MOLECULAR ECOLOGY

Molecular Ecology (2011) **20**, 560–572 doi: 10.1111/j.1365-294X.2010.04963.x

Interpopulation patterns of divergence and selection across the transcriptome of the copepod *Tigriopus californicus*

FELIPE S. BARRETO, GARY W. MOY and RONALD S. BURTON

Marine Biology Research Division, Scripps Institution of Oceanography, University of California, San Diego, La Jolla, CA 92037, USA

Abstract

The accumulation of genetic incompatibilities between isolated populations is thought to lead to the evolution of intrinsic postzygotic isolation. The molecular basis for these mechanisms, however, remains poorly understood. The intertidal copepod *Tigriopus californicus* provides unique opportunities for addressing mechanistic questions regarding the early stages of speciation; hybrids between highly divergent populations are fertile and viable, but exhibit reduced fitness at the F_2 or later generations. Given the current scarcity of genomic information in taxa at incipient stages of reproductive isolation, we utilize high-throughout 454 pyrosequencing to characterize a substantial fraction of protein-coding regions (the transcriptome) of *T. californicus*. Our sequencing effort was divided equally between two divergent populations in order to estimate levels of divergence and to reveal patterns of selection across the transcriptome. Assembly of sequences generated over 40 000 putatively unique transcripts (unigenes) for each population, 19 622 of which were orthologous between populations. BLAST searches of public databases determined protein identity and functional features for 15 402 and 12 670 unigenes, respectively. Based on rates of nonsynonymous and synonymous substitutions in 5897 interpopulation orthologs (those >150 bp and with at least 2X coverage), we identified 229 potential targets of positive selection. Many of these genes are predicted to be involved in several metabolic processes, and to function in hydrolase, peptidase and binding activities. The library of *T. californicus* coding regions, annotated with their predicted functions and level of divergence, will serve as an invaluable resource for elucidating molecular mechanisms underlying the early stages of speciation.

Keywords: comparative transcriptomics, next-generation sequencing, population divergence, positive selection

Received 11 August 2010; revision received 31 October 2010; accepted 9 November 2010

Introduction

Complete reproductive isolation between populations is often preceded by a stage of reduced hybrid fitness, brought upon by genetic divergence due to selection, mutation and drift. Such hybrid breakdown normally occurs only at the F_2 and later generations, with F_1 hybrids showing similar or sometimes superior fitness to that of their parents (Endler 1977; Harrison 1990). Postzygotic barriers of this kind are generally explained by the Dobzhansky-Muller model (Dobzhanksy 1936; Muller 1942). According to this model, different positive epistatic interactions evolve in isolated populations, but the divergent allelic states are incompatible when they recombine in the F_2. Dobzhansky-Muller incompatibilities have been invoked to explain patterns of postzygotic isolating barriers in numerous systems (Arntzen *et al*. 2009; Moyle & Nakazato 2009; Stelkens *et al*. 2009; reviewed in Coyne & Orr 2004); the molecular mechanisms underlying these

Correspondence: Felipe Barreto, Fax: +1 858 534 7313;
E-mail: fbarreto@ucsd.edu

patterns, however, are still poorly understood (Coyne & Orr 2004).

Among animals, the intertidal copepod *Tigriopus californicus* provides both a well-described example of hybrid breakdown and a highly tractable system for addressing mechanistic questions about the early stages of reproductive isolation. This species is generally abundant throughout its range, from central Baja California, Mexico to Alaska, but gene flow is highly restricted among populations, as evidenced by mitochondrial DNA (mtDNA) divergence in excess of 18% (Burton & Lee 1994; Burton 1998; Edmands 2001). Interpopulation crosses in the laboratory reveal that, while F_1 hybrids have similar or higher fitness than their parents, subsequent generations experience significant fitness impairments, such as decreased tolerance to hyperosmotic stress (Burton 1986), slower development (Burton 1990) and lower fecundity and viability (Edmands 1999). Furthermore, recent studies have shown that hybrid breakdown in *T. californicus* occurs at least partly because of nuclear-mitochondrial incompatibilities exposed in recombinant genomes (Edmands & Burton 1999; Ellison & Burton 2008a; Rawson & Burton 2002). Reduced viability of certain hybrid crosses during salinity stress, for instance, seems to be associated with a disruption of the mitochondrial transcriptional machinery (Ellison & Burton 2008b). These studies suggest that population-specific intergenomic coadaptation may be driven largely by rapid mtDNA evolution, and that coadaptation may be broken down by hybridization at the F_2 or later generations (reviewed in Burton *et al.* 2006).

The mitonuclear interactions responsible for hybrid breakdown are likely not restricted to a few sets of interacting genes. As many as 700–1500 nuclear-encoded proteins function in the mitochondrion (Marcotte *et al.* 2000; Taylor *et al.* 2003; Reichert & Neupert 2004), providing numerous opportunities for coadaptation. In addition, there is evidence that nuclear–nuclear epistasis is also involved in hybrid breakdown in *T. californicus* (Edmands *et al.* 2009). Nucleotide divergence across the entire mitochondrial genome is known to be high (~22%; Burton *et al.* 2007), but levels of divergence have been investigated in relatively few nuclear genes (Rawson *et al.* 2000; Willett & Burton 2003, 2004). Discovery of functionally important genetic changes would hence be facilitated by genomic information, which is currently lacking for *T. californicus*.

In recent years, advances in high-throughput sequencing and bioinformatics have made genome-level information from nonmodel organisms increasingly accessible (Hudson 2008). Transcriptome sequencing, in particular, has been successfully accomplished in several nonmodel eukaryotes, enabling detection of single nucleotide polymorphisms (SNPs), discovery and annotation of genes, genome-wide estimates of diversity, and functional studies of gene expression (e.g.: Vera *et al.* 2008; Hahn *et al.* 2009; Meyer *et al.* 2009; Roeding *et al.* 2009; Schwarz *et al.* 2009; Elmer *et al.* 2010; Renaut *et al.* 2010). A useful initial approach is to perform whole-transcriptome sequencing, which can be annotated based on similarity to current protein databases (e.g. GenBank and EMBL). The resulting custom database of protein-coding loci can then be mined for functions of interest (Schwarz *et al.* 2009; Zagrobelny *et al.* 2009) or scanned for regions undergoing accelerated differentiation (Elmer *et al.* 2010; Künstner *et al.* 2010; Renaut *et al.* 2010).

In the current study, we employ high-throughput 454 pyrosequencing (Roche Life Sciences; Margulies *et al.* 2005) to characterize a large portion of the transcriptome of two geographically isolated populations of *T. californicus* (San Diego and Santa Cruz, CA, USA), between which hybrid breakdown occurs. This technology generates, per run, more than 500 000 single-pass sequences that are sufficiently long (50–500 bp) to permit *de novo* assembly of nonmodel transcriptomes (Vera *et al.* 2008; Kristiansson *et al.* 2009; Elmer *et al.* 2010; Renaut *et al.* 2010; reviewed in Wheat 2010). After the initial task of annotating thousands of coding regions with their putative functions, we examine the extent of nucleotide and protein sequence divergence throughout the genome, as well as identify genes and gene functions that may be under positive selection. To our knowledge, our study is one of the first to provide comparison of transcriptome-wide sequences between intraspecific populations of a nonmodel species, and hence, it provides an essential foundation for investigating the molecular mechanisms at the early stages of speciation.

