



The genome of *Anoplarchus purpurescens* (Stichaeidae) reflects its carnivorous diet

Ninh Le^{1,4} · Joseph Heras² · Michelle J. Herrera³ · Donovan P. German³ · Lisa T. Crummett¹

Received: 29 March 2023 / Accepted: 28 August 2023
© The Author(s) 2023

Abstract

Digestion is driven by digestive enzymes and digestive enzyme gene copy number can provide insights on the genomic underpinnings of dietary specialization. The “Adaptive Modulation Hypothesis” (AMH) proposes that digestive enzyme activity, which increases with increased gene copy number, should correlate with substrate quantity in the diet. To test the AMH and reveal some of the genetics of herbivory vs carnivory, we sequenced, assembled, and annotated the genome of *Anoplarchus purpurescens*, a carnivorous prickleback fish in the family Stichaeidae, and compared the gene copy number for key digestive enzymes to that of *Cebidichthys violaceus*, a herbivorous fish from the same family. A highly contiguous genome assembly of high quality (N50 = 10.6 Mb) was produced for *A. purpurescens*, using combined long-read and short-read technology, with an estimated 33,842 protein-coding genes. The digestive enzymes that we examined include pancreatic α -amylase, carboxyl ester lipase, alanyl aminopeptidase, trypsin, and chymotrypsin. *Anoplarchus purpurescens* had fewer copies of pancreatic α -amylase (carbohydrate digestion) than *C. violaceus* (1 vs. 3 copies). Moreover, *A. purpurescens* had one fewer copy of carboxyl ester lipase (plant lipid digestion) than *C. violaceus* (4 vs. 5). We observed an expansion in copy number for several protein digestion genes in *A. purpurescens* compared to *C. violaceus*, including trypsin (5 vs. 3) and total aminopeptidases (6 vs. 5). Collectively, these genomic differences coincide with measured digestive enzyme activities (phenotypes) in the two species and they support the AMH. Moreover, this genomic resource is now available to better understand fish biology and dietary specialization.

Keywords Genomics · Gene copy number · Nutrition · Digestive enzyme · Long read · Short read · Feeding ecology

Communicated by Joan Cerdá.

Donovan P. German and Lisa T. Crummett authors contributed equally to this article.

✉ Donovan P. German
dgerman@uci.edu

✉ Lisa T. Crummett
lcrummett@soka.edu

¹ Life Sciences Concentration, Soka University of America, Aliso Viejo, CA 92656, USA

² Department of Biology, California State University, San Bernardino, CA 92407, USA

³ Department of Ecology and Evolutionary Biology, University of California, Irvine, CA 92697, USA

⁴ Department of Epigenetics and Molecular Carcinogenesis, The University of Texas M.D. Anderson Cancer Center, Houston, TX 77030, USA

Introduction

Comparing the genomes of closely related animals that have evolved different specializations offers opportunities to understand how differences in their physiology are attributed to differences on the genomic level. Whether it is epistatic interactions of suites of genes affecting phenotypes (e.g., Chen et al. 2022), changes in gene copy number impacting gene expression, which in turn affects enzyme activity levels (Axelsson et al. 2013; Heras et al. 2020; Perry et al. 2007), or molecular convergence of genes underlying unique phenotypic traits (Protas et al. 2005), genomic evidence of adaptation to environmental variables abound in the comparative genomics literature (Fan et al. 2020; Lamichhaney et al. 2015; Taylor et al. 2021; Yuan et al. 2018). As the nutrient supply organ that interfaces directly with items consumed from the environment, the digestive tract has become an attractive system in which to examine how animals can specialize on specific resources (Brun et al. 2020; Karasov

and Douglas 2013). But how does dietary specialization impact an animal's genome? With whole genome sequencing becoming more affordable and genomic analyses becoming more accessible, we are now better equipped to explore genomic adaptations to dietary specializations such as carnivory or herbivory (e.g., Axelsson et al. 2013; Heras et al. 2020; Wang et al. 2015).

Depending on the approach, one can start at the genomic level, find areas of a genome that may be under selection, and then scale up to discern what traits are impacted by the molecular changes identified (Axelsson et al. 2013; Yuan et al. 2018). Conversely, one can take a well-known system with well-studied physiological and biochemical processes and dig down to the genetic level to identify the underpinnings of specific phenotypes. This latter approach was done successfully in a recent examination of the herbivorous fish, *Cebidichthys violaceus* (Heras et al. 2020). Years of ecological, physiological, and biochemical data (Fris and Horn 1993; Gawlicka and Horn 2006; German et al. 2016; German et al. 2004; German et al. 2015; Horn et al. 1986; Kim et al. 2014) were used to make a priori predictions about how this herbivore is able to thrive on an algal diet. A detailed genomic analysis of *C. violaceus* largely confirmed these predictions and provided the genetic underpinnings of known herbivorous phenotypes (Heras et al. 2020). For instance, elevated amylolytic and lipolytic activities in the guts of the fish were attributed to increased gene copy number coding for those proteins.

In the context of digestive specialization, it is important to note that digestion is a chemical process, and the agents of that process are digestive enzymes. Thus, digestive enzymes play a crucial role in digestion and have been studied extensively in the context of diet (e.g., Brun et al. 2020; Vonk and Western 1984; Karasov and Douglas 2013; German et al. 2015, 2016; Schondube et al. 2001; Skea et al. 2005, 2007). The Adaptive Modulation Hypothesis (AMH) posits that digestive enzyme activity should correlate with substrate quantity in an animal's diet because protein synthesis would be wasted on enzymes targeting less abundant substrates (Karasov 1992; Karasov and Martínez del Rio 2007). To target abundant substrates, digestive enzyme activity can increase with increased expression level of the enzyme genes (Choi and Yamazaki 1994; Cockell et al. 1989; Gawlicka and Horn 2006; German et al. 2016; Howard et al. 1989; Ma et al. 2004; Wiebe et al. 2007), increased gene copy number of the enzyme genes, which increases expression (Axelsson et al. 2013; German et al. 2016; Gout et al. 2010; Qian and Zhang 2014; Springer et al. 2010), or expression of additional enzyme isoforms or gene variants (German et al. 2016). For instance, humans from agrarian backgrounds and domestic dogs (which have been consuming grains for nearly as long as humans) have expanded gene copy number of amylase genes, which correlates with increased

amylase expression and higher enzymatic activity against starch (Axelsson et al. 2013; Perry et al. 2007). Similarly, in an examination of amylase genes and enzyme activity in stichaeid fishes, the herbivorous *C. violaceus* was found to have an extra copy of the amylase gene, elevated expression of amylase genes, and elevated amylase activity in comparison to carnivorous stichaeids (German et al. 2016).

In this study, we set out to test predictions anchored in the AMH for how gene copy numbers for specific digestive enzymes will differ between a carnivorous fish, *Anoplarchus purpureus*, and a sympatric herbivorous fish, *C. violaceus*, both in the family Stichaeidae (German and Horn 2006; Kim et al. 2014; Fig. 1). *Cebidichthys violaceus* and *A. purpureus* represent separate intertidal invasions within the family Stichaeidae, where herbivory evolved in the former, but not the latter (Heras et al. 2020; Herrera et al. 2022; Fig. 1). The geographic range of *A. purpureus* overlaps significantly with that of *C. violaceus*, with *A. purpureus* extending

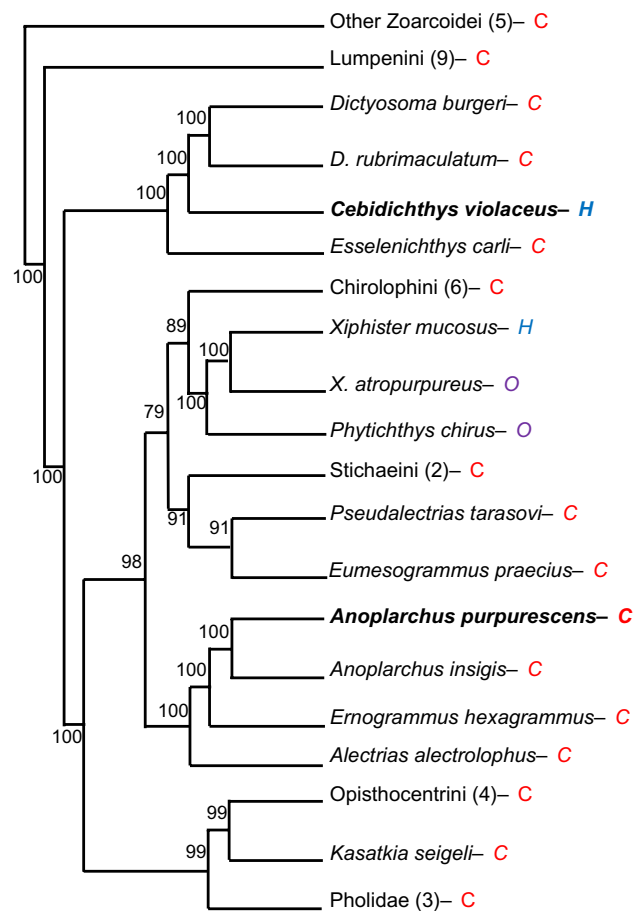


Fig. 1 Summarized phylogenetic relationships of the family Stichaeidae based on 2100 bp of *cytb*, *16 s*, and *tomo4c4* genes (Kim et al. 2014). Bayesian posterior probabilities indicated on nodes are from Kim et al. (2014). Species used in this study bolded. H=herbivory, O=omnivory, C=carnivory. Numbers in parentheses show number of taxa evaluated at that branch

from southern California to the Aleutian Islands to the north (Stoddard 1985). Based on the AMH, we predicted that, in comparison with the herbivorous *C. violaceus*, the carnivorous *A. purpurescens* will exhibit a reduction in gene copy number for enzymes that digest carbohydrates (carbohydrases), an expansion in gene copy number for enzymes associated with animal lipid digestion (triacylglycerol lipase) and/or a reduction in gene copy number for enzymes associated with plant lipid digestion (carboxyl ester lipase). We also predicted an expansion of gene copy number in *A. purpurescens* for one or more enzymes associated with protein digestion (proteases) that would correlate with increased dietary protein concentration and concomitant proteolytic activity in *A. purpurescens* relative to *C. violaceus* (German et al. 2004, 2015). This comparison will help uncover genomic differences associated with herbivory and carnivory in the family Stichaeidae, and vertebrates more broadly, and may have broad implications for aquaculture and fisheries management. We used long- and short-read technology to sequence, assemble, and annotate the genome of *A. purpurescens*, yielding even better genome assembly metrics (e.g., N50 of 10.6 Mb) than the *C. violaceus* genome (N50 of 6.7 Mb; Heras et al. 2020). Currently, there is little known about genome size or chromosome number for *A. purpurescens*, which we assumed would be similar to *C. violaceus*.

