive diversifying selection in oyster, mussel, sea urchin and
729fruit fly (Mah 2004; Moy *et al.* 2008; Findlay *et al.* 2009; Lima &
730McCartney 2013). Histones play an important role in DNA binding and
731gene regulation, and have also been suggested to underlie genomic
732coadaptation in mammal species (Nowick *et al.* 2013; Carneiro *et al.*
7332014). Surprisingly, other genes that are known to evolve under
734diversifying selection in other organisms, such as immune genes
735(Obbard *et al.* 2009) are not present in this list.

736 Together, these results show that despite an important
737contribution of gene-specific evolutionary constraints, positive
738selection likely contributes to the repeated patterns of genomic
739divergence observed in *Tigriopus californicus*. Positive selection
740targeting specific genes during allopatric evolution can be both caused
741by adaptation to an extrinsic ecological environment (e.g. in response
742to abiotic stressors), or adaptation to an intrinsic genetic environment
743(e.g. in response to compensatory coadaptation among genes). *In*
744*vitro* experiments of isolated proteins (Rawson & Burton 2002; Harrison
745& Burton 2006) and recombinant inbred lines (Barreto & Burton 2013b)
746have been successfully used in *T. californicus* to establish links
747between accelerated evolution in specific genes and the phenotypes

748affected by them, offering unique insights into the adaptive regimes
749operating in this system (Burton et al. 2013). Follow up studies can
750elucidate the function of the candidate genes identified here and
751provide important insights into the selective regimes causing the
752observed pattern of repeated accelerated evolution.

753

754**Concluding remarks:**

755Here, we focused on multiple allopatric populations of the copepod
756*Tigriopus californicus* to understand the evolutionary processes
757underlying the emergent pattern of repeated genomic divergence
758commonly observed among different species. A phylogenetic tree
759based on transcriptome data confirms that the evolutionary history of
760these populations is largely influenced by geographic isolation and that
761allopatric divergence as resulted in three stages of divergence (Fig. 1).
762Polymorphism data show that genetic drift is an important driver of
763allopatric evolution, leading to an exponential fixation of alleles
764between diverging populations (Fig. 2). Synonymous and non-
765synonymous amino-acid substitutions show that mutations tend to
766accumulate in a specific subset of proteins, some of which with
767molecular signatures of accelerated evolution, leading to repeated and
768predictable pattern of divergence at the gene-level (Fig. 4). Although
769this pattern can partially be explained by gene-specific evolutionary
770constraints such as drift (Fig. 5), we find a strong contribution of
771positive selection (Fig. 6, Table 1). These results show that, although
772initial stages of allopatric evolution is largely conditioned by neutral
773processes, adaptive divergence is an important contributor to repeated
774patterns of genomic divergence.

775

776 Acknowledgements

777 This study was supported by grants from the U.S. National Science
778 Foundation (DEB1051057 and IOS1155030 to R.S.B). M.C. was
779 supported by POPH-QREN funds from the European Social Fund and
780 Portuguese MCTES (FCT Investigator grant IF/00283/2014; FCT post-
781 doc grant SFRH/BPD/72343/2010). We thank A. Geraldes, P. Ralph, S.
782 Singhal, and two anonymous reviewers for their comments on the early
783 version of this manuscript.

784

785 References

- 786 Arnegard ME, McGee MD, Matthews B *et al.* (2014) Genetics of
787 ecological divergence during speciation. *Nature*, **511**, 307–311.
- 788 Baker L, Brown T, Maiden MC, Drobniowski F (2004) Silent nucleotide
789 polymorphisms and a phylogeny for *Mycobacterium tuberculosis*.
790 *Emerging Infectious Diseases*, **10**, 1568–1577.
- 791 Barreto FS, Burton RS (2013a) Evidence for Compensatory Evolution of
792 Ribosomal Proteins in Response to Rapid Divergence of
793 Mitochondrial rRNA. *Molecular Biology and Evolution*, **30**, 310–314.
- 794 Barreto FS, Burton RS (2013b) Elevated oxidative damage is correlated
795 with reduced fitness in interpopulation hybrids of a marine
796 copepod. *Proceedings Of The Royal Society B-Biological Sciences*,
797 **280**, 20131521–20131521.
- 798 Barreto FS, Moy GW, Burton RS (2011) Interpopulation patterns of
799 divergence and selection across the transcriptome of the copepod
800 *Tigriopus californicus*. *Molecular Ecology*, **20**, 560–572.
- 801 Brandt U (2006) Energy converting NADH: quinone oxidoreductase
802 (complex I). *Annual review of biochemistry*, **75**, 69–92.
- 803 Burton RS (1998) Intraspecific phylogeography across the Point
804 Conception biogeographic boundary. *Evolution*, **52**, 734–745.
- 805 Burton RS, Barreto FS (2012) A disproportionate role for mtDNA in
806 Dobzhansky-Muller incompatibilities? *Molecular Ecology*, **21**, 4942–
807 4957.
- 808 Burton RS, Byrne RJ, Rawson PD (2007) Three divergent mitochondrial
809 genomes from California populations of the copepod *Tigriopus*
810 *californicus*. *Gene*, **403**, 53–59.
- 811 Burton RS, Ellison CK, Harrison JS (2006) The sorry state of F2 hybrids:
812 consequences of rapid mitochondrial DNA evolution in allopatric
813 populations. *The American Naturalist*, **168**, S14–24.

