BUSCO

An interactive guide

Matthew Berkeley, 14th October 2024

Introduction

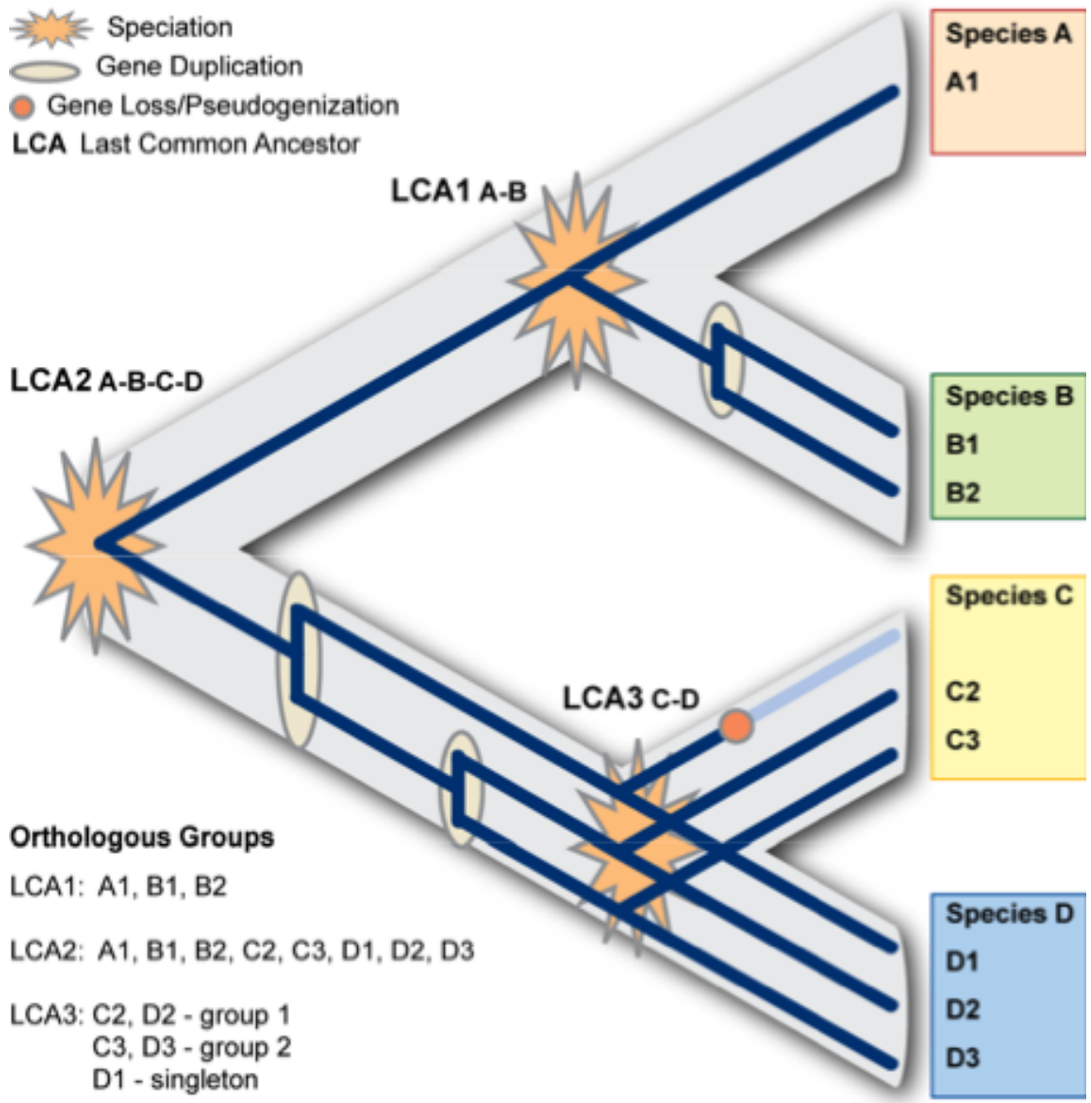
Background

- **2012** BA Physics & Astrophysics, Trinity College Dublin
- **2013** MSc Space Studies, International Space University
  - Summer internship at NASA GSFC, X-ray Astrophysics
- **2015** MS Physics, Catholic University of America
- **2018** PhD Physics, Catholic University of America
  - Research with NASA GSFC Cosmology Division
- **2018 - present** Software Developer, Swiss Institute of Bioinformatics
  - Lead developer for BUSCO

Orthology

Orthology is a relative concept

referring to *the last common ancestor (e.g. of the considered species)*



A group of genes that can trace their origin to a single ancestral gene form an **orthologous group**

Rob Waterhouse at BGA 2023
https://www.youtube.com/watch?v=Q0T4DLGFSNQ

https://www.orthodb.org

VPN Update

OrthoDB v12.0

UNIVERSITÉ DE GENÈVE
FACULTÉ DE MÉDECINE

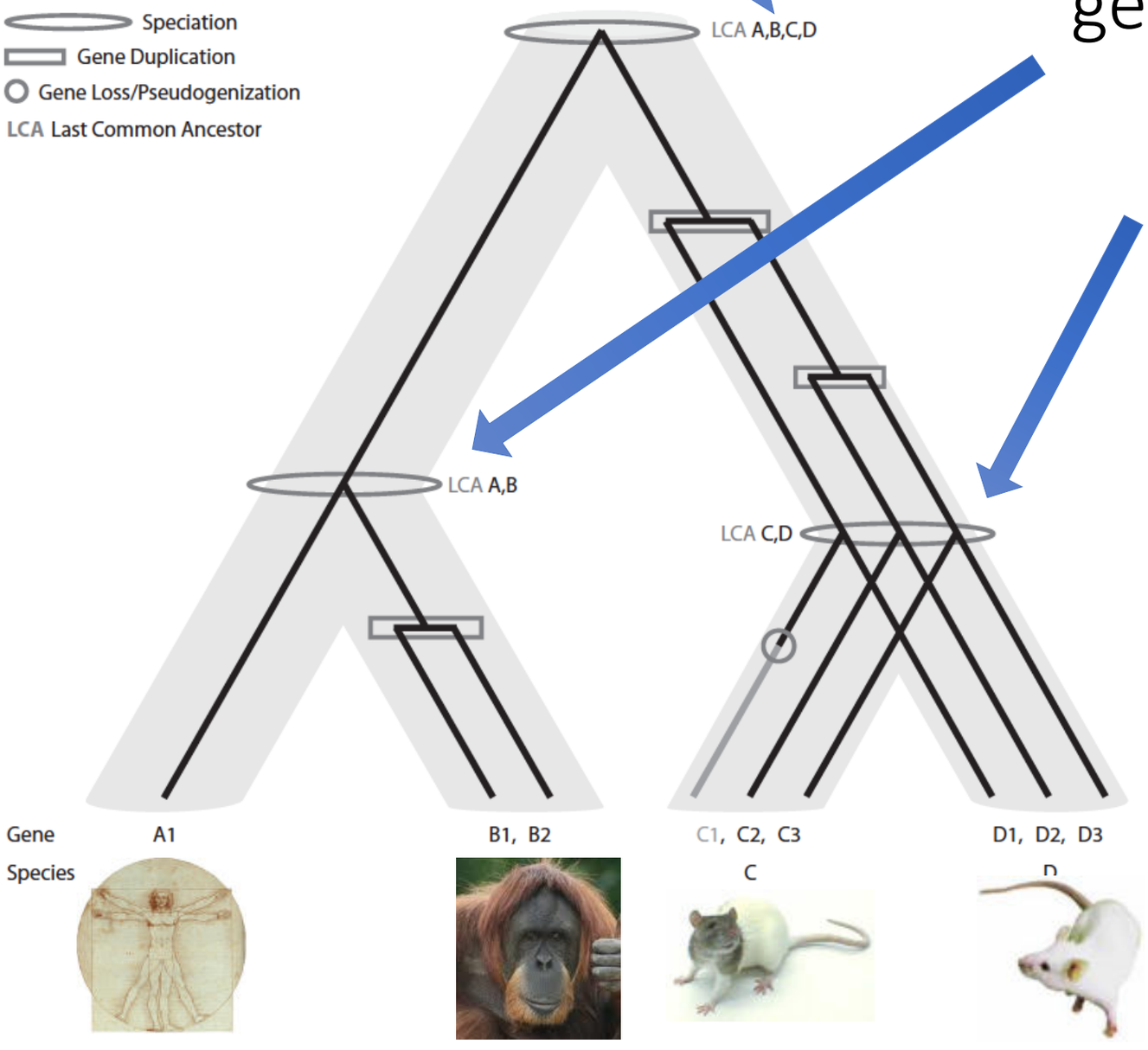
SIB Swiss Institute Bioinformatics

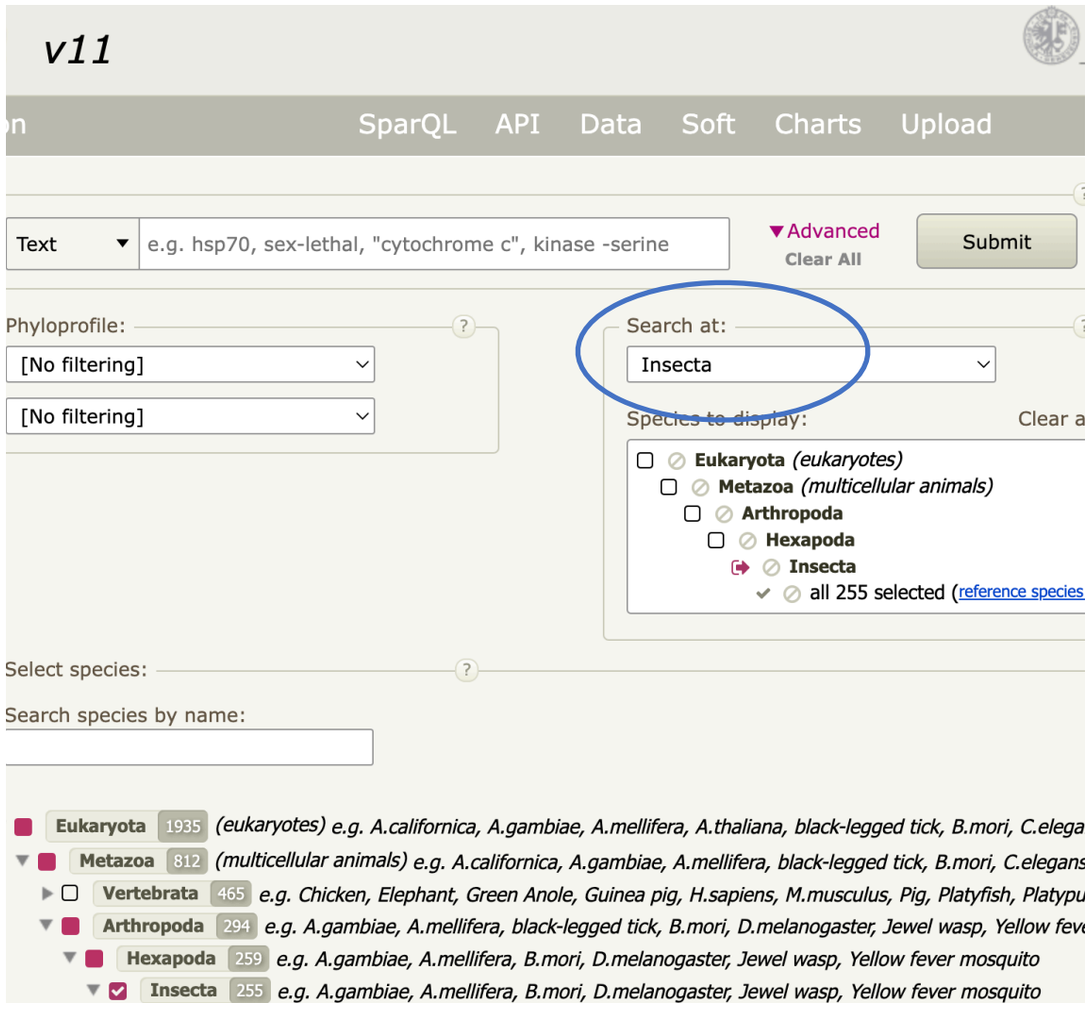
Text | e.g. hsp70, sex-lethal, "cytochrome c", kinase -serine | ▶ Advanced | Submit

