DNA Research, 2018, 25(1), 11–23
doi: 10.1093/dnares/dsx034
Advance Access Publication Date: 6 September 2017
Full Paper

Full Paper

Comparative transcriptomics of cyprinid minnows and carp in a common wild setting: a resource for ecological genomics in freshwater communities

Trevor J. Krabbenhoft[*,†] **and Thomas F. Turner**

Department of Biology and Museum of Southwestern Biology, University of New Mexico, Albuquerque, NM 87131, USA

*To whom correspondence should be addressed. Tel. 716-645-2363. Fax. 716-645-2975. Email: krabbent.j@gmail.com

[†]Present address: Department of Biological Sciences and Research and Education in eNergy, Environment and Water Program (RENEW), University at Buffalo, Buffalo, NY 14260-1300, USA

Edited by Prof. Masahira Hattori

Received 18 January 2017; Editorial decision 10 August 2017; Accepted 12 August 2017

Abstract

Comparative transcriptomics can now be conducted on organisms in natural settings, which has greatly enhanced understanding of genome–environment interactions. Here, we demonstrate the utility and potential pitfalls of comparative transcriptomics of wild organisms, with an example from three cyprinid fish species (Teleostei:Cypriniformes). We present extensively filtered and annotated transcriptome assemblies that provide a valuable resource for studies of genome evolution (e.g. polyploidy), ecological and morphological diversification, speciation, and shared and unique responses to environmental variation in cyprinid fishes. Our results and analyses address the following points: (i) 'essential developmental genes' are shown to be ubiquitously expressed in a diverse suite of tissues across later ontogenetic stages (i.e. juveniles and adults), making these genes are useful for assessing the quality of transcriptome assemblies, (ii) the influence of microbiomes and other exogenous DNA, (iii) potentially novel, species-specific genes, and (iv) genomic rearrangements (e.g. whole genome duplication). The data we present provide a resource for future comparative work in cypriniform fishes and other taxa across a variety of sub-disciplines, including stress response, morphological diversification, community ecology, ecotoxicology, and climate change.

Key words: RNA-seq, essential genes, *Cyprinus carpio*, carp, gene silencing

1. Introduction

High-throughput sequencing has dramatically accelerated the pace of genomic research.[1,2] While once restricted to model species in laboratory settings, genomic methods are being widely applied to non-model species in nature,[3–8] rapidly illuminating the black box of the genome and giving rise to the field of ecological genomics.[9–11] Reduced sequencing costs have made it feasible to study transcriptomes of co-occurring species in a community ecology context (i.e. 'community transcriptomics'),[12] as well as comparative studies of transcriptome evolution across diverse clades (i.e. 'comparative transcriptomics').[13–16] While genomic

© The Author 2017. Published by Oxford University Press on behalf of Kazusa DNA Research Institute.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

data from model species can be informative for the biology of related organisms, not all species are the same in terms of their ecology, genetics, and morphology. For example, research on the zebrafish (*Danio rerio*, family Cyprinidae) can be relevant for closely related species, but cannot explain the tremendous ecological and morphological diversity in this clade, as studies of single species are insufficient for understanding dynamic interactions among species and their respective genomes in a macroevolutionary context. In order to understand the causes and consequences of those interactions, as well as the origin of ecological novelty (e.g. new genes[17]), we must examine genomes across species that reflect that diversity. Recent studies across a diverse suite of teleost fish lineages have focused on functionally-important genetic variation in non-model species in nature.[18–22]

Transcriptomics of species in the wild has enormous potential to advance our understanding of mechanisms underlying molecular adaptation, evolutionary diversification, ecotoxicology, and community ecology.[4,23–26] In this context, several important questions arise. For example, What are the proximate and ultimate mechanisms underlying phylogenetic, ecological, and morphological divergence? How have ancestral genomes been molded by divergent natural selection and other evolutionary forces into myriad forms that exist today? How does genomic architecture constrain or promote diversification? How important are genome duplication events in adaptive radiations? What role do genomes play in underlying the ecological dynamics of community assembly (e.g. competition, abundance, spatial and temporal dynamics, physiological constraint, etc.)? A necessary first step in addressing these questions is the generation of databases reflecting the genomic or transcriptomic variation among species, which we provide in the current study.

Here we present transcriptomic resources for three members of the freshwater fish family Cyprinidae (Teleostei: Cypriniformes), one of the most speciose vertebrate clades, with over 2,000 species.[27] In addition to remarkable species diversity, the clade includes extensive ecological, genetic, and morphological diversity. Cyprinid fishes (minnows and carp) comprise an important component of freshwater fish communities throughout North America, Asia, Europe and Africa.[28] They are often the dominant fish taxa in numerical abundance and biomass and play an important functional role in aquatic ecosystems.[29–31]

The ecological and taxonomic diversity of cyprinids is particularly interesting in light of the history of genome evolution in this clade. Cyprinids were part of the radiation that occurred after the teleost-specific genome duplication event, known as the '3R hypothesis',[32,33] that preceded and perhaps facilitated the diversification of teleost fishes.[34] In addition, several cyprinid lineages have independently undergone additional rounds of genome duplications.[35–38] For example, the common carp (*Cyprinus carpio*) lineage had a fourth round of genome duplication approximately 5.6–11.3[39] or 8.2[40] million years ago (Ma), which we refer to as 'Cc4R'. Pairs of genes arising from whole genome duplication, referred to as 'Ohnologs', were theoretically present for all genes immediately following the Cc4R genome duplication. Varying levels of subsequent gene-silencing and 're-diploidization' have since occurred in polyploid lineages making cyprinids ideal for comparative studies of genome evolution.[41,42]

Despite the ecological importance of cyprinids in freshwater systems worldwide and dynamic lineage-specific patterns of genomic expansions and contractions, most species have little or no genomic resources available for investigating their molecular ecology or genome evolution. Three notable exceptions are zebrafish (*Danio rerio*), fathead minnow (*Pimephales promelas*), and common carp (*Cyprinus carpio*). The family includes zebrafish (*Danio rerio*),[43–45] a model species with a comprehensively-annotated genome.[46] Zebrafish is an important model organism in developmental biology and disease research,[43–45] due to its semitransparent embryos and ease of laboratory culture, as well as its comprehensively annotated genome.[46] Fathead minnow is widely used as an indicator species in ecotoxicology studies for which microarrays have been developed[47,48] and a draft genome sequence is now available.[49] Common carp is an important food fish, especially in Asia, and is produced extensively in aquaculture.[50] The carp transcriptome has been studied elsewhere[39,51,52] and recently a draft genome sequence was published.[40]

We used the extensive zebrafish genomic resources available to annotate transcriptomes of three evolutionarily related, but non-model species that co-occur in parts of the west-central United States: *Cyprinella lutrensis* (red shiner), *Platygobio gracilis* (flathead chub) and *Cyprinus carpio* (common carp). These species were selected to reflect phylogenetic breadth, but also because their distributions overlap and occupy identical dryland river habitats (i.e. the Rio Grande, New Mexico), where they are exposed to similar biotic and abiotic conditions. *Cyprinus carpio* is native to Asia and Europe, but was introduced into North America, perhaps as early as 1831,[53] and enthusiastically stocked throughout the US thereafter as a food fish, including New Mexico as early as 1889.[54] *Cyprinella lutrensis* and *Platygobio gracilis* are both native to central and western North America, from the Mississippi River basin to the Rio Grande in New Mexico. Both *C. lutrensis* and *C. carpio* are highly tolerant of a wide range of environmental conditions and are highly invasive in areas outside of their natural range,[55,56] whereas *Platygobio gracilis* is sensitive to environmental disturbance and imperiled or declining in several parts of its range.[57]

Transcriptomes of *Cyprinella lutrensis* and *Platygobio gracilis* have not been published to our knowledge, whereas genomic and transcriptomic data are available for *Cyprinus carpio*.[39,40,51] *Cyprinella lutrensis* and *Platygobio gracilis* are diploid ($2n = 50$), while *Cyprinus carpio* is allotetraploid $2n = 100$,[37] with some duplicated genes silenced after a lineage-specific whole genome duplication (i.e. Cc4R). Our aims in this study were to: (i) succinctly summarize and compare genes and functional annotation information obtained from various databases; (ii) test whether *Cyprinus carpio* expresses additional copies of particular genes compared to the two diploid species (*Cyprinella lutrensis* and *Platygobio gracilis*); (iii) identify potentially novel genes present in the three cyprinids that may underlie their unique ecological and morphological novelty and (iv) to assess evolutionary conservation of essential genes for development.

In zebrafish, 307 genes are known to be essential for development. Knockout mutations in these genes are embryonic lethal according to experiments by Amsterdam *et al.*[58] with subsequent revisions by Chen *et al.*[59] and updates to the ENSEMBL database.[60] These genes are highly conserved across extremely deep phylogenetic splits (e.g. yeast, fly, zebrafish, and human) due to their essential roles in development.[58] Despite their importance, essential genes have not been studied in the context of comparative molecular ecology or ecological genomics of co-occurring species. Using transcriptome data presented in this study, we assessed the evolutionary conservation of the 307 zebrafish essential genes across four cyprinid lineages. We predicted that these genes would be highly conserved across all species, consistent with their critical functional roles, as compared to non-essential genes.[61] If this is the case, then differences among species should be found in non-essential genes, such as lineage- or species-specific genes. We also tested whether both copies of duplicated genes in *C. carpio* (i.e. Cc4R Ohnologs) were retained and expressed in duplicate or whether one copy was evolutionarily lost.[38,42] One mechanism for the loss of Ohnologs is 'pseudogenization',[62] wherein a gene accumulates one or more internal stop codons that prevent formation of a functional protein product and thus becomes a pseudogene.[63] If having redundant copies of essential genes were important for survival (e.g. due to loss-of-function mutations in one copy), then evolutionary retention of duplicates would likely be favored in *C. carpio*. Conversely, if regulation of proper gene expression levels were important in the context of functional pathways, then duplicated essential genes would likely be silenced at roughly the same rate as non-essential genes (although regulatory changes could also fine-tune expression patterns). We tested these hypotheses using expression data for the three cyprinid transcriptomes as compared to zebrafish. These sequences will provide resources for more detailed studies of the evolution and functional constraint of these critical genes, particularly in the context of genome expansions and reductions.

