



Structural Annotation of Marine Genomes from ATLASea

Jean-Marc Aury, Genoscope

JOBIM 2026, Mini-Symposium



jmaury@genoscope.cns.fr

Agenda

- The ATLASea program
- Structural annotation
- Annotation workflow
- First results on marine eukaryotic genomes



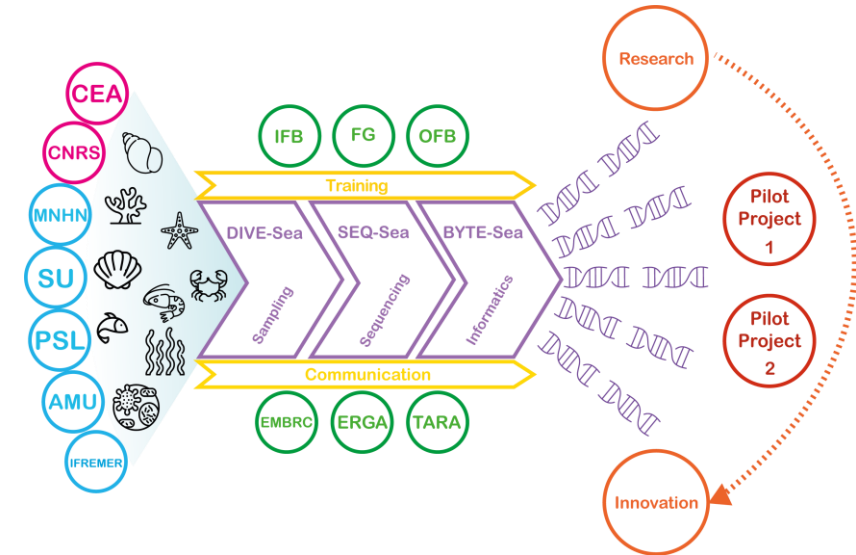
ATLASEa program



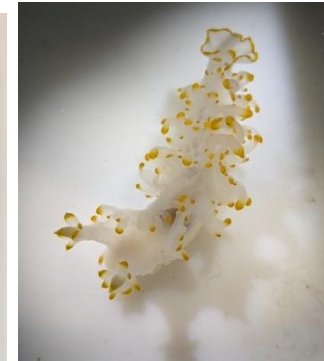
PROGRAMME
DE RECHERCHE
GÉNOMES MARINS



- An 8-year program with €41 million in funding under the France 2030 framework
- Directed by the CNRS and the CEA
- To sample and sequence 4500 species in the French EEZ
- To bank taxonomically identified frozen samples in a dedicated ATLASEa collection
- To provide data freely in a dedicated portal with comparative genomics tools
- To train students and researchers in taxonomy, computational genomics through a PhD programme, Master bursaries and summer
- To establish and fund consortia to exploit data in scientific pilot projects



DIVE-Sea sampling expedition in Roscoff – April 2026



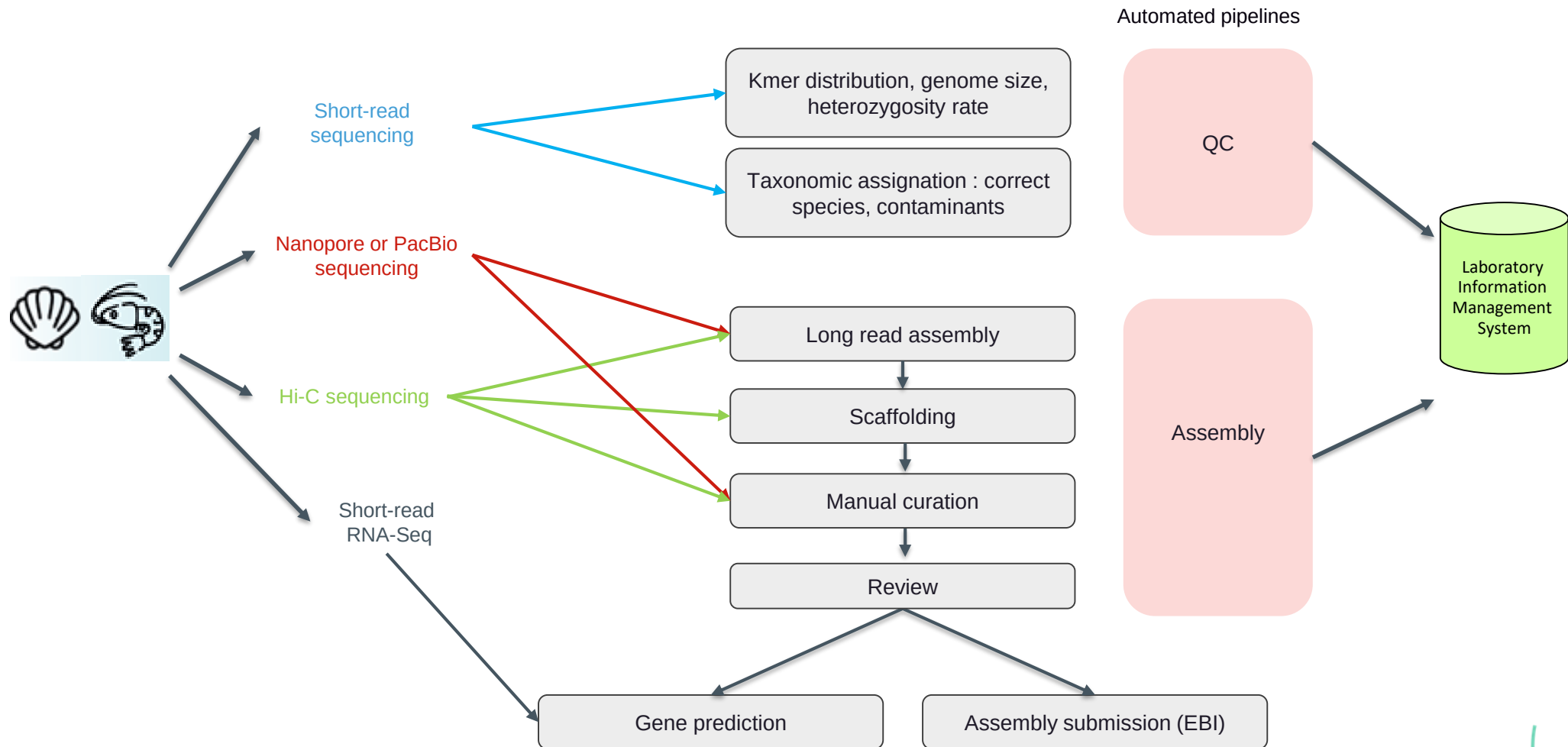
Station Biologique de Roscoff (CNRS/SU)

28/03/2026 – 11/04/2026

631 species sampled



SEQ-Sea strategy



ATLASEa: Where are we today?

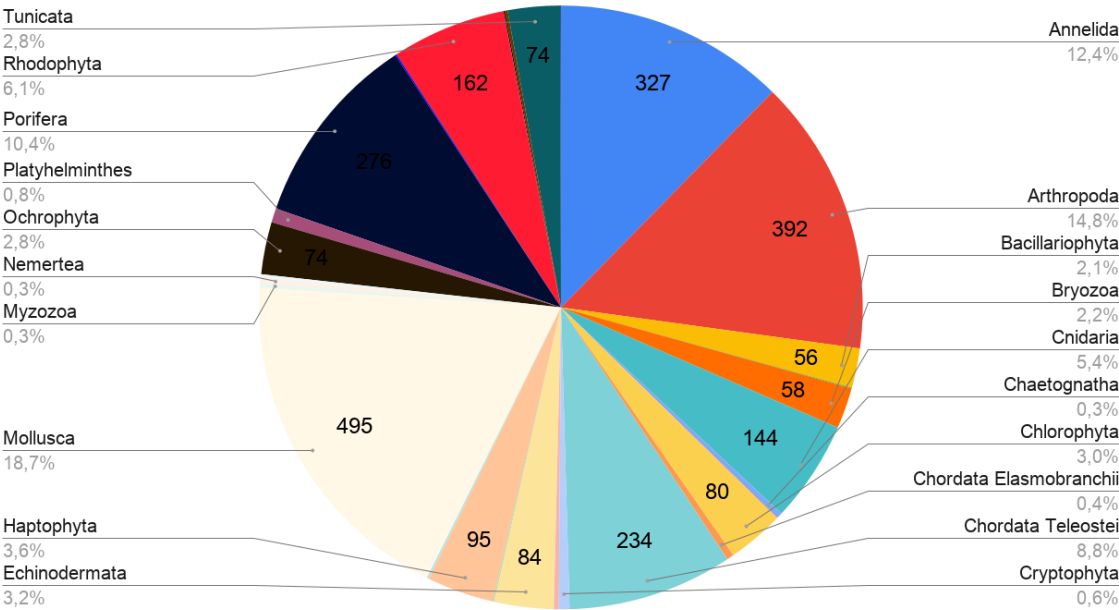
sample
acquired

Progression of samples from

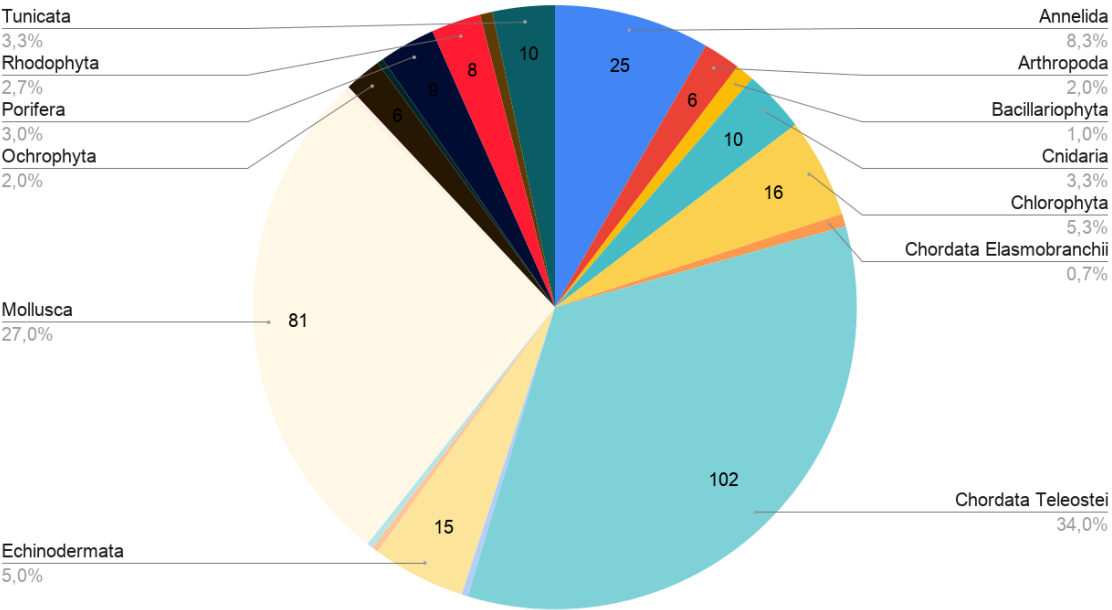
curated
assembly

Number of samples	Validated species	Short-read data	Validated samples	Long-read data	Long-read assemblies	Hi-C data	Chromosome-scale assemblies	Curated assemblies	RNA-Seq	Submitted assemblies
2510	1608	1075	860	563	529	620	388	330	175	300

ATLASEa samples

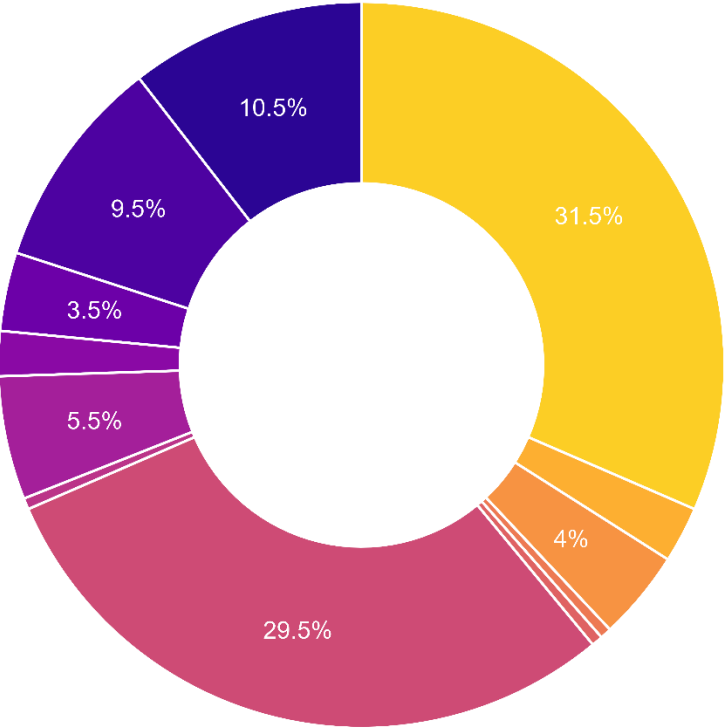


Submitted assemblies

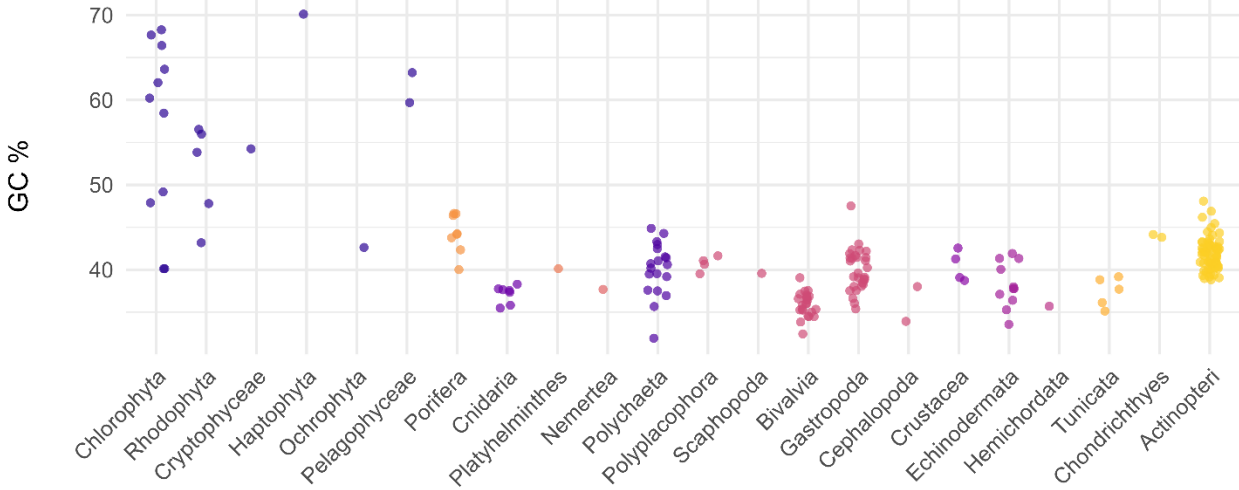


ATLASEa: Where are we today?