The hierarchical catalog of orthologs
mapping genomics to functional data

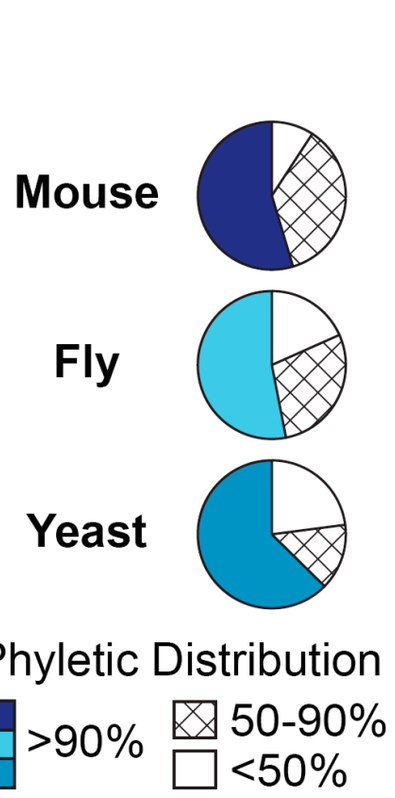
Eukaryotes	Prokaryotes	Viruses	Genes
5,827	**18,158**	**7,962**	**162M**

Hierarchical – means we resolve gene orthology at different levels

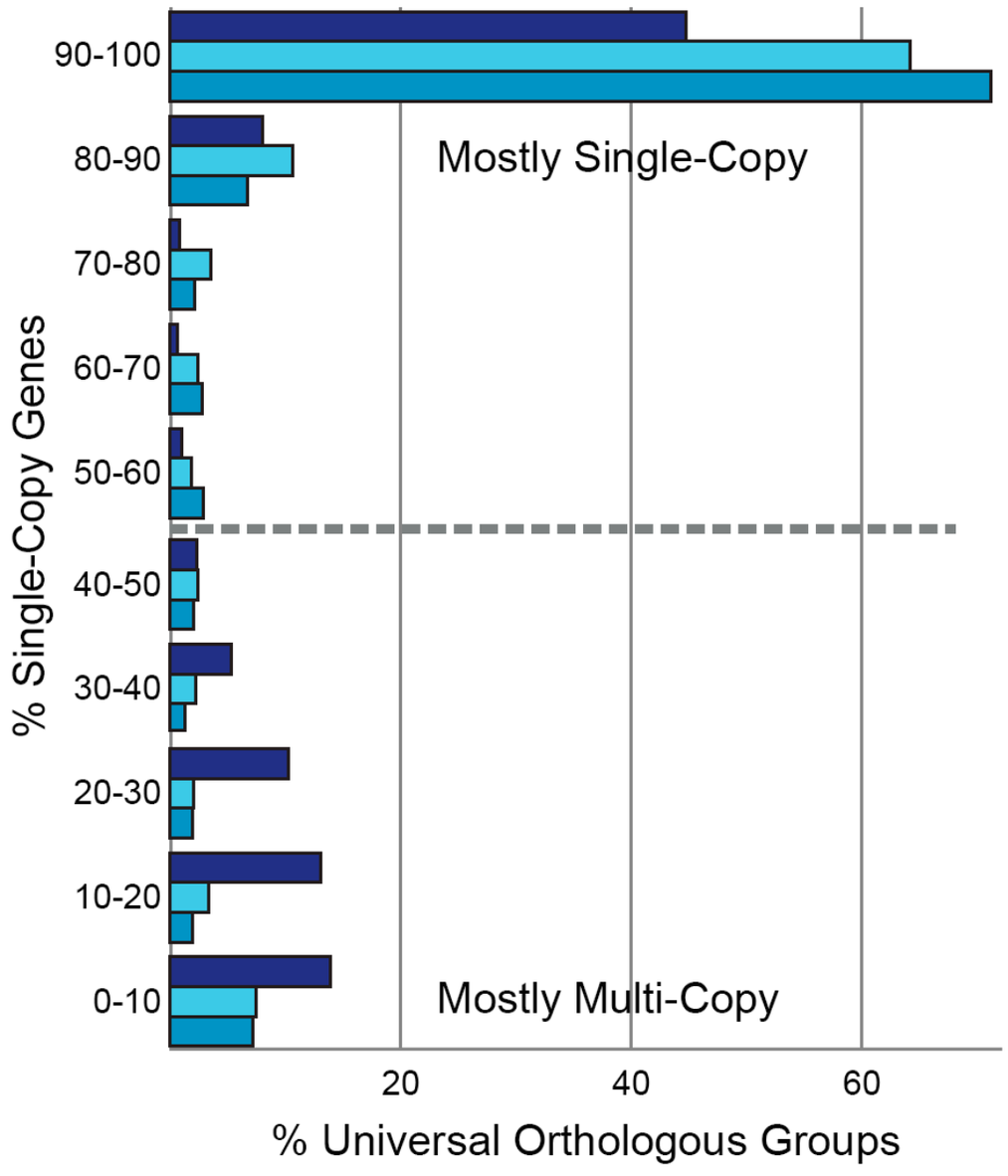




Copy-number distribution



Over half of OGs are universal

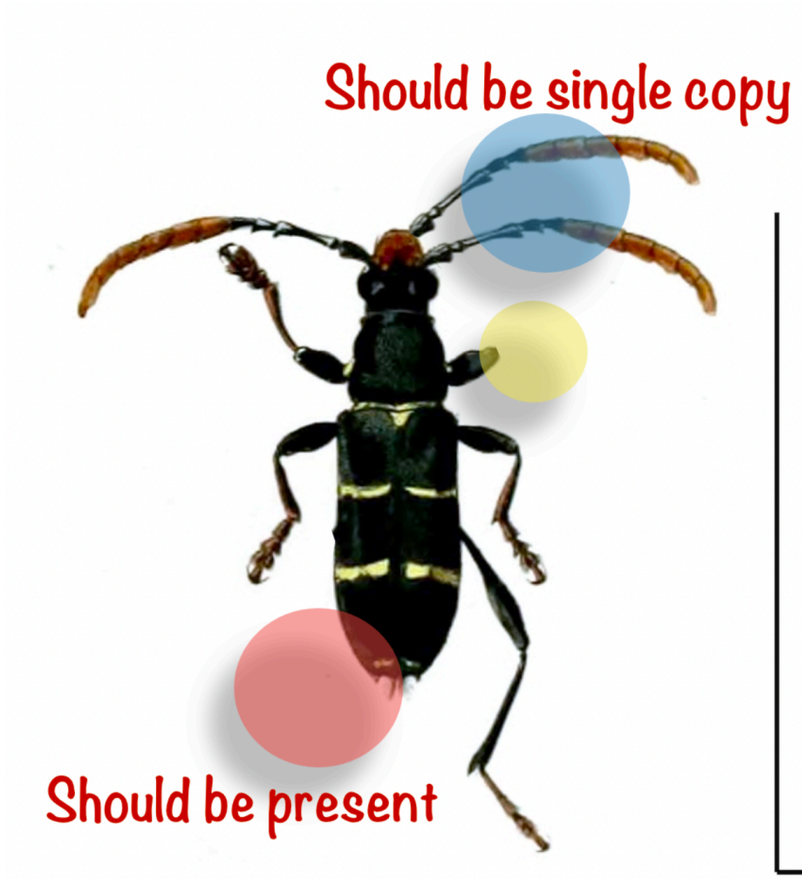


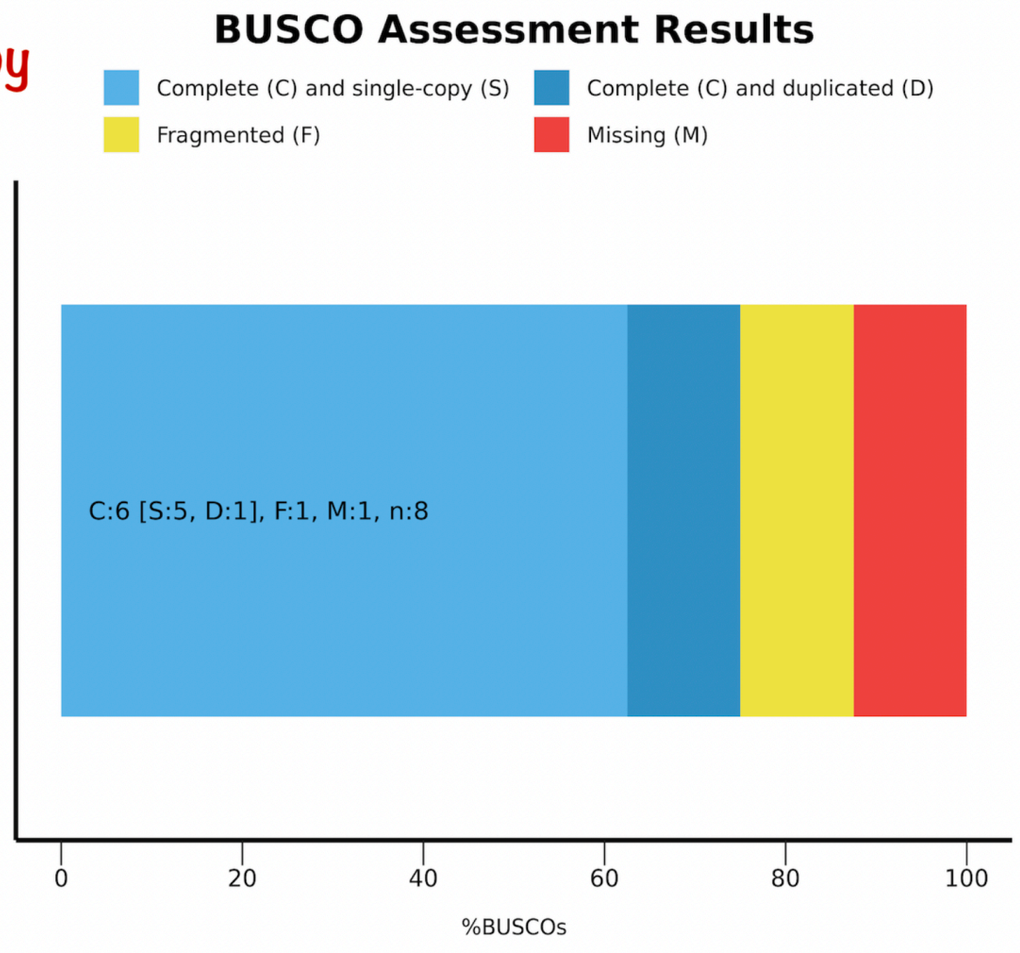
Waterhouse RM, Zdobnov EM, Kriventseva EV. **Genome Biol Evol.** 2011 Jan;3:75-86.

