

1 ANALYTICAL CONSIDERATIONS FOR COMPARATIVE TRANSCRIPTOMICS OF WILD ORGANISMS

2

3 Trevor J. Krabbenhoft<sup>1,2</sup> & Thomas F. Turner<sup>1</sup>

4 <sup>1</sup>Department of Biology and Museum of Southwestern Biology, University of New Mexico,  
5 Albuquerque, NM 87131.

6 <sup>2</sup>Present address: Department of Biological Sciences, 5047 Gullen Mall, Wayne State University,  
7 Detroit, MI 48202.

8

9 Keywords: ecological genomics, RNA-seq, essential genes, gene silencing, microbiome, non-  
10 model

11

12 Corresponding Author: Trevor Krabbenhoft, Department of Biological Sciences, 5047 Gullen  
13 Mall, Wayne State University, Detroit, MI 48202. Email: [Krabbenhoft@wayne.edu](mailto:Krabbenhoft@wayne.edu).

14

15 RUNNING TITLE: COMPARATIVE TRANSCRIPTOMICS IN THE WILD

16

17 ABSTRACT

18 Comparative transcriptomics can now be conducted on organisms in natural settings,  
19 which has greatly enhanced understanding of genome-environment interactions. However,  
20 important data handling and quality control challenges remain, particularly when working with  
21 non-model species outside of a controlled laboratory environment. Here, we demonstrate the  
22 utility and potential pitfalls of comparative transcriptomics of wild organisms, with an example  
23 from three cyprinid fish species (Teleostei: Cypriniformes). We present computational solutions  
24 for processing, annotating and summarizing comparative transcriptome data for assessing  
25 genome-environment interactions across species. The resulting bioinformatics pipeline  
26 addresses the following points: (1) the potential importance of “essential genes”, (2) the  
27 influence of microbiomes and other exogenous DNA, (3) potentially novel, species-specific  
28 genes, and (4) genomic rearrangements (e.g., whole genome duplication). Quantitative  
29 consideration of these points contributes to a firmer foundation for future comparative work  
30 across distantly related taxa for a variety of sub-disciplines, including stress and immune  
31 response, community ecology, ecotoxicology, and climate change.

32

33 INTRODUCTION

34 High throughput sequencing has dramatically accelerated the pace of genomic research.  
35 While once restricted to model species in laboratory settings, genomic methods are being widely  
36 applied to non-model species in nature, rapidly illuminating the black box of the genome and  
37 giving rise to the field of ecological genomics. Reduced sequencing costs have made it feasible  
38 to study transcriptomes of co-occurring species in a community ecology context (i.e.,

39 “community transcriptomics”), as well as comparative studies of transcriptome evolution across  
40 diverse clades (i.e., “comparative transcriptomics”). While genomic data from model species  
41 can be informative for the biology of related organisms, not all species are the same in terms of  
42 their ecology, genetics, and morphology. For example, research on the zebrafish (*Danio rerio*,  
43 family Cyprinidae) can be relevant for other members of this hyper-diverse clade, but cannot  
44 explain the tremendous ecological and morphological diversity in this clade, as studies of single  
45 species are insufficient for understanding dynamic interactions among species and their  
46 respective genomes in a macroevolutionary context. In order to understand the causes and  
47 consequences of those interactions, as well as the origin of evolutionary novelty (e.g., new  
48 genes), we must examine genomes across species that reflect that diversity.

49 Comparative transcriptomics of co-occurring species in the wild has enormous potential  
50 to advance our understanding of mechanisms underlying molecular adaptation, evolutionary  
51 diversification, and community ecology. In this context, several important questions arise. For  
52 example: What are the proximate and ultimate mechanisms underlying phylogenetic, ecological,  
53 and morphological divergence? How have ancestral genomes been molded by divergent natural  
54 selection and other evolutionary forces into myriad forms that exist today? How does genomic  
55 architecture constrain or promote diversification? How important are genome duplication events  
56 in adaptive radiations? What role do genomes play in underlying the ecological dynamics of  
57 community assembly (e.g., competition, abundance, spatial and temporal dynamics,  
58 physiological constraint, etc.)?

59 Comparative study of transcriptomes in the wild presents important bioinformatic  
60 challenges that must be addressed in order to produce quality assemblies and annotations. For

61 example, the degree to which information from model-organism databases can be extended to  
62 related non-model species is unknown and sources of genetic material in a given sample need to  
63 be partitioned correctly (e.g., target species versus microbiome). In this paper, we illustrate a  
64 computational strategy for comparative analysis of transcriptomes from a hyper-diverse lineage  
65 (>2,000 species) of vertebrates, the family Cyprinidae. The family includes zebrafish (*Danio*  
66 *rerio*) (Briggs 2002; Dooley & Zon 2000; Lieschke & Currie 2007), a model species with a  
67 comprehensively-annotated genome (Howe *et al.* 2013). We evaluated the application and  
68 limitations of zebrafish resources for annotating transcriptomes of three evolutionarily related,  
69 but non-model species: *Cyprinella lutrensis* (red shiner), *Platygobio gracilis* (flathead chub) and  
70 *Cyprinus carpio* (common carp).

71 Transcriptomes of *Cyprinella lutrensis* and *Platygobio gracilis* have not been published  
72 to our knowledge, whereas genomic and transcriptomic data are available for *Cyprinus carpio* (Ji  
73 *et al.* 2012; Wang *et al.* 2012; Xu *et al.* 2014). *Cyprinella lutrensis* and *Platygobio gracilis* are  
74 diploid ( $2n = 50$ ), while *Cyprinus carpio* is allotetraploid ( $2n = 100$ ; Ohno *et al.* 1967), with  
75 some duplicated genes silenced after a lineage specific whole genome duplication (referred to  
76 here as “Cc4R”). Our aims in this study were to: (1) succinctly summarize and compare genes  
77 and functional annotation information obtained from various databases; (2) test whether  
78 *Cyprinus carpio* expresses additional copies of particular genes compared to the two diploid  
79 species (*Cyprinella lutrensis* and *Platygobio gracilis*); (3) identify potentially novel genes  
80 present in the three cyprinids but not found in zebrafish or other genome databases; and (4) to  
81 assess evolutionary conservation of essential genes for development.

82           In zebrafish, 307 genes are known to be essential for development. Knockout mutations  
83   in these genes are embryonic lethal according to experiments by Amsterdam *et al.* (2004), with  
84   subsequent revisions by Chen *et al.* (2012) and updates to the ENSEMBL database (Flicek *et al.*  
85   2014). These genes are highly conserved across extremely deep phylogenetic branches (e.g.,  
86   yeast, fly, zebrafish, and human) due to their essential roles in development (Amsterdam *et al.*  
87   2004). Despite their importance, essential genes have not been studied in the context of  
88   comparative molecular ecology or ecological genomics of co-occurring species. Using  
89   transcriptome data presented in this study, we assessed the evolutionary conservation of the 307  
90   zebrafish essential genes across three additional cyprinid lineages. We predicted that these genes  
91   would be highly conserved across all species, consistent with their critical functional roles. We  
92   also tested whether both copies of duplicated genes in *C. carpio* (i.e., Cc4R Ohnologs) were  
93   retained and expressed in duplicate or whether one copy was evolutionarily lost.

94

95

#### MATERIALS AND METHODS

96           Fish (n = 3 per species) were collected with a seine on 6 July 2012 from a field site on the  
97   Rio Grande, approximately 40 km south of Socorro, New Mexico (33.690556°N,  
98   106.993042°W). Whole fish samples were immediately frozen in liquid nitrogen and transported  
99   to the laboratory. Skin, gill, gut, and kidney tissues were dissected as they thawed, placed in  
100   TRIZol (Invitrogen), and mechanically homogenized. Total RNA was isolated using Purelink  
101   RNA Mini kits (Ambion) following manufacturer's protocol, along with DNase treatment to  
102   reduce genomic DNA contamination. Purified total RNA was sent to the National Center for  
103   Genome Resources (Santa Fe, New Mexico, USA) for quantification, quality assessment, cDNA

104 library preparation and sequencing. RNA integrity and purity was assessed with a Bioanalyzer  
105 2100 instrument (Agilent Technologies). Thirty-six Illumina libraries were constructed (3  
106 species x 4 tissues x 3 biological replicates) from the total RNA samples using Illumina TruSeq  
107 DNA prep kits according to the manufacturer's protocol. Libraries were barcoded using standard  
108 six base pair Illumina oligonucleotides, and six libraries were pooled for each lane of Illumina  
109 HiSeq 2000 (V3 chemistry) for a total of six lanes of 2 x 100 bp paired-end sequencing.