814Burton RS, Pereira RJ, Barreto FS (2013) Cytonuclear Genomic
 815 Interactions and Hybrid Breakdown. *Annual Review Of Ecology*
 816 *Evolution And Systematics*, **44**, 281–302.
 817Carneiro M, Albert FW, Afonso S *et al.* (2014) The genomic architecture
 818 of population divergence between subspecies of the european
 819 rabbit (JL Feder, Ed,). *PLoS Genetics*, **10**, e1003519.
 820Charlesworth B, Charlesworth D (2010) *Elements of Evolutionary*
 821 *Genetics*. Roberts Publishers.
 822Christin P-A, Weinreich DM, Besnard G (2015) Causes and evolutionary
 823 significance of genetic convergence. *Trends in Genetics*, **26**, 400–
 824 405.
 825Conesa A, Gotz S, Garcia-Gomez JM *et al.* (2005) Blast2GO: a universal
 826 tool for annotation, visualization and analysis in functional
 827 genomics research. *Bioinformatics*, **21**, 3674–3676.
 828Conte GL, Arnegard ME, Peichel CL, Schluter D (2012) The probability
 829 of genetic parallelism and convergence in natural populations.
 830 *Proceedings Of The Royal Society B-Biological Sciences*, **279**, 5039–
 831 5047.
 832Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH (2005) Why
 833 highly expressed proteins evolve slowly. *Proceedings of the*
 834 *National Academy of Sciences of the United States of America*, **102**,
 835 14338–14343.
 836Edgar RC (2004) MUSCLE: multiple sequence alignment with high
 837 accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
 838Edmands S (1999) Heterosis and outbreeding depression in
 839 interpopulation crosses spanning a wide range of divergence.
 840 *Evolution*, **53**, 1757–1768.
 841Edmands S, Burton R (1999) Cytochrome C oxidase activity in
 842 interpopulation hybrids of a marine copepod: A test for nuclear-
 843 nuclear or nuclear-cytoplasmic coadaptation. *Evolution*, **53**, 1972–
 844 1978.
 845Ellison CK, Burton RS (2008a) Genotype-dependent variation of
 846 mitochondrial transcriptional profiles in interpopulation hybrids.
 847 *Proceedings of the National Academy of Sciences*, **105**, 15831.
 848Ellison CK, Burton RS (2008b) Interpopulation hybrid breakdown maps
 849 to the mitochondrial genome. *Evolution*, **62**, 631–638.
 850Findlay GD, MacCoss MJ, Swanson WJ (2009) Proteomic discovery of
 851 previously unannotated, rapidly evolving seminal fluid genes in
 852 *Drosophila*. *Genome Research*, **19**, 886–896.
 853Foley BR, Rose CG, Rundle DE *et al.* (2011) A gene-based SNP resource
 854 and linkage map for the copepod *Tigriopus californicus*. *BMC*
 855 *genomics*, **12**.
 856Gagnaire P-A, Pavey SA, Normandeau E, Bernatchez L (2013) The
 857 genetic architecture of reproductive isolation during speciation-
 858 with-gene-flow in lake whitefish species pairs assessed by RAD
 859 sequencing. *Evolution*, **67**, 2483–2497.

860 Gayral P, Melo-Ferreira J, Glémin S *et al.* (2013) Reference-Free
861 Population Genomics from Next-Generation Transcriptome Data and
862 the Vertebrate-Invertebrate Gap (JJ Welch, Ed.). *PLoS Genetics*, **9**,
863 e1003457.

864 Gould SJ (2002) *The structure of evolutionary theory*. The Belknap
865 University Press, Cambridge, Massachusetts.

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

869 Haas BJ, Papanicolaou A, Yassour M *et al.* (2013) De novo transcript
870 sequence reconstruction from RNA-seq using the Trinity platform
871 for reference generation and analysis. *Nature Protocols*, **8**, 1494–
872 1512.

873 Harrison JS, Burton RS (2006) Tracing hybrid incompatibilities to single
874 amino acid substitutions. *Mol. Biol. Evol.*, **23**, 559–564.

875 Harvey PH, Pagel MD (1991) *The comparative method in evolutionary*
876 *biology*. Oxford university press, Oxford.

877 Huang X, Madan A (1999) CAP3: A DNA Sequence Assembly Program.
878 *Genome Research*, **9**, 868–877.

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

881 Kimura M (1968) Evolutionary rate at the molecular level. *Nature*, **217**,
882 624–626.

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

885 Li H, Handsaker B, Wysoker A *et al.* (2009) The Sequence
886 Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–
887 2079.

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

891 Lima TG, McCartney MA (2013) Adaptive Evolution of M3 Lysin - A
892 Candidate Gamete Recognition Protein in the *Mytilus edulis* Species
893 Complex. *Molecular Biology and Evolution*, **30**, 2688–2698.

