Genomic and biochemical evidence of dietary adaptation in a marine herbivorous fish

Joseph Heras, Mahul Chakraborty, J. J. Emerson and Donovan P. German

Department of Ecology & Evolutionary Biology, University of California, Irvine, CA 92697-2525, USA

Adopting a new diet is a significant evolutionary change, and can profoundly affect an animal's physiology, biochemistry, ecology and genome. To study this evolutionary transition, we investigated the physiology and genomics of digestion of a derived herbivorous fish, Cebidichthys violaceus. We sequenced and assembled its genome (N50 = 6.7 Mb) and digestive transcriptome, and revealed the molecular changes related to digestive enzymes (carbohydrases, proteases and lipases), finding abundant evidence of molecular adaptation. Specifically, two gene families experienced expansion in copy number and adaptive amino acid substitutions: amylase and carboxyl ester lipase (cel), which are involved in the digestion of carbohydrates and lipids, respectively. Both show elevated levels of gene expression and increased enzyme activity. Because carbohydrates are abundant in the prickleback's diet and lipids are rare, these findings suggest that such dietary specialization involves both exploiting abundant resources and scavenging rare ones, especially essential nutrients, like essential fatty acids.

1. Introduction

Populations exposed to new environments often experience strong selection [1]. Thus, comparing closely related species has been effective in pinpointing changes that drive adaptation [2]. The comparative method can link genetic variation to molecular phenotypes that in turn change animal performance, revealing for example adaptation to new abiotic factors [3,4], changes in diet [5,6] and exposure to pollution [7]. Animal digestion is an ideal model phenotype because it is central to fitness, is understood in many species at genetic, molecular, biochemical and physiological levels and is variable throughout animal evolution [8–11]. Untangling the genetic basis of digestion is contingent on quality genomic resources [12], which have traditionally been lacking in non-model species [13,14]. Advances in genome technology have improved genomic quality, stimulating the production of genome assemblies and catalysing genetic discoveries in non-model organisms [15,16].

The digestive process requires digestive enzymes [8]. In the field of nutritional physiology, two main hypotheses are invoked to explain digestive enzyme activities in relation to ingested substrate concentrations in the animal’s diet (electronic supplementary material, figure S1). The adaptive modulation hypothesis [17] suggests a positive correlation between digestive enzyme activities and ingested quantity of the substrate for those enzymes. Based on economic principles, this hypothesis is well supported in the literature for carbohydrases, as herbivorous and omnivorous animals tend to have elevated amylase activities in their guts [11,18–23], and achieve these activities largely via gene duplications [22]. The nutrient balancing hypothesis [24] suggests that there can be elevated expression of enzymes towards limiting dietary resources to ensure acquisition of essential nutrients, like essential fatty acids. Indeed, elevated carboxyl ester lipase activities are observed in fishes consuming low-lipid, high-fibre foods [11,25,26]. In this investigation, we sequenced the genome of the herbivorous fish Cebidichthys violaceus, which revealed extensive genetic variation and adaptive amino acid

Electronic supplementary material is available online at https://doi.org/10.6084/m9.ﬁgshare.c.4838775.
variation for amylase and carboxyl ester lipase, suggesting multiple mechanisms underlying the novel derived dietary physiology in *C. violaceus*, and simultaneously supporting the adaptive modulation and nutrient balancing hypotheses in the same organism.

We generated a physiological genomics dataset for this non-model species, including a highly contiguous and complete genome, the transcriptomes of digestive and hepatic tissues, and digestive enzyme activity levels. The *C. violaceus* genome had a N50, a measure of genome assembly quality, of 6.7 MB, placing it among the most contiguous teleost assemblies (figure 1). The ecological and evolutionary positions of *C. violaceus* makes these resources important for unravelling the acquisition of herbivory. Members of the family Stichaeidae, including *C. violaceus*, have independently invaded the intertidal zone multiple times, and in two cases began consuming significant amounts of algae (electronic supplementary material, figure S2; [11]). For each section, more detailed methods can be found in the Supplemental Methods in the electronic supplementary material (see electronic supplementary material).

