



EOSC-Life Demonstrator Project: Marine Eukaryote Genomics Portal

A genome annotation platform for small pelagic fishes

**Cymon J. Cox*, Bruno Louro, Gianluca De Moro, and Adelino V.M.
Canário**

Centro de Ciências do Mar (CCMAR), Universidade do Algarve

11th October 2018

Overview

- A demonstrator project for the EOSC-Life project
- Implement a genome annotation platform for small pelagic fishes - Clupeiforms: 400 species, 20 of which provide third of fisheries catches worldwide
- provide access to annotation tools, and a novel reciprocal annotation tool to exchange annotations between taxa



European pilchard - **Sardina pilchardus**



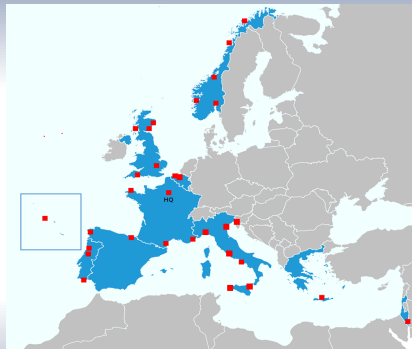
Allis shad - **Alosa alosa**



Atlantic herring - **Clupea harengus**

European Marine Biological Resource Centre (EMBRC)

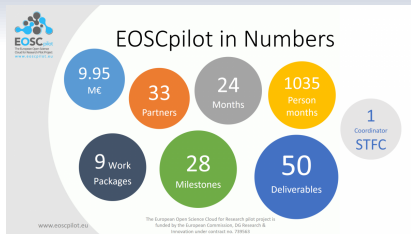
The EMBRC is a pan-European Research Infrastructure for marine biology and ecology research. With its services, it aims to answer fundamental questions regarding the health of oceanic ecosystems in a changing environment, enable new technologies to further our investigation capabilities, support life-science breakthrough discoveries with the use of marine biological models, and continue long-term marine monitoring efforts. EMBRC is a driver in the development of blue biotechnologies, supporting both fundamental and applied research activities for sustainable solutions in the food, health and environmental sectors



CCMAR (Algarve), ACOI (Coimbra), CIIMAR (Porto), and IMAR (Açores)

ESOC-Life: INFRAEOSC-04-2018

- A Research and Innovation Action (RIA) of H2020
- **“Providing an open collaborative space for digital biology in Europe“**
- Call for Demonstrator projects in 2018 to be completed in 2019



Demonstrator Proposal

Platform designed to address the fragmented research landscape for genome annotation of marine organisms

- focus for post-assembly genomic workflows and data access
- portal for access to services such as EMBRIC Configurator, Elixir ontologies, and meta-data standards
- community-driven annotation platform for marine eukaryotes (initially small pelagic fishes)

EMBRIC Configurator Service

Entry point into data resources in molecular biology. Help to design new marine projects with a project-specific informatics configuration which includes a description of the elements of infrastructure (such as databases, standards, formats, curation groups, analysis methods, and cloud compute capacity), and advice on accessing and setting these elements up for the project and data management guidelines. Funding ends June 2019.



The banner features the EMBRIC logo (European Marine Biological Research Infrastructure Cluster) at the top left. Below it are three circular images: a rack of laboratory flasks, a blue fish, and a green plant. The word "Configurator" is written in white on a dark blue background on the right. Below the images, the text reads: "Do you need help with your marine data planning, management & interpretation?" followed by a small network icon.