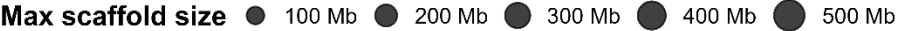
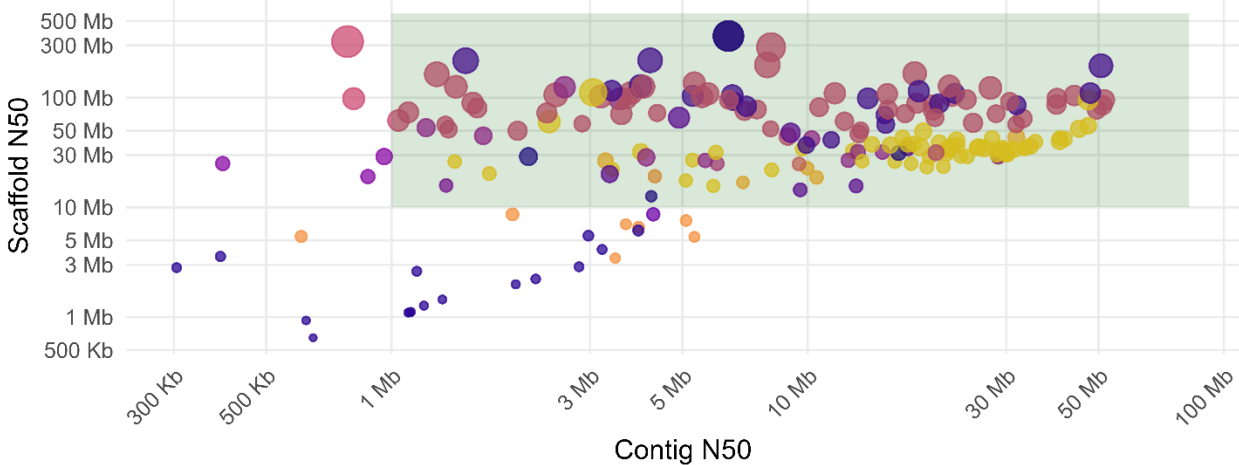
A



B

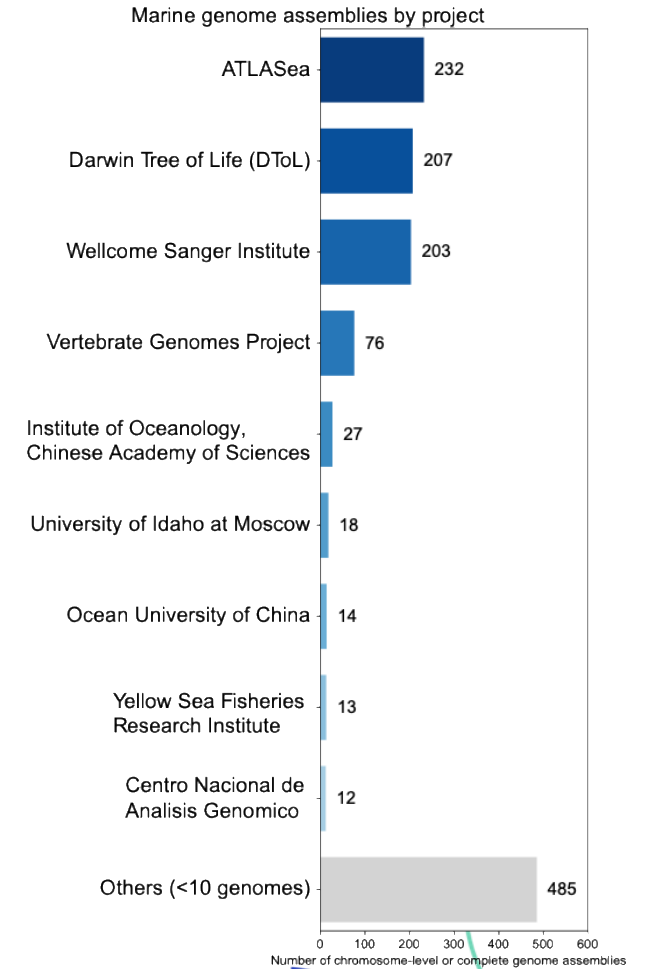
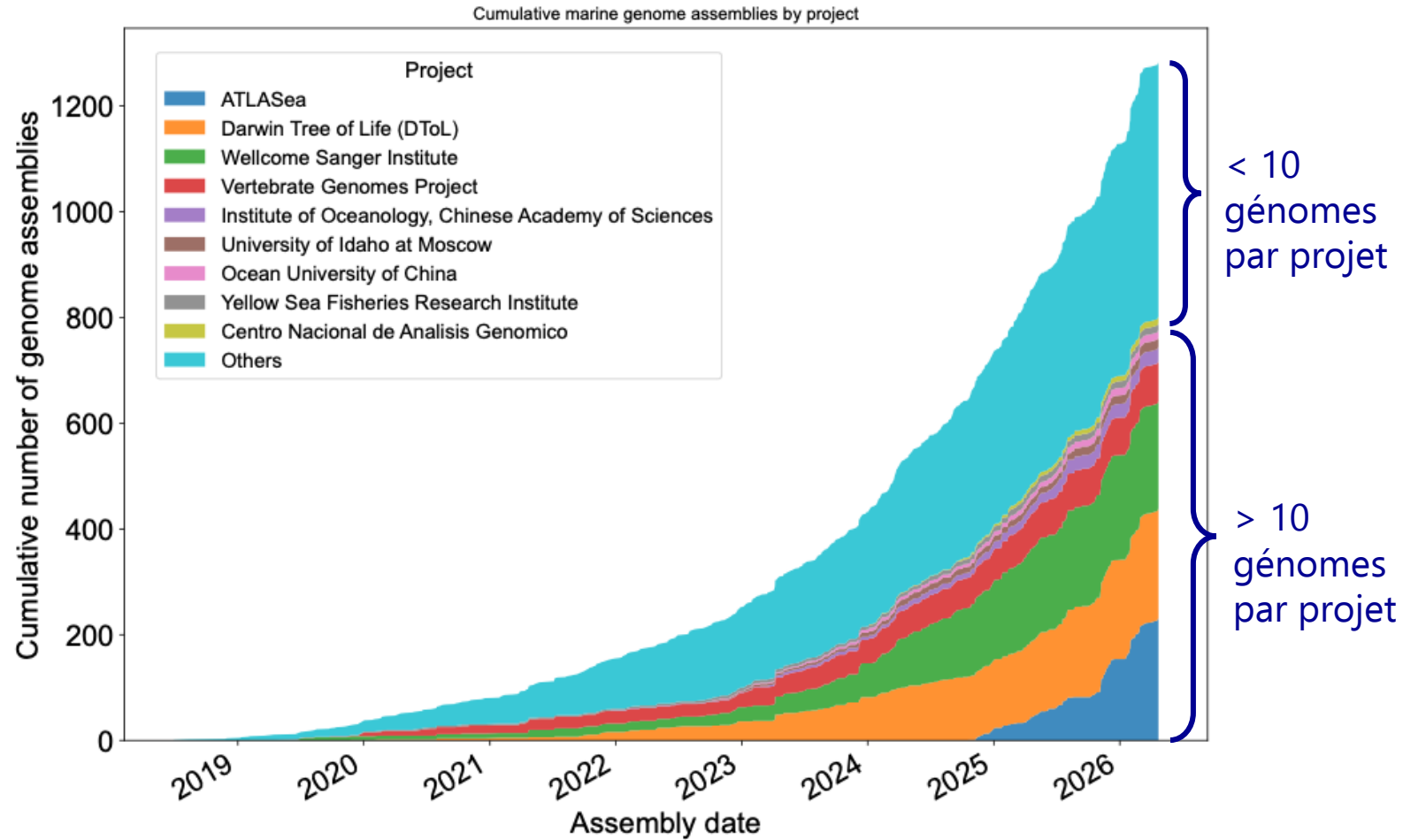


C



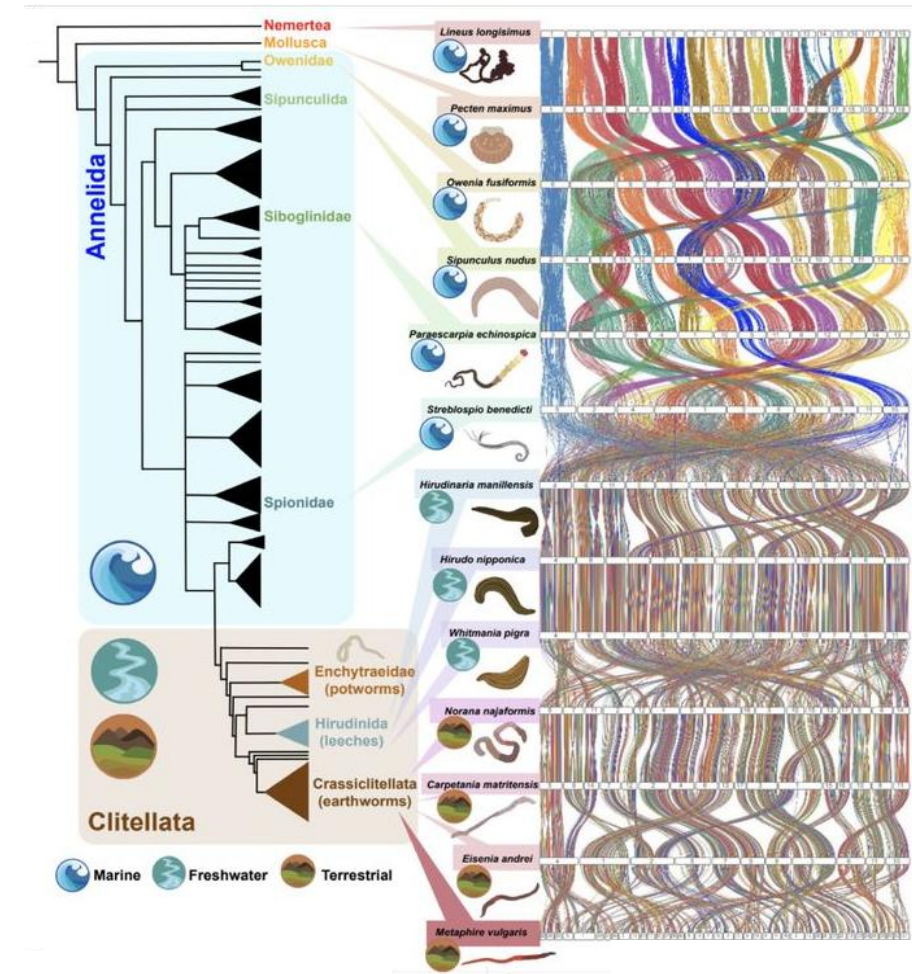
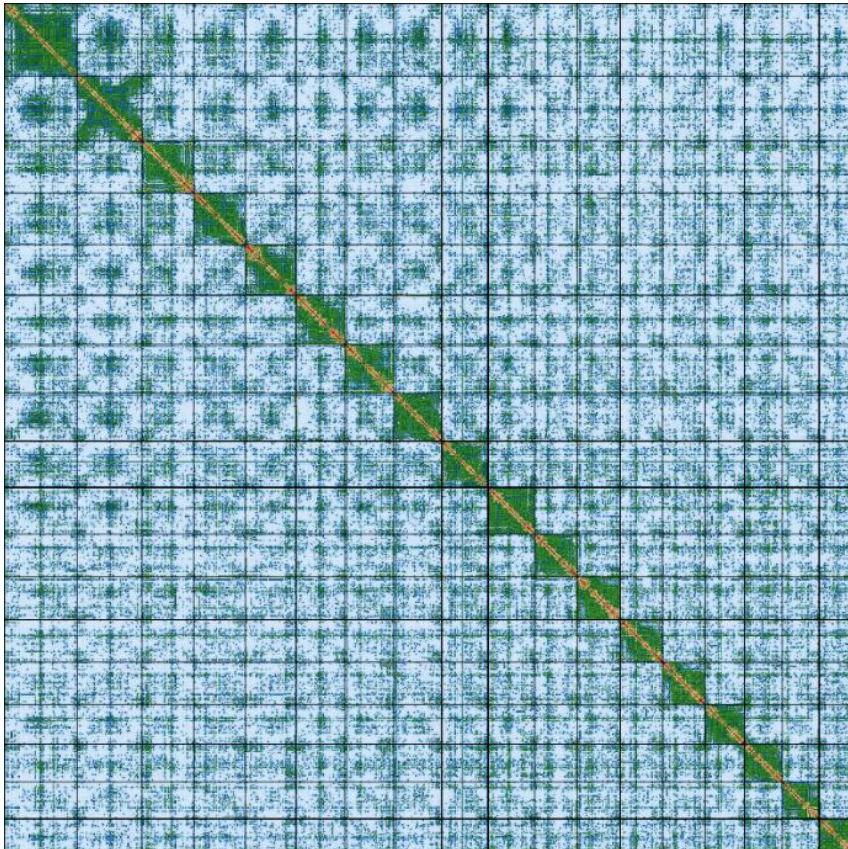
ATLASea: Where are we today?

