JOBIM 2023
27>30 JUIN
Édition multisite

BEAURIS

An automated, modular, and FAIR system for large-scale genome data management

Matéo BOUDET, Loraine BRILLET-GUÉGUEN, Arthur LE BARS, Karine MASSAU, Laura LEROI,
Alexandre CORMIER, Patrick DURAND, Erwan CORRE and Anthony BRETAUDEAU

BIPAA
BioInformatics Platform for Agroecosystem Arthropods

GenOuest

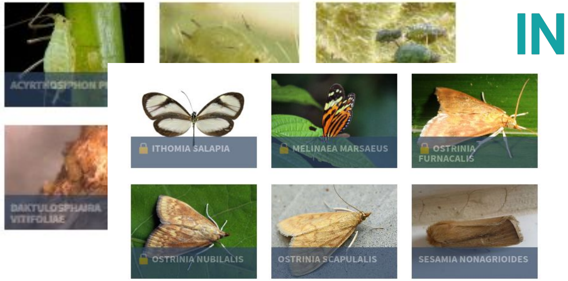
ABiMS

SeBiMER
Bioinformatique marine

Challenges: how to value the mass of genomic data?

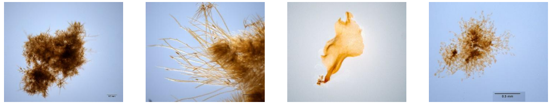
Increasingly high volume & diversity of genomic data produced

aphids, parasitoïd wasps, lepidoptera, Brassicaceae…



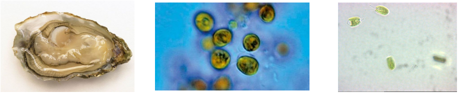
INRAE

brown algae, red algae, fungi…



CNRS • SORBONNE UNIVERSITÉ
Station Biologique
de Roscoff

Vibrio, Cassostrea, Tisochrysis…

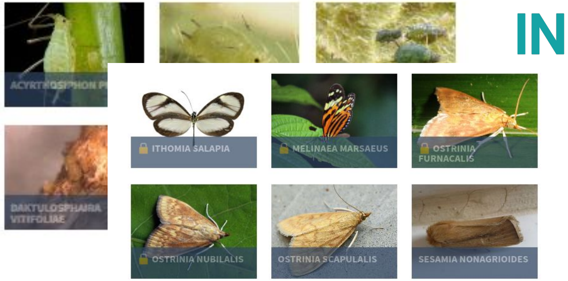


Ifremer

Challenges: how to value the mass of genomic data?

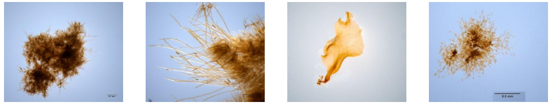
Increasingly high volume & diversity of genomic data produced

aphids, parasitoïd wasps, lepidoptera, Brassicaceae…



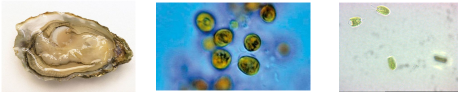
INRAE

brown algae, red algae, fungi…

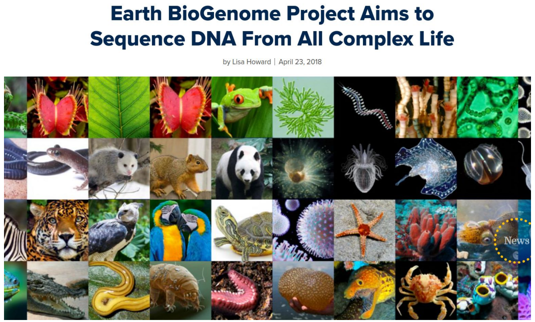


CNRS • SORBONNE UNIVERSITÉ
Station Biologique de Roscoff

Vibrio, Cassostrea, Tisochrysis…



Ifremer



Earth BioGenome Project Aims to Sequence DNA From All Complex Life

by Lisa Howard | April 23, 2018

i5k

OUR MISSION

Sequencing reference-quality genomes for all European species

VGP
VERTEBRATE GENOMES PROJECT
A PROJECT OF THE G10K CONSORTIUM

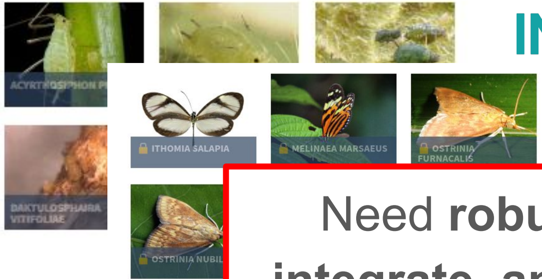
Darwin TREE of LIFE

Sequencing the genomes of 70,000 plants, fungi, animals and protists in Britain and Ireland

Challenges: how to value the mass of genomic data?

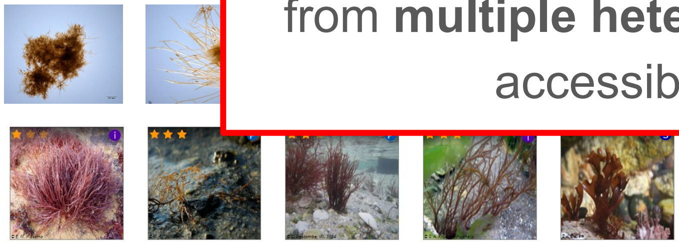
Increasingly high volume & diversity of genomic data produced

aphids, parasitoïd wasps, lepidoptera, Brassicaceae…

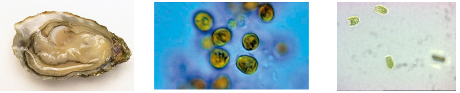


INRAE

brown algae, red algae

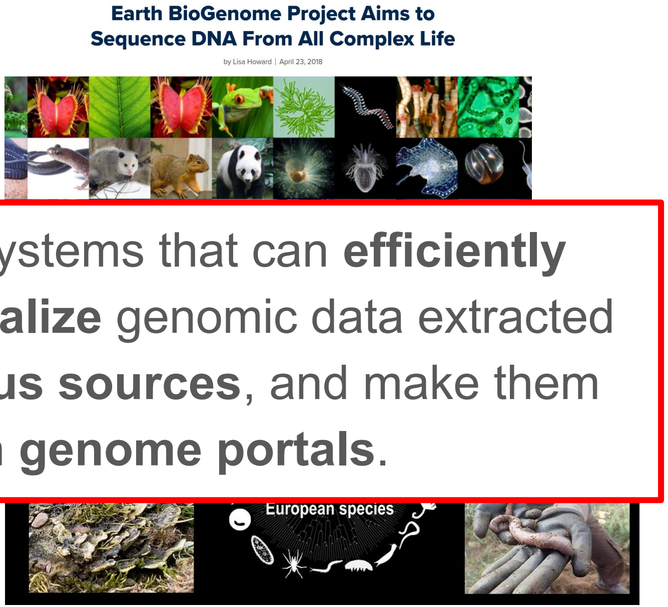


Vibrio, Cassostrea, Tisochrysis…



Ifremer

Need **robust** information systems that can **efficiently integrate**, **analyze**, and **visualize** genomic data extracted from **multiple heterogeneous sources**, and make them accessible through **genome portals**.



Earth BioGenome Project Aims to Sequence DNA From All Complex Life

by Lisa Howard | April 23, 2018

i5k

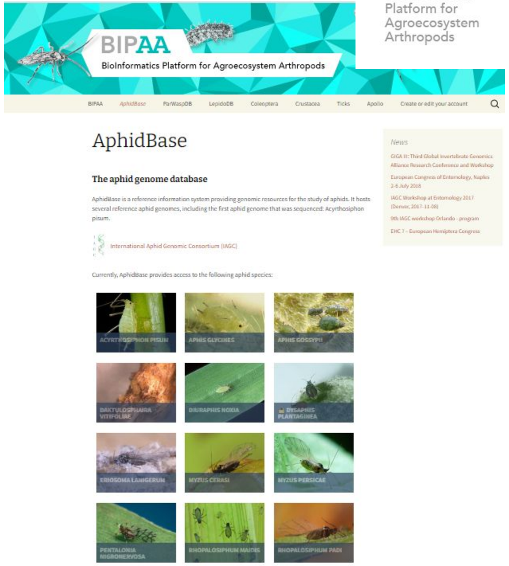
VGP VERTEBRATE GENOMES PROJECT

A PROJECT OF THE G10K CONSORTIUM

Darwin Tree of Life

Sequencing the genomes of 70,000 plants, fungi, animals and protists in Britain and Ireland

Challenges: deploy thematic genome portals



BIPAA
BioInformatics Platform for Agroecosystem Arthropods

AphidBase

The aphid genome database

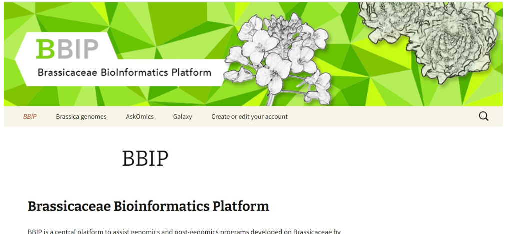
AphidBase is a reference information system providing genomic resources for the study of aphids. It hosts several reference aphid genomes, including the first aphid genome that was sequenced: Acyrthosiphon pisum.

