MAG reconstruction, quality assessment and taxonomic annotation

Jeferyd Yepes García

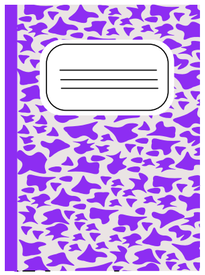
Swiss Institute of Bioinformatics

UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

Overview

Assembly/Binning

- Assembly.
- Binning.



Quality control

- Metrics.
- Tools.

Pipelines

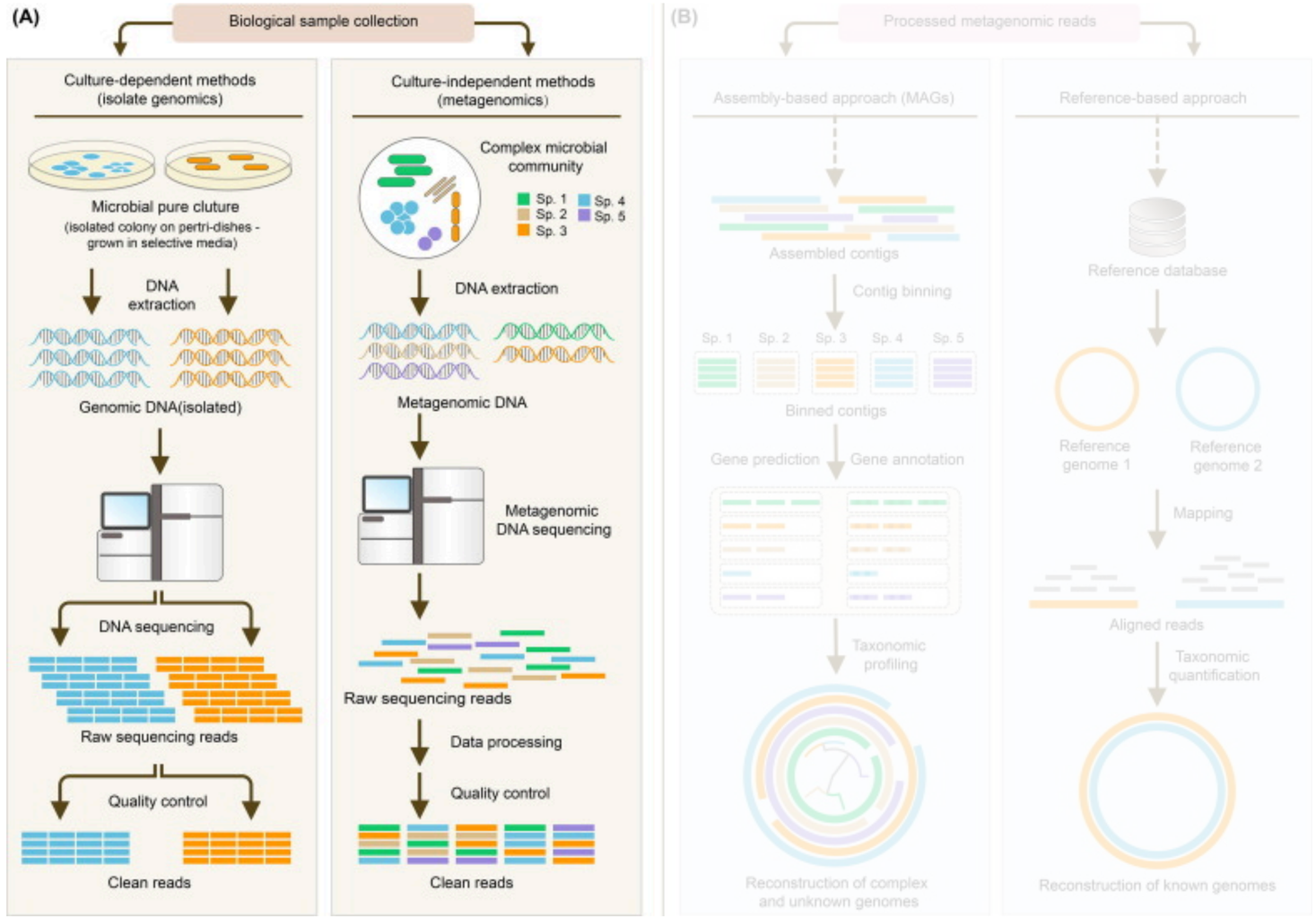
- Analysis workflow.
- Workflow managers.
- Nextflow.

Taxonomic annotation

- Tools.
- Pipelines.

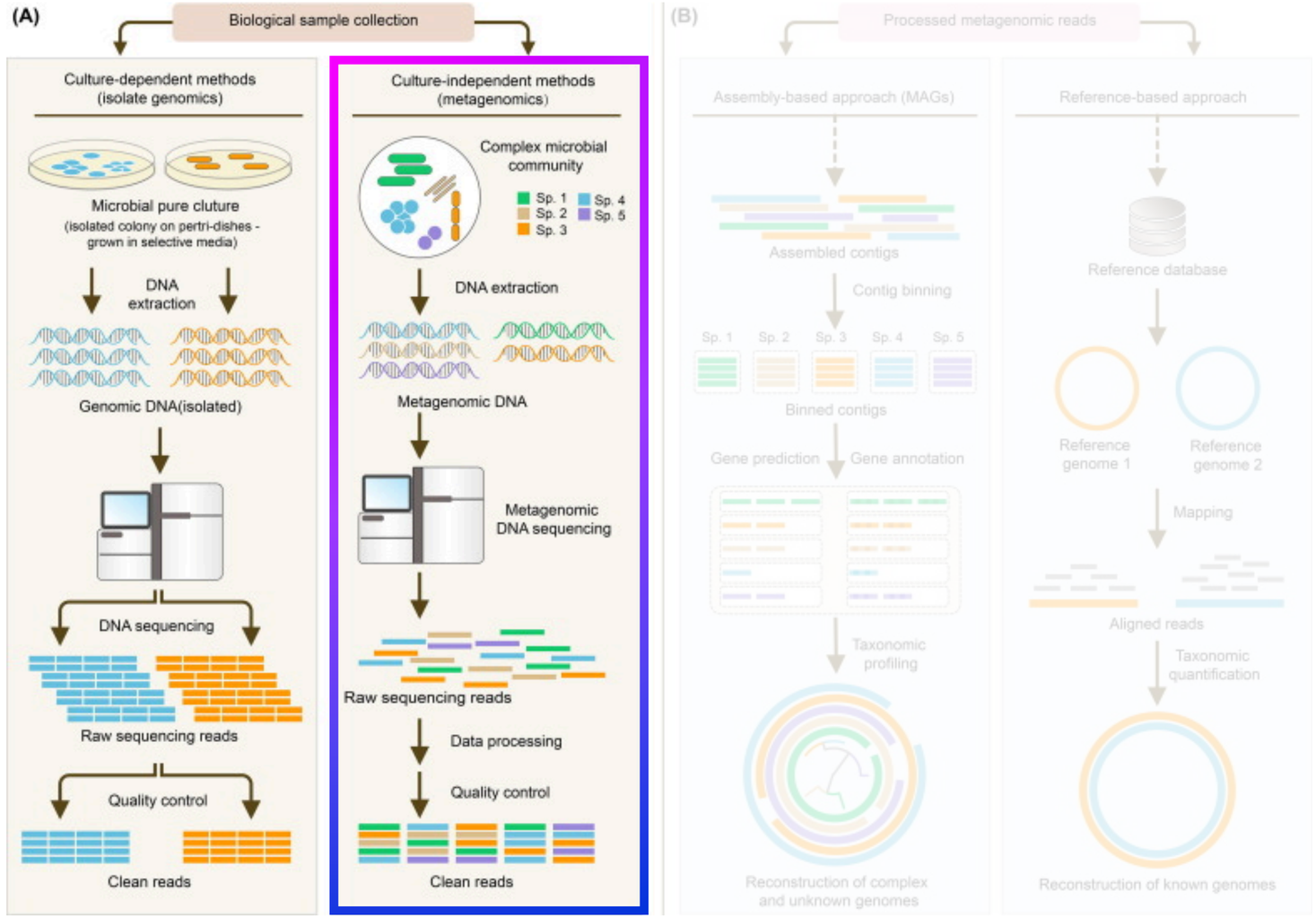
Assembly/Binning

Assembly/Binning



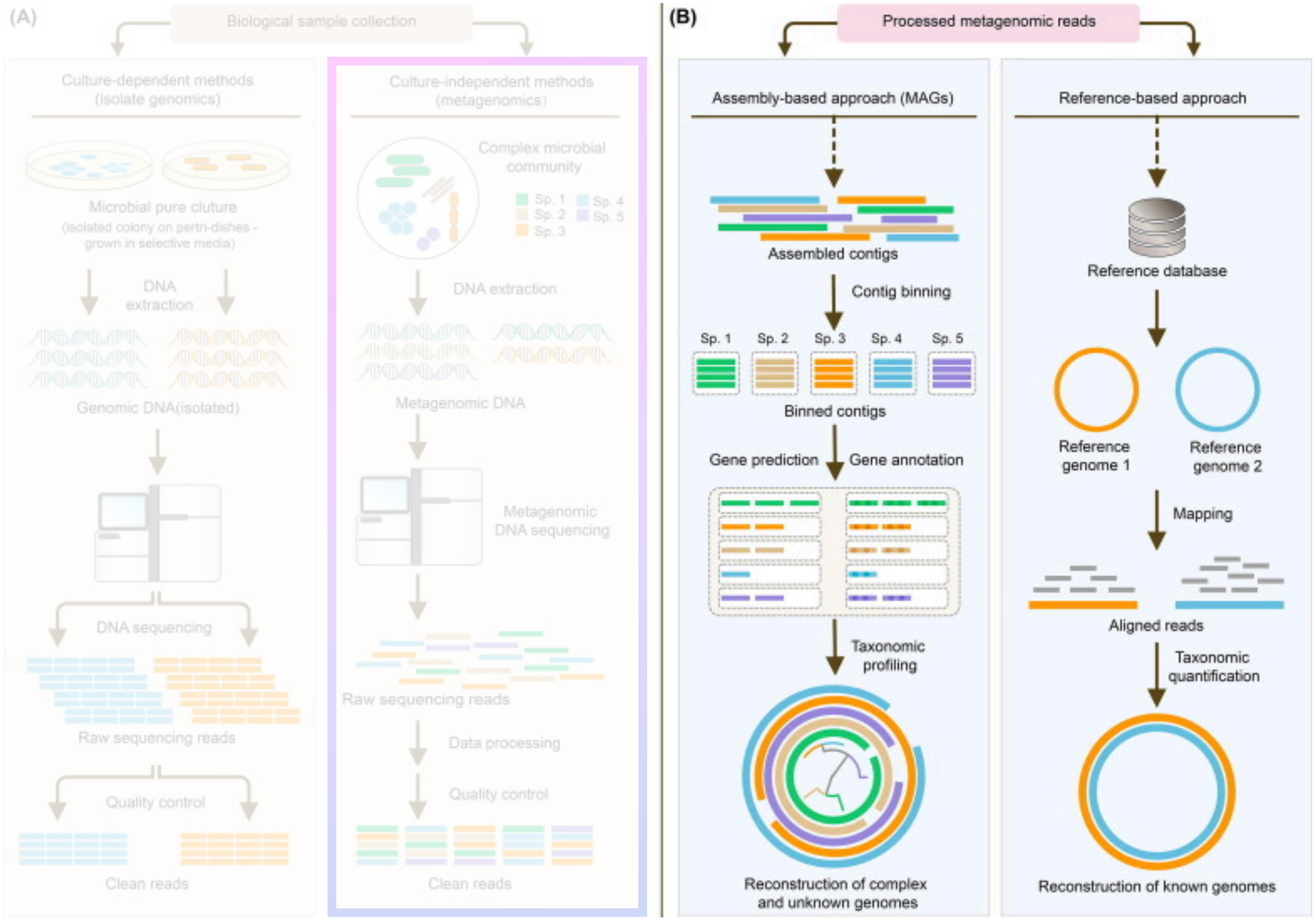
Yang *et al.* (2021)