**Figure 1.** Phylogenetic relationships from 13 fishes with sequenced genomes including *Cebidichthys violaceus*. (a) Illustration of *Cebidichthys violaceus*. (b) A maximum-likelihood (ML) tree constructed with 1000 bootstrap replicates in PHYML v3.1 using alignments of concatenated 30 protein coding genes from 13 fish species. *Cebidichthys violaceus* is highlighted in the grey box. Genomes from the following 12 species were acquired from ENSEMBL (release 91): *Oryzias latipes*, *Gasterosteus aculeatus*, *Takifugu rubripes*, *Astyanax mexicanus*, *Poecilia formosa*, *Danio rerio*, *Lepisosteus oculatus* and *Latimeria chalumnae*. Genome size and contig N50, and BUSCO v3 complete genes identified out of 2586 BUSCO groups are represented for all taxa.

<table>
<thead>
<tr>
<th>genome size (MB)</th>
<th>contig N50 (MB)</th>
<th>BUSCO score (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>749</td>
<td>0.057</td>
<td>97.1</td>
</tr>
<tr>
<td>704</td>
<td>9.200</td>
<td>90.9</td>
</tr>
<tr>
<td>744</td>
<td>3.500</td>
<td>94.3</td>
</tr>
<tr>
<td>1010</td>
<td>3.100</td>
<td>98.3</td>
</tr>
<tr>
<td>657</td>
<td>6.690</td>
<td>97.0</td>
</tr>
<tr>
<td>447</td>
<td>0.089</td>
<td>96.6</td>
</tr>
<tr>
<td>342</td>
<td>0.029</td>
<td>87.8</td>
</tr>
<tr>
<td>393</td>
<td>&lt;0.001</td>
<td>96.0</td>
</tr>
<tr>
<td>824</td>
<td>0.002</td>
<td>86.6</td>
</tr>
<tr>
<td>1335</td>
<td>1.800</td>
<td>90.6</td>
</tr>
<tr>
<td>1679</td>
<td>0.850</td>
<td>95.5</td>
</tr>
<tr>
<td>869</td>
<td>0.068</td>
<td>94.3</td>
</tr>
<tr>
<td>2860</td>
<td>0.012</td>
<td>87.4</td>
</tr>
</tbody>
</table>
We used 1.21 g of skin and muscle tissue to extract genomic DNA using a Genomic DNA and RNA purification kit (Macherey-Nagel, Düren, Germany). After extraction, the DNA samples were sheared and separated into large DNA molecules. We used Pacific Biosciences (PacBio) and Illumina platforms for sequencing. For PacBio sequencing, genomic DNA was sized selected with BluePippin with a 15 kb size cut-off, and 40 SMRT cells were sequenced with the PacBio RS II. In addition, from the same gDNA extraction, a multiplex gDNA-Seq Illumina sequencing library was prepared from size selected fragments which ranged from 500 to 700 bp and sequenced on two lanes that resulted in short reads (100 bp paired-end) on an Illumina HiSeq 2500. All genomic sequencing was completed at the University of California, Irvine (UCI) Genomics High-Throughput Facility (GHTF).

We implemented multiple bioinformatic assembly programs to generate our final assembly of the *C. violaceus* genome (electronic supplementary material, figure S3 and tables S2–S4). Sequence data generated from two lanes of Illumina HiSeq 2500 were assembled through PLATANUS v. 1.2.1 [29]. Contigs assembled from PLATANUS and reads from 40 SMRT cells of PacBio sequencing were assembled with a hybrid assembler DR2OLC v. 1.0 [30]. Without Illumina sequence reads, we also conducted a PacBio reads only assembly with FALCON v. 0.3.0 (https://github.com/Pacific-Biosciences/FCALCON). We then used the outputs from FALCON and DR2OLC as input for QUICKMERGE v. 1.0 [31], a metassembler and assembly gap filler developed for long molecule-based assemblies. We then polished the assembly with two rounds of QUERER [32] and then used PHONG v. 1.16 [33] to make final improvements. N25, N50 and N75 was estimated with a perl script (Joseph Fass—http://bioinformatics.ucdavis.edu) from our final *C. violaceus* genome assembly. We processed the final genome assembly through REPEATMASKER v. 4.0.6 [34] and we used BUSCO v. 3 [35] to estimate the completeness of our genome assembly with the Vertebrata and Actinopterygii gene set.