International Aphid Genomic Consortium (IAGC)

Currently, AphidBase provides access to the following aphid species:

News

- GIGA III: Third Global Invertebrate Genomics Alliance Research Conference and Workshop
- European Congress of Entomology, Naples 2-6 July 2018
- IAGC Workshop at Entomology 2017 (Denver, 2017-11-08)
- 9th IAGC workshop Orlando - program
- EHC 7 – European Hemiptera Congress

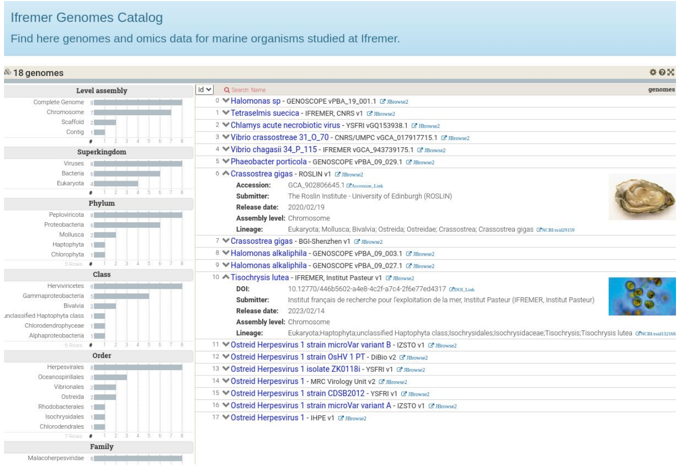


BBIP

Brassicaceae Bioinformatics Platform

BBIP is a central platform to assist genomics and post-genomics programs developed on Brassicaceae by

SeBiMER
Bioinformatique marine



Ifremer Genomes Catalog

Find here genomes and omics data for marine organisms studied at Ifremer.

18 genomes

Level assembly
Complete Genome 8
Chromosome 7
Scaffold 2
Contig 1

Superkingdom
Viruses 8
Bacteria 6
Eukaryota 4

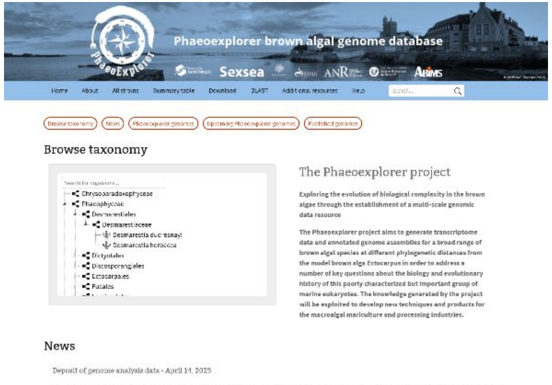
Phylum
Peploviricota 8
Proteobacteria 6
Mollusca 2
Haptophyta 1
Chlorophyta 1
5 Rows

Class
Herviviricetes 8
Gammaproteobacteria 5
Bivalvia 2
unclassified Haptophyta class 1
Chlorodendrophyceae 1
Alphaproteobacteria 1
6 Rows

Order
Herpesvirales 8
Oceanospirillales 3
Vibrionales 2
Ostreida 2
Rhodobacterales 1
Isochrysidales 1
Chlorodendrales 1
7 Rows

Family
Malacoherpesviridae 8

0 Halomonas sp - GENOSCOPE vPBA_19_001.1 JBrowse2
1 Tetraselmis suecica - IFREMER, CNRS v1 JBrowse2
2 Chlamys acute necrobiotic virus - YSFRI vGQ153938.1 JBrowse2
3 Vibrio crassostreae 31_O_70 - CNRS/UMPC vGCA_017917715.1 JBrowse2
4 Vibrio chagasii 34_P_115 - IFREMER vGCA_943739175.1 JBrowse2
5 Phaeobacter porticola - GENOSCOPE vPBA_09_029.1 JBrowse2
6 Crassostrea gigas - ROSLIN v1 JBrowse2
Accession: GCA_902806645.1 Accession_Link
Submitter: The Roslin Institute - University of Edinburgh (ROSLIN)
Release date: 2020/02/19
Assembly level: Chromosome
Lineage: Eukaryota; Mollusca; Bivalvia; Ostreida; Ostreidae; Crassostrea; Crassostrea gigas NCBI:txid29159
7 Crassostrea gigas - BGI-Shenzhen v1 JBrowse2
8 Halomonas alkaliphila - GENOSCOPE vPBA_09_003.1 JBrowse2
9 Halomonas alkaliphila - GENOSCOPE vPBA_09_027.1 JBrowse2
10 Tisochrysis lutea - IFREMER, Institut Pasteur v1 JBrowse2
DOI: 10.12770/446b5602-a4e8-4c2f-a7c4-2f6e77ed4317 DOI_Link
Submitter: Institut français de recherche pour l'exploitation de la mer, Institut Pasteur (IFREMER, Institut Pasteur)
Release date: 2023/02/14
Assembly level: Chromosome
Lineage: Eukaryota;Haptophyta;unclassified Haptophyta class;Isochrysidales;Isochrysidaceae;Tisochrysis;Tisochrysis lutea NCBI:txid132166
11 Ostreid Herpesvirus 1 strain microVar variant B - IZSTO v1 JBrowse2
12 Ostreid Herpesvirus 1 strain OsHV 1 PT - DiBio v2 JBrowse2
13 Ostreid Herpesvirus 1 isolate ZK0118i - YSFRI v1 JBrowse2
14 Ostreid Herpesvirus 1 - MRC Virology Unit v2 JBrowse2
15 Ostreid Herpesvirus 1 strain CDSB2012 - YSFRI v1 JBrowse2
16 Ostreid Herpesvirus 1 strain microVar variant A - IZSTO v1 JBrowse2
17 Ostreid Herpesvirus 1 - IHPE v1 JBrowse2



Browse taxonomy

The Phaeoexplorer project

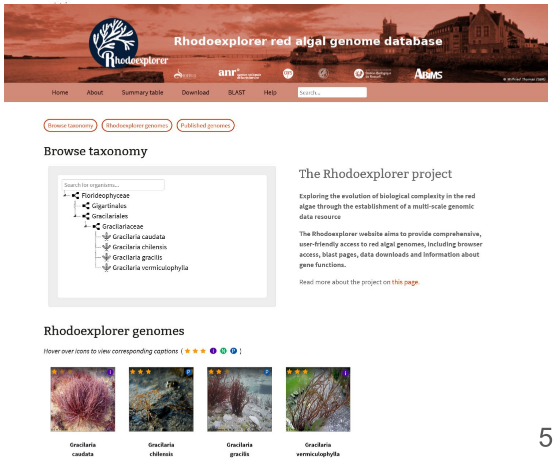
Exploring the evolution of biological complexity in the brown algae through the establishment of a multi-scale genomic data resource

The Phaeoexplorer project aims to generate transcriptome data and annotated genome assemblies for a broad range of brown algal species at different phylogenetic distances from the model brown alga Ectocarpus in order to address a number of key questions about the biology and evolutionary history of this poorly characterized but important group of marine eukaryotes. The knowledge generated by the project will be exploited to develop new techniques and products for the macroalgal mariculture and processing industries.

News

Deposit of genome analysis data - April 14, 2023

Message to the consortium: a web page has been created to help you deposit the results of your analyses on the Roscoff server (tabs: Help, Deposit your



Browse taxonomy

The Rhodoexplorer project

Exploring the evolution of biological complexity in the red algae through the establishment of a multi-scale genomic data resource