BUSCO

BUSCO

Benchmarking Universal Single Copy Orthologs





BUSCO

Benchmarking **U**niversal **S**ingle **C**opy **O**rthologs

- Provides Completeness assessment of newly assembled genomes, proteomes and transcriptomes
- Evolutionary-based approach, built on data from OrthoDB
- First published in 2015
- Has become an integral part of all major genome assembly pipelines

BUSCO

Google Scholar

BUSCO@EZlab

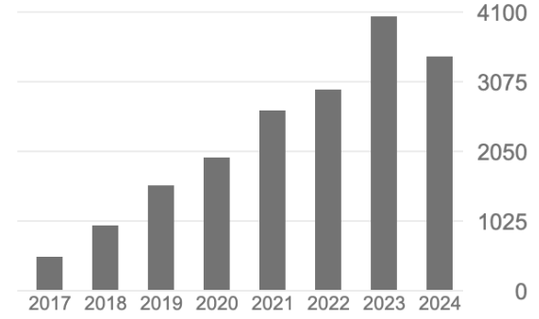
Unknown affiliation
No verified email

FOLLOW

TITLE	CITED BY	YEAR
BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs FA Simão, RM Waterhouse, P Ioannidis, EV Kriventseva, EM Zdobnov Bioinformatics 31 (19), 3210-3212	10970	2015
BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes M Manni, MR Berkeley, M Seppey, FA Simão, EM Zdobnov Molecular Biology and Evolution	3037	2021
BUSCO applications from quality assessments to gene prediction and phylogenomics RM Waterhouse, M Seppey, FA Simão, M Manni, P Ioannidis, ... Molecular biology and evolution 35 (3), 543-548	1968	2018
BUSCO: Assessing Genome Assembly and Annotation Completeness M Seppey, M Manni, EM Zdobnov Kollmar M. (eds) Gene Prediction. Methods Mol Biol. 1962	1745	2019
BUSCO: assessing genomic data quality and beyond M Manni, MR Berkeley, M Seppey, EM Zdobnov Current Protocols 1 (12), e323	526	2021
Using BUSCO to assess insect genomic resources RM Waterhouse, M Seppey, FA Simão, EM Zdobnov Insect Genomics: Methods and Protocols, 59-74	29	2019

Cited by

	All	Since 2019
Citations	18275	16580
h-index	6	6
i10-index	6	6



BUSCO

https://busco.ezlab.org/busco_userguide.html

Setting up BUSCO - Gitpod

Exercise 1 - Bacterial gene set in Protein mode

HMMER: biosequence analysis using profile hidden Markov models

- Statistical model that can be used to model biological sequences
- Hidden states give rise to a probability distribution that can be used to predict sequence continuation
- Biological sequences display codon biases which can be modelled
- Similarly, possible to predict protein domains

BUSCO filters

- HMM profile matches must exceed threshold scores and lengths
  - Exceed score and length thresholds -> Complete
  - Exceed score but not length threshold -> Fragmented
  - Does not exceed score threshold -> Discarded
- Remove overlapping matches
  - If possible, save a match by discarding unused exons
- Remove duplications
  - e.g. if a BUSCO marker matches as both Complete and Fragmented, remove Fragmented match
  - if a gene matches multiple BUSCO markers keep only the highest scoring match
- Remove any low scoring multi-copy matches
  - If a BUSCO marker matches to multiple input genes, remove any matches that score < 85% of the best scoring match

BUSCO Pipelines

Pipelines

Three run modes

Genome

Transcriptome

Protein

Gene Predictors

- Needed to translate nucleotide input into protein gene sets
- Computationally demanding
- New and better tools continue to be developed
- Each tool has different requirements, approaches, outputs

National Library of Medicine
National Center for Biotechnology Information

Log in

BLAST ®

Home Recent Results Saved Strategies Help

Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. Learn more

NEWS

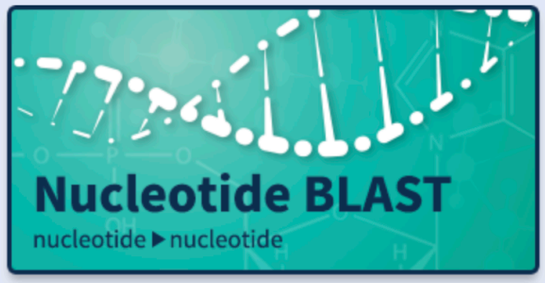
BLAST+ 2.15.0 is here!

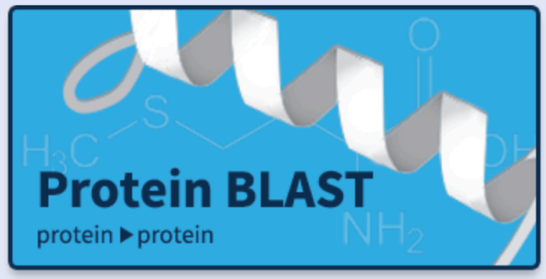
We have included two exciting new features in the latest BLAST+ release

Tue, 28 Nov 2023

More BLAST news...

Web BLAST





Augustus [gene prediction]

Bioinformatics Group of the Institute for Mathematics and Computer Science of the University of Greifswald

web interface WebAUGUSTUS accuracy results download AUGUSTUS data sets predictions references

AUGUSTUS is a program that predicts genes in eukaryotic genomic sequences. It can be run on this web server, on a new web server for larger input files or be downloaded and run locally. It is open source so you can compile it for your computing platform. You can now run AUGUSTUS on the German MediGRID. This enables you to submit larger sequence files and allows to use protein homology information in the prediction. The MediGRID requires an instant easy registration by email for first-time users.

Pipelines

Three run modes

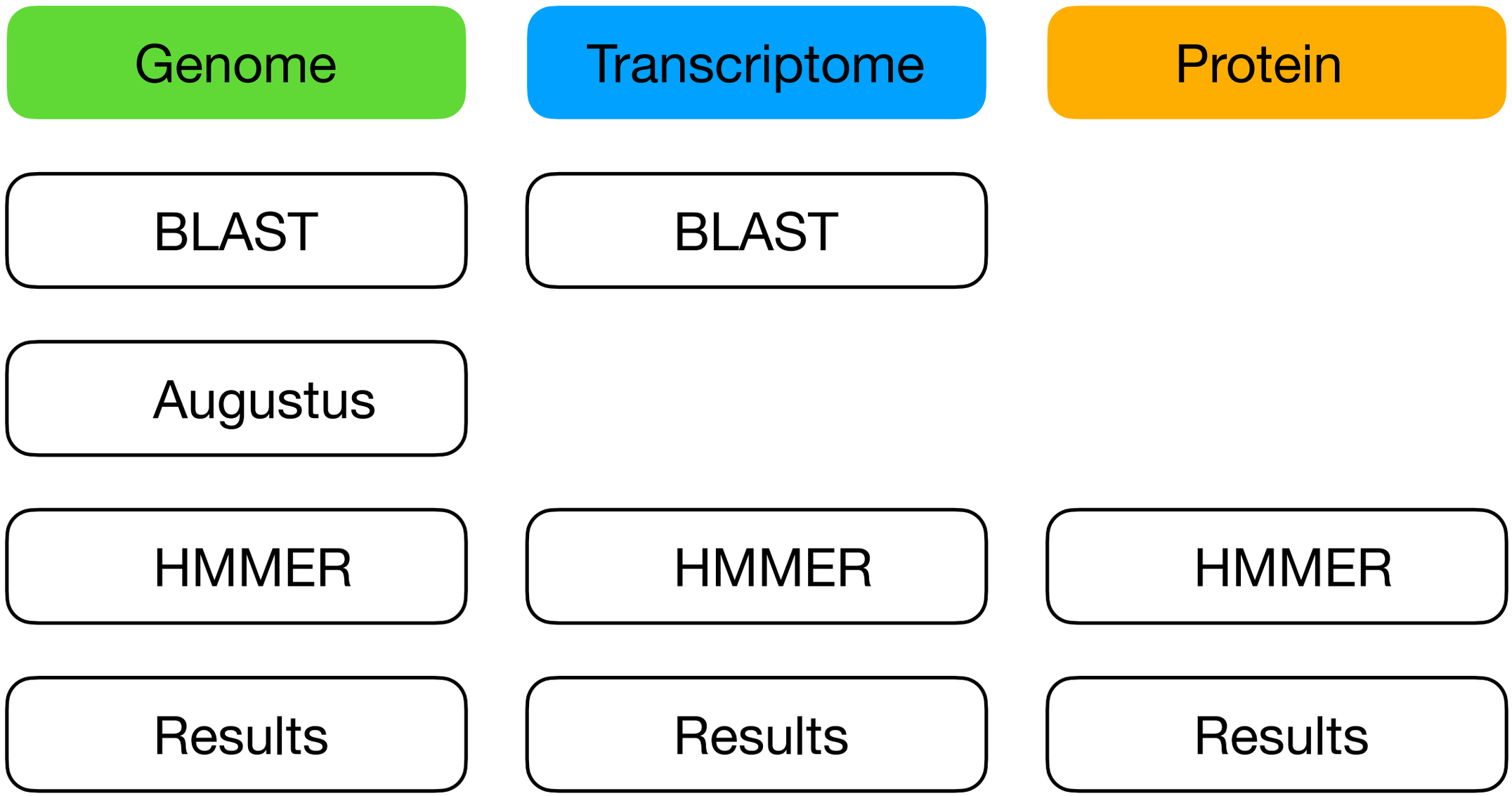
Genome

Transcriptome

Protein

Pipelines

Three run modes



Software | Open access | Published: 08 March 2010

Prodigal: prokaryotic gene recognition and translation initiation site identification

Doug Hyatt ✉, Gwo-Liang Chen, Philip F LoCascio, Miriam L Land, Frank W Larimer & Loren J Hauser