2. Materials and methods

Fish ($n = 3$ per species) were collected with a seine on 6 July 2012 from a field site on the Rio Grande, approximately 40 km south of Socorro, New Mexico (33.690556°N, 106.993042°W). Whole fish samples (juveniles or non-spawning adults) were immediately frozen in liquid nitrogen and transported to the laboratory. Skin, gill, gut, and kidney tissues were dissected and removed from frozen fish (outer layers were only slightly thawed by the time dissection was completed; <5 minutes total time), placed in TRIzol (Invitrogen), and mechanically homogenized. Total RNA was isolated using Purelink RNA Mini kits (Ambion) following manufacturer's protocol, along with DNase treatment to reduce genomic DNA contamination. Purified total RNA was sent to the National Center for Genome Resources (Santa Fe, New Mexico, USA) for quantification, quality assessment, cDNA library preparation and sequencing. RNA integrity and purity was assessed with a Bioanalyzer 2100 instrument (Agilent Technologies). Thirty-six Illumina libraries were constructed (3 species × 4 tissues × 3 biological replicates) from the total RNA samples using Illumina TruSeq DNA prep kits according to the manufacturer's protocol. Libraries were barcoded using standard six base pair Illumina oligonucleotides, and six libraries were pooled for each lane of Illumina HiSeq 2000 (V3 chemistry) for a total of six lanes of 2 × 100 bp paired-end sequencing.

2.1. Bioinformatics

We used the bioinformatics pipeline outlined in Figure 1 for analyzing transcriptomic data in three main steps: *de novo* assembly, gene annotation, and analysis of expression of duplicated genes. Adapters and barcode sequences were removed from raw reads, and reads were trimmed using Trimmomatic[64] with parameter settings as follows: leading quality = 5; trailing quality = 5; minimum trimmed read length = 36). Reads were normalized *in silico* to maximum read coverage of 50×. Clipped and trimmed reads were assembled, *de novo*, for each species separately using Trinity version 2014-04-13,[65,66] with minimum contig length set to 200 bp. Libraries were pooled within a species for *de novo* assembly, to maximize the number of genes included. Trinity assembles reads into contigs ('Trinity transcripts'), places similar transcripts in groups loosely referred to as 'genes', and groups similar 'genes' into gene clusters.

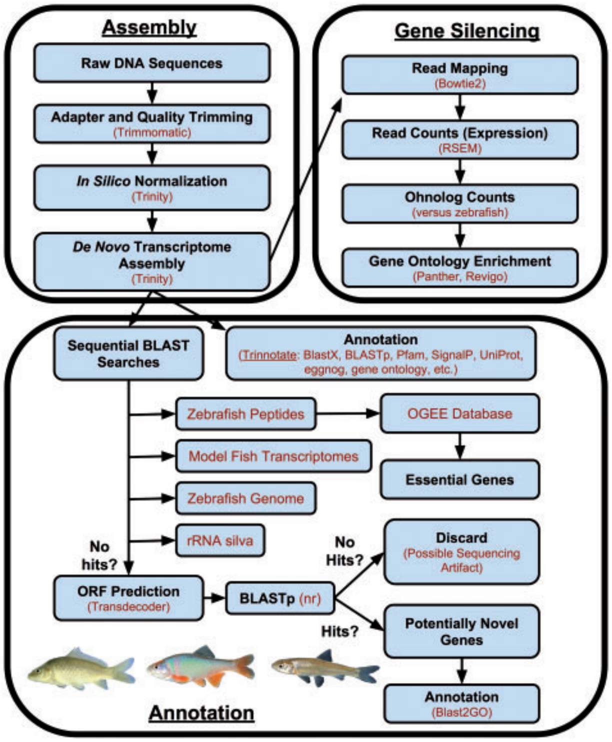

Figure 1. Flow diagram illustrating our bioinformatic pipeline. Analyses consist of three main steps: assembly, annotation, and analysis of gene silencing patterns. Databases queried and software packages used are listed.

Putative protein coding genes were also identified by BLASTx searches of contigs against zebrafish (*Danio rerio*) peptide sequences (database build Zv9) obtained from Ensembl 78.[60] Significant BLAST hits were identified based on the following parameter settings: E-value < 0.0001; gap open penalty = 11; gap extend = 1; wordsize = 3. After extensive testing, this parameter combination was found to give the optimal balance between finding matches for large numbers of contigs, while minimizing spurious hits. For most genes a 1–1 match was expected between zebrafish versus *Platygobio gracilis* or *Cyprinella lutrensis*, whereas zebrafish and *Cyprinus carpio* should have either 1-2 or 1-1 due to partial diploidy in carp. We used this expectation in determining the threshold E-value (i.e. $E < 0.0001$ in this study) to use. In practice, more stringent E-value thresholds (e.g. $E < 1e{-6}$) had very little effect on the number of significant BLAST hits.

Contigs with no significant BLAST hits against the zebrafish transcriptome were subjected to a series of stepwise BLASTn searches until significant hits were found (or not) in order to identify the possible sources of those sequences (e.g. microbiome[7]) or to identify novel genes not present in the zebrafish genome. First, remaining contigs lacking significant hits against the zebrafish transcriptome were queried against the rRNA silva database (SSU Ref 119 NR99 and LSU Parc 119), which contains bacterial and eukaryotic rRNA sequences.[67] Contigs with still no significant BLAST hits were then queried against a database containing all nine additional teleost fish transcriptomes (Amazon molly, *Poecilia formosa*; cavefish, *Astyanax mexicanus*; cod, *Gadus morhua*; fugu, *Takifugu rubripes*; medaka, *Oryzias latipes*; platyfish, *Xiphophorus maculates*; stickleback, *Gasterosteus aculeatus*; tetraodon, *Tetraodon nigroviridis*; tilapia, *Oreochromis niloticus*) from Ensembl 78. Contigs with no BLAST hits at this point were then BLASTed against the zebrafish genome (Zv9) using the 'Top Level' sequences from Ensembl to identify possible genomic DNA contamination. Remaining contigs with no significant blast hits in any of these databases were piped to TransDecoder[66] to identify open reading frames (ORFs) that represent potentially novel genes. Default parameter settings were used with TransDecoder. The software generates predicted peptide sequences for contigs with ORFs. Predicted peptide sequences for the contigs with ORFs but no BLAST hits to the aforementioned databases were queried (BLASTp; *E*-value < 0.001) against the NCBI nr database. Blast2GO version 3.0,[68] was used to identify top species hits for those predicted proteins with significant hits against nr. The remaining sequences with no hits to databases and no ORFs were discarded as likely non-protein coding, genomic DNA contamination with sufficient divergence from zebrafish to render genomic BLASTn searches ineffective.

2.2. Genome duplication, diploidization and gene silencing

Trimmed sequence reads were mapped to Trinity contigs using Bowtie2 version 2.2.2.3[69] and corresponding gene expression was quantified with RSEM version 1.2.13.[70] Because RSEM is incompatible with indel, local, and discordant alignments, parameter settings were chosen to avoid these alignments. The following RSEM parameters were used: –sensitive; –dpad 0; –gbar 99999999; –mp 1,1 –np 1 –score-min L,0,-0.1; –no-mixed; –no-discordant. Normalized expression for Trinity genes was calculated by standardizing by total mapped reads across libraries and summed across alternate Trinity transcripts (isoforms) for each locus. Networks of co-expressed genes were identified for the three species using the WGCNA package in R.[71] In order to assess the expression of duplicated genes in *Cyprinus carpio* arising from the Cc4R duplication event, we quantified the number of Trinity genes present in each species relative to zebrafish genes, as well as their expression levels. We used an arbitrary threshold of ten sequence reads per gene per tissue, summed across all three individuals, for a given gene to be considered 'expressed' in a particular tissue. Note that we are comparing whether or not a gene is expressed beyond a certain threshold, as opposed to quantifying levels of expression (i.e. RNA-seq). This approach was aimed at reducing the influence of unique reads (e.g. sequencing artifacts). Most of the contigs excluded as a result were contigs represented only by singleton reads in one library.

For *C. carpio*, we tested whether certain functional classes of genes were preferentially expressed in duplicate (i.e. the case where neither ohnolog is silenced). For this analysis, we used Panther[72] to test for statistical overrepresentation of GO-slim Biological Processes, with Bonferroni correction. The test genes consisted of the list of *C. carpio* ohnologs expressed in duplicate, while the list of all *C. carpio* genes present in the assembly was used as the reference set. GO terminology was based on the zebrafish database. Results of the overrepresentation analysis were visualized with Revigo.[73]

2.3. Essential genes

To test the hypothesis of evolutionary conservation of essential genes among cyprinid fishes, we used zebrafish genes present in the Online Gene Essentiality Database OGEE;[59] and identified orthologs in the three transcriptomes from BLASTx searches described above. Of the 307 essential genes in zebrafish,[58,59] one (ENSDARG00000038423) has been retired from ENSEMBL and one (ENSDARG00000045605) is an unprocessed pseudogene with no protein product. We searched for the remaining 305 genes in the three transcriptome assemblies to assess their conservation across cyprinids.[58]

3. Results

3.1. Sequencing and transcriptome assemblies

Six lanes of Illumina sequencing produced more than 1.2 billion paired-end reads, including 420.5-, 413.9-, and 385.3-million sequences in *Cyprinus carpio*, *Cyprinella lutrensis*, and *Platygobio gracilis*, respectively. *De novo* assembly resulted in high quality transcriptomes for all three species (Table 1). The *C. carpio* assembly had the largest number of contigs ('Trinity transcripts') and genes ('Trinity genes'), while *P. gracilis* had the fewest. In contrast, metrics for contig length (N25, N50, N75, median contig length, average contig length) were all longer in *P. gracilis* than the other two species (Table 1; Fig. 2). Overall, the *P. gracilis* transcriptome assembly was more complete despite fewer raw sequence reads. TransDecoder predicted ORFs in about half of all Trinity contigs (not shown), with the remainder comprised mainly of genomic DNA contamination that was filtered out of the final dataset. The N50 of predicted ORFs was only slightly shorter in the three species (i.e. 1,299–1,572 bp) than in zebrafish (CDS N50 = 2,037 bp), and similar to the recently published draft *C. carpio* genome 1,487 bp.[40] Removal of microbiome and genomic DNA contamination from the final assembly resulted in fewer, but longer contigs (see filtering of the final dataset, below), and an overall higher-quality assembly.