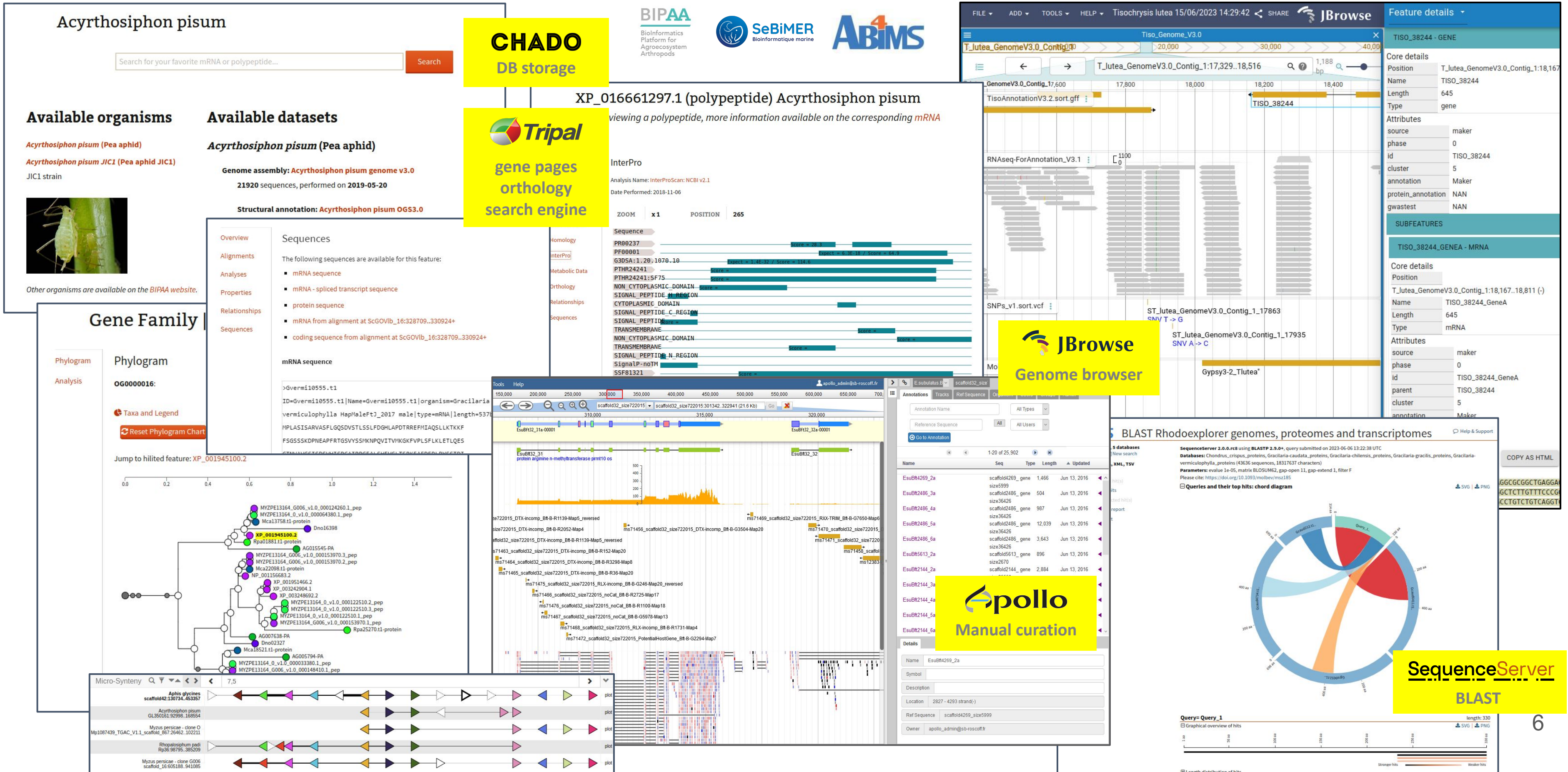
The Rhodoexplorer website aims to provide comprehensive, user-friendly access to red algal genomes, including browser access, blast pages, data downloads and information about gene functions.

Read more about the project on this page.

Rhodoexplorer genomes

Hover over icons to view corresponding captions (★ ★ ★ i N P)

Challenges: deploy visualisation, analysis and annotation tools



CHADO
DB storage

Tripal
gene pages
orthology
search engine

JBrowse
Genome browser

Apollo
Manual curation

SequenceServer
BLAST

Acyrthosiphon pisum

Available organisms

Acyrthosiphon pisum (Pea aphid)

Acyrthosiphon pisum JIC1 (Pea aphid JIC1)

JIC1 strain

Other organisms are available on the BIPAA website.

Available datasets

Acyrthosiphon pisum (Pea aphid)

Genome assembly: Acyrthosiphon pisum genome v3.0

21920 sequences, performed on 2019-05-20

Structural annotation: Acyrthosiphon pisum OGS3.0

XP_016661297.1 (polypeptide) Acyrthosiphon pisum

viewing a polypeptide, more information available on the corresponding mRNA

InterPro

Analysis Name: InterProScan: NCBI v2.1

Date Performed: 2018-11-06

Sequences

The following sequences are available for this feature:

- mRNA sequence
- mRNA - spliced transcript sequence
- protein sequence
- mRNA from alignment at ScGOVlb_16:328709..330924+
- coding sequence from alignment at ScGOVlb_16:328709..330924+

mRNA sequence

Gene Family

Phylogram

OG0000016:

Jump to hilited feature: XP_001945100.2

Micro-Synteny

Tisochrysis lutea 15/06/2023 14:29:42

Feature details

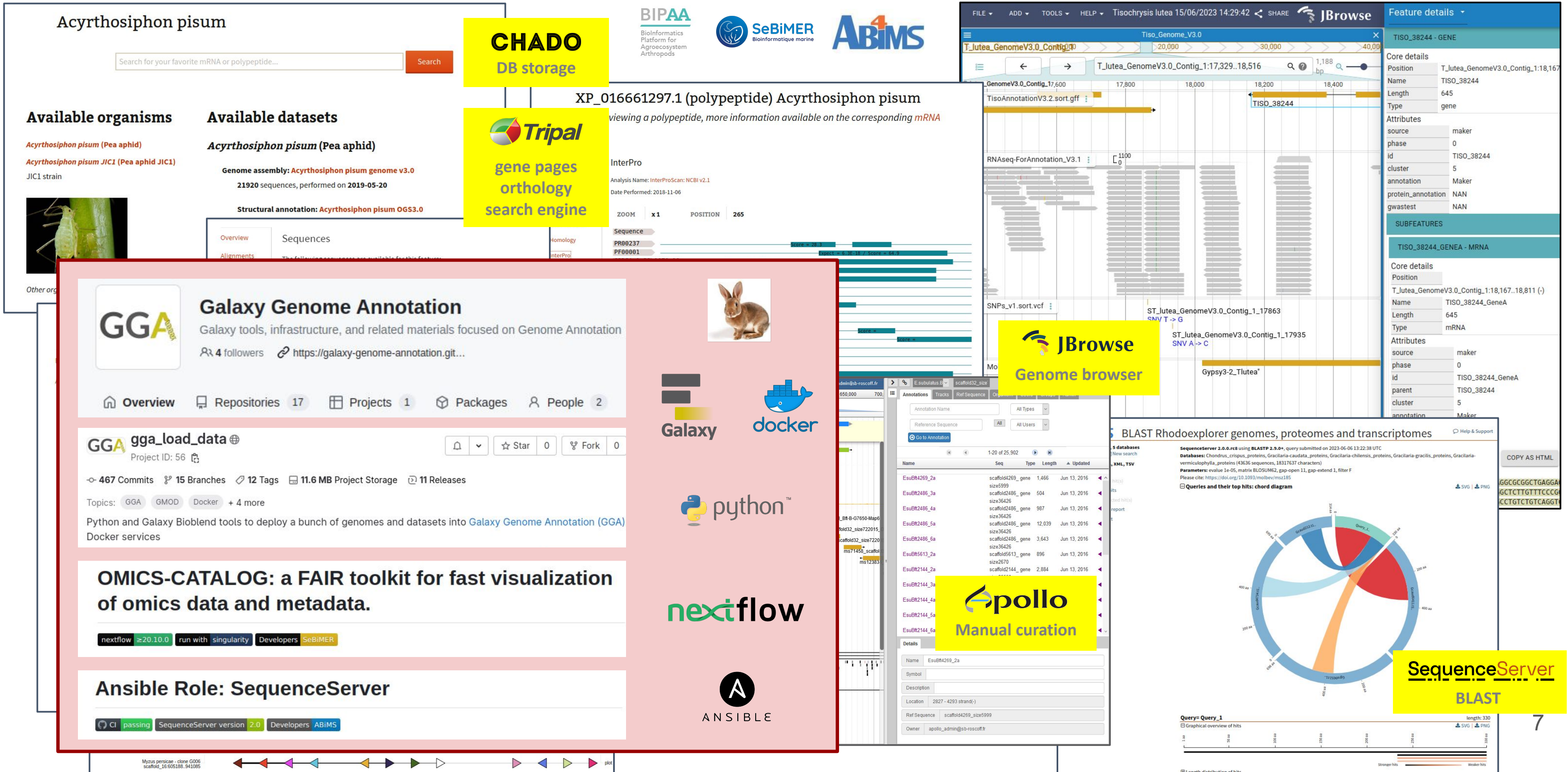
Core details	
Position	T_lutea_GenomeV3.0_Contig_1:18,167
Name	TISO_38244
Length	645
Type	gene

Attributes	
source	maker
phase	0
id	TISO_38244
cluster	5
annotation	Maker
protein_annotation	NAN
gwastest	NAN

BLAST Rhodoexplorer genomes, proteomes and transcriptomes

Queries and their top hits: chord diagram

Challenges: towards automated deployment



Acyrthosiphon pisum

Search for your favorite mRNA or polypeptide... Search

Available organisms

Acyrthosiphon pisum (Pea aphid)

Acyrthosiphon pisum JIC1 (Pea aphid JIC1)

JIC1 strain

Available datasets

Acyrthosiphon pisum **(Pea aphid)**

Genome assembly: Acyrthosiphon pisum genome v3.0

21920 sequences, performed on 2019-05-20

Structural annotation: Acyrthosiphon pisum OGS3.0

CHADO
DB storage

Tripal
gene pages
orthology
search engine

BIPAA
Bioinformatics Platform for Agroecosystem Arthropods

SeBiMER
Bioinformatique marine

ABiMS

XP_016661297.1 (polypeptide) Acyrthosiphon pisum

viewing a polypeptide, more information available on the corresponding mRNA

InterPro

Analysis Name: InterProScan: NCBI v2.1

Date Performed: 2018-11-06

Galaxy Genome Annotation

Galaxy tools, infrastructure, and related materials focused on Genome Annotation

4 followers https://galaxy-genome-annotation.git...