Materials and methods

Sample preparation and sequencing

Copepods were collected from high intertidal rocky pools in San Diego (SD: 32°45′N, 117°15′W) and Santa Cruz (SC: 36°57′N, 122°03′W), California, and kept *en masse* in 1-litre beakers of 0.2-μm-filtered seawater at 20°C with a 12-hour light:dark cycle. The SD sample was split into three roughly equal samples and subjected to the following treatments: (i) 100% seawater treatment (control); (ii) 30 min hyperosmotic stress (two day acclimation to 50% seawater followed by transfer to 100% seawater for 30 min before RNA extraction); and (iii) 24-h hyperosmotic stress (the same 50–100% seawater transfer but for 24 h). The SC sample was maintained in control 100% seawater. Copepods at all developmental stages were pooled for extraction. Total RNA was extracted from each of the four samples

(>0.5 g wet weight each) using Tri Reagent (Sigma-Aldrich), and mRNA was extracted from the total RNA using Poly(A) Purist kit (Ambion). Complementary DNA (cDNA) libraries were prepared using Roche's GS FLX Titanium Rapid Library Preparation Kit, with the three SD samples barcoded with multiplex identifier adaptors. These libraries were not normalized because we intend to use the resulting sequence reads in another study aimed at documenting differential gene expression. The barcoded SD samples were pooled and run on half of the sequencing plate; the SC sample was run on the second half of the plate. Sequencing was performed by Roche Life Sciences.

De novo assemblies and annotation

We used the CLC Genomics Workbench 3.7 (CLC Bio) software to trim adapter sequences and low-quality bases from all reads, and then to assemble reads into contigs for each population (assembly parameters: similarity = 0.9, fraction length = 0.5). Prior to assembly, reads shorter than 60 bp were excluded, and remaining reads were screened for potential vector contamination by BLAST searches against the UniVec database (NCBI). Any remaining singletons (i.e. reads that were not incorporated into any contig) were retained in the data set, because they are likely fragments of low-expression transcripts (Vera *et al.* 2008; Meyer *et al.* 2009; Wheat 2010). Assembled contigs and singletons were pooled for subsequent analyses, and are hereafter referred to as 'unigenes'.

Because the SC data set represents the largest set of reads from a single RNA sample (the SD sample consisted of three pooled treatments), the SC unigenes were used for the functional annotation steps. Putative gene products were determined by BLASTX searches against NCBI's nr (nonredundant) protein database via the BLAST2GO (Conesa *et al.* 2005; Götz *et al.* 2008) pipeline, retaining up to 10 hits with E-value $\leq 10^{-3}$. The predicted gene name from the highest scoring BLAST hit was then assigned to its respective SC query sequence. Gene Ontology (GO; The Gene Ontology Consortium 2000) classification terms, which describe the cellular components, molecular functions and biological processes of known genes, were retrieved from BLAST hits at a more stringent E-value threshold of 10^{-6}.

In order to evaluate the coverage of our sequencing effort and the completeness of our transcriptome library, we employed a sample rarefaction method. Our strategy involved determining how many unique genes were identified by successively larger fractions of the data. A pool of ~50 000 randomly selected reads from the SC population were assembled as above, and the set of unigenes was compiled. Next, ~50 000 reads were added to the previous pool of raw reads, a new assembly performed and unigenes saved. This process was repeated using increments of ~50 000 until all reads in the SC data set were included. The resulting sets of unigenes were then compared (BLASTX, $E \leq 10^{-3}$) separately against the complete set of peptides from *Tribolium castaneum* (Tribolium Genome Sequencing Consortium 2008), and the number of unique hits identified. The BLAST searches were repeated by using *Apis mellifera* (Honeybee Genome Sequencing Consortium 2006) and *Pediculus humanus* (Lawson *et al.* 2009) protein databases as references. These species were chosen as reference genomes for this analysis because they were the three most common taxa among top BLAST hits (see *Functional annotation* below).

Interpopulation divergence

In order to describe genome-wide levels of coding sequence evolution and to identify genes undergoing accelerated divergence, we estimated rates of nonsynonymous (d_N) and synonymous (d_S) substitutions between SD and SC orthologs. The ratio between these rates is frequently used as indicator of the intensity and mode of selection under which a protein-coding sequence is evolving, with $d_N/d_S > 1$ and $d_N/d_S < 1$ generally interpreted as signatures of, respectively, positive and purifying selections (Miyata & Yasunaga 1980; Nei & Kumar 2000). When the estimate of d_N/d_S is calculated across the entire sequence, however, a criterion of $d_N/d_S > 1$ as evidence for positive selection is extremely stringent (Swanson *et al.* 2001a); an overall d_N/d_S threshold of 0.5 has been shown to consistently identify genes subjected to adaptive evolution (Swanson *et al.* 2001b, 2004).

We employed a reciprocal best BLAST hit strategy to identify orthologous sequences between SD and SC unigenes. Using standalone executable scripts (NCBI), we created BLAST-searchable databases with each set of unigenes and then performed BLASTN searches between the two data sets. The output was parsed so that only pairs of sequences that were each other's best hit, with $E \leq 10^{-20}$, were retained as putatively homologous. To reduce the chance of including sequencing errors in the d_N/d_S estimates, we included only sequences that had mean coverage of 2 reads/bp or greater. Finally, we kept only pairs for which the SC sequence was successfully annotated with a unique metazoan accession in the BLAST analysis. While our contig assembly parameters do not account for intrapopulation polymorphisms, previous studies in *Tigriopus californicus* suggest this level of variation is likely to be very low compared to between-population divergence (Burton & Lee 1994; Burton 1998; Edmands 2001; Willett & Burton 2004;

Rawson & Burton 2006; Burton *et al.* 2007; Willett & Berkowitz 2007). We hence assume fixed allelic states within populations during the divergence analyses.

The most probable translated region for each contig was determined by first setting the sequences to the sense strand specified by the best BLASTX hit, and then obtaining the longest open reading frame (ORF) along that set sense using scripts from the EMBOSS package (Rice *et al.* 2000). A custom Perl script was used to automate the pairwise alignments in ClustalW (Larkin *et al.* 2007). Alignments were checked manually for errors, and those shorter than 150 bp were excluded from further analysis. Rates of synonymous and non-synonymous substitutions were estimated with the method of Yang & Nielsen (2000), implemented in the PAML package (version 4.4; Yang 2007). This method takes into account biases in the transition/transversion rate and in codon usage (Yang & Nielsen 2000). Overall d_N/d_S was calculated as mean of d_N divided by mean of d_S. This allowed us to include genes that had $d_S = 0$, as d_N/d_S for these regions could not be calculated. Confidence interval (95%) for the overall mean was then estimated by bootstrapping (1000 repetitions) in R 2.62 (R Development Core Team).