ATLASea is currently the EBP project that has generated the largest number of marine genomes



The importance of obtaining chromosome-scale assemblies

Chromosomal-scale and validated assemblies



Article | Published: 18 June 2025

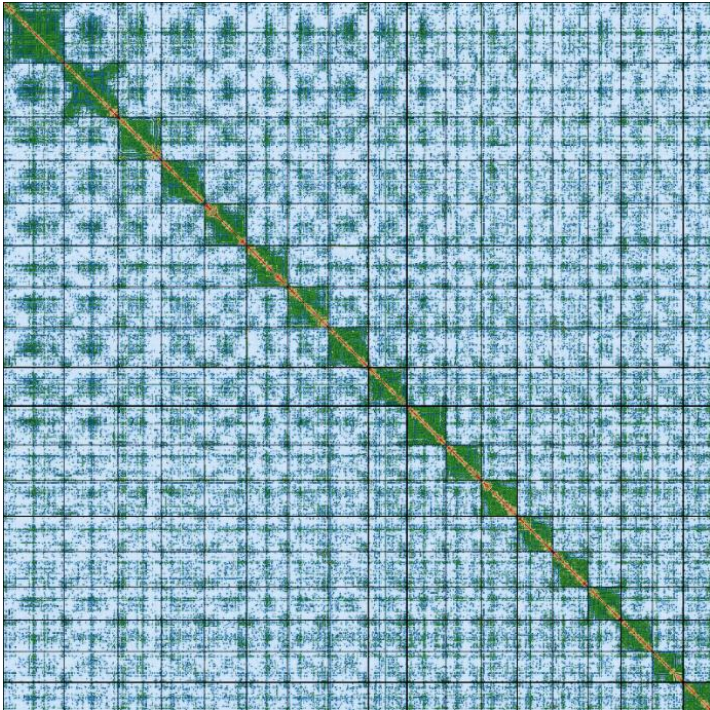
An episodic burst of massive genomic rearrangements and the origin of non-marine annelids

Carlos Vargas-Chávez, Lisandra Benítez-Álvarez, Gemma I. Martínez-Redondo, Lucía Álvarez-González, Judit Salces-Ortiz, Klara Eleftheriadi, Nuria Escudero, Nadège Guiguielmoni, Jean-François Flot, Marta Novo, Aurora Ruiz-Herrera, Aoife McLysaght & Rosa Fernández

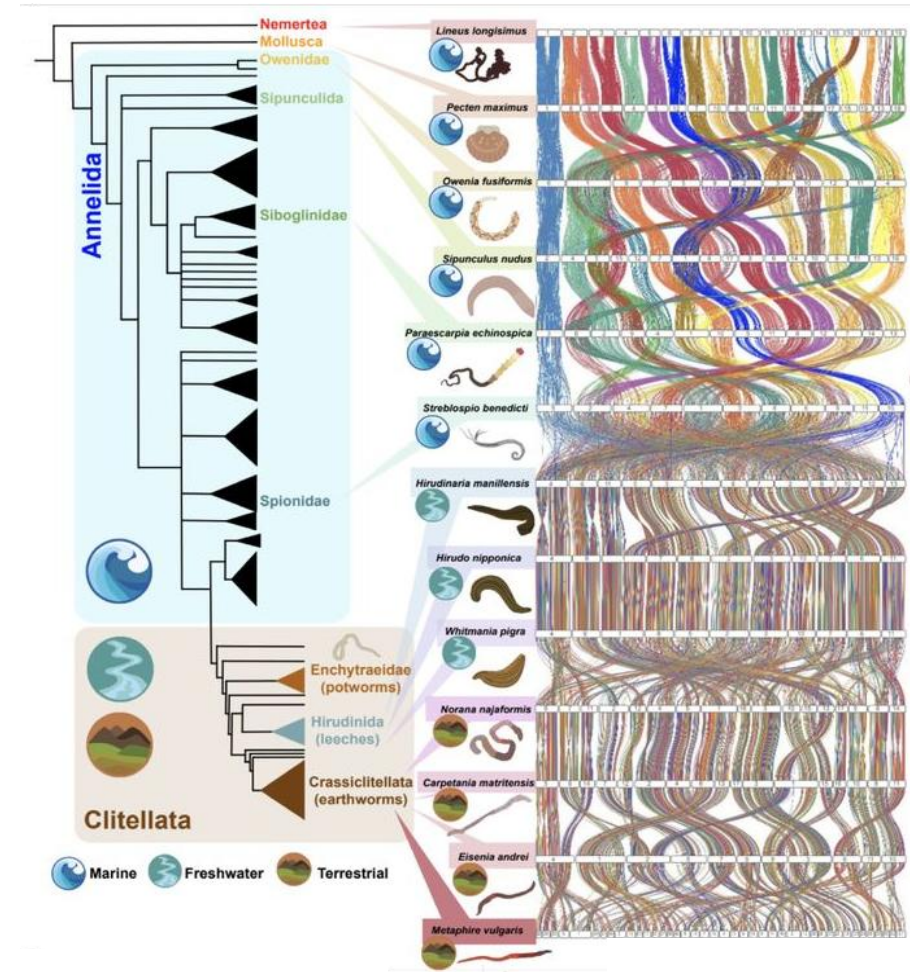
Nature Ecology & Evolution 9, 1263–1279 (2025) | [Cite this article](#)

The importance of obtaining chromosome-scale assemblies

Chromosomal-scale and
validated assemblies



+
gene
annotation



Article | Published: 18 June 2025

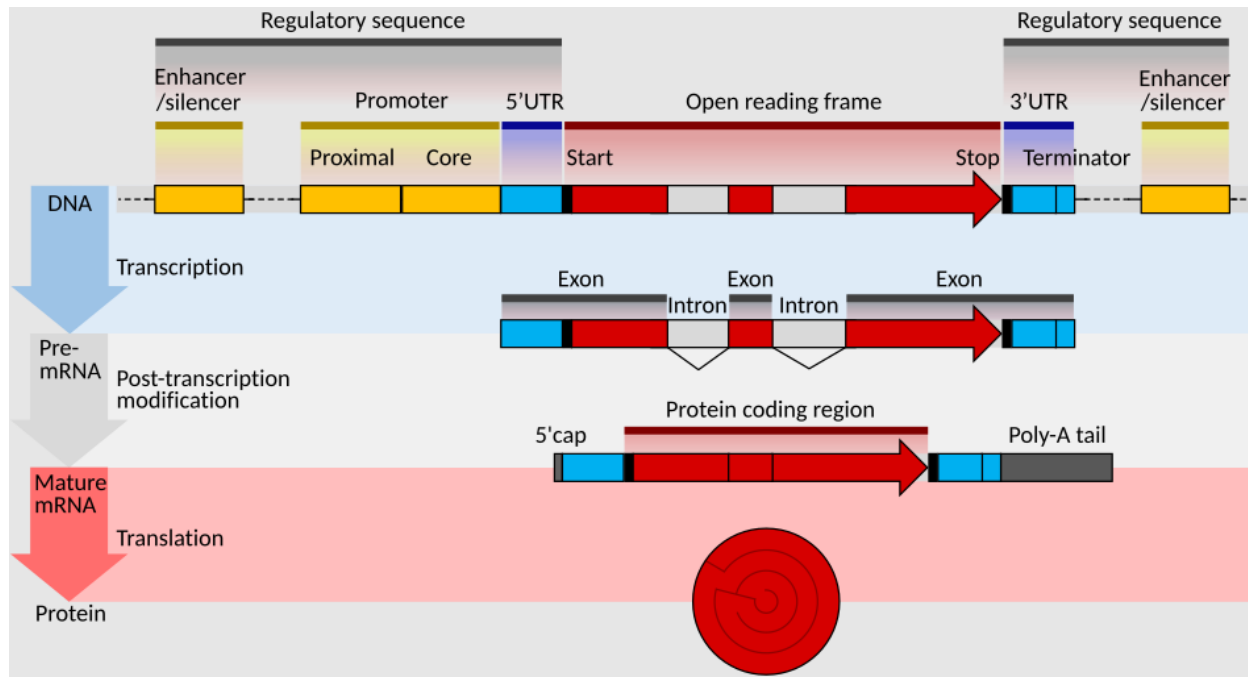
An episodic burst of massive genomic rearrangements and the origin of non-marine annelids

[Carlos Vargas-Chávez](#), [Lisandra Benítez-Álvarez](#), [Gemma I. Martínez-Redondo](#), [Lucía Álvarez-González](#), [Judith Salces-Ortiz](#), [Klara Eleftheriadi](#), [Nuria Escudero](#), [Nadège Guiguielmoni](#), [Jean-François Flot](#), [Marta Novo](#), [Aurora Ruiz-Herrera](#), [Aoife McLysaght](#) & [Rosa Fernández](#)

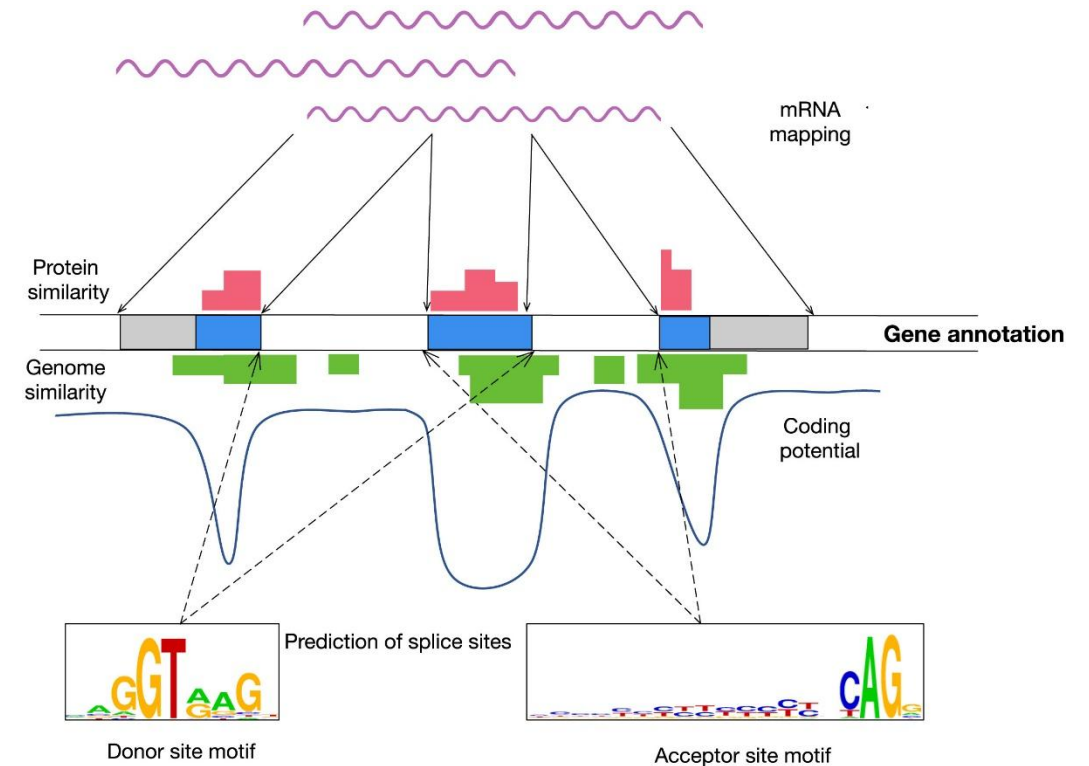
[Nature Ecology & Evolution](#) 9, 1263–1279 (2025) | [Cite this article](#)

Structural annotation in eukaryotic genomes

- Subdivided into exon and intron (spliced out during post-transcriptional processing)
- Transposable elements, non-coding RNAs, pseudogenes



<https://commons.wikimedia.org/>



Guigó R. **Genome annotation: From human genetics to biodiversity genomics.** *Cell Genomics*, 2023

=> Gene prediction of eukaryotic genomes requires dedicated workflows



Eukaryotic gene prediction is challenging

The "average" eukaryotic gene does not exist

Every characteristic of a gene can vary by several orders of magnitude

Feature	Typical range
Gene length	100 bp → >2 Mb
Number of exons	1 → >300
Exon length	<10 bp → several kb
Intron length	<30 bp → >1 Mb
Isoforms	1 → dozens
UTR length	0 → several kb
Expression	very low → extremely high