110

111 *Bioinformatics.*—We developed a bioinformatics pipeline (Fig. 1) for analyzing transcriptomic  
112 data in three main steps: *de novo* assembly, gene annotation, and analysis of expression of  
113 duplicated genes. Adapters and barcode sequences were removed from raw reads, and reads  
114 were trimmed using TRIMMOMATIC (Bolger *et al.* 2014) with parameter settings as follows:  
115 leading quality = 5; trailing quality = 5; minimum trimmed read length = 36). Reads were  
116 normalized *in silico* to maximum read coverage of 50X. Clipped and trimmed reads were  
117 assembled, *de novo*, for each species separately using TRINITY (version 2014-04-13; Grabherr *et*  
118 *al.* 2011; Haas *et al.* 2013), with minimum contig length set to 200 bp. TRINITY assembles reads  
119 into contigs (“TRINITY transcripts”), places similar transcripts in groups loosely referred to as  
120 “genes”, and groups similar “genes” into gene clusters. *De novo* TRINITY assemblies were  
121 annotated using TRINOTATE, a comprehensive annotation package distributed with the TRINITY  
122 package suite. TRINOTATE was used to query contigs against the following databases or search  
123 tools: BLASTx, BLASTp, Pfam, SignalP, Uniprot, eggnoG, and gene ontology.

124 Putative protein coding genes were also identified by BLASTx searches of contigs  
125 against zebrafish (*Danio rerio*) peptide sequences (database build Zv9) obtained from Ensembl

126 78 (Flicek *et al.* 2014). Significant BLAST hits were identified based on the following parameter  
127 settings: E-value < 0.0001; gap open penalty = 11; gap extend = 1; wordsize = 3. After  
128 extensive testing, this parameter combination was found to give the optimal balance between  
129 finding matches for large numbers of contigs, while minimizing spurious hits. For most genes a  
130 1-1 match was expected between zebrafish versus *Platygobio gracilis* or *Cyprinella lutrensis*,  
131 whereas zebrafish and *Cyprinus carpio* should have either 1-2 or 1-1 due to partial diploidy in  
132 carp. We used this expectation in deciding the threshold E-value to use. In practice, more  
133 stringent E-value thresholds (e.g.,  $E < 1e-6$ ) had very little effect on the number of significant  
134 BLAST hits.

135 Contigs with no significant BLAST hits against the zebrafish transcriptome were  
136 subjected to a series of stepwise BLASTn searches until significant hits were found (or not) in  
137 order to identify the possible sources of those sequences (e.g., microbiome) or to identify novel  
138 genes not present in the zebrafish genome. First, remaining contigs lacking significant hits  
139 against the zebrafish transcriptome were queried against the rRNA silva database (SSU Ref 119  
140 NR99 and LSU Parc 119), which contains bacterial and eukaryotic rRNA sequences (Pruesse *et*  
141 *al.* 2007). Contigs with still no significant BLAST hits were then queried against a database  
142 containing all nine additional teleost fish transcriptomes (Amazon molly, *Poecilia formosa*;  
143 cavefish, *Astyanax mexicanus*; cod, *Gadus morhua*; fugu, *Takifugu rubripes*; medaka, *Oryzias*  
144 *latipes*; platyfish, *Xiphophorus maculatus*; stickleback, *Gasterosteus aculeatus*; tetraodon,  
145 *Tetraodon nigroviridis*; tilapia, *Oreochromis niloticus*) from Ensembl 78. Contigs with no  
146 BLAST hits at this point were then BLASTed against the zebrafish genome (Zv9) using the “Top  
147 Level” sequences from Ensembl to identify possible genomic DNA contamination. Remaining

148 contigs with no significant blast hits in any of these databases were piped to TRANSDCODER  
149 (Haas *et al.* 2013) to identify open reading frames (ORFs) that represent potentially novel genes.  
150 Default parameter settings were used with TRANSDCODER. The software generates predicted  
151 peptide sequences for contigs with ORFs. Predicted peptide sequences for the contigs with  
152 ORFs but no BLAST hits to the aforementioned databases were queried (BLASTp; e-value <  
153 0.001) against the NCBI nr database. BLAST2GO (version 3.0, Conesa *et al.* 2005) was used to  
154 identify top species hits for those predicted proteins with significant hits against nr. The  
155 remaining sequences with no hits to databases and no ORFs were discarded as likely non-protein  
156 coding, genomic DNA contamination with sufficient evolutionary divergence from zebrafish to  
157 render genomic BLASTn searches ineffective.

158

159 *Genome duplication, diploidization and gene silencing.*—Trimmed sequence reads were mapped  
160 to TRINITY contigs using BOWTIE2 (version 2.2.2.3; Langmead & Salzberg 2012) and  
161 corresponding gene expression was quantified with RSEM (version 1.2.13; Li & Dewey 2011).  
162 Because RSEM is incompatible with indel, local, and discordant alignments, parameter settings  
163 were chosen to avoid these alignments. The following RSEM parameters were used: --sensitive;  
164 --dpad 0; --gbar 99999999; --mp 1,1 --np 1 --score-min L,0,-0.1; --no-mixed; --no-discordant.  
165 Normalized expression for TRINITY genes was calculated by standardizing by total mapped reads  
166 across libraries and summed across alternate TRINITY transcripts (isoforms) for each locus. In  
167 order to assess the expression of duplicated genes in *Cyprinus carpio* arising from the carp-  
168 specific whole genome duplication (“Cc4R”), we quantified the number of TRINITY genes  
169 present in each species relative to zebrafish genes, as well as their expression levels. We used an



170 arbitrary threshold of ten sequence reads per gene per tissue, summed across all three  
171 individuals, for a given gene to be considered “expressed” in a particular tissue. This approach  
172 was aimed at reducing the influence of unique or nearly unique reads (e.g., sequencing artifacts).  
173 Most of the contigs excluded as a result of this filtering were contigs represented only by  
174 singleton reads in one library.

175 For *C. carpio*, we tested whether certain functional classes of genes were preferentially  
176 expressed in duplicate (i.e., the case where neither ohnolog silenced). For this analysis, we used  
177 PANTHER (MI *ET AL.* 2013) to test for statistical overrepresentation of GO-slim Biological  
178 Processes, with Bonferroni correction. The test genes consisted of the list of *C. carpio* ohnologs  
179 expressed in duplicate, while the list of all *C. carpio* genes present in the assembly was used as  
180 the reference set. GO terminology was based on the zebrafish database. Results of the  
181 overrepresentation analysis were visualized with REVIGO (Supek *et al.* 2011).

182

183 *Essential genes.*—To test the hypothesis of evolutionary conservation of essential genes among  
184 cyprinid fishes, we used zebrafish genes present in the Online Gene Essentiality Database  
185 (OGEE; Chen *et al.* 2012) and identified orthologs in the three transcriptomes from BLASTx  
186 searches described above. Of the 307 essential genes in zebrafish (Amsterdam *et al.* 2004; Chen  
187 *et al.* 2012), one (ENSDARG00000038423) has been retired from ENSEMBL and one  
188 (ENSDARG00000045605) is an unprocessed pseudogene with no protein product. We searched  
189 for the remaining 305 genes in the three transcriptome assemblies to assess their conservation  
190 across the cyprinid phylogeny. We predicted that expression would be conserved due to their  
191 high rate of conservation seen in other organisms (Amsterdam *et al.* 2004).

192

193 *Data accessibility.*—Raw sequence reads were uploaded to the NCBI Sequence Read Archive  
194 (Acc. Nos. SRXXXX, SRXXXX, SRXXXX). The three TRINITY assemblies were archived as  
195 separate bam files in NCBI Transcriptome Shotgun Assembly Database (Acc. Nos. XXXXX,  
196 XXXXX, XXXXX). Because these raw assemblies also contain microbiome and genomic  
197 sequences, we also deposited the final, filtered, high quality transcriptome fasta files in Dryad  
198 (Acc. Nos. XXXXX). These files include TRINITY contigs with significant BLASTx hits for  
199 zebrafish or the nine other fish transcriptomes, as well as companion fasta files containing the  
200 potentially novel genes (i.e., ORFs present but no significant BLAST hits in nr or the other  
201 databases). Annotations files from TRINOTATE were also deposited in Dryad.