Materials and methods

Collection and preparation

One individual of *A. purpurescens* (85 mm standard length) was collected in January 2021 from San Simeon, California (35.6525°N, 121.2417°W). The individual was euthanized in tricaine methanesulfonate (Syndel, Ferndale, Washington, USA; 1 g l⁻¹), dissected for internal organ removal, decapitated and preserved in liquid nitrogen. Fish handling from capture to euthanization was conducted under approved protocol 2021–012 of the Institutional Animal Care and Use Committee (IACUC) at the University of California, Irvine. Genomic DNA was extracted from 250 mg of skin and muscle tissue using a Qiagen Blood & Cell Culture DNA Midi Kit (Qiagen, Redwood City, California, USA), following the manufacturer's instructions. Following extraction, the DNA samples were sheared and separated into high molecular weight DNA fragments by a pulse field electrophoresis. We used Pacific Biosciences (PacBio) and Illumina platforms for sequencing. For PacBio sequencing, genomic DNA was size-selected, with a 15 kb size cut-off, using a BluePippin Size Selection System. A PacBio Sequel II was used to sequence one Single Molecule Real Time (SMRT) cell, which can produce up to 100 Gb of sequence data, and in our case, 74 Gb. Additionally, from the same gDNA extraction,

a multiplex gDNA-Seq Illumina sequencing library was prepared from size-selected fragments ranging from 500 to 700 bp, and sequenced on two lanes on an Illumina HiSeq 4000, which resulted in short reads (100 bp paired-end). All genomic sequencing was completed at the University of California, Irvine (UCI) Genomics Research and Technology Hub (GRTH) following Heras et al. (2020).

Assembly of Illumina short reads and PacBio long reads

Computational and bioinformatics analyses were all conducted on the High-Performance Computing (HPC) Cluster at UCI. Paired-end Illumina sequence data was examined for quality control using FastQC version 0.11.9 (Andrews 2010). Trimming and adapter removal was performed by BBDuk version 38.87 with the following parameters: ref = adapters, phix ktrim = r mink = 11 hdist = 1 qtrim = rl trimq = 10 (Bushnell 2014; Supplemental Fig. S1). Genome size and heterozygosity percentage were estimated from unassembled Illumina reads with GenomeScope1.0 (Maçais and Kingsford 2011; Vurture et al. 2017) and ploidy was confirmed with Smudgeplot version 0.1.3 (Ranallo-Benavidez et al. 2020; Supplemental Fig. S1). The k-mer size was set to 21 for these programs. Illumina sequencing data were assembled with Platanus version 1.2.4 (Kajitani et al. 2014) with the following parameters: platanus assemble -t 24 -m 240 -k 17 -s 2; where initial k-mer size was set at 17 and step size was set to 2 (Supplemental Fig. S1, S2).

We conducted a PacBio SMRT Sequencing read-only assembly using Flye version 2.9-b1787 (Kolmogorov et al. 2019) with “scaffolding mode” on and “keep-haplotype” mode off (Supplemental Fig. S1).

Combined hybrid assembly

To produce a hybrid assembly, we input the contigs generated by Platanus and the raw long-read sequencing data into DBG2OLC (Ye et al. 2016; Supplemental Fig. S1) with the following parameters: k 17 KmerCovTh 2 MinOverlap 150 AdaptiveTh 0.02 LD1 0 and RemoveChimera 1. To reduce the size of the data set, DBG2OLC “compressed” each long read into an ordered set of Platanus contigs that mapped to it (above certain thresholds set by AdaptiveTh and KmerCovTh). The cleaned compressed long reads were then used to construct a best overlap graph, from which backbone sequences were generated. Finally, a consensus module, Pbdagcon version 0.3 (Chin et al. 2013), was utilized with default parameters to align reads to each backbone to produce the polished final assembly (Supplemental Fig. S1).

We used the long-read assembly from Flye as the reference assembly and the hybrid assembly from DBG2OLC as the query assembly for Quickmerge v.1.0 (Chakraborty

et al. 2016), which is both a meta-assembler and assembly gap filler program originally developed for long read assemblies (Supplemental Fig. S1). Quickmerge clusters contigs between the query and the reference assemblies based on their high confidence overlap (HCO), a metric score that quantifies how well a contig overlaps with another. Different parameters were tested until the most contiguous assembly was obtained with the following parameters: -hco 5, -c 1.5, -l 1,000,000, -ml 5000. We conducted a second round of Quickmerge using the output assembly from the first round of Quickmerge as the reference assembly, and the Flye assembly was used as the query assembly, with the following parameters: -hco 5, -c 1.5, -l 3,900,000, -ml 5000.

Hybrid assembly polishing, purging, and repeat masking

We polished the *A. purpurescens* hybrid genome assembly (Supplemental Fig. S1) through two rounds of Arrow, executed through gcpp version 2.0.2 (<https://github.com/PacificBiosciences/gcpp>), where long reads were aligned to our quickmerge assembly by Pbmm2 version 1.4.0 (<https://github.com/PacificBiosciences/pbmm2>). We used purge_dups (Supplemental Fig. S1), which utilizes sequence similarity and read depth to resolve haplotigs and homologous chromosome overlaps (Guan et al. 2020). To mask repetitive elements, the final genome assembly was processed through Repeatmasker version 4.1.2 (Smit et al. 2015) with the parameters -e ncbi -pa 8 -species teleostei -s -xsmall, using the complete Dfam library (Supplemental Fig. S1).

Hybrid assembly quality assessment

The genome size and the N50 value for the hybrid assemblies (Table 1) were computed using a Perl script (Bradnam et al. 2013), while the heterozygosity rates were estimated with a pipeline consisting of Burrows–Wheeler aligners version 0.7.8 (Li and Durbin 2009), SAMtools version 1.10 (Li et al. 2009), BCFtools version 1.14 (Li et al. 2009), and a Python script provided by Dr. John Bracht (Asalone et al.

2020; Supplemental Fig. S1). Heterozygosity estimates of assemblies were corroborated with analyses of spectra copy number plots generated by Merqury (Rhie et al. 2020) and Meryl version 1.3 (Walenz 2020), the latter providing 21-mer count histograms for the former (Supplemental Fig. S1, S2). The completeness of our *A. purpurescens* genome was evaluated using BUSCO (Benchmarking Universal Single Copy Orthologs) version 5.3.0, using the Vertebrata and Actinopterygii gene sets (Simão et al. 2015).

Genome synteny analyses

After concatenating contigs less than 1 MB into one contig per genome, we ordered the genome of *A. purpurescens* against that of *C. violaceus* using SyMAP v5.2.0 (Synteny Mapping and Analysis Program; Soderlund et al. 2011). CIRCOS and synteny blocks between the two species were computed with SyMAP v5.2.0 at default parameters.

Structural gene annotation

We used the BRAKER2 pipeline (Altschul et al. 1990; Brůna et al. 2021; Camacho et al. 2009; Hoff et al. 2016, 2019; Stanke et al. 2006, 2008) to perform structural gene annotations on the genomes of both *A. purpurescens* and *C. violaceus*, the latter sequenced and assembled by Heras et al. (2020). The Vertebrata section of the OrthoDB database (Kriventseva et al. 2019) was used to generate protein hints by BRAKER2, and was processed by the ProHint pipeline (Brůna et al. 2020; Buchfink et al. 2015; Gotoh et al. 2014; Iwata and Gotoh 2012; Lomsadze et al. 2005). RNA sequence data for *A. purpurescens* and *C. violaceus* were taken from Herrera et al. (2022) and Heras et al. (2020) respectively. The RNA sequence data was trimmed by TrimGalore version 0.6.6 (<https://github.com/FelixKrueger/TrimGalore>). We aligned the RNA-seq data against their respective genomes using HISAT2 version 2.1.0 (Kim et al. 2019), which provided spliced alignments for BRAKER (Barnett et al. 2011; Li et al. 2009). The alignments were utilized by GeneMark-ET to generate a training gene set

Table 1 Genome assembly programs, types of reads used, and statistics used to assemble the genome of *Anoplarchus purpurescens*

Assembly	Programs used	Reads used	Assembly size (bp)	Scaffold/Contig number	N50 value (bp)
1	Platanus	Illumina	667,076,046	2,599,093	1,207
2	DBG2OLC, Blasr, pbdagcon	Illumina PacBio	598,594,787	976	1,980,038
3	Flye	PacBio	582,412,549	1,331	3,956,814
4	Quickmerge (round 1)	Illumina PacBio	593,045,980	732	6,841,580
5	Quickmerge (round 2)	Illumina PacBio	585,237,394	1,210	10,452,245
6	Pbmm2, Arrow	Illumina PacBio	586,262,929	1,210	10,464,639
7	Purge_dup	Illumina PacBio	567,389,083	489	10,617,371

for AUGUSTUS (version 3.5; Lomsadze et al. 2014). We used the RNA-seq data of the liver, mid-intestine and pyloric caeca tissues of wild-type *A. purpurascens* (Herrera et al. 2022), and that of the spleen, mid-intestine, gonads, pyloric caeca, heart, brain, liver, proximal intestine, and gill tissues of *C. violaceus* (Heras et al. 2020). The unassembled RNA-seq data of both species were included in all replicates, if replicates were available, to retain and increase coverage information of each splice site, optimizing GeneMark-ET performance (Hoff et al. 2019). The final genome annotation was uploaded as project PRJNA950117 at NCBI (<https://www.ncbi.nlm.nih.gov/bioproject/>).

Functional gene annotation

The transcripts predicted by AUGUSTUS, from the BRAKER2 pipeline, were functionally annotated by BLAST2GO from OmicsBox version 2.2.4 (Götz et al. 2008). Within BLAST2GO, we used blastx-fast, with default parameters, to search the Vertebrata subset of the non-redundant protein sequence database, version 5 (Götz et al. 2008). Gene identities were derived from BLAST2GO consensus descriptions.

Creation of gene synteny maps among fish species

We compared the genomic regions surrounding specific digestive enzymes, among *A. purpurascens*, *C. violaceus*, and several non-stichaeid fish, including *Danio rerio*, *Oryzias latipes*, *Gasterosteus aculeatus*, and *Oreochromis niloticus*. Multiple sequence alignments of specific genomic regions, among six fish species, were performed with MUSCLE version 3.8.425 (Madeira et al. 2022), and phylogenetic trees were made with 1,000 bootstrap replicates using PhyML 3.0, where AIC helped to determine the best model (Guindon et al. 2010; Lefort et al. 2017).