No single prediction model can accurately describe all eukaryotic genes

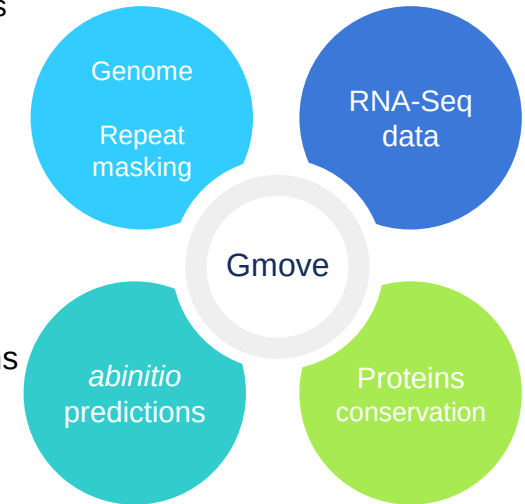
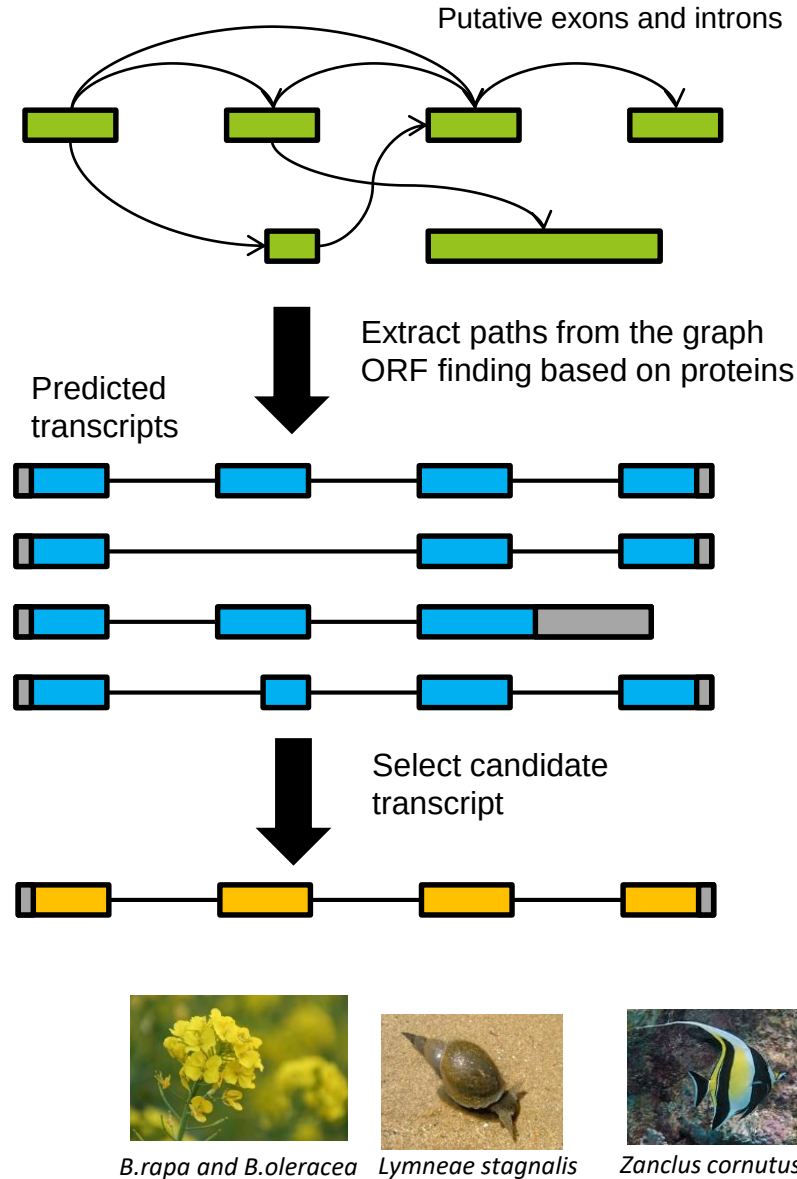


Genoscope workflow for eukaryotic gene prediction

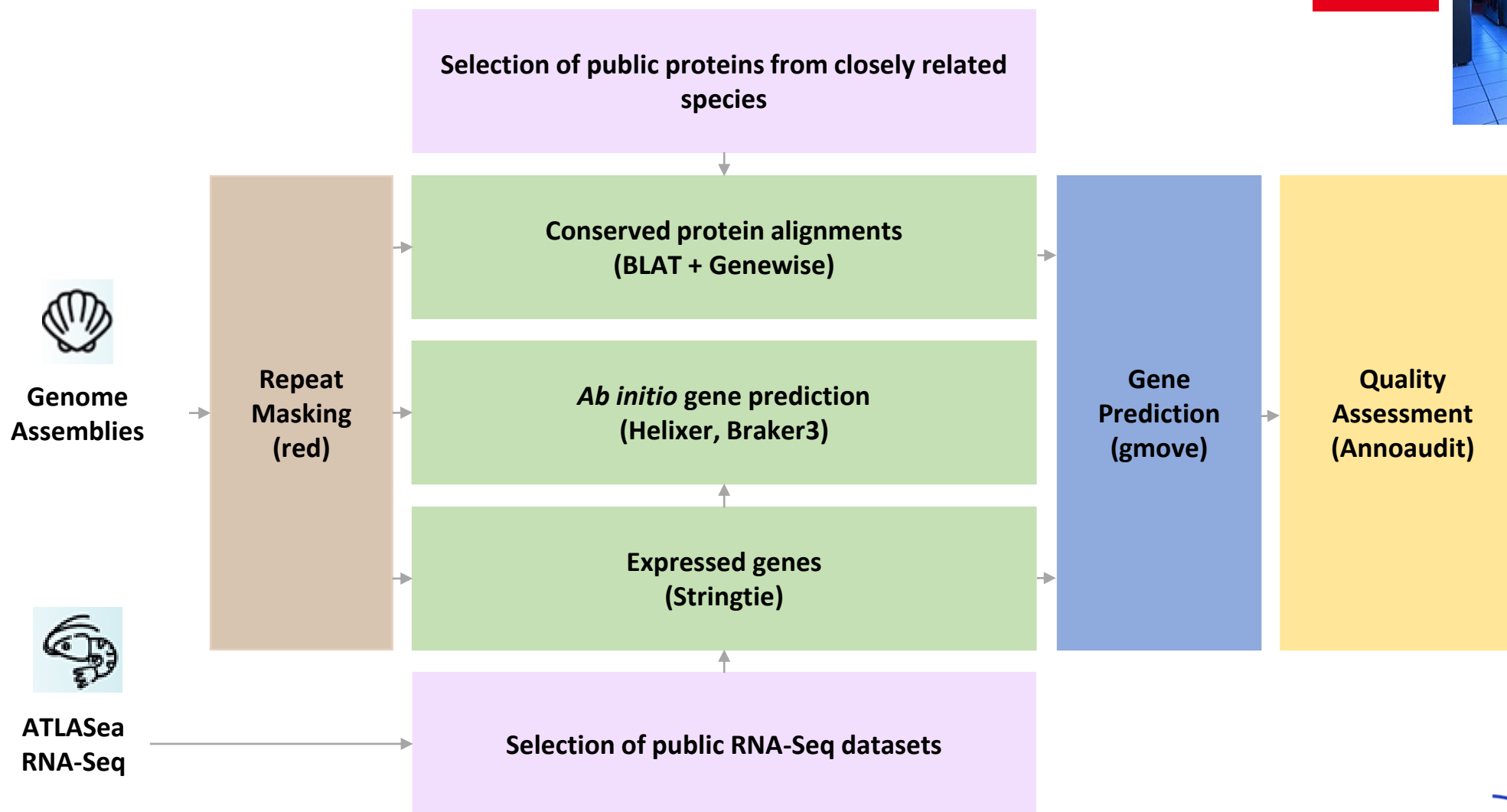
- We developed a workflow to predict coding genes in eukaryotic genomes
- Data integration is performed using in-house tool, named gmove
- Gmove is calibration-free and can be applied to different organisms



<http://www.genoscope.cns.fr/gmove>



Genoscope workflow for eukaryotic gene prediction



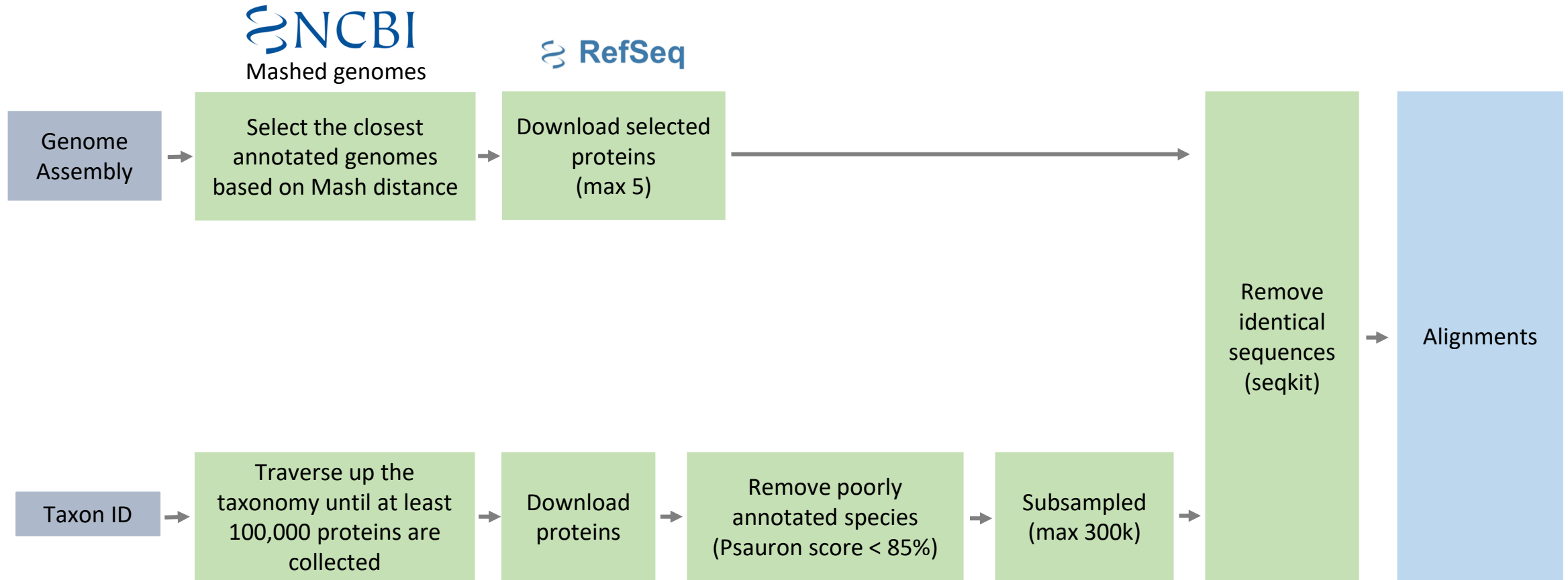
Computing



TGCC
127,360 CPUs, 995 nodes
9,600 GPUs, 75 nodes



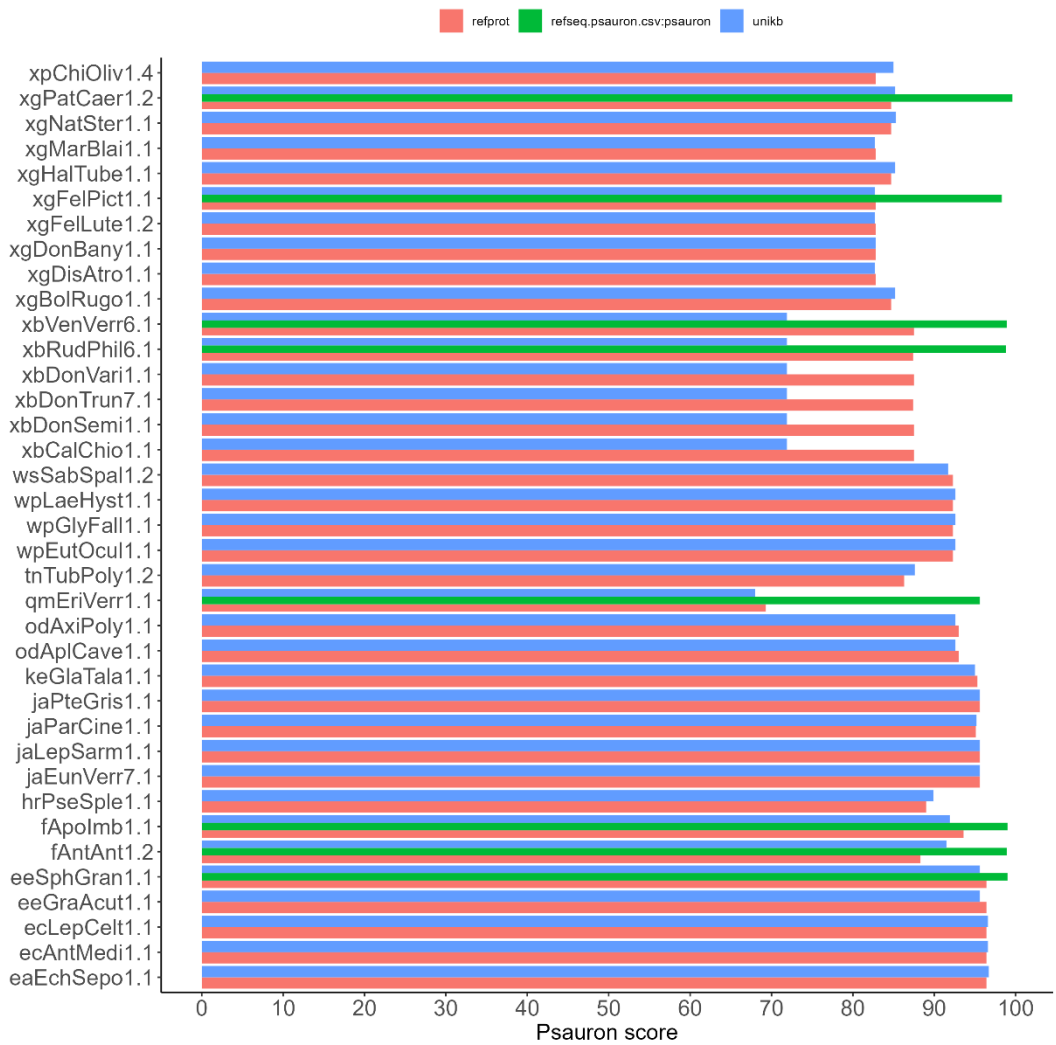
Selection of protein evidences to identify conserved genes



