Genome assembly and annotation for temperate and Antarctic icefishes

This dataset contains the genome assembly and associated annotation for two icefish species (Channichthyidae), the Antarctic mackerel icefish (*Champsocephalus gunnari*) and the temperate pike icefish (*Champsocephalus esox*), alongside a temperate outgroup, the Patagonian Blennie (*Eleginops maclovinus*). It is associated with the folllowing publication:

> Rivera-Colón et al. (2022) **Genomics of secondarily temperate adaptation in the only non-Antarctic icefish**. *bioRxiv*. https://doi.org/10.1101/2022.08.13.503862 (In review at *Molecular Biology & Evolution*)

Abstract

White-blooded Antarctic icefishes are an example of extreme biological specialization to both the chronic cold of the Southern Ocean and life without hemoglobin. As a result, icefishes display derived physiology that limits them to the cold and highly oxygenated Antarctic waters. Against these constraints, remarkably one species, the pike icefish *Champsocephalus esox*, successfully colonized temperate South American waters. To study the genetic mechanisms underlying secondarily temperate adaptation of *C. esox*, we generated chromosome-level genome assemblies of both *C. esox* and its Antarctic sister species, *Champsocephalus gunnari*. The *C. esox* genome is similar in structure and organization to that of its Antarctic congener; however, we observe evidence of chromosomal rearrangements coinciding with regions of elevated genetic divergence in pike icefish populations. We also find several key biological pathways under selection, including genes related to mitochondria and vision, highlighting candidates behind temperate adaptation in *C. esox*. Substantial antifreeze glycoprotein (*AFGP*) pseudogenization has occurred in the pike icefish, likely due to relaxed selection following ancestral escape from Antarctica. The canonical *AFGP* locus organization is conserved in *C. esox* and *C. gunnari*, but both show a translocation of two *AFGP* copies to a separate locus, previously unobserved in cryonotothenioids. Our study presents the first whole genome characterization of a secondarily temperate notothenioid to date, providing a key resource and platform for understanding the adaptive potential of the highly vulnerable icefishes, and of cryonotothenioids in general, in face of a warming Antarctic environment.

Methods

Genomes for *C. esox* and *C. gunnari* were assembled using `flye` v2.5. The *E. maclovinus* genome was assembled with `wtdbg2`. All genomes were scaffolded with `juicer` v1.6.2. Annotation generated using BRAKER v2.1.6 and TSEBRA v1.0.1. A more in depth description is available on the README file and in the publication (Rivera-Colón et al. 2022).

Usage Notes

All files are gzipped, but are otherwise standard bioinformatic formats (i.e., FASTA for genome assembly and coding/amino acid sequences), GTF for annotation, AGP for scaffolding).

See links for a description of the FASTA (http://www.ncbi.nlm.nih.gov/blast/fasta.shtml), and GTF (https://useast.ensembl.org/info/website/upload/gff.html), and AGP (https://www.ncbi.nlm.nih.gov/assembly/agp/AGP_Specification/) file format specifications.

File format specifications

File Suffix*	**Description**
`*.fa`	Genome assembly in nucleotide FASTA format.
`*.agp`	Assembly structure in AGP format.
`*.gtf`	Genome annotation in GTF format.
`*.cds.fa`	Genomic sequence for all annotated protein-coding genes in nucleotide FASTA format.
`*.protein.fa`	Protein sequence for all annotated protein-coding genes in amino acid FASTA format.

> *Does not include the gzipped compression suffix (`.gz`).

C. esox

Assembly and Annotation

The *C. esox* assembly and associated annotation have the label `ceso.ftc.fv8`, which denotes the species (`ceso`), the annotation (`ftc`, `flye`+`TSEBRA` Curated), and integration (`fv8`, `flye` assembly, 8th iteration).

A *C. esox* specimen was collected from the Puerto Natales, Chile in January 2018. For CLR sequencing, HMW DNA was extractd and sequenced using a PacBio Sequel II. After filtering for coverage and read length, raw reads were assembled with `flye` v2.5 (Kolgomorov *et al.* 2019). For scaffolding, Hi-C libraries were constructed and sequenced, after which the genome was scaffolded into super-scaffolds with `juicer` v1.6.2 (Durand *et al.* 2016). PacBio and HiC raw data is available under NCBI BioProject PRJNA857989, BioSample SAMN29660769.

For annotation, RNA was extracted from pooled *C. gunnari* specimens, made into a RNAseq library and sequenced. After alignment, the transcript-level evidence was used to annotate protein-coding genes using `BRAKER` v2.1.6 (Brůna *et al.* 2021), which were then further processed using `TSEBRA` v1.0.1 (Gabriel *et al.* 2021). Using a custom Python script, we curated the `TSEBRA` output to guarantee consistency in the naming of genes and transcripts, as well as incorporating gene names and description based on the corresponding zebrafish orthologs. Manual curation was performed for the annotation of Antifreeze Glycoprotein (AFGP) coding sequences.