3.2. BLAST searches: zebrafish transcriptome

Top BLASTx hits of Trinity contigs against zebrafish peptides included approximately 20,000 unique genes (ENSDARG) and 11,000 protein families (ENSFAM) present in each of the three species (Fig. 3), suggesting similar annotation efficiency and transcriptome representation for each species. However, after pooling isoforms, the number of Trinity

Table 1. *De novo* transcriptome assembly results. Zebrafish (*Danio rerio*) data is included as an example of a well-assembled and complete transcriptome based primarily on Sanger sequencing

	Cyprinus carpio	*Cyprinella lutrensis*	*Platygobio gracilis*	*Danio rerio*
Trinity 'genes' (=Clusters of contigs)	309,921	255,863	180,130	30,651
Trinity 'transcripts' (=Assembled contigs)	440,696	382,504	262,969	43,153
GC content	42.45	43.25	42.67	49.60
N25 (bp)	3,327	3,069	3,644	3,465
N50	1,841	1,666	1,972	2,037
N75	704	679	788	1,179
Median contig length	418	439	450	1,080
Average contig length	907	886	978	1,501
Total assembled bases	399,790,412	339,160,955	257,217,466	64,757,328

genes that significantly matched these ~20,000 zebrafish genes varied among species: 66,447 in *Cyprinus carpio*, 60,990 in *Cyprinella lutrensis*, and 39,915 in *Platygobio gracilis* (Table 2, top row). Zebrafish genes were well covered, with more than 15,000 unique zebrafish genes covered over at least 70% of their length in corresponding contigs from each of the three cyprinids, consistent with the N50 data presented above. In general, zebrafish proteins were more completely covered by *P. gracilis* contigs than *C. carpio* or *C. lutrensis*. For example, zebrafish genes were more than 90% covered (i.e. the alignment covers >90% of bases of a gene) by sequences in 50.3% (12,489 of 24,817 genes) of *P. gracilis* genes with significant zebrafish peptide hits, versus 49.9% (13,453 of 26,963) for *C. carpio*, and 46.8% (12,538 of 26,817) in *C. lutrensis*. A large number of Trinity contigs did not significantly match (BLASTx) zebrafish peptide sequences and were subsequently queried against several additional databases.

3.3. BLAST searches: other databases

Contigs lacking significant BLASTx hits against zebrafish peptides were queried (BLASTn) iteratively against rRNA silva microbiome database, nine teleost transcriptomes, and the zebrafish genome databases (Table 2). For contigs lacking hits against zebrafish peptides, BLASTn searches versus the rRNA silva database revealed a small number of significant hits (i.e. <400 contigs; Table 2). BLASTn searches of the remaining unmatched contigs versus the nine teleost fish transcriptomes identified approximately 1,500–4,500 additional hits (Table 2), far fewer than the evolutionarily more closely related zebrafish transcriptome. BLASTn searches of the remaining unidentified contigs against the zebrafish genome revealed a large number of significant hits (>30,000 per species), suggesting these reads were the result of low levels of background genomic DNA contamination in the cDNA libraries, a common occurrence resulting from the hypersensitivity of Illumina sequencing. Conservation of sequences across deep evolutionary lineages suggests functional importance, such as regulatory regions.

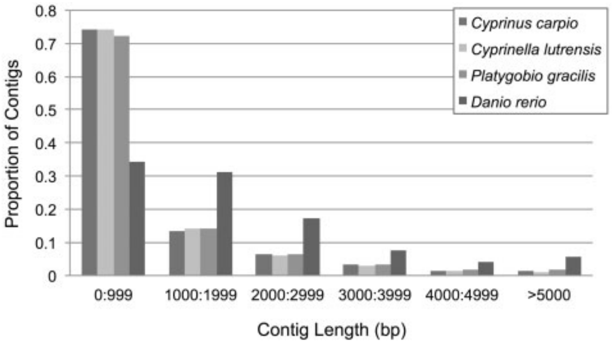

Figure 2. Contig length histogram of three cyprinids in this study and zebrafish, *Danio rerio*. By leveraging high throughput sequencing and bioinformatic filtering, we were able to generate high quality transcriptomes at a fraction of the cost and research effort used for zebrafish. As expected, *de novo* Trinity assemblies resulted in proportionally fewer contigs longer than 1000 bp, as compared to those of a well-assembled transcriptome, zebrafish (*Danio rerio*). However, note that we only used canonical transcripts for zebrafish and not the shorter isoforms, which skews the distribution toward longer transcripts for that species.

Despite extensive BLAST searches, a large number of Trinity contigs (>100,000 in each species or more than 50% of all contigs) did not have significant hits in any of the databases. These contigs are short in length (i.e. 200 bp) and have few reads mapping to them (e.g. single-read contigs). These could represent endogenous genomic DNA contamination of cDNA libraries and have sufficient evolutionary divergence from zebrafish to render BLASTn searches ineffective. A large number are also expected to be non-rRNA sequences from the microbiome, which were not present in target databases. Of contigs with no BLAST hits in the aforementioned databases, TransDecoder predicted ORFs in 8,652 (*Cyprinus carpio*), 9,215 (*Cyprinella lutrensis*), and 3,011 (*Platygobio gracilis*) contigs

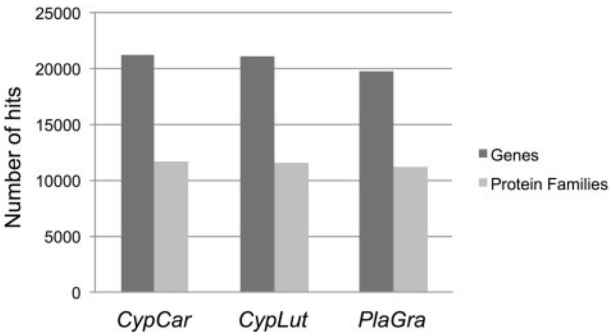

Figure 3. Unique genes and protein families from BLASTx searches (E-value threshold = 0.0001) against zebrafish (*Danio rerio*) peptide sequences.

Table 2. Significant BLAST hits for Trinity 'genes' versus various databases and number of ORFs present. BLAST searches were done in stepwise fashion: all Trinity genes were queried against zebrafish peptides but only genes without zebrafish peptide hits were queried against rRNA silva, and so on until all of the databases were queried. Summary of open reading frames (ORFs) identified in Trinity contigs with no significant BLAST hits against databases listed ('No significant BLAST hits'). Some of the ORFs lacking similar proteins in the nr database may represent novel genes or genes with divergent sequences and function, while many are likely spurious results from the sequencing and assembly process or are from unidentified microbes

	Cyprinus carpio	*Cyprinella lutrensis*	*Platygobio gracilis*
Zebrafish peptides	66,447	60,990	39,915
rRNA Silva (microbiome)	140	306	87
Teleost fish transcriptomes	4,572	2,923	1,561
Zebrafish genome	48,527	38,199	31,955
No significant BLAST hits	190,235	153,445	106,612
Total contigs	309,921	255,863	180,130
Predicted ORFs present	8,652	9,215	3,011
ORFs with nr BLASTp hits	3,789	4,154	1,548
ORFs without nr BLASTp hits (i.e. potentially novel genes)	4,863	5,061	1,463