EMBRIC: European Marine Biological Research Infrastructure Cluster

Draft genome assembly of the sardine

Statistics

- 759 million 150bp paired-end Illumina reads using Chromium system
- Estimated haploid genome size of 949.62Mb (1.43% heterozygosity)
- Estimated genome completeness: 83-92%



Article in submission

A haplotype-resolved draft genome of the European sardine (***Sardina pilchardus***). *Bruno Louro, Gianluca De Moro, Carlos Garcia, Cymon J. Cox, Ana Veríssimo, Stephen J. Sabatino, António M. Santos, and Adelino V. M. Canário*

Annotating the sardine genome

Statistics

- **ab initio** gene prediction, protein homology (InterProScan, NCBI BlastX), 12X RNA-Seq transcriptome assembly
- 30,169 gene models and 17,199 functional annotations 57%
- at least 12,970 gene models remain unannotated and **all** need manual verification and curation



Annotating the sardine genome

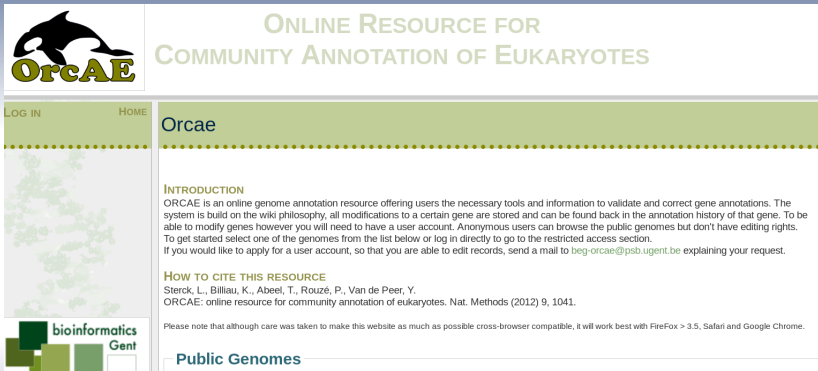
Statistics

- **ab initio** gene prediction, protein homology (InterProScan, NCBI BlastX), 12X RNA-Seq transcriptome assembly
- 30,169 gene models and 17,199 functional annotations 57%
- at least 12,970 gene models remain unannotated and **all** need manual verification and curation



So what happens to the other 43% of predicted genes that remain unannotated?

Community-driven annotation platforms



The screenshot shows the Orcae website interface. At the top left is the Orcae logo, which features a stylized orca silhouette above the text "Orcae". To the right of the logo, the text "ONLINE RESOURCE FOR COMMUNITY ANNOTATION OF EUKARYOTES" is displayed in a light green font. Below the logo, there are two buttons: "LOG IN" and "HOME". The main content area is titled "Orcae" in a large, bold, light green font. Underneath the title, there is an "INTRODUCTION" section followed by a paragraph of text. Below that is a "HOW TO CITE THIS RESOURCE" section with a citation. At the bottom of the page, there is a "Public Genomes" section with a search bar.

Orcae

INTRODUCTION

ORCAE is an online genome annotation resource offering users the necessary tools and information to validate and correct gene annotations. The system is built on the wiki philosophy, all modifications to a certain gene are stored and can be found back in the annotation history of that gene. To be able to modify genes however you will need to have a user account. Anonymous users can browse the public genomes but don't have editing rights. To get started select one of the genomes from the list below or log in directly to go to the restricted access section. If you would like to apply for a user account, so that you are able to edit records, send a mail to beg-orcae@psb.ugent.be explaining your request.

HOW TO CITE THIS RESOURCE

Sterck, L., Billiau, K., Abeel, T., Rouzé, P., Van de Peer, Y.
ORCAE: online resource for community annotation of eukaryotes. *Nat. Methods* (2012) 9, 1041.

Please note that although care was taken to make this website as much as possible cross-browser compatible, it will work best with Firefox > 3.5, Safari and Google Chrome.