Enrichment analyses

We took advantage of our annotation results to assess whether certain functional categories tended to be found among divergent or conserved genes. Genes with $d_N/d_S \geq 0.5$ were grouped as putatively fast-evolving, while those with $d_N/d_S \leq 0.1$ were considered conserved. A Fisher's exact test was then used to test for the overabundance of GO terms between the two sets of sequences, with significance also assessed by means of the False Discovery Rate method (FDR; Benjamini & Hochberg 1995) to control for multiple comparisons. The analysis was carried out with the GOSSIP package (Blüthgen *et al.* 2005) implemented in BLAST2GO (Conesa *et al.* 2005).

Given the strong evidence for nuclear-mitochondrial incompatibilities in *T. californicus* (reviewed in Burton *et al.* 2006), we sought to assess whether nuclear-encoded mitochondrial proteins were over-represented among the fast-evolving genes. We kept the entire set of highly divergent sequences ($d_N/d_S \geq 0.5$) as above, but we randomly picked an equal number of sequences from the conserved set. Because many of our assembled sequences are likely to not represent full-length transcripts, predictions of subcellular localization were performed with methods that do not rely solely on N-terminal targeting sequences. Initial screening of putative mitochondrial-targeting proteins (mTPs) was accomplished with MITOPRED (Guda *et al.* 2004). To prevent false-positives, only predictions confirmed by another method (CELLO; Yu *et al.* 2006) were retained as mTPs. A Fisher's exact test was then used to compare the frequencies of mTPs between fast-evolving and conserved genes.

Results

Sequencing and assembly

We obtained a combined total of 1 196 464 sequence reads, ranging in length from 21 to 823 bp (mean = 360 bp). Sequences were deposited in the NCBI Sequence Read Archive (Accession number SRA024709.1). After excluding very short (length < 60 bp) and poor quality reads, the number and length of sequences available for assembly were very similar between the SD and SC samples (Table 1). Assembly of transcripts was efficient, with over 96% of reads within each sample successfully incorporated into 22 262 and 23 580 contigs for SD and SC, respectively. The corresponding number of remaining singletons (4% of total) was relatively low in comparison with other nonmodel 454 sequencing studies (Vera *et al.* 2008; Meyer *et al.* 2009; Roeding *et al.* 2009; Schwarz *et al.* 2009). While mean contig length was similar between the two population samples, the longest SD contig was over 1200 bp longer than that of SC. This could likely be explained by overabundance of certain transcripts in the SD sequence pool brought forth by the hyperosmotic stress treatments. Mean depth of coverage in both assemblies was high and consistent with that of other studies with comparable 454 sequencing effort in nonmodel organisms (Vera *et al.* 2008; Roeding *et al.* 2009; Künstner *et al.* 2010; Renaut *et al.* 2010). Coverage was also extremely variable in both samples (SD: 1.1–9764 reads/bp; SC:

Table 1 Summary of sequencing output and *de novo* assembly of 454 pyrosequencing reads in two populations of *Tigriopus californicus*

	San Diego	Santa Cruz
Total number of reads (after initial filtering)	579 995	576 965
Mean read length (bp)	362	384
Assembly results		
Number of reads assembled	559 784	554 808
Number of contigs	22 262	23 580
Mean contig length (bp)	925	872
Maximum contig length (bp)	8807	6599
Mean coverage (reads/bp)	8.26	8.94
Number of singletons	20 211	22 157
Total number of unigenes	42 473	45 737

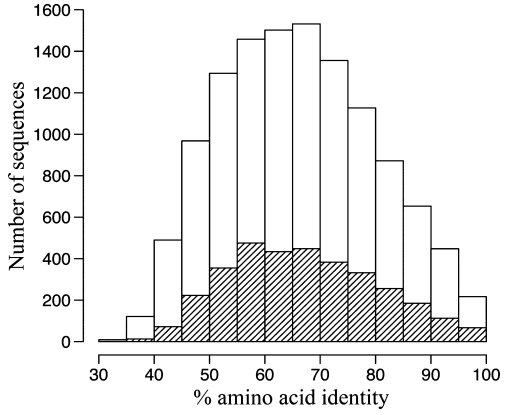

Fig. 1 Distribution of alignment similarities between *Tigriopus californicus* unigenes and their best BLASTX hits. Contigs (white bars) and singletons (shaded bars) were compared to NCBI's nr protein database, and only unique metazoan hits were retained (N = 15 402).

1.1–5623 reads/bp), and this was likely due to the non-normalization of cDNA libraries.

We retained singleton sequences in our data set on the basis that many are likely to be fragments of true transcripts of genes expressed at low levels, and hence may represent unique sequences (Wheat 2010). This prediction was supported by BLASTX searches against the nonredundant protein database (NCBI). A total of 5955 singleton sequences had significant matches to eukaryote proteins, and over half of these (n = 3355) matched unique metazoan accessions not identified by contigs (details in *Functional annotation* below). Moreover, the range and distribution of best-hit amino acid identity of singleton queries were very similar to those of higher coverage contigs (Fig. 1).

Rarefaction analysis suggests our 454 sequencing effort has reached a phase of diminishing returns. Although the overall number of unigenes may still increase with additional sequencing runs, the proportion of singletons has continually decreased (Fig. 2a). This suggests that additional reads are more likely to become incorporated into contigs than to remain as singletons. Moreover, even as the total number of unigenes continues to increase, fewer novel proteins are identified through BLASTX searches of public databases (Fig. 2b). The same pattern was observed with all three reference genomes, albeit with different number of genes identified. Therefore, many unigenes are likely nonoverlapping fragments of the same transcripts; increasing our sequencing effort hence would likely improve transcript length coverage but not uncover new transcripts. Owing to the evolutionary divergence between *Tigriopus californicus* and the reference species, we do not expect these analyses to estimate the true number of genes in *T. californicus*.

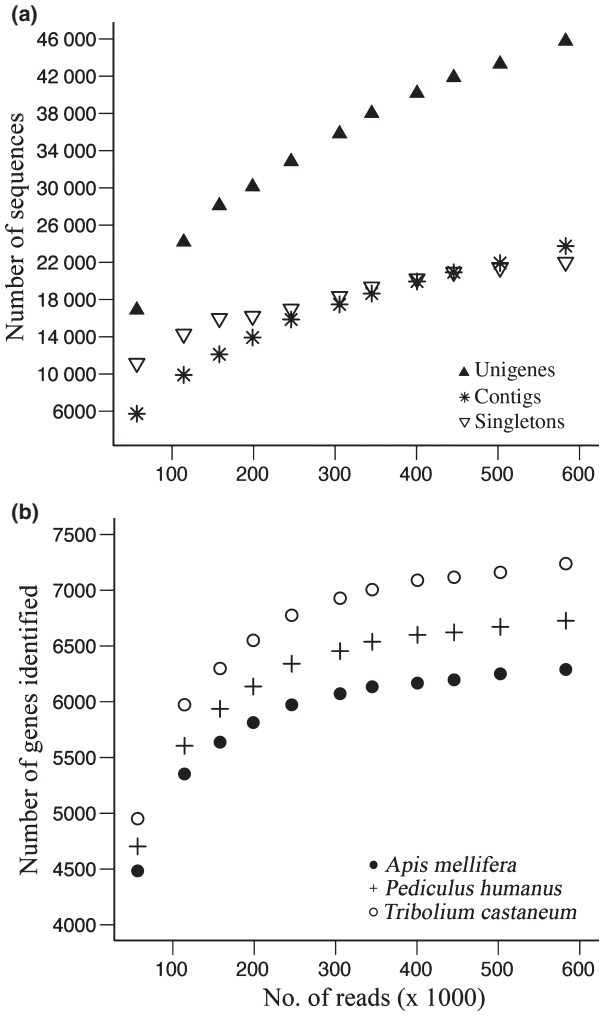

Fig. 2 Assessment of 454 assembly coverage in *Tigriopus californicus*. (a) Rarefaction of assemblies, assessed by assembling successively larger fractions of 454 reads; (b) Rarefaction of annotated gene discovery according to sequencing effort. Assemblies generated in (a) were used in BLASTX searches of three reference arthropod genomes.