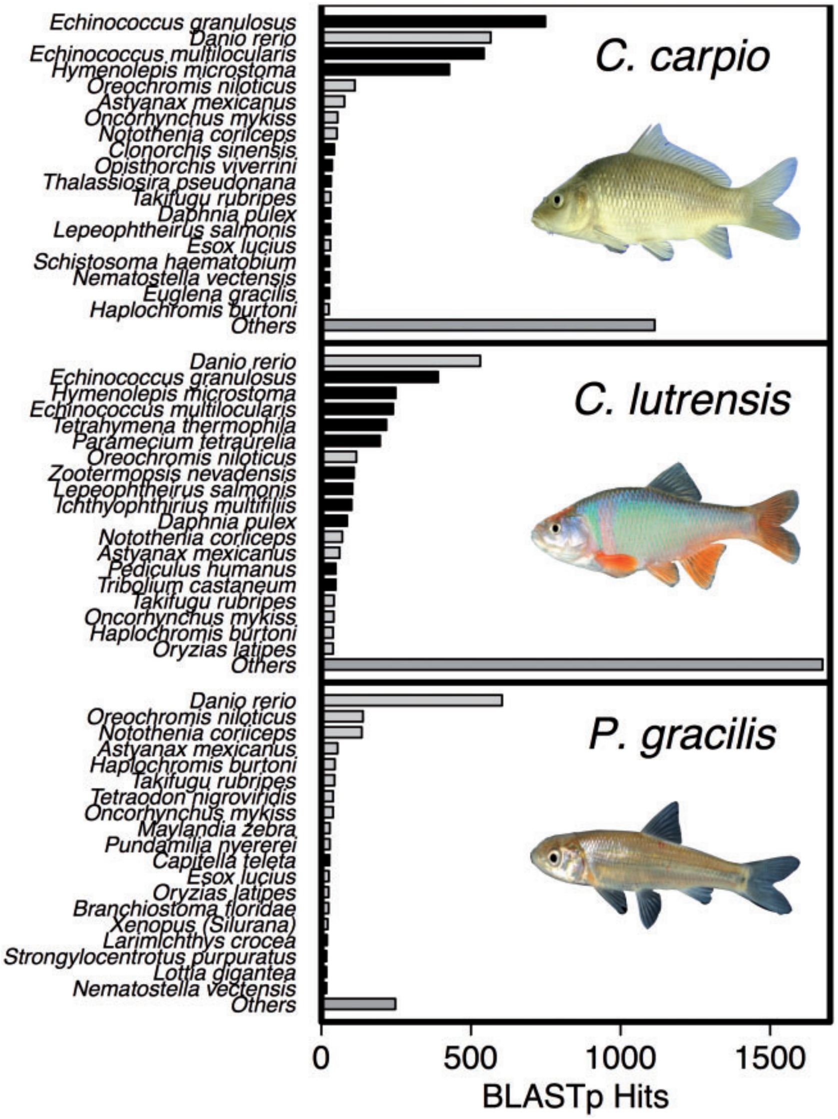

Figure 4. Top-species BLASTp hits for predicted open reading frame (ORF) peptide sequences queried against the nr database. Query sequences only included ORFs from contigs that lacked significant BLAST hits (see Table 2). Grey bars represent fish or other chordates, while black bars represent non-chordate taxa.

(Table 2). Roughly half of the predicted ORFs had significant BLASTp hits against the nr protein database (3,789, 4,154, and 1,548 contigs, respectively). Conversely, there were 4,863 (*C. carpio*), 5,061 (*C. lutrensis*), and 1,463 (*P. gracilis*) predicted ORFs had no significant hits against nr (Table 2). These ORFs could include novel genes not present in zebrafish or other teleost models, genes present in zebrafish but with significantly divergent sequences to cause BLAST searches to miss them, or could include genes from the microbiome that are not present in sequence databases.

For ORFs with nr hits, zebrafish was the top-hit species for a large portion (Fig. 4), somewhat paradoxically given the lack of significant BLAST hits against zebrafish peptide and genome sequences discussed above. This appears to be due to the fact that TransDecoder-predicted ORFs exclude 5' and 3' untranslated regions (UTRs) which diverge more rapidly than ORFs over evolutionary time. In C. *carpio* and C. *lutrensis*, many of these ORFs are from a diverse microbiome with many sequences sharing significant similarity to cyclophyllid tapeworms (e.g. *Echinococcus*, *Hymenolepis*) and protozoans (e.g. *Tetrahymena*, *Parameceum*). Conversely, in *P. gracilis* the ORFs appear to be endogenous genes with high similarity to zebrafish (Fig. 4), i.e. a less diverse microbiome is present. Contigs with predicted ORFs but no BLAST hits to any of the databases possibly represent novel or functionally divergent genes in these species that warrant further study.

3.4. Filtering and the final assembly datasets

After filtering and removal of genomic DNA and microbiome reads, the final *de novo* assembly datasets contained only Trinity contigs falling into one of the following categories: (i) contigs with significant BLAST hits against zebrafish or the nine other teleost transcriptomes; or (ii) contigs with no matches against any of the databases but with predicted ORFs present, i.e. potentially novel genes. All other contigs were removed via bioinformatic filtering. While it is possible that some of the 'microbiome' hits are actually external contamination, we expect this to be a minor component given the diverse nature of

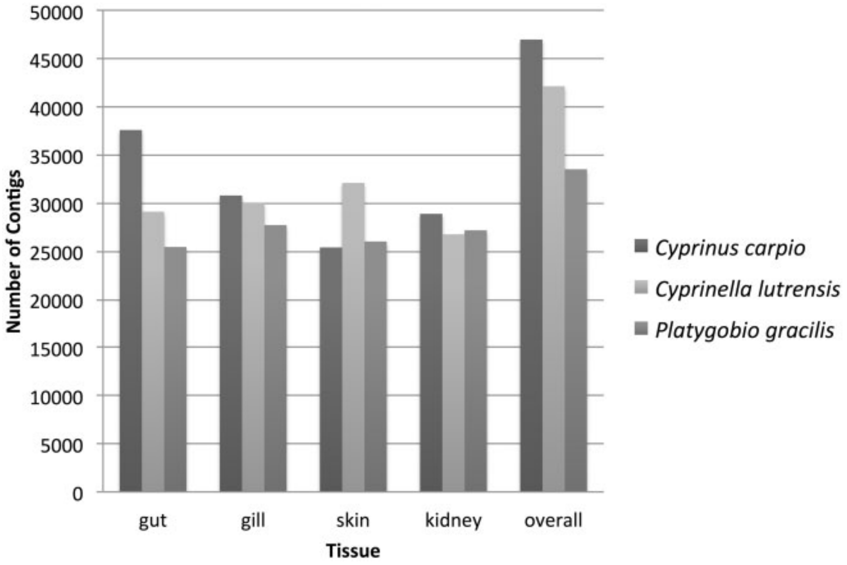

Figure 5. Number of Trinity genes (contigs) expressed in each of four tissue types, as well as all tissues pooled. Contigs only include those with significant BLASTx hits versus zebrafish peptides.

these sequences in terms of top-hit organism (Fig. 4). It is also possible that some of the genes that significantly align against zebrafish are actually microbiome or contaminant reads, though these genes being target species DNA is a more parsimonious conclusion. The final datasets are significantly smaller than the raw *de novo* assembly but present much more reliable sequence information, i.e. transcriptome sequences rather than microbiome or genomic DNA contamination.

3.5. Genome duplication, diploidization and gene silencing

Transcriptome annotation and comparison with zebrafish revealed that *Cyprinus carpio* expresses more genes than *Cyprinella lutrensis* and *Platygobio gracilis*, due to the Cc4R duplication (Fig. 5). *Cyprinus carpio* expressed about 41% more genes overall than *P. gracilis* and 11% more than *C. lutrensis*. The number of duplicate genes expressed varied dramatically among tissue types (Fig. 5). In all tissues except skin, *C. carpio* expressed more genes than the other two species (i.e. 3–48% more). In skin, both *C. lutrensis* and *P. gracilis* expressed more genes than *C. carpio* (26 and 2%, respectively). Using higher thresholds for 'expression' had moderate impact on the inferred percentage of duplicates expressed: a threshold of 100 reads instead of 10 resulted in different estimates of duplicated genes expressed in *C. carpio* versus *P. gracilis* (18% more in *C. carpio*) and *C. lutrensis* (8% more in *C. carpio*), i.e. retained expression of Cc4R duplicates. The disparity in these results could be driven in part by different assembly qualities (e.g. a better assembled *P. gracilis* transcriptome). WGCNA analysis revealed broadly similar patterns of blocks of co-expressed genes across species, consistent with the phylogenetic relatedness of the three species (Fig. 6).

Genes with retained duplicate expression (i.e. Ohnologs) in *C. carpio* represented a suite of functional groups: gene ontology terms that were significantly enriched in the 'retained duplicates' list were diverse (Fig. 7, top panel). One functional grouping that was a predominant contributor in the Revigo analysis was 'anatomical structure morphogenesis,' of interest because common carp attain much larger body size than the other two species (Fig. 7, bottom panel).

3.6. Expression of essential genes

Genes that are essential for embryonic development in *D. rerio* were nearly all present in the three cyprinids: 285 (*Platygobio gracilis*), 301 (*Cyprinella lutrensis*), and 301 (*Cyprinus carpio*) genes were expressed out of 305 zebrafish essential genes (i.e. 93.4–97.8%). Of the 20 essential genes that we did not detect in *P. gracilis*, only one was also missing in *C. lutrensis*, and two were shared with *C. carpio*. No missing essential genes were shared between *C. lutrensis* and *C. carpio*, of the four missing in each species. Essential genes missing in one or more species were generally expressed at low levels in the other species. Essential genes were nearly ubiquitously expressed across all four tissue types (skin, gill, gut, kidney), with low levels of tissue specificity (Fig. 8), in contrast to non-essential genes which generally exhibited higher levels of tissue specificity. A few essential genes do exhibit patterns of tissue specificity or species-specificity. For example, *C. carpio* expresses more essential genes in the gut than the other two species, including genes such as *wdr46* and *exosc8*, which are missing in both of the other species. Normalized levels of expression were higher in *C. carpio* than *P. gracilis* and *C. lutrensis* for 165 and 204 out of 305 genes, respectively. This pattern was not due to *C. carpio* expressing more loci per zebrafish gene (e.g. Ohnologs) than the other two species. Only slightly more loci (e.g. $n = 2$ contigs) were expressed per essential gene in the recently duplicated *C. carpio* genome (Fig. 9) whereas most duplicated essential genes in *C. carpio* are not transcribed and have either been lost evolutionarily, e.g. pseudogenes, or are expressed in other developmental stages or tissues.