Assembly/Binning



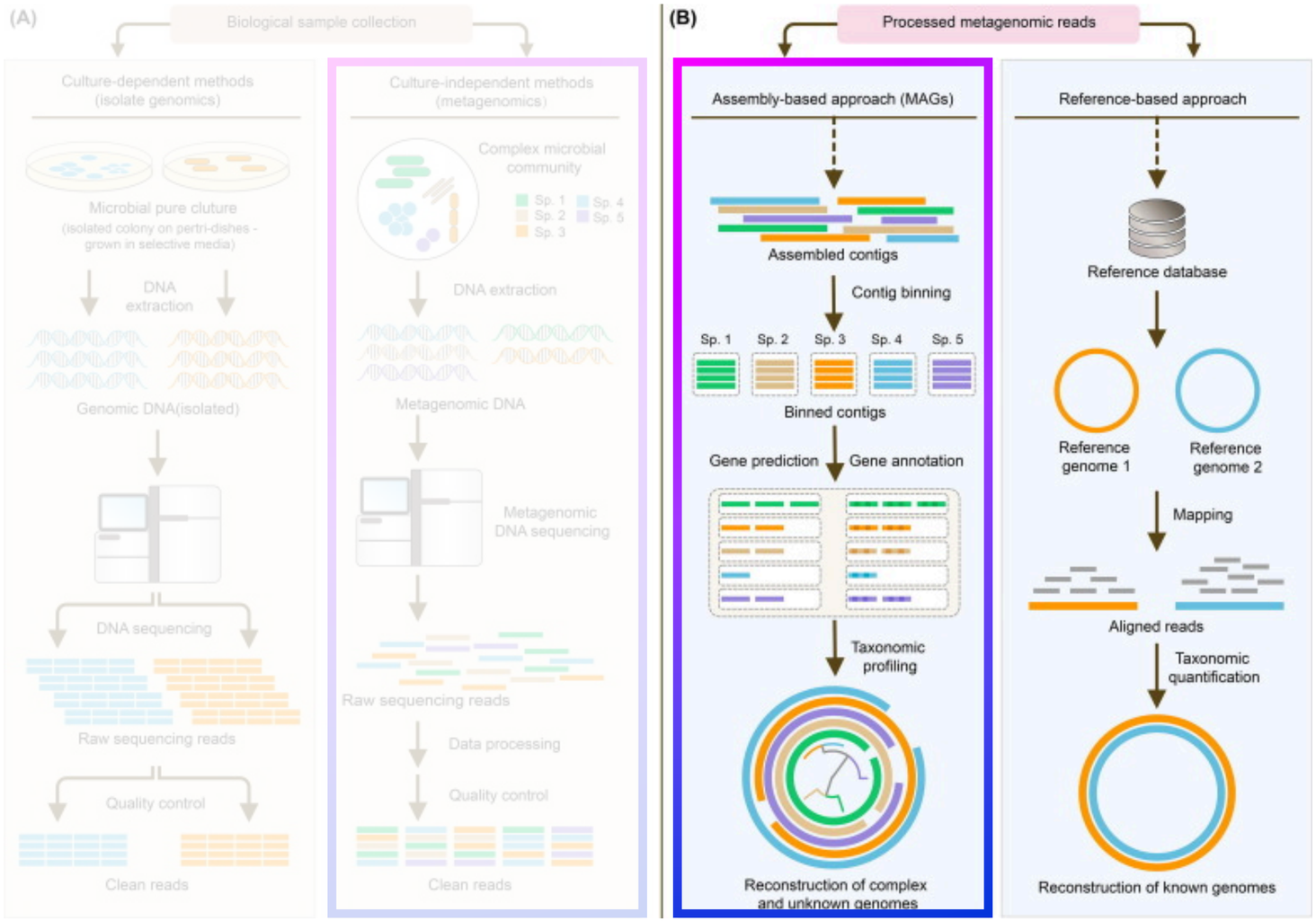
Yang *et al.* (2021)

Assembly/Binning



Yang *et al.* (2021)

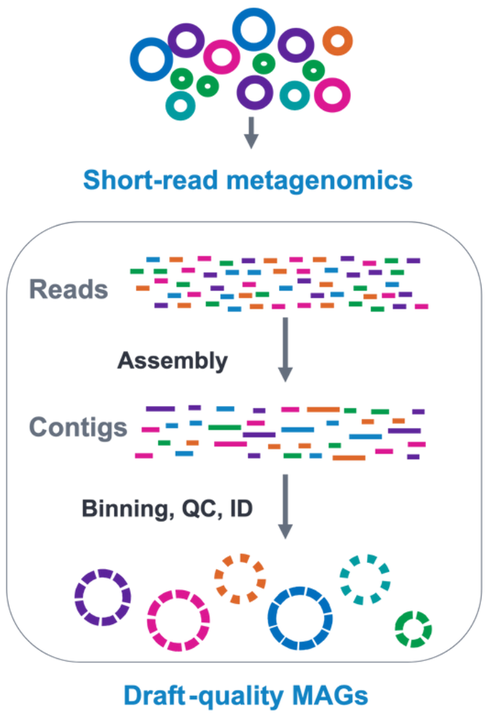
Assembly/Binning



Yang *et al.* (2021)

Assembly/Binning

Criteria to define a MAG:



Minimum information about a Metagenome-Assembled Genome (MIMAG)

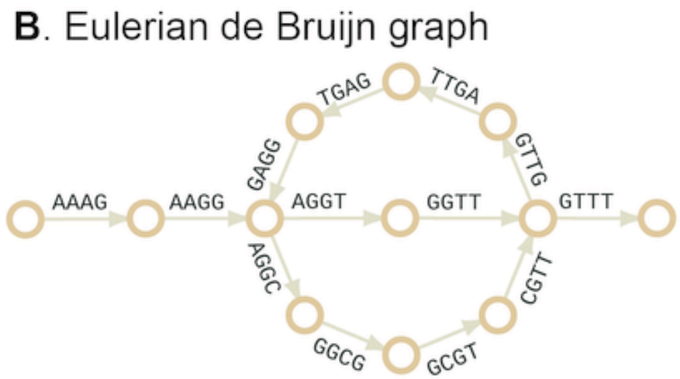
Table 1 Genome reporting standards for SAGs and MAGs

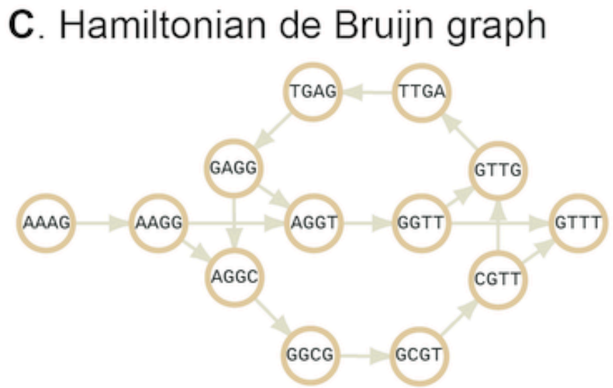
Criterion	Description
Finished (SAG/MAG)	
Assembly quality[a]	Single contiguous sequence without gaps or ambiguities with a consensus error rate equivalent to Q50 or better
High-quality draft (SAG/MAG)	
Assembly quality[a]	Multiple fragments where gaps span repetitive regions. Presence of the 23S, 16S, and 5S rRNA genes and at least 18 tRNAs.
Completion[b]	>90%
Contamination[c]	<5%
Medium-quality draft (SAG/MAG)	
Assembly quality[a]	Many fragments with little to no review of assembly other than reporting of standard assembly statistics.
Completion[b]	≥50%
Contamination[c]	<10%
Low-quality draft (SAG/MAG)	
Assembly quality[a]	Many fragments with little to no review of assembly other than reporting of standard assembly statistics.
Completion[b]	<50%
Contamination[c]	<10%

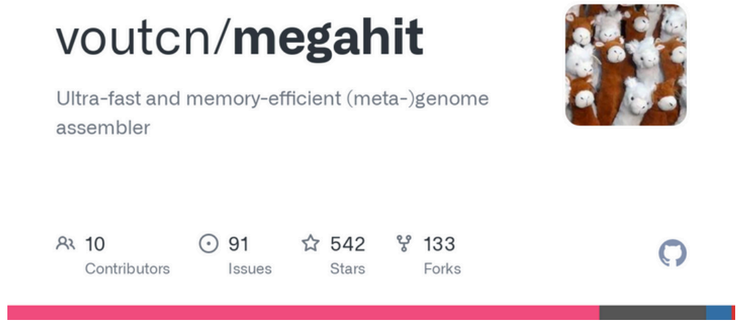
Bowers *et al.* (2017)

Assembly/Binning

Assembly:

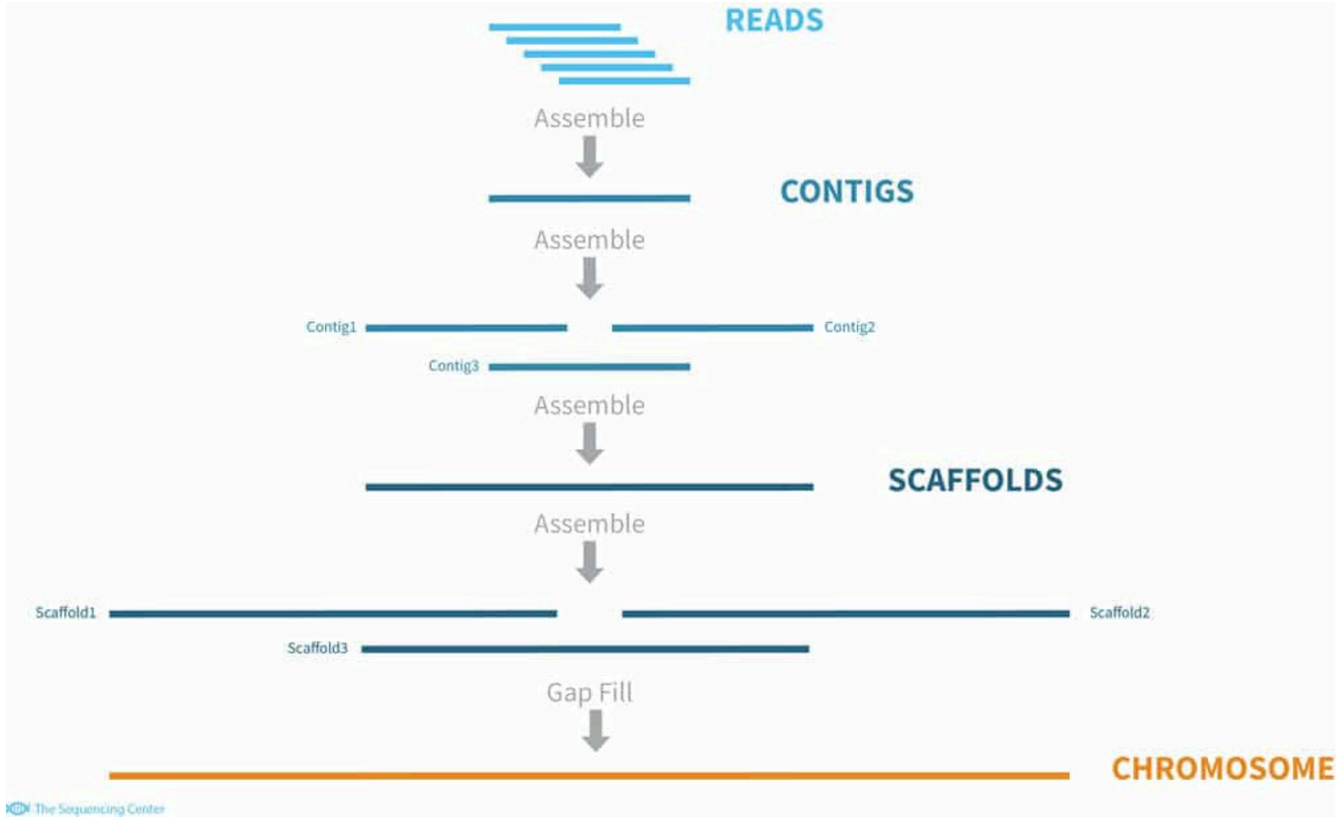






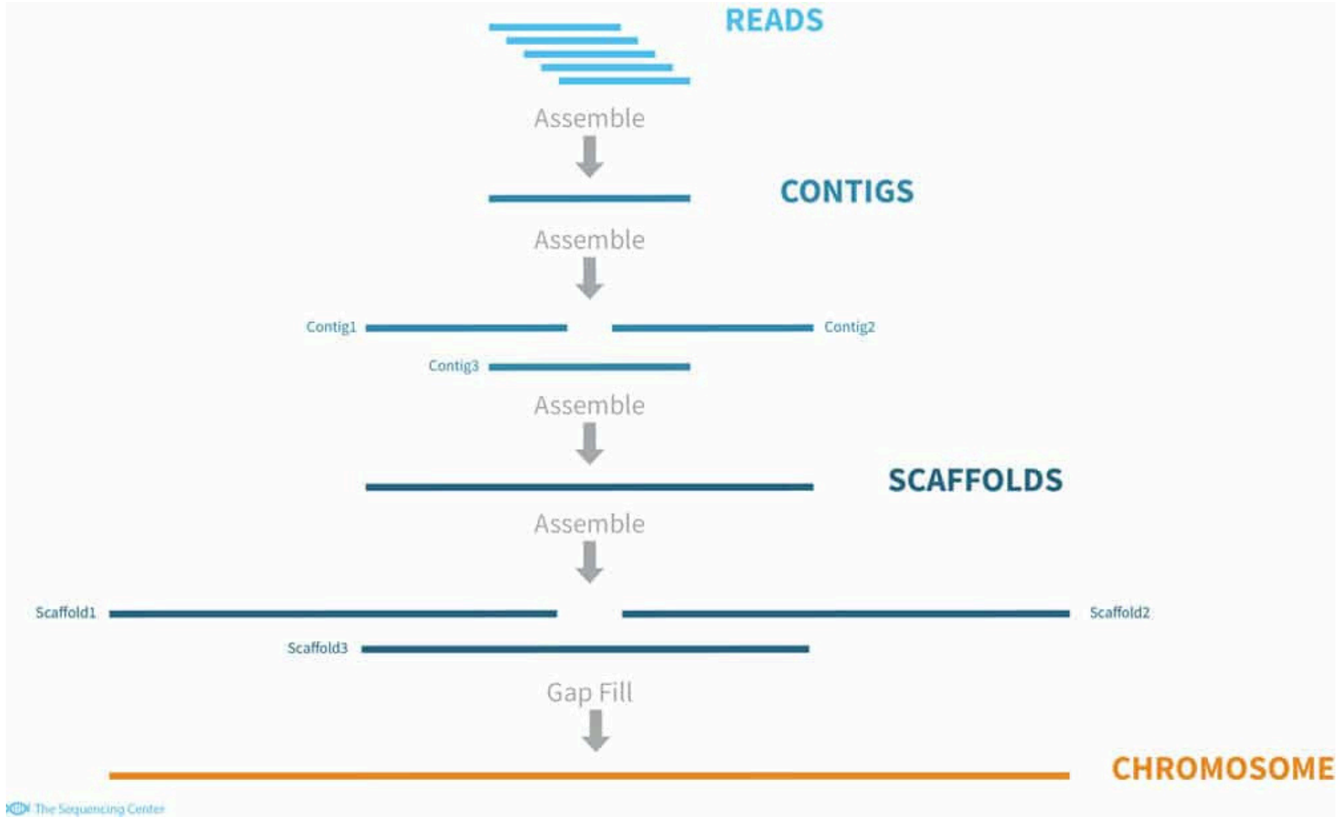
Assembly/Binning

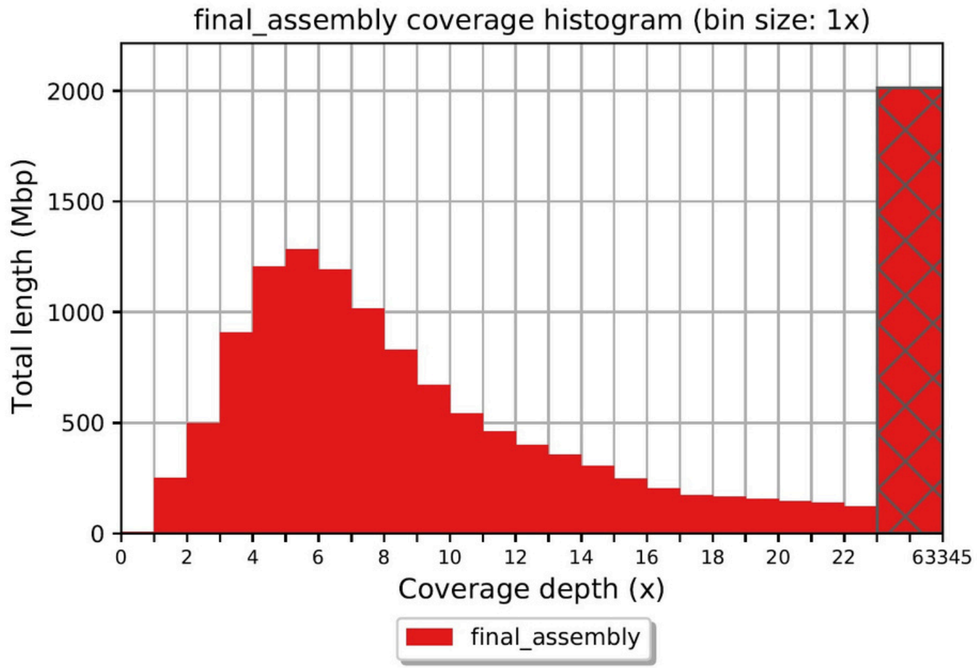
Assembly:



Assembly/Binning

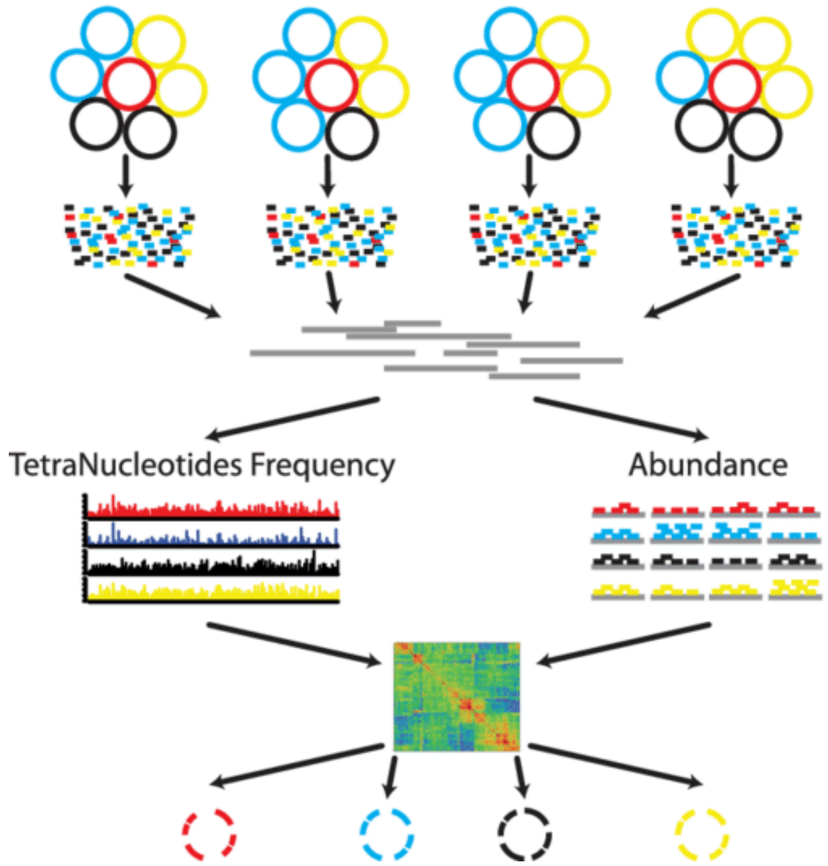
Assembly:





Assembly/Binning

Binning:



Preprocessing

1. Samples from multiple sites or times
2. Metagenome libraries
3. Initial de-novo assembly using the combined library

MetaBAT

4. Calculate TNF for each contig
5. Calculate Abundance per library for each contig
6. Calculate the pairwise distance matrix using pre-trained probabilistic models
7. Forming genome bins iteratively

BigDataBiology/
SemiBin

SemiBin: metagenomics binning with self-supervised deep learning

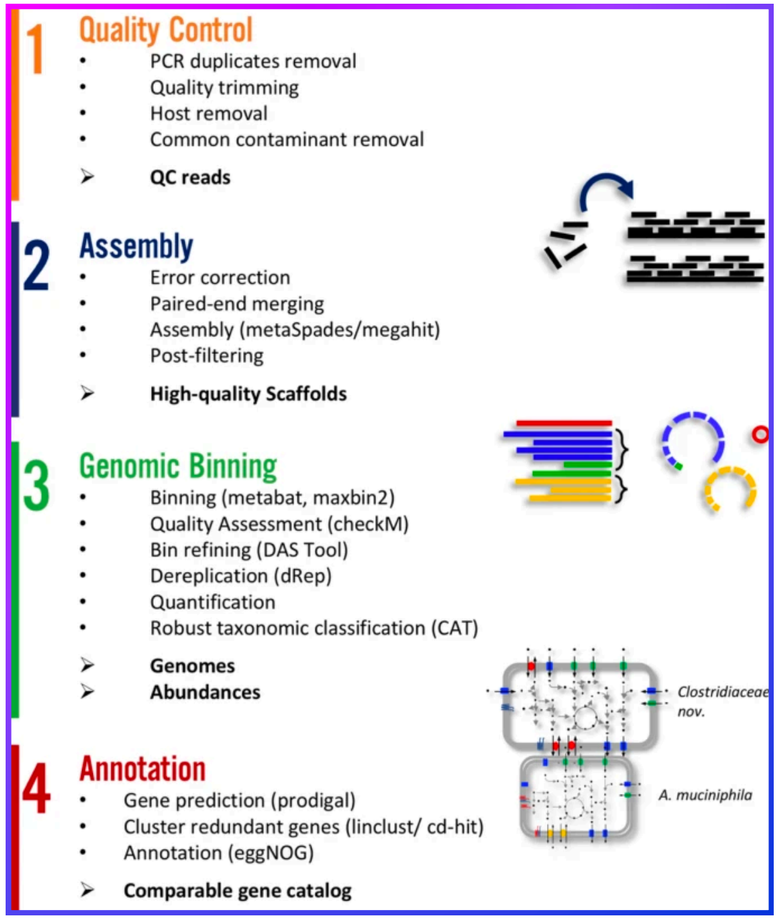
5 Contributors · 12 Issues · 1 Discussion · 102 Stars · 11 Forks

ziyewang/
MetaBinner

2 Contributors · 8 Issues · 39 Stars · 5 Forks

Pipelines

Orchestrated workflow:



Pipelines

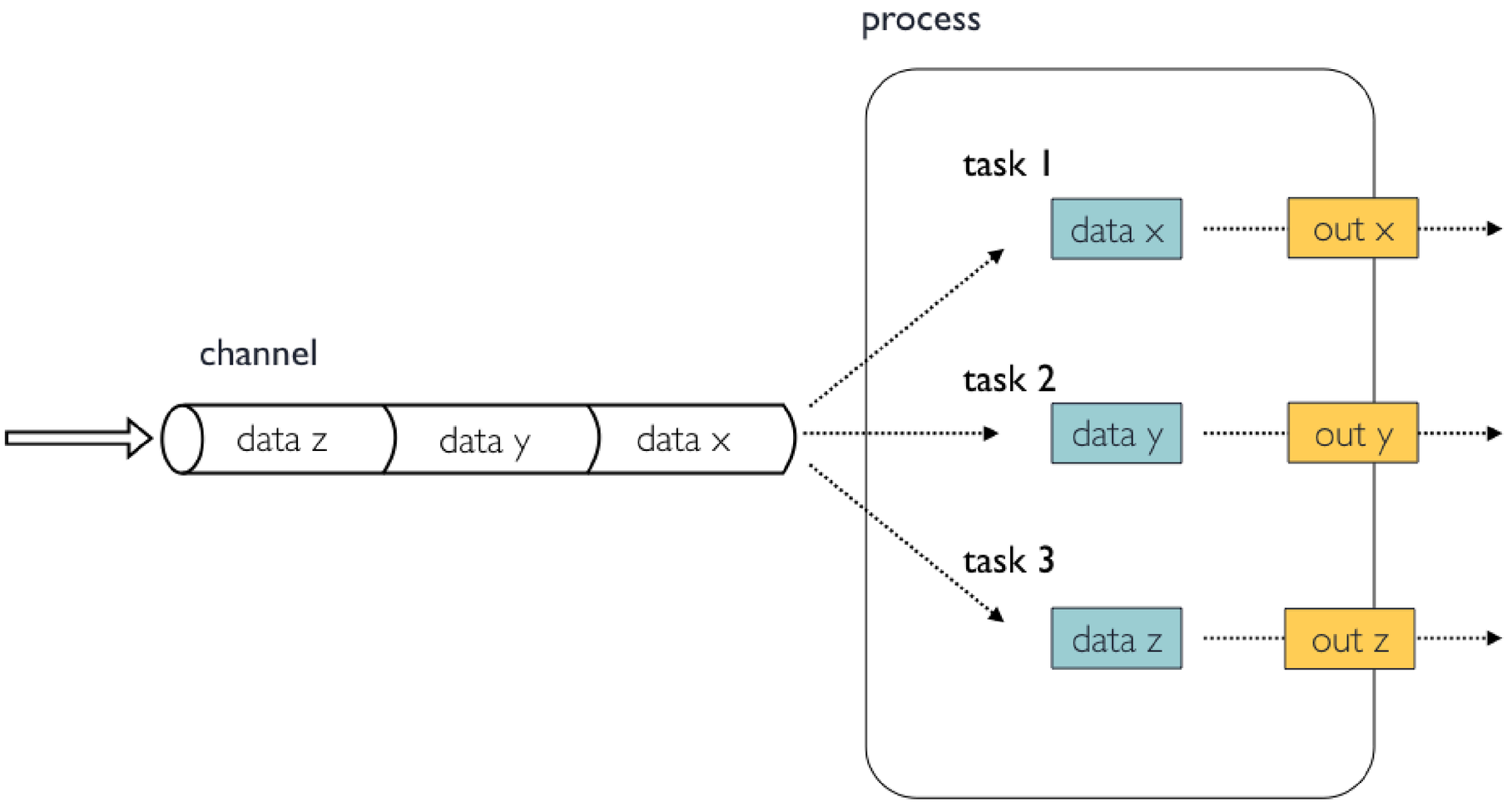
Workflow managers:

Table 1 | Overview of workflow managers for bioinformatics (top, editable version; bottom, image version)

Tool	Class	Ease of use[a]	Expressiveness[b]	Portability[c]	Scalability[d]	Learning resources[e]	Pipeline initiatives[f]
Galaxy	Graphical	●●●	●○○	●●●	●●●	●●●	●●○
KNIME	Graphical	●●●	●○○	○○○	●●◐	●●●	●●○
Nextflow	DSL	●●○	●●●	●●●	●●●	●●●	●●●
Snakemake	DSL	●●○	●●●	●●◐	●●●	●●○	●●●
GenPipes	DSL	●●○	●●●	●●○	●●○	●●○	●●○
bPipe	DSL	●●○	●●●	●●○	●●◐	●●○	●○○
Pachyderm	DSL	●●○	●●●	●○○	●●○	●●●	○○○
SciPipe	Library	●●○	●●●	○○○	○○○	●●○	○○○
Luigi	Library	●●○	●●●	●○○	●●◐	●●○	○○○
Cromwell + WDL	Execution + workflow specification	●○○	●●○	●●●	●●◐	●●○	●●○
cwltool + CWL	Execution + workflow specification	●○○	●●○	●●◐	○○○	●●●	●●○
Toil + CWL/ WDL/Python	Execution + workflow specification	●○○	●●●	●◐○	●●●	●●○	●●○