Functional annotation

A total of 45 737 unigenes (contigs plus singletons) from the SC assembly were queried against the nr protein database from NCBI, with 45% (14 736 contigs + 5955 singletons = 20 691) significantly matching ($E \leq 10^{-3}$) eukaryotic proteins. We further filtered the BLAST output for only sequences with best hits to metazoan accessions, and retained 18 545 unigenes. Because assembled sequences may represent nonoverlapping fragments of the full transcript, different unigenes often matched the same accession. We reduced this redundancy by keeping only the highest scoring query-subject pair among those with identical top BLAST hit. Among the resulting 15 402 unigenes, over 50% are present in

Table 2 Distribution of 12 most common arthropod taxa among best BLAST hits to *Tigriopus californicus* transcriptome sequences. Similarity searches of NCBI's nr database were performed with the BLASTX method using an E-value threshold of 10^{-3}

Species	Taxonomic group	No. of hits (% total*)	Mean % identity (SD)	Median score
Tribolium castaneum	Insecta	1572 (10.2)	68.6 (13.2)	165.43
Pediculus humanus	Insecta	1238 (8.0)	67.7 (12.8)	145.01
Drosophila†	Insecta	1215 (7.9)	65.1 (13.9)	107.07
Apis mellifera	Insecta	993 (6.4)	68.3 (12.9)	164.45
Nasonia vitripennis	Insecta	943 (6.1)	67.9 (12.8)	159.46
Lepeophtheirus salmonis	Copepoda	847 (5.5)	75.9 (13.1)	218
Caligus‡	Copepoda	658 (4.3)	74.4 (13)	215.89
Acyrthosiphon pisum	Insecta	564 (3.7)	66.8 (13.5)	120.36
Ixodes scapularis	Arachnida	540 (3.5)	66.8 (13.2)	114
Aedes aegypti	Insecta	456 (3.0)	66.3 (13.3)	136.73
Anopheles gambiae	Insecta	394 (2.6)	67 (14)	128.64
Culex quinquefasciatus	Insecta	381 (2.5)	65.2 (13.9)	113.23
	Total	9801 (63.6)		

*Percentage based on the total number of unique metazoan hits (N = 15 402).
†Hits to the 12 *Drosophila* species with sequenced genomes were pooled.
‡Hits to *Caligus clemensi* and *C. rogercresseyi* were pooled.

10 arthropod taxa with completed genome sequences (Table 2). The beetle *Tribolium castaneum* was the most common best-hit taxon, followed by *Pediculus humanus* and by a pool of 12 *Drosophila* species. Three parasitic copepods, with large expressed sequence tag (EST) databases, were also among the most common metazoan hits, having the highest BLAST alignment similarities among them (Table 2; Fig. S1, Supporting information).

Gene Ontology annotations were performed at a more stringent E-value, yet most (n = 12 670) of the unigenes with distinct metazoan best hits were assigned at least one GO term. Within the 'biological process' designation, 16 level-2 (i.e. most inclusive) categories were assigned, with the majority of sequences implicated in metabolic and cellular processes (Fig. S2, Supporting information). The most common molecular functions assigned to these sequences were binding and catalytic activity, with an additional 8 level-2 functional categories being represented (Fig. S2, Supporting information). These unigenes were predicted to function across 10 different cellular components, being commonly located on cell parts or in organelles (Fig. S2, Supporting information). Finally, a total of 2475 Enzyme Commission (EC) codes were assigned to annotated unigenes.

Interpopulation divergence

Reciprocal BLAST searches identified 19 622 pairs of putatively orthologous unigenes between SD and SC samples. Local alignment lengths ranged from 50 to 6535 bp, with a median of 503 bp. Comparison of nucleotide similarity along the aligned regions revealed a wide range of genetic divergences (Fig. 3). While median difference was moderate (2.71%), over 13% of orthologous pairs showed nucleotide divergence ≥5%, with a high of 20.3%. Nearly 55% of orthologous sequences were annotated through matches to unique metazoan hits, but fewer (n = 5897) had sufficient depth of coverage for estimation of d_N and d_S.

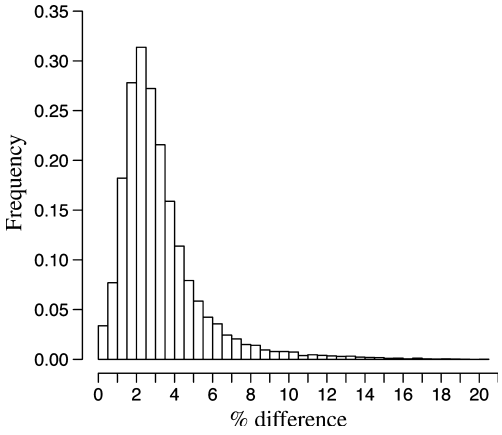

Fig. 3 Nucleotide sequence divergence between SD and SC orthologous unigenes. Alignments were based on reciprocal BLASTN searches.

Based on alignments of 5897 pairs of predicted ORFs, we estimated that the majority of coding regions (54%) had a constrained rate of nonsynonymous substitution ($d_N/d_S \leq 0.1$). Mean d_N and d_S across all genes were, respectively, 0.0075 and 0.062, for a transcriptome-wide ratio ($\overline{d_N}/\overline{d_S}$) of 0.120 (95% confidence interval: 0.116–0.124). While only seven orthologous contigs exhibited d_N/d_S slightly greater than 1 (maximum of 1.63), an additional 222 were above the less conservative thresh-