Overview · Repositories 17 · Projects 1 · Packages · People 2

gga_load_data

Project ID: 56

Star 0 · Fork 0

467 Commits · 15 Branches · 12 Tags · 11.6 MB Project Storage · 11 Releases

Topics: GGA · GMOD · Docker · + 4 more

Python and Galaxy Bioblend tools to deploy a bunch of genomes and datasets into Galaxy Genome Annotation (GGA) Docker services

OMICS-CATALOG: a FAIR toolkit for fast visualization of omics data and metadata.

nextflow ≥20.10.0 · run with singularity · Developers SeBiMER

Ansible Role: SequenceServer

CI passing · SequenceServer version 2.0 · Developers ABiMS

Galaxy · docker · python · nextflow · ANSIBLE

FILE · ADD · TOOLS · HELP · Tisochrysis lutea 15/06/2023 14:29:42 · SHARE · JBrowse

Tiso_Genome_V3.0

T_lutea_GenomeV3.0_Contig_1:17,329..18,516

TisoAnnotationV3.2.sort.gff · TISO_38244

RNAseq-ForAnnotation_V3.1

SNPs_v1.sort.vcf

ST_lutea_GenomeV3.0_Contig_1_17863
SNV T -> G

ST_lutea_GenomeV3.0_Contig_1_17935
SNV A -> C

Gypsy3-2_Tlutea"

Feature details

TISO_38244 - GENE

Core details

Position	T_lutea_GenomeV3.0_Contig_1:18,167
Name	TISO_38244
Length	645
Type	gene

Attributes

source	maker
phase	0
id	TISO_38244
cluster	5
annotation	Maker
protein_annotation	NAN
gwastest	NAN

SUBFEATURES

TISO_38244_GENEA - MRNA

Core details

Position
T_lutea_GenomeV3.0_Contig_1:18,167..18,811 (-)

Name	TISO_38244_GeneA
Length	645
Type	mRNA

Attributes

source	maker
phase	0
id	TISO_38244_GeneA
parent	TISO_38244
cluster	5
annotation	Maker

COPY AS HTML

JBrowse
Genome browser

Apollo
Manual curation

BLAST Rhodoexplorer genomes, proteomes and transcriptomes

SequenceServer 2.0.0.rc8 using **BLASTP 2.9.0+**, query submitted on 2023-06-06 13:22:38 UTC

Databases: Chondrus_crispus_proteins, Gracilaria-caudata_proteins, Gracilaria-chilensis_proteins, Gracilaria-gracilis_proteins, Gracilaria-vermiculophylla_proteins (43636 sequences, 18317637 characters)

Parameters: evalue 1e-05, matrix BLOSUM62, gap-open 11, gap-extend 1, filter F

Please cite: https://doi.org/10.1093/molbev/msz185

Queries and their top hits: chord diagram

Query= Query_1

SequenceServer
BLAST

BEAURIS: How does it work?

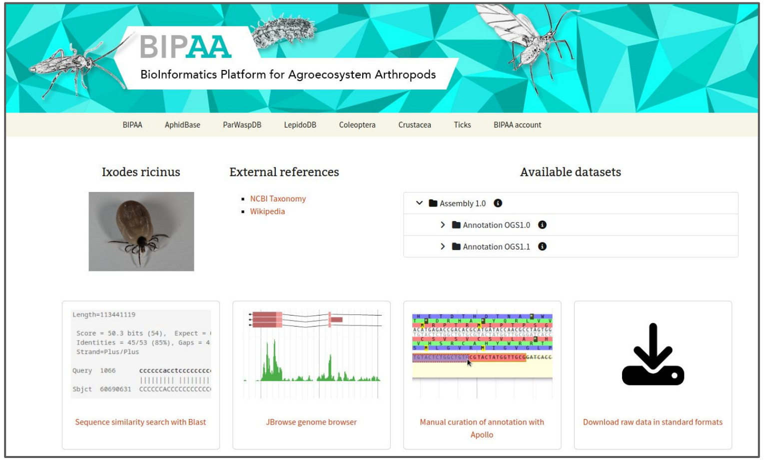
Biological data

fasta, gff, bam, tsv

Tedious & slow to do *by hand*

Automated, modular and FAIR deployment ?

Web interfaces



BIPAA

BioInformatics Platform for Agroecosystem Arthropods

BIPAA AphidBase ParWaspDB LepidoDB Coleoptera Crustacea Ticks BIPAA account

Ixodes ricinus

External references

- NCBI Taxonomy
- Wikipedia

Available datasets

- Assembly 1.0
  - Annotation OGS1.0
  - Annotation OGS1.1

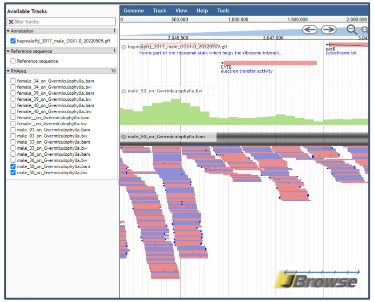
```
Length=113441119

 Score = 50.3 bits (54),  Expect = 
 Identities = 45/53 (85%), Gaps = 4
 Strand=Plus/Plus

Query  1066      cccccacctcccccccc
                 ||||||||| ||||||||
Sbjct  60690631  CCCCCACCCCCCCCCCC
```

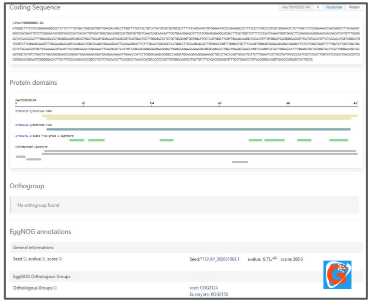
Sequence similarity search with Blast

JBrowse genome browser

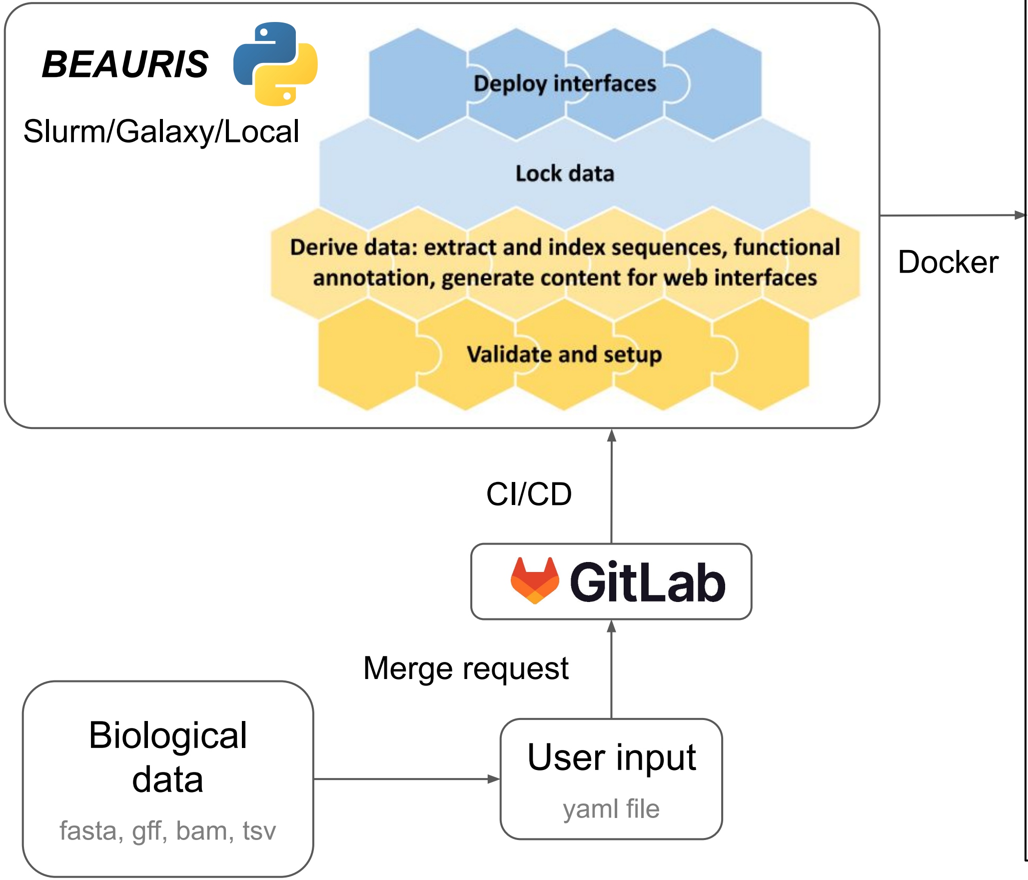
Manual curation of annotation with Apollo

Download raw data in standard formats





BEAURIS: How does it work?