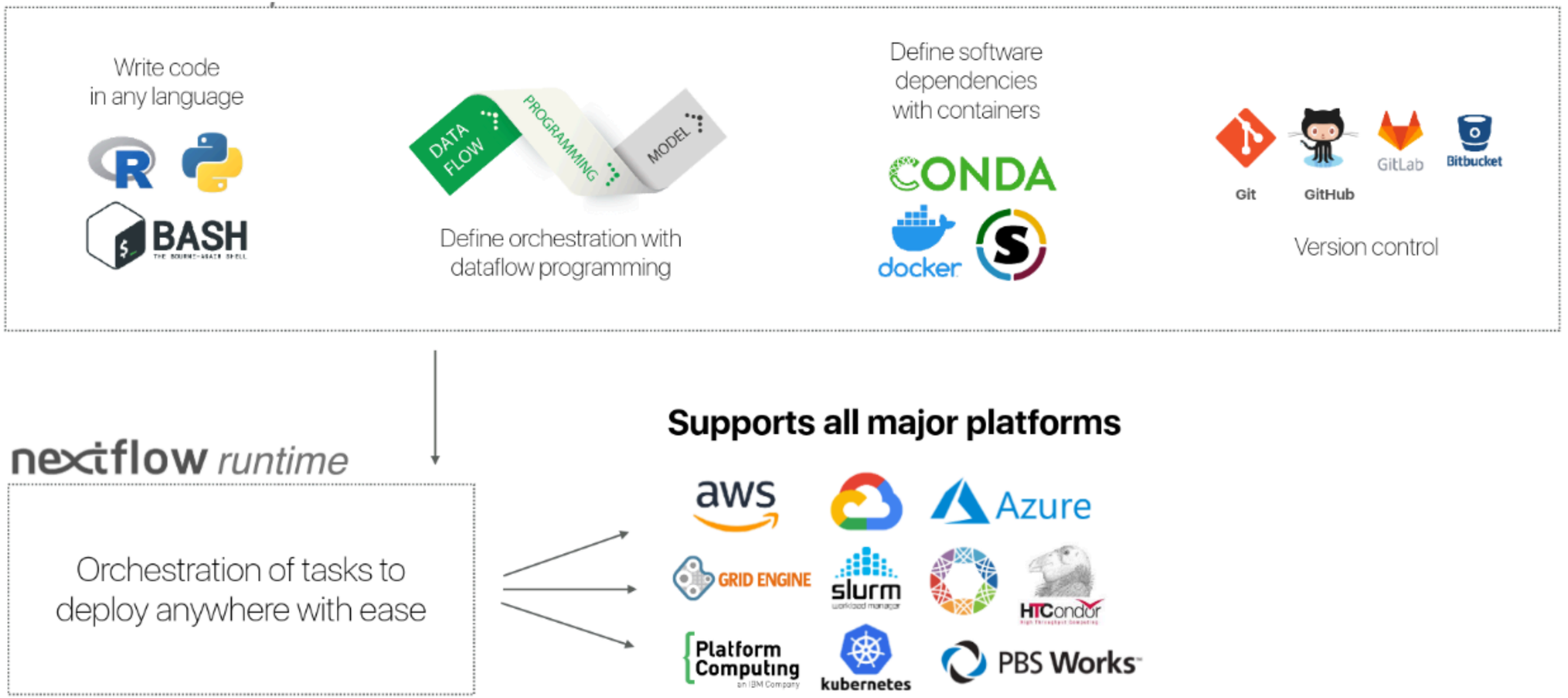
Pipelines

Nextflow:



Pipelines

Nextflow:



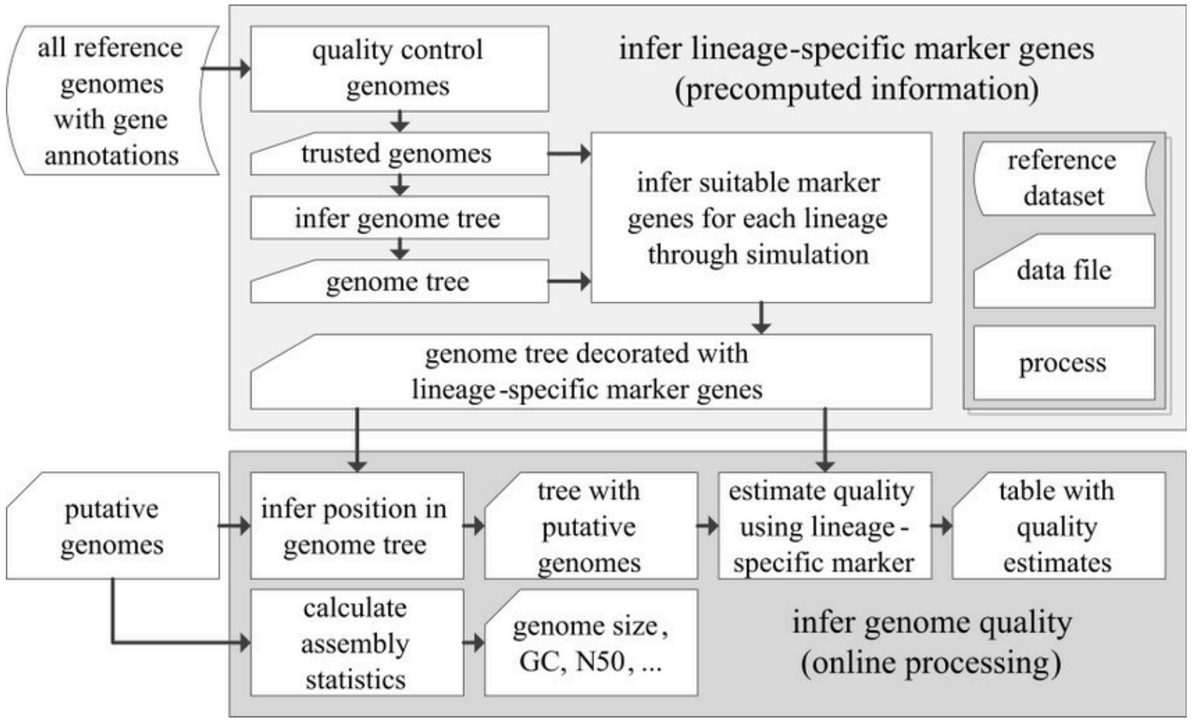
Pipelines

Examples and exercise:

Scan me!