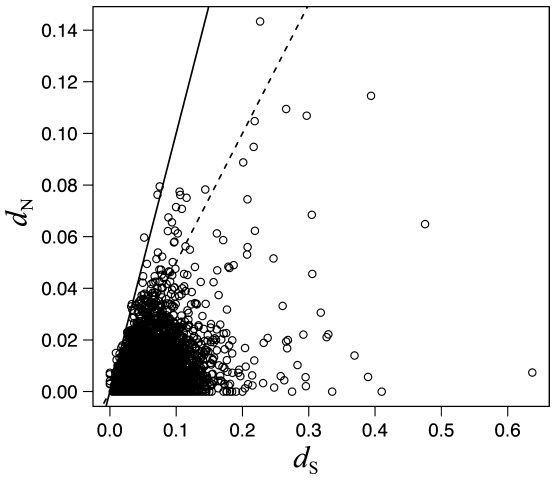

Fig. 4 Relationship between the number of nonsynonymous substitutions per nonsynonymous site (d_N) and the number of synonymous substitutions per synonymous site (d_S) for 5897 protein-coding regions of *Tigriopus californicus*. Analysis was performed on alignments of San Diego and Santa Cruz orthologous sequences, using the method of Yang & Nielsen (2000). The solid line shows the threshold of $d_N/d_S = 1$, while the dashed line marks the less conservative threshold of $d_N/d_S = 0.5$.

old of 0.5 (Fig. 4). Among these 229 gene contigs with high d_N/d_S, 124 were annotated with GO terms. Fifty-eight level 3 biological process and molecular function categories are represented in this subset of genes, with biosynthetic and metabolic processes, as well as hydrolase activity and binding functions being the most common (Table S1, Supporting information). Terms with more specificity (i.e. lower level, child terms) regarding function and process were successfully assigned to most genes, with some genes reaching GO level 11 (Table S1, Supporting information). Comparison of GO terms between divergent and conserved sets of genes revealed several GO categories that were over-represented. Four functional categories were found to be in significant excess among genes showing evidence of positive selection, while as many as 37 others were significantly over-represented among the conserved genes (Table 3). When correction for multiple testing was applied, all of the terms in the latter set retained statistical significance (at FDR = 0.05). Finally, 11 genes of the 229 were predicted to function in the mitochondria, but this proportion was not significantly over-represented relative to the set of conserved genes (Fisher's exact test, $P = 0.41$).

Discussion

Intrinsic postzygotic isolation is thought to occur primarily through the accumulation of genetic incompatibilities between isolated populations, causing interpopulation hybrids to have reduced fitness (Dobzhanksy 1936; Muller 1942). Ultimately, hybridization between species typically produce no offspring or F_1 offspring with significant fitness reduction (including complete sterility or inviability). Before this stage, however, incompatibilities may already be present, but these might only be exposed when recessive alleles are recombined in later hybrid generations. The intertidal copepod *Tigriopus californicus* hence provides a highly suitable system for investigating the early stages of reproductive isolation brought upon by divergence in allopatry, because interpopulation crosses are viable and fertile despite high levels of population genetic differentiation (Burton & Lee 1994; Burton *et al.* 2007). In addition, F_2 hybrid breakdown occurs at different levels of biological integration (reviewed in Burton *et al.* 2006), with a genetic system of nuclear-mitochondrial incompatibility being proposed as a key mechanism for this phenomenon (Burton *et al.* 2006; Ellison & Burton 2008a). Elucidating details of hybrid breakdown at the molecular level, however, has been hindered by the lack of genome-level information in this species.

This study had two primary goals: (i) to characterize a substantial portion of the *T. californicus* transcriptome by linking DNA sequence of coding regions to their predicted protein products and functions; and (ii) to reveal genome-wide patterns of sequence evolution in this species by comparing the identified genes between two divergent populations. As a result of recent advances in high-throughput sequencing technology, there has been a surge of studies describing the transcriptome of nonmodel organisms (Vera *et al.* 2008; Hahn *et al.* 2009; Kristiansson *et al.* 2009; Schwarz *et al.* 2009; Zagrobelny *et al.* 2009; Ferguson *et al.* 2010). Only a few studies have used these methods to provide interspecific comparisons of sequence evolution at genomic levels normally reserved for established model systems (Künstner *et al.* 2010; Wolf *et al.* 2010b; reviewed in Ellegren 2008). Our study is among a few that perform comparisons between populations of a species at the incipient stages of reproductive isolation (Elmer *et al.* 2010; Renaut *et al.* 2010). Raw reads obtained through 454 pyrosequencing were sufficiently long to permit efficient *de novo* assembly of ~20 000 contigs, with mean coverage greater than 8 reads/bp, for each population's cDNA sample. Most of these, along with nearly 6000 unassembled singleton reads, exhibited significant amino acid sequence homology to known eukaryotic genes. This represents a significant increase in the number and diversity of coding regions currently available for this and other nonmodel crustaceans.

As we included whole individuals across different developmental stages for cDNA library preparations, the pool of sequences we obtained provides a global view of the transcriptome in *T. californicus*. Assuming

Table 3 Gene Ontology terms significantly over-represented among divergent ($d_N/d_S \geq 0.5$) or conserved ($d_N/d_S \leq 0.1$) genes in *Tigriopus californicus*. Bold and normal fonts indicate terms over-represented in the divergent and conserved subsets, respectively

Gene ontology (GO) term	GO term ID	% of genes within group*		
		$d_N/d_S \geq 0.5$	$d_N/d_S \leq 0.1$	*P*-value†
Biological process				
Amine metabolic process	0009308	2.42	8.61	0.0056
Carboxylic acid metabolic process	0019752	3.23	9.52	0.0077
Cell cycle process	0022402	0	2.69	0.0370
Cellular amino acid metabolic process	0006520	2.42	7.62	0.0145
Cellular aromatic compound metabolic process	0006725	0	3.77	0.0097
Cellular carbohydrate metabolic process	0044262	0	3.35	0.0162
Cellular component organization	0016043	0	10.47	0.0000
Cellular ketone metabolic process	0042180	3.23	9.73	0.0063
Coenzyme metabolic process	0006732	0	2.81	0.0317
Cofactor metabolic process	0051186	0	3.15	0.0210
Cytoskeleton organization	0007010	0	2.65	0.0390
Developmental process	0032502	4.03	10.14	0.0124
Macromolecular complex assembly	0065003	0	3.27	0.0180
Macromolecular complex subunit organization	0043933	0	3.52	0.0132
Nitrogen compound metabolic process	0006807	9.68	18.54	0.0059
Nucleobase, nucleoside, nucleotide metabolic process	0055086	0	4.10	0.0064
Organelle organization	0006996	0	4.59	0.0034
Organic acid metabolic process	0006082	3.23	9.52	0.0077
Oxoacid metabolic process	0043436	3.23	9.52	0.0077
Primary metabolic process	0044238	22.58	35.31	0.0019
Protein complex assembly	0006461	0	2.52	0.0455
Regulation of cellular metabolic process	0031323	4.84	11.01	0.0151
Regulation of macromolecule metabolic process	0060255	5.65	11.34	0.0267
Regulation of primary metabolic process	0080090	4.84	11.26	0.0124
Response to chemical stimulus	0042221	0	3.06	0.0233
Ribonucleoprotein complex biogenesis	0022613	0	2.57	0.0432
Ribosome biogenesis	0042254	0	2.48	0.0479
Small molecule metabolic process	0044281	4.84	14.98	0.0004
Vasculature development	**0001944**	**1.61**	**0.21**	**0.0423**
Molecular function				
Adenyl nucleotide binding	0030554	2.42	11.38	0.0003
Adenyl ribonucleotide binding	0032559	2.42	10.47	0.0008
ATP binding	0005524	2.42	10.26	0.0011
Binding	0005488	50.00	62.04	0.0051
Cysteine-type peptidase activity	**0008234**	**3.23**	**0.70**	**0.0170**
Helicase activity	**0004386**	**4.03**	**1.49**	**0.0469**
Ligand-gated ion channel activity	**0015276**	**1.61**	**0.17**	**0.0312**
Nucleotide binding	0000166	8.06	19.08	0.0007
Purine nucleoside binding	0001883	2.42	11.38	0.0003
Purine nucleotide binding	0017076	4.03	14.49	0.0002
Structural molecule activity	0005198	0	3.68	0.0107
Transferase activity	0016740	8.06	13.66	0.0433