4. Discussion

Next-generation transcriptome sequencing has revolutionized the field of molecular ecology over the past decade.[4,74] One outcome is increased appreciation for the molecular complexity underlying the evolution of basic ecological traits.[75,76] Here we present transcriptomic resources for comparative study of non-model cyprinid fishes in a natural 'common-garden' setting. Previous work, along with our bioinformatic analyses demonstrate that careful processing and filtering is needed to assess the sources of DNA fragments, which can be

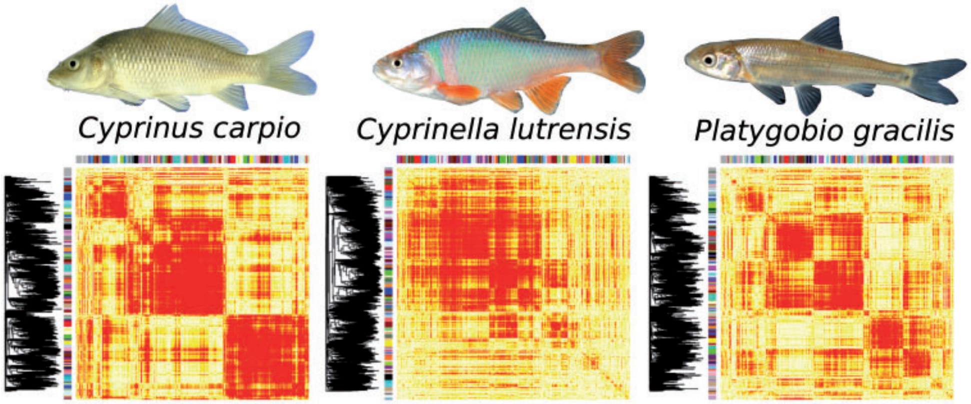

Figure 6. Results of WGCNA analysis for three cyprinid species. Dendrograms represent results of hierarchical cluster analysis of co-expression patterns of genes. Colored bars to the left and top of the heatmap show modules of co-expressed genes and pairs of genes with higher co-expression show (darker) coloration in the heatmap.

endogenous target transcriptome sequences, genomic DNA 'contamination' from the study organism, or DNA from the microbiome or diet items. Assessment of transcriptome quality also requires careful consideration.[77,78] Traditional measures of assembled read lengths such as N50 are largely meaningless for transcriptomes without additional context. We advocate combining N50 and/or histograms of contig lengths with explicit comparisons to well-studied transcriptomes of model organisms, when available. For example, we compared our *de novo* transcriptomes to zebrafish, which yielded valuable insight into progress made in our target species. Finally, positive identification of nearly all zebrafish essential genes in our transcriptomes provides additional evidence of the utility of our annotation procedures. Using the bioinformatics pipeline presented in Fig. 1, we obtained high quality transcriptome data from three species of cyprinid fishes with distinctly different evolutionary histories.

Our specific aims in this study were to sequence, annotate, and assemble the transcriptomes of co-occurring fishes with the goal of developing resources for ongoing studies of the evolution and molecular ecology of North American cyprinids. This comparative transcriptome dataset offers tools to construct assays to pose and test hypotheses related to differences in DNA sequences, functional pathways, and expression patterns among organisms that are more or less closely related (i.e. comparative approach), but that also cooccur in nature and experience similar biotic and abiotic conditions, including exposure to similar suites of pathogens and water quality conditions, for example. These data are also a resource for identifying single nucleotide polymorphisms (SNPs) in transcribed genes,[79] which could be used to explore functional or phenotypic variation within and among species.

There are several key findings in this study, including: (i) high-quality transcriptome assemblies for cyprinid fishes that reveal broad similarities and evolutionary conservation of genes with zebrafish, but with some key differences; (ii) several potentially novel genes not identified in zebrafish that are candidates for studies of ecological and morphological novelty; (iii) diverse microbiomes that vary substantially among species, despite origin from a single collection locality; (iv) ubiquitous expression of essential genes for development in later ontogenetic stages (i.e. juveniles and adults) across a broad array of tissue types; (v) a large number of duplicate genes expressed in the tetraploid, *Cyprinus carpio*, representing a diverse suite of biological processes or gene ontologies. We discuss each of these findings in greater detail below.

4.1. Assembly results

There are important considerations associated with conducting transcriptome analysis in a non-laboratory setting and in species lacking high-quality, well-annotated genomes.[4,80,81] For example, it is necessary to identify ways to maximize the quality and completeness of *de novo* assemblies.[77,80,82] Our assemblies are somewhat less complete than the zebrafish reference, but this was expected because zebrafish has been sequenced extensively at the genomic DNA level, empirically validated with RNA-seq, and refined by years of manual curation.

Trinity assemblies resulted in proportionally fewer long contigs (e.g. > 1,000 bp) compared to zebrafish. Four factors account for this result. First, the microbiome is present in these sequences and many of the contigs are not endogenous, as reflected by top species hits in BLAST searches (Fig. 4). Second, a small amount of genomic DNA contamination persists despite DNase treatment during library preparation. Genomic contamination tends to be observed as short (e.g. 200 bp), shallow contigs often comprised of single-reads. Third, the *de novo* assemblies are more fragmented due to the short read technology employed, with multiple contigs often representing non-overlapping fragments of the same gene. This effect is particularly acute in genes with short sequence repeats (e.g. microsatellites). Finally, we only used the canonical zebrafish transcripts in this study, which excludes the shorter isoforms present in many genes and biases the zebrafish distribution toward longer sequences. Transcriptomes presented here represent an improvement (i.e. more sequences, higher coverage; longer relative N50) over earlier work on sequencing and assembling the common carp transcriptome using Roche 454 sequencing,[39,51] due to the higher throughput, Illumina paired-end sequencing approach we employed. The bioinformatic approach we presented to identify and filter non-target sequences from the final dataset resulted in high quality and well annotated assemblies.

4.2. Potentially novel genes

Results of BLAST searches and ORF predictions helped us identify candidate genes that may represent novel species- or taxon-specific genes. Our interest in these genes lies in the idea that they may contain some of the functional elements responsible for extensive ecological and phylogenetic diversity present in the Cyprinidae, as in previous studies of lineage-specific gene family expansions.[83–85] For example, expansion of the *patristacin* gene family in pipefish may be an important driver in the evolution of male pregnancy in that

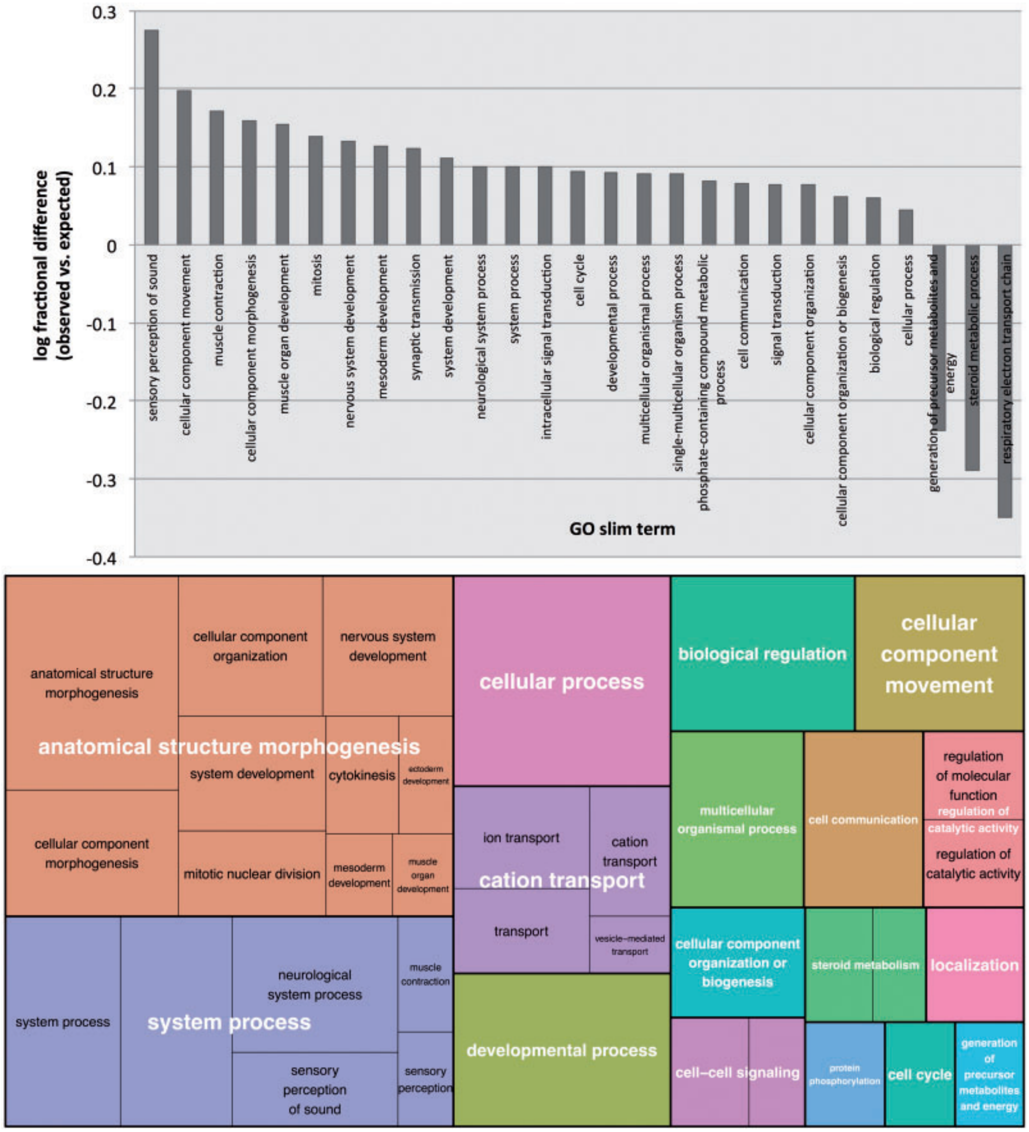

Figure 7. Top panel: Gene-ontology terms that are over- or under-represented (y-axis) in the list of genes retained as duplicates in the common carp transcriptome as compared to all expressed genes in common carp. Bottom panel: Summary of groups of biological processes overrepresented in the retained-duplicates in common carp. Box size is proportional to the number of genes with particular gene ontology terms, which may suggest a dosage effect in common carp.

lineage.[85] Many of the potentially novel genes we identified may prove to be false positives as more fish genomes are sequenced and annotated; however, these candidates would be an excellent starting point for researchers interested in targeted searches for genes or proteins underlying ecological novelty in cyprinids that may have arisen through local gene duplications, exon shuffling, horizontal transfer, or other mechanisms.