BEAURIS

Slurm/Galaxy/Local

Deploy interfaces

Lock data

Derive data: extract and index sequences, functional annotation, generate content for web interfaces

Validate and setup

Docker

CI/CD

GitLab

Merge request

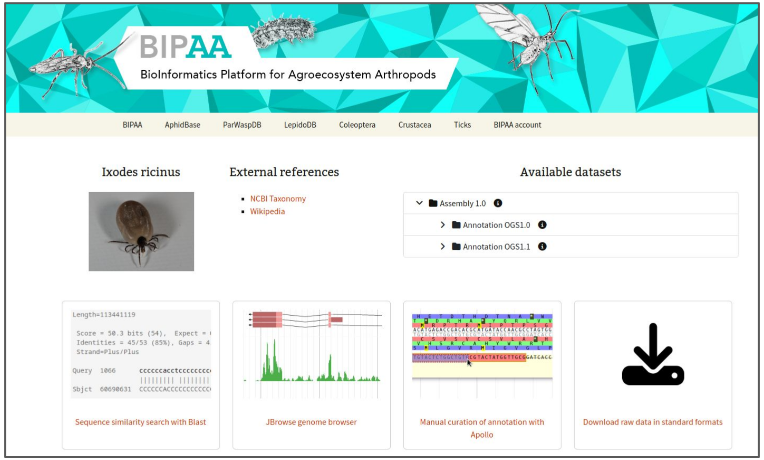
Biological data

fasta, gff, bam, tsv

User input

yaml file

Web interfaces



BIPAA

BioInformatics Platform for Agroecosystem Arthropods

Ixodes ricinus

External references

- NCBI Taxonomy
- Wikipedia

Available datasets

- Assembly 1.0
  - Annotation OGS1.0
  - Annotation OGS1.1

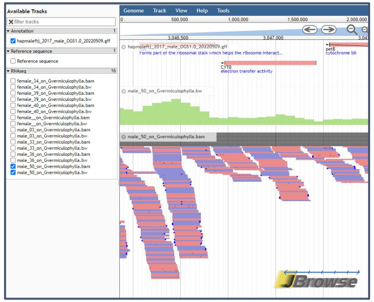
```
Length=113441119

 Score = 50.3 bits (54),  Expect = 
 Identities = 45/53 (85%), Gaps = 4
 Strand=Plus/Plus

Query  1066       cccccacctcccccccc
Sbjct  60690631   CCCCCACCCCCCCCCCC
```

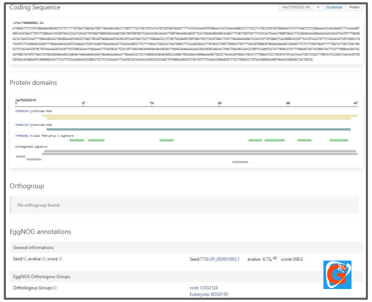
Sequence similarity search with Blast

JBrowse genome browser

Manual curation of annotation with Apollo

Download raw data in standard formats





BEAURIS: Input files

Users submit ***yaml*** files describing their datasets through a ***merge request***

Add Hyposoter didymator

ROBIN Stephanie requested to merge `slittoralis` into `master` 5 months ago

```
genus: Hyposoter
species: didymator
common_name: Hyposoter didymator
restricted_to: didydb
xrefs:
  ncbi_taxonomy: 260305
  wikipedia: Hyposoter
picture:
  author: Picture copyright 2011 J.K. Lindsey
  source_url: http://www.commanster.eu/commanster/Insects/Bees/ABees/Hyposoter.didymator.html
assemblies:
  - version: "2.0"
    file:
      path: /home/foobar/assembly/v2.0/Hd_genome_2.0.fa
      type: fasta
    date: 2022-09-22
    source: BIPAA
    description: |
      Rescaffolding of assembly 1.0
    annotations:
      - version: "OGS2.0"
        file:
          path: /home/foobar/annotation/v2.0/OGS2.0_20230316.gff3
          type: gff
        date: 2023-03-16
        source: BIPAA
        description: |
          Helixer annotation
    tracks:
      - name: Transcriptome of Calyx-Oviduct at stage 1
        type: rnaseq
        file:
          path: /home/foobar/rnaseq/genome_v2.0/1_calyx_oviduct.markdup.sorted.bam
          type: bam
```

BEAURIS: Input file (yml)

“Base” metadata

- Genus
- Species
- Common name
- *More…*

Required “Data” values

- Assemblies
- Annotations

Optional “Data” values

- Proteomes
- Transcriptomes
- Jbrowse tracks

“Customisation” values

- Job parameters
- Web services
- Access restriction

The input file follows a **modular & evolving** *validation schema*

⊗ failed **Job lint-yml** triggered 2 weeks ago by **BRETAUDEAU Anthony**

BEAURIS: Staging & Production

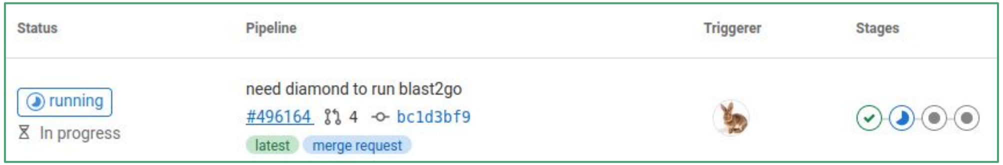
Two distinct working environments

- Staging environment: when opening a *Merge Request*
  - Web interfaces are admin-restricted (for validation)
  - Apollo **staging** instance
- Production environment: after merging the MR
  - Datasets are “locked” for safekeeping
  - Interfaces are accessible (public / restricted access)
  - Apollo **production** instance

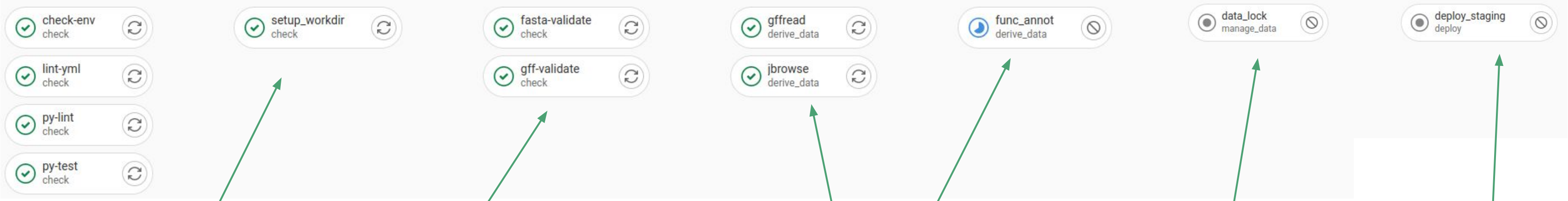
The *production* pipeline will **reuse** the *staging* pipeline results: no waste of resources!

Easy sanity check for data quality

BEAURIS: Steps



Status	Pipeline	Triggerer	Stages
running In progress	need diamond to run blast2go #496164 ⇅ 4 -○- bc1d3bf9 latest merge request		✓ ◔ ● ●



Quick checks

Input data check

Derive data

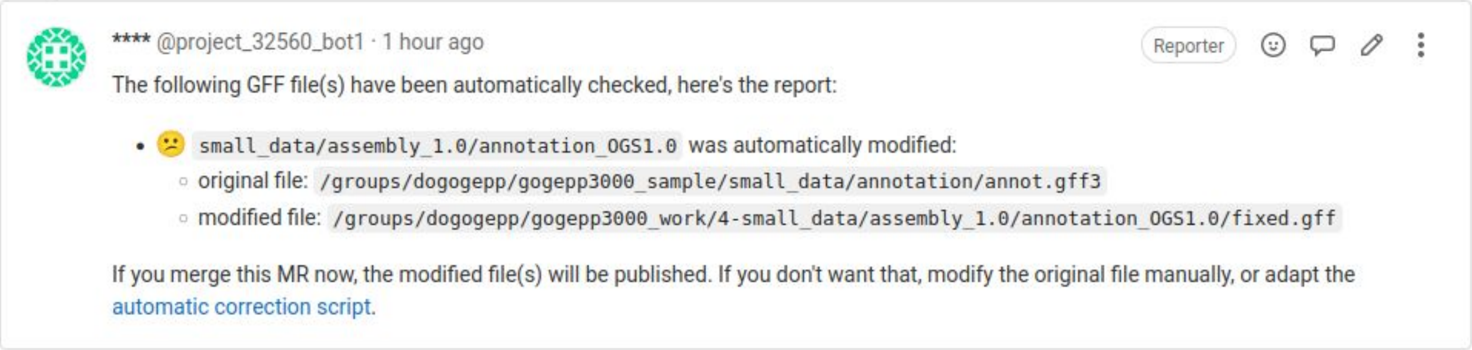
Lock data

Interface deployment

BEAURIS: Data validation & setup

First step: Checking input data

- Yaml file structure check
- Setting up job environment
- Fasta & Tracks validation
- GFF validation & ***cleanup***

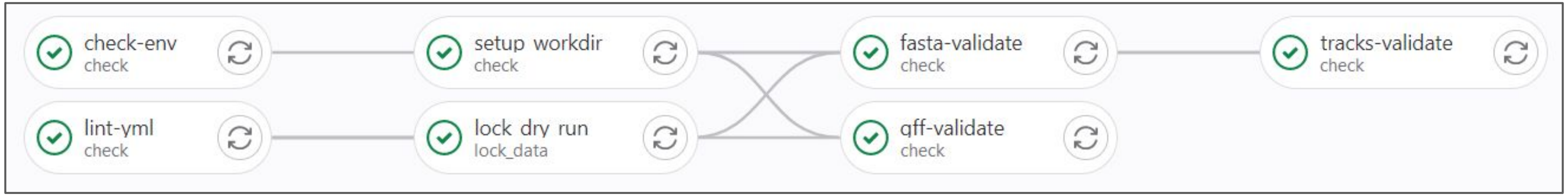


**** @project_32560_bot1 · 1 hour ago

The following GFF file(s) have been automatically checked, here's the report:

- 😕 `small_data/assembly_1.0/annotation_OGS1.0` was automatically modified:
  - original file: `/groups/dogogepp/gogepp3000_sample/small_data/annotation/annot.gff3`
  - modified file: `/groups/dogogepp/gogepp3000_work/4-small_data/assembly_1.0/annotation_OGS1.0/fixed.gff`