894 Liu L, Yu L (2011) Estimating Species Trees from Unrooted Gene Trees.
895 *Systematic Biology*, **60**, 661–667.

896 Mah SA (2004) Positive Selection in the Carbohydrate Recognition
897 Domains of Sea Urchin Sperm Receptor for Egg Jelly (suREJ)
898 Proteins. *Molecular Biology and Evolution*, **22**, 533–541.

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

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

905 Nowick K, Carneiro M, Faria R (2013) A prominent role of KRAB-ZNF

transcription factors in mammalian speciation? *TIG*, **29**, 130–139.

Obbard DJ, Welch JJ, Kim K-W, Jiggins FM (2009) Quantifying Adaptive Evolution in the *Drosophila* Immune System (DJ Begun, Ed.). *PLoS Genetics*, **5**, e1000698–13.

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 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.

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

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. *Mol. Biol. Evol.*, **13**, 735–748.

Rand DM, Haney RA, Fry AJ (2004) Cytonuclear coevolution: the genomics of cooperation. *TREE*, **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., **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

952 population genomics in animals uncovers the determinants of
 953 genetic diversity. *Nature*, **515**, 261–263.
 954 Schmieder R, Edwards R (2011) Quality control and preprocessing of
 955 metagenomic datasets. *Bioinformatics*, **27**, 863–864.
 956 Shaw TI, Ruan Z, Glenn TC, Liu L (2013) STRAW: Species TRee Analysis
 957 Web server. *Nucleic Acids Res.*, **41**, W238–W241.
 958 Smeitink J, van den Heuvel L, DiMauro S (2001) The genetics and
 959 pathology of oxidative phosphorylation. *Nature Reviews Genetics*,
 960 **2**, 342–352.
 961 Soria-Carrasco V, Gompert Z, Comeault AA *et al.* (2014) Stick Insect
 962 Genomes Reveal Natural Selection's Role in Parallel Speciation.
 963 *Science*, **344**, 738–742.
 964 Stern DL, Orgogozo V (2009) Is genetic evolution predictable? *Science*,
 965 **323**, 746–751.
 966 Stoletzki N, Eyre-Walker A (2011) Estimation of the neutrality index.
 967 *Molecular Biology and Evolution*, **28**, 63–70.
 968 Subramanian S (2004) Gene Expression Intensity Shapes Evolutionary
 969 Rates of the Proteins Encoded by the Vertebrate Genome. *Genetics*,
 970 **168**, 373–381.
 971 Swanson WJ, Clark AG, Waldrip-Dail HM, Wolfner MF, Aquadro CF
 972 (2001a) Evolutionary EST analysis identifies rapidly evolving male
 973 reproductive proteins in *Drosophila*. *Proceedings of the National*
 974 *Academy of Sciences of the United States of America*, **98**, 7375–
 975 7379.
 976 Swanson WJ, Wong A, Wolfner MF, Aquadro CF (2004) Evolutionary
 977 expressed sequence tag analysis of *Drosophila* female reproductive
 978 tracts identifies genes subjected to positive selection. *Genetics*,
 979 **168**, 1457–1465.
 980 Swanson WJ, Zhang ZH, Wolfner MF, Aquadro CF (2001b) Positive
 981 Darwinian selection drives the evolution of several female
 982 reproductive proteins in mammals. *Proceedings of the National*
 983 *Academy of Sciences of the United States of America*, **98**, 2509–
 984 2514.
 985 Tsagkogeorga G, Cahais V, Galtier N (2012) The Population Genomics
 986 of a Fast Evolver: High Levels of Diversity, Functional Constraint,
 987 and Molecular Adaptation in the Tunicate *Ciona intestinalis*.
 988 *Genome Biology and Evolution*, **4**, 740–749.
 989 Wake DB (1991) Homoplasy: the result of natural selection, or
 990 evidence of design limitations? *American Naturalist*, **138**, 543–567.
 991 Willett CS (2012) Quantifying the Elevation of Mitochondrial DNA
 992 Evolutionary Substitution Rates Over Nuclear Rates in the Intertidal
 993 Copepod *Tigriopus californicus*. *Journal of Molecular Evolution*, **74**,
 994 310–318.
 995 Willett CS, Burton RS (2004) Evolution of interacting proteins in the
 996 mitochondrial electron transport system in a marine copepod. *Mol.*
 997 *Biol. Evol.*, **21**, 443–453.

998 Willett CS, Ladner JT (2009) Investigations of fine-scale
999 phylogeography in *Tigriopus californicus* reveal historical patterns
1000 of population divergence. *BMC Evolutionary Biology*, **9**, 139.
1001 Wolf JBW, Künstner A, Nam K, Jakobsson M, Ellegren H (2009)
1002 Nonlinear dynamics of nonsynonymous (dN) and synonymous (dS)
1003 substitution rates affects inference of selection. *Genome Biology*
1004 *and Evolution*, **1**, 308–319.
1005 Yang Z (2007) PAML 4: Phylogenetic Analysis by Maximum Likelihood.
1006 *Mol. Biol. Evol.*, **24**, 1586–1591.
1007 Yang Z, Nielsen R (2000) Estimating synonymous and nonsynonymous
1008 substitution rates under realistic evolutionary models. *Mol. Biol.*
1009 *Evol.*, **17**, 32–43.
1010
1011

1012 **Data Accessibility**

1013 Raw reads were deposited in the NCBI Sequence Read Archive (SRQ
1014 Acc. ###). The dryad archive (doi: ####) contains: all de novo
1015 assemblies, list of homologous genes, complete polymorphism data
1016 set, alignments for mitochondrial and autosomal genes, bootstrapped
1017 alignments and resulting tree file, p-distance estimates, estimates of
1018 dN, dS, omega and DoS, read counts and RPKM estimates. Parsing
1019 scripts are available in a public GitHub depository
1020 (<https://github.com/bluegenes/repeatability>).
1021

1023**Table 1.** Positive selection in proteins with repeated accelerated evolution.