*Based on the number of annotated genes: high d_N/d_S – 124; low d_N/d_S – 2416.
†Uncorrected *P*-values from a Fisher's exact test.

tissue- or age-specific biases are not substantial in our library of transcripts, we can provide an initial estimate of the minimum number of protein-coding loci in this species. After retaining only unigenes that were matched to unique accessions among metazoans, we arrived at a value of 15 402 putatively distinct transcripts. This should be considered an upper-bound within our library, however, as many unigenes may represent nonoverlapping portions of the same transcript and hence may match different accessions. An alternative approach is to retain unigenes according to NCBI-annotated gene names. Among all unigenes with metazoan hits, we retained a nonredundant list 10 922 gene names (after excluding uninformative names such

as 'predicted protein', 'unknown', and 'hypothetical protein'). This is most likely an underestimate of the number of genes in *T. californicus*, as our sequencing effort has approached but not yet reached full saturation in gene discovery (Fig. 1b). In addition, many unigenes lack matches to public databases probably because they are either too short to permit appropriate alignments, or represent highly divergent, previously uncharacterized regions. Approximately 3000 (18%) of the 16 404 predicted genes from *Tribolium castaneum*, the most common best-hit taxon among our unigene annotations, showed no homology to other metazoan genes (Tribolium Genome Sequencing Consortium 2008). If we assume that a similar proportion of *Tigriopus* genes would find no BLAST matches in public databases, we can adjust the minimum number of genes expressed in this species to roughly 13 300 (10 922/0.82).

Annotation of transcripts according to GO terminology (The Gene Ontology Consortium 2000) provides a valuable resource for future functional studies in *Tigriopus* and closely related crustaceans. These annotations can be used, for instance, to find suites of genes associated with stress response. The intertidal habitats where *T. californicus* is found undergo daily fluctuations in several abiotic factors, notably temperature, salinity, and pH. Given its tractability for laboratory culturing and experimentation, this species provides a suitable system for investigating eco-physiological mechanisms of stress response in the marine environment. A search through the catalog of *T. californicus* GO annotations revealed several genes involved in response to stress ($n = 45$), with fewer more specific assignments to response to temperature (7) and osmotic stimuli (7).

Stress response genes in *T. californicus* may also be of interest in studies of their role in local adaptation. Despite the highly variable physical conditions in their intertidal habitats, gene flow between neighbouring populations is very limited (Burton & Lee 1994; Edmands 2001; Willett & Ladner 2009), and some evidence exists for adaptation to locally varying osmotic (Burton & Feldman 1983; Burton 1986) and temperature (Willett 2010) regimes. Besides showing 1–8% structural differentiation between SD and SC, these genes are suitable candidates for studies of local adaptation at the level of gene regulation. Investigations in a variety of taxa have shown that differences in gene expression patterns may play a role in species divergence (Wittkopp *et al.* 2008; reviewed in Wolf *et al.* 2010a), and are often associated with ecologically relevant phenotypes (Abzhanov *et al.* 2006; Shapiro *et al.* 2006). In addition to providing a library of candidate genes, the characterized transcriptome can serve as reference 'genome' on which to map genome-wide gene expression profiles gathered with new high-throughput methods (i.e., RNA-seq, Wang *et al.* 2009).

We found evidence for substantial interpopulation sequence diversity across much of the *T. californicus* transctiptome. While mean sequence divergence was moderate (2.7%), only a relatively small proportion of orthologous regions had no nucleotide changes, even while accounting for sequencing errors (Wheat 2010). This level of sequence variation allowed us to estimate the adjusted rate of nonsynonymous substitution (d_N/d_S) for over 5000 pairs of putatively orthologous sequences. This analysis identified 229 genes with elevated d_N/d_S, suggestive of adaptive evolution, as well 3222 that are potentially under strong purifying selection (i.e. with very low d_N/d_S). Nearly 40 functional categories, comprised mostly of metabolic processes and binding functions, were detected to be in significant excess among the conserved genes. The fast-evolving genes, in contrast, were not concentrated in particular gene classes; they were distributed across a wide array of processes and functions, with only four GO terms being over-represented. While accelerated differentiation in most of these functional classes of genes is likely to be specific to *T. californicus*, some may be more generally associated with the early stages of population divergence. Twenty specific GO terms, for instance, were shared between *T. californicus* and other systems of recently diverged populations or species for which a similar transcriptomic screen (i.e. based on high d_N/d_S) was reported (Table 4). The generality and biological significance of these findings merits further investigation.

Among the set of genes potentially under positive selection, we found eleven that are predicted to be mitochondrial-targeting proteins (mTPs) encoded in the nucleus (Table S2, Supporting information). Although the functional relevance of these genes in *T. californicus* divergence is currently unclear, these regions are suitable candidates for future studies in this species, given accumulated evidence that mitochondrial performance and evolution play a role in hybrid breakdown (Edmands & Burton 1999; Ellison & Burton 2006, 2008a). Our finding that this proportion of mTPs is not different from that found among conserved genes may simply suggest adaptive evolution should not be expected to occur throughout the entire mitochondrial machinery. Adaptive divergence may, for instance, be restricted to mTPs interacting directly with protein complexes of the electron transport system (ETS) (Willett & Burton 2004; Mishmar *et al.* 2006).

The ratio of the rates of nonsynonymous to synonymous substitutions (d_N/d_S) is commonly used in comparative genomic studies as a beacon for fast-evolving regions (reviewed in Ellegren 2008). When applied to

Table 4 Gene Ontology terms shared between this and other studies of recently diverged populations or species. Terms were annotated to genes with $d_N/d_S \geq 0.5$, identified through transcriptomic scans.