UniProtKB
Ref. Proteomes

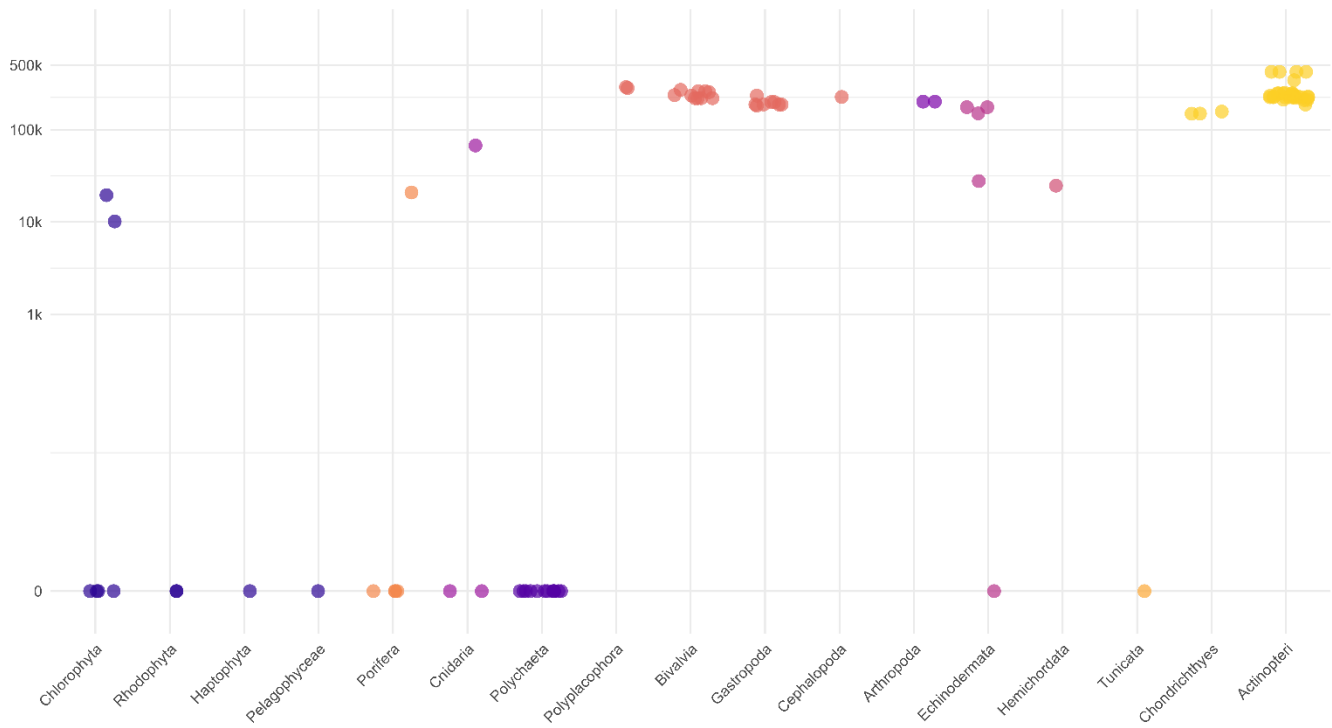


Selected protein sequences for gene prediction



RefSeq annotations are generally of higher quality

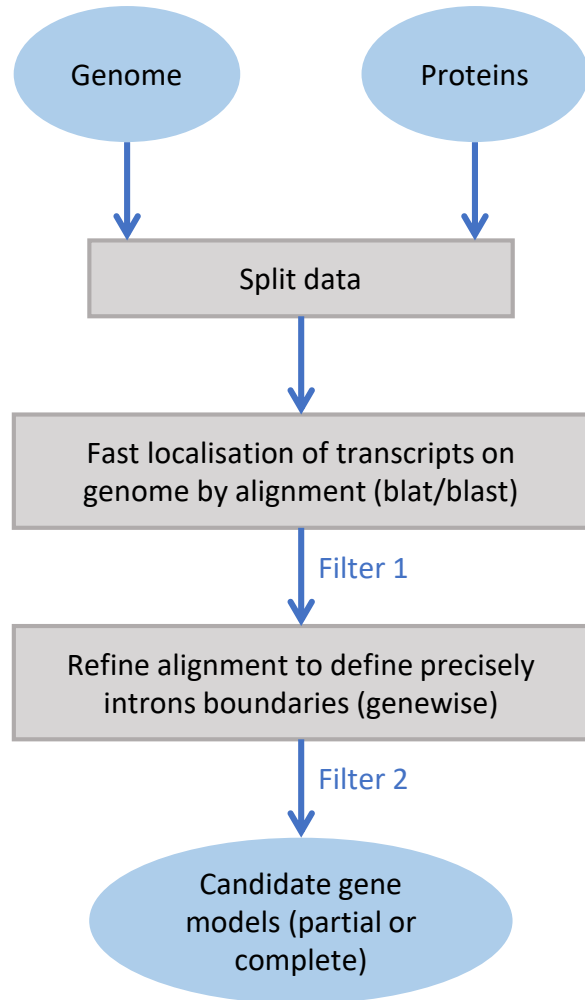
Number of downloaded RefSeq proteins



Some phylogenetic groups have few RefSeq annotations



Alignment of selected proteins



Alignment of selected proteins

nature
plants

Letter | OPEN | Published: 18 June 2018

Oak genome reveals facets of long lifespan

Christophe Plomion✉, Jean-Marc Aury, Joëlle Amselem, Thibault Leroy, Florent Murat, Sébastien Duplessis, Sébastien Faye, Nicolas Francillon, Karine Labadie, Grégoire Le Provost, Isabelle Lesur,

Noel et al. *Genome Biology* (2023) 24:123
https://doi.org/10.1186/s13059-023-02960-7

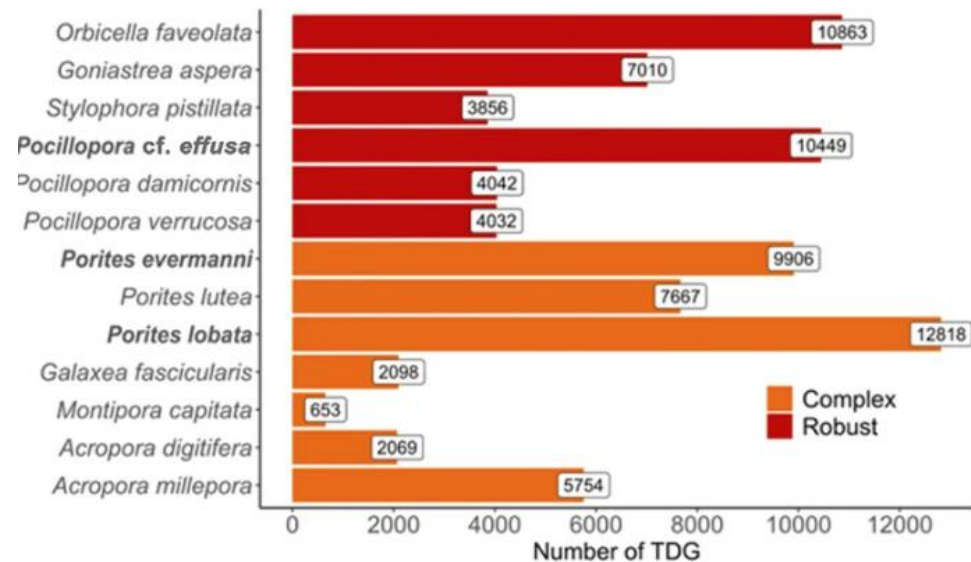
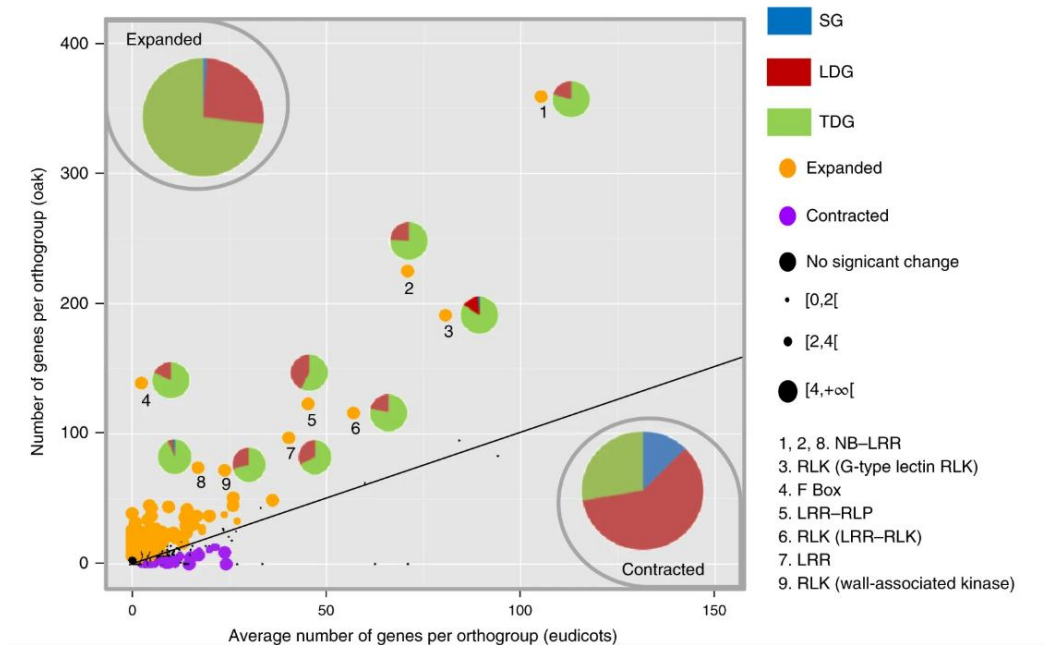
Genome Biology

RESEARCH

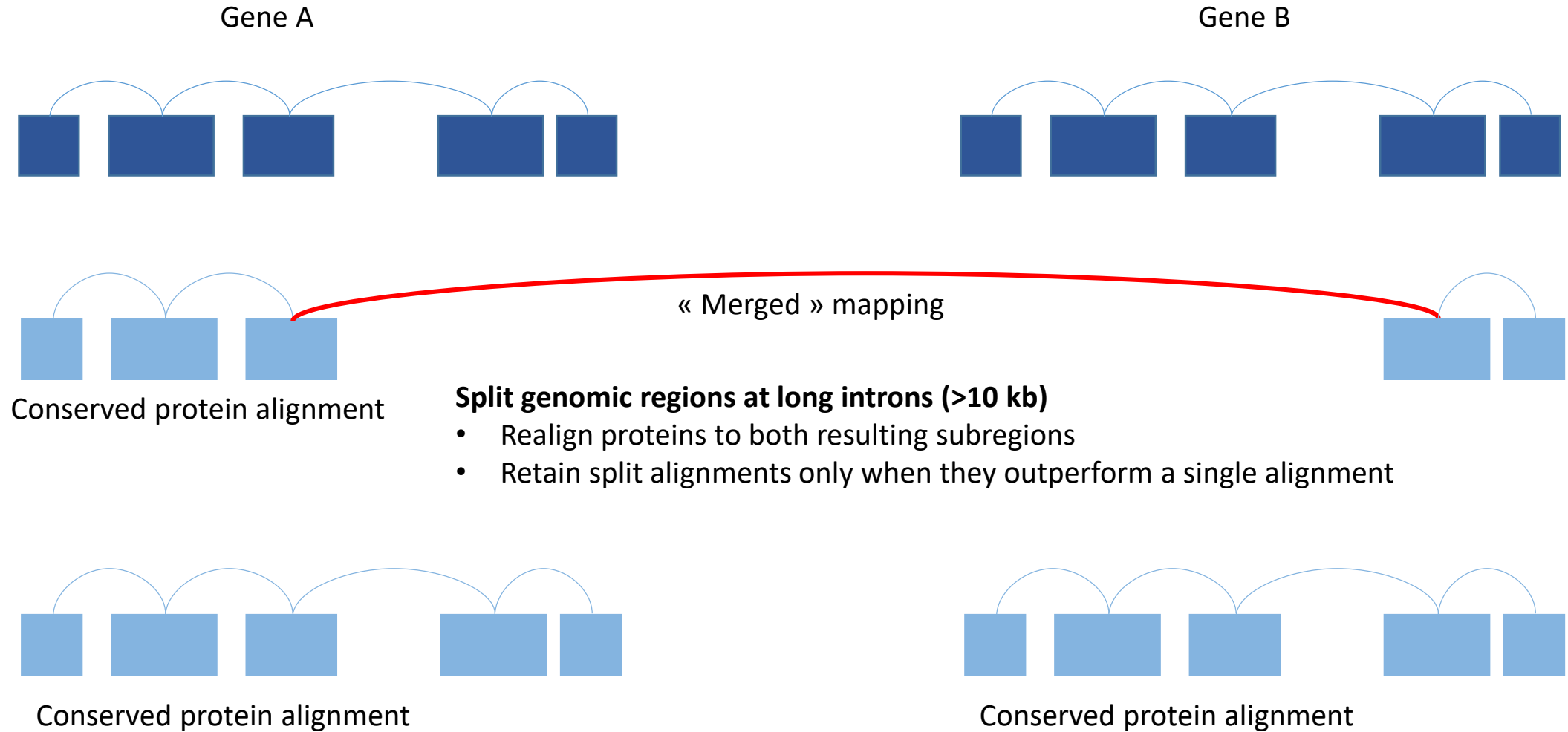
Open Access

Pervasive tandem duplications and convergent evolution shape coral genomes

Benjamin Noel^{1,2†}✉, France Denoeud^{1,2†}✉, Alice Rouan^{3,4}✉, Carol Buitrago-López⁵✉, Laura Capasso^{4,6,7}✉, Julie Poulain^{1,2}✉, Emilie Boissin⁸✉, Mélanie Pousse^{3,4}✉, Corinne Da Silva^{1,2}✉, Arnaud Couloux^{1,2}✉, Eric Armstrong^{1,2}✉, Quentin Carradec^{1,2}✉, Corinne Cruaud^{2,9}✉, Karine Labadie^{2,9}✉, Julie Lê-Hoang^{1,2}✉, Sylvie Tambutti^{4,6}✉, Valérie Barbe^{1,2}✉, Clémentine Moulin^{2,10}✉, Guillaume Bourdin¹¹✉, Guillaume Iwankow⁸✉, Sarah Romac¹²✉, Sylvain Agostini¹³✉, Bernard Banaigs⁸✉, Emmanuel Boss¹¹✉, Chris Bowler^{2,14}✉, Colomán de Vargas^{2,12}✉, Eric Douville¹⁵✉, J. Michel Flores¹⁶✉, Didier Forcioli^{3,4}✉, Paola Furla^{3,4}✉, Pierre E. Galand^{2,17}✉, Fabien Lombard^{2,18,19}✉, Stéphane Pesant²⁰✉, Stéphanie Reynaud^{4,6}✉, Matthew B. Sullivan²¹✉, Shinichi Sunagawa²²✉, Olivier P. Thomas²³✉, Romain Troublé^{2,10}✉, Rebecca Vega Thurber²⁴✉, Denis Allemand^{4,6}✉, Serge Planes^{2,8}✉, Eric Gilson^{3,4,25}✉, Didier Zoccola^{4,6}✉, Patrick Wincker^{1,2}✉, Christian R. Voolstra⁵✉ and Jean-Marc Aury^{1,2†}✉