<i>Protein name</i>	<i>omega overlap</i>	<i>Alignme nt size (bp)</i>	<i>likelihood</i>		<i>p- value</i>	<i>p</i>	ω
			<i>neutr al</i>	<i>selecti on</i>			
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		3378					
LOC101241641	10%		-6764.7	-6761.0	0.023	0.01	8.48

1024Omega overlap = presence among the top 5 or 10% on each interpopulation comparison according to ω
1025 values averaged for the entire protein alignment; proteins in bold are referred in the text. p =
1026 proportion of sites estimated to be under positive selection with $\omega > 1$; ω = value of ω in those sites.

1027 **Figure Captions**

1028

1029 **Figure 1. Geographic isolation in *Tigriopus californicus* results**
1030 **in a continuum of population divergence. A.** Population tree (NJst;
1031 Liu & Yu 2011). Branch lengths reflect average genetic divergence
1032 (Table S1) and numbers represent bootstrap support for topology.
1033 “Early”, “middle” and “late” refer to the three stages of population
1034 divergence tested for repeatability of genomic divergence. **B.** Sampling
1035 of allopatric populations of copepods.

1036 **Figure 2. Interpopulation variability changes exponentially**
1037 **with time since divergence.** Points represent all pairwise
1038 comparisons between the six populations for fixed (**A**) and shared (**B**)
1039 mutations. Solid lines represent the exponential model of SNP variation
1040 relative to the percentage of genetic divergence (proxy for time);
1041 dashed lines represent the linear model.

1042 **Figure 3. Population differentiation is largely affected by**
1043 **genetic drift. A.** Relationship between π_N/π_S and π_S . Grey symbols
1044 refer to published data in non-model species (Romiguier et al. 2014),
1045 while black symbols refer to the six allopatric populations of *T.*
1046 *californicus*. Lines reflect the linear models for all data ($r^2 = 0.35$, $p <$
1047 10^{-8}). **B.** Genetic diversity within populations of *T. californicus* shown as
1048 the fraction of polymorphic SNPs in neutral sites (π_S).

1049 **Figure 4. Proteins with accelerated evolution are predictable at**
1050 **independent vicariant events and along the continuum of**
1051 **species formation.** Based on ω estimates in the early a comparison
1052 (SCN-PES), 5,677 genes were classified either as candidate genes
1053 under accelerated evolution (mean $\omega \geq 0.5$) or neutral (mean $\omega < 0.5$).
1054 We tested whether these candidate genes have higher signature of
1055 rapid evolution (ω) than the rest of the genome at an independent

vicariant event with similar divergence (early b: SD-BR) and at latter 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; $p < 2 \times 10^{-6}$).

Figure 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 less than 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 ($p\text{-values} < 10^{-10}$). Box plots show differences between overlap and the remaining genes; $p\text{-values}$ are denoted above.

Figure 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; $p\text{-values}$ are denoted.

1075

Figure 1

A. Population tree

B. Geographic sampling

Figure 2

1092

Figure 3

1093

1094

1095

1096

1097

1098

A. Genetic diversity between species

B. Genetic diversity within populations

Figure 4

Figure 5:

1115

Figure 6:

1116

1117

1118

1119 **Supplementary Information Supplementary Figure Legends**

1120 **Figure S1. Quality of the de novo assemblies for the**
1121 **orthologous genes. A.** Density plot of contig size relative to the SD
1122 **de novo assembly. B.** Number of mapped reads (log scale) per contig,
1123 normalized by library size and gene length. Contigs are ordered by
1124 coverage in SD population.

1125 **Figure S2. Genetic variability sorts exponentially with time**
1126 **since population divergence in independent population**
1127 **comparisons.** Points represent comparisons between SD-BR, SCN-BB,
1128 and AB-PES, respectively for early, middle and late population
1129 divergence. Solid lines represent the exponential model of SNP
1130 variation relative to the percentage of genetic divergence (proxy for
1131 time); dashed lines represent the linear model.

1132 **Figure S3. Genome-wide protein divergence during the**
1133 **continuum of species formation in *Tigriopus californicus*.** d_N
1134 refers to the rate of non-synonymous substitutions on each of the
1135 7,018 proteins used on each population comparison. Each plot refers to
1136 a population comparison, order by increasing stages of divergence (left
1137 to right). Percentages refer to average pairwise divergence across the
1138 11,560 nuclear loci larger than 100 bp. The distribution of d_N for the
1139 independent population pairs at each stage of divergence are not
1140 statistically different from each other (test of equality; $P_{\text{early a} - \text{early b}} =$
1141 0.44 , $P_{\text{late a} - \text{late b}} = 0.53$), but distributions of d_N for population pairs at
1142 consecutive stages of divergence are statistically different (test of
1143 equality; $P_{\text{early a} - \text{middle}} = 0.01$, $P_{\text{early b} -$
1144 $\text{middle}} = 0.01$, $P_{\text{middle} - \text{late a}} < 0.01$, $P_{\text{middle} - \text{late b}} < 0.01$).