202

203

## RESULTS

204 *Sequencing and transcriptome assemblies.*—Six lanes of Illumina sequencing produced more  
205 than 1.2 billion paired-end reads, including 420.5-, 413.9-, and 385.3-million sequences in  
206 *Cyprinus carpio*, *Cyprinella lutrensis*, and *Platygobio gracilis*, respectively. *De novo* assembly  
207 resulted in high quality transcriptomes for all three species (Table 1). The *C. carpio* assembly  
208 had the largest number of contigs (“TRINITY transcripts”) and genes (“TRINITY genes”), while *P.*  
209 *gracilis* had the fewest. In contrast, metrics for contig length (N25, N50, N75, median contig  
210 length, average contig length) were all longer in *P. gracilis* than the other two species (Table 1).  
211 Overall, the *P. gracilis* transcriptome assembly was more complete despite fewer raw sequence  
212 reads. TRANSDCODER predicted ORFs in about half of all TRINITY contigs (Table 2), with the  
213 remainder comprised mainly of genomic DNA contamination that was filtered out of the final

214 dataset. The N50 of predicted ORFs was only slightly shorter in the three species (i.e., 1,299 –  
215 1,572 bp) than in zebrafish (CDS N50 = 2,037 bp), and similar to the recently published draft *C.*  
216 *carpio* genome (1,487 bp; Xu *et al.* 2014). Removal of microbiome and genomic DNA  
217 contamination from the final assembly resulted in fewer, but longer contigs (see filtering of the  
218 final dataset, below), and an overall higher-quality assembly.

219  
220 *BLAST searches: zebrafish transcriptome.*—Top BLASTx hits of TRINITY contigs against  
221 zebrafish peptides included approximately 20,000 unique genes (ENSDARG) and 11,000 protein  
222 families (ENSFAM) present in each of the three species (Fig. 3), suggesting similar annotation  
223 efficiency and transcriptome representation for each species. However, after pooling isoforms,  
224 the number of TRINITY genes that significantly matched these ~20,000 zebrafish genes varied  
225 among species: 66,447 in *Cyprinus carpio*, 60,990 in *Cyprinella lutrensis*, and 39,915 in  
226 *Platygobio gracilis* (Table 3, top row). Zebrafish genes were well covered, with more than  
227 15,000 unique zebrafish genes covered over at least 70% of their length in corresponding contigs  
228 from each of the three cyprinids, consistent with the N50 data presented above. In general,  
229 zebrafish proteins were more completely covered by *P. gracilis* contigs than *C. carpio* or *C.*  
230 *lutrensis*. For example, zebrafish genes were more than 90% covered by sequences in 50.3%  
231 (12,489 of 24,817 genes) of *P. gracilis* genes with significant zebrafish peptide hits versus 49.9%  
232 (13,453 of 26,963) for *C. carpio* and 46.8% (12,538 of 26,817) in *C. lutrensis*. A large number  
233 of TRINITY contigs did not significantly match (BLASTx) zebrafish peptide sequences and were  
234 subsequently queried against several additional databases.

235

236 *BLAST searches: other databases.*—Contigs lacking significant BLASTx hits against zebrafish  
237 peptides were queried (BLASTn) iteratively against rRNA silva microbiome database, nine  
238 teleost transcriptomes, and the zebrafish genome databases (Table 3). For contigs lacking hits  
239 against zebrafish peptides, BLASTn searches versus the rRNA silva database revealed a small  
240 number of significant hits (i.e., <400 contigs; Table 3). BLASTn searches of the remaining  
241 unmatched contigs versus the nine teleost fish transcriptomes identified approximately 1,500 –  
242 4,500 additional hits (Table 3), far fewer than the evolutionarily more closely related zebrafish  
243 transcriptome. BLASTn searches of the remaining unidentified contigs against the zebrafish  
244 genome revealed a large number of significant hits (>30,000 per species), suggesting these reads  
245 were the result of low levels of background genomic DNA contamination in the cDNA libraries,  
246 a common occurrence resulting from the hypersensitivity of Illumina sequencing. Conservation  
247 of sequences across deep evolutionary lineages suggests functional importance, such as non-  
248 coding regulatory regions.

249         Despite extensive BLAST searches, a large number of TRINITY contigs (>100,000 in each  
250 species or more than 50% of all contigs) did not have significant hits in any of the databases.  
251 These, contigs are short in length (i.e., 200 bp) and have few reads mapping to them (e.g., single-  
252 read contigs). These could represent endogenous genomic DNA contamination of cDNA  
253 libraries and have sufficient evolutionary divergence from zebrafish to render BLASTn searches  
254 ineffective. A large number are also expected to be non-rRNA sequences from the microbiome,  
255 which were not present in target databases. Of contigs with no BLAST hits in the  
256 aforementioned databases, TRANSDCODER predicted ORFs in 8,652 (*Cyprinus carpio*), 9,215  
257 (*Cyprinella lutrensis*), and 3,011 (*Platygobio gracilis*) contigs (Table 4). Roughly half of the

258 predicted ORFs had significant BLASTp hits against the nr protein database (3,789, 4,154, and  
259 1,548 contigs, respectively). Conversely, there were 4,863 (*C. carpio*), 5,061 (*C. lutrensis*), and  
260 1,463 (*P. gracilis*) predicted ORFs had no significant hits against nr (Table 4). These ORFs  
261 could include novel genes not present in zebrafish or other teleost models, genes present in  
262 zebrafish but with significantly divergent sequences or exon structure to cause BLAST searches  
263 to miss them, or could include genes from the microbiome that are not present in sequence  
264 databases.

265 For ORFs with nr hits, zebrafish was the top-hit species for a large portion (Fig. 4),  
266 somewhat paradoxically given the lack of significant BLAST hits against zebrafish peptide and  
267 genome sequences discussed above. This appears to be due to the fact that TRANSDCODER-  
268 predicted ORFs exclude 5' and 3' untranslated regions (UTRs) which diverge more rapidly than  
269 ORFs over evolutionary time. In *C. carpio* and *C. lutrensis*, many of these ORFs are from the  
270 microbiome and share significant similarity to cyclophyllid tapeworms (e.g., *Echinococcus*,  
271 *Hymenolepis*) and protozoans (e.g., *Tetrahymena*, *Paramecium*). Conversely, in *P. gracilis* the  
272 ORFs appear to be endogenous genes with high similarity to zebrafish (Fig. 4). Contigs with  
273 predicted ORFs but no BLAST hits to any of the databases possibly represent novel or  
274 functionally divergent genes in these species that warrant further study. These sequences are  
275 available as fasta files in Dryad (XXXXXXXXX).

276

277 *Filtering and the final assembly datasets.*—After filtering and removal of genomic DNA and  
278 microbiome reads, the final *de novo* assembly datasets contained only TRINITY contigs falling  
279 into one of the following categories: (1) contigs with significant BLAST hits against zebrafish

280 or the nine other teleost transcriptomes; or (2) contigs with no matches against any of the  
281 databases but with predicted ORFs present, i.e., potentially novel genes. All other contigs were  
282 removed via bioinformatic filtering. The final datasets are significantly smaller than the raw *de*  
283 *novo* assembly deposited in NCBI TSA, but present much more reliable sequence information,  
284 i.e., actual transcriptome sequences rather than microbiome or genomic DNA contamination.

285

286 *Genome duplication, diploidization and gene silencing.*—Transcriptome annotation and  
287 comparison with zebrafish revealed that *Cyprinus carpio* expresses more genes than *Cyprinella*  
288 *lutrensis* and *Platygobio gracilis*, due to the carp-specific whole genome duplication (Fig. 5).  
289 *Cyprinus carpio* expressed about 41% more genes overall than *P. gracilis* and 11% more than *C.*  
290 *lutrensis*. The number of duplicate genes expressed varied dramatically among tissue types (Fig.  
291 5). In all tissues except skin, *C. carpio* expressed more genes than the other two species (i.e., 3-  
292 48% more). In skin, both *C. lutrensis* and *P. gracilis* expressed more genes than *C. carpio* (26  
293 and 2%, respectively). Using higher thresholds for “expression” had moderate impact on the  
294 inferred percentage of duplicates expressed: a threshold of 100 reads instead of 10 resulted in  
295 different estimates of duplicated genes expressed in *C. carpio* versus *P. gracilis* (18% more in *C.*  
296 *carpio*) and *C. lutrensis* (8% more in *C. carpio*), i.e., retained expression of Cc4R duplicates.  
297 The disparity in these results could be driven in part by different assembly qualities (e.g., a better  
298 assembled *P. gracilis* transcriptome).

299 Genes with retained duplicate expression (i.e., Ohnologs) in *C. carpio* represented a  
300 diverse suite of functional groups: gene ontology terms that were significantly enriched in the  
301 ‘retained duplicates’ list were diverse (Fig. 6, top panel). One functional grouping that was a

302 predominant contributor in the REVIGO analysis was “anatomical structure morphogenesis,” of  
303 interest because common carp attain much larger body size than the other two species (Fig. 6,  
304 bottom panel).