See Supplemental Methods in the electronic supplementary material for information on transcript assembly, annotation and heatmap generation of all tissues and genes associated with diet.

We selected the following candidate genes involved in the digestion of dietary carbohydrates, proteins, and lipids to identify gene copy number and estimate positive diversifying selection: the carbohydrate amylase; the proteolytic enzymes aminopeptidase a, aminopeptidase b, aminopeptidase Ey, aminopeptidase N, aminopeptidase Ey-like, chymotrypsin A, chymotrypsin B, chymotrypsin-like, and trypsin; the lipolytic enzymes phospholipase B1, group XIIb secretory phospholipase A2-like protein, carboxyl ester lipase and carboxyl ester lipase-like enzyme.

To evaluate amylase gene copies, we used previously published variants of *C. violaceus* amylase amyl2a and amyl2b [28] deposited on NCBI (KT920438 and KT920439) to search our assembled genome using both MUMMER v. 3.23 [36] and BLAST [37].
3. Results and discussion

(a) Genome assembly, quality and size

PacBio Single Molecule Real Time sequencing generated approximately 30 Gb long reads (approx. 37X based on genome size = 792 Mb for C. violaceus [48]), whereas the Illumina effort generated 84.5 Gb (approx. 107X) paired end reads. The Illumina-only assembly was highly fragmented (N50 = 2760 bp), consistent with other fish genome assemblies [49] (www.ensembl.org). Using the Illumina contigs in concert with long reads yielded a more contiguous hybrid assembly (N50 = 2.21 Mb), whereas an assembly of the long reads alone yielded similar results (N50 = 2.45 Mb). Merging the hybrid assembly with the long read only assembly [31] (see Methods) yielded a highly contiguous assembly (N50 = 6.69 Mb), ranking it among the most contiguous teleost genomes, and, to our knowledge, the most contiguous among herbivorous fishes (electronic supplementary material, figure S4). The universal single copy orthologues (BUSCOs) [50] show that our assembly (>97%) is comparable to or better than many model reference fish genomes (86.6–98.3%) (figure 1b). Based on an average of four k-mer size counts (k-mer: 25, 27, 29, 31), JELLYFISH estimates the genome size at 656 598 967 base pairs (standard deviation of 4 138 853 base pairs; electronic supplementary material, figure S5 and table S7), which is close to the C. violaceus genome size estimate (792 Mb [48]), and other fish genomes (figure 1b) [49]. We estimate that approximately 21.5% of the genome (128 Mb) is repetitive, with approximately 17% (100 Mb) and 4.5% (27.5 Mb) of the genome occupied by transposable elements and simple repeats, respectively. Only 7.9% of these TE’s could be detected using the fugu TE database, so the majority of the TEs we detect in the C. violaceus genome are likely novel. We also were able to identify about 38 448 repeated loci where the repetitive sequence (period) range from 1 to 1983 and number of repeats identified were 1.8 to 14 140.8 bps (electronic supplementary material, figures S6 and S7).

The genome of C. violaceus is most similar to that of stickleback, Gasterosteus aculeatus (electronic supplementary material, figure S8) and less so with those of Oryzias latipes, Danio rerio and Lepisosteus oculatus (electronic supplementary material, figures S9–S11). Although the synteny of many of the comparisons are rearranged, these comparisons nevertheless show the relative completeness of our draft genome in comparison to these model systems.