4.3. Microbiome diversity

Another valuable aspect of transcriptome sequencing of samples taken from nature is the simultaneous generation of quantifiable data on the microbiome.[86] These data are applicable to study of host-parasite dynamics, immune response, paired comparative population genetics or phylogeographic analysis of host and microbiota. When generating *de novo* transcriptome assemblies for focal species, it is imperative that microbiome sequences are identified and filtered out of final assemblies.[87] Genome-scale sequence data is often lacking for the bacterial and metazoan microbiota on vertebrate samples, which complicates attempts at removal. We used an iterative and successive filtering approach to address this issue (Fig. 1) that provides valuable information on the likely source (e.g. exogenous or endogenous) of particular sequences or contigs. Transcriptome characterization studies often do not attempt to remove exogenous microbiome and genomic DNA contamination. Researchers should be cautious when using unfiltered sequence reads, particularly when they are compiled into massive databases that lack appropriate metadata.

4.4. Conservation of essential genes

Genes that are essential for embryonic development present interesting targets for studying genome evolution due to their critical functional importance.[58,88,89] Essential genes also bear biomedical

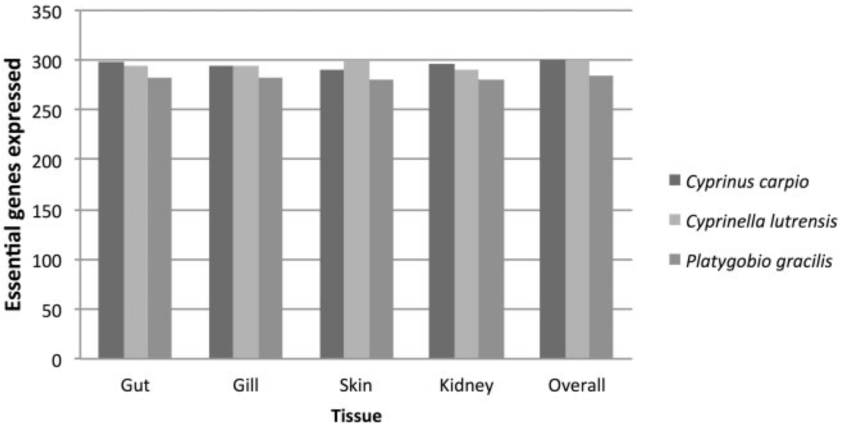

Figure 8. Expression of essential developmental genes by tissue type in three cyprinid fishes compared to 305 essential genes expressed across all tissues in zebrafish (*Danio rerio*).

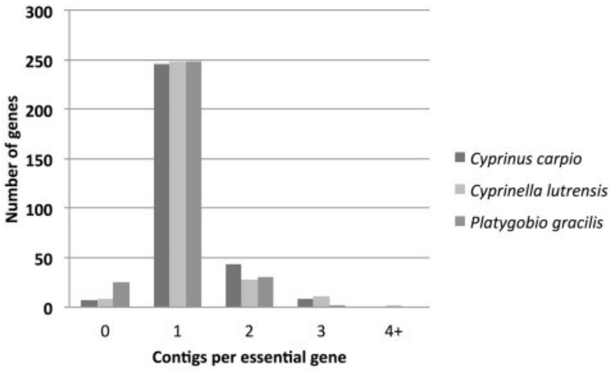

Figure 9. Number of loci (Trinity genes) expressed per zebrafish essential gene. Only slightly more (e.g. $n = 2$) were expressed per essential gene in the recently duplicated genome of *Cyprinus carpio*. Most duplicated essential genes in *C. carpio* are not transcribed and have either been lost evolutionarily, e.g. pseudogenes, or are expressed in other developmental stages or tissues.

significance as many have been implicated in human diseases and developmental abnormalities.[88] Our data demonstrate that essential 'developmental' genes previously identified in larval zebrafish[58,59] are almost all ubiquitously expressed in juvenile or adults across a broad range of tissues, suggesting their importance is not simply limited to early ontogenetic stages or particular tissues. Previous work has shown that many of these genes are critical for basic cell function, which may underlie their ubiquitous expression.[58] Based on the critical functions they perform, these genes are candidates for future studies looking at the cause of high rates of genetic inviability and mortality in cyprinids and other organisms with type-III life histories.[90] From a practical standpoint, however, we suggest that the ubiquitous expression of these genes makes their sequencing coverage and completeness useful metrics that should be used to assess the quality and completeness of *de novo* transcriptome assemblies, analogous to the use of 'housekeeping' genes as positive controls in qPCR studies.[91] The presence of nearly all essential genes across these four cyprinid species (representing more than 100 million years of evolutionary divergence[92]) is consistent with the hypothesis of broad evolutionary and functional conservation.[58,61] The few essential genes not detected may still be present in the genome, but were missed due to assembly errors or are expressed transiently at larval or juvenile developmental stages. We propose that the number of essential genes expressed could be used as a metric to complement other measures of assembly quality and completeness,[77] in addition to comparing transcript length histograms to closely related model species (see Fig. 2). While beyond the scope of the present study, future work should compare the utility of these metrics for assessing transcriptome assemblies.

4.5. Tetraploidy and expression of duplicated genes

Our results indicated that a large number of duplicate genes are expressed in *Cyprinus carpio*, representing a diverse suite of biological processes or gene ontologies, similar to previous studies.[39,40,51] For genes where both Ohnologs were expressed in *C. carpio*, there was enrichment in several different functional pathways, but many genes were associated with 'anatomical structure morphogenesis' in particular (Fig. 7). Functional duplicates at these genes correlate with large body size and rapid growth in *C. carpio* as compared to *C. lutrensis* and *P. gracilis* and a potential dosage effect. Using a different set of tissues, Wang *et al.*[39] identified enrichment of retained expression of duplicates in gene ontology pathways involved in metabolic and immune functions using 454 transcriptome sequencing and EST data mining. The availability of a (draft) genome for common carp[40] will eventually help identify Ohnologs that are silenced because pseudogenes of silenced genes may still be present in genomic DNA sequences; currently, the incomplete annotation of that genome precludes analysis of gene silencing at the genomic DNA level. Ultimately, knowledge of which genes are retained and expressed in duplicate in tetraploids as compared to related diploid species can provide insight into the role that whole genome duplication plays in the molecular ecology and phylogenetic diversification of organisms.[93,94] Note that our analyses and those of Wang *et al.*[39] are based only on expressed genes in particular tissues at a single time point, rather than genomic DNA sequences and consequently would not include Ohnologs expressed only in different tissues or at different time points. The recent Cc4R allotetraploidy event[35,95,96] complicates transcriptome assembly because there has been little time for divergence of Ohnologs e.g, 8.2 million years.[40] In autopolyploid salmonids, the fourth round of whole genome duplication is much older i.e. 90–102 ma;[97] yet many Ohnologous loci are difficult to separate via bioinformatic approaches. Some loci even maintain tetrasomic inheritance because of the autopolyploid nature of the duplication.[98] These factors need to be explicitly considered when conducting analyses that require orthologous alignments, such as

RNA-seq and syntenic mapping, when working with polyploid or partially diploidized species.[38]

4.6. Summary

Results from short read sequences yield high-quality transcriptome resources for comparative study of cyprinids, a hyper-diverse clade of fishes. We used a variety of bioinformatic tools for assembly quality assessment, gene annotation, orthology assignment, and identification and partitioning of exogenous DNA in wild cyprinid fishes. This approach facilitates technology transfer from a model organism (zebrafish) to a group of related species that fill diverse and critical roles in these ecosystems and comprise an important component of biodiversity. Conserved expression of essential developmental genes across a broad phylogenetic scope, later ontogenetic stages, and array of tissue types, illustrates their utility as benchmarks for assessing coverage in *de novo* assemblies. Moreover, their ubiquitous expression further supports the hypothesis that these genes are required for the basic biology of cyprinid fish and are candidate loci for developmental abnormalities and disease. Finally, comparative transcriptomics must contend with genome duplications and other genomic 'events' that affect gene identity and expression. Nonetheless, comparative approaches could provide enormous power to identify shared and unique physiological pathways that respond to common environmental stressors in a natural setting.

4.7. Data Availability

Raw sequence reads were uploaded to the NCBI Sequence Read Archive (SRA: SRP107991: SRR5601334-SRR560133469. BIOPROJECT: PRJNA383604. BIOSAMPLES: SAMN07166458-SAMN0716493). TRINITY-assembled transcriptomes are available via FigShare. BLAST results and the list of contigs corresponding to potentially novel genes are available in the Supplementary data.

Acknowledgements

This project was supported by the National Institute of General Medical Sciences (8P20GM103451-12), New Mexico IDeA Networks of Biomedical Research Excellence (NMINBRE_A2_Jan_2013), and the Center for Evolutionary and Theoretical Immunology. Samples were collected under New Mexico Department of Game and Fish permit #3015. This research was approved by Institutional Animal Care and Use Committee Protocol #10-100468-MCC and #10-100492-MCC. We thank E. Loker, R. Miller, J. Kavka, and G. Rosenberg for research and technical support. Thanks to Z. Ren and L. Hao for assistance with the Database of Essential Genes. Fish images were provided T. Kennedy (red shiner, flathead chub) and C. Thomas (common carp). This research benefitted from insight and technical assistance provided by F. Schilkey, N. Devitt, P. Mena, T. Ramaraj, I. Lindquist, A. Snyder, M. Osborne, and C. Krabbenhoft. We thank the anonymous reviewers for helpful comments on the manuscript. The authors have no competing interests.

Accession numbers

NCBI Genbank SRA: SRP107991: SRR5601334-SRR560133469

Conflict of interest

None declared.