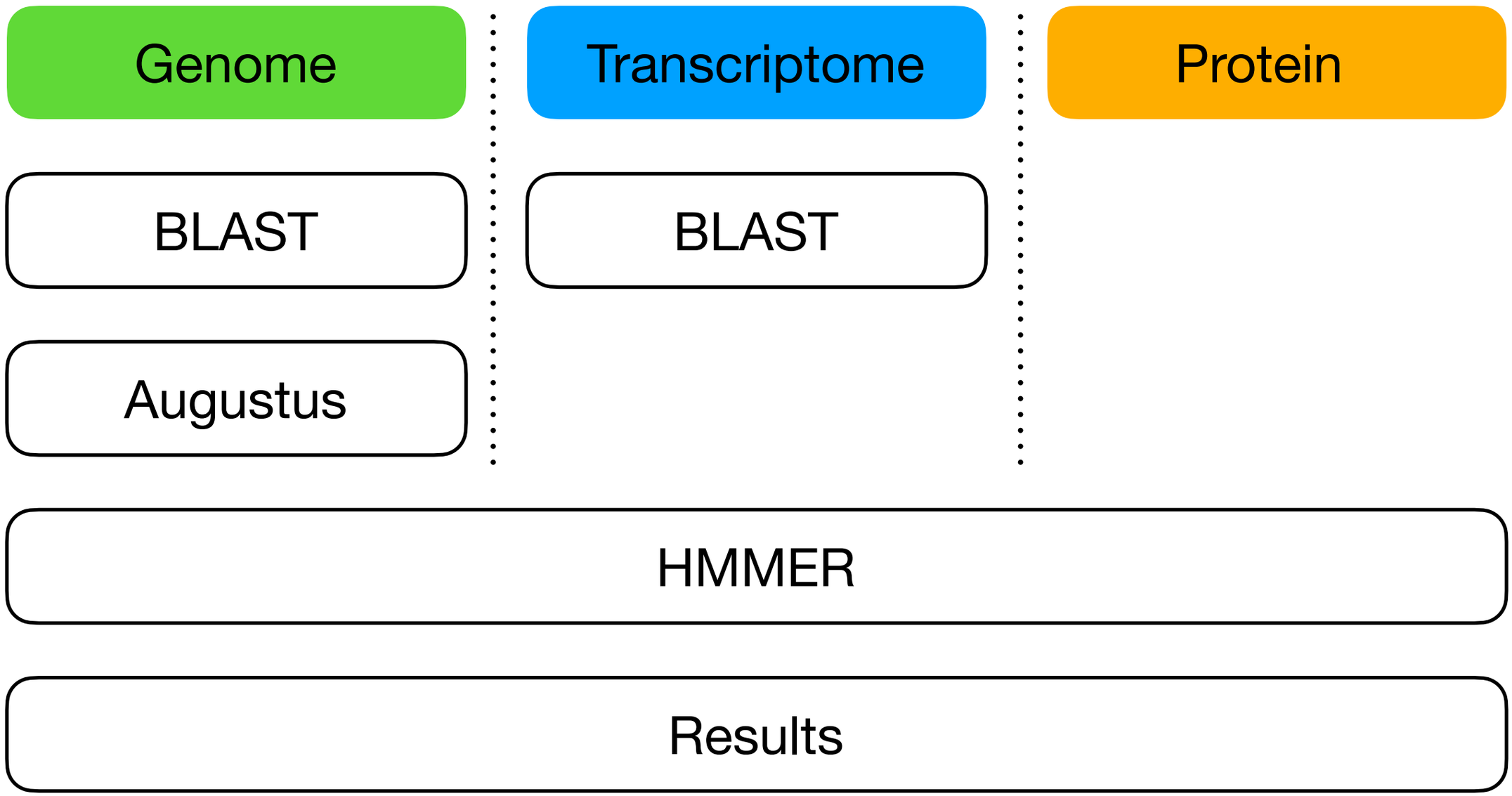
BMC Bioinformatics **11**, Article number: 119 (2010) | Cite this article

-

- Ab initio gene prediction
- Option to provide translation table

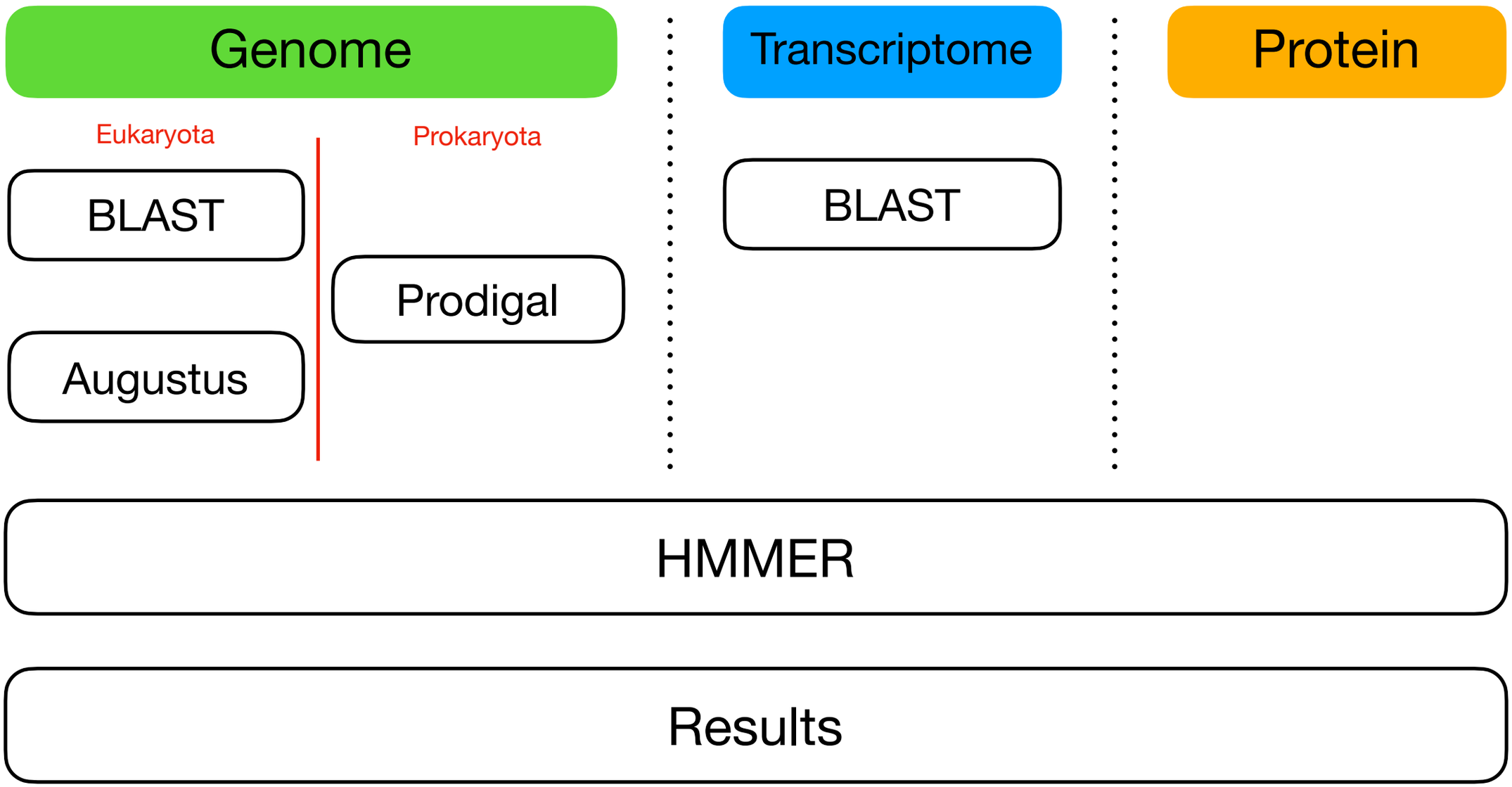
Pipelines

Three run modes



Pipelines

Three run modes



Exercise 2

Bacterial genome in Genome mode

README GPL-3.0 license

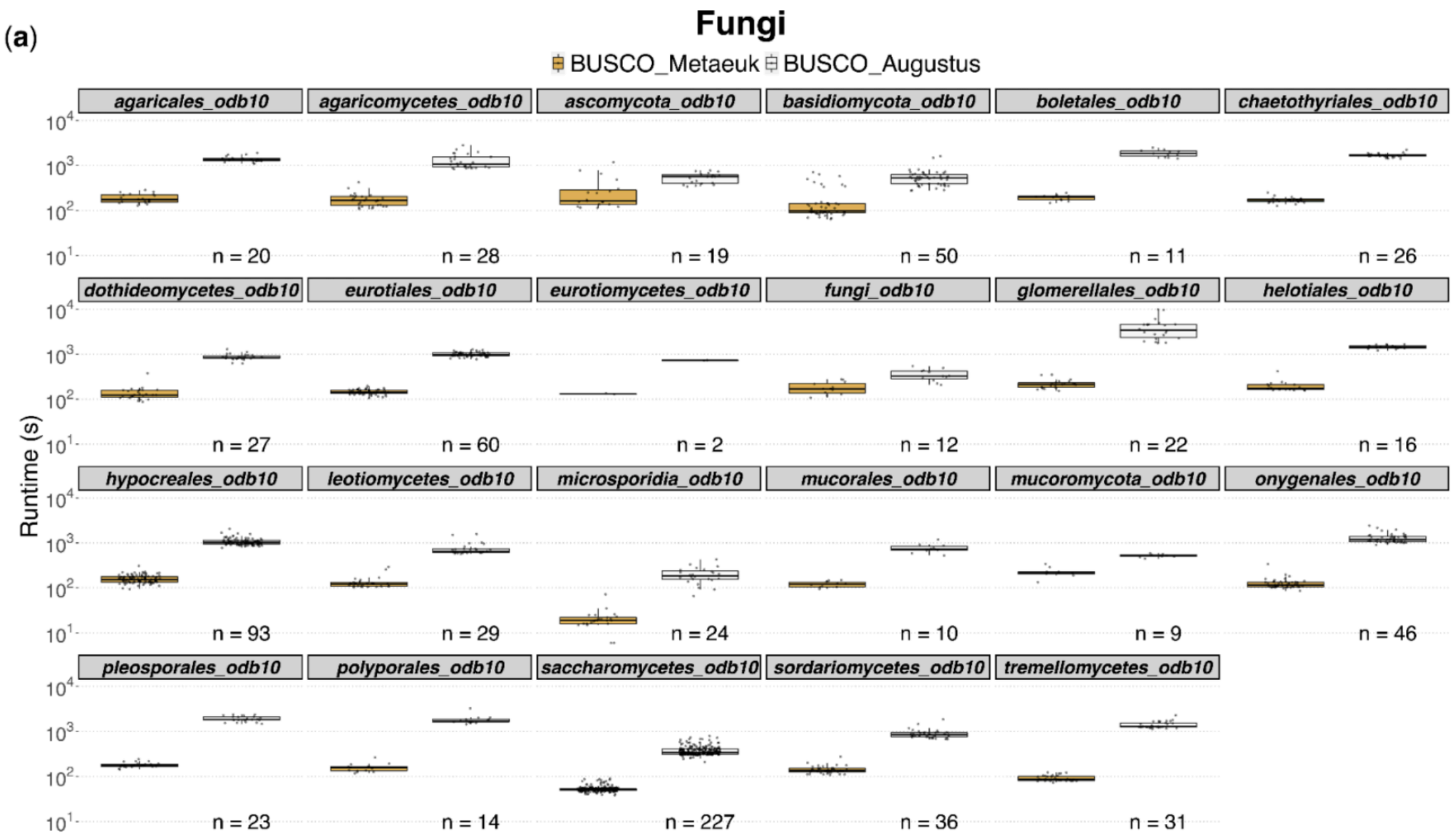
MetaEuk - sensitive, high-throughput gene discovery and annotation for large-scale eukaryotic metagenomics

MetaEuk is a modular toolkit designed for large-scale gene discovery and annotation in eukaryotic metagenomic contigs. MetaEuk combines the fast and sensitive homology search capabilities of MMseqs2 with a dynamic programming procedure to recover optimal exons sets. It reduces redundancies in multiple discoveries of the same gene and resolves conflicting gene predictions on the same strand. MetaEuk is GPLv3-licensed open source software that is implemented in C++ and available for Linux and macOS. The software is designed to run efficiently on multiple cores.