Public Genomes

Sardine Annotation Platform

Sardina pilchardus

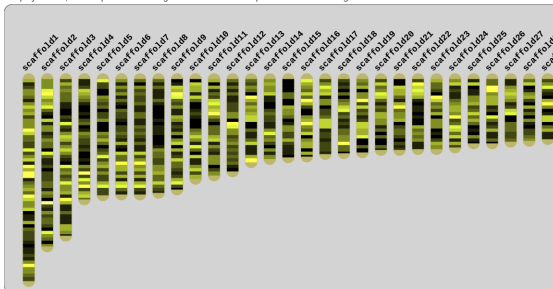


Navigation

▀ BLAST ▀ SEARCH ▀ WIKI ▀ DOWNLOAD ▀ HELP

Browse

The brighter the color, the higher the gene-density in that region. Click on a region to go to that location in the browser. Only contigs larger than 10Kb are displayed here, the complete list of contigs is available in the dropdown menu from the genome browser.



<http://bioinformatics.psb.ugent.be/orcae/overview/Spil>

Sardine Annotation Platform

The screenshot displays the Sardine Annotation Platform interface. On the left, a 'TRACK SELECTION' panel is visible, containing a tree view of track categories: 'Active Tracks' (Gene Models), 'Inactive Tracks' (Annotation models), and 'Messenger RNA' (Transcript sequences). The 'Gene Models' track is currently active. The main 'TRACKS' panel shows the selected track 'Gene Models' for the organism 'ORCAE: Sardina pilchardus'. The interface includes a search bar with 'Splil_000001g0067.1' entered, a 'Dragmode: browse' dropdown, a '60:1' zoom level, and a 'scaffold1' dropdown. The main visualization area shows a genomic track with two gene models: 'Splil_000001g0066.1' and 'Splil_000001g0067.1'. The 'Splil_000001g0067.1' model is highlighted with a red arrow. A coordinate '4340' is visible on the left side of the track.

Sardine Annotation Platform

Sardina pilchardus



Showing region from base-position 1865428 to 2109484 (244.1 Kb)



Gene ID	Spil_000001g0067.1
Locus	Spil_000001g0067.1
Functional Description	n/a
Gene Type	protein-coding gene
Contig	scaffold1
Last Modified On	27 June 2018 18h00
History	No history available v

Sardine Annotation Platform

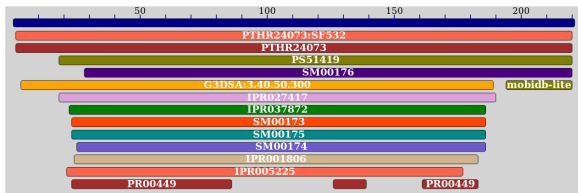
Gene Ontology ⊕

Top

Cellular Component	n/a
Molecular Function	1. GO:0003924 GTPase activity 2. GO:0005525 GTP binding
Biological Process	n/a

Protein Domains ⊕

Top



Domain ID	Description	Database
SM00173	n/a	SMART
IPR037872	Rab3	InterPro
IPR027417	P-loop containing nucleoside triphosphate hydrolase	InterPro
mobidb-lite	consensus disorder prediction	MobiDBLite
SM00175	n/a	SMART
SM00176	n/a	SMART
SM00174	n/a	SMART
G3DSA:3.40.50.300	n/a	Gene3D
IPR001806	Small GTPase	InterPro
IPR005225	Small GTP-binding protein domain	InterPro
PS51419	small GTPase Rab1 family profile.	ProSiteProfiles
PR00449	Transforming protein P21 ras signature	PRINTS
PTHR24073:SF532	n/a	PANTHER
PTHR24073	n/a	PANTHER