1145 **Figure S4. Null distributions of gene overlap.** Simulations for 5
1146 and 10% gene overlap

1147 out of a total of 7,018 genes and across three population comparisons.

1148 **Figure S5. Pairwise estimation of adaptive divergence during**
1149 **the continuum of species formation in *Tigriopus californicus*.**
1150 Values of adaptive divergence (ω) were calculated for 5,677 proteins
1151 in the five focal interpopulation comparisons ($\omega = d_N/d_S$, where d_N is
1152 the rate of non-synonymous substitutions and d_S is the rate of
1153 synonymous

1154 substitutions). **A.** Histograms of ω for each pair wise comparison. **B.**
1155 Relationship between d_N and d_S at each comparison.

1156 **Figure S6. The number of genes with signature of accelerated**
1157 **evolution decreases with time since divergence. A.** Counts of
1158 candidate genes under selection, using a threshold of mean ω per
1159 locus > 0.5 (open squares) or 1 (close squares). **B.** Accumulation of
1160 synonymous (d_S , close circles) and non-synonymous substitutions (d_N ,
1161 open circles) averaged across 7,019 genes; r^2 for linear models forced
1162 through the origin is 0.99 ($p < 0.01$).

1163 **Figure S7. Correlation between gene expression and rates of**
1164 **protein divergence in populations of *Tigriopus californicus*.** Dot
1165 plots refer to values of expression (RPKM) and divergence (d_N/d_S);
1166 larger dots refer to values in the genes with overlap of d_N/d_S across the
1167 five focal population comparisons (i.e. candidate genes under adaptive
1168 evolution); lines refer to linear models on each correlation (p -values $<$
1169 10^{-10}). Box plots shows differences between overlap and the remaining
1170 genes; p -values are denoted above.

1171

1172

1173

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

	Populations					
	SD	BR	AB	SCN	PES	BB
reads' curation:						
paired raw reads	63,968,351	80,018,732	54,517,600	24,059,252	85,913,595	107,003,046
paired filtered reads	61,278,883	75,984,145	54,507,073	22,230,685	85,902,726	103,817,588
de novo assemblies for nDNA:						
contigs	76,216	108,509	80,879	55,252	105,344	10,6311
min. size (bp)	201	201	201	201	201	201
max. size (bp)	24,210	24,972	24,433	17,357	19,100	27,638
CEGMA completeness (% of partial genes)	99.6	99.6	99.2	99.6	98.8	98.0
orthologs:						
contigs	12,573	12,573	12,573	12,573	12,573	12,573
N50	2,941	3,204	2,750	2,640	2,575	2,914
max. size (bp)	18,221	24,001	16,675	16,587	18,947	17,180
mtDNA						
% N	0	0.15	0.15	0.03	0.05	4.62
mapping SNPs						
% mapped reads	98.72	98.33	98.96	97.72	99.08	99.35
% uniquely mapped	66.64	57.94	68.49	64.98	68.93	69.33

1174

1175

Table S2. Average genetic divergence between all pairwise population comparisons of *Tigriopus californicus*. Upper diagonal contains averages of uncorrected percentage divergence based of 11,560 nuclear loci with a length of at least 100bp. Lower Diagonal contains averages of uncorrected percentage divergence based on the 13 mitochondrial genes. Colors highlight the focal pairwise population comparisons that describe the three levels of divergence: early (blue), middle (green), and late (red).

	SD	BR	AB	SCN	PES	BB	
SD		0.69	2.30	2.77	2.70	2.76	% average nDNA
BR	9.94		2.36	2.81	2.76	2.82	
AB	21.28	21.22		2.54	2.47	2.55	
SCN	22.17	22.02	21.25		0.65	1.36	
PES	22.27	22.12	21.38	1.26		1.30	
BB	22.98	23.09	22.25	16.06	16.05		
% average mtDNA							

Table S3. Genetic diversity among populations of *Tigriopus californicus*. Data based on 12,044 nuclear genes; the alignment across the six populations covers 16,415,579 base pairs (bp). The comparisons on top refer to counts of SNPs, while the comparisons on the bottom refer to the percentage relative to the total number of SNPs in that comparison.