(b) Transcriptomics

We constructed a transcriptomics dataset for nine tissues (electronic supplementary material, figures S12–S17) in C. violaceus. Focusing on the liver, pyloric caeca (which includes pancreatic tissue [9,22]), proximal intestine, mid intestine, and spleen (figure 2; electronic supplementary material, tables S8–S12), we identified genes associated with metabolism (figure 2a), and digestion (figure 2a). Like many herbivorous mammals, C. violaceus has an active microbial community in their hindgut that ferments dietary substrates to short chain fatty acids (SCFAs) (electronic supplementary material, figure S2) [11]. SCFAs are absorbed by the host and used to generate ATP in various tissues [8,52]. Because the two most commonly produced SCFAs acetate and propionate are largely metabolized to ketones in the liver, animals reliant on hindgut fermentation tend to have active ketogenic pathways in their tissues [8,51]. Indeed, genes coding for proteins in ketone synthesis and degradation are upregulated in most tissues in C. violaceus, especially in the liver (figure 2h). There are also clear expression patterns for carbohydrate, proteases, and lipases, confirming the suite of enzymes necessary to digest a range of nutrients (figure 2b). Cebidichthys violaceus appears to express four chymotrypsin genes and two trypsin genes (electronic supplementary material, figures S18–S20). This may be consistent with their herbivorous diet, as carnivores (e.g. salmon) may invest more in trypsin expression [53], whereas herbivores (e.g. grass carp) may express more chymotrypsin [54]. Cebidichthys violaceus efficiently digests protein from algae [10]. Fishes appear to possess five aminopeptidase genes that appear to be retained following fish whole genome duplication events (electronic supplementary material, figures S21–S24 and supplemental discussion), and signatures of selection were observed for these genes.

The pyloric caecal tissue of C. violaceus and other pricklebacks is recognized as ‘pancreatic’ because it is sheathed in acinar cells [9] and shows elevated activity levels of pancreatic enzymes [11,22]. However, this tissue only has two differentially expressed genes in comparison to the mid intestine (electronic supplementary material, figure S15), which is a highly absorptive region of the fish gut [55,56]. Although pyloric caeca have been documented as absorptive (i.e. similar function to the mid intestine [57]), the mid intestine is rarely recognized as also having pancreatic function. In fishes, pancreatic tissue can be embedded in the liver (forming a hepatopancreas) or diffuse along the proximal intestine, particularly in fishes with pyloric caeca [58]. Our transcriptomic and biochemical data suggest an absorptive function of the pyloric caeca, and that the acinar cells are distributed at least down to the mid intestine and not restricted to the pyloric caecal region in C. violaceus (electronic supplementary material, figures S16 and S17).

(c) Physiological genomics of digestive enzymes

Digestive enzyme activity levels reveal what substrates are readily digested in an animal’s digestive tract, highlighting the enzyme genes that are potential targets of selection for efficient digestion [8]. Within the field of nutritional physiology, the adaptive modulation hypothesis predicts a match between the amount of an ingested substrate (e.g. starch) and digestive enzyme expression and activity levels to digest such substrates (e.g. amylase) based on economic principles [8]. Pancreatic amylase activity is elevated in the guts of herbivores and omnivores in comparison to carnivores (especially in prickleback fishes [22,59]), matching the higher intake of soluble carbohydrates in these animals [11,21–23,60]. Cebidichthys violaceus has three tandem pancreatic amylase genes: two copies of amyl2a and one copy of amyl2b (figure 3). The three C. violaceus amyl genes in tandem differs from other pricklebacks (and most other fishes for which genetic data are available), which tend to have one or two identical copies of amyl2 (figure 3) [22]. The two amyl2a copies in C. violaceus are supported by three spanning reads, emphasizing the correct assembly of the amyl2a tandem duplicates (electronic supplementary material, figure S25). Each amylase gene is preceded by a 4.3 Kb DNA element encoding a transposase (figure 3; electronic supplementary material, figure S26), hinting at a role of this transposable element (TE) in gene duplications in this region [61,62]. Additionally, the amyl2b gene has a
transcript per Million mapped reads (FPKM). When testing all 11 branches for the seven prickleback taxa, only one branch was under episodic diversifying selection [62]. When testing all 11 branches for the seven prickleback expression of the amylase gene copies because both of these transcription start site and within first introns could modulate the terminal repeat retrotransposon (electronic supplementary