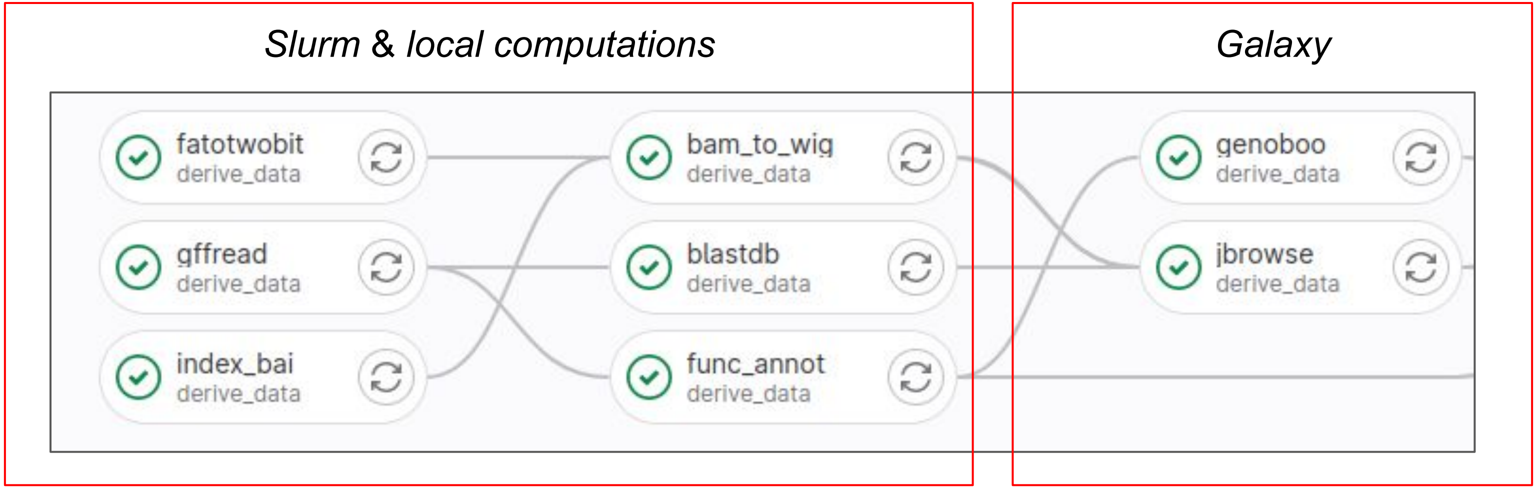
If you merge this MR now, the modified file(s) will be published. If you don't want that, modify the original file manually, or adapt the automatic correction script.



BEAURIS: Deriving data

Second step: Functional annotation & generating content for web interfaces

- Currently using internal workflows, but new ones can be easily added
- Web interface content is generated on a Galaxy server, and UIs are deployed at a later step



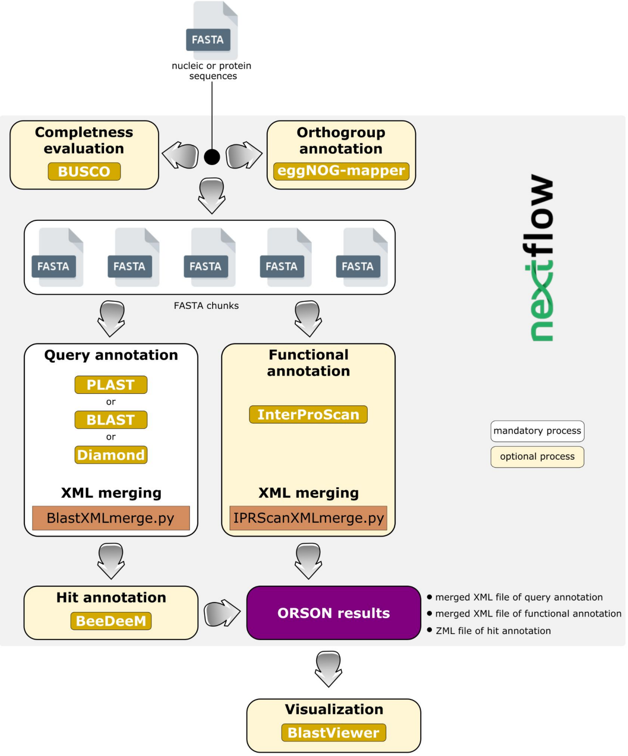
BEAURIS: Functional annotation with the ORSON module

ORSON: workflow for prOteome and tRanScriptome functiOnal aNnotation.

ORSON version 1.2.0 | nextflow ≥20.10.0 | run with singularity | Developers SeBiMER

- Runs state-of-the-art annotation tools:
    - eggNOG-mapper
    - InterProScan
    - Diamond
- Uses NextFlow workflow management system with:
    - singularity containers (reproducibility)
    - jobs submitted on a SLURM HPC

```
job_specs:
  drmaa:
    func_annot_orson:
      # SCRATCH_WORK_DIR is supposed to be a scratch storage area for nextflow that can be c
      env : >
        source /etc/profile.d/modules.sh; module load nextflow/22.10.0 graphviz/2.40.1;
        export ORSON_PATH=/path/to/orson/;
        export BLAST_DB_PATH=/path/to/uniref90/diamond/uniref90_2022_03/uniref90.dmnd;
        export CLUSTER_CONFIG_PATH=/path/to/cluster.config;
        export SCRATCH_WORK_DIR=/path/to/scratch/workdir
```



- merged XML file of query annotation
- merged XML file of functional annotation
- ZML file of hit annotation

BEAURIS: Data locking

Third step: raw / generated datasets are labelled & stored for safekeeping & reuse

Metadata:

- Genus
- Species
- Strain
- Assembly version
- Annotation version
- Tool version
- File md5

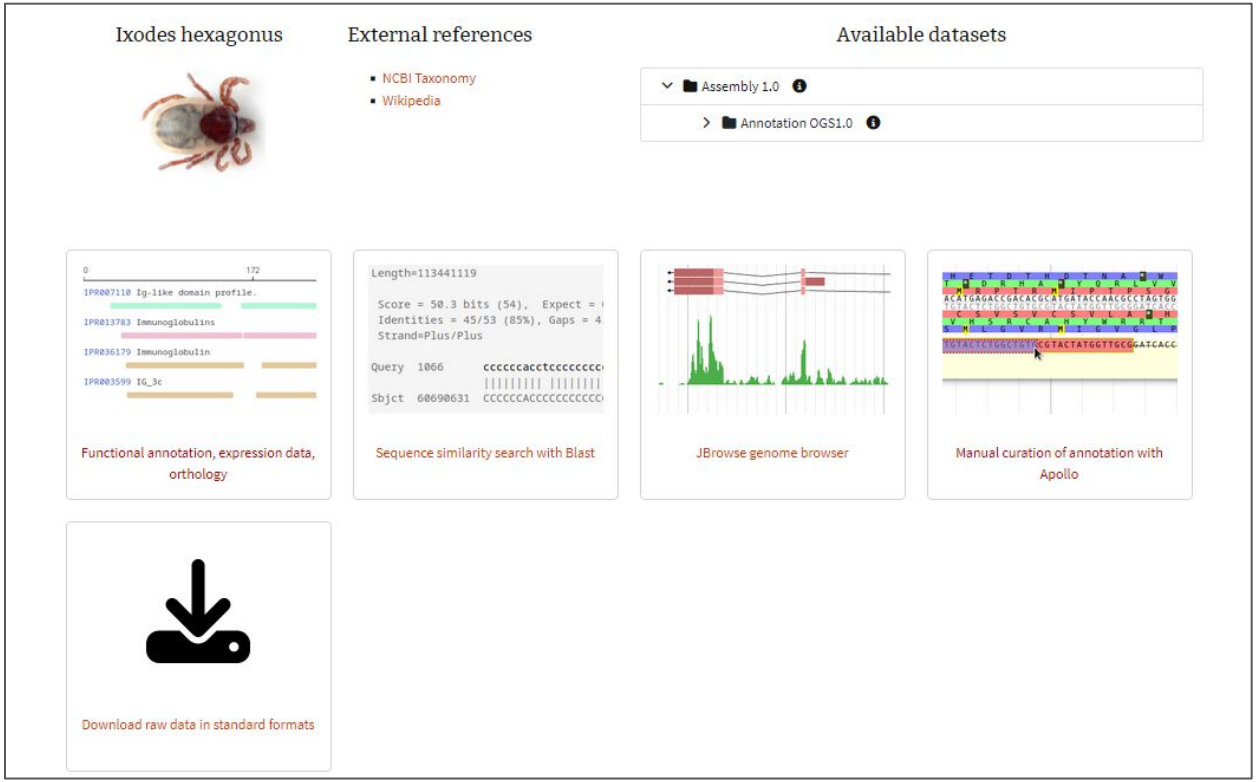
```
data_locker:
  method: dir
  options:
    target_dir: "$WORK_DIR/locked/"
    base_pattern: "{genus}/{species}/{strain}/"
    pattern_input: "{type}/{version}/{filename}"
    pattern_derived: "{task_id}/{tool_version}-{date}-{revision}/{filename}"
    locked_yml_dir: "$CI_PROJECT_DIR/locked/"
    locked_yml_dir_future: "$CI_PROJECT_DIR/future_locked/"
```

```
├── gff
│   └── OGS2.0
│       ├── OGS1.0_20230329.gff3
│       └── OGS1.0_20230329.gff3.metadata
├── gffread
│   └── 0.12.7-unknown-0
│       ├── carcania_ass1.0_annotOGS2.0_cds.fa
│       ├── carcania_ass1.0_annotOGS2.0_cds.fa.metadata
│       ├── carcania_ass1.0_annotOGS2.0_proteins.fa
│       ├── carcania_ass1.0_annotOGS2.0_proteins.fa.metadata
│       ├── carcania_ass1.0_annotOGS2.0_transcripts.fa
│       └── carcania_ass1.0_annotOGS2.0_transcripts.fa.metadata
```