Sardine Annotation Platform

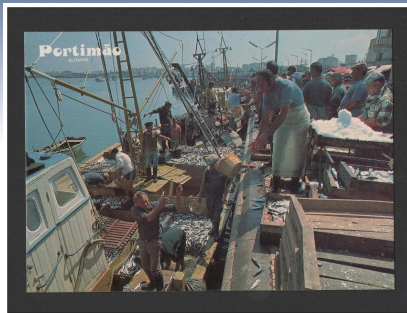
Best10	NCBI	Self	SwissP		
ProteinID	Description / BlastScore			Database	Actions
XP_018601460	PREDICTED: ras-related protein Rab-3A [<i>Scleropages formosus</i>] Value: 8.73e-153 Bitscore: 432 Aln-length = 220, Identities = 92%, Positives = 95%			NCBI	SHOW BLAST
KPP79140	ras-related protein Rab-3A-like [<i>Scleropages formosus</i>] Value: 1.3e-152 Bitscore: 432 Aln-length = 220, Identities = 92%, Positives = 95%			NCBI	SHOW BLAST
XP_023699494	ras-related protein Rab-3A [<i>Paramormyrops kingsleyae</i>] Value: 1.78e-152 Bitscore: 431 Aln-length = 220, Identities = 92%, Positives = 95%			NCBI	SHOW BLAST
Spil_001841g0001.1	(251) ;mRNA; r:8411-17681 Value: 5.01e-155 Bitscore: 429 Aln-length = 220, Identities = 92%, Positives = 94%			Self	SHOW BLAST
XP_012683307	PREDICTED: ras-related protein Rab-3A [<i>Clupea harengus</i>] Value: 5.4e-152 Bitscore: 429 Aln-length = 220, Identities = 92%, Positives = 95%			NCBI	SHOW BLAST
XP_012683691	PREDICTED: ras-related protein Rab-3A [<i>Clupea harengus</i>] Value: 1.37e-151 Bitscore: 429 Aln-length = 220, Identities = 92%, Positives = 94%			NCBI	SHOW BLAST
XP_017542287	PREDICTED: ras-related protein Rab-3A-like [<i>Pygocentrus nattereri</i>] Value: 1.75e-151 Bitscore: 428 Aln-length = 220, Identities = 91%, Positives = 95%			NCBI	SHOW BLAST
XP_017333842	PREDICTED: ras-related protein Rab-3A [<i>Ictalurus punctatus</i>] Value: 2.01e-151 Bitscore: 428 Aln-length = 220, Identities = 92%, Positives = 94%			NCBI	SHOW BLAST
AWP04288	putative ras-related protein Rab-3A-like [<i>Scophthalmus maximus</i>] Value: 6.49e-151 Bitscore: 427 Aln-length = 220, Identities = 92%, Positives = 95%			NCBI	SHOW BLAST
XP_016331917	PREDICTED: ras-related protein Rab-3A [<i>Sinocyclocheilus anshuiensis</i>] Value: 6.96e-151 Bitscore: 427 Aln-length = 220, Identities = 92%, Positives = 94%			NCBI	SHOW BLAST

Demonstrator Project: A genome annotation platform for small pelagic fishes

Novelty of platform

- dedicated instance of Orcae for small pelagic fishes (Culpidea) in the EOSC
- transfer annotation between closely related small pelagic fishes
- automated system for suggesting new annotations based on reciprocal genome analysis

Acknowledgements

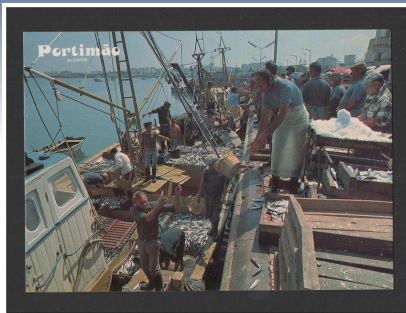


gustaf_collectors

www.delcampe.net

- FCT - project UID/Multi/04326/2013
- INCD funded by FCT and FEDER under project 22153-01/SAICT/2016
- BioData.pt consortium
- EMBRIC Configurator service at EMBL/EBI
- Mario David and João Pina at LIP for access to INCD/Ingrid

Acknowledgements



luis8an_collections

www.delcampe.net



- FCT - project UID/Multi/04326/2013
- INCD funded by FCT and FEDER under project 22153-01/SAICT/2016
- BioData.pt consortium
- EMBRIC Configurator service at EMBL/EBI
- Mario David and João Pina at LIP for access to INCD/Ingrid