Alignment of selected proteins: the case of tandemly duplicated genes



Alignment of selected proteins

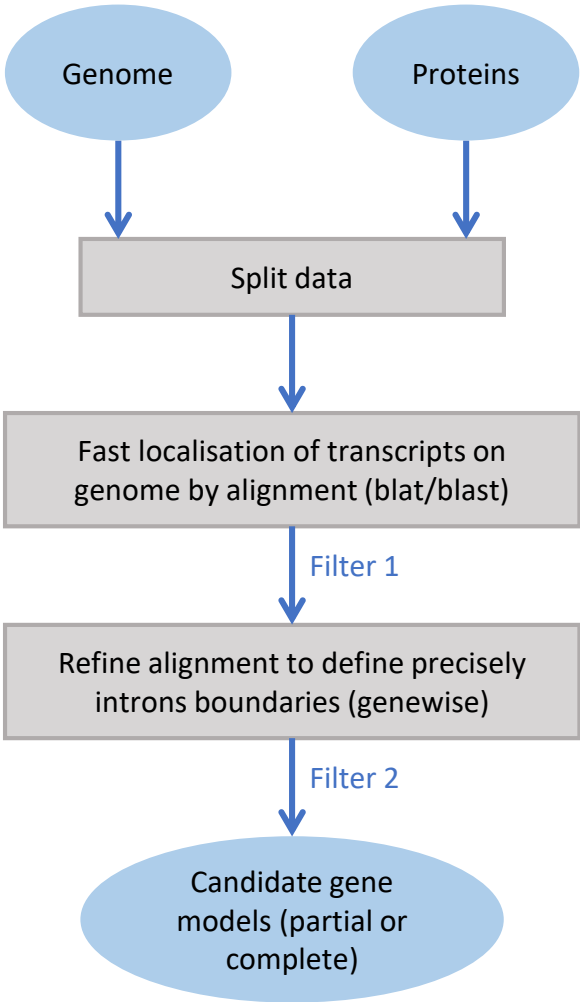
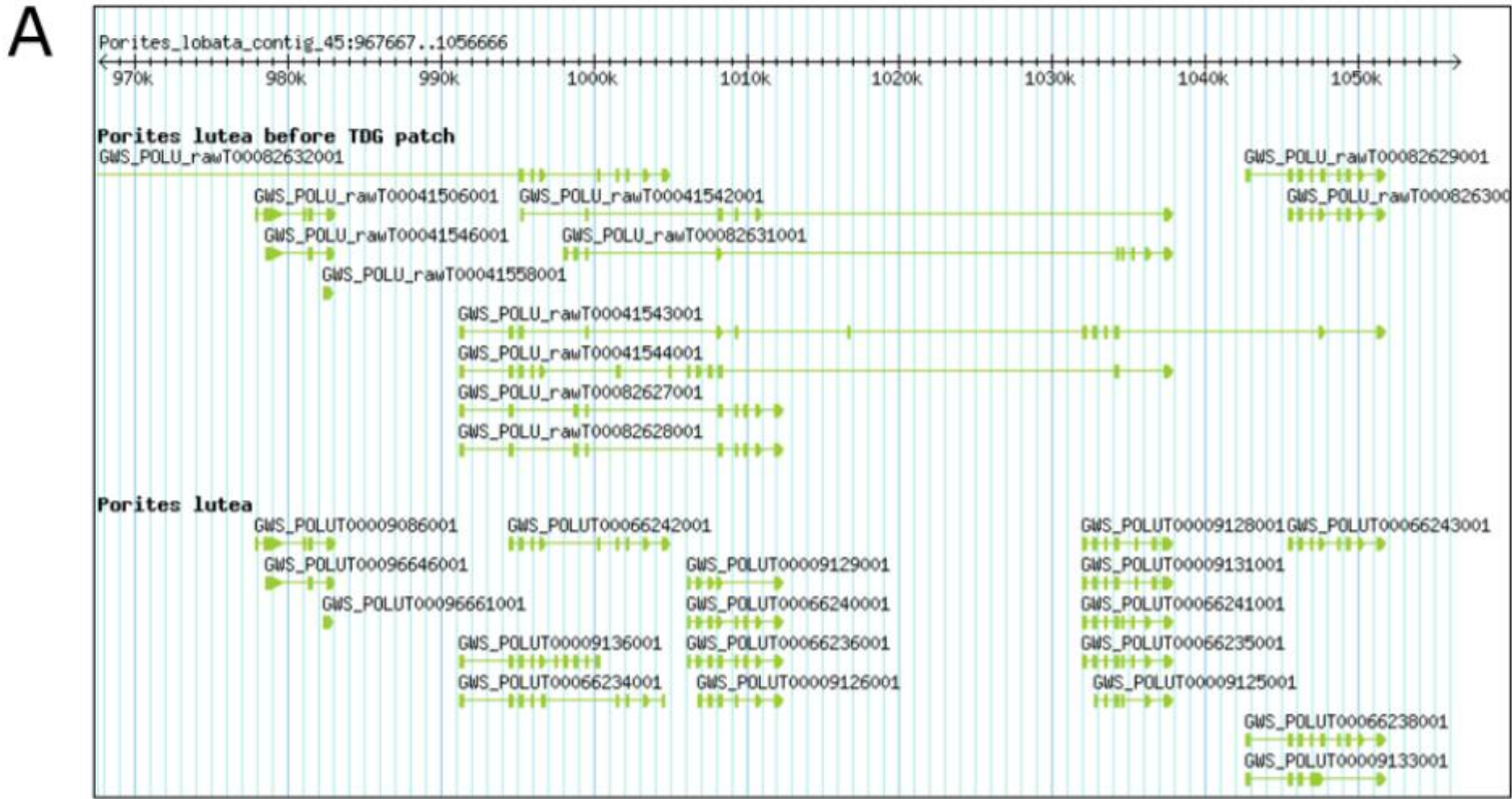
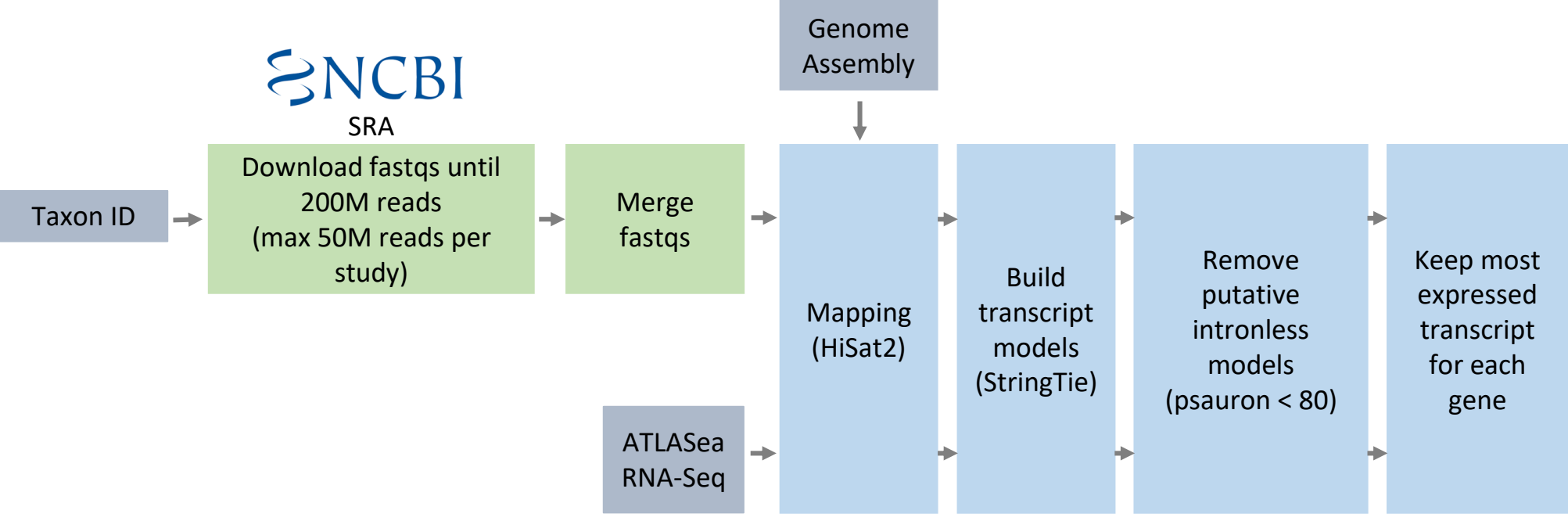


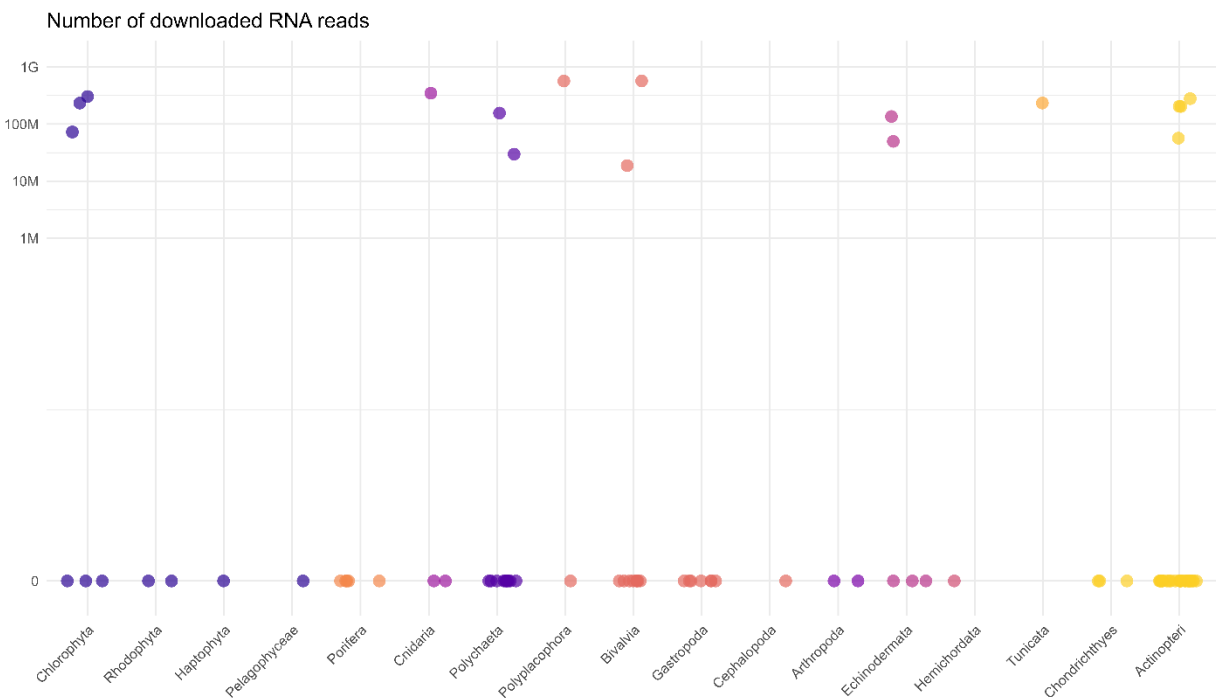
Figure S10. Two *P. lobata* genomic regions that illustrate proteins (A, green tracks) and transcripts (B, orange tracks) alignments before (middle track) or after (down track) the adaptation of gene prediction workflow for tandemly duplicated genes (“TDG patch”).