A. Fixed polymorphisms

	SD	BR	AB	SCN	PES	BB
SD	-	90,322	326,546	390,523	388,618	395,992
BR	75.22%	-	339,669	402,943	401,528	408,766
AB	92.48%	91.98%	-	363,635	360,339	369,688
SCN	90.83%	90.47%	90.37%	-	89,015	191,750
PES	90.38%	90.04%	89.86%	62.74%	-	189,505
BB	91.67%	91.27%	91.27%	79.92%	79.09%	-

B. Shared polymorphisms

	SD	BR	AB	SCN	PES	BB
SD	-	715	205	162	164	228
BR	0.60%	-	128	107	113	129
AB	0.06%	0.03%	-	112	140	136
SCN	0.04%	0.02%	0.03%	-	783	158
PES	0.04%	0.03%	0.03%	0.55%	-	186
BB	0.05%	0.03%	0.03%	0.07%	0.08%	-

C. Polymorphisms within population A (left column)

	SD	BR	AB	SCN	PES	BB
SD	-	13,026	13,536	13,579	13,577	13,513
BR	10.85%	-	16,606	16,627	16,621	16,605
AB	3.83%	4.50%	-	12,887	12,859	12,863
SCN	3.16%	3.73%	3.20%	-	25,059	25,684
PES	3.16%	3.73%	3.21%	17.66%	-	27,613
BB	3.13%	3.71%	3.18%	10.70%	11.52%	-

D. Polymorphisms within population B (top row)

	SD	BR	AB	SCN	PES	BB
SD	-	16,019	12,794	25,680	27,635	22,264
BR	13.34%	-	12,871	25,735	27,686	22,363
AB	3.62%	3.49%	-	25,730	27,659	22,356
SCN	5.97%	5.78%	6.39%	-	27,016	22,334
PES	6.43%	6.21%	6.90%	19.04%	-	22,306
BB	5.15%	4.99%	5.52%	9.31%	9.31%	-

Table S4. Overlaps in fast-evolving proteins across the five focal population comparisons. Contig name refers to the orthologous sequence in the population SD; protein name refers to identification according to BLAST; percentile of overlap refers to whether that protein is consistently among the top 5 or 10% on each comparison; d_N values refer to the number of the non-synonymous substitutions averaged across the entire gene alignment. Proteins with higher mutation rate (overlap among the top 10% d_S) are demarked with *.

Contig name	Protein name	percentile of overlap	dN				
			early a (SCN-PES)	early b (SD-BR)	middle (SCN-BB)	late a (SD-AB)	late b (BR-BB)
comp46825_c0_seq2_33874	kunitz bovine pancreatic trypsin inhibitor domain containing protein	5%	0.019	0.019	0.016	0.027	0.026
comp49487_c0_seq1_39816	pcna-interacting partner-like	5%	0.009	0.013	0.016	0.027	0.027
Contig1153_3051	g-protein coupled receptor mth2-like	5%	0.026	0.007	0.023	0.027	0.029
comp24755_c0_seq1_13190	serine protease 27-like	5%	0.014	0.010	0.015	0.030	0.029
comp47547_c0_seq1_35336	pan domain-containing	5%	0.010	0.012	0.020	0.027	0.029
Contig880_2805	fha domain protein	5%	0.008	0.009	0.021	0.030	0.032
comp49915_c0_seq2_40969	type-2 ice-structuring	5%	0.009	0.010	0.016	0.022	0.032
Contig4124_5788	low-density lipoprotein receptor-related protein 2	5%	0.007	0.013	0.018	0.034	0.032
comp40389_c1_seq1_25220	hypothetical protein CAPTEDRAFT_207764	5%	0.007	0.015	0.020	0.028	0.033
comp42303_c0_seq1_27516	thap domain-containing protein 2-like	5%	0.009	0.020	0.015	0.024	0.034
Contig435_434	predicted protein	5%	0.017	0.039	0.044	0.036	0.035
comp49189_c0_seq1_39026	juvenile hormone-inducible protein	5%	0.023	0.014	0.025	0.031	0.035
comp49346_c1_seq1_39479	vesicle-associated membrane protein-associated protein a-like	5%	0.012	0.007	0.024	0.025	0.036
comp47301_c1_seq2_34848	cg13310 cg13310-pa *	5%	0.015	0.016	0.037	0.030	0.038
Contig5739_7292	cg2206-pb-like protein	5%	0.021	0.039	0.034	0.049	0.039
Contig1491_3362	thap domain-containing protein 9	5%	0.007	0.009	0.014	0.034	0.039
comp24404_c0_seq1_12881	g patch domain-containing protein 4	5%	0.012	0.008	0.021	0.026	0.040
comp50358_c0_seq1_42674	protein 5nuc	5%	0.009	0.010	0.027	0.043	0.041
comp49150_c0_seq1_38915	dna repair protein xp-c	5%	0.010	0.010	0.019	0.036	0.041
comp26646_c0_seq1_14205	kelch domain-containing protein 8b	5%	0.055	0.045	0.041	0.034	0.042
mtDNA_cox3	mtDNA_cox3 *	5%	0.013	0.022	0.031	0.044	0.043
comp45359_c0_seq1_31481	histone h1-delta	5%	0.010	0.014	0.026	0.039	0.048
comp41370_c0_seq1_26377	upf0462 protein c4orf33 homolog isoform 1	5%	0.022	0.063	0.029	0.037	0.051
comp44900_c1_seq1_30822	c-type lectin domain family member f	5%	0.018	0.010	0.025	0.039	0.052
comp47900_c0_seq1_36216	hypothetical protein DAPPUDRAFT_262406	5%	0.013	0.017	0.023	0.070	0.055
comp39026_c0_seq1_23636	pms1 protein homolog 1	5%	0.011	0.011	0.022	0.047	0.057
comp50791_c0_seq4_44466	krab-a domain-containing protein *	5%	0.008	0.033	0.043	0.045	0.057
Contig5703_7258	receptor-type tyrosine-protein phosphatase alpha-like	5%	0.009	0.030	0.026	0.046	0.063
comp50452_c0_seq2_42994	cg13310 cg13310-pa	5%	0.019	0.015	0.037	0.026	0.067
comp44784_c0_seq1_30670	aleurain-like protease	5%	0.046	0.024	0.029	0.056	0.072
comp56160_c0_seq1_44688	oryzain gamma chain precursor	5%	0.032	0.012	0.039	0.079	0.073
comp42532_c0_seq1_27781	kelch-like 4 isoform cra_c *	5%	0.091	0.067	0.054	0.057	0.078
Contig3238_4951	e3 ubiquitin-protein ligase rnf8	5%	0.018	0.016	0.023	0.092	0.080
comp48875_c0_seq2_38290	cysteine protease	5%	0.010	0.028	0.019	0.092	0.091
comp28601_c0_seq1_15372	poly protein	5%	0.038	0.012	0.069	0.038	0.092
comp44512_c0_seq1_30303	hypothetical protein BRAFLDRAFT_120870	5%	0.025	0.017	0.042	0.046	0.122
mtDNA_nad2	mtDNA_nad2 *	5%	0.009	0.046	0.037	0.147	0.132
mtDNA_nad5	mtDNA_nad5 *	5%	0.009	0.035	0.068	0.134	0.142
mtDNA_nd4L	mtDNA_nd4L *	5%	0.010	0.019	0.181	0.133	0.229
comp45715_c0_seq1_32043	crossover junction endonuclease mus81	10%	0.008	0.006	0.016	0.024	0.019
comp43057_c0_seq1_28420	gtp-binding protein 5-like	10%	0.005	0.008	0.013	0.020	0.020
comp41251_c0_seq1_26229	hypothetical protein EAI_15411	10%	0.011	0.011	0.010	0.017	0.020
Contig776_2709	mki67 fha domain-interacting nucleolar phospho	10%	0.006	0.006	0.011	0.020	0.020
comp43626_c0_seq1_29206	core 1 udp-galactose:n-acetylglactosamine-alpha-r beta -	10%	0.015	0.011	0.025	0.017	0.020
Contig853_2780	zinc finger ccch-type with g patch domain-containing	10%	0.005	0.006	0.013	0.017	0.020
comp48292_c0_seq1_36935	broad-complex core protein isoform 6; zinc finger and BTB domain	10%	0.013	0.017	0.013	0.032	0.021
comp48362_c0_seq1_37103	kelch domain protein	10%	0.006	0.006	0.014	0.017	0.021
comp47098_c0_seq1_34449	Fucoatlectin	10%	0.015	0.011	0.015	0.026	0.021
comp868721_c0_seq1_51456	retrovirus-related pol polyprotein from	10%	0.014	0.045	0.022	0.061	0.021

