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

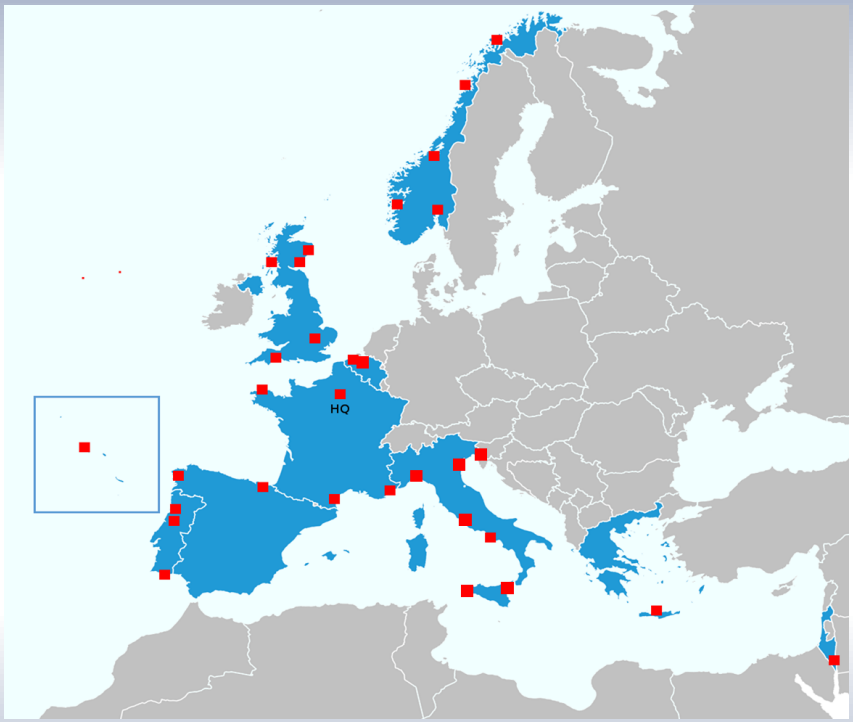
European pilchard - **Sardina pilchardus**

Allis shad - **Alosa alosa**

Atlantic herring - **Culpea harengus**

European Marine Biological Resource Centre (EMBRC)

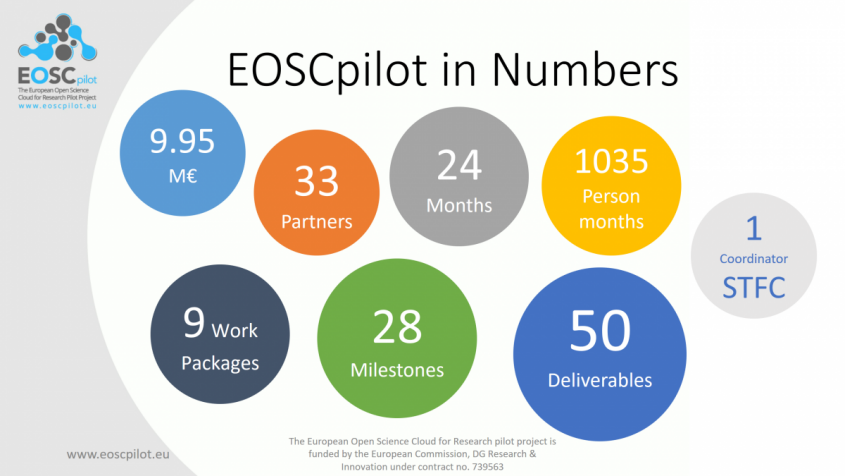
The EMBRC is a pan-European Research Infrastructure for marine biology and ecology research. With its services, it aims to answer fundamental questions regarding the health of oceanic ecosystems in a changing environment, enable new technologies to further our investigation capabilities, support life-science 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.