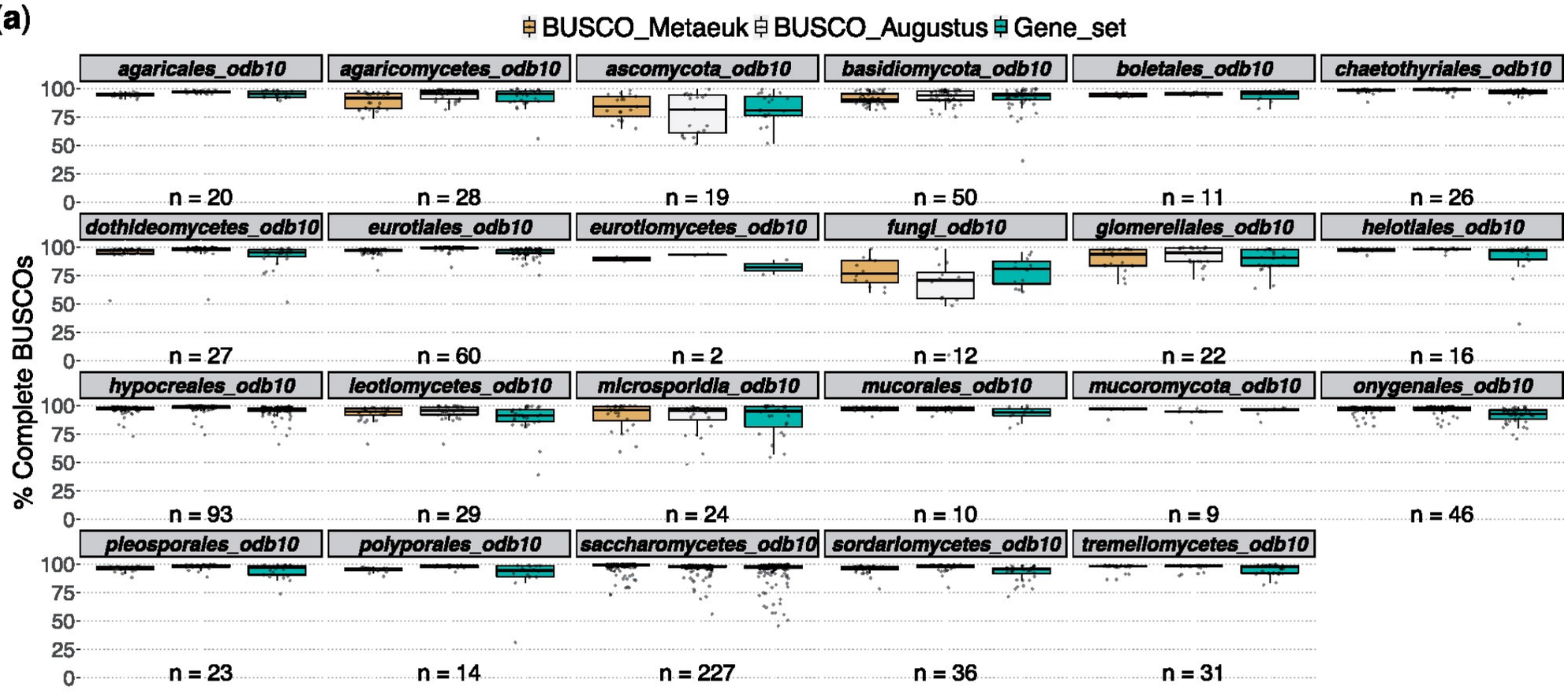
MMseqs2: ultra fast and sensitive sequence search and clustering suite

MMseqs2 (Many-against-Many sequence searching) is a software suite to search and cluster huge protein and nucleotide sequence sets. MMseqs2 is open source GPL-licensed software implemented in C++ for Linux, MacOS, and (as beta version, via cygwin) Windows. The software is designed to run on multiple cores and servers and exhibits very good scalability. MMseqs2 can run 10000 times faster than BLAST. At 100 times its speed it achieves almost the same sensitivity. It can perform profile searches with the same sensitivity as PSI-BLAST at over 400 times its speed.

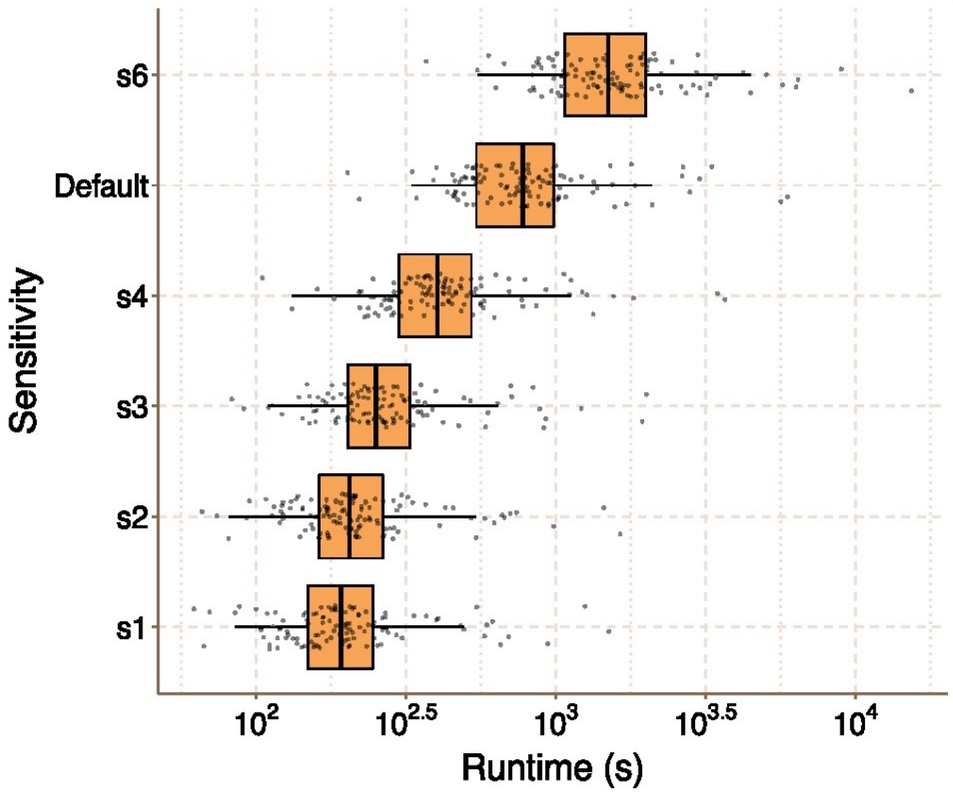
Metaeuk


(a)
Fungi
BUSCO_Metaeuk
BUSCO_Augustus
Runtime (s)
agaricales_odb10
agaricomycetes_odb10
ascomycota_odb10
basidiomycota_odb10
boletales_odb10
chaetothyriales_odb10
n = 20
n = 28
n = 19
n = 50
n = 11
n = 26
dothideomycetes_odb10
eurotiales_odb10
eurotiomycetes_odb10
fungi_odb10
glomerellales_odb10
helotiales_odb10
n = 27
n = 60
n = 2
n = 12
n = 22
n = 16
hypocreales_odb10
leotiomycetes_odb10
microsporidia_odb10
mucorales_odb10
mucoromycota_odb10
onygenales_odb10
n = 93
n = 29
n = 24
n = 10
n = 9
n = 46
pleosporales_odb10
polyporales_odb10
saccharomycetes_odb10
sordariomycetes_odb10
tremellomycetes_odb10
n = 23
n = 14
n = 227
n = 36
n = 31

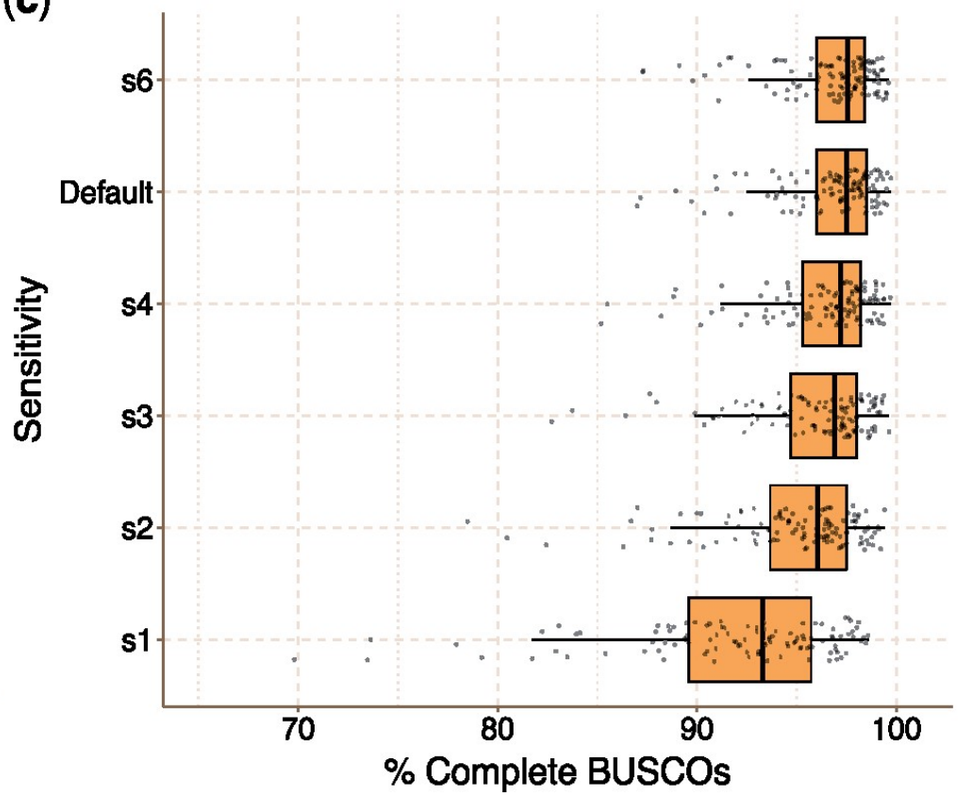
(a)


BUSCO_Metaeuk
BUSCO_Augustus
Gene_set
agaricales_odb10
agaricomycetes_odb10
ascomycota_odb10
basidiomycota_odb10
boletales_odb10
chaetothyriales_odb10
dothideomycetes_odb10
eurotiales_odb10
eurotiomycetes_odb10
fungi_odb10
glomerellales_odb10
helotiales_odb10
hypocreales_odb10
leotiomycetes_odb10
microsporidia_odb10
mucorales_odb10
mucoromycota_odb10
onygenales_odb10
pleosporales_odb10
polyporales_odb10
saccharomycetes_odb10
sordariomycetes_odb10
tremellomycetes_odb10
n = 20
n = 28
n = 19
n = 50
n = 11
n = 26
n = 27
n = 60
n = 2
n = 12
n = 22
n = 16
n = 93
n = 29
n = 24
n = 10
n = 9
n = 46
n = 23
n = 14
n = 227
n = 36
n = 31
% Complete BUSCOs
0
25
50
75
100

(b)