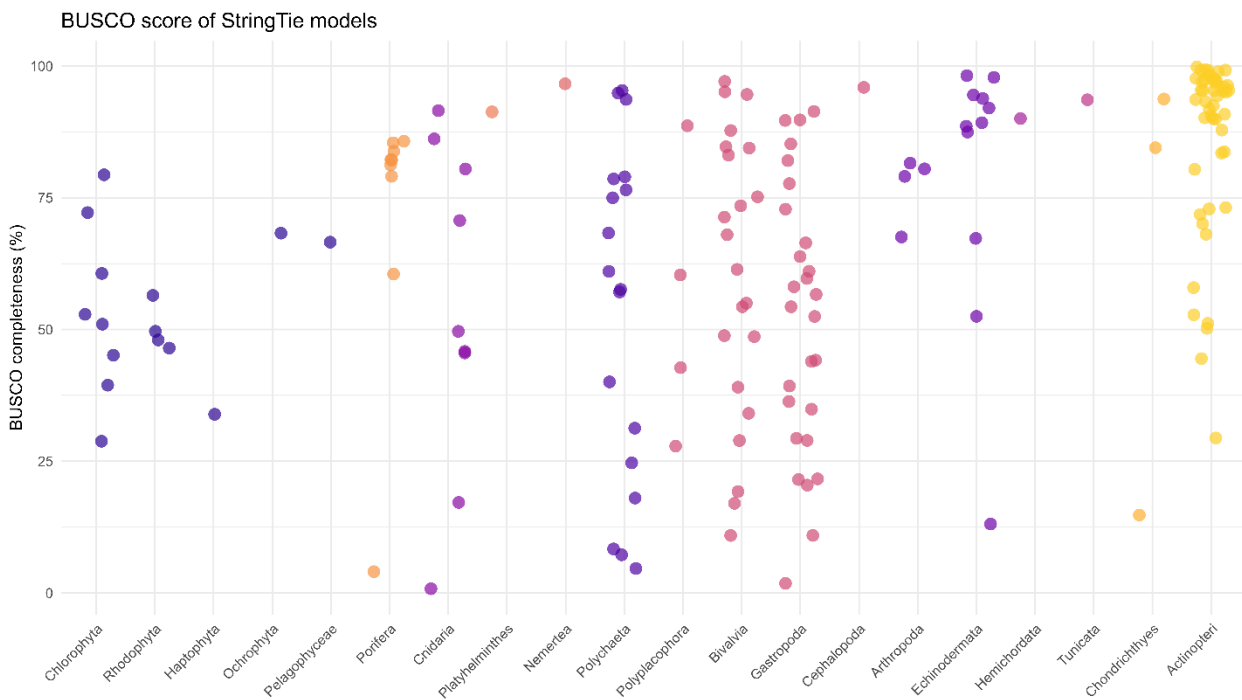
RNAseq data to detect expressed genes



RNAseq data to detect expressed genes



Only a few genomes have publicly available RNA-Seq data

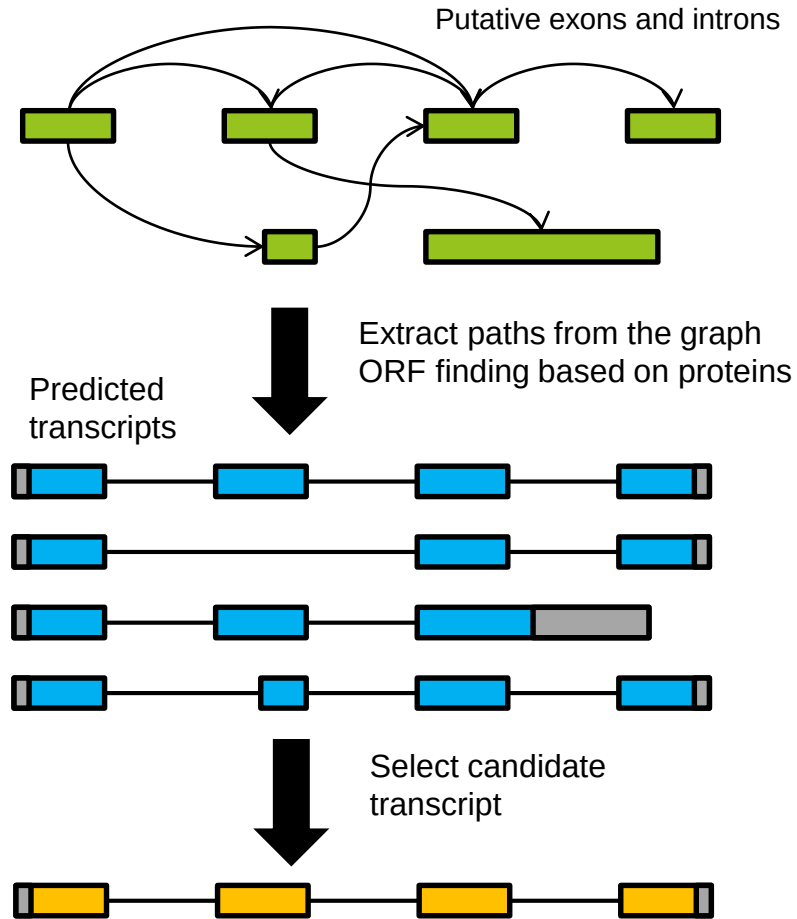
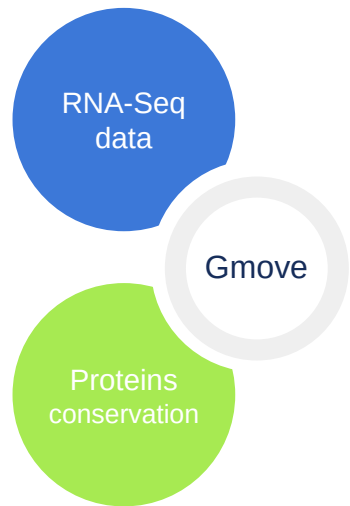


RNA-Seq data quality is low for some phylogenetic groups



Gmove – Gene Modeling using Various Evidence

<http://www.genoscope.cns.fr/gmove>



Protein-coding gene models (≥ 2 criteria required)

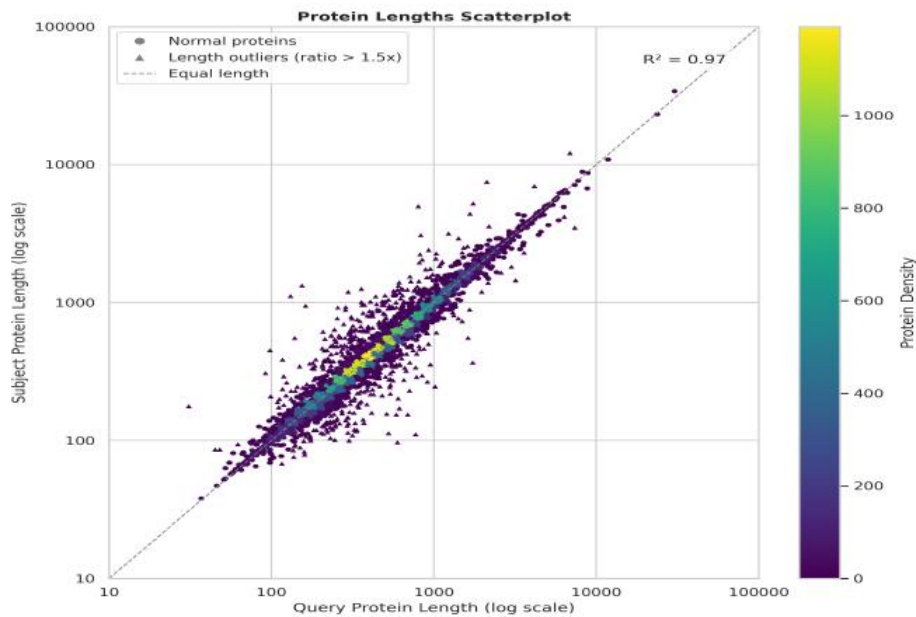
- CDS ≥ 300 nt
- Coding ratio $\geq 20\%$
- ≥ 1 Pfam domain



Quality assessment of gene predictions

Genome Annotation Report

Taxon Info	
TxID	357811
Species	Apogon imberbis
Class	Actinopteri
Order	Kurtiformes



Annoaudit workflow



<https://github.com/ERGA-consortium/AnnoAudit>

- General metrics (number of genes, coding sizes...)
- Orthologous genes (BUSCO, OMArk, BRH)
- Coding assessment (psauro)
- RNAseq support

JOURNAL ARTICLE

A General Tendency for Conservation of Protein Length Across Eukaryotic Kingdoms FREE

[Save](#) [Related Papers](#) [Chat with paper](#)

[Daryi Wang](#), [Mufen Hsieh](#), [Wen-Hsiung Li](#)

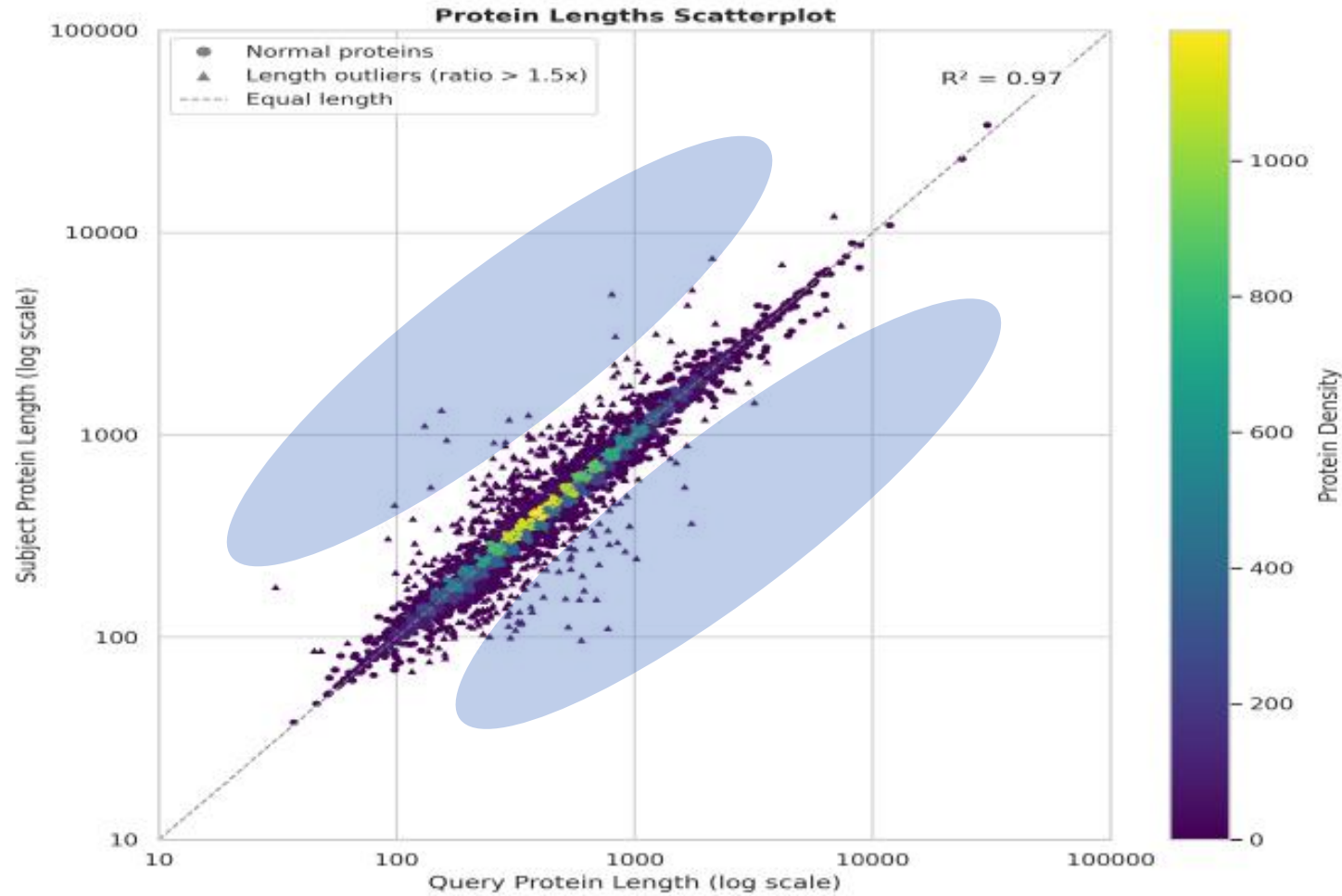
Molecular Biology and Evolution, Volume 22, Issue 1, January 2005, Pages 142–147,

<https://doi.org/10.1093/molbev/msh263>

Published: 15 September 2004 [Article history](#)



Quality assessment of gene predictions



Estimate proportion of potential split and fused genes

Reference annotation if:

- Psauron ≥ 0.90
- Δ BUSCO (complete) ≥ -10
- Split + fusion $\leq 30\%$

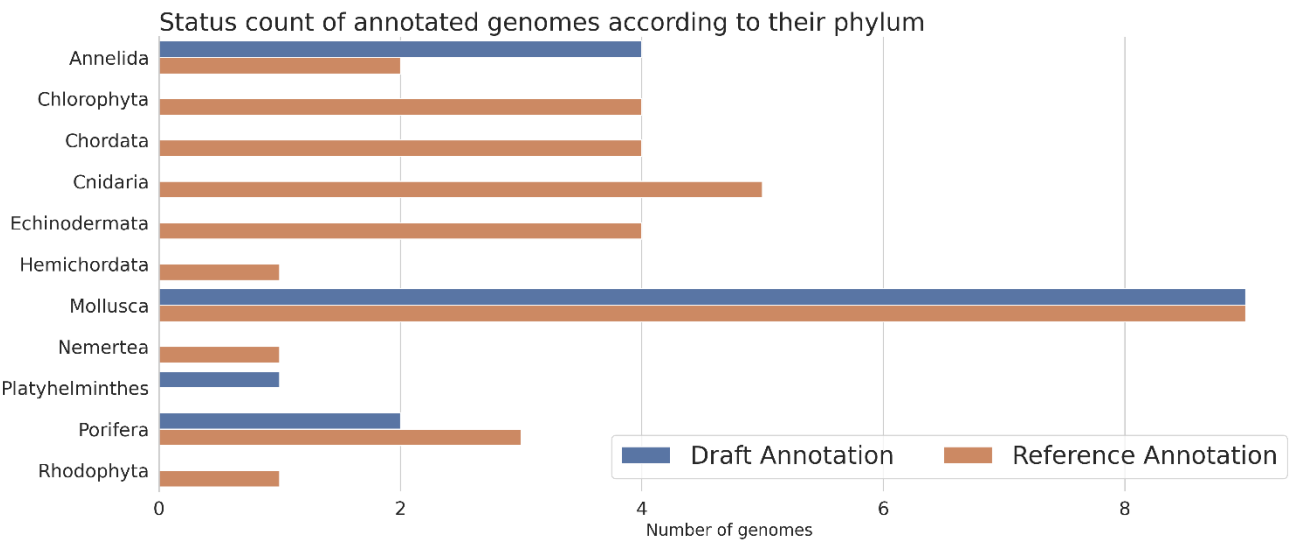
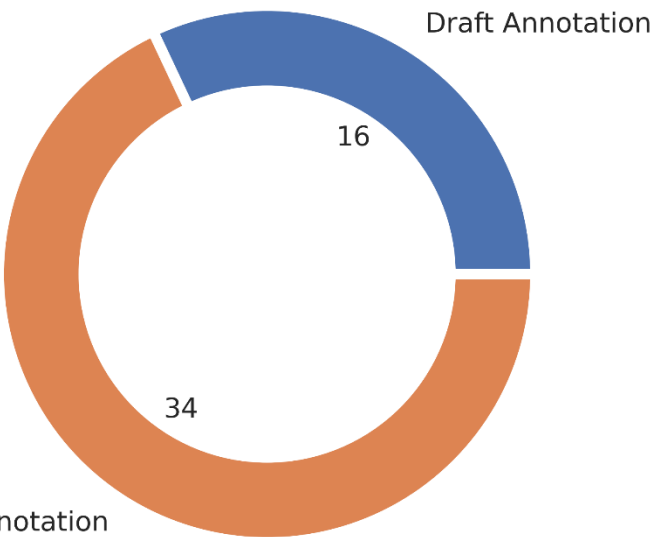
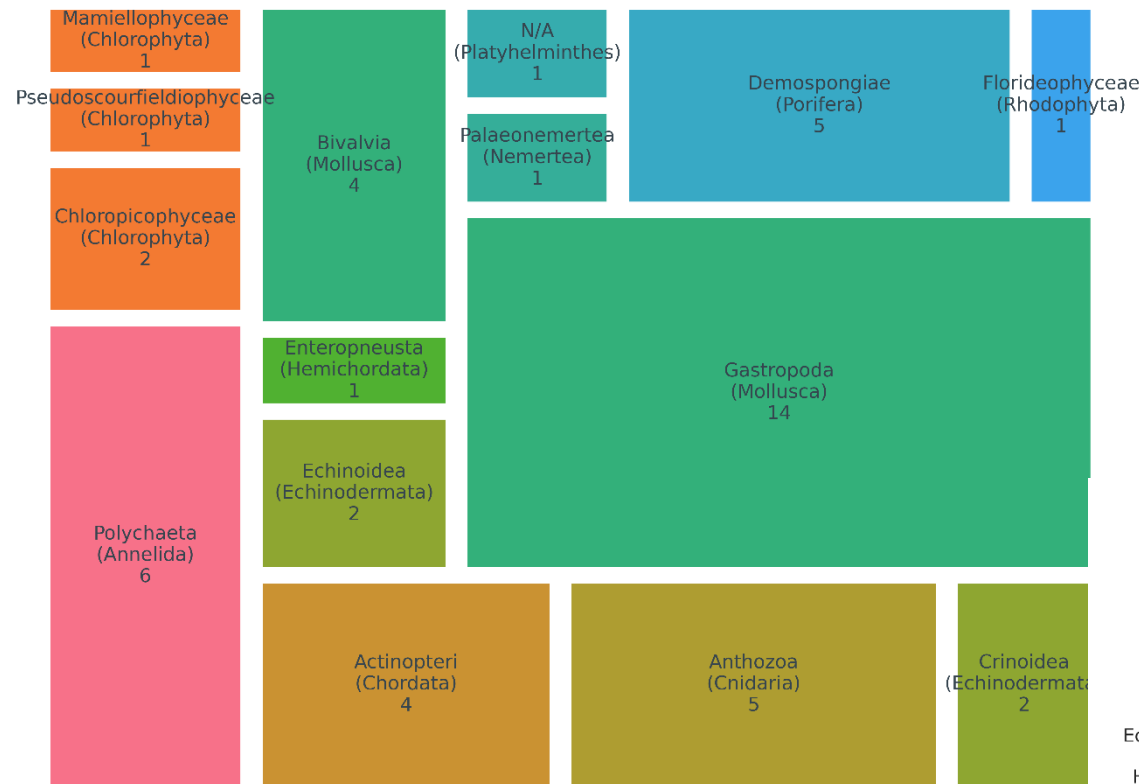
Otherwise: draft annotation



ATLASea: Where are we today?

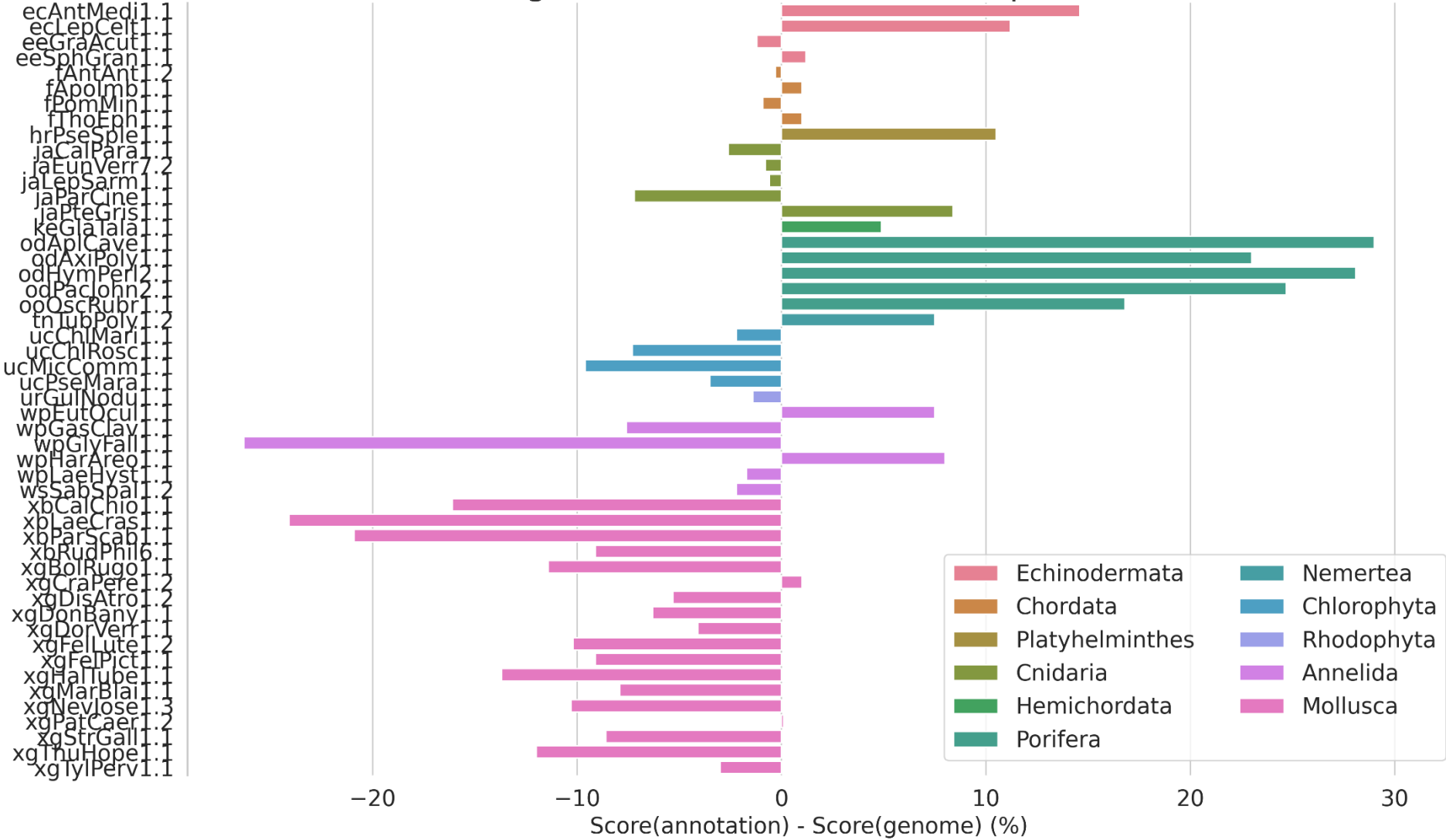
Number of annotated genomes : 50

Count of annotated clades



ATLASea: Where are we today?

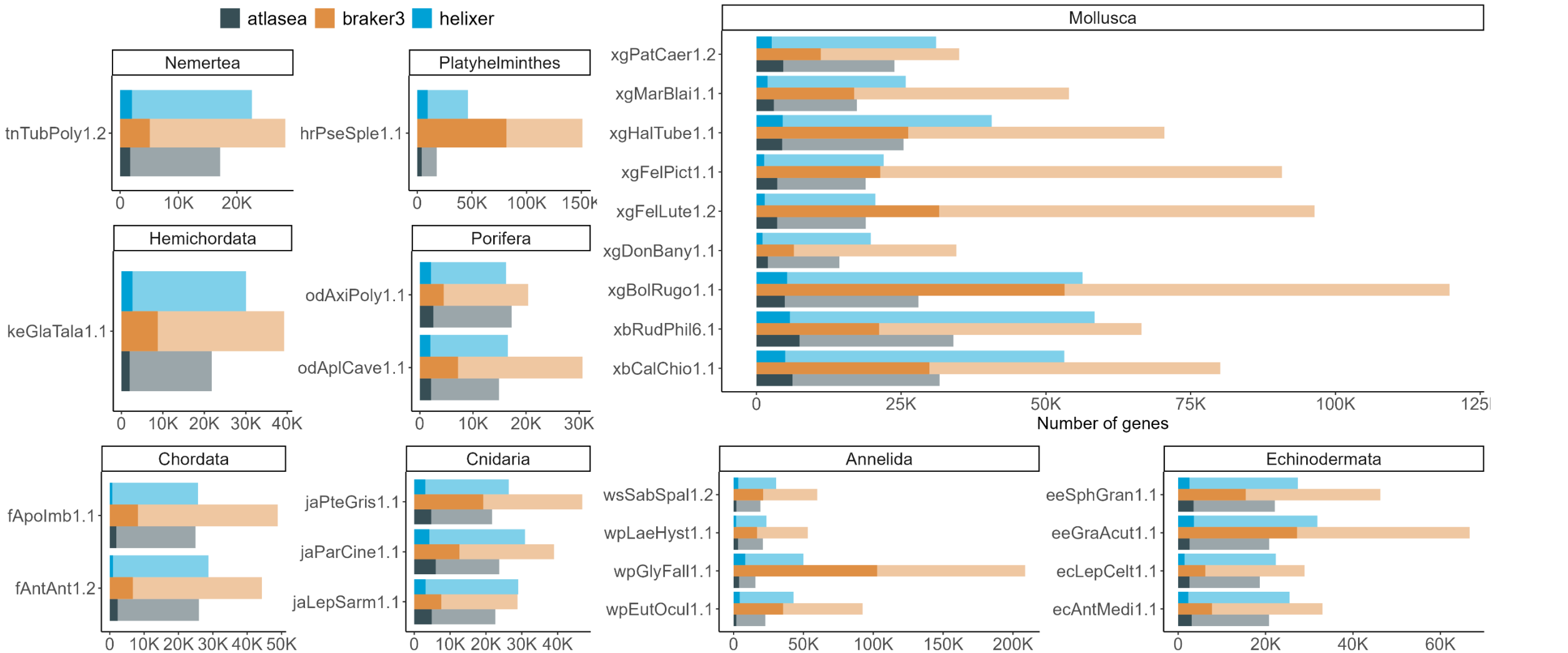
Difference between genome and annotation complete busco scores



Quality of a gene dataset seems to be correlated to the phylum of the annotated species



ATLASea: Where are we today?



Take home message

- Over 2,500 samples received at Genoscope
- Over 1,000 species sequenced using short reads technology
- 300 genome assemblies submitted to public repositories
- Annotation workflow has been validated on the first 50 genome assemblies

Next steps:

- Annotate genomes as soon as both the genome assembly and RNA-Seq data are available
- Release annotated genomes through the ATLASEa data portal
- Implement large-scale reannotation workflows to leverage the growing collection of ATLASEa genomes



Acknowledgements



Benjamin Noel



Khaoula Ziane



France Denoeud



Corinne Da Silva



Marc Wessner



Phuong Doan



www.atlasea.fr

Coordination committee:

H. Roest Crolius
P. Wincker
B. Bed'Hom

J-M. Aury
E. Corre
C. Chapperon

Administration (CNRS):

C. Ferras
S. Djian

S. Capelle
N. Plantey
C. Bousquet

WHEEL-Sea:

H. Roest Crolius
P. Wincker
S. Bertrand
S. Mona
M. Origas

DIVE-Sea:

B. Bed'Hom
L. Le Gall
T. Perez
S. Bertrand
B. Ernande
L. Guidi

M. Travers-Trollet
G. Illien
F. Clément
F. Bellugeon
B. Girard
S. Chagnoux
I. Probert

K. Boujbara
J-L Jung
P. Gourvil
C. Bachy
R. Garnier
D. Vaultot

SEQ-Sea:

J-M. Aury
C. Scarpelli
P. Oliveira
J. Rigonato
C. Menguy
E. Jacoby
K. Labadie
C. Cruaud
A. Moussy

J. Poulain
F. Gavory
S. Mangenot
E. Teodori
S. Engelen
A. Ndar
B. Istace
L. Demirdjian

A. Couloux
B. Noel
S. Duprat
Y. Dussert
P. Doan
P. Mielle
K. Ziane
.....

BYTE-Sea:

E. Corre
P. Durand
A. Breteau
A. Louis
L. Guegen

A. Cormier
O. Sallou
G. Le Corguille
O. Collin
L. Leroi
P. Aufret



PROGRAMME
DE RECHERCHE
GÉNOMES MARINS



Questions ?