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

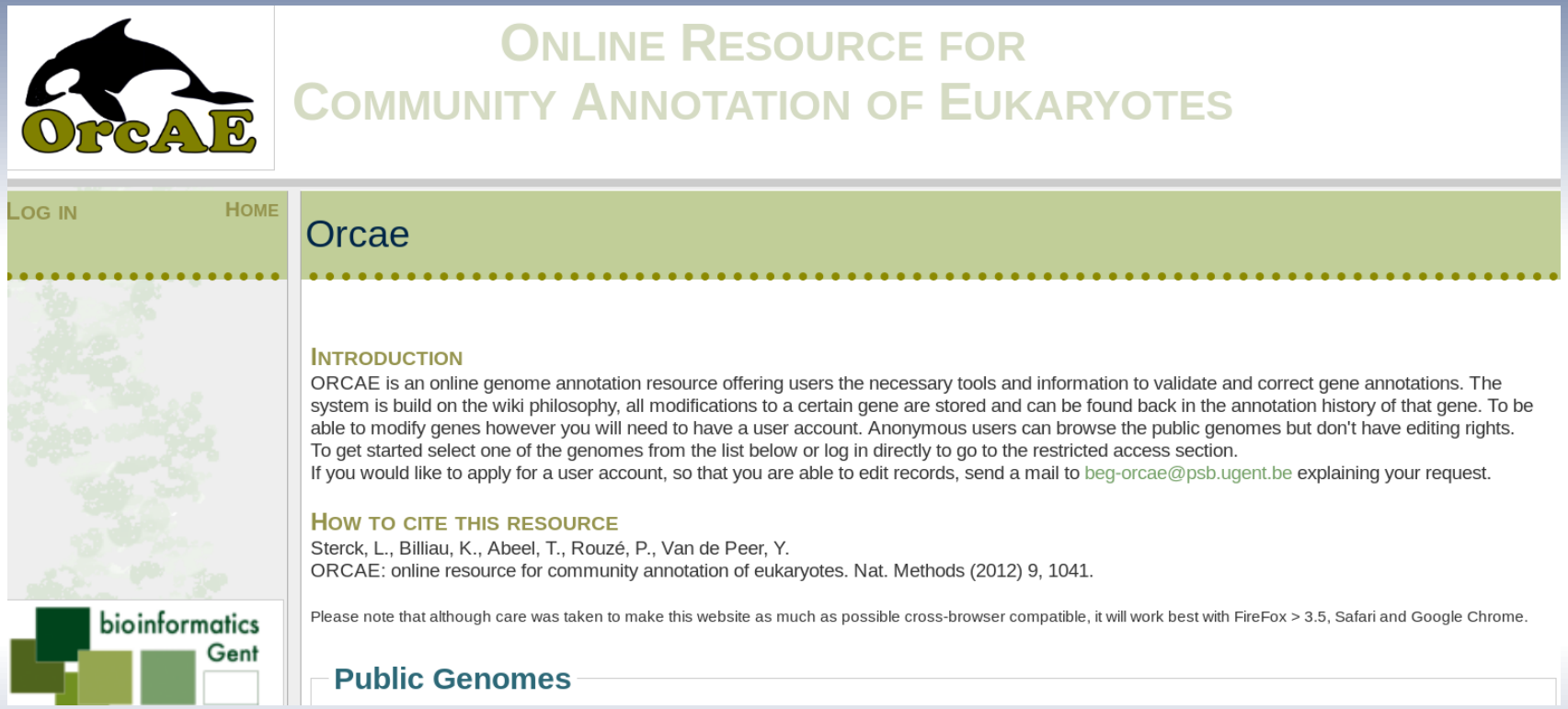
Annotating the sardine genome

Statistics

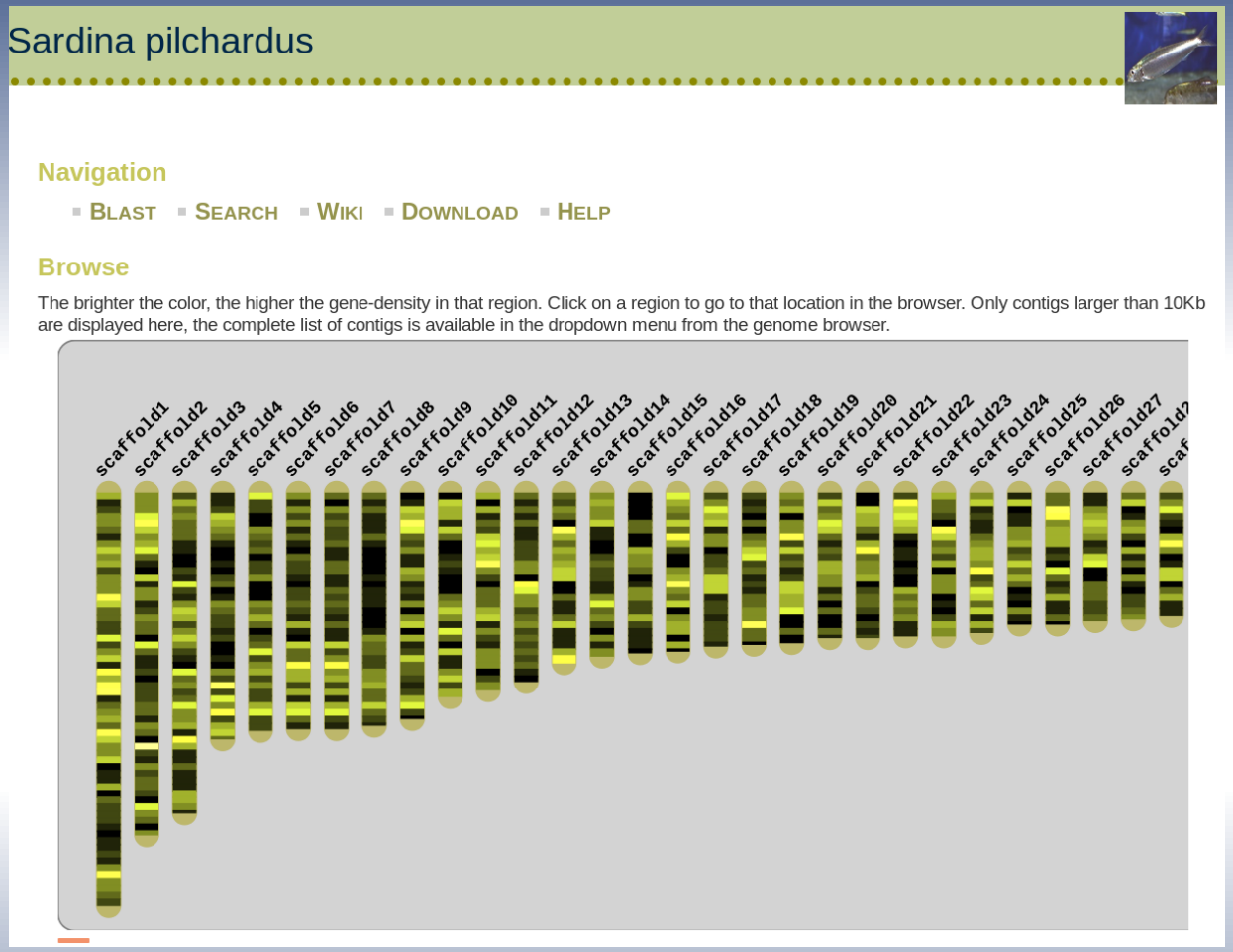
- **ab initio** gene prediction, protein homology (InterProScan, NCBI BlastX), 12X RNA-Seq transcriptome assembly
- 30,169 gene models and 17,199 functional annotations 57%
- at least 12,970 gene models remain unannotated and **all** need manual verification and curation