Figure 2. Gene expression profiles in tissues of Cebidichthys violaceus. We used brain, gill, gonads (testes), heart, liver, pyloric caeca (PC), proximal intestine (PI), middle intestine (MI) and spleen tissues from C. violaceus to represent the transcriptome. Only gene expression profiles of liver, PC, PI, MI and spleen are shown. Low to high expression is shown on a gradient scale from violet (darker) to yellow (lighter) respectively. Unit of expression is measured as fragments per kilobase of transcript per Million mapped reads (FPKM). (a) Three heatmaps were generated for transcripts representing glycolytic pathways (blue box), transcripts for enzymes associated with lipid synthesis/glucogenes (green box), and ketone degradation pathway transcripts (red/pink box). (b) Three heatmaps of digestive enzymes were generated which include carbohydrases (blue boxes), proteases (green box) and lipases (red/pink box). (c) A diagram of the C. violaceus gut, which includes the stomach, pyloric caeca, proximal intestine, middle intestine and distal intestine. Asterisk shows candidate genes for identifying positive selection. (Online version in colour.)

2025 bp long interspersed nuclear element (LINE) inside the 2nd intron of the gene. All three copies of amylase gene copies possess an approximately 470 bp fragment of a long terminal repeat retrotransposon (electronic supplementary material, figure S26). Insertion of the TEs proximal to the transcription start site and within first introns could modulate the expression of the amylase gene copies because both of these regions are typically enriched with cis-regulatory elements [62]. When testing all 11 branches for the seven prickleback taxa, only one branch was under episodic diversifying selection (C. violaceus, amy2b; figure 3c). TheAMY2A andAMY2B proteins have different isoelectric points (7.86 versus 6.82 [22]), which hints that they may be active in different parts of the gut, and the transcriptomic data show that amy2a is expressed at a fairly constant level throughout the proximal GI tract (including the pyloric caeca), whereas amy2b is expressed mostly in the mid intestine (figure 2).

Amylolytic activities in the guts of C. violaceus are similar to those in the two species of Xiphister (figure 3a), yet the Xiphister taxa have two copies of amy2a and no amy2b [22], and C. violaceus and X. mucosus digest algal starch with similar efficiencies [10]. Thus, the phenotype of elevated amylase activity can be achieved via increased gene copy number leading to increased expression of the genes or by increased expression of fewer genes with similar performance outcomes at the whole animal level [10,22].

Herbivores consume a food that is simultaneously low in lipid [63], has more plant-derived galactolipids and betaine lipids than animal material [64,65], is high in fibre [66], and fibre binds to lipid, impeding its digestion [67]. Thus, carboxyl ester lipase (cel) [68] represents an important digestive enzyme gene for herbivorous fishes because lipids (especially essential fatty acids) are crucial for survival. In fishes, CEL is the most important lipolytic enzyme due to its broad specificity [68], and ability to hydrolyse galactolipids [69,70] (in fact, fishes lack a pancreatic lipase–colipase system like mammals have, further emphasizing CEL as a key lipase [68–70]). In pricklebacks, lipolytic activities are elevated in herbivores [11,25]. Moreover, Danio rerio fed a high-fibre, low-lipid diet, analogous to a herbivorous diet in the laboratory, had elevated lipase activities in their guts [26]. Thus, it appears that herbivorous pricklebacks, and other herbivorous fishes [11,25,26] invest in lipase expression to ensure lipid digestion from their algal diet, consistent with the Nutrient Balancing Hypothesis, under which animals invest in the synthesis of digestive enzymes to acquire limiting nutrients [24].
In terms of lipid composition, algae have mostly galactolipids, betaine lipids, and phosphatidylcholine in their tissues, whereas animal material mostly contains triacylglycerides; each contains phospholipids in membranes [67,68]. CEL would be the main fish lipase to hydrolyse galactolipids into free fatty acids [65,69,70]. In mammals, omnivores consuming plant material (i.e. more galactolipids) have elevated lipolytic activities towards galactolipids, whereas carnivores do not [71]. This is also true in insects [72]. Galactolipid assimilation has been confirmed in Danio rerio [73]. Here, we see an expansion in cel gene copy number in an animal consuming more galactolipids (figure 4), whereas we did not observe any expansion in gene copy numbers of phospholipases or triacylglycerol lipase (figure 2b; electronic supplementary material, figures S27 and S28).