Results

Quality and coverage of sequence data from Illumina and PacBio platforms

From one PacBio SMRT cell sequencing, we generated 74 Gb long reads, with approximately 70X coverage. Illumina generated 36 Gb of 100 bp paired-end reads, approximately 24X coverage. The quality of Illumina reads was excellent, with reads having an average QC score of 34–36 in all positions, which translates to a base call accuracy of 99.97%.

Estimated characteristics of the *A. purpurascens* genome

At a k-mer size of 21, GenomeScope estimated, with high confidence, the haploid genome size of *A. purpurascens* to be 538,951,370 bp, which is similar to other fish genomes (Heras et al. 2020). There were an estimated 33,842 protein-coding genes in the genome. GenomeScope estimated the heterozygosity percentage of the *A. purpurascens* genome to be 0.879% (Supplemental Fig. S2). The diploid nature of the *A. purpurascens* genome was confirmed by Smudgeplot (Ranallo-Benavidez et al. 2020; Supplemental Fig. S3). RepeatMasker identified 32.02% of the genome assembly as repetitive sequences, including 4.48% as retroelements, and 8.75% as DNA transposons (Supplemental Table S1).

Quality of the final assembly

Utilizing both the Illumina contigs and long reads from PacBio SMRT sequencing, the hybrid assembly produced a more contiguous assembly than using the short reads, alone (N50 = 1.99 Mb vs. N50 = 1,207 bp; Table 1). The PacBio long-read assembly, using Flye, yielded an N50 value that was more than double the N50 value from the original hybrid assembly (N50 = 3.96 Mb vs. 1.99 Mb; Table 1). Two rounds of merging the hybrid assembly with the long-read assembly produced a highly contiguous assembly with the highest N50 value (10.46 Mb) and 1210 scaffolds (Table 1). Reducing the number of scaffolds and contigs, via purging, increased the N50 value further to 10.62 Mb and decreased the scaffold count to 489 (Table 1). Purging also brought the hybrid genome assembly size (567 Mb) closer to the haploid genome size estimated by GenomeScope, 538 Mb (Table 1). BUSCO showed that the final hybrid genome assembly is 97% complete with 33,862 protein-coding genes.

Digestive enzyme gene copy number in *A. purpurascens* vs. *C. violaceus*

The genomic comparisons of *A. purpurascens* and *C. violaceus* revealed highly syntenic genomes (Supplemental Fig. S4) with differences in gene copy number among key digestive enzyme genes. *A. purpurascens* has a reduced number of pancreatic α -amylase (*amy2*) genes compared to *C. violaceus* (1 vs. 3 copies; Table 2; Fig. 2); The AMY2 enzyme digests dietary starches (carbohydrates). *A. purpurascens* had a reduced number of carboxyl ester lipase (*cel*) genes compared to *C. violaceus* (4 vs. 5 copies; Table 2; Fig. 3 and Supplemental Fig. S5). The CEL enzyme efficiently digests plant lipids. Both species had one copy each of triacylglycerol lipase (*lipc*) and hepatic triacylglycerol lipase (*lipf*), which efficiently digests triacylglycerol, the main constituent of body fat in animals (Sahaka et al. 2020;

Table 2 Gene copy numbers of pancreatic α -amylase, carboxyl ester lipase, chymotrypsinogen, trypsinogen, alanyl aminopeptidases, gastric and hepatic triacylglycerol lipase in *Anoplarchus purpureus* and *Cebidichthys violaceus*

Gene full name	Gene acronym	Gene copy number	
		<i>A. purpureus</i>	<i>C. violaceus</i>
Pancreatic α -amylase	<i>amy2</i>	1	3
Carboxyl ester lipase (total)		4	5
Carboxyl ester lipase 1	<i>cel 1</i>	2	3
Carboxyl ester lipase 2	<i>cel 2</i>	1	1
Carboxyl ester lipase like	<i>cel-like</i>	1	1
Chymotrypsinogen (total)		4	4
Chymotrypsinogen B1	<i>ctrb 1</i>	1	2
Chymotrypsinogen B2	<i>ctrb 2</i>	1	1
Chymotrypsinogen-like	<i>ctrl</i>	2	1
Trypsinogen (total)		5	3
Trypsinogen 1	<i>prss 1</i>	4	2
Trypsinogen 2	<i>prss 2</i>	1	1
Aminopeptidase (total)		6	5
Aminopeptidase A	<i>anpep a</i>	1	1
Aminopeptidase B	<i>anpep b</i>	1	1
Aminopeptidase N	<i>anpep N</i>	1	1
Aminopeptidase Ey	<i>anpep Ey</i>	1	1
Aminopeptidase Ey-like	<i>anpep Ey-like</i>	2	1
Gastric triacylglycerol lipase	<i>lipc</i>	1	1
Hepatic triacylglycerol lipase	<i>lipf</i>	1	1

Table 2). We observed an expansion in gene copy number for two proteolytic enzymes in *A. purpureus* (Table 2). *A. purpureus* had more copies of aminopeptidase Ey-like (2 vs. 1 copies; Table 2; Figs. 4 and 5) than *C. violaceus*, whereas *A. purpureus* had more copies of trypsinogen (5 vs. 3 copies; Table 2; Fig. 6; Supplemental Fig. S8). While the total number of chymotrypsin genes is the same between *A. purpureus* and *C. violaceus*, the species differ in the number of gene copies of chymotrypsin B1 (1 vs. 2 copies; Table 2; Fig. 7a) and chymotrypsin-like genes (2 vs. 1 copies; Table 2; Fig. 7b).

Discussion

We assembled the genome for the carnivorous prickleback fish, *Anoplarchus purpureus* with the intent of comparing it to that of the herbivorous *C. violaceus* so that we could observe potential dietary adaptations on the level of digestive enzyme genes. Matching with gene expression (Gawlicka and Horn 2006; German et al. 2016; Heras et al. 2020; Herrera et al. 2022; Kim et al. 2014) and digestive enzyme

activity data (German et al. 2004, 2015) gathered in many previous studies of these fish species, we found support in the form of gene copy number differences in digestive enzyme genes that match with dietary-related phenotypes in these species. Hence, we provide clear support for the AMH (Karasov 1992; Karasov and Martínez del Río 2007) on the genomic level to support previously measured phenotypes.

The N50 value of the *A. purpureus* genome, at 10.6 Mb, makes it one of the most contiguous fish genomes in the literature (Lu and Luo 2020; Lu et al. 2020) (and see Supplemental Table S1 in Heras et al. 2020). The BUSCO score of 97% also illustrates a nearly complete assembly and annotation. The size of the *A. purpureus* genome (539 Mb) shows that it is similar in size to other fish genomes spanning a deep diversity of fisheries and aquaculture species (Lu and Luo 2020), as well as those of *C. violaceus* and *Siniperca kneri* (Heras et al. 2020; Lu and Luo 2020; Lu et al. 2020). The success of this genome assembly further supports that merging multiple assemblies with Quickmerge (Chakraborty et al. 2016) can be done for species that are not model organisms (e.g., Baldwin-Brown et al. 2018; Heras et al. 2020). Moreover, the *A. purpureus* genome is highly syntenic with that of *C. violaceus* (Supplemental Fig. S4), allowing us to make comparisons between these two closely related fish species with different diets (German et al. 2015; Herrera et al. 2022).

Based on the AMH (Karasov 1992; Karasov and Martínez del Río 2007), we predicted that *A. purpureus* would have fewer amylase (*amy2*) gene copies than *C. violaceus* as carnivorous fishes consume much less starch in their diet compared to herbivorous fishes and increased gene copy number often results in increased gene expression and increased protein activity (Axelsson et al. 2013; German et al. 2016; Gout et al. 2010; Heras et al. 2020; Perry et al. 2007; Qian and Zhang 2014). This prediction was supported (Fig. 2) and this finding is in line with the observation that *C. violaceus* has gut amylolytic activity that is more than an order of magnitude higher than that of *A. purpureus*, reflecting the higher starch content in the diet of *C. violaceus* (German et al. 2004, 2015; Neighbors and Horn 1991). In general, carnivorous fishes have lower gut amylase activities than herbivorous or omnivorous fishes (Chakraborty et al. 1995; Chan et al. 2004; Fernandez et al. 2001; Hidalgo et al. 1999; Sabapathy and Teo 1993). Moreover, the observation of elevated amylolytic activity in the guts of animals consuming more starch is well-known across nearly all animal clades (see German et al. 2016 for a detailed discussion).

Our prediction of fewer gene copies of carboxyl ester lipase (*cel*) in *A. purpureus* compared to *C. violaceus* was supported (Fig. 3), whereas there was no difference in gene copy number for gastric triacylglycerol lipase and hepatic lipase between the two species (Table 2). Heras et al. (2020) confirmed that the *cel* genes are expressed in the gut and

Fig. 2 Synteny map for pancreatic α -amylase genes (*amy2*) from *D. rerio*, *O. latipes*, *C. violaceus*, and *A. purpurescens*. See Supplemental Table S2 for information on genetic resources for each species