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

Community-driven annotation platforms



Sardine Annotation Platform



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

Sardine Annotation Platform

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

Sardine Annotation Platform

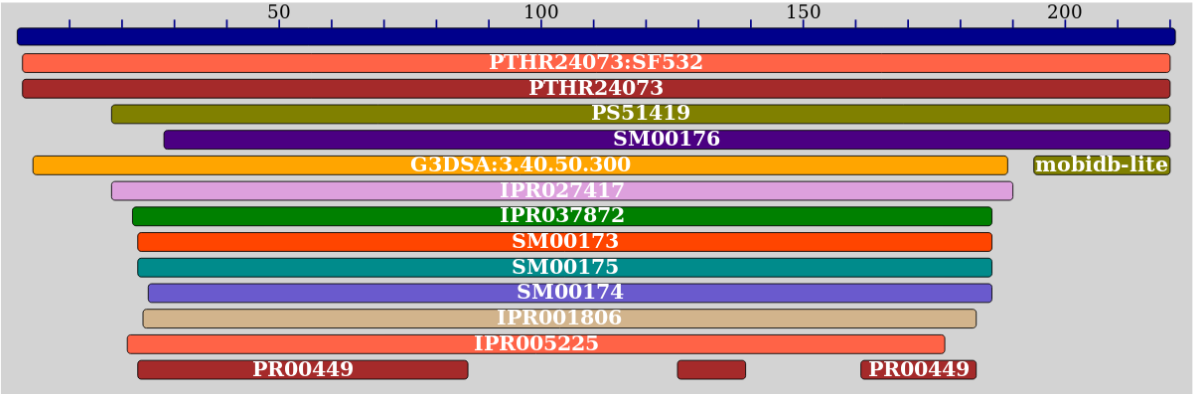
Gene Ontology

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

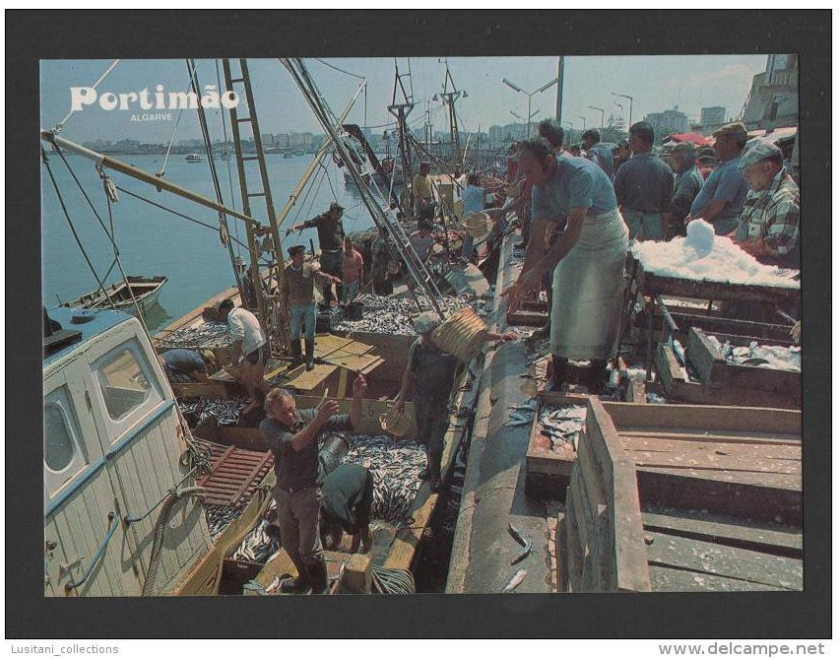
ProteinID	Description / BlastScore	Database	Actions
XP_018601460	PREDICTED: ras-related protein Rab-3A [Scleropages formosus] Evalue: **8.73e-153** | Bitscore: 432 Aln-length = 220, Identities = 92%, Positives = 95%	NCBI	Show Blast
KPP79140	ras-related protein Rab-3A-like [Scleropages formosus] Evalue: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

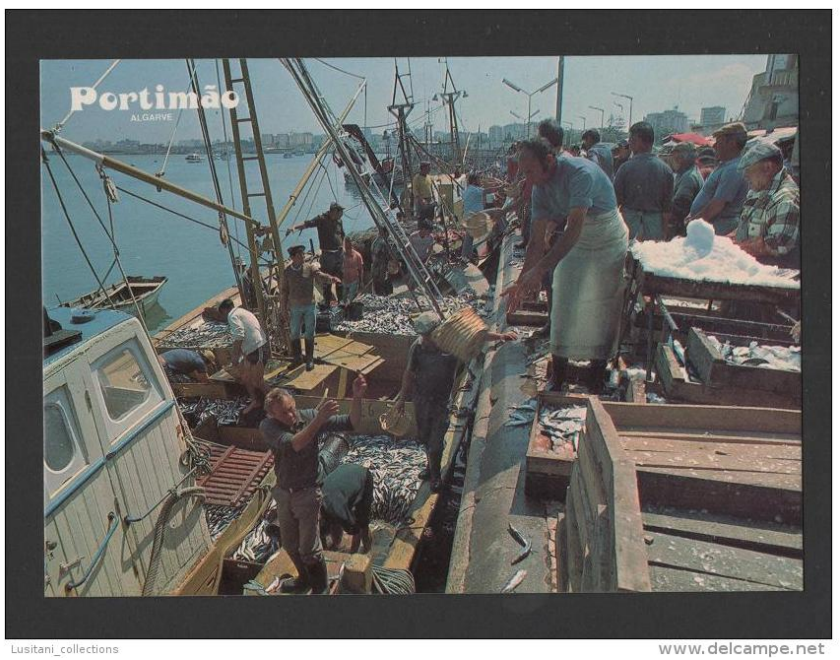
- dedicated instance of Orcae for small pelagic fishes (Culpidea) in the EOSC
- transfer annotation between closely related small pelagic fishes
- automated system for suggesting new annotations based on reciprocal genome analysis

Acknowledgements



* 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



* 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