305

306 *Expression of essential genes.*—Genes that are essential for embryonic development in *D. rerio*  
307 were nearly all present in the three cyprinids: 285 (*Platygobio gracilis*), 301 (*Cyprinella*  
308 *lutrensis*), and 301 (*Cyprinus carpio*) genes were expressed out of 305 zebrafish essential genes  
309 (i.e., 93.4 – 97.8%). Essential genes were nearly ubiquitously expressed across all four tissue  
310 types (skin, gill, gut, kidney), with low levels of tissue specificity (Fig. 7), in contrast to non-  
311 essential genes which generally exhibited higher levels of tissue specificity. Normalized levels  
312 of expression were higher in *C. carpio* than *P. gracilis* and *C. lutrensis* for 165 and 204 out of  
313 305 genes, respectively. This pattern was not due to *C. carpio* expressing more loci per  
314 zebrafish gene (e.g., Ohnologs) than the other two species. Only slightly more loci (e.g., n=2  
315 contigs) were expressed per essential gene in the recently duplicated *C. carpio* genome (Fig. 8)  
316 whereas most duplicated essential genes in *C. carpio* are not transcribed and have either been  
317 lost evolutionarily, e.g., pseudogenes, or are expressed in other developmental stages or tissues.

318

319

## DISCUSSION

320 Next-generation transcriptome sequencing has revolutionized the field of molecular  
321 ecology over the past decade. One outcome is increased appreciation for the molecular  
322 complexity underlying the evolution of basic ecological traits. Here we demonstrated the utility  
323 and challenges associated with comparative study of transcriptomes in non-model organisms in a

324 natural setting. Bioinformatic analyses requires careful processing and filtering to assess the  
325 sources of DNA fragments, which can be endogenous target transcriptome sequences, genomic  
326 DNA ‘contamination’ from the study organism, or DNA from the microbiome or diet items.  
327 Assessment of transcriptome quality also requires careful consideration. Traditional measures of  
328 assembled read lengths such as N50 are largely meaningless for transcriptomes without  
329 additional context. We advocate combining N50 and/or histograms of read lengths with explicit  
330 comparisons to well-studied transcriptomes of model organisms, when available. For example,  
331 we compared our *de novo* transcriptomes to zebrafish, which yielded valuable insight into  
332 progress made in our target species. Finally, positive identification of nearly all zebrafish  
333 essential genes in our transcriptomes is an additional test of our annotation procedures. Using the  
334 bioinformatics pipeline presented in Fig. 1, we demonstrate that high quality transcriptome data  
335 can be obtained from wild-caught organisms, which provide valuable tools for molecular  
336 ecologists studying functional genomics in a comparative context and natural settings.

337         We identified several key findings in this study, including: (1) high-quality transcriptome  
338 assemblies that reveal broad similarities and evolutionary conservation of genes with zebrafish,  
339 but with some key differences; (2) several potentially novel genes not identified in zebrafish that  
340 are candidates for studies of ecological novelty; (3) diverse microbiomes that vary strongly  
341 among the three species, despite their origin from a single collection locality; (4) extreme  
342 conservation of expression of essential genes for development; (5) a large number of duplicate  
343 genes expressed in the tetraploid, *Cyprinus carpio*, representing a diverse suite of biological  
344 processes or gene ontologies. We discuss each of these findings in greater detail below.

345



346 *Assembly results.*—There are important considerations associated with conducting transcriptome  
347 analysis in a non-laboratory setting and in species lacking high-quality, well-annotated genomes.  
348 For example, it is necessary to identify ways to maximize the quality and completeness of *de*  
349 *novo* assemblies. Our assemblies are somewhat less complete than the zebrafish reference, as  
350 expected because zebrafish has been sequenced extensively at the genomic DNA level,  
351 empirically validated with RNA-seq, and refined by years of manual curation.

352 TRINITY assemblies resulted in proportionally fewer long contigs (e.g., > 1000 bp)  
353 compared to zebrafish. Four factors account for this result. First, the microbiome is present in  
354 these sequences and many of the contigs are not endogenous, as reflected by top species hits in  
355 BLAST searches (Fig. 4). Second, a small amount of genomic DNA contamination persists  
356 despite DNase treatment during library preparation. Genomic contamination tends to be observed  
357 as short (e.g., 200 bp), shallow contigs often comprised of single-reads. Third, the *de novo*  
358 assemblies are more fragmented due to the short read technology employed, with multiple  
359 contigs often representing non-overlapping fragments of the same gene. This effect is  
360 particularly acute in genes with short sequence repeats, such as microsatellites. Finally, we only  
361 used the canonical zebrafish transcripts in this study, which excludes the shorter isoforms present  
362 in many genes and biases the zebrafish distribution toward longer sequences. Transcriptomes  
363 presented here represent an improvement (i.e., more sequences, higher coverage; longer relative  
364 N50) over earlier work on sequencing and assembling the common carp transcriptome using  
365 Roche 454 sequencing (Ji *et al.* 2012; Wang *et al.* 2012), due to the higher throughput, Illumina  
366 paired-end sequencing approach we employed. The bioinformatic approach we presented to

367 identify and filter non-target sequences from the final dataset resulted in high quality and well  
368 annotated assemblies.

369

370 *Potentially novel genes.*—Results of BLAST searches and ORF predictions helped us identify  
371 candidate genes that may represent novel species- or taxon-specific genes. Our interest in these  
372 genes lies in the idea that they may contain some of the functional elements responsible for  
373 extensive ecological and phylogenetic diversity present in Cyprinidae. Many candidates may  
374 prove to be false positives as more fish genomes are sequenced and annotated; however, these  
375 candidates would be an excellent starting point for researchers interested in targeted searches for  
376 genes or proteins underlying ecological novelty in cyprinids that may have arisen through local  
377 gene duplications, exon shuffling, horizontal transfer, or other mechanisms.

378

379 *Microbiome diversity.*— Another valuable aspect of transcriptome sequencing of samples taken  
380 from nature is the simultaneous generation of quantifiable data on the microbiome. These data  
381 are applicable to study of host-parasite dynamics, immune response, paired comparative  
382 population genetics or phylogeographic analysis of host and microbiota. When generating *de*  
383 *novo* transcriptome assemblies for focal species, it is imperative that microbiome sequences are  
384 identified and filtered out of final assemblies. Genome-scale sequence data is often lacking for  
385 the bacterial and metazoan microbiota on vertebrate samples, which complicates attempts at  
386 removal. We used an iterative and successive filtering approach to address this issue (Fig. 1) that  
387 provides valuable information on the likely source (e.g., exogenous or endogenous) of particular  
388 sequences or contigs. Transcriptome characterization studies often do not attempt to remove

389 exogenous microbiome and genomic DNA contamination. Researchers should be cautious when  
390 using unfiltered sequence reads, particularly when they are compiled into massive databases that  
391 lack appropriate metadata.

392

393 *Conservation of essential genes.*—Genes that are essential for embryonic development present  
394 interesting targets for studying genome evolution due to their critical functional importance.  
395 Zebrafish essential genes in the Online Gene Essentiality Database (OGEE; Chen *et al.* 2012)  
396 were widely transcribed in all three cyprinids and with low levels of tissue specificity. Coverage  
397 of essential genes is a metric that should be used to assess the quality and coverage of *de novo*  
398 transcriptome assemblies, analogous to the use of “housekeeping” genes as positive controls in  
399 qPCR studies. The presence of nearly all essential genes across these four cyprinid species  
400 (representing more than 100 million years of evolutionary time; Saitoh *et al.* 2011) is consistent  
401 with the hypothesis of broad evolutionary and functional conservation. The few essential genes  
402 not detected may still be present, but expressed transiently at larval or juvenile developmental  
403 stages. We propose that the number of essential genes expressed could be used as a metric to  
404 complement other measures of assembly quality and completeness, such as comparing transcript  
405 length histograms to closely related model species (see Fig. 2).