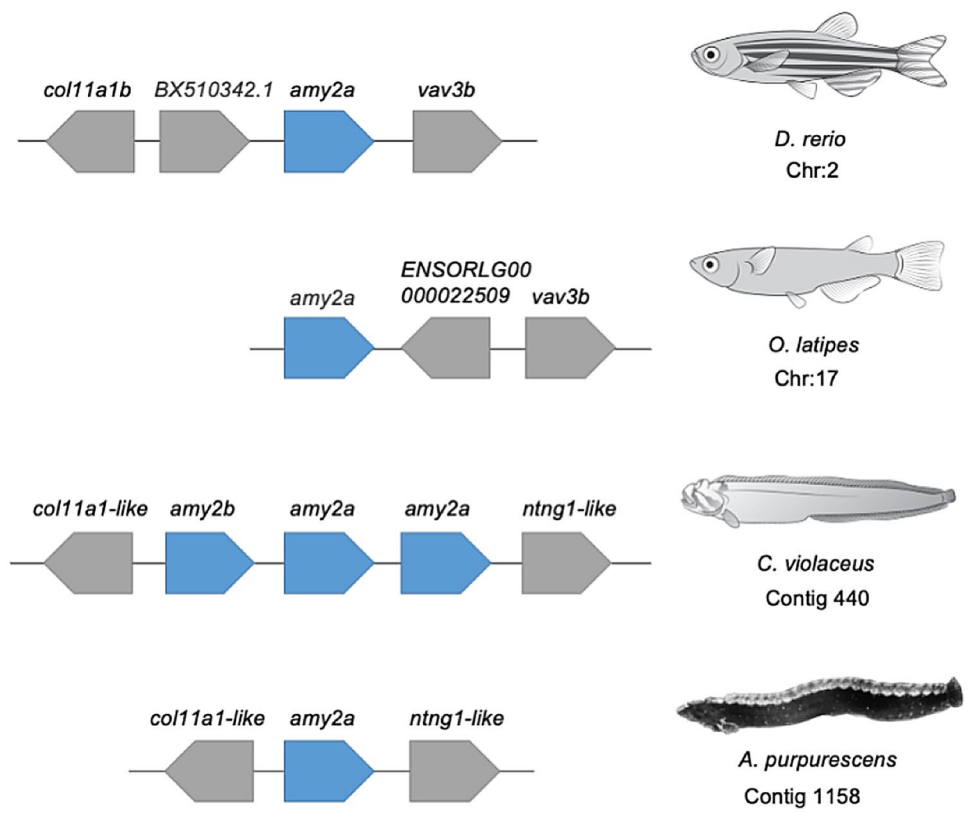
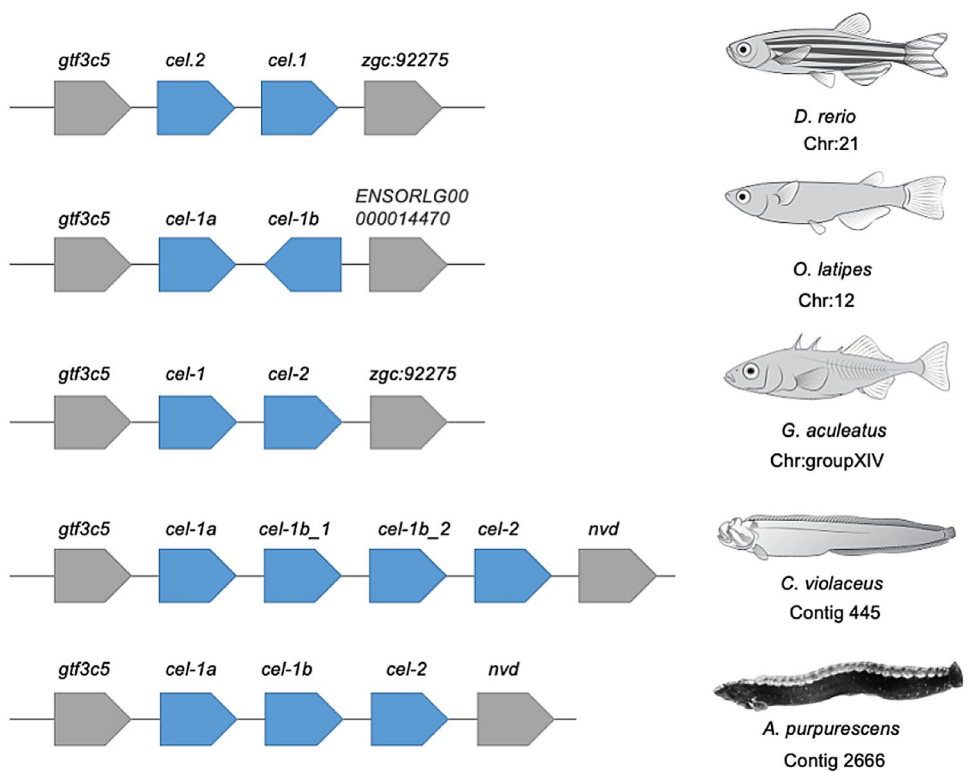


Fig. 3 Synteny map for carboxyl ester lipase genes (*cel*) from *D. rerio*, *O. latipes*, *G. aculeatus*, *C. violaceus*, and *A. purpurescens*. See Supplemental Table S2 for information on genetic resources for each species



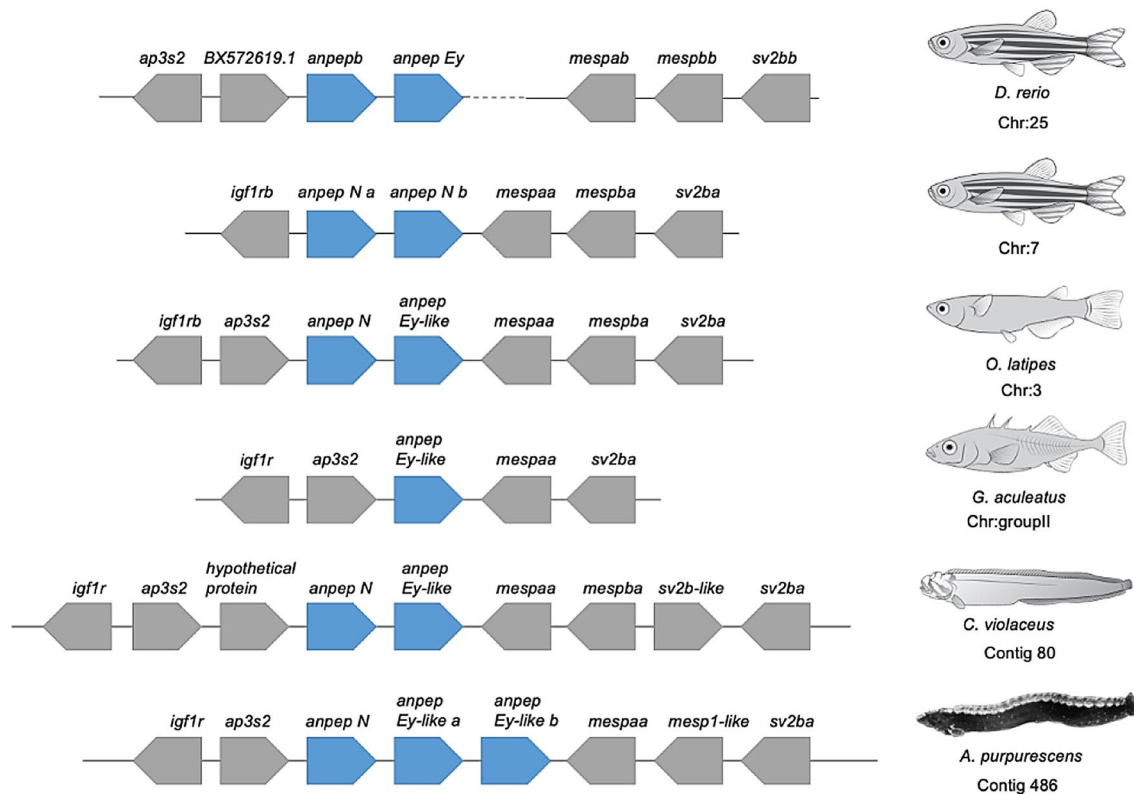


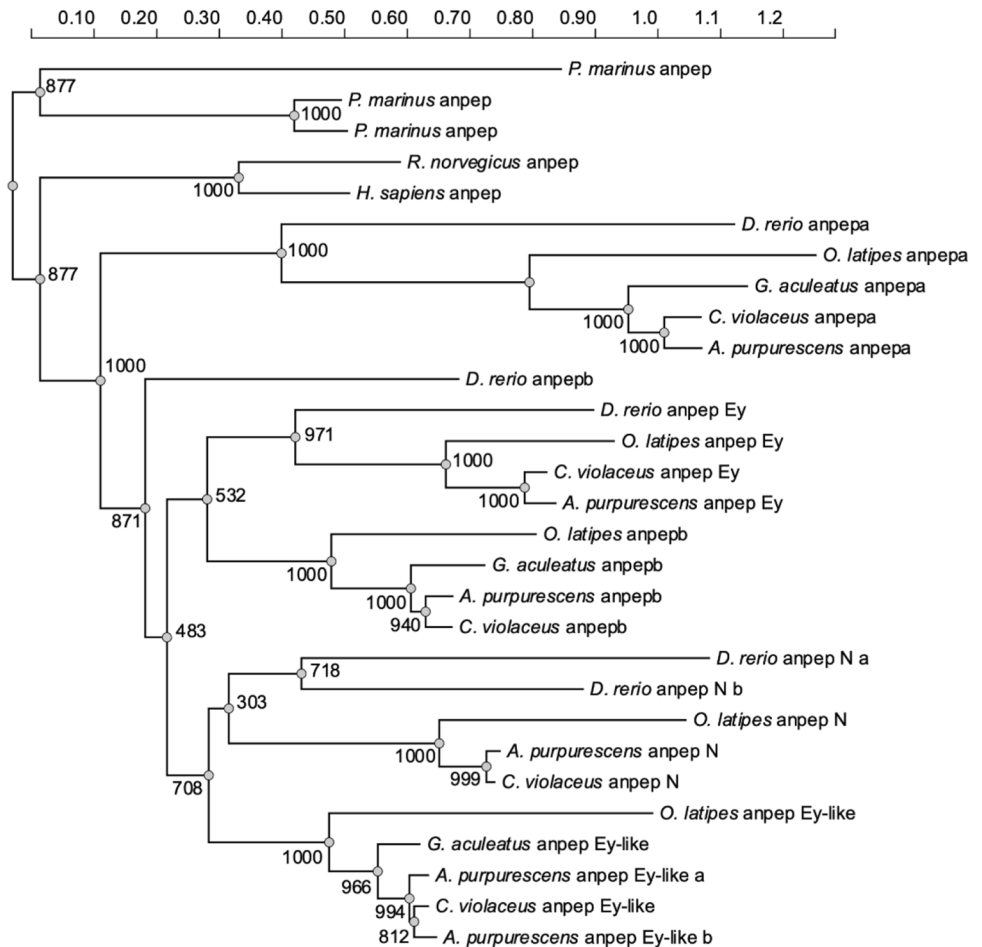
Fig. 4 Synteny map for aminopeptidase N and Ey-like genes (*anpep N* and *anpep Ey-like*) from *D. rerio*, *O. latipes*, *G. aculeatus*, *C. violaceus*, and *A. purpureus*. See Supplemental Table S2 for information on genetic resources for each species

code for the carboxyl ester lipase (CEL) enzyme, which is the primary intestinal lipase in fishes and has broad specificity, especially toward algal and plant lipids (Li-Beisson et al. 2019; Murray et al. 2003; Olsen and Ringø 1997; Sæle et al. 2010). Although an animal-based diet contains more total lipids (Neighbors and Horn 1991), those lipids are more of the triacylglycerol variety (Sahaka et al. 2020). Plant material, although lower in lipid than animal tissue, is abundant in galacto- and betaine lipids (Kato et al. 1996; Li-Beisson et al. 2019; Sahaka et al. 2020), which CEL efficiently digests. Data for lipolytic activity against different lipid types is lacking for fishes, but herbivorous and omnivorous insects and mammals hydrolyze galactolipids at a higher rate than carnivores, whereas triacylglycerides are hydrolyzed at higher rates in carnivorous insects and mammals (Amara et al. 2010; Christeller et al. 2011). Animal material and plant material both contain phospholipids (German et al. 1996; Murray et al. 2003), and phospholipase gene copy number has not been found to vary among the fish species that have been examined thus far (Castro-Ruiz et al. 2021; Heras et al. 2020). While fishes may not possess the lipase diversity found in mammals (Murray et al. 2003; Olsen and Ringø 1997; Sæle et al. 2010; Tang et al. 2022), they do possess several *cel* genes that are worth investigating (Tang