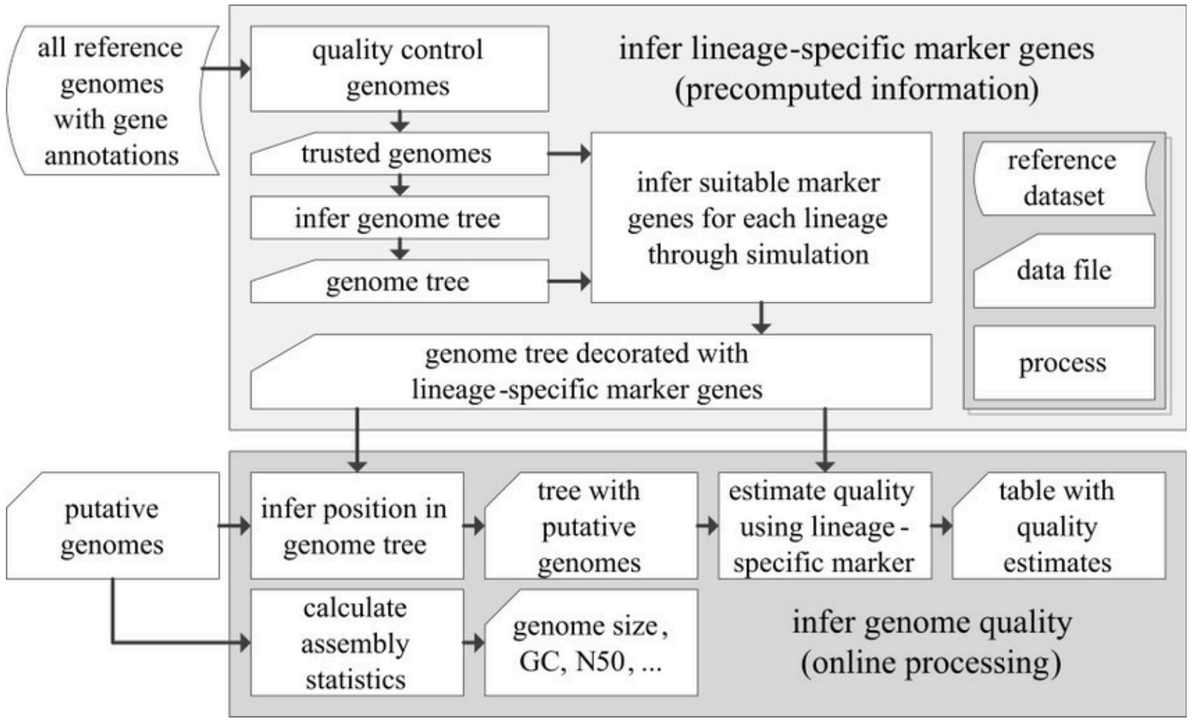
Quality and annotation

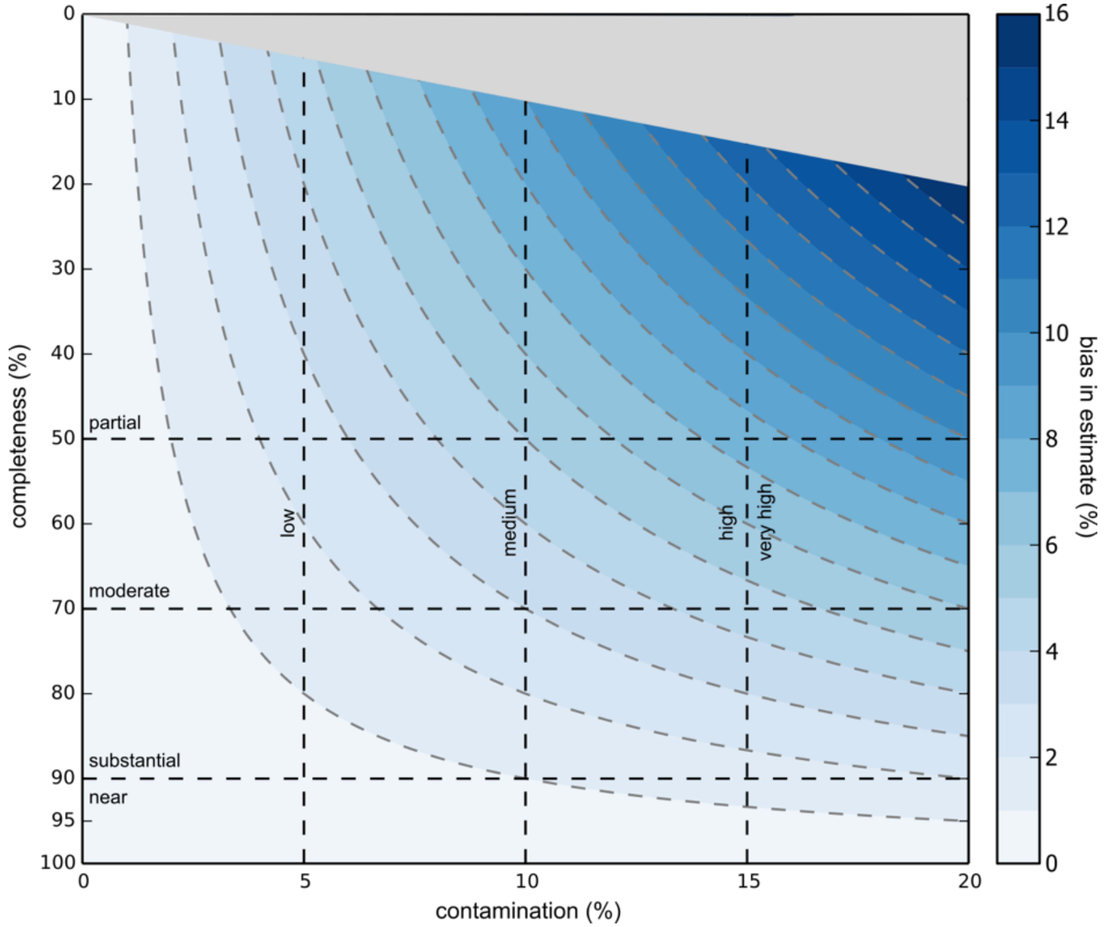
Tools to measure quality: CheckM



Quality and annotation

Tools to measure quality: **CheckM**





Functional Annotation

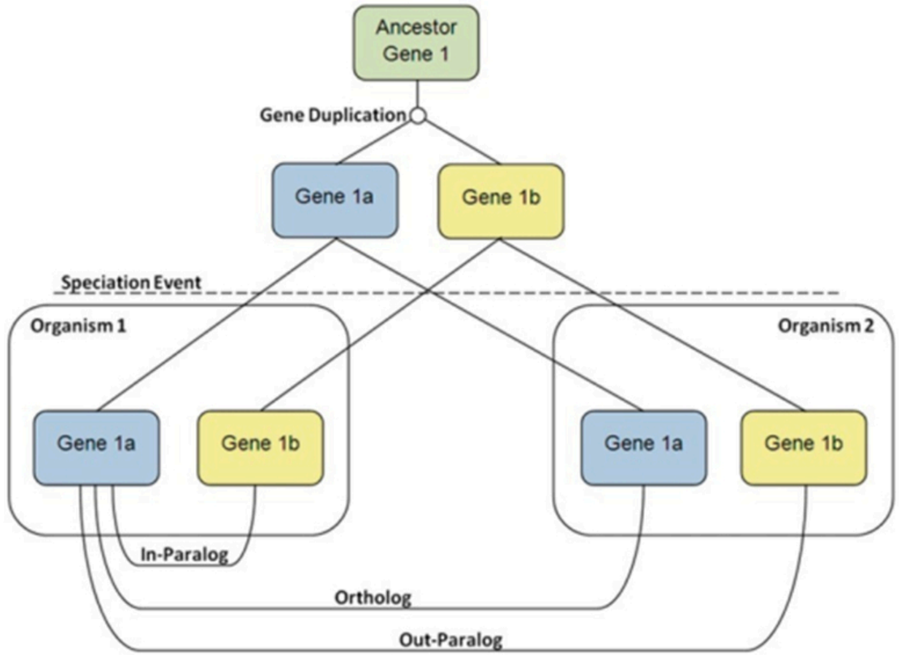
Homology:

Two genes that are similar:

Homologs

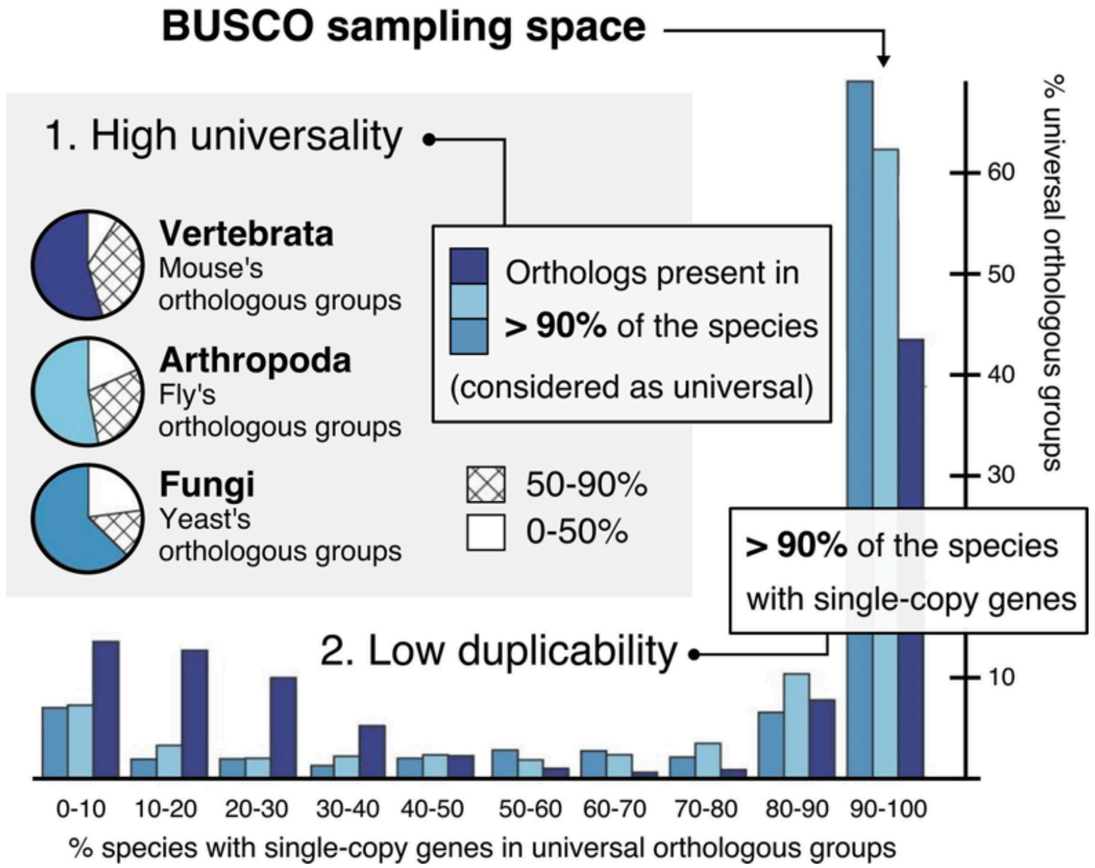
- **Orthologs: Origin of a common ancestor** → **Retain the same function**
- **Paralogs: Gene duplication within the same organism** → **Can retain function or have new function**