406

407 *Tetraploidy and expression of duplicated genes.*— Our analyses suggested that a large number of  
408 duplicate genes are expressed in *Cyprinus carpio*, representing a diverse suite of biological  
409 processes or gene ontologies. Note that this analysis is based only on expressed genes in these  
410 tissues at a particular time point, rather than genomic DNA sequences and consequently would

411 not include Ohnologs expressed only in different tissues or at different time points. Short reads  
412 in overlapping Ohnologous regions may not contain sufficient sequence divergence in recently  
413 duplicated genomes such as common carp, resulting in assembly of non-orthologous reads in the  
414 *de novo* assembly. The relatively more fragmented transcriptome of *Cyprinus carpio* in this  
415 study may reflect the challenge of assembling a recently duplicated transcriptome with little time  
416 for divergence in Ohnologs (e.g, 8.2 million years, Xu *et al.* 2014). In salmonids, although the  
417 fourth round of whole genome duplication is much older (i.e., 90-102 ma; Berthelot *et al.* 2014);  
418 many Ohnologous loci are still difficult to separate and some even maintain tetrasomic  
419 inheritance (Timusk *et al.* 2011). The draft genome sequence for common carp (Xu *et al.* 2014)  
420 should help identify Ohnologs that are in fact silenced by identifying pseudogenes in genomic  
421 DNA sequence. Previously, Wang *et al.* (2012) identified enrichment of retained duplicates in  
422 gene ontology pathways involved in immune function. Conversely, our analysis suggests for  
423 genes where both Ohnologs were expressed in *C. carpio*, there was enrichment in several  
424 different functional pathways, but in particular many genes were associated with “anatomical  
425 structure morphogenesis” (Fig. 6). This difference may be due to the large body size and rapid  
426 growth in *C. carpio* as compared to *C. lutrensis* and *P. gracilis* and an associated dosage effect.  
427 Ultimately, knowledge of which genes are retained and expressed in duplicate in tetraploids as  
428 compared to related diploid species can provide insight into the role that whole genome  
429 duplication plays in the molecular ecology and phylogenetic diversification of organisms.

430

431 **Summary:** We present a bioinformatic analysis of short read sequences that yields high-quality  
432 transcriptome assemblies for non-model, ecologically-relevant organisms in a natural setting.

433 The bioinformatics solutions presented here refine the process of gene annotation, assembly  
434 quality assessment, orthology assignment, and identification and partitioning of exogenous DNA  
435 in a single informatics pipeline. This approach facilitates technology transfer from a model  
436 organism (zebrafish) to a group of related species that fill diverse and important roles in these  
437 ecosystems and comprise an important component of biodiversity. The conserved expression of  
438 essential developmental genes across a broad phylogenetic scope and array of tissue types, and  
439 illustrated their utility as benchmarks for assessing coverage in *de novo* assemblies. More  
440 broadly, we demonstrate the promise and challenges associated with adapting model organism  
441 data to non-model, wild-caught samples.

442

443

444 ACKNOWLEDGEMENTS

445 This project was supported by the National Institute of General Medical Sciences  
446 (8P20GM103451-12), New Mexico IDeA Networks of Biomedical Research Excellence  
447 (NMINBRE\_A2\_Jan\_2013), and the Center for Evolutionary and Theoretical Immunology.  
448 Samples were collected under New Mexico Department of Game and Fish permit #3015. This  
449 research was approved by Institutional Animal Care and Use Committee Protocol #10-100468-  
450 MCC and #10-100492-MCC. We thank E. Loker, R. Miller, J. Kavka, and G. Rosenberg for  
451 research support. Thanks to Z. Ren and L. Hao for assistance with the Database of Essential  
452 Genes. Fish images were provided T. Kennedy (red shiner, flathead chub) and C. Thomas  
453 (common carp). This research benefitted from insight and technical assistance provided by F.  
454 Schilkey, N. Devitt, P. Mena, T. Ramaraj, I. Lindquist, A. Snyder, M. Osborne, and C.  
455 Krabbenhoft.

456

457 REFERENCES

- 458 Amsterdam A, Nissen RM, Sun Z, *et al.* (2004) Identification of 315 genes essential for early zebrafish  
459 development. *Proc Natl Acad Sci U S A* **101**, 12792-12797.
- 460 Berthelot C, Brunet F, Chalopin D, *et al.* (2014) The rainbow trout genome provides novel insights into  
461 evolution after whole-genome duplication in vertebrates. *Nat Commun* **5**, 3657.
- 462 Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data.  
463 *Bioinformatics*, btu170.
- 464 Briggs JP (2002) The zebrafish: a new model organism for integrative physiology. *American Journal of*  
465 *Physiology-Regulatory, Integrative and Comparative Physiology* **282**, R3-R9.
- 466 Chen W-H, Minguez P, Lercher MJ, Bork P (2012) OGEE: an online gene essentiality database. *Nucleic*  
467 *Acids Res* **40**, D901-D906.
- 468 Conesa A, Gotz S, Garcia-Gomez JM, *et al.* (2005) Blast2GO: a universal tool for annotation, visualization  
469 and analysis in functional genomics research. *Bioinformatics* **21**, 3674-3676.
- 470 Dooley K, Zon LI (2000) Zebrafish: a model system for the study of human disease. *Curr Opin Genet Dev*  
471 **10**, 252-256.
- 472 Flicek P, Amode MR, Barrell D, *et al.* (2014) Ensembl 2014. *Nucleic Acids Res* **42**, D749-755.
- 473 Grabherr MG, Haas BJ, Yassour M, *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data  
474 without a reference genome. *Nat Biotechnol* **29**, 644-652.

- 475 Haas BJ, Papanicolaou A, Yassour M, *et al.* (2013) De novo transcript sequence reconstruction from RNA-  
476 seq using the Trinity platform for reference generation and analysis. *Nature protocols* **8**, 1494-  
477 1512.
- 478 Howe K, Clark MD, Torroja CF, *et al.* (2013) The zebrafish reference genome sequence and its  
479 relationship to the human genome. *Nature* **496**, 498-503.
- 480 Ji P, Liu G, Xu J, *et al.* (2012) Characterization of common carp transcriptome: *de novo* sequencing,  
481 assembly, annotation and comparative genomics. *PLoS One* **7**, 1-9.
- 482 Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359.
- 483 Li B, Dewey CN (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a  
484 reference genome. *BMC Bioinformatics* **12**, 323.
- 485 Lieschke GJ, Currie PD (2007) Animal models of human disease: zebrafish swim into view. *Nature*  
486 *Reviews Genetics* **8**, 353-367.
- 487 Mi H, Muruganujan A, Thomas PD (2013) PANTHER in 2013: modeling the evolution of gene function,  
488 and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Res* **41**, D377-386.
- 489 Ohno S, Muramoto J, Christian L, Atkin NB (1967) Diploid-tetraploid relationship among old-world  
490 members of the fish family Cyprinidae. *Chromosoma* **23**, 1-9.
- 491 Pruesse E, Quast C, Knittel K, *et al.* (2007) SILVA: a comprehensive online resource for quality checked  
492 and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res* **35**, 7188-  
493 7196.
- 494 Saitoh K, Sado T, Doosey MH, *et al.* (2011) Evidence from mitochondrial genomics supports the lower  
495 Mesozoic of South Asia as the time and place of basal divergence of cypriniform fishes  
496 (Actinopterygii: Ostariophysii). *Zoological Journal of the Linnean Society* **161**, 633-662.
- 497 Supek F, Bosnjak M, Skunca N, Smuc T (2011) REVIGO summarizes and visualizes long lists of gene  
498 ontology terms. *PLoS One* **6**, e21800.
- 499 Timusk ER, Ferguson MM, Moghadam HK, *et al.* (2011) Genome evolution in the fish family salmonidae:  
500 generation of a brook charr genetic map and comparisons among charrs (Arctic charr and brook  
501 charr) with rainbow trout. *BMC Genet* **12**, 1-15.
- 502 Wang J-T, Li J-T, Zhang X-F, Sun X-W (2012) Transcriptome analysis reveals the time of the fourth round  
503 of genome duplication in common carp (*Cyprinus carpio*). *BMC Genomics* **13**, 96.
- 504 Xu P, Zhang X, Wang X, *et al.* (2014) Genome sequence and genetic diversity of the common carp,  
505 *Cyprinus carpio*. *Nat Genet* **46**, 1212-1219.

506

507 *Data Accessibility.* Raw DNA Sequences: NCBI SRA (SRXXXX, SRXXXX, and SRXXXX).

508 Transcriptome assemblies: NCBI TSA (XXXXXXXX, XXXXXXXX, and XXXXXXXX). Final,

509 filtered assemblies, fasta files of contigs representing to potentially novel genes, and TRINOTATE

510 annotation output files are available from Dryad (Acc. Nos. XXXXXXX, XXXXXXX, and

511 XXXXXXX).

512

513 *Author Contributions.* TJK and TFT designed and performed the research, TJK conducted the  
514 analyses, and TJK and TFT wrote the manuscript.