et al. 2022; Fig. 3, Supplemental Figs. S5 and S6). There are generally two *cel* loci in fishes: one that contains *cel-1* and *cel-2* (and their copies), and a different one for *cel-like* (Tang et al. 2022; Fig. 3, Supplemental Figs. S5 and S6). Each of these *cel* genes group separately in gene phylogenetic trees, showing that they have their own properties (Tang et al. 2022; Supplemental Fig. S6), although Tang et al. (2022) haphazardly named their *cel* genes as bile salt activated lipase (*bsal*) with random numbers, which we attempted to salvage based on the phylogenetic relationships of the genes themselves (Supplemental Fig. S6). Like *anpep*, the naming of *cel* genes in fishes deserves more attention. The extra copy of *cel 1* in the *C. violaceus* genome could help explain the observation that total lipolytic activity (across the whole gut) in *C. violaceus* is more than twice that of *A. purpureus* (German et al. 2004, 2015). However, more specific analyses are needed, including a pH stat method (Amara et al. 2010; Christeller et al. 2011; Sahaka et al. 2020), to differentiate between different types of lipase activities and to determine what dietary substrates the lipases in *C. violaceus* and *A. purpureus* can hydrolyze.

Perhaps one of the most intriguing findings in this study is that *A. purpureus* has only one additional copy of alanine aminopeptidase (Figs. 4 and 5, Supplemental Fig. S6),

Fig. 5 Phylogenetic relationship of alanyl aminopeptidase genes (*anpep*) in fishes (including *A. purpurescens*). A maximum likelihood (ML) tree was constructed with 1000 bootstrap replicates in PhyML v3.0 based on alanyl aminopeptidase sequences from *A. purpurescens*, *C. violaceus*, *G. aculeatus*, *O. latipes*, *D. rerio*, *Homo sapiens* and *Rattus norvegicus*. Alanyl aminopeptidase sequences from *Petromyzon marinus* were used as an outgroup. See Supplemental Table S2 for information on genetic resources for each species

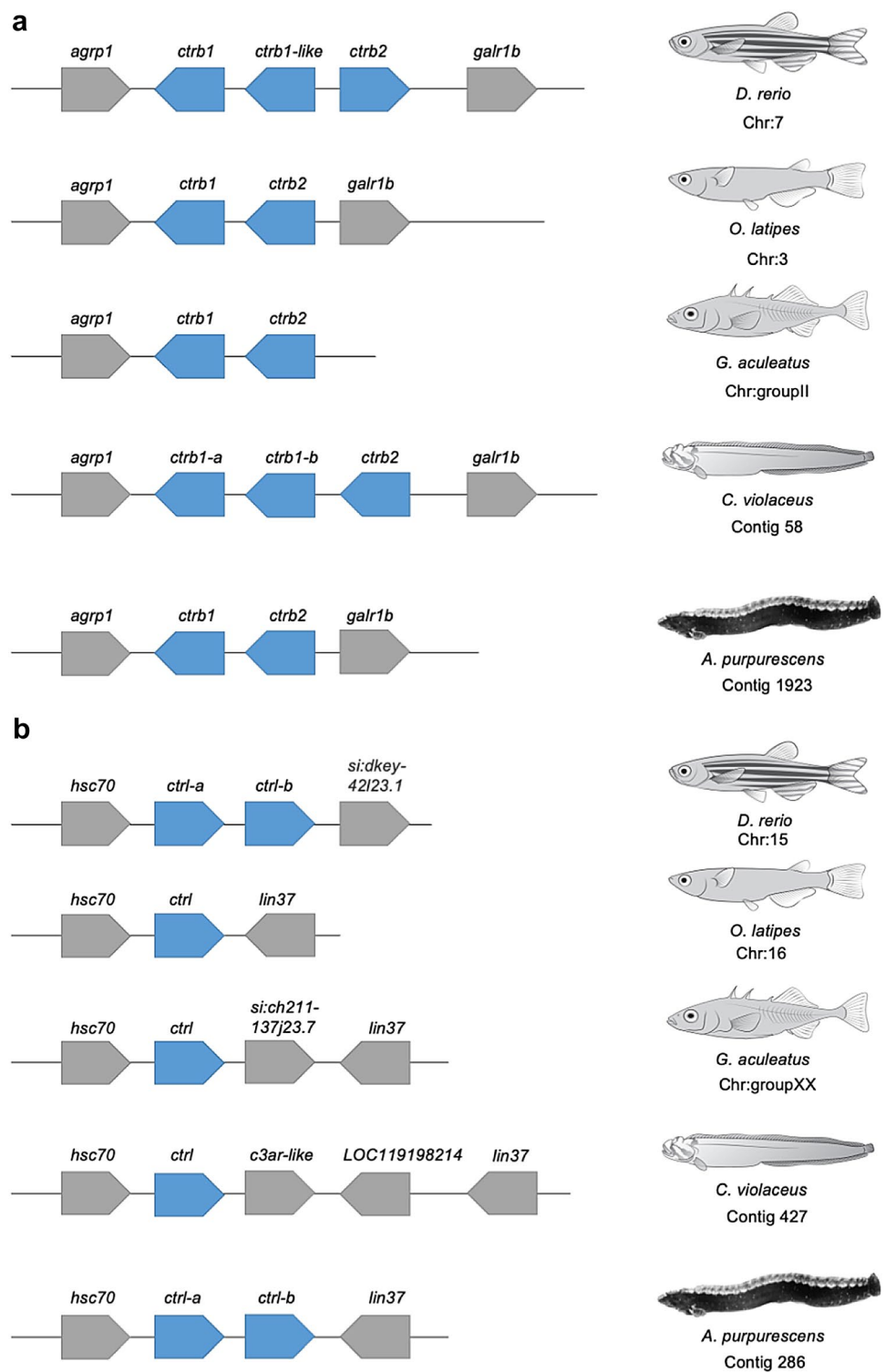


compared to *C. violaceus*. This finding was surprising given that, on a per-gram tissue basis, the carnivorous *A. purpurescens* has approximately doubled the aminopeptidase activity in its gut compared to the herbivorous *C. violaceus* (German et al. 2004, 2015). The aminopeptidase enzyme hydrolyzes peptides prior to amino acid absorption in the intestine (Karasov and Douglas 2013). The additional copy of aminopeptidase in *A. purpurescens* is at the *anpep Ey-like* locus (Fig. 4), but there are five different copies of alanyl-aminopeptidase in each species (*anpep a*, *anpep b*, *anpep N*, *anpep Ey*, *anpep Ey-like*), that are also found in other fishes (Figs. 4 and 5, Supplemental Fig. S7). This motivated us to follow up on the phylogenetic analysis of *anpep* genes that Heras et al. (2020) first reported. The three ancestral vertebrate *anpep* genes in *Petromyzon marinus* are most similar to the *anpep N* gene in mammals, and in turn, these are sister to the fish aminopeptidases (with the limited number we analyzed; Fig. 5). The most ancestral teleost aminopeptidase is *anpep a*. In vertebrate evolution, there were two rounds of whole genome duplication (WGD), followed by a teleost-fish-specific WGD event (Christoffels et al. 2004; Glasauer and Neuhauss 2014; Kasahara 2007; Ohno 1970). Genes retained from WGDs are known as ohnologs (Ohno 1970).

According to the website <http://ohnologs.curie.fr>, which predicts ohnologs, *anpep a* and *anpep b* are ohnologs from one of the vertebrate WGD events, whereas *anpep b* and *anpep N* are ohnologs from the teleost-fish-specific WGD (Heras et al. 2020; Kasahara 2007; Ohno 1970). There is similarity of the surrounding genes, in the respective loci, of these *anpep* genes (Heras et al. 2020). It appears that *anpep Ey* is a paralog of *anpep b*, whereas *anpep Ey-like* is a paralog of *anpep N* (Fig. 5, Supplemental Fig. S7). Each of these is found in other teleost fishes, with some variations (e.g., *G. aculeatus* apparently lacks *anpep N* and *D. rerio* lacks *anpep Ey-like*; Figs. 4 and 5, Supplemental Fig. S7; Heras et al. 2020). Each of these *anpep* genes shows strong gut expression in pricklebacks (Heras et al. 2020; Herrera et al. 2022). ANPEP enzyme activity in *A. purpurescens* should be explored in more detail to determine if the extra copy of *anpep Ey-like* is causing elevated enzymatic activity in this species, and/or some other *anpep* gene is being expressed at a higher level (e.g., Brun et al. 2021) and boosting ANPEP activity.

We have observed some inconsistencies with the number of, and naming of, vertebrate alanyl-aminopeptidases. What is known as *anpep N* in humans and other mammals

Fig. 7 a Synteny map for chymotrypsin genes (*ctr*) from *D. rerio*, *O. latipes*, *G. aculeatus*, *C. violaceus*, and *A. purpureus* **b** Synteny map for chymotrypsin-like genes (*ctrl*) from *D. rerio*, *O. latipes*, *G. aculeatus*, *C. violaceus*, and *A. purpureus*. See Supplemental Table S2 for information on genetic resources for each species



gene in a genome, and the dosage effect of the protein isn't necessary (Chen et al. 1997).

The final protease that showed gene copy number differences between *A. purpureus* and *C. violaceus* is chymotrypsinogen (*ctr*) and chymotrypsinogen-like (*ctrl*) (Fig. 7, Supplemental Fig. S10). Chymotrypsin is a pancreatic

serine protease, like trypsin, but there is some debate as to whether chymotrypsin matters more for fishes consuming more plant material (Gioda et al. 2017; Heras et al. 2020; Ruan et al. 2010; Rungruangsak-Torrissen et al. 2006). Chymotrypsin cleaves different peptide bonds than trypsin does (phenylalanine, tyrosine, and tryptophan as opposed

to lysine and arginine for trypsin; Ma et al. 2005), and has shown different temperature optima than trypsin, perhaps providing different activity under different circumstances for an animal (Navarro-Guillén et al. 2022). What is intriguing here is that *C. violaceus* has an extra copy at the *ctrl* locus, whereas *A. purpurascens* has an extra copy at the *ctrl* locus (Fig. 7, Supplemental Fig. S10). We have never measured the activity of chymotrypsin in the guts of pricklebacks. Chymotrypsin requires more attention to discern its role in the digestive process in fishes consuming different diets (Castro-Ruiz et al. 2019, 2021; Heras et al. 2020; Navarro-Guillén et al. 2022).