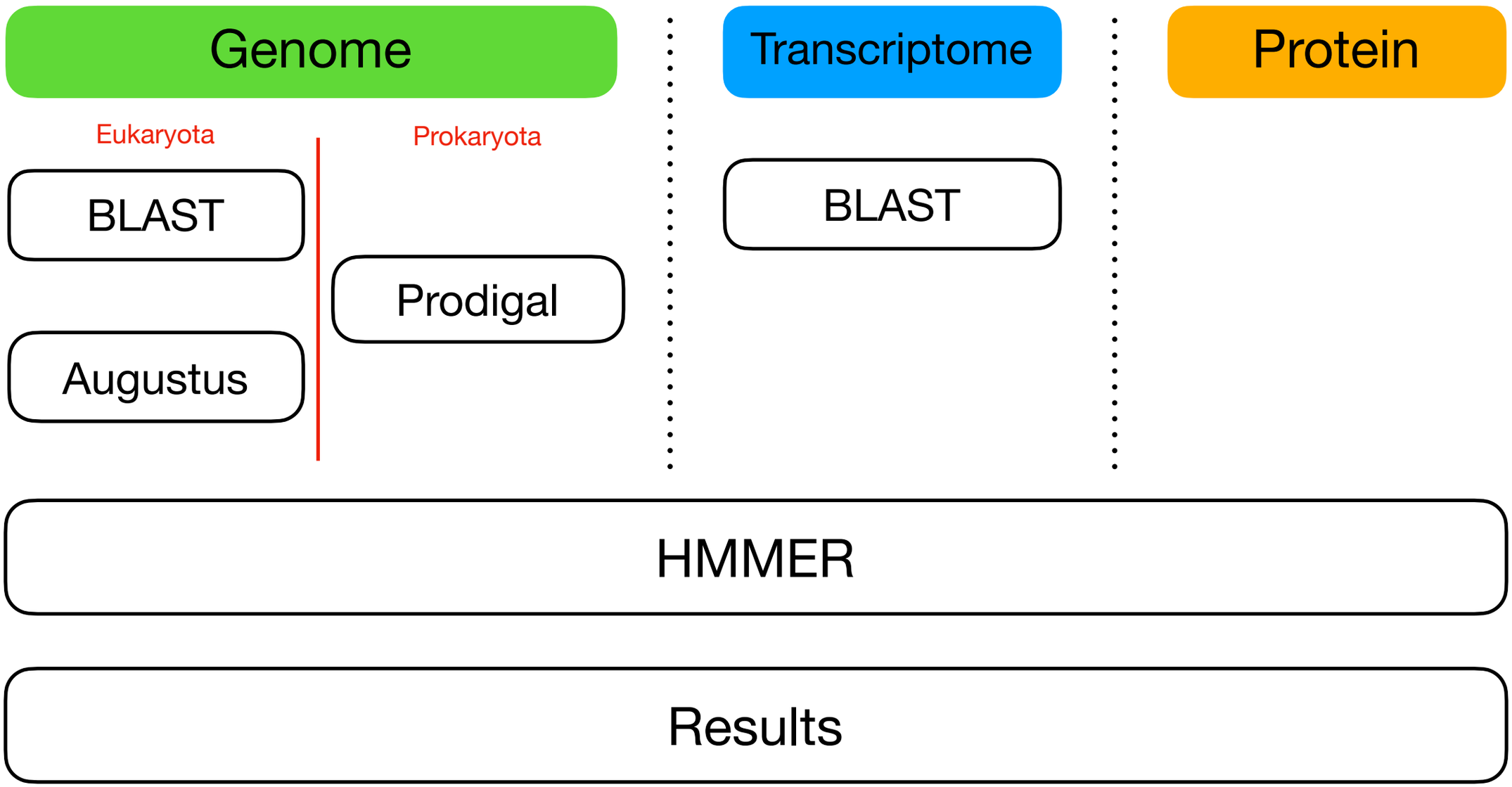

Sensitivity
s6
Default
s4
s3
s2
s1
10^2
$10^{2.5}$
10^3
$10^{3.5}$
10^4
Runtime (s)

(c)


Sensitivity
s6
Default
s4
s3
s2
s1
70
80
90
100
% Complete BUSCOs

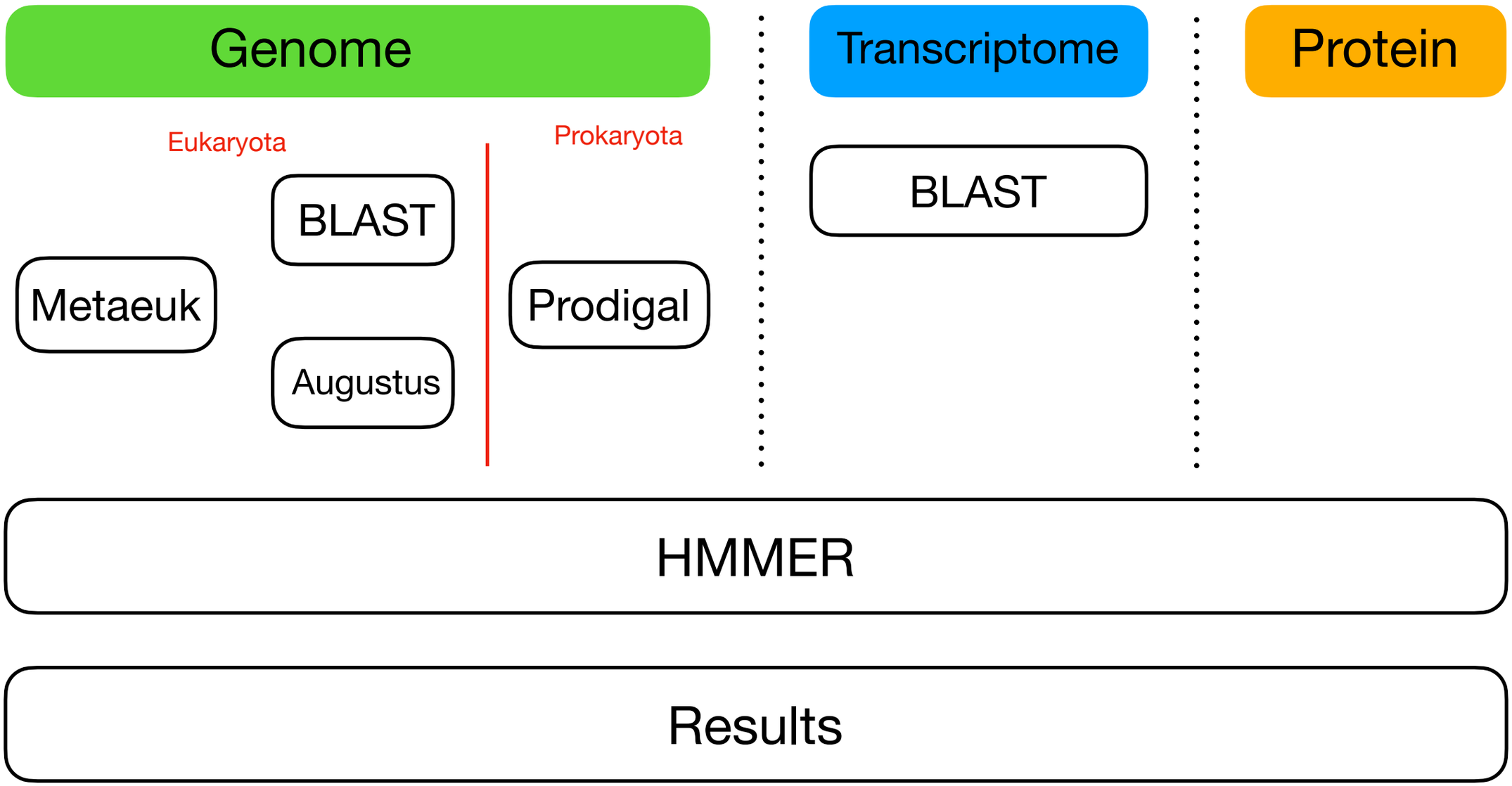
Pipelines

Three run modes



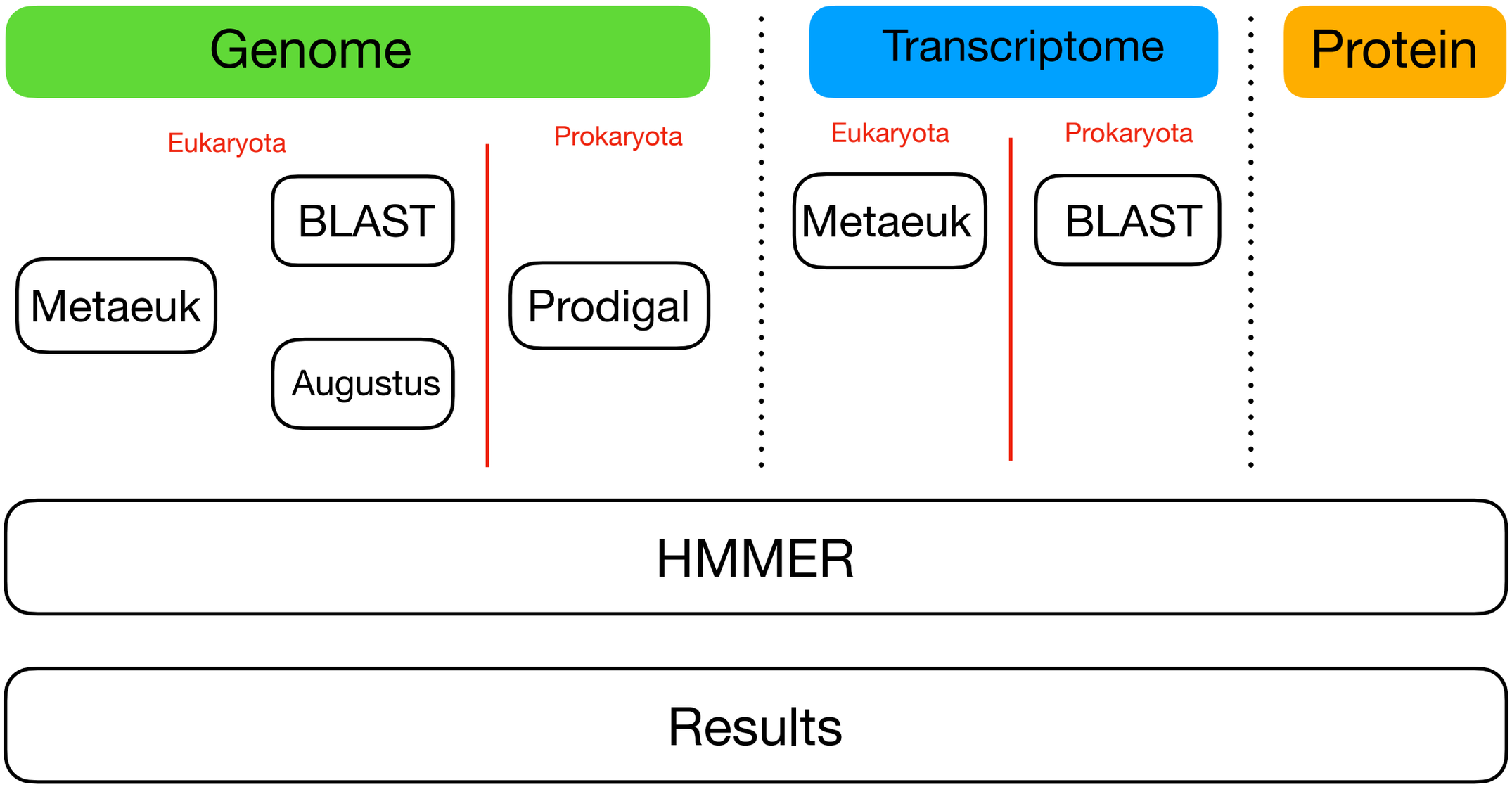
Pipelines

Three run modes



Pipelines

Three run modes



Genome analysis

Protein-to-genome alignment with miniprot

Heng Li [1,2]

[1]Department of Data Sciences, Dana-Farber Cancer Institute, Boston, MA 02215, USA and [2]Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02115, USA

Received on October 13, 2022; revised on December 25, 2022; editorial decision on January 9, 2023