515 Table 1. *De novo* transcriptome assembly results. Zebrafish (*Danio rerio*) data is included as an  
516 example of a well-assembled and complete transcriptome based primarily on Sanger sequencing.  
517

	<i>Cyprinus carpio</i>	<i>Cyprinella lutrensis</i>	<i>Platygobio gracilis</i>	<i>Danio rerio</i>
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

518

519 Table 2. Assembly results and predicted coding sequence (CDS) in three study species and  
520 zebrafish. Zebrafish data is from ENSEMBL 78 (Flicek *et al.* 2014).

521

	<i>Cyprinus carpio</i>	<i>Cyprinella lutrensis</i>	<i>Platygobio gracilis</i>	<i>Danio rerio</i>
CDS (predicted)	145,288	130,632	86,877	43,153
CDS total length (bp)	152,746,710	128,684,241	99,482,505	64,757,328
CDS GC content (%)	49.94	51.02	50.25	46.18
CDS N50 (bp)	1,410	1,299	1,572	2,037
CDS longest (bp)	39,605	24,870	17,805	93,656

522

523

524 Table 3. Significant BLAST hits for TRINITY “genes” versus zebrafish peptides, rRNA silva,  
525 (non-zebrafish) teleost fish transcriptomes, and the zebrafish genome. BLAST searches were  
526 done in stepwise fashion: all TRINITY genes were queried against zebrafish peptides but only  
527 genes without zebrafish peptide hits were queried against rRNA silva, and so on until all of the  
528 databases were queried.  
529

	<i>Cyprinus carpio</i>	<i>Cyprinella lutrensis</i>	<i>Platygobio gracilis</i>
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
<b>Total Contigs</b>	<b>309,921</b>	<b>255,863</b>	<b>180,130</b>

530

531

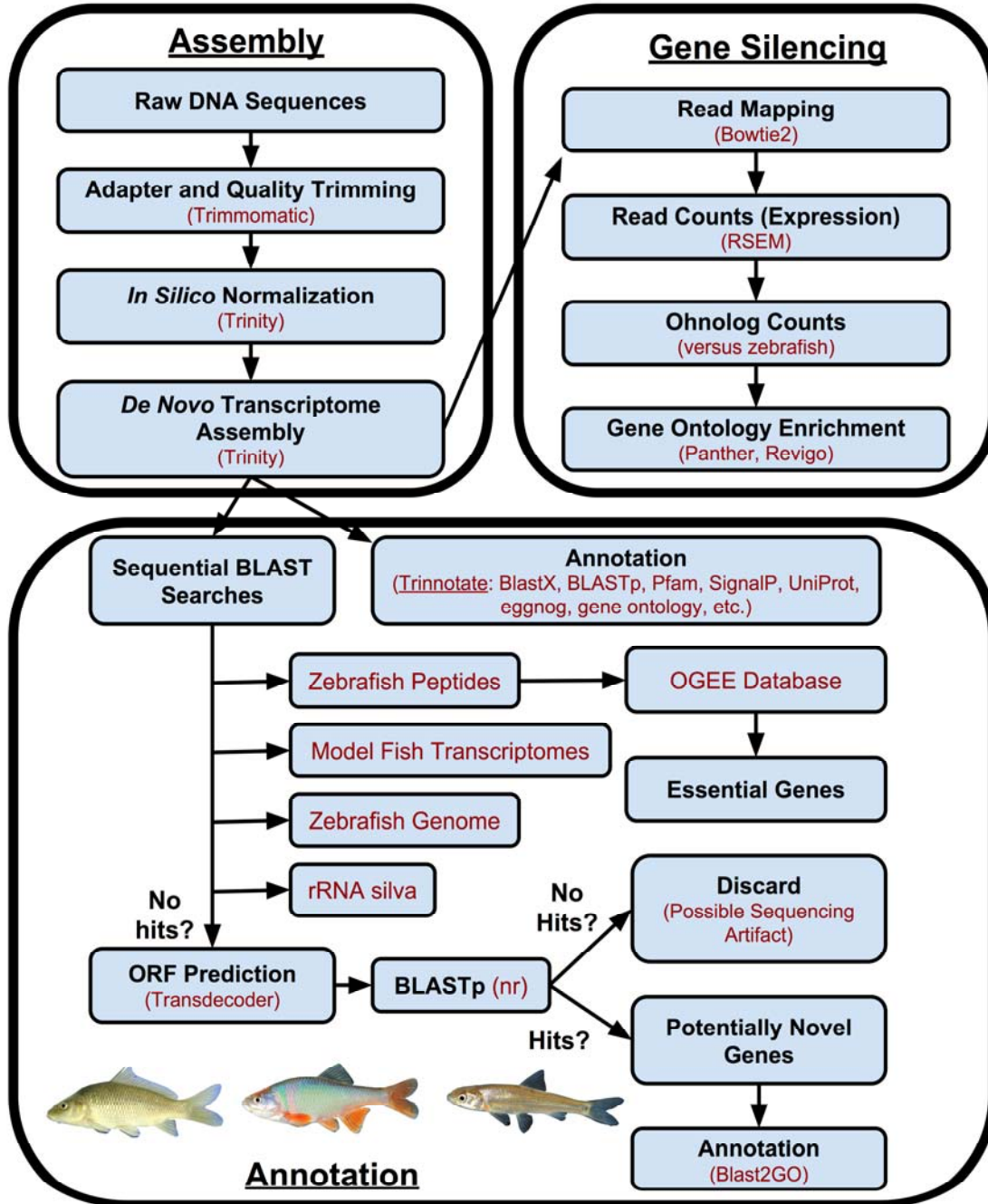
532 Table 4. Summary of open reading frames (ORFs) identified in TRINITY contigs with no  
533 significant BLAST hits against databases listed in Table 3 (“No significant BLAST hits”).  
534 Roughly half of these predicted ORFs had significant protein level BLAST hits in the nr  
535 database. Some of the ORFs lacking similar proteins in the nr database may represent novel  
536 genes or genes with divergent sequences and function, while many are likely spurious results  
537 from the sequencing and assembly process or are from unidentified microbiota.  
538

	<i>Cyprinus carpio</i>	<i>Cyprinella lutrensis</i>	<i>Platygobio gracilis</i>
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

539

540

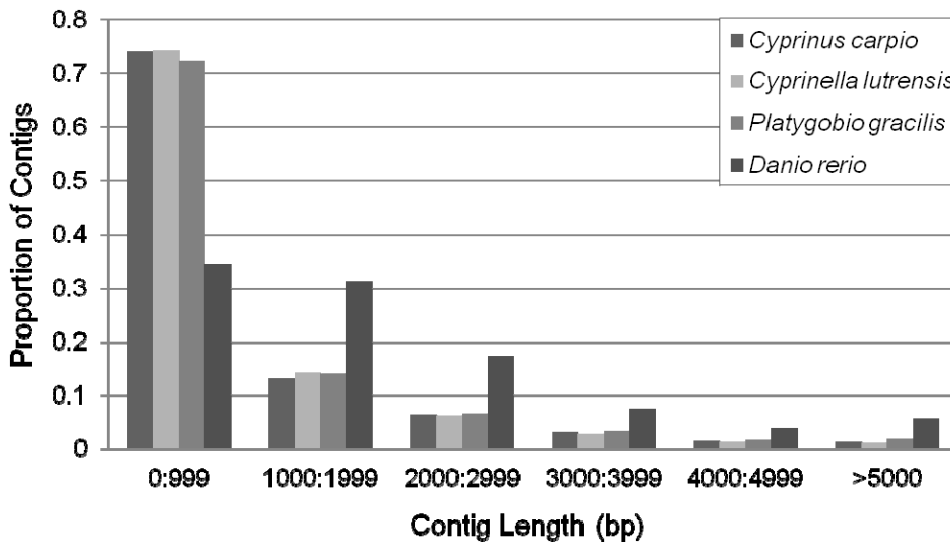
541 Figure 1. Flow diagram illustrating our bioinformatic pipeline. Analyses consist of three main  
542 steps: assembly, annotation, and analysis of gene silencing patterns. Databases queried and  
543 software packages used are in red font.  
544