Orthologous groups used to identify gene function between different species



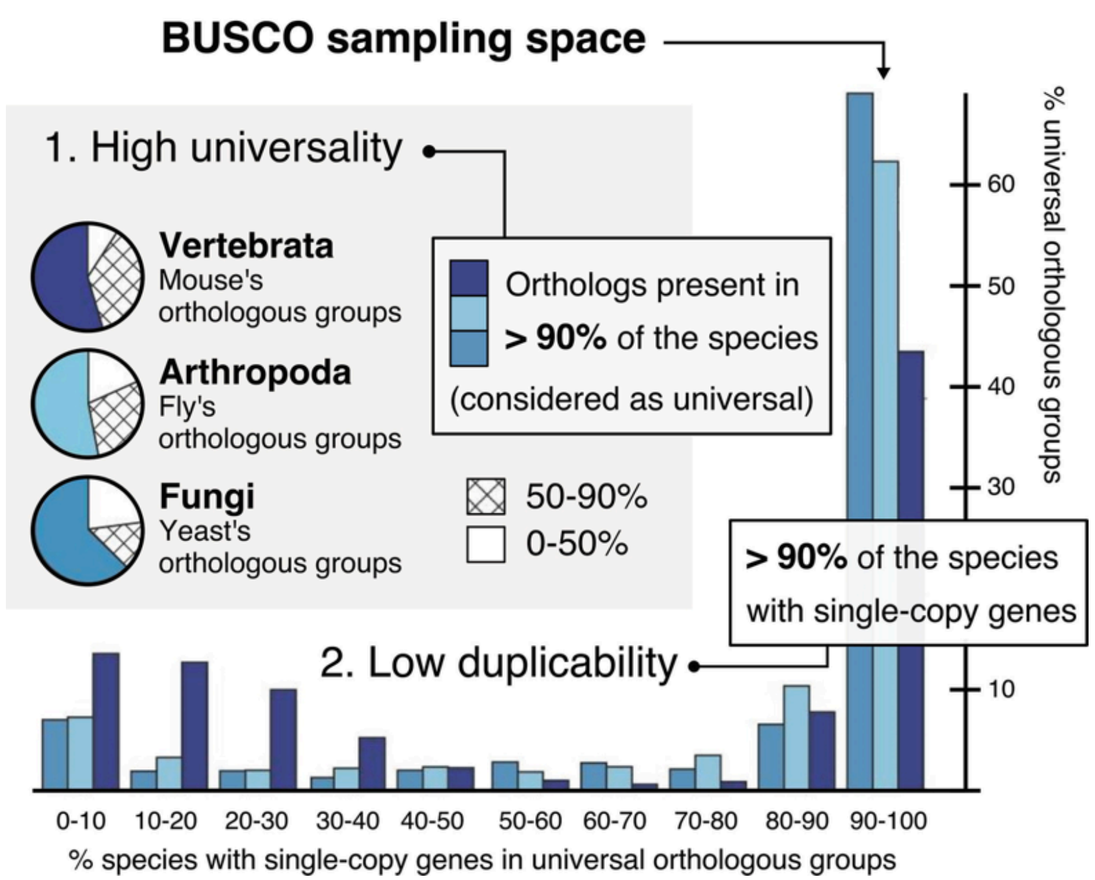
Quality and annotation

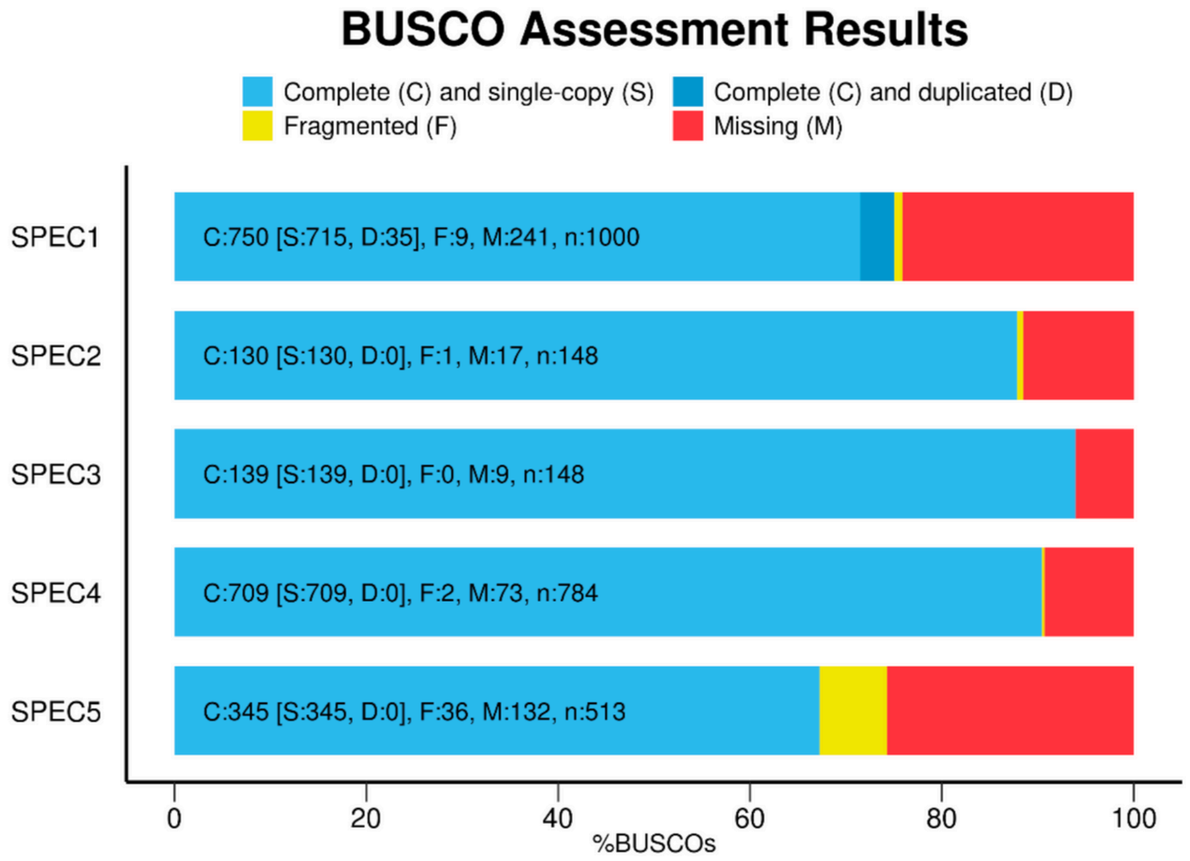
Tools to measure quality: **BUSCO**



Quality and annotation

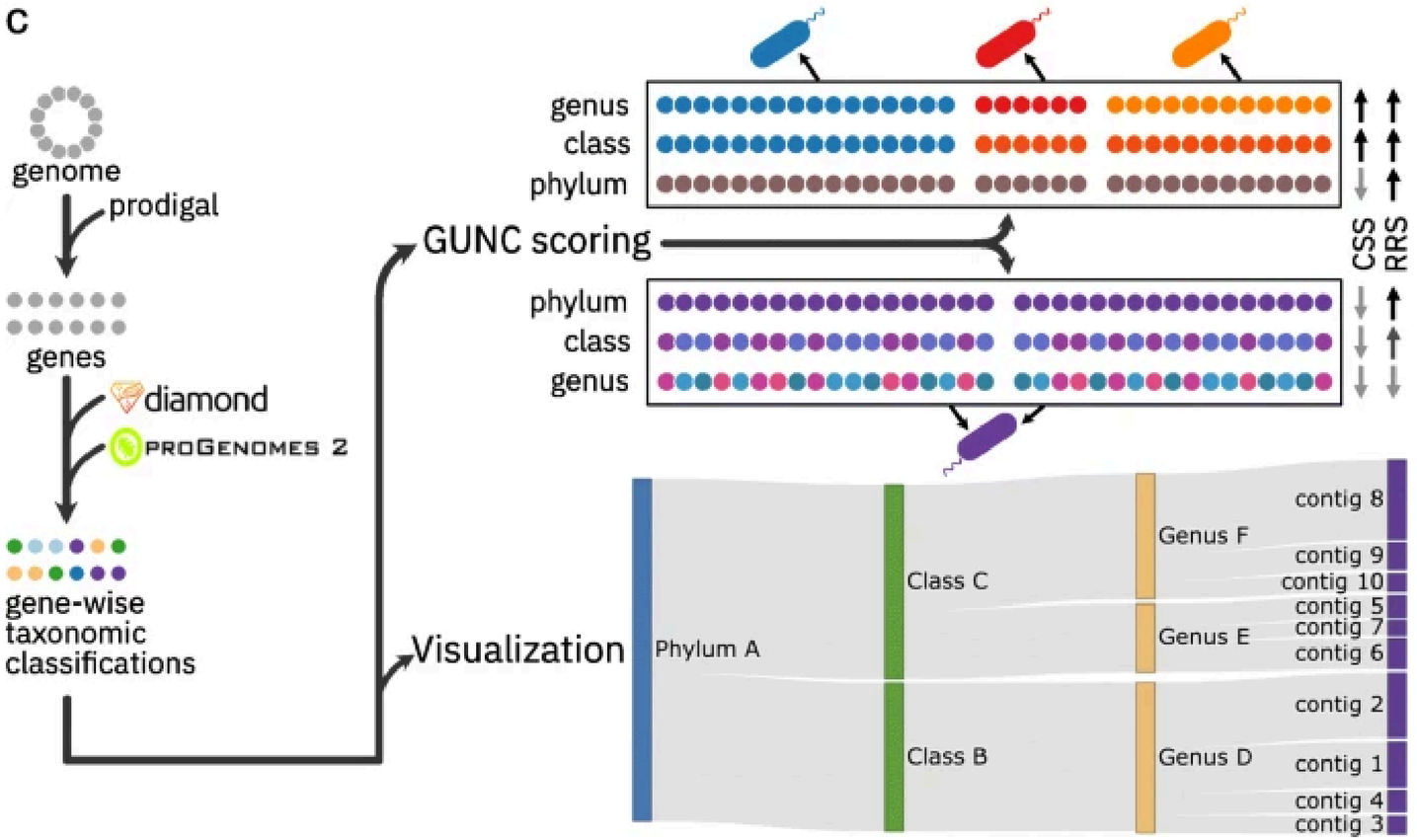
Tools to measure quality: **BUSCO**





Quality and annotation

Tools to measure quality: **GUNC**



The **Clade Separation Score** (CSS) quantifies the degree to which a genome is a chimeric mixture of distinct lineages following non-random distributions across contigs.

Quality and annotation

Tools to measure quality: **QUAST**

Aligned to "BCep_ref" | 8 605 945 bp | 4 fragments | 66.61% G+C
7705 genomic features

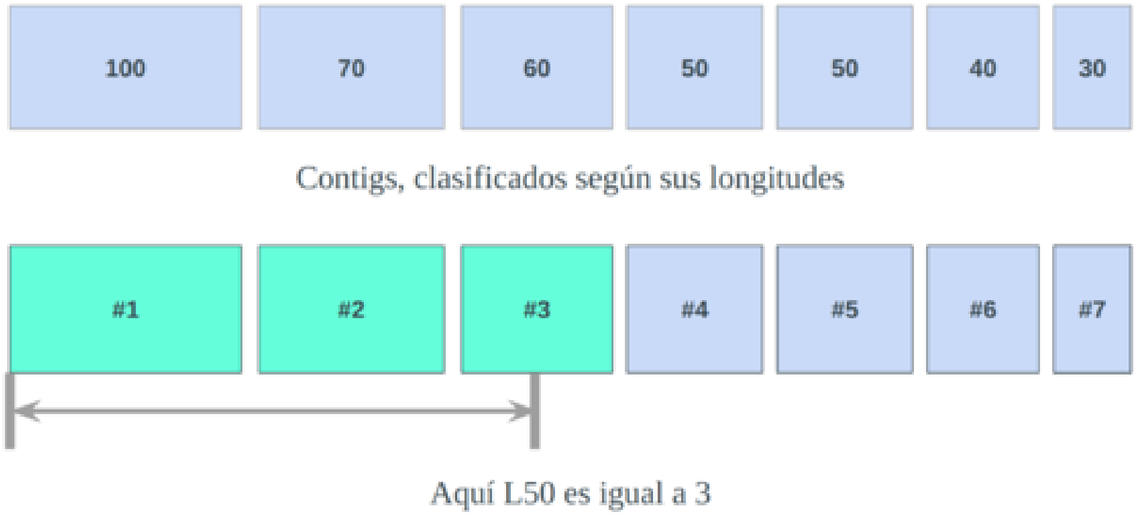
Worst Median Best ☑ Show heatmap

Genome statistics	spades_default	spades_kmers_careful	megahit_default	megahit_min_count_3
Genome fraction (%)	98.14	98.421	98.6	98.603
Duplication ratio	1	1	1.001	1.001
# genomic features	7539 + 75 part	7563 + 62 part	7540 + 105 part	7540 + 104 part
Largest alignment	455 950	505 898	350 746	350 746
Total aligned length	8 436 553	8 470 789	8 492 473	8 493 177
NGA50	143 431	198 969	144 083	125 159
LGA50	18	13	20	21
Misassemblies				
# misassemblies	6	6	9	8
Misassembled contigs length	1 469 048	1 719 134	1 200 775	1 050 989
Mismatches				
# mismatches per 100 kbp	3.85	2.72	2.38	2.3
# indels per 100 kbp	0.67	0.45	0.32	0.28
# N's per 100 kbp	0	0	0	0
Statistics without reference				
# contigs	132	91	156	158
Largest contig	754 490	961 949	539 126	539 126
Total length	8 447 218	8 472 540	8 492 975	8 493 797
Total length (>= 1000 bp)	8 447 218	8 472 540	8 492 975	8 493 797
Total length (>= 10000 bp)	8 324 069	8 384 754	8 296 360	8 308 919
Total length (>= 50000 bp)	7 438 644	7 723 080	6 917 725	6 910 273