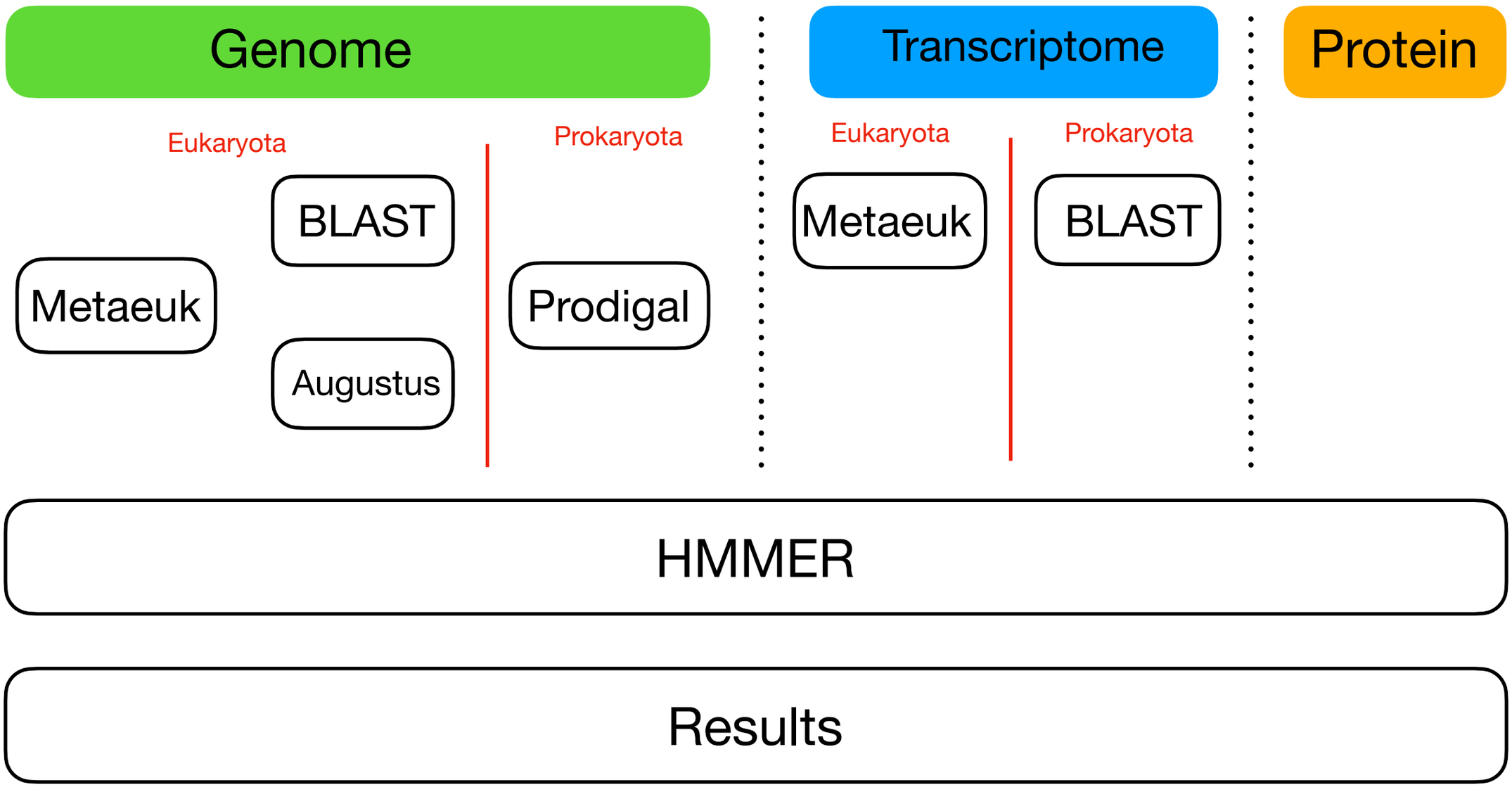
Abstract

Motivation: Protein-to-genome alignment is critical to annotating genes in non-model organisms. While there are a few tools for this purpose, all of them were developed over 10 years ago and did not incorporate the latest advances in alignment algorithms. They are inefficient and could not keep up with the rapid production of new genomes and quickly growing protein databases.

Results: Here, we describe miniprot, a new aligner for mapping protein sequences to a complete genome. Miniprot integrates recent techniques such as k-mer sketch and vectorized dynamic programming. It is tens of times faster than existing tools while achieving comparable accuracy on real data.

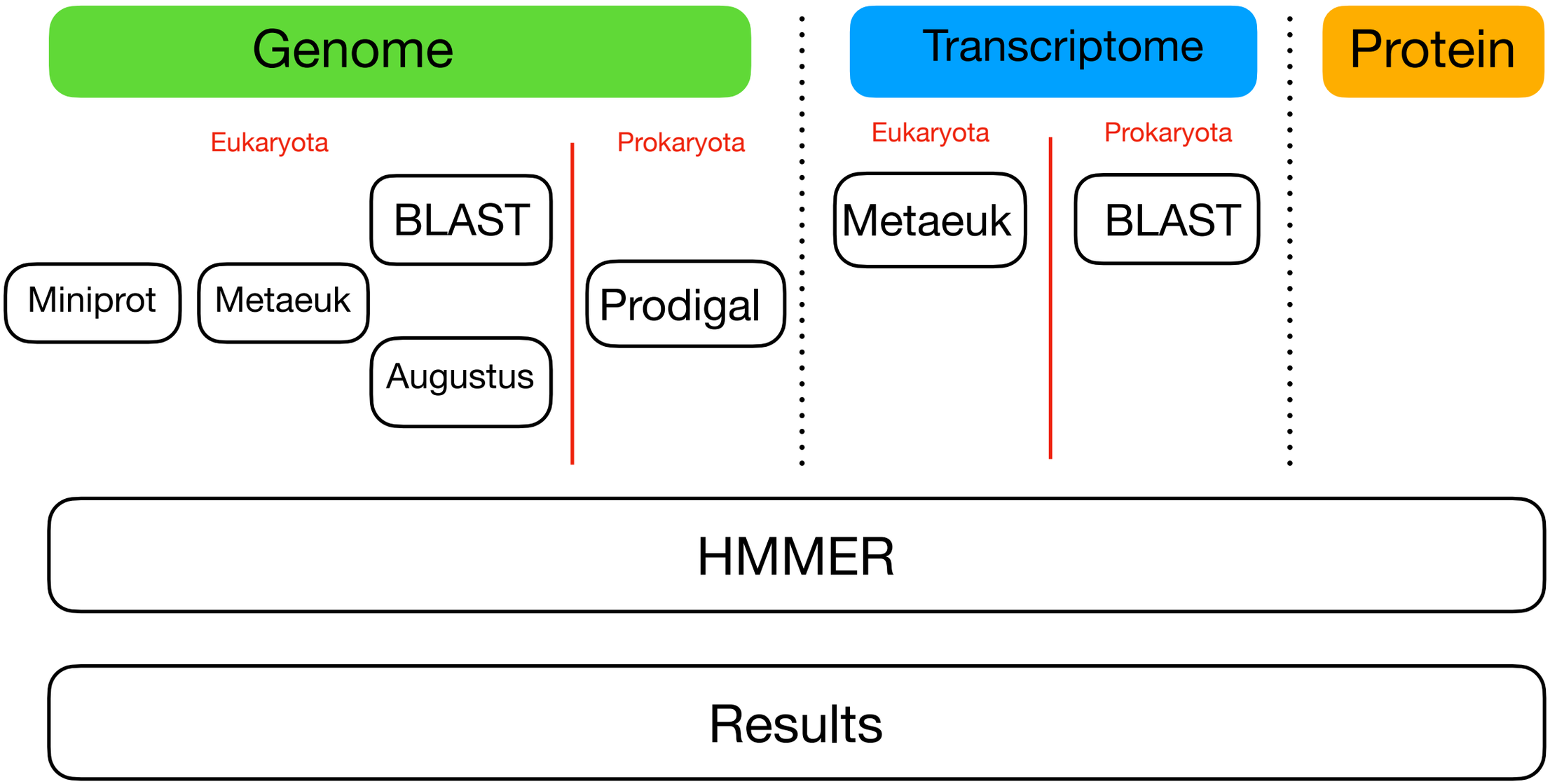
Pipelines

Three run modes



Pipelines

Three run modes



Exercise 3

Genome pipelines comparison for eukaryota

BUSCO for Metagenomics

- Metagenomics: sampling an environment and sequencing everything
- Resulting sequence assemblies may contain unknown organisms
- BUSCO v4: Auto-Select Lineage
- Three phylogenetic domains: Archaea, Bacteria, Eukaryota
- Phylogenetic Tree placement
- BUSCO is the only tool able to evaluate both microbial eukaryotes and prokaryotes
- Viral datasets have been introduced

Biocomputing 2012, pp. 247-258 (2011)

SEPP: SATé-Enabled Phylogenetic Placement

S. MIRARAB, N. NGUYEN, and T. WARNOW

https://doi.org/10.1142/9789814366496_0024 | **Cited by:** 37 **(Source: Crossref)**

Introduction to SEPP

SEPP [1] stands for SATé-Enabled Phylogenetic Placement and addresses the problem of phylogenetic placement for meta-genomic short reads. More precisely, SEPP addresses the following problem.

- **Input**:
  i) *backbone* tree `T` and backbone alignment `A` for a set of full-length gene sequences ii) the set `X` of fragmentary sequences from the same gene as the backbone
- **Output**: the placement of each fragment in `X` onto the tree T and the alignment of all fragment in `X` to the alignment `A`.

Input Assembly/Batch

Lineage specified

e.g. Diptera

Autoselect Lineage

Archaea

Bacteria

Eukaryota

Dataset

Prokaryota/ Viruses

Gene Predictor

Prodigal

--augustus

Eukaryota

BLAST

Augustus

Miniprot/ Metaeuk

HMMER

Evaluation

Genome

Gene Sets/Proteins

Transcriptome

BUSCO v3 pipeline

BUSCO Assessment Results

Complete (C) and single-copy (S)

Complete (C) and duplicated (D)

Fragmented (F)

Missing (M)

0 20 40 60 80 100

%BUSCOs

Final Result

Archaea

Bacteria

Eukaryota

Diptera

e.g.

SEPP

Eukaryota

e.g.

Evaluation

Exercise 4

Auto-lineage mode

Reporting parasitic-reduced genomes in fungi

```
---------------------------------------------------
|Results from dataset fungi_odb10                 |
---------------------------------------------------
|C:15.4%[S:14.5%,D:0.9%],F:2.9%,M:81.7%,n:758,E:3.4%     |
|117    Complete BUSCOs (C)    (of which 4 contain       |
|internal stop codons)                            |
|110    Complete and single-copy BUSCOs (S)       |
|7    Complete and duplicated BUSCOs (D)          |
|22    Fragmented BUSCOs (F)                      |
|619    Missing BUSCOs (M)                        |
|758    Total BUSCO groups searched               |
|                                                 |
|!!! The missing BUSCOs match the pattern of a    |
|parasitic-reduced genome. 81.4% of your missing BUSCOs  |
|are typically missing in these. A corrected score would |
|be:                                              |
|C:46.3%[S:43.5%,D:2.8%],F:8.7%,M:45.0%,n:253     |
|                                                 |
|                                                 |
|Consider using the auto-lineage mode to select a more   |
|specific lineage.                                |
---------------------------------------------------
```

BUSCO command line

Mandatory Options

```
-i  # Input sequence file or folder
-m  # BUSCO analysis mode to run. Can be 'genome', 'protein' or 'transcriptome'.
```

Recommended Options

```
-c  # Number of threads/cores to use
-l  # Specify the BUSCO lineage dataset to be used for scoring
-o  # Specify the name of the output folder
```

Pipeline-specific options

Augustus pipeline

```
--augustus  # Invoke the BLAST/Augustus pipeline
-e  # E-value cutoff for BLAST searches.
--limit  # How many BLAST candidate regions to consider per BUSCO (default: 3)
--long  # Optimization Augustus self-training mode (Default: Off)
--augustus_parameters  # "--PARAM1=VALUE1,--PARAM2=VALUE2"
--augustus_species AUGUSTUS_SPECIES  # Specify a species for Augustus training
```

Metaeuk pipeline

```
--metaeuk  # Invoke the Metaeuk pipeline
--metaeuk_parameters  # "--PARAM1=VALUE1,--PARAM2=VALUE2"
--metaeuk_rerun_parameters  # "--PARAM1=VALUE1,--PARAM2=VALUE2"
```