545

546

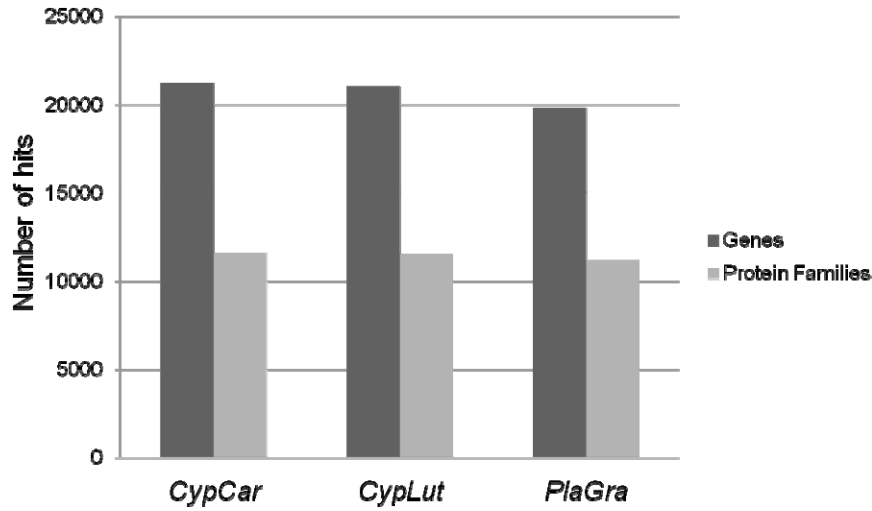
547 Figure 2. Contig length histogram of three cyprinids in this study and zebrafish, *Danio rerio*. By  
548 leveraging high throughput sequencing and bioinformatic filtering, we were able to generate high  
549 quality transcriptomes at a fraction of the cost and research effort used for zebrafish. As  
550 expected, *de novo* TRINITY assemblies resulted in proportionally fewer contigs longer than 1000  
551 bp, as compared to those of a well-assembled transcriptome, zebrafish (*Danio rerio*). However,  
552 note that we only used canonical transcripts for zebrafish and not the shorter isoforms, which  
553 skews the distribution toward longer transcripts for that species.  
554



555

556

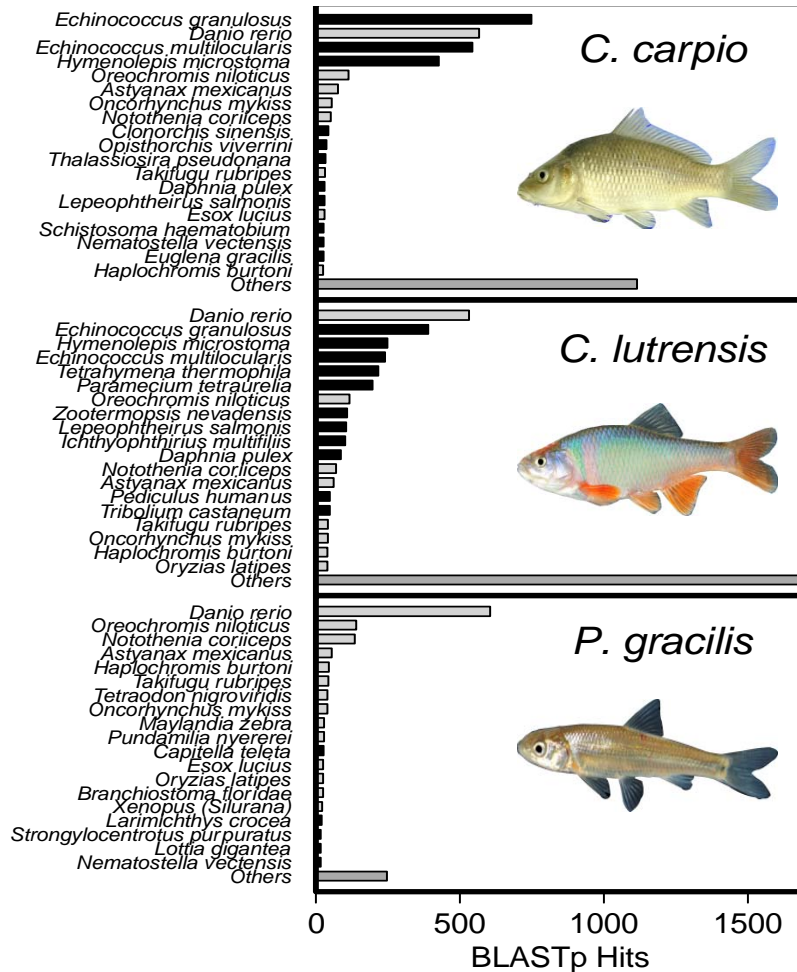
557 Figure 3: Unique genes and protein families from BLASTx searches (E-value threshold =  
558 0.0001) against zebrafish (*Danio rerio*) peptide sequences. Similar numbers of zebrafish genes  
559 (~20,000) and protein families (~11,000) were identified across the three species, suggesting  
560 comparable assembly and annotation efficiency across these species.  
561



562

563

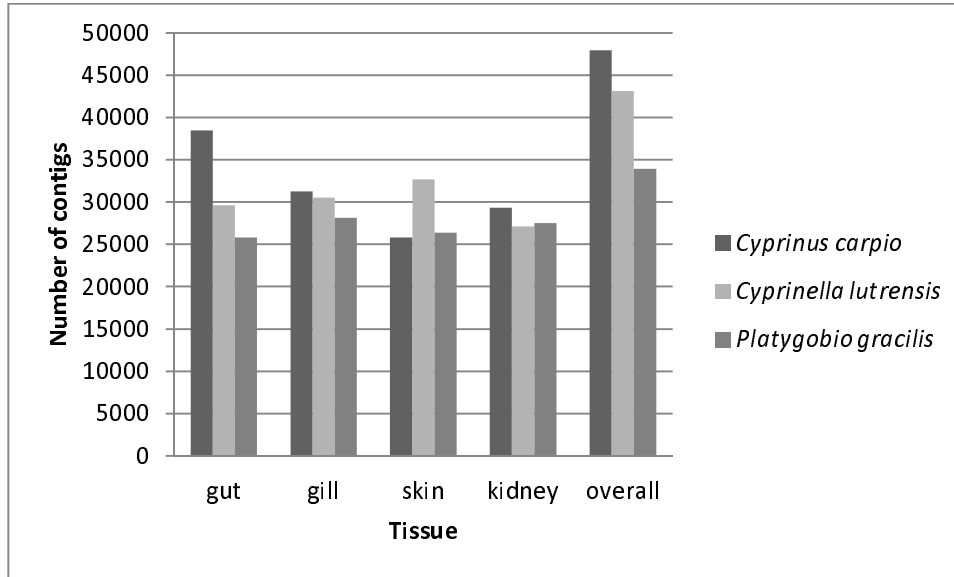
564 Figure 4. Top-species BLASTp hits for predicted open reading frame (ORF) peptide sequences  
 565 queried against the nr database. Query sequences only included ORFs from contigs that lacked  
 566 significant BLAST hits (see Tables 3 and 4). Grey bars represent fish or other chordates, while  
 567 black bars represent non-chordate taxa. In *Cyprinus carpio* and *Cyprinella lutrensis*, many of  
 568 these ORFs are likely from the microbiome (black bars) as they share significant similarity to  
 569 cyclophyllid tapeworms (e.g., *Echinococcus*, *Hymenolepis*) and protozoans (e.g., *Tetrahymena*,  
 570 *Paramecium*). Conversely, in *Platygobio gracilis* most of the ORFs appear to be endogenous  
 571 genes with high similarity to zebrafish and other teleost fish.  
 572



573



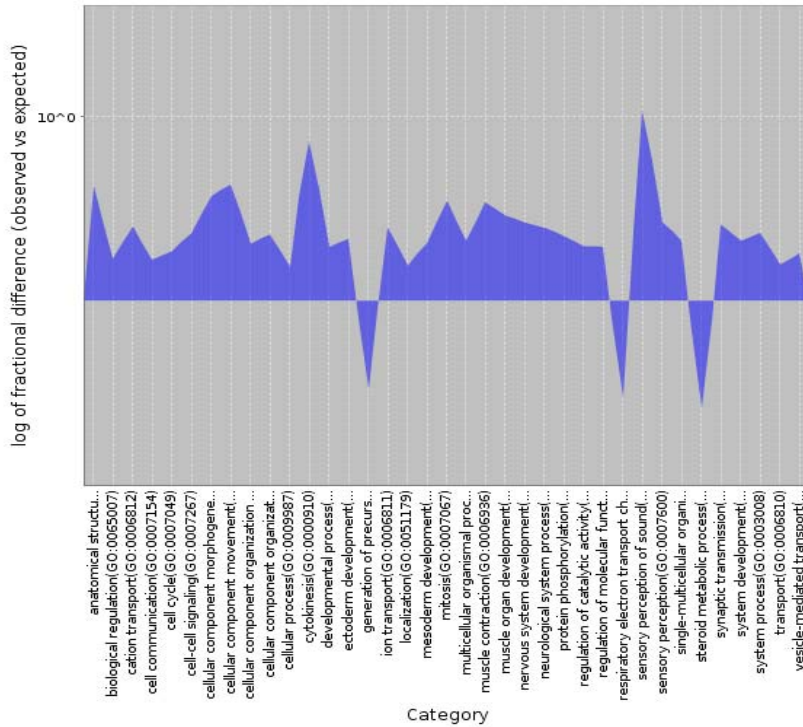
574 Figure 5. Number of TRINITY genes expressed in each of four tissue types, as well as all tissues  
575 pooled. Contigs only include those with significant BLASTx hits versus zebrafish peptides. Due  
576 to the carp-specific genome duplication event (Cc4R), *Cyprinus carpio* generally expresses more  
577 genes in a given tissue type than the other species, except in skin.  
578