Gene Ontology term
Biological process
Amino acid metabolic process[1]
Biosynthetic process[1]
Cell redox homeostais[1]
Cognition[1]
Immune response[1]
Lipid metabolic process[1]
Nervous system development[1]
Proteolysis[1]
Tissue remodelling[1]
Ubiquitin-dependent protein catabolic process[1]
Carbohydrate metabolic process[2]
Cell adhesion[2]
mRNA processing[2]
Protein amino acid phosphorylation[2]
RNA processing[2]
Translation[2]
Molecular function
ATP binding[1]
Isomerase activity[1]
Nucleotide binding[1]
Serine-type peptidase activity[1]

[1]Elmer *et al.* (2010).
[2]Renaut *et al.* (2010).

data generated with novel high-throughput sequencing methods, this approach provides a powerful way of quickly identifying potential targets of positive selection. While the subset of genes we have earmarked through this approach are suitable subjects for future research, more rigorous statistical tests of positive selection, using multiple sequence samples, are required to confirm current assignments as well as to detect specific codons undergoing adaptive change (Swanson *et al.* 2004; Bielawski & Yang 2005). The *Drosophila* 12 Genomes Consortium (2007) have shown that, among *Drosophila* species, as few as 2% of codons in regions with elevated d_N/d_S are actually targets of positive selection, while an average across the entire gene exhibits $d_N/d_S << 1$ (Swanson *et al.* 2001a, 2004; Clark *et al.* 2007). Interacting amino acid regions among peptides of the ETS are prime candidates of adaptive divergence and coadaptation and might provide a mechanism of postzygotic isolation in *T. californicus* (Burton *et al.* 2006; Willett & Burton 2004).

Conclusion

In their discussion of 'speciation genes', Orr *et al.* (2004) reached three conclusions: (i) the factors that cause postzygotic reproductive isolation are often ordinary genes that have normal function within species (i.e., there is no evidence that speciation genes belong to a single functional class such as transcription factors); (ii) the responsible genes are evolving rapidly; and (iii) *most importantly* (their emphasis), the rapid evolution is driven by positive Darwinian selection (i.e., diversifying selection). Although we do not agree that positive selection is necessary for the evolution of reproductive isolation (Burton *et al.* 2006), our genome-wide scan for evidence of positive selection has revealed a set of new rapidly evolving candidate genes, across a variety of fundamental functional classes, to add to those previously identified on the basis of biochemical physiology (i.e., the nuclear and mitochondrial genes encoding components of the ETS). In addition to identifying this new set of candidate genes, the transcriptome studies initiated here permit future work focusing on both structural gene differentiation and changes in gene expression underlying hybrid breakdown in *Tigriopus californicus*.

Acknowledgements

We thank S. Schoville and two anonymous reviewers for insightful comments on earlier versions of the manuscript. We especially thank Jane Hutchinson (Roche Applied Science) for facilitating the sequencing work. This study was funded in part by a grant from the National Science Foundation (DEB 0717178) to RSB.

References

Abzhanov A, Kuo WP, Hartmann C, Grant BR, Grant PR, Tabin CJ (2006) The calmodulin pathway and evolution of elongated beak morphology in Darwin's finches. *Nature*, **442**, 563–567.

Arntzen JW, Jehle R, Bardakci F, Burke T, Wallis GP (2009) Asymmetric viability of reciprocal-cross hybrids between crested and marbled newts (*Triturus cristatus* and *T. marmoratus*). *Evolution*, **63**, 1191–1202.

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, **57**, 289–300.

Bielawski JP, Yang Z (2005) Maximum likelihood methods for detecting adaptive protein evolution. In: *Statistical Methods in Molecular Evolution* (ed. Nielsen R), pp. 103–124. Springer, New York.

Blüthgen N, Brand K, Cajavec B, Swat M, Herzel H, Beule D (2005) Biological profiling of gene 494 groups utilizing gene ontology. *Genome Inform*, **16**, 106–115.

Burton RS (1986) Evolutionary consequences of restricted gene flow in the intertidal copepod *Tigriopus californicus*. *Bulletin of Marine Science*, **39**, 526–535.

Burton RS (1990) Hybrid breakdown in developmental time in the copepod *Tigriopus californicus*. *Evolution*, **44**, 1814–1822.

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

Burton RS, Feldman MW (1983) Physiological effects of an allozyme polymorphism: Glutamate-pyruvate transaminase and response to hyperosmotic stress in the copepod *Tigriopus californicus*. *Biochemical Genetics*, **21**, 239–251.

Burton RS, Lee B-N (1994) Nuclear and mitochondrial gene genealogies and allozyme polymorphism across a major phylogeographic break in *Tigriopus californicus*. *Proceedings of the National Academy of Sciences of the United States of America*, **91**, 5197–5201.

Burton RS, Ellison CK, Harrison JS (2006) The sorry state of F_2 hybrids: consequences of rapid mitochondrial DNA evolution in allopatric populations. *American Naturalist*, **168**, S14–S24.

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

Drosophila 12 Genomes Consortium (2007) Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature*, **450**, 203–218.

Conesa A, Götz S, García-Gómez JM, Terol J, Tálon M, Robles M (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, **21**, 3674–3676.

Coyne JA, Orr HA (2004) *Speciation*. Sinauer, Sunderland, MA.

Dobzhanksy T (1936) Studies on hybrid sterility. II. Localization of sterility factors in Drosophila pseudoobscura hybrids. *Genetics*, **21**, 113–135.

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

Edmands S (2001) Phylogeography of the intertidal copepod *Tigriopus californicus* reveals substantially reduced population differentiation at northern latitudes. *Molecular Ecology*, **10**, 1743–1750.

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

Edmands S, Northrup SL, Hwang AS (2009) Maladapted gene complexes within populations of the intertidal copepod *Tigriopus californicus*? *Evolution*, **63**, 2184–2192.

Ellegren H (2008) Comparative genomics and the study of evolution by natural selection. *Molecular Ecology*, **17**, 4586–4596.

Ellison CK, Burton RS (2006) Disruption of mitochondrial function in interpopulation hybrids of *Tigriopus californicus*. *Evolution*, **60**, 1382–1391.

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

Ellison CK, Burton RS (2008b) Genotype-dependent variation of mitochondrial transcriptional profiles in interpopulation hybrids. *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 15831–15836.

Elmer KR, Fan S, Gunter HM *et al.* (2010) Rapid evolution and selection inferred from the transcriptomes of sympatric crater lake cichlid fishes. *Molecular Ecology*, **19**(Suppl. 1), 197–211.

Endler JA (1977) *Geographic Variation, Speciation, and Clines*. Princeton University Press, Princeton, NJ.

Ferguson L *et al.* (2010) Characterization of a hotspot for mimicry: assembly of a butterfly wing transcriptome to genomic sequence at the *HmYb/Sb* locus. *Molecular Ecology*, **19**(S1), 240–254.

Götz S, García-Gómez JM, Terol J *et al.* (2008) High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Research*, **36**, 3420–3435.

Guda C, Fahy E, Subramaniam S (2004) MITOPRED: a genome-scale method for prediction of nucleus-encoded mitochondrial proteins. *Bioinformatics*, **20**, 1785–1794.

Hahn DA, Ragland GJ, Shoemaker DD, Denlinger DL (2009) Gene discovery using massive parallel pyrosequencing to develop ESTs for the flesh fly *Sarcophaga crassipalpis*. *BMC Genomics*, **10**, 234.

Harrison RG (1990) Hybrid zones: Windows on evolutionary process. In: *Oxford Surveys in Evolutionary Biology* (eds Futuyma DJ, Antonovics J), pp. 69–128. Oxford University Press, New York.

Honeybee Genome Sequencing Consortium (2006) Insights into social insects from the genome of the honeybee *Apis mellifera*. *Nature*, **443**, 931–949.