In conclusion, we produced a high-quality fish genome, and analyzed it in the context of a growing literature on the nutritional physiology of *A. purpurascens* and other prickleback fishes. Interestingly, the vast majority of sequenced fish genomes are for carnivorous species, since these are largely the ones we culture for human consumption or ornamental use (Heras et al. 2020). In this case, because of the recent publication of the *C. violaceus* genome (Heras et al. 2020), we were able to compare the carnivorous *A. purpurascens* genome to that of a closely related, sympatric, herbivorous species. Based on years of ecological (Horn et al. 1986), physiological (Fris and Horn 1993), biochemical (Chan et al. 2004; German et al. 2016, 2004, 2015), and molecular data (Gawlicka and Horn 2006; Heras et al. 2020; Herrera et al. 2022; Kim et al. 2014), we made a priori predictions about gene copy number for specific digestive enzymes among *A. purpurascens* and *C. violaceus*, and our results largely support the AMH (Karasov 1992; Karasov and Martinez del Rio 2007) from the genomic to the phenotypic level. This powerful physiological genomics approach provides new ways forward in nutritional physiological research, generating new hypotheses on how animals specialize to use different resources (Brun et al. 2020). Indeed, comparative genomics is becoming more common and can lead to more informed understanding of the biology of various taxa, particularly if other data are known about the studied species (Axelsson et al. 2013; Heras et al. 2020; Lamichhaney et al. 2015; Protas et al. 2005; Taylor et al. 2021). Our focus on digestive enzymes is because enzymes are the agents of chemical digestion (Vonk and Western 1984; Karasov and Douglas 2013), and diversity at this key step in nutrient acquisition is likely to inform how animals use various resources (Brun et al. 2020; Heras et al. 2020) although focusing on the liver can also inform about how fishes metabolize various nutrients once they are absorbed (e.g., Heras et al. 2020; Herrera et al. 2022). Finally, given that *C. violaceus* and *A. purpurascens* are commonly found in Marine Protected Areas on the west coast of the United States, our data will also have application for conservation of these species.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s00438-023-02067-5>.

Acknowledgements We would like to thank Mahul Chakraborty and JJ Emerson for providing reagents for the genomic DNA extraction and allowing us to use their pulse-field gel electrophoresis rig. We thank the staff at the Genomics Research and Technology Hub at UC Irvine. We thank A. Dingeldin for the fish illustrations used in our figures. This work was funded by NSF Grant IOS-1355224 and UC Irvine School of Biological Sciences Bridge Funds (to D.P.G.), and Soka University Research Funds (to L.T.C.).

Data availability The final genome annotation was uploaded as project PRJNA950117 at NCBI (<https://www.ncbi.nlm.nih.gov/bioproject/>).

Declarations

Conflict of interest The authors declare no competing interests.

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

References

- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215(3):403–410. [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2)
- Amara S, Barouh N, Lecomte J, Lafont D, Robert S, Villeneuve P, De Caro A, Carrière F (2010) Lipolysis of natural long chain and synthetic medium chain galactolipids by pancreatic lipase-related protein 2. *Biochimica et Biophysica Acta (BBA) Mol Cell Biol Lipids* 1801 (4):508–516. <https://doi.org/10.1016/j.bbalip.2010.01.003>
- Andrews S (2010) FastQC: A quality control tool for high throughput sequence data. <https://www.bioinformaticsbabraham.ac.uk/projects/fastqc/>
- Asalone KC, Ryan KM, Yamadi M, Cohen AL, Farmer WG, George DJ, Joppert C, Kim KH, Mughal MF, Said R, Toksoz-Exley M, Bisk E, Bracht JR (2020) Regional sequence expansion or collapse in heterozygous genome assemblies. *PLOS Comput Biol* 16(7):e1008104. <https://doi.org/10.1371/journal.pcbi.1008104>
- Axelsson E, Ratnakumar A, Arendt ML, Maqbool K, Webster MT, Perloski M, Liberg O, Arnemo JM, Hedhammar A, Lindblad-Toh K (2013) The genomic signature of dog domestication reveals adaptation to a starch-rich diet. *Nature* 495:360–364
- Baldwin-Brown JG, Weeks SC, Long AD (2018) A new standard for Crustacean genomes: the highly contiguous, annotated genome assembly of the clam shrimp *Eulimnadia texana* reveals HOX gene order and identifies the sex chromosome. *Genome Biol Evol* 10(1):143–156. <https://doi.org/10.1093/gbe/evx280>
- Barnett DW, Garrison EK, Quinlan AR, Strömberg MP, Marth GT (2011) BamTools *Bioinform* 27(12):1691–1692. <https://doi.org/10.1093/bioinformatics/btr174>

- Bradnam KR, Fass JN, Alexandrov A, Baranay P, Bechner M, Birol I, Boisvert S, Chapman JA, Chapuis G, Chikhi R, Chitsaz H, Chou W-C, Corbeil J, Del Fabbro C, Docking TR, Durbin R, Earl D, Emrich S, Fedotov P, Fonseca NA, Ganapathy G, Gibbs RA, Gnerre S, Godzaridis É, Goldstein S, Haimel M, Hall G, Haussler D, Hiatt JB, Ho IY, Howard JJ, Hunt M, Jackman SD, Jaffe DB, Jarvis ED, Jiang HY, Kazakov S, Kersey PJ, Kitzman JO, Knight JR, Koren S, Lam T-W, Lavenier D, Laviolette F, Li Y, Li Z, Liu B, Liu Y, Luo R, MacCallum I, MacManes MD, Maillet N, Melnikov S, Naquin D, Ning Z, Otto TD, Paten B, Paulo OS, Phillippy AM, Pina-Martins F, Place M, Przybylski D, Qin X, Qu C, Ribeiro FJ, Richards S, Rokhsar DS, Ruby JG, Scalabrin S, Schatz MC, Schwartz DC, Sergushichev A, Sharpe T, Shaw TI, Shendure J, Shi Y, Simpson JT, Song H, Tsarev F, Vezzi F, Vicedomini R, Vieira BM, Wang J, Worley KC, Yin S, Yiu S-M, Yuan J, Zhang G, Zhang H, Zhou S, Korf IF (2013) Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience* 2(1):10. <https://doi.org/10.1186/2047-217X-2-10>
- Brun A, Mendez-Aranda D, Magallanes ME, Karasov WH, Martínez del Río C, Baldwin MW, Caviedes-Vidal E (2020) Duplications and functional convergence of intestinal carbohydrate-digesting enzymes. *Mol Biol Evol* 37(6):1657–1666. <https://doi.org/10.1093/molbev/msaa034>
- Brun A, Magallanes ME, Karasov WH, Caviedes-Vidal E (2021) Rapid and parallel changes in activity and mRNA of intestinal peptidase to match altered dietary protein levels in juvenile house sparrows (*Passer domesticus*). *J Experim Biol* 224(3):2jeb234708. <https://doi.org/10.1242/jeb.234708>
- Brůna T, Lomsadze A, Borodovsky M (2020) GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins. *NAR Genom Bioinform*. <https://doi.org/10.1093/nargab/lqaa026>
- Brůna T, Hoff KJ, Lomsadze A, Stanke M, Borodovsky M (2021) BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genom Bioinform*. <https://doi.org/10.1093/nargab/lqaa108>
- Buchfink B, Xie C, Huson DH (2015) Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 12(1):59–60. <https://doi.org/10.1038/nmeth.3176>
- Bushnell Brian (2014) “BBMap: A Fast, Accurate, Splice-Aware Aligner”. United States. <https://www.osti.gov/servlets/purl/1241166>
- Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: architecture and applications. *BMC Bioinform* 10(1):421. <https://doi.org/10.1186/1471-2105-10-421>
- Castro-Ruiz D, Mozanzadeh MT, Fernández-Méndez C, Andree KB, García-Dávila C, Cahu CL, Gisbert E, Darias MJ (2019) Ontogeny of the digestive enzyme activity of the Amazonian pimelodid catfish *Pseudoplatystoma punctifer* (Castelna, 1855). *Aquaculture* 504:210–218. <https://doi.org/10.1016/j.aquaculture.2019.01.059>
- Castro-Ruiz D, Andree KB, Blondeau-Bidet E, Fernández-Méndez C, García-Dávila C, Gisbert E, Darias MJ (2021) Isolation, identification, and gene expression analysis of the main digestive enzymes during ontogeny of the Neotropical catfish *Pseudoplatystoma punctifer* (Castelna, 1855). *Aquaculture* 543:737031. <https://doi.org/10.1016/j.aquaculture.2021.737031>
- Chakrabarti I, Gani MA, Chaki KK, Sur R, Misra KK (1995) Digestive enzymes in 11 freshwater teleost fish species in relation to food habit and niche segregation. *Compar Biochem Physiol A* 112:167–177
- Chakraborty M, Baldwin-Brown JG, Long AD, Emerson JJ (2016) Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. *Nucleic Acids Res* 44(19):e147–e147. <https://doi.org/10.1093/nar/gkw654>
- Chan AS, Horn MH, Dickson KA, Gawlicka A (2004) Digestive enzyme activity in carnivores and herbivores: comparisons among four closely related pricklyback fishes (Teleostei: Stichaeidae) from a California rocky intertidal habitat. *J Fish Biol* 65:848–858
- Chen L, DeVries AL, Cheng C-HC (1997) Evolution of antifreeze glycoprotein gene from a trypsinogen gene in antarctic notothenioid fish. *Proc Natl Acad Sci USA* 94(8):3811–3816
- Chen Z, Huang X, Fu R, Zhan A (2022) Neighbours matter: Effects of genomic organization on gene expression plasticity in response to environmental stresses during biological invasions. *Compar Biochem Physiol Part D Genom Proteom* 42:100992. <https://doi.org/10.1016/j.cbd.2022.100992>
- Chin C-S, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, Turner SW, Korlach J (2013) Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods* 10(6):563–569. <https://doi.org/10.1038/nmeth.2474>
- Choi JI, Yamazaki T (1994) Molecular analysis of cis-regulatory sequences of the alpha-amylase gene in *D. melanogaster*: a short 5'-flanking region of Amy distal gene is required for full expression of Amy proximal gene. *Japan J Genet* 69(6):619–635. <https://doi.org/10.1266/jgg.69.619>
- Christeller JT, Amara S, Carrière F (2011) Galactolipase, phospholipase and triacylglycerol lipase activities in the midgut of six species of lepidopteran larvae feeding on different lipid diets. *J Insect Physiol* 57:1232–1239
- Christoffels A, Koh EGL, Chia JM, Brenner S, Aparicio S, Venkatesh B (2004) Fugu genome analysis provides evidence for a whole-genome duplication early during the evolution of ray-finned fishes. *Mol Biol Evol* 21(6):1146–1151. <https://doi.org/10.1093/molbev/msh114>
- Cockell M, Stevenson BJ, Strubin M, Hagenbüchle O, Wellauer PK (1989) Identification of a cell-specific DNA-binding activity that interacts with a transcriptional activator of genes expressed in the acinar pancreas. *Mol Cell Biol* 9(6):2464–2476. <https://doi.org/10.1128/mcb.9.6.2464>
- Erşahin C, Szpaderska AM, Orawski AT, Simmons WH (2005) Aminopeptidase P isozyme expression in human tissues and peripheral blood mononuclear cell fractions. *Arch Biochem Biophys* 435(2):303–310. <https://doi.org/10.1016/j.abb.2004.12.023>
- Fan G, Song Y, Yang L, Huang X, Zhang SY, Zhang M, Yang X, Chang Y, Zhang H, Li Y, Liu S, Yu L, Chu J, Seim I, Feng C, Near TJ, Wing RA, Wang W, Wang K, Wang J, Xu X, Yang H, Liu X, Chen N, He S (2020) Initial data release and announcement of the 10,000 Fish Genomes Project (Fish10K). *GigaScience*. <https://doi.org/10.1093/gigascience/giaa080>
- Fernandez I, Moyano FJ, Diaz M, Martinez T (2001) Characterization of α -amylase activity in five species of Mediterranean sparid fishes (Sparidae, Teleostei). *J Exp Mar Biol Ecol* 262:1–12
- Fris MB, Horn MH (1993) Effects of diets of different protein content on food consumption, gut retention, protein conversion, and growth of *Cebidichthys violaceus* (Girard), an herbivorous fish of temperate zone marine waters. *J Exp Mar Biol Ecol* 166:185–202
- Gawlicka A, Horn MH (2006) Trypsin gene expression by quantitative in situ hybridization in carnivorous and herbivorous pricklyback fishes (Teleostei : Stichaeidae): ontogenetic, dietary, and phylogenetic effects. *Physiol Biochem Zool* 79:120–132
- German DP, Horn MH (2006) Gut length and mass in herbivorous and carnivorous pricklyback fishes (Teleostei: Stichaeidae): ontogenetic, dietary, and phylogenetic effects. *Mar Biol* 148:1123–1134
- German JB, Xu R, Walzem R, Kinsella JE, Knuckles B, Nakamura M, Yokoyama WH (1996) Effect of dietary fats and barley fiber