In the C. violaceus genome, we identified four tandem copies of cel on contig 445 (three to cel-1 and one cel-2; figure 4; electronic supplementary material, figure S29), and a cel-like locus on contig 138 (electronic supplementary material, figure S30). Each of the individual CEL proteins contains the requisite bile salt binding motif [70] (electronic supplementary material, figure S31). Two of the cel-1 genes (cel-1b and cel-1c) are more similar to each other in intron/exon arrangements than either are cel-1a or cel-2, suggesting that cel-1b and cel-1c copies are more recent gene duplications. Interestingly, cel-1a and cel-2 possess two different LINEs in their first and last introns, respectively, contributing to the structural diversification of the two gene copies (electronic supplementary material, figure S29). We see positive episodic selection for cel-1 (figure 4c; electronic supplementary material, table S13).

The cel-1 gene is expressed across tissues, except for brain, whereas cel-2 shows primarily gut expression (figure 2b). The protein sequences of the cel-1 copies are similar (pairwise distance of 0.014–0.054; Poisson corrected). Whereas C. violaceus has four copies of cel (cel-1 and cel-2) in their genome, other fishes for which the genome has been sequenced, and other stichaeids (based on transcriptomes of their relevant tissues), appear to have two cel genes (figure 4b). cel-2 shows evidence of recombination (figure 4c); there is no evidence of positive diversifying selection. Phylogenetic analyses of cel loci show that cel-1 and cel-2 are represented in G. aculeatus, whereas O. latipes and D. rerio do not have a divergent cel-2 locus (electronic supplementary material, figure S30 and table S14).
Cebidichthys violaceus digests algal lipid with consistent efficiency across a range of lipid concentrations, whereas *X. mucosus* shows decreasing lipid digestibility with decreasing dietary lipid content [10], suggesting that CEL diversity may affect lipid digestibility. Close examination of the CEL function in fishes with different diets is clearly warranted, especially using a pH-stat method, which allows for the examination of the hydrolysis of specific types of lipids [71,72].

Genomic scans of humans with lipid-rich diets (Nganasans and Yakuts), shows that they have experienced selection on lipases and proteins involved in lipid metabolism [6], consistent with the adaptive modulation hypothesis [17]. Although lipids compose a proportionally small part of algal mass (less than 10%), galactolipids are a major component of these lipids [64,65]. Thus, on a gross scale, elevated lipolytic activity in the guts of herbivorous fishes aligns with the nutrient balancing hypothesis [24], but on a more detailed level, an enzyme in the guts of herbivorous fishes with different diets is clearly warranted, especially when considering the hydrolysis of specific types of lipids [71,72].

In conclusion, we produced a highly contiguous fish genome, and coupled with a rich literature on the nutritional physiology of *C. violaceus* and other stichaeid fishes, we were able to analyse our genomic data in a nutritional physiology context. Our results show that both, the adaptive modulation hypothesis [17], and the nutrient balancing hypothesis [24], can have genetic underpinnings within the same organism. This powerful physiological genomics approach will provide a model for nutritional physiological research. There is strong interest in using more plant-based aquaculture feeds, including plant lipid. The diversity of cel genes in *C. violaceus* may provide utility in genetically modified aquaculture fishes. Finally, given that *C. violaceus* is commonly found in Marine Protected Areas on the west coast of the United States, and is targeted for aquaculture in northern California where it is a delicacy, our data will also have application for conservation and better culturing techniques for this species.

**Ethics.** All handling of fish from capture to euthanization was conducted under approved protocol no. 2011-2989 of the Institutional Animal Care and Use Committee (IACUC) at the University of California, Irvine.

**Data accessibility.** All sequence data was deposited to NCBI’s GenBank under the bioproject ID no. PRJNA384078. Included under this bioproject are the Illumina genomic and tissue transcriptomic objects under SAMN06857690-SAMN0687699, 40 SMRT cells of PacBio sequencing under SAMN06857690, and the final genome assembly (NJBE00000000).

**Competing interests.** We declare we have no competing interests.

**Funding.** This project was supported by the National Science Foundation grant no. IOS-1355224 (to D.P.G.). The work was supported in part by US National Institutes of Health (NIH) grant R01GM123303 (to J.E.E.).
References