Contig1109_1108	transposon tnt 1-94						
comp48034_c0_seq1_36476	engulfment and cell motility protein 1	10%	0.013	0.014	0.016	0.034	0.021
comp48388_c0_seq6_37152	dna replication licensing factor mcm9	10%	0.005	0.007	0.010	0.024	0.021
Contig4123_5787	upf0532 protein cg3570-like	10%	0.008	0.007	0.010	0.020	0.021
comp23575_c0_seq1_12535	hypothetical protein Phum_PHUM300620	10%	0.005	0.005	0.014	0.018	0.021
	leucine-rich repeat-containing protein c10orf11-like protein	10%	0.007	0.005	0.014	0.017	0.022
comp34904_c1_seq1_19830	lactosylceramide 4-alpha-galactosyltransferase	10%	0.027	0.008	0.025	0.023	0.022
comp50732_c0_seq1_44240	hypothetical protein TRIADDRAFT_60504	10%	0.005	0.006	0.012	0.018	0.022
Contig1146_3045	mdm2-binding protein	10%	0.005	0.006	0.017	0.018	0.023
comp48320_c0_seq1_37012	PREDICTED: hypothetical protein LOC100741340 isoform 1	10%	0.007	0.008	0.014	0.024	0.024
Contig3489_5193	zinc finger protein 91-like	10%	0.045	0.010	0.045	0.018	0.025
comp44353_c0_seq1_30118	zinc finger protein 346	10%	0.007	0.014	0.011	0.025	0.025
comp29107_c0_seq1_15682	PREDICTED: hypothetical protein LOC100551595	10%	0.014	0.006	0.026	0.033	0.025
Contig1788_3633	gag-pol polyprotein	10%	0.072	0.006	0.032	0.023	0.025
comp74067_c0_seq1_44890	copper transporter isoform a	10%	0.006	0.006	0.012	0.022	0.025
comp37705_c0_seq1_22192	peptidyl-tRNA hydrolase mitochondrial-like	10%	0.005	0.009	0.012	0.026	0.026
comp44168_c0_seq1_29856	nurim homolog	10%	0.005	0.009	0.011	0.016	0.026
comp49045_c0_seq2_38677	transcriptional regulator atrx	10%	0.007	0.012	0.016	0.018	0.026
comp44591_c0_seq4_30421	bone morphogenetic protein receptor type-2	10%	0.008	0.008	0.015	0.016	0.026
comp42689_c0_seq1_27954	thap domain-containing protein 4	10%	0.007	0.007	0.010	0.035	0.027
comp49057_c0_seq1_38695	ben, isoform b	10%	0.010	0.007	0.016	0.020	0.027
Contig5957_7496	glyceraldehyde-3-phosphate dehydrogenase	10%	0.007	0.006	0.010	0.027	0.028
Contig2875_4623	nose resistant to fluoxetine protein 6-like	10%	0.011	0.007	0.014	0.021	0.028
comp38061_c0_seq1_22564	platelet glycoprotein 4	10%	0.005	0.011	0.010	0.016	0.028
Contig5968_7505	microtubule-associated protein futsch	10%	0.006	0.012	0.016	0.040	0.030
comp24463_c0_seq1_12934	39s ribosomal protein mitochondrial-like	10%	0.010	0.007	0.015	0.023	0.031
comp30894_c0_seq1_16699	alpha1,6-fucosyltransferase	10%	0.018	0.007	0.044	0.038	0.032
comp36541_c0_seq1_21195	unnamed protein product	10%	0.006	0.008	0.024	0.019	0.032
comp48146_c0_seq1_36699	PREDICTED: uncharacterized protein LOC100180758	10%	0.006	0.036	0.024	0.044	0.033
Contig2258_4071	thap domain-containing protein 4	10%	0.007	0.006	0.021	0.027	0.034