- on total cholesterol and lipoprotein cholesterol distribution in plasma of hamsters. *Nutr Res* 16(7):1239–1249. [https://doi.org/10.1016/0271-5317\(96\)00127-3](https://doi.org/10.1016/0271-5317(96)00127-3)
- German DP, Horn MH, Gawlicka A (2004) Digestive enzyme activities in herbivorous and carnivorous prickleback fishes (Teleostei: Stichaeidae): ontogenetic, dietary, and phylogenetic effects. *Physiol Biochem Zool* 77:789–804
- German DP, Sung A, Jhaveri PK, Agnihotri A (2015) More than one way to be an herbivore: convergent evolution of herbivory using different digestive strategies in prickleback fishes (family Stichaeidae). *Zoology* 118:161–170
- German DP, Foti DM, Heras J, Amerkhanian H, Lockwood BL (2016) Elevated gene copy number does not always explain elevated amylase activities in fishes. *Physiol Biochem Zool* 89:277–293. <https://doi.org/10.1086/687288>
- Gioda CR, Pretto A, de Souza FC, Leitemperger J, Loro VL, Lazzari R, Lissner LA, Baldissarotto B, Salbego J (2017) Different feeding habits influence the activity of digestive enzymes in freshwater fish. *Ciência Rural, Santa Maria* 47(03):e20160113
- Glasauer SMK, Neuhauss SCF (2014) Whole-genome duplication in teleost fishes and its evolutionary consequences. *Mol Genet Genom* 289(6):1045–1060. <https://doi.org/10.1007/s00438-014-0889-2>
- Gotoh O, Morita M, Nelson DR (2014) Assessment and refinement of eukaryotic gene structure prediction with gene-structure-aware multiple protein sequence alignment. *BMC Bioinform* 15(1):189. <https://doi.org/10.1186/1471-2105-15-189>
- Götz S, García-Gómez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, Robles M, Talón M, Dopazo J, Conesa A (2008) High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res* 36(10):3420–3435. <https://doi.org/10.1093/nar/gkn176>
- Gout JF, Kahn D, Duret L, Consortium PP-G (2010) The relationship among gene expression, the evolution of gene dosage, and the rate of protein evolution. *PLoS Genet* 6(5):e1000944. <https://doi.org/10.1371/journal.pgen.1000944>
- Guan D, McCarthy SA, Wood J, Howe K, Wang Y, Durbin R (2020) Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* 36(9):2896–2898. <https://doi.org/10.1093/bioinformatics/btaa025>
- Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *System Biol* 59(3):307–321. <https://doi.org/10.1093/sysbio/syq010>
- Heras J, Chakraborty M, Emerson JJ, German DP (2020) Genomic and biochemical evidence of dietary adaptation in a marine herbivorous fish. *Proc Royal Soc B Biol Sci* 287:20192327
- Herrera MJ, Heras J, German DP (2022) Comparative transcriptomics reveal tissue level specialization towards diet in prickleback fishes. *J Comp Physiol B* 192:275–295
- Hidalgo MC, Urea E, Sanz A (1999) Comparative study of digestive enzymes in fish with different nutritional habits. *Proteolytic Amylase Activit Aquac* 170:267–283
- Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M (2016) BRAKER1: unsupervised RNA-seq-based genome annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics* 32(5):767–769. <https://doi.org/10.1093/bioinformatics/btv661>
- Hoff KJ, Lomsadze A, Borodovsky M, Stanke M (2019) Whole-genome annotation with BRAKER. In: Kollmar M (Ed) *Gene Prediction: Methods and Protocols*. Springer New York, New York, NY, pp 65–95. https://doi.org/10.1007/978-1-4939-9173-0_5
- Horn MH (1989) Biology of marine herbivorous fishes. *Oceanography Marine Biol Annual Rev* 27:167–272
- Horn MH, Neighbors MA, Murray SN (1986) Herbivore responses to a seasonally fluctuating food supply: growth potential of two temperate intertidal fishes based on the protein and energy assimilated from their macroalgal diets. *J Exp Mar Biol Ecol* 103:217–234
- Howard G, Keller PR, Johnson TM, Meisler MH (1989) Binding of a pancreatic nuclear protein is correlated with amylase enhancer activity. *Nucleic Acids Res* 17(20):8185–8194. <https://doi.org/10.1093/nar/17.20.8185>
- Iwata H, Gotoh O (2012) Benchmarking spliced alignment programs including Spaln2, an extended version of Spaln that incorporates additional species-specific features. *Nucleic Acids Res* 40(20):e161–e161. <https://doi.org/10.1093/nar/gks708>
- Kajitani R, Toshimoto K, Noguchi H, Toyoda A, Ogura Y, Okuno M, Yabana M, Harada M, Nagayasu E, Maruyama H, Kohara Y, Fujiyama A, Hayashi T, Itoh T (2014) Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Res* 24(8):1384–1395. <https://doi.org/10.1101/gr.170720.113>
- Karasov WH (1992) Tests of the adaptive modulation hypothesis for dietary control of intestinal nutrient transport. *Am J Physiol* 263:R496–R502
- Karasov WH, Douglas AE (2013) *Comparative Digestive Physiology*. Comprehensive. Physiology 3:741–783
- Karasov WH, Martínez del Río C (2007) *Physiological ecology: how animals process energy, nutrients, and toxins*. Princeton University Press, Princeton, NJ USA
- Kasahara M (2007) The 2R hypothesis: an update. *Curr Opin Immunol* 19(5):547–552. <https://doi.org/10.1016/j.coi.2007.07.009>
- Kato M, Sakai M, Adachi K, Ikemoto H, Sano H (1996) Distribution of betaine lipids in marine algae. *Phytochemistry* 42(5):1341–1345. [https://doi.org/10.1016/0031-9422\(96\)00115-X](https://doi.org/10.1016/0031-9422(96)00115-X)
- Kim KH, Horn MH, Sosa AE, German DP (2014) Sequence and expression of an α -amylase gene in four related species of prickleback fishes (Teleostei: Stichaeidae): ontogenetic, dietary, and species-level effects. *J Comp Physiol B* 184:221–234
- Kim D, Paggi JM, Park C, Bennett C, Salzberg SL (2019) Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* 37(8):907–915. <https://doi.org/10.1038/s41587-019-0201-4>
- Kolmogorov M, Yuan J, Lin Y, Pevzner PA (2019) Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol* 37(5):540–546. <https://doi.org/10.1038/s41587-019-0072-8>
- Kriventseva EV, Kuznetsov D, Tegenfeldt F, Manni M, Dias R, Simão FA, Zdobnov EM (2019) OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Res* 47(D1):D807–D811. <https://doi.org/10.1093/nar/gky1053>
- Lamichhane S, Berglund J, Almén MS, Maqbool K, Grabherr MG, Martínez-Barrio A, Promerová M, Rubin C-J, Wang C, Zamani N, Grant BR, Grant PR, Webster MT, Andersson L (2015) Evolution of Darwin's finches and their beaks revealed by genome sequencing. *Nature* 518(7539):371–375. <https://doi.org/10.1038/nature14181>
- Lefort V, Longueville J-E, Gascuel O (2017) SMS: Smart Model Selection in PhyML. *Mol Biol Evol* 34(9):2422–2424. <https://doi.org/10.1093/molbev/msx149>
- Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25(14):1754–1760. <https://doi.org/10.1093/bioinformatics/btp324>
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16):2078–2079. <https://doi.org/10.1093/bioinformatics/btp352>
- Li-Beisson Y, Thelen JJ, Fedosejevs E, Harwood JL (2019) The lipid biochemistry of eukaryotic algae. *Prog Lipid Res* 74:31–68