Supplementary data

Supplementary data are available at *DNARES* online.

References

1. Mardis, E. R. 2008, The impact of next-generation sequencing technology on genetics. *Trends Genet.*, **24**, 133–41.
2. Shendure, J. and Ji, H. 2008, Next-generation DNA sequencing. *Nat. Biotechnol.*, **26**, 1135–45.
3. Ekblom, R. and Galindo, J. 2011, Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity*, **107**, 1–15.
4. Alvarez, M., Schrey, A. W. and Richards, C. L. 2015, Ten years of transcriptomics in wild populations: what have we learned about their ecology and evolution? *Mol. Ecol.*, **24**, 710–25.
5. Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G. and Hohenlohe, P. A. 2016, Harnessing the power of RADseq for ecological and evolutionary genomics. *Nat. Rev. Genet.*, **17**, 81–92.
6. Barshis, D. J., Ladner, J. T., Oliver, T. A., Seneca, F. O., Traylor-Knowles, N. and Palumbi, S. R. 2013, Genomic basis for coral resilience to climate change. *Proc. Natl. Acad. Sci.*, **110**, 1387–92.
7. Schunter, C., Vollmer, S. V., Macpherson, E. and Pascual, M. 2014, Transcriptome analyses and differential gene expression in a non-model fish species with alternative mating tactics. *BMC Genomics*, **15**, 167.
8. Verbruggen, B., Bickley, L. K., Santos, E. M., et al. 2015, De novo assembly of the *Carcinus maenas* transcriptome and characterization of innate immune system pathways. *BMC Genomics*, **16**, 1.
9. Pavey, S. A., Bernatchez, L., Aubin-Horth, N. and Landry, C. R. 2012, What is needed for next-generation ecological and evolutionary genomics? *Trends Ecol. Evol.*, **27**, 673–78.
10. Van Straalen, N. M. and Roelofs, D. 2012, *An Introduction to Ecological Genomics*. Oxford University Press.
11. Landry, C. R. and Aubin-Horth, N. 2014, Recent advances in ecological genomics: from phenotypic plasticity to convergent and adaptive evolution and speciation. *Adv. Exp. Med. Biol.*, **781**, 1–5.
12. Goltsman, D. S. A., Comolli, L. R., Thomas, B. C. and Banfield, J. F. 2015, Community transcriptomics reveals unexpected high microbial diversity in acidophilic biofilm communities. *ISME J.*, **9**, 1014–23.
13. DeBiasse, M. B. and Kelly, M. W. 2016, Plastic and evolved responses to global change: what can we learn from comparative transcriptomics? *J. Hered.*, **107**, 71–81.
14. Eo, S. H., Doyle, J. M., Hale, M. C., Marra, N. J., Ruhl, J. D. and DeWoody, J. A. 2012, Comparative transcriptomics and gene expression in larval tiger salamander (*Ambystoma tigrinum*) gill and lung tissues as revealed by pyrosequencing. *Gene*, **492**, 329–38.
15. Whitehead, A., Triant, D., Champlin, D. and Nacci, D. 2010, Comparative transcriptomics implicates mechanisms of evolved pollution tolerance in a killifish population. *Mol. Ecol.*, **19**, 5186–203.
16. Cohen, D., Bogeat-Triboulot, M.-B., Tisserant, E., et al. 2010, Comparative transcriptomics of drought responses in Populus: a meta-analysis of genome-wide expression profiling in mature leaves and root apices across two genotypes. *BMC Genomics*, **11**, 1.
17. Des Marais, D. L. and Rausher, M. D. 2008, Escape from adaptive conflict after duplication in an anthocyanin pathway gene. *Nature*, **454**, 762–65.
18. Reid, N. M., Proestou, D. A., Clark, B. W., et al. 2016, The genomic landscape of rapid repeated evolutionary adaptation to toxic pollution in wild fish. *Science*, **354**, 1305–8.
19. Thompson, A. W. and Ortí, G. 2016, Annual killifish transcriptomics and candidate genes for metazoan diapause. *Mol. Biol. Evol.*, **33**, 2391–95.
20. Dion-Côté, A.-M., Renaut, S., Normandeau, E. and Bernatchez, L. 2014, RNA-seq reveals transcriptomic shock involving transposable elements reactivation in hybrids of young lake whitefish species. *Mol. Biol. Evol.*, **31**, 1188–99.
21. Kavembe, G. D., Franchini, P., Irisarri, I., Machado-Schiaffino, G. and Meyer, A. 2015, Genomics of adaptation to multiple concurrent stresses: insights from comparative transcriptomics of a Cichlid fish from one of

earth's most extreme environments, the Hypersaline Soda Lake Magadi in Kenya, East Africa. *J. Mol. Evol.*, **81**, 90–109.

22. Kelley, J. L., Passow, C. N., Plath, M., Arias Rodriguez, L., Yee, M. C. and Tobler, M. 2012, Genomic resources for a model in adaptation and speciation research: characterization of the *Poecilia mexicana* transcriptome. *BMC Genomics*, **13**, 652.
23. Deyholos, M. K. 2010, Making the most of drought and salinity transcriptomics. *Plant Cell Environ.*, **33**, 648–54.
24. Ramachandran, V. K., East, A. K., Karunakaran, R., Downie, J. A. and Poole, P. S. 2011, Adaptation of *Rhizobium leguminosarum* to pea, alfalfa and sugar beet rhizospheres investigated by comparative transcriptomics. *Genome Biol.*, **12**, 1.
25. Cheviron, Z. A., Connaty, A. D., McClelland, G. B. and Storz, J. F. 2014, Functional genomics of adaptation to hypoxic cold-stress in high-altitude deer mice: transcriptomic plasticity and thermogenic performance *Evolution*, **68**, 48–62.
26. Webster, T. M. U. and Santos, E. M. 2015, Global transcriptomic profiling demonstrates induction of oxidative stress and of compensatory cellular stress responses in brown trout exposed to glyphosate and Roundup. *BMC Genomics*, **16**, 1.
27. Nelson, J. S. 2006, *Fishes of the World*. John Wiley & Sons: Hoboken, NJ.
28. Berra, T. M. 2001, *Freshwater fish distribution*. Academic Press: San Diego, CA.
29. Winfield, I. and Townsend, C. 1991, The role of cyprinids in ecosystems. In: Winfield, I. and Nelson, J. (eds), *Cyprinid Fishes*, Springer, 552–71.
30. Sterner, R. W. and George, N. B. 2000, Carbon, nitrogen, and phosphorous stoichiometry of cyprinid fishes. *Ecology*, **81**, 127–40.
31. Power, M. E., Matthews, W. J. and Stewart, A. J. 1985, Grazing minnows, piscivorous bass, and stream algae: dynamics of a strong interaction. *Ecology*, **66**, 1448–56.
32. Amores, A., Force, A., Yan, Y.-L., et al. 1998, Zebrafish hox clusters and vertebrate genome evolution. *Science*, **282**, 1711–14.
33. Postlethwait, J. H., Yan, Y.-L., Gates, M. A., et al. 1998, Vertebrate genome evolution and the zebrafish gene map. *Nature Genet.*, **18**, 345–49.
34. Meyer, A. and Van de Peer, Y. 2005, From 2R to 3R: evidence for a fish-specific genome duplication (FSGD). *BioEssays*, **27**, 937–45.
35. David, L., Blum, S., Feldman, M. W., Lavi, U. and Hillel, J. 2003, Recent duplication of the common carp (*Cyprinus carpio* L.) genome as revealed by analyses of microsatellite loci. *Mol. Biol. Evol.*, **20**, 1425–34.
36. Machordom, A. and Doadrio, I. 2001, Evolutionary history and speciation modes in the cyprinid genus *Barbus*. *Proc. Roy. Soc. Lond B Biol. Sci.*, **268**, 1297–306.
37. Ohno, S., Muramoto, J., Christian, L. and Atkin, N. B. 1967, Diploid-tetraploid relationship among old-world members of the fish family Cyprinidae. *Chromosoma*, **23**, 1–9.
38. Yang, L., Sado, T., Hirt, M. V., et al. 2015, Phylogeny and polyploidy: Resolving the classification of cyprinine fishes (Teleostei: Cypriniformes). *Mol. Phylogenet. Evol.*, **85**, 97–116.
39. Wang, J.-T., Li, J.-T., Zhang, X.-F. and Sun, X.-W. 2012, Transcriptome analysis reveals the time of the fourth round of genome duplication in common carp (*Cyprinus carpio*). *BMC Genomics*, **13**, 96.
40. Xu, P., Zhang, X., Wang, X., et al. 2014, Genome sequence and genetic diversity of the common carp, *Cyprinus carpio*. *Nature Genet.*, **46**, 1212–1219.
41. Woods, T. D. and Buth, D. G. 1984, High level of gene silencing in the tetraploid goldfish. *Biochem. Syst. Ecol.*, **12**, 415–21.
42. Ferris, S. D. and Whitt, G. S. 1977, The evolution of duplicate gene expression in the carp (*Cyprinus carpio*). *Experientia*, **33**, 1299–1301.
43. Briggs, J. P. 2002, The zebrafish: a new model organism for integrative physiology. *Am. J. Physiol. Regul. Integr. Comp. Physiol.*, **282**, R3–R9.
44. Dooley, K. and Zon, L. I. 2000, Zebrafish: a model system for the study of human disease. *Curr. Opinion Genet. Dev.*, **10**, 252–56.
45. Lieschke, G. J. and Currie, P. D. 2007, Animal models of human disease: zebrafish swim into view. *Nat. Rev. Genet.*, **8**, 353–367.
46. Howe, K., Clark, M. D., Torroja, C. F., et al. 2013, The zebrafish reference genome sequence and its relationship to the human genome. *Nature*, **496**, 498–503.
47. Ankley, G. T. and Villeneuve, D. L. 2006, The fathead minnow in aquatic toxicology: past, present and future. *Aquatic Toxicol.*, **78**, 91–102.
48. Klaper, R., Carter, B. J., Richter, C., Drevnick, P., Sandheinrich, M. and Tillitt, D. 2008, Use of a 15 k gene microarray to determine gene expression changes in response to acute and chronic methylmercury exposure in the fathead minnow *Pimephales promelas* Rafinesque. *J. Fish Biol.*, **72**, 2207–80.
49. Burns, F. R., Cogburn, A. L., Ankley, G. T., et al. 2016, Sequencing and de novo draft assemblies of a fathead minnow (*Pimephales promelas*) reference genome. *Environ. Toxicol. Chem.*, **35**, 212–217.
50. FAO 2014, *FISHSTAT: Global Aquaculture Production*. FAO, Rome, Italy.
51. Ji, P., Liu, G., Xu, J., et al. 2012, Characterization of common carp transcriptome: *de novo* sequencing, assembly, annotation and comparative genomics. *PloS One*, **7**, 1–9.
52. Li, G., Zhao, Y., Liu, Z., et al. 2015, De novo assembly and characterization of the spleen transcriptome of common carp (*Cyprinus carpio*) using illumina paired-end sequencing. *Fish Shellfish Immunol.*
53. Dekay, J. E. 1842, *Zoology of New-York, or, the New-York fauna; comprising detailed descriptions of all the animals hitherto observed within the state of New-York, with brief notices of those occasionally found near its borders, and accompanied by appropriate illustrations. Part IV. Fishes*. W. & A. White & J. Visscher: Albany, NY.
54. McDonald, M. 1893, Report of the commissioner of fish and fisheries for the fiscal years 1889-90 and 1890-91. Part XVII. *U.S. Commission of Fish and Fisheries. Washington, DC*, 1–96.
55. Walters, D. M., Blum, M. J., Rashleigh, B., Freeman, B. J., Porter, B. A. and Burkhead, N. M. 2008, Red shiner invasion and hybridization with blacktail shiner in the upper Coosa River, USA. *Biol. Invas.*, **10**, 1229–1242.
56. Marsh-Matthews, E. and Matthews, W. J. 2000, Spatial variation in relative abundance of a widespread, numerically dominant fish species and its effect on fish assemblage structure. *Oecologia*, **125**, 283–292.
57. Hoagstrom, C. W., DeWitte, A. C., Gosch, N. J. and Berry, C. R., Jr 2007, Historical fish assemblage flux in the Cheyenne River below Angostura Dam. *J. Freshwater Ecol.*, **22**, 219–29.
58. Amsterdam, A., Nissen, R. M., Sun, Z., Swindell, E. C., Farrington, S. and Hopkins, N. 2004, Identification of 315 genes essential for early zebrafish development. *Proc. Natl. Acad. Sci USA*, **101**, 12792–97.
59. Chen, W.-H., Minguez, P., Lercher, M. J. and Bork, P. 2012, OGEE: an online gene essentiality database. *Nucleic Acids Res.*, **40**, D901–D906.
60. Flicek, P., Amode, M. R., Barrell, D., et al. 2014, Ensembl 2014. *Nucleic Acids Res.*, **42**, D749–755.
61. Hahn, M. W. and Kern, A. D. 2005, Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Mol. Biol. Evol.*, **22**, 803–6.
62. Innan, H. and Kondrashov, F. 2010, The evolution of gene duplications: classifying and distinguishing between models. *Nat. Rev. Genet.*, **11**, 97–108.
63. Chain, F. J., Dushoff, J. and Evans, B. J. 2011, The odds of duplicate gene persistence after polyploidization. *BMC Genomics*, **12**, 1.
64. Bolger, A. M., Lohse, M. and Usadel, B. 2014, Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, btu170.
65. Grabherr, M. G., Haas, B. J., Yassour, M., et al. 2011, Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.*, **29**, 644–52.
66. Haas, B. J., Papanicolaou, A., Yassour, M., et al. 2013, De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protocols*, **8**, 1494–512.
67. Pruesse, E., Quast, C., Knittel, K., et al. 2007, SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res.*, **35**, 7188–96.
68. Conesa, A., Gotz, S., Garcia-Gomez, J. M., Terol, J., Talon, M. and Robles, M. 2005, Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, **21**, 3674–76.
69. Langmead, B. and Salzberg, S. L. 2012, Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–59.