BEAURIS: Interfaces

Fourth step: Interfaces deployment

Several interfaces are available, accessible through a *customizable* landing page



Ixodes hexagonus

External references

- NCBI Taxonomy
- Wikipedia

Available datasets

Functional annotation, expression data, orthology

```
Length=113441119

 Score = 50.3 bits (54),  Expect =
 Identities = 45/53 (85%), Gaps = 4
 Strand=Plus/Plus

Query  1066      ccccccacctcccccccc
                 |||||||||| ||||||||
Sbjct  60690631  CCCCCCACCCCCCCCCCC
```

Sequence similarity search with Blast

JBrowse genome browser

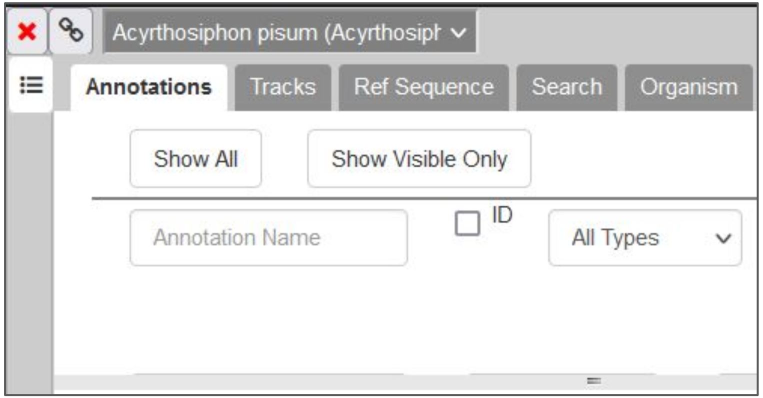
Manual curation of annotation with Apollo

Download raw data in standard formats

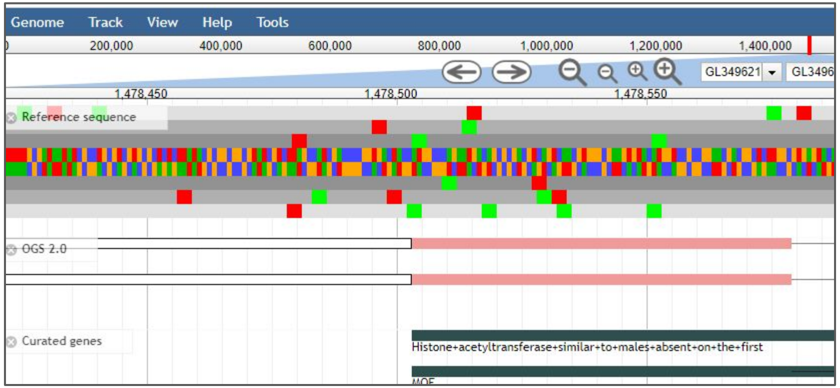
BEAURIS: Interfaces *(modular & optional)*

Available interfaces (for now)

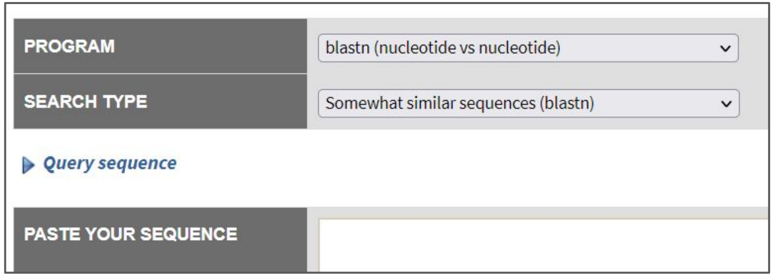
- Apollo (Existing server)



- Jbrowse



- Blast sequence search



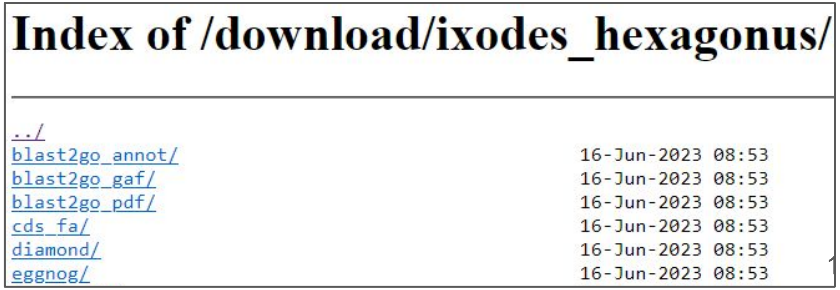
PROGRAM blastn (nucleotide vs nucleotide)

SEARCH TYPE Somewhat similar sequences (blastn)

Query sequence

PASTE YOUR SEQUENCE

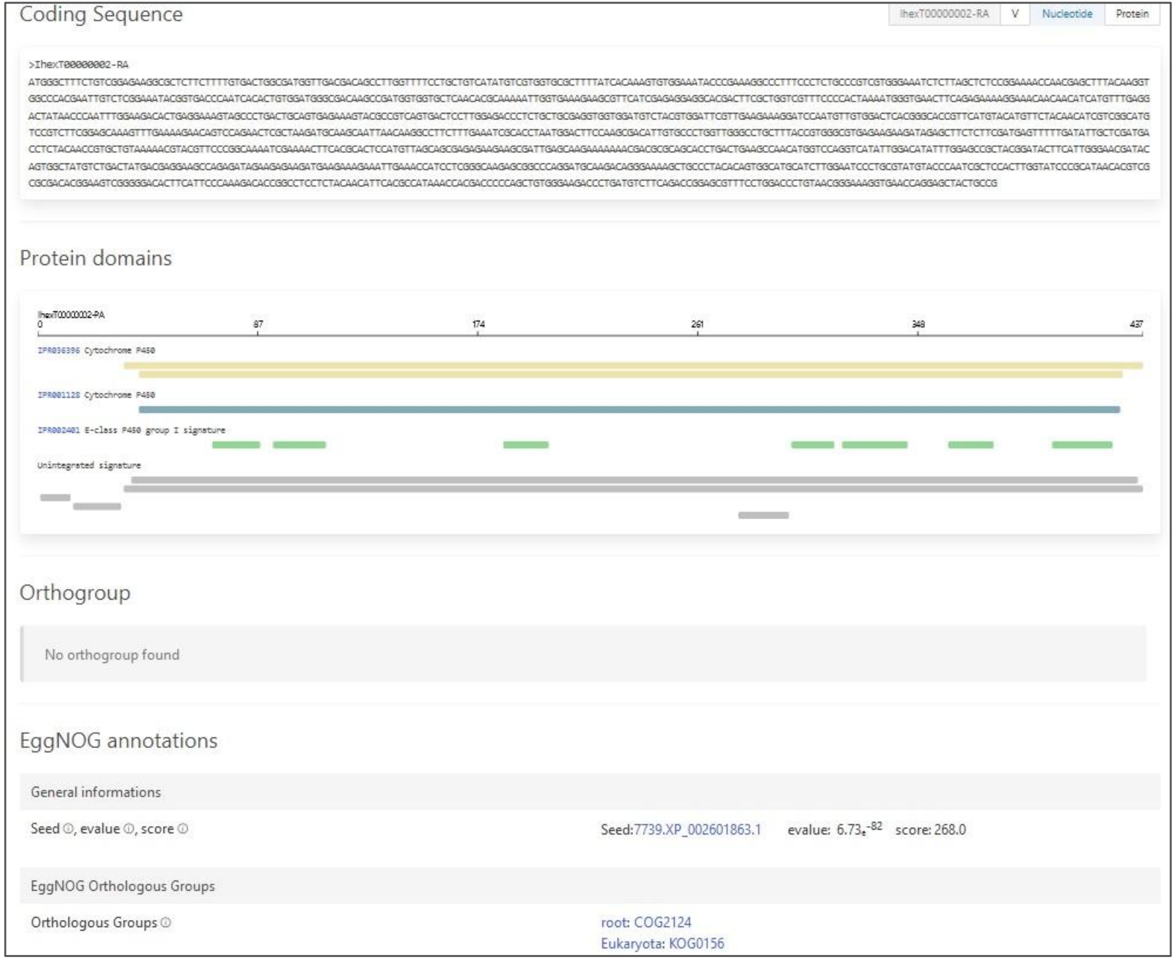
- Data download



Index of /download/ixodes_hexagonus/

```
../
blast2go_annot/                                   16-Jun-2023 08:53
blast2go_gaf/                                     16-Jun-2023 08:53
blast2go_pdf/                                     16-Jun-2023 08:53
cds_fa/                                           16-Jun-2023 08:53
diamond/                                          16-Jun-2023 08:53
eggnog/                                           16-Jun-2023 08:53
```