- Lomsadze A, Ter-Hovhannissyan V, Chernoff YO, Borodovsky M (2005) Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res* 33(20):6494–6506. <https://doi.org/10.1093/nar/gki937>
- Lomsadze A, Burns PD, Borodovsky M (2014) Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Res* 42(15):e119–e119. <https://doi.org/10.1093/nar/gku557>
- Lu L, Zhao J, Li C (2020) High-quality genome assembly and annotation of the big-eye mandarin fish (*Siniperca kneri*). *G3 (Bethesda)* 10 (3):877–880. <https://doi.org/10.1534/g3.119.400930>
- Lu G, Luo M (2020) Genomes of major fishes in world fisheries and aquaculture: Status, application and perspective. *Aquaculture and Fisheries* 5(4):163–173. <https://doi.org/10.1016/j.aaf.2020.05.004>
- Ma P, Liu Y, Reddy KP, Chan WK, Lam TJ (2004) Characterization of the seabass pancreatic α -amylase gene and promoter. *Gen Comp Endocrinol* 137(1):78–88. <https://doi.org/10.1016/j.ygcen.2004.02.018>
- Ma W, Tang C, Lai L (2005) Specificity of trypsin and chymotrypsin: loop-motion-controlled dynamic correlation as a determinant. *Biophys J* 89(2):1183–1193. <https://doi.org/10.1529/biophysj.104.057158>
- Madeira F, Pearce M, Tivey ARN, Basutkar P, Lee J, Edbali O, Madhusoodanan N, Kolesnikov A, Lopez R (2022) Search and sequence analysis tools services from EMBL-EBI in 2022. *Nucleic Acids Res* 50(W1):W276–W279. <https://doi.org/10.1093/nar/gkac240>
- Marçais G, Kingsford C (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 27(6):764–770. <https://doi.org/10.1093/bioinformatics/btr011>
- Murray H, Gallant J, Perez-Casanova J, Johnson S, Douglas S (2003) Ontogeny of lipase expression in winter flounder. *J Fish Biol* 62:816–833
- Navarro-Guillén C, Yúfera M, Perera E (2022) Biochemical features and modulation of digestive enzymes by environmental temperature in the greater amberjack, *Seriola Dumerili*. *Front Marine Sci* 9:960746. <https://doi.org/10.3389/fmars.2022.960746>
- Neighbors MA, Horn MH (1991) Nutritional quality of macrophytes eaten and not eaten by two temperate-zone herbivorous fishes: a multivariate comparison. *Mar Biol* 108:471–476
- Ohno S (1970) *Evolution by gene duplication*. Springer-Verlag, New York
- Olsen RE, Ringø E (1997) Lipid Digestibility in Fish: a Review Recent Research Developments in Lipid Research 1:199–264
- Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, Werner J, Villanea FA, Mountain JL, Misra R, Carter NP, Lee C, Stone AC (2007) Diet and the evolution of human amylase gene copy number variation. *Nat Genet* 39:1256–1260
- Protas ME, Hersey C, Kochanek D, Zhou Y, Wilkens H, Jeffery WR, Zon LI, Borowsky R, Tabin CJ (2005) Genetic analysis of cavefish reveals molecular convergence in the evolution of albinism. *Nat Genet* 38:107. <https://doi.org/10.1038/ng1700> <https://www.nature.com/articles/ng1700#supplementary-information>
- Qian W, Zhang J (2014) Genomic evidence for adaptation by gene duplication. *Genome Res* 24(8):1356–1362. <https://doi.org/10.1101/gr.172098.114>
- Ranallo-Benavidez TR, Jaron KS, Schatz MC (2020) GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat Commun* 11(1):1432. <https://doi.org/10.1038/s41467-020-14998-3>
- Rhie A, Walenz BP, Koren S, Phillippy AM (2020) Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol* 21(1):245. <https://doi.org/10.1186/s13059-020-02134-9>
- Ruan G-L, Li Y, Gao Z-X, Wang H-L, Wang W-M (2010) Molecular characterization of trypsinogens and development of trypsinogen gene expression and tryptic activities in grass carp (*Ctenopharyngodon idellus*) and topmouth culter (*Culter alburnus*). *Comp Biochem Physiol B Biochem Mol Biol* 155:77–85. <https://doi.org/10.1016/j.cbpb.2009.10.005>
- Rungruangsak-Torrissen K, Moss R, Andresen LH, Berg A, Waagbø R (2006) Different expressions of trypsin and chymotrypsin in relation to growth in Atlantic salmon (*Salmo salar* L.). *Fish Physiol Biochem* 32(1):7–23. <https://doi.org/10.1007/s10695-005-0630-5>
- Sabapathy U, Teo LH (1993) A quantitative study of some digestive enzymes in the rabbitfish, *Siganus canaliculatus* and the sea bass, *Lates calcarifer*. *J Fish Biol* 42:595–602
- Sæle Ø, Nordgreen A, Olsvik PA, Hamre K (2010) Characterization and expression of digestive neutral lipases during ontogeny of Atlantic cod (*Gadus morhua*). *Comp Biochem Physiol Part A Mol Integr Physiol* 157:252–259. <https://doi.org/10.1016/j.cbpa.2010.07.003>
- Sahaka M, Amara S, Wattanakul J, Gedi MA, Aldai N, Parsieglia G, Lecomte J, Christeller JT, Gray D, Gontero B, Villeneuve P, Carrière F (2020) The digestion of galactolipids and its ubiquitous function in Nature for the uptake of the essential α -linolenic acid. *Food Funct* 11(8):6710–6744. <https://doi.org/10.1039/D0FO01040E>
- Schondube J, Herrera L, Martinez Del Rio C (2001) Diet and the evolution of digestion and renal function in phyllostomid bats. *Zoology* 104:59–73
- Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31(19):3210–3212. <https://doi.org/10.1093/bioinformatics/btv351>
- Skea G, Mountfort D, Clements KD (2005) Gut carbohydrases from the New Zealand marine herbivorous fishes *Kyphosus sydneyanus* (Kyphosidae), *Aplodactylus arcitidens* (Aplodactylidae), and *Odax pullus* (Labridae). *Comp Biochem Physiol B* 140:259–269
- Skea G, Mountfort D, Clements KD (2007) Contrasting digestive strategies in four New Zealand herbivorous fishes as reflected by carbohydrase activity profiles. *Comp Biochem Physiol B* 146:63–70
- Smit AFA, Hubley R, Green P (2015) RepeatMasker Open-4.0. <https://www.repeatmasker.org>
- Soderlund C, Bomhoff M, Nelson WM (2011) SyMAP v3.4 a turnkey synteny system with application to plant genomes. *Nucleic Acids Res* 39(10):68–68. <https://doi.org/10.1093/nar/gkr123>
- Springer M, Weissman JS, Kirschner MW (2010) A general lack of compensation for gene dosage in yeast. *Mol Syst Biol* 6(1):368. <https://doi.org/10.1038/msb.2010.19>
- Stanke M, Schöffmann O, Morgenstern B, Waack S (2006) Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinform* 7(1):62. <https://doi.org/10.1186/1471-2105-7-62>
- Stanke M, Diekhans M, Baertsch R, Haussler D (2008) Using native and syntetically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* 24(5):637–644. <https://doi.org/10.1093/bioinformatics/btn013>
- Stoddard KM (1985) A phylogenetic analysis of some prickleback fishes (Teleostei, Stichaeidae, Xiphisterinae) from the north Pacific Ocean, with a discussion of their biogeography. MS Thesis, Biological Science, California State University Fullerton, Fullerton, CA, USA. 88 pages
- Tang J, Qu F, Tang X, Zhao Q, Wang Y, Zhou Y, Feng J, Lu S, Hou D, Liu Z (2016) Molecular characterization and dietary regulation of aminopeptidase N (APN) in the grass carp (*Ctenopharyngodon idella*). *Gene* 582(1):77–84. <https://doi.org/10.1016/j.gene.2016.01.046>

- Tang S-L, Liang X-F, He S, Li L, Alam MS, Wu J (2022) Comparative Study of the Molecular Characterization, Evolution, and Structure Modeling of Digestive Lipase Genes Reveals the Different Evolutionary Selection Between Mammals and Fishes. *Front Genet.* <https://doi.org/10.3389/fgene.2022.909091>
- Taylor RS, Jensen EL, Coltman DW, Foote AD, Lamichhaney S (2021) Seeing the whole picture: What molecular ecology is gaining from whole genomes. *Mol Ecol* 30(23):5917–5922. <https://doi.org/10.1111/mec.16282>
- Voet D, Voet J (1995) *Biochemistry*. John Wiley and Sons Inc, New York, NY
- Vonk HJ, Western JRH (1984) *Comparative Biochemistry and Physiology of Enzymatic Digestion*. Academic Press, London
- Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, Schatz MC (2017) GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* 33(14):2202–2204. <https://doi.org/10.1093/bioinformatics/btx153>
- Walenz B (2020) A genomic k-mer counter (and sequence utility) with nice features. <https://github.com/marbl/meryl>
- Wang YX, Lu Y, Zhang YG, Ning Z, Li Y, Zhao Q, Lu H, Huang R, Xia X, Feng Q, Liang XF, Liu K, Zhang L, Lu T, Huang T, Fan D, Weng Q, Zhu C, Lu Y, Li W, Wen Z, Zhou C, Tian Q, Kang X, Shi M, Zhang W, Jang S, Du F, He S, Liao L, Li Y, Gui B, He H, Ning Z, Yang C, He L, Luo L, Yang R, Luo Q, Liu X, Li S, Huang W, Xiao L, Lin H, Han B, Zhu Z (2015) The draft genome of the grass carp (*Ctenopharyngodon idella*) provides insights into its evolution and vegetarian adaptation. *Nat Genet* 47(6):625–631. <https://doi.org/10.1038/ng.3280>
- Wiebe PO, Kormish JD, Roper VT, Fujitani Y, Alston NI, Zaret KS, Wright CVE, Stein RW, Gannon M (2007) Ptf1a Binds to and activates area III, a highly conserved region of the Pdx1 promoter that mediates early pancreas-wide Pdx1 expression. *Mol Cell Biol* 27:4093–4104. <https://doi.org/10.1128/mcb.01978-06>
- Ye C, Hill CM, Wu S, Ruan J, Ma Z (2016) DBG2OLC: efficient assembly of large genomes using long erroneous reads of the third generation sequencing technologies. *Sci Rep* 6:31900. <https://doi.org/10.1038/srep31900>
- Yuan Z, Liu S, Zhou T, Tian C, Bao L, Dunham R, Liu Z (2018) Comparative genome analysis of 52 fish species suggests differential associations of repetitive elements with their living aquatic environments. *BMC Genom* 19:141. <https://doi.org/10.1186/s12864-018-4516-1>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.