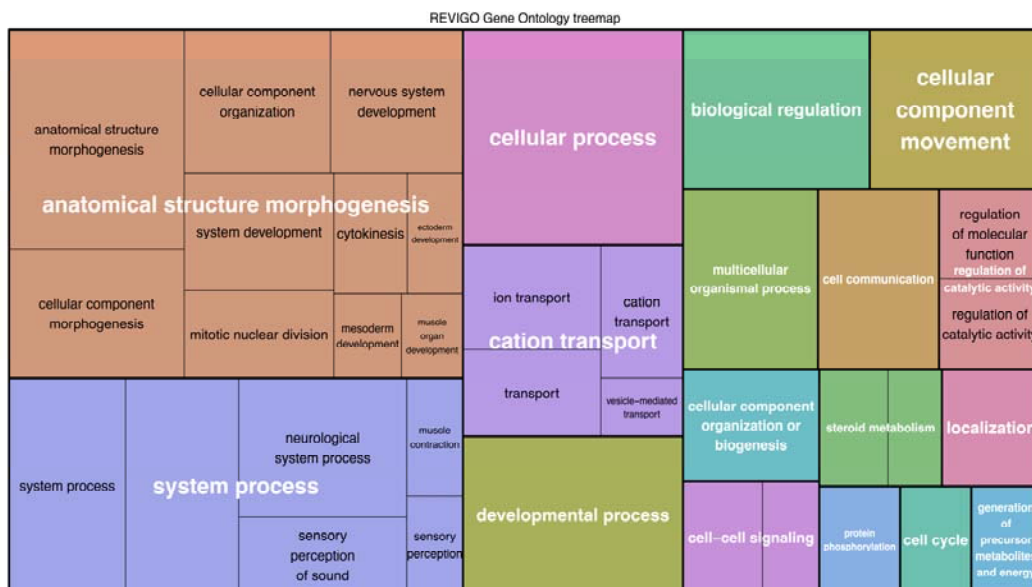
579

580

581 Figure 6. Top panel: Gene-ontology terms that are over- or under-represented (y-axis) in the list  
 582 of genes retained as duplicates in the common carp transcriptome as compared to all expressed  
 583 genes in common carp. Bottom panel: Summary of groups of biological processes  
 584 overrepresented in the retained-duplicates in common carp. Box size is proportional to the  
 585 number of genes with particular gene ontology terms, which may suggest a dosage effect in  
 586 common carp.

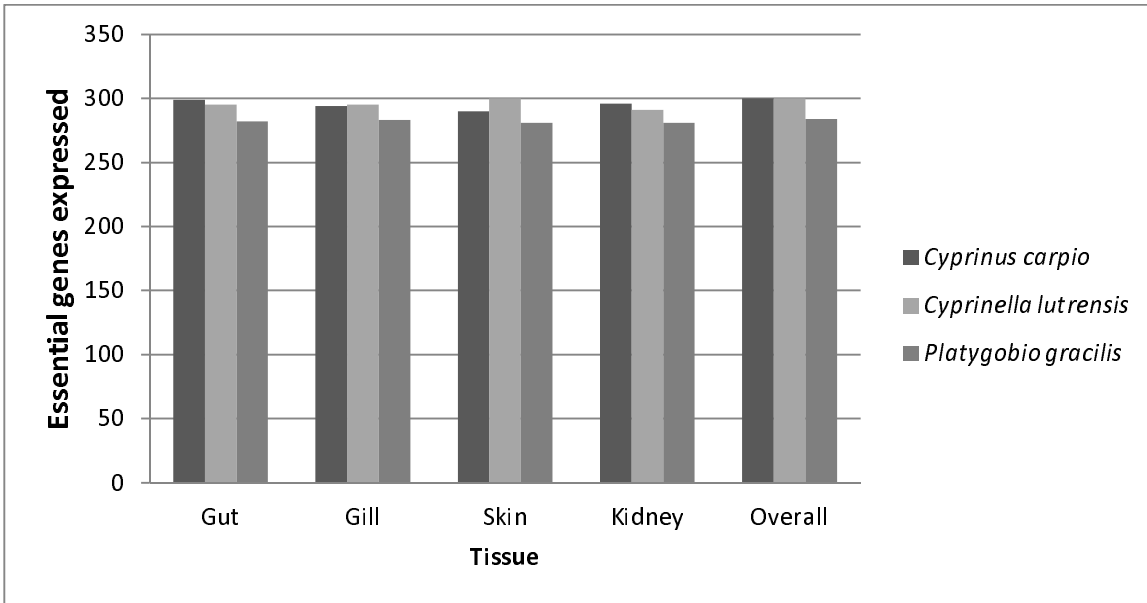


587



588

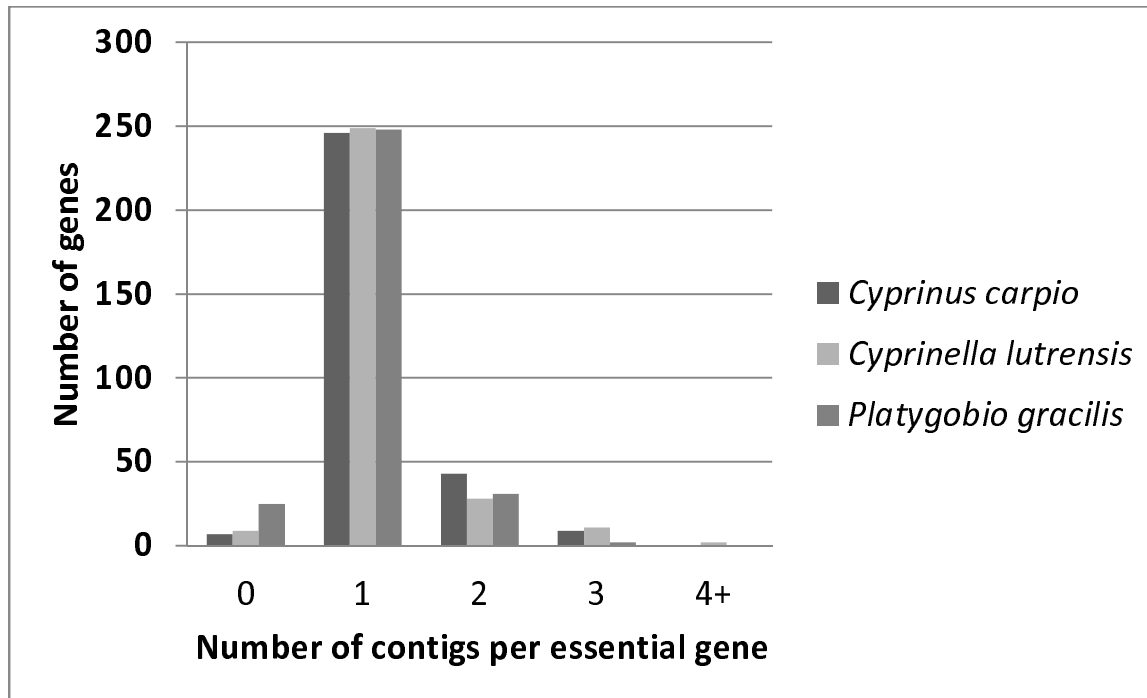
589 Figure 7. Expression of essential developmental genes by tissue type in three cyprinid fishes  
590 compared to 305 essential genes expressed across all tissues in zebrafish (*Danio rerio*). Essential  
591 genes were nearly ubiquitously expressed in all tissues.  
592



593

594

595 Figure 8. Number of loci (TRINITY genes) expressed per zebrafish essential gene. Only slightly  
596 more (e.g., n=2) were expressed per essential gene in the recently duplicated genome of  
597 *Cyprinus carpio*. Most duplicated essential genes in *C. carpio* are not transcribed and have  
598 either been lost evolutionarily, e.g., pseudogenes, or are expressed in other developmental stages  
599 or tissues.  
600



601