comp47479_c0_seq1_35212	zinc finger protein 510	10%	0.007	0.007	0.014	0.021	0.034
comp46413_c0_seq1_33115	thap domain-containing protein 9-like	10%	0.008	0.007	0.022	0.017	0.034
comp47604_c0_seq3_35466	beta-1,4-galactosyltransferase	10%	0.008	0.011	0.013	0.018	0.035
comp24876_c0_seq1_13251	trans-golgi network integral membrane protein 2	10%	0.006	0.006	0.013	0.028	0.035
Contig393_2367	peptidoglycan-recognition protein precursor	10%	0.011	0.007	0.013	0.038	0.038
comp43267_c0_seq1_28724	poly(adp-ribose) glycohydrolase	10%	0.005	0.007	0.014	0.044	0.038
Contig1667_1666	zinc finger protein 26	10%	0.009	0.005	0.022	0.020	0.039
comp49432_c0_seq1_39678	e3 ubiquitin-protein ligase rad18	10%	0.009	0.012	0.017	0.017	0.040
comp41527_c0_seq1_26555	e3 ubiquitin-protein ligase dzp3	10%	0.005	0.007	0.019	0.026	0.040
comp48692_c0_seq1_37843	tecd-inducible-parp-like domain-containing protein	10%	0.008	0.007	0.015	0.042	0.041
comp50064_c1_seq6_41474	zinc finger protein 707	10%	0.010	0.012	0.012	0.033	0.041
comp47708_c0_seq1_35795	adenosine deaminase	10%	0.006	0.011	0.016	0.031	0.047
comp7605_c0_seq1_10669	PREDICTED: uncharacterized protein LOC100893948	10%	0.006	0.012	0.012	0.022	0.050
Contig184_183	hypothetical protein TRIADDRAFT_51346	10%	0.006	0.018	0.026	0.051	0.054
comp46009_c0_seq1_32468	conserved hypothetical protein	10%	0.007	0.007	0.049	0.039	0.055
Contig1922_1921	cysteine proteinase aalp	10%	0.041	0.006	0.089	0.058	0.057
comp46243_c0_seq1_32838	lipoma-preferred partner homolog *	10%	0.006	0.008	0.049	0.026	0.058
Contig1921_1920	isoform a	10%	0.005	0.006	0.030	0.039	0.059
comp35187_c0_seq1_20070	hypothetical protein DAPPUDRAFT_316213	10%	0.033	0.007	0.034	0.032	0.061
Contig1378_1377	loc553381 protein	10%	0.044	0.006	0.065	0.052	0.062
Contig935_2857	protein dep- isoform b	10%	0.023	0.006	0.035	0.052	0.064
mtDNA_nad1	mtDNA_nad1 *	10%	0.006	0.020	0.033	0.072	0.076
comp42340_c0_seq1_27563	sh3 domain-binding protein 5-like protein	10%	0.007	0.017	0.055	0.126	0.124
mtDNA_nad4	mtDNA_nad4 *	10%	0.006	0.043	0.071	0.133	0.135
mtDNA_nad6	mtDNA_nad6 *	10%	0.006	0.030	0.056	0.128	0.149

1180

1181 **Supplementary Figure 1**

A. size symmetry across orthologs

1182

B. coverage across orthologs

1183 **Supplementary Figure 2**

A. fixed mutations

B. shared mutations

1184

% genetic divergence

1185 **Supplementary Figure 3**

1186

1187 [Supplementary Figure 4](#)

1188

1189 [A.](#)

1190 [Supplementary Figure 5](#)

1191

1192 **B.**

1193

1194 **Supplementary Figure 6**

A. Candidate genes under positive selection

B. Divergence at all genes

1195

1196 **Supplementary Figure 7**

1197
1198