A conserved synteny analysis using `synolog` (Catchen *et al.* 2009; Small *et al.* 2016) was employed for the manual curation of the assemblies. For example, we identifying missassemblies in structural variants limited to contig boundaries or merged scaffolds belonging to the same chromosome sequences. We used a custom Python program to propagate these changes through the constituent assembly files.

Files

File	**Description**

File	Description
`ceso.ftc.fv8.fa.gz`	Genome assembly in FASTA format. File contains 2,067 sequences, including 24 chromosome-scale scaffolds. Total size 987.1 Mbp. Contig and scaffold N50 2.6 and 43.6 Mbp, respectively.
`ceso.ftc.fv8.agp.gz`	Assembly structure file in AGP format.
`ceso.ftc.fv8.gtf.gz`	Genome annotation in GTF format.
`ceso.ftc.fv8.cds.fa.gz`	CDS for all annotated protein-coding genes in FASTA format.
`ceso.ftc.fv8.protein.fa.gz`	Amino acid sequences for all annotated protein-coding genes in FASTA format.

C. gunnari

Assembly and Annotation

The *C. gunnari* assembly and associated annotation have the label `cgun.ftc.fv8`, which denotes the species (`cgun`), the annotation (`ftc`, `flye`+`TSEBRA` Curated), and integration (`fv8`, `flye` assembly, 8th iteration).

A male *C. gunnari* specimen was collected from the West Antarctic Peninsula in 2014. HMW DNA was extracted and sequenced using PacBio Sequel II and a Hi-C library, after which genome assembly, scaffolding, and annotation was performed as described above for *C. esox*. PacBio and HiC raw data is available under NCBI BioProject PRJNA857989, BioSample SAMN29660770.

Files

File	Description
`cgun.ftc.fv8.fa.gz`	Genome assembly in FASTA format. File contains 1,536 sequences, including 24 chromosome-scale scaffolds. Total size 994.2 Mbp. Contig and scaffold N50 3.2 and 44.1 Mbp, respectively.
`cgun.ftc.fv8.agp.gz`	Assembly structure file in AGP format.
`cgun.ftc.fv8.gtf.gz`	Genome annotation in GTF format.
`cgun.ftc.fv8.cds.fa.gz`	CDS for all annotated protein-coding genes in FASTA format.
`cgun.ftc.fv8.protein.fa.gz`	Amino acid sequences for all annotated protein-coding genes in FASTA format.

E. maclovinus

Assembly and Annotation

The *E. maclovinus* assembly and associated annotation have the label `emac.rtc.rv5`, which denotes the species (`emac`), the annotation (`rtc`, `redbean`+`TSEBRA` Curated), and integration (`rv8`, `redbean` assembly, 5th iteration).

A *E. maclovinus* specimen was collected from the Puerto Natales, Chile in January 2018. HMW DNA was extracted and sequenced using PacBio Sequel II and a Hi-C library. A contig-level genome assembly was first generated using `wtdgb2` (*a.k.a.* `redbean`) v2.5 (Ruan & Li 2020), and scaffolded with `juicer` v1.6.2. PacBio and HiC raw data is available under NCBI BioProject PRJNA857989, BioSample SAMNXXXXXXXX. For annotation, the RNA-seq data generated by Bilyk et al. (2018) was aligned to the genome, and processed using BRAKER v2.1.6 and TSEBRA v1.0.1. Both the assembly and annotations were curated as described above for *C. esox*.

Files

File	Description
`emac.rtc.rv5.fa.gz`	Genome assembly in FASTA format. File contains 26sequences, including 24 chromosome-scale scaffolds. Total size 606.3 Mbp. Contig and scaffold N50 7.6 and 26.7 Mbp, respectively.
`emac.rtc.rv5.agp.gz`	Assembly structure file in AGP format.
`emac.rtc.rv5.gtf.gz`	Genome annotation in GTF format.
`emac.rtc.rv5.cds.fa.gz`	CDS for all annotated protein-coding genes in FASTA format.
`emac.rtc.rv5.protein.fa.gz`	Amino acid sequences for all annotated protein-coding genes in FASTA format.

Code Availability

All the scripts and custom code used to generate these assemblies are available in a BitBucket repository.

Authors

Angel G. Rivera-Colon
Department of Evolution, Ecology, and Behavior
University of Illinois at Urbana-Champaign
angelgr2@illinois.edu

Julian M. Catchen
Department of Evolution, Ecology, and Behavior
University of Illinois at Urbana-Champaign
jcatchen@illinois.edu

C-H Christina Cheng
Department of Evolution, Ecology, and Behavior
University of Illinois at Urbana-Champaign
c-cheng@illinois.edu