Miniprot pipeline

```
--miniprot  # Invoke the miniprot pipeline (default for eukaryota, option for prokaryota)
```

Genome Mode (all)

```
--skip_bbtools  # Skip BBTools for assembly statistics
--scaffold_composition  # Writes ACGTN content per scaffold to a file scaffold_composition.txt
--contig_break n  #Number of contiguous Ns to signify a break between contigs. Default is n=10.
```

Other options

```
--config CONFIG_FILE   # Provide a config file
--download [dataset ...]  # Can be a dataset name, "all", "prokaryota", "eukaryota" or "virus"
--download_base_url DOWNLOAD_BASE_URL  # Location from which to take the datsets
--download_path DOWNLOAD_PATH  # Set a custom local location for downloaded files
-f  # Force overwrite an output directory
-r  # Restart a previous incomplete run
--list-datasets  # List all available BUSCO datasets
--offline  # Indicate BUSCO should not attempt to download files
-q  # Quiet mode
--tar  # Compress some output subdirectories
```

Generate Plots

Exercise 5

Batch mode

BUSCO use case examples

Protocols Paper

https://currentprotocols.onlinelibrary.wiley.com/doi/pdf/10.1002/cpz1.323

PROTOCOL | Open Access

BUSCO: Assessing Genomic Data Quality and Beyond

Mosè Manni, Matthew R. Berkeley, Mathieu Seppey, Evgeny M. Zdobnov

Protocols Paper

https://currentprotocols.onlinelibrary.wiley.com/doi/pdf/10.1002/cpz1.323

- https://gitlab.com/ezlab/busco_protocol
- BUSCO for viewing syntenies
- BUSCO for constructing phylogenetic trees

Datasets

OrthoDB v12.0

https://www.orthodb.org/

	OrthoDB.v12	OMA	OrthoDB.v11	eggNOG.v6	KEGG-OC	OrthoDB.v10	eggNOG.v5
Release	2024	2024	2022	2022	2019	2018	2018
Eukaryota	5,827	713	1,952	1322	456	1,271	477
Bacteria	17,551	1,965	17,551	10,756	4,880	5,609	4,445
Archaea	607	173	607	457	278	404	168

New Datasets

	odb9	odb10	odb12
Archaea	0	16	32
Bacteria	16	83	344
Eukaryota	33	67	113
Virus	0	27	27
Total	**49**	**193**	**516**

Dataset Creation

Data selection

- Obtain all data from OrthoDB
- Run all species against the odb10 parent dataset (archaea, bacteria, eukaryota)
    - Remove all species that have high (>20%) duplication
    - Remove species that score below either 90% Complete or clade median Complete score less two standard deviations
- Propagate species to include from leaf nodes up to parent nodes
- Discard clades with fewer than 10 species remaining
- Take only the best scoring 100 species from each family
- Take only the best scoring 5 species from each genus

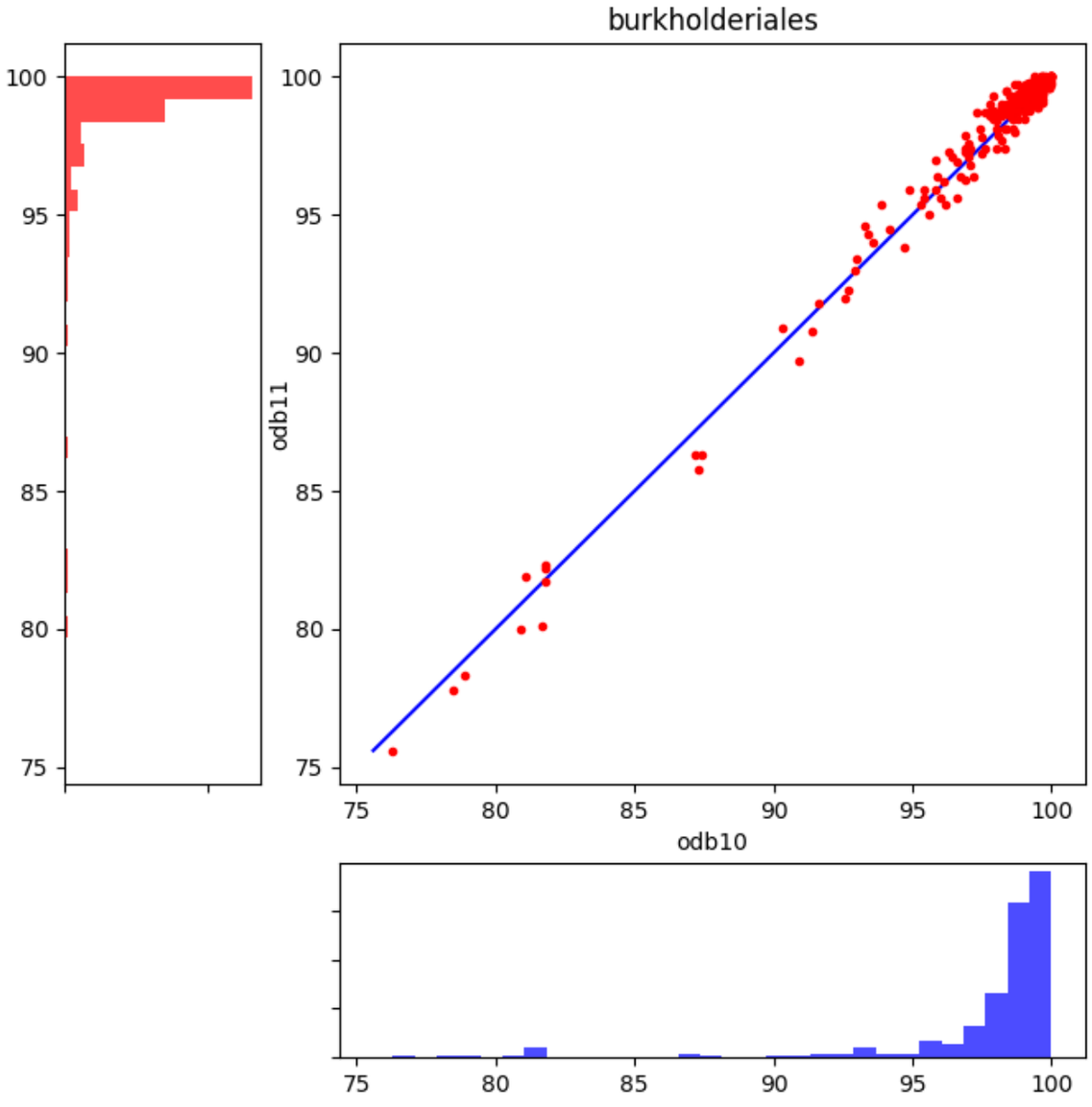
Dataset Creation

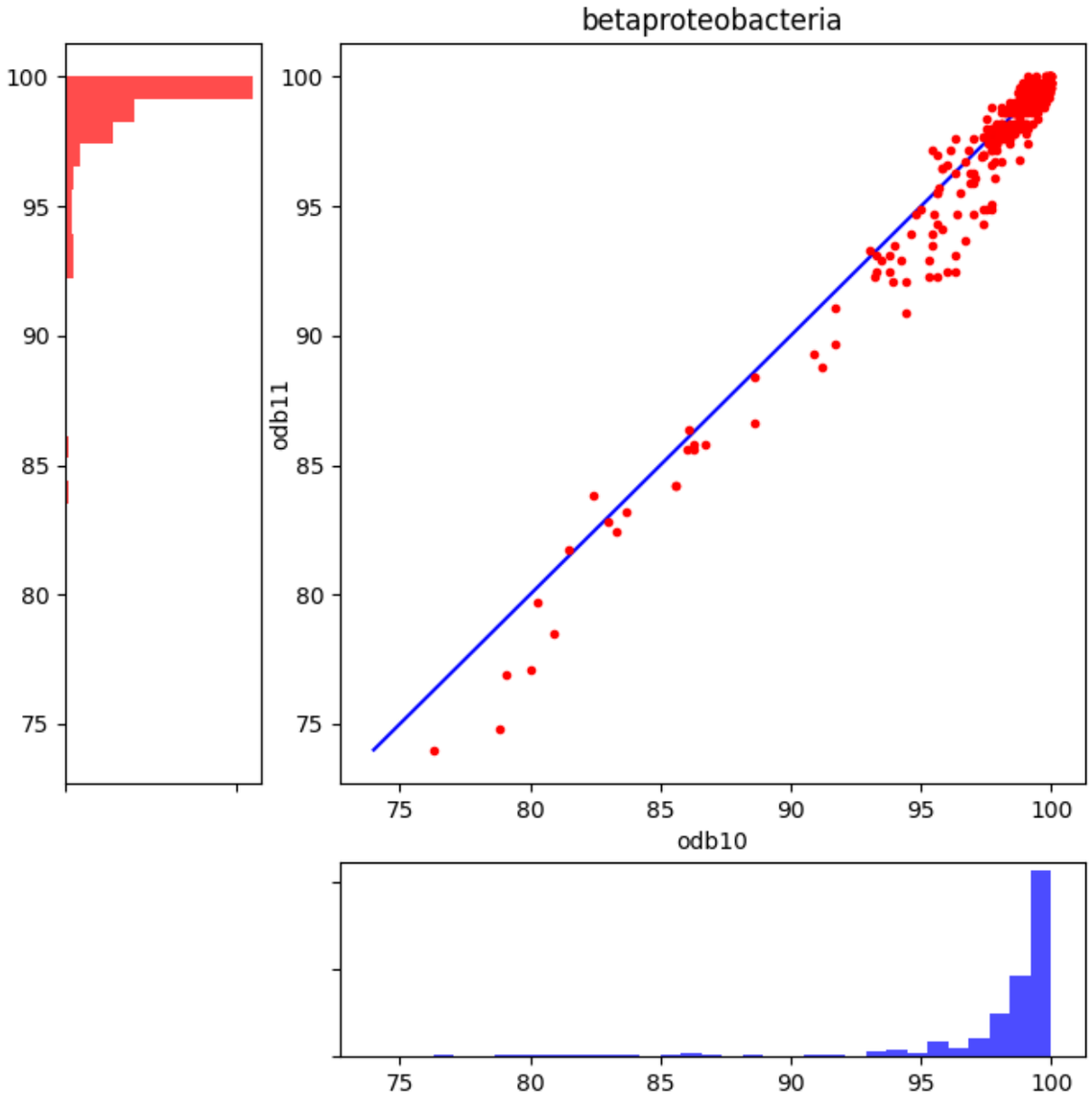
Pipeline

- Create alignment of genes with Clustalo
- Trim alignments with Trimal
- Build HMM profiles with HMMER hmmbuild
- Run HMMER hmmsearch and determine appropriate threshold scores for matches
- Run all species in clade against the dataset and remove markers that perform poorly (low f1 score)
- Remove datasets that are very similar to a parent

Dataset Creation

Preview of odb12 datasets





Nucleic Acids Research paper

OrthoDB and BUSCO update: annotation of orthologs with wider sampling of genomes

Fredrik Tegenfeldt[#], Dmitry Kuznetsov[#], Mosè Manni[#], Matthew Berkeley[#], Evgeny M. Zdobnov*, and Evgenia V. Kriventseva*

Department of Genetic Medicine and Development, University of Geneva Medical School, rue Michel-Servet 1, 1211 Geneva, Switzerland, and Swiss Institute of Bioinformatics, rue Michel-Servet 1, 1211 Geneva, Switzerland.

[#] Contributed equally to this work.

Thank you