Hudson ME (2008) Sequencing breakthroughs for genomic ecology and evolutionary biology. *Molecular Ecology Resources*, **8**, 3–17.

Kristiansson E, Asker N, Förlin L, Larsson DGJ (2009) Characterization of the *Zoarces viviparus* liver transcriptome using massive parallel pyrosequencing. *BMC Genomics*, **10**, 345.

Künstner A, Wolf JBW, Backström N *et al.* (2010) Comparative genomics based on massive parallel transcriptome sequencing reveals patterns of substitution and selection across 10 bird species. *Molecular Ecology*, **19**(S1), 266–276.

Larkin MA, Blackshields G, Brown NP *et al.* (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**, 2947–2948.

Lawson D *et al.* (2009) VectorBase: a data resource for invertebrate vector genomics. *Nucleic Acids Research*, **37**, D583–D587.

Marcotte EM, Xenarios I, van Der Bliek AM, Eisenberg D (2000) Localizing proteins in the cell from their phylogenetic profiles. *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 12115–12120.

Margulies M, Egholm M, Altman WE *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.

Meyer E, Aglyamova GV, Wang S *et al.* (2009) Sequencing and *de novo* analysis of a coral larval transcriptome using 454 GSFlx. *BMC Genomics*, **10**, 219.

Mishmar D, Ruiz-Pesini E, Mondragon-Palomino M, Proccacio V, Gaut B, Wallace DC (2006) Adaptive selection of mitochondrial complex I subunits during primate radiation. *Gene*, **378**, 11–18.

Miyata T, Yasunaga T (1980) Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its applications. *Journal of Molecular Evolution*, **16**, 23–36.

Moyle LC, Nakazato T (2009) Complex epistasis for Dobzhansky-Muller Hybrid incompatibility in *Solanum*. *Genetics*, **181**, 347–351.

Muller HJ (1942) Isolating mechanisms, evolution, and temperature. *Biological Symposia*, **6**, 71–125.

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

Orr HA, Masly JP, Presgraves DC (2004) Speciation genes. *Current Opinion in Genetics & Development*, **14**, 675–679.

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

Rawson PD, Burton RS (2006) Molecular evolution at the cytochrome oxidase subunit 2 gene among divergent populations of the intertidal copepod, *Tigriopus californicus*. *Journal of Molecular Evolution*, **62**, 753–764.

Rawson PD, Brazeau DA, Burton RS (2000) Isolation and characterization of cytochrome *c* from the marine copepod, *Tigriopus californicus*. *Gene*, **248**, 15–22.

Reichert AS, Neupert W (2004) Mitochondriomics or what makes use breathe. *Trends in Genetics*, **20**, 555–562.

Renaut S, Nolte AW, Bernatchez L (2010) Mining transcriptome sequences towards identifying adaptive single nucleotide polymorphisms in lake whitefish species pairs (*Coregonus* spp. Salmonidae). *Molecular Ecology*, **19**(S1), 115–131.

Rice P, Longden I, Bleasby A (2000) EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics*, **16**, 276–277.

Roeding F, Borner J, Kube M, Klages S, Reinhardt R, Burmester T (2009) A 454 sequencing approach for large scale phylogenomic analysis of the common emperor scorpion (*Pandinus imperator*). *Molecular Phylogenetics and Evolution*, **53**, 826–834.

Schwarz D, Robertson HM, Feder JL *et al.* (2009) Sympatric ecological speciation meets pyrosequencing: sampling the transcriptome of the apple maggot *Rhagoletis pomonella*. *BMC Genomics*, **10**, 633.

Shapiro MD, Bell MA, Kingsley DM (2006) Parallel genetic origins of pelvic reduction in vertebrates. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 13753–13758.

Stelkens RB, Young KA, Seehausen O (2009) The accumulation of reproductive incompatibilities in African cichlid fish. *Evolution*, **64**, 617–633.

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

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

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

Taylor SW, Fahy E, Zhang B *et al.* (2003) The mitochondrial proteome of normal human heart muscle. *Nature, Biotechnology*, **21**, 281–286.

The Gene Ontology Consortium (2000) Gene ontology: tool for the unification of biology. *Nature Genetics*, **25**, 25–29.

Tribolium Genome Sequencing Consortium (2008) The genome of the model beetle and pest *Tribolium castaneum*. *Nature*, **452**, 949–955.

Vera JC, Wheat CW, Fescemyer HW *et al.* (2008) Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Molecular Ecology*, **17**, 1636–1647.

Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, **10**, 57–63.

Wheat CW (2010) Rapidly developing functional genomics in ecological model systems via 454 transcriptome sequencing. *Genetica*, **138**, 433–451.

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

Willett CS, Berkowitz JN (2007) Viability effects and not meiotic drive cause dramatic departures from Mendelian inheritance for malic enzyme in hybrids of *Tigriopus californicus* populations. *Journal of Evolutionary Biology*, **20**, 1196–1205.

Willett CS, Burton RS (2003) Characterization of the glutamate dehydrogenase gene and its regulation in a euryhaline copepod. *Comparative Biochemistry and Physiology Part B*, **135**, 639–646.

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

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

Wittkopp PJ, Haerum BK, Clark AG (2008) Regulatory changes underlying expression differences within and between *Drosophila* species. *Nature Genetics*, **40**, 346–350.

Wolf JBW, Lindell J, Backström N (2010a) Speciation genetics: current status and evolving approaches. *Philosophical Transactions of the Royal Society B*, **365**, 1717–1733.

Wolf JBW, Bayer T, Haubold B, Schilhabel M, Rosenstiel P, Tautz D (2010b) Nucleotide divergence vs. gene expression differentiation: comparative transcriptome sequencing in natural isolates from the carrion crow and its hybrid zone with the hooded crow. *Molecular Ecology*, **19**(S1), 162–175.

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

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

Yu C-S, Chen Y-C, Lu C-H, Hwang J-K (2006) Prediction of protein subcellular localization. *Proteins: Structure, Function, and Bioinformatics*, **64**, 643–651.

Zagrobelny M, Scheibye-Alsing K, Jensen NB, Møller BL, Gorodkin J, Bak S (2009) 454 pyrosequencing based transcriptome analysis of *Zygaena filipendulae* with focus on genes involved in biosynthesis of cyanogenic glucosides. *BMC Genomics*, **10**, 574.

F.S.B. is a postdoctoral scholar in the Burton laboratory, and his research interests include speciation, phylogeography, and reproductive behaviour in marine animals. G.W.M. is a

research associate interested in using molecular methods to address questions about the evolution of copepods.

Supporting information

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

Table S1 Gene Ontology term hierarchy for 124 Tigriopus californicus contigs with $d_N/d_S \geq 0.5$.

Table S2 Mitochondrial-targeting proteins (mTPs) showing elevated divergence ($d_N/d_S \geq 0.5$) between San Diego and Santa Cruz populations of *Tigriopus californicus*.

Fig. S1 Distribution of amino acid similarity between *Tigriopus californicus* and 12 arthropod taxa.

Fig. S2 Distributions of Gene Ontology assignments for 12 670 *Tigriopus californicus* transcripts.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.