BEAURIS: Interfaces (Genenotebook)



Coding Sequence

```
>IhexT00000002-RA
ATGGGCTTTCTGTCGGAGAAGGCGCTCTTCTTTTGTGACTGGCGATGGTTGACGACAGCCTTGGTTTTCCTGCTGTCATATGTCGTGGTGCGCTTTTATCACAAAGTGTGGAAATACCCGAAAGGCCCTTTCCCTCTGCCCGTCGTGGGAAATCTCTTAGCTCTCCGGAAAACCAACGAGCTTTACAAGGT
GGCCCACGAATTGTCTCGGAAATACGGTGACCCAATCACACTGTGGATGGGCGACAAGCCGATGGTGGTGCTCAACACGCAAAAATTGGTGAAAGAAGCGTTCATCGAGAGGAGGCACGACTTCGCTGGTCGTTTCCCCACTAAAATGGGTGAACTTCAGAGAAAAGGAAACAACAACATCATGTTTGAGG
ACTATAACCCAATTTGGAAGACACTGAGGAAAGTAGCCCTGACTGCAGTGAGAAAGTACGCCGTCAGTGACTCCTTGGAGACCCTCTGCTGCGAGGTGGTGGATGTCTACGTGGATTCGTTGAAGAAAGGATCCAATGTTGTGGACTCACGGGCACCGTTCATGTACATGTTCTACAACATCGTCGGCATG
TCCGTCTTCGGAGCAAAGTTTGAAAAGAACAGTCCAGAACTCGCTAAGATGCAAGCAATTAACAAGGCCTTCTTTGAAATCGCACCTAATGGACTTCCAAGCGACATTGTGCCCTGGTTGGGCCTGCTTTACCGTGGGCGTGAGAAGAAGATAGAGCTTCTCTTCGATGAGTTTTTGATATTGCTCGATGA
CCTCTACAACCGTGCTGTAAAAACGTACGTTCCCGGCAAAATCGAAAACTTCACGCACTCCATGTTAGCAGCGAGAGAAGAAGCGATTGAGCAAGAAAAAAACGACGCGCAGCACCTGACTGAAGCCAACATGGTCCAGGTCATATTGGACATATTTGGAGCCGCTACGGATACTTCATTGGGAACGATAC
AGTGGCTATGTCTGACTATGACGAGGAAGCCAGAGATAGAAGAGAAGATGAAGAAAGAAATTGAAACCATCCTCGGGCAAGAGCGGCCCAGGATGCAAGACAGGGAAAAGCTGCCCTACACAGTGGCATGCATCTTGGAATCCCTGCGTATGTACCCAATCGCTCCACTTGGTATCCCGCATAACACGTCG
CGCGACACGGAAGTCGGGGGACACTTCATTCCCAAAGACACCGGCCTCCTCTACAACATTCACGCCATAAACCACGACCCCCAGCTGTGGGAAGACCCTGATGTCTTCAGACCGGAGCGTTTCCTGGACCCTGTAACGGGAAAGGTGAACCAGGAGCTACTGCCG
```

Protein domains

IPR036396 Cytochrome P450

IPR001128 Cytochrome P450

IPR002401 E-class P450 group I signature

Unintegrated signature

Orthogroup

No orthogroup found

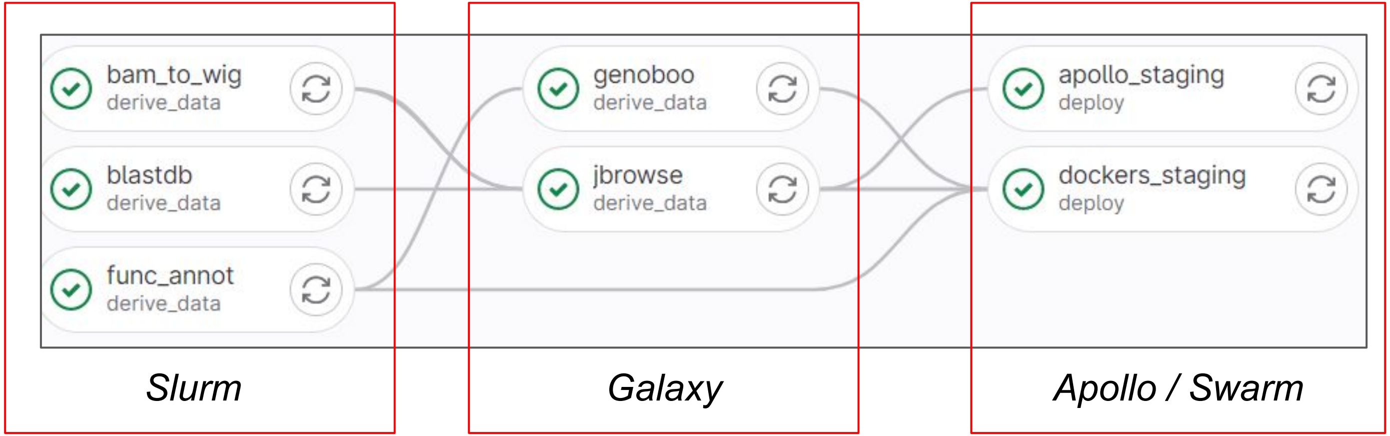
EggNOG annotations

General informations	
Seed ⓘ, evalue ⓘ, score ⓘ	Seed:7739.XP_002601863.1 evalue: 6.73_{e}^{-82} score: 268.0

EggNOG Orthologous Groups	
Orthologous Groups ⓘ	root: COG2124 Eukaryota: KOG0156

BEAURIS: Technical stuff (what do you need?)

- **Gitlab Runner**, able to:
  - Access a computing cluster (Slurm) / Have local computing resources
  - Read access your data, and write access for the jobs and locks
  - Access an hosting node for web interfaces (if needed) via Ansible
- **Galaxy account** (*If using Jbrowse, Apollo or GeneNoteBook*)
- **Apollo instance**(s) (*If using Apollo*)
- **Docker Swarm** cluster & *Traefik (If deploying any interface)*

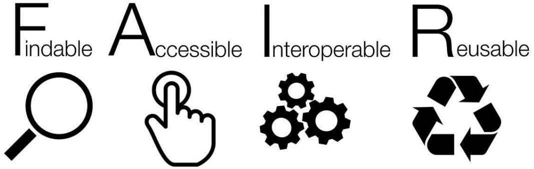


Conclusion

C'EST BEAU(RIS) ET C'EST FAIR !

Benefits:

- Complete automation: saves **time** and improves **reliability**
- Structured **data catalog** to capitalize on
- Completely **modular**, adaptable and deployable wherever you want
- **Reproducible** analyses & web portal deployment
- Open code (contributions encouraged!):
  - https://gitlab.com/beaur1s/beauris
  - https://beauris.readthedocs.io



Perspectives

- Improve **portability** of the solution
- Support a wide-range of **data types** (e.g. genomic variants, orthology, synteny, phenotypic data)
- Support **more web interfaces** (e.g. JBrowse2, AskOmics, synteny viewer)
- Use **data standards**: Schema.org, RO-Crate, ...
- **Data submission** to public data repositories (ENA, recherche.data.gouv.fr, …)
- Automatic integration of **genomes from public banks**

Perspectives

- Improve **portability** of the solution
- Support a wide-range of **data types** (e.g. genomic variants, orthology, synteny, phenotypic data)
- Support **more web interfaces** (e.g. JBrowse2, AskOmics, synteny viewer)
- Use **data standards**: Schema.org, RO-Crate, ...
- **Data submission** to public data repositories (ENA, recherche.data.gouv.fr, …)
- Automatic integration of **genomes from public banks**

© Thierry PEREZ/CNRS Photothèque

ATLASea

Atlas of marine genomes: from massive data to innovation

5000 GENOMES

PC3 - BYTE-SEA: numerical infrastructure

Coordinated by the Institut Français de Bioinformatique, this project will improve and store the computerized annotation of the DNA in order to locate genes, retrace their evolutionary history and assign functions to them. The genomes will then be stored in open databases accessible to the international community.

IFB INSTITUT FRANÇAIS DE BIOINFORMATIQUE GenOuest ABiMS SeBiMER Bioinformatique marine IBENS Institut de Biologie de l'Ecole Normale Supérieure

Collaborators

- Alizée Bardon (SeBiMER)
- Matéo Boudet (Genouest BIPAA)
- Anthony Bretaudeau (Genouest BIPAA)
- Loraine Brillet-Guéguen (ABiMS)
- Alexandre Cormier (SeBiMER)
- Erwan Corre (ABiMS)
- Patrick Durand (SeBiMER)
- Ludovic Fourteau (Genouest BIPAA)
- Arthur Le Bars (ABiMS, IFB)
- Laura Leroi (SeBiMER)
- Karine Massau (ABiMS)
- Cyril Noël (SeBiMER)

Funders

INRAE

cnrs

Ifremer

SORBONNE UNIVERSITÉ
CRÉATEURS DE FUTURS
DEPUIS 1257

IFB
INSTITUT FRANÇAIS DE BIOINFORMATIQUE

With the support of the
Erasmus+ Programme
of the European Union

Project (`2020-1-NL01-KA203-064717`) funded with the support of the Erasmus+ programme of the European Union. Their funding has supported a large number of tutorials within the GTN across a wide array of topics.

Thank you for your attention!