70. Li, B. and Dewey, C. N. 2011, RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinform.*, **12**, 323.
71. Langfelder, P. and Horvath, S. 2008, WGCNA: an R package for weighted correlation network analysis. *BMC Bioinform.*, **9**, 559.
72. Mi, H., Muruganujan, A. and Thomas, P. D. 2013, PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Res.*, **41**, D377–386.
73. Supek, F., Bosnjak, M., Skunca, N. and Smuc, T. 2011, REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One*, **6**, e21800.
74. van Dijk, E. L., Auger, H., Jaszczyszyn, Y. and Thermes, C. 2014, Ten years of next-generation sequencing technology. *Trends Genet.*, **30**, 418–26.
75. Bernatchez, L. 2016, On the maintenance of genetic variation and adaptation to environmental change: considerations from population genomics in fishes. *J. Fish Biol.*, **89**, 2519–56.
76. Martin, J. A. and Wang, Z. 2011, Next-generation transcriptome assembly. *Nat. Rev. Genet.*, **12**, 671–82.
77. Li, B., Fillmore, N., Bai, Y., et al. 2014, Evaluation of de novo transcriptome assemblies from RNA-Seq data. *Genome Biol.*, **15**, 1.
78. O'Neil, S. T. and Emrich, S. J. 2013, Assessing De Novo transcriptome assembly metrics for consistency and utility. *BMC Genomics*, **14**, 465.
79. Lopez-Maestre, H., Brinza, L., Marchet, C., et al. 2016, SNP calling from RNA-seq data without a reference genome: identification, quantification, differential analysis and impact on the protein sequence. *Nucleic Acids Res.*, **44**, e148–e148.
80. DeWoody, J. A., Abts, K. C., Fahey, A. L., et al. 2013, Of contigs and quagmires: next-generation sequencing pitfalls associated with transcriptomic studies. *Mol. Ecol. Resour.*, **13**, 551–58.
81. Wolf, J. B. 2013, Principles of transcriptome analysis and gene expression quantification: an RNA-seq tutorial. *Mol. Ecol. Resour.*, **13**, 559–72.
82. Vijay, N., Poelstra, J. W., Künstner, A. and Wolf, J. B. 2013, Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive in silico assessment of RNA-seq experiments. *Mol. Ecol.*, **22**, 620–34.
83. Aravind, L., Watanabe, H., Lipman, D. J. and Koonin, E. V. 2000, Lineage-specific loss and divergence of functionally linked genes in eukaryotes. *Proc. Natl. Acad. Sci.*, **97**, 11319–24.
84. Stein, C., Caccamo, M., Laird, G. and Leptin, M. 2007, Conservation and divergence of gene families encoding components of innate immune response systems in zebrafish. *Genome Biol.*, **8**, 1.
85. Small, C., Bassham, S., Catchen, J., et al. 2016, The genome of the Gulf pipefish enables understanding of evolutionary innovations. *Genome Biol.*, **17**, 258.
86. Di Bella, J. M., Bao, Y., Gloor, G. B., Burton, J. P. and Reid, G. 2013, High throughput sequencing methods and analysis for microbiome research. *J. Microbiol. Methods*, **95**, 401–14.
87. Coco, J. R., Flemington, E. K. and Taylor, C. M. 2011, PARSES: a pipeline for analysis of RNA-Seq exogenous sequences. *BICoB*, 196–200.
88. Dickerson, J. E., Zhu, A., Robertson, D. L. and Hentges, K. E. 2011, Defining the role of essential genes in human disease. *PLoS One*, **6**, e27368.
89. Luo, H., Lin, Y., Gao, F., Zhang, C.-T. and Zhang, R. 2014, DEG 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements. *Nucleic Acids Res.*, **42**, D574–D580.
90. Plough, L., Shin, G. and Hedgecock, D. 2016, Genetic inviability is a major driver of type-III survivorship in experimental families of a highly fecund marine bivalve. *Mol. Ecol.*
91. McCurley, A. T. and Callard, G. V. 2008, Characterization of housekeeping genes in zebrafish: male-female differences and effects of tissue type, developmental stage and chemical treatment. *BMC Mol. Biol.*, **9**, 1.
92. Saitoh, K., Sado, T., Doosey, M. H., et al. 2011, Evidence from mitochondrial genomics supports the lower Mesozoic of South Asia as the time and place of basal divergence of cypriniform fishes (Actinopterygii: Ostariophysi). *Zool. J. Linnean Soc.*, **161**, 633–62.
93. Postlethwait, J. H. 2007, The zebrafish genome in context: ohnologs gone missing. *J. Exp. Zool. B Mol. Dev. Evol.*, **308**, 563–77.
94. Ferris, S. D. and Whitt, G. S. 1977, Loss of duplicate gene expression after polyploidisation. *Nature*, **265**, 258–60.
95. Larhammar, D. and Risinger, C. 1994, Molecular genetic aspects of tetraploidy in the common carp Cyprinus carpio. *Mol. Phylogenet. Evol.*, **3**, 59–68.
96. Ohno, S. 1970, *Evolution by Gene Duplication*. Springer: New York, NY.
97. Berthelot, C., Brunet, F., Chalopin, D., et al. 2014, The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. *Nat. Commun.*, **5**, 3657.
98. Timusk, E. R., Ferguson, M. M., Moghadam, H. K., Norman, J. D., Wilson, C. C. and Danzmann, R. G. 2011, Genome evolution in the fish family salmonidae: generation of a brook charr genetic map and comparisons among charrs (Arctic charr and brook charr) with rainbow trout. *BMC Genet.*, **12**, 1–15.