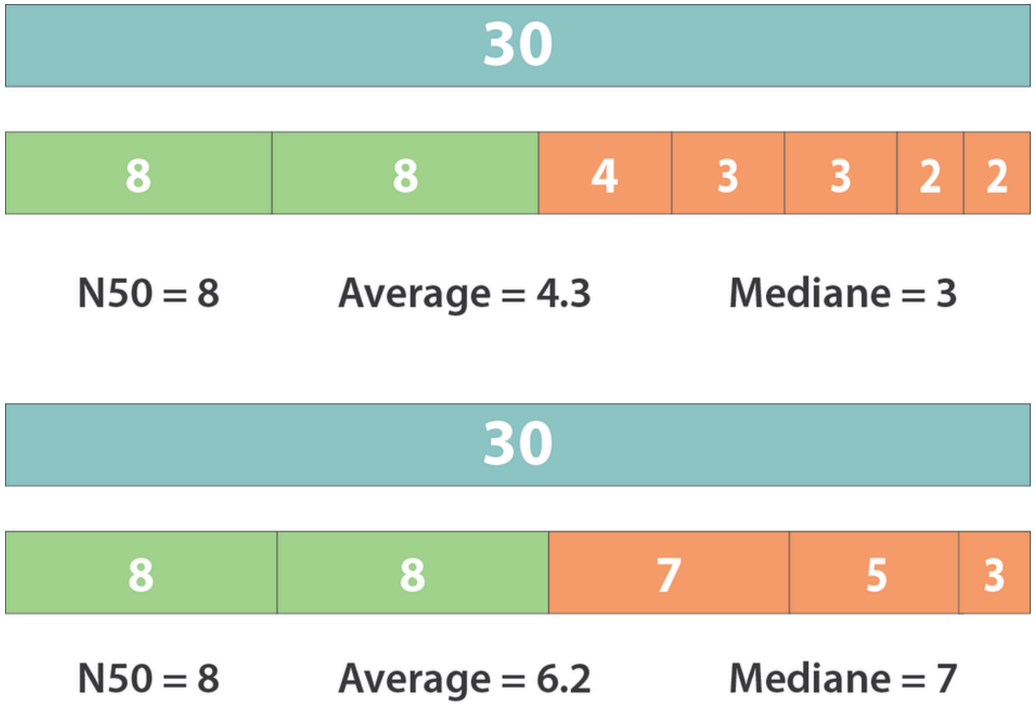
Quality and annotation

Tools to measure quality: **QUAST**

L50

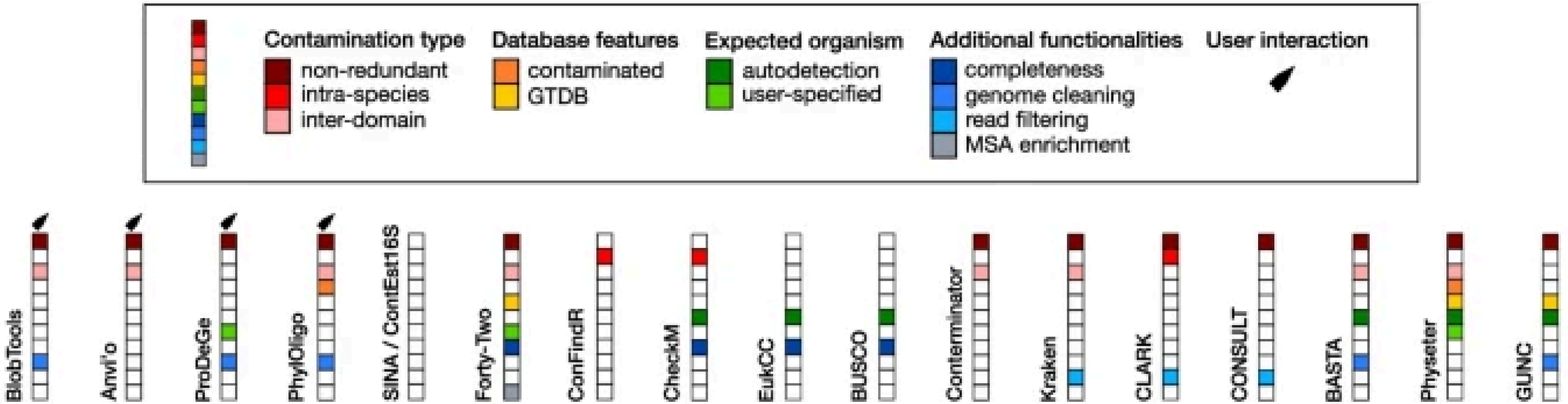


N50



Quality control

Tools to measure quality:

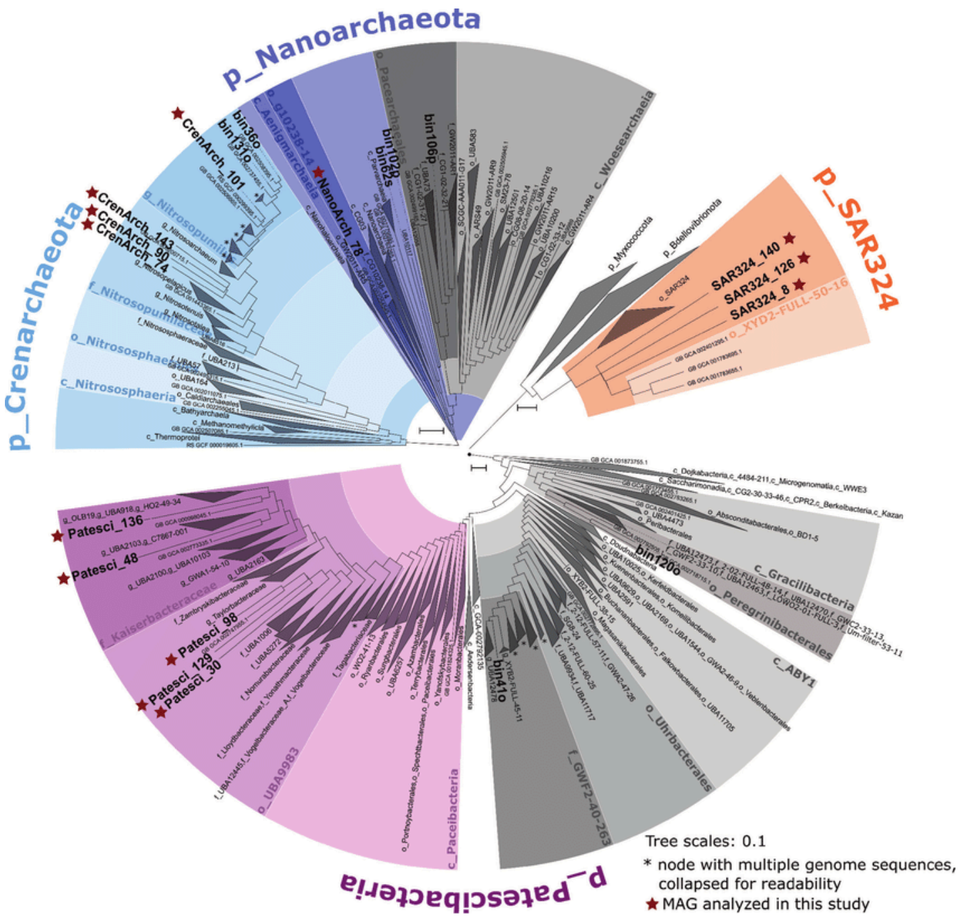


Taxonomic annotation

Taxonomical annotation of MAGs: **GTDB**

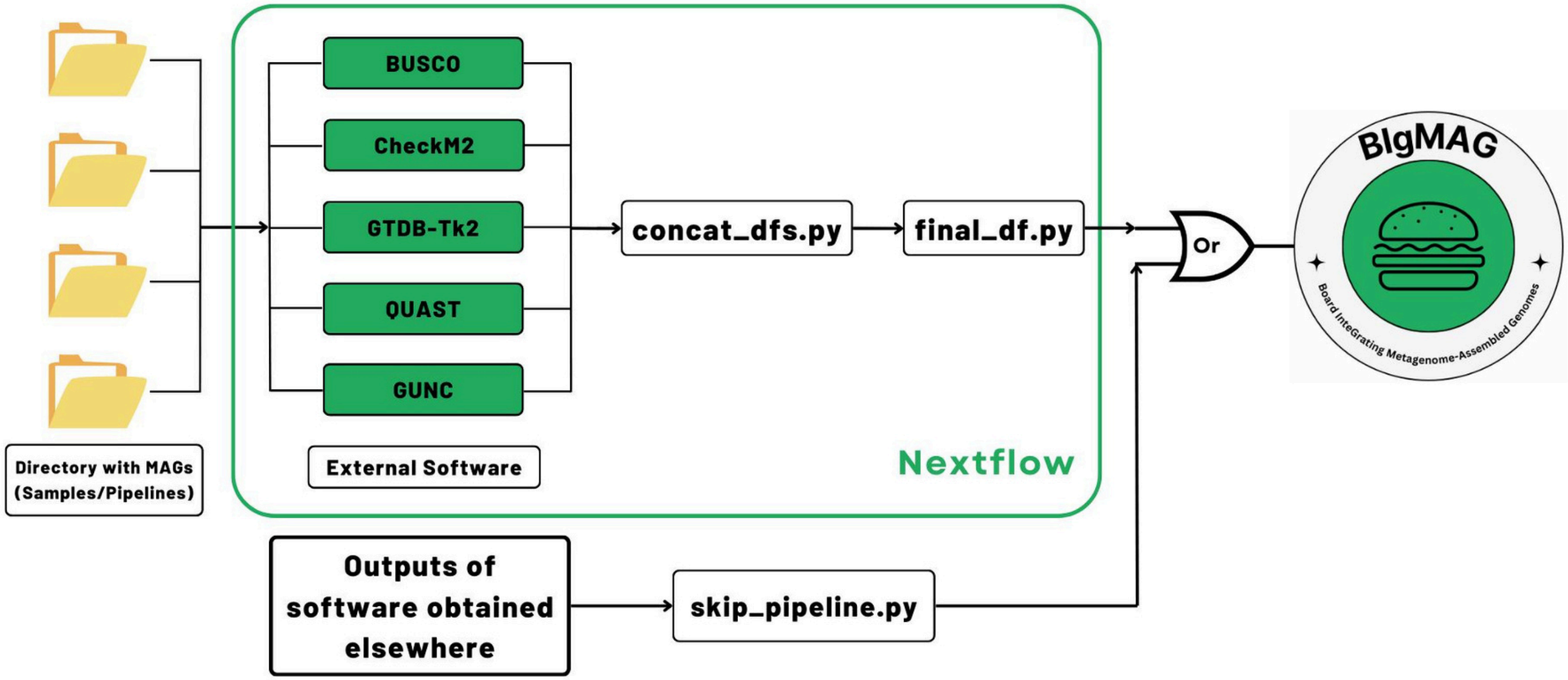
Taxonomic annotation

Taxonomical annotation of MAGs: **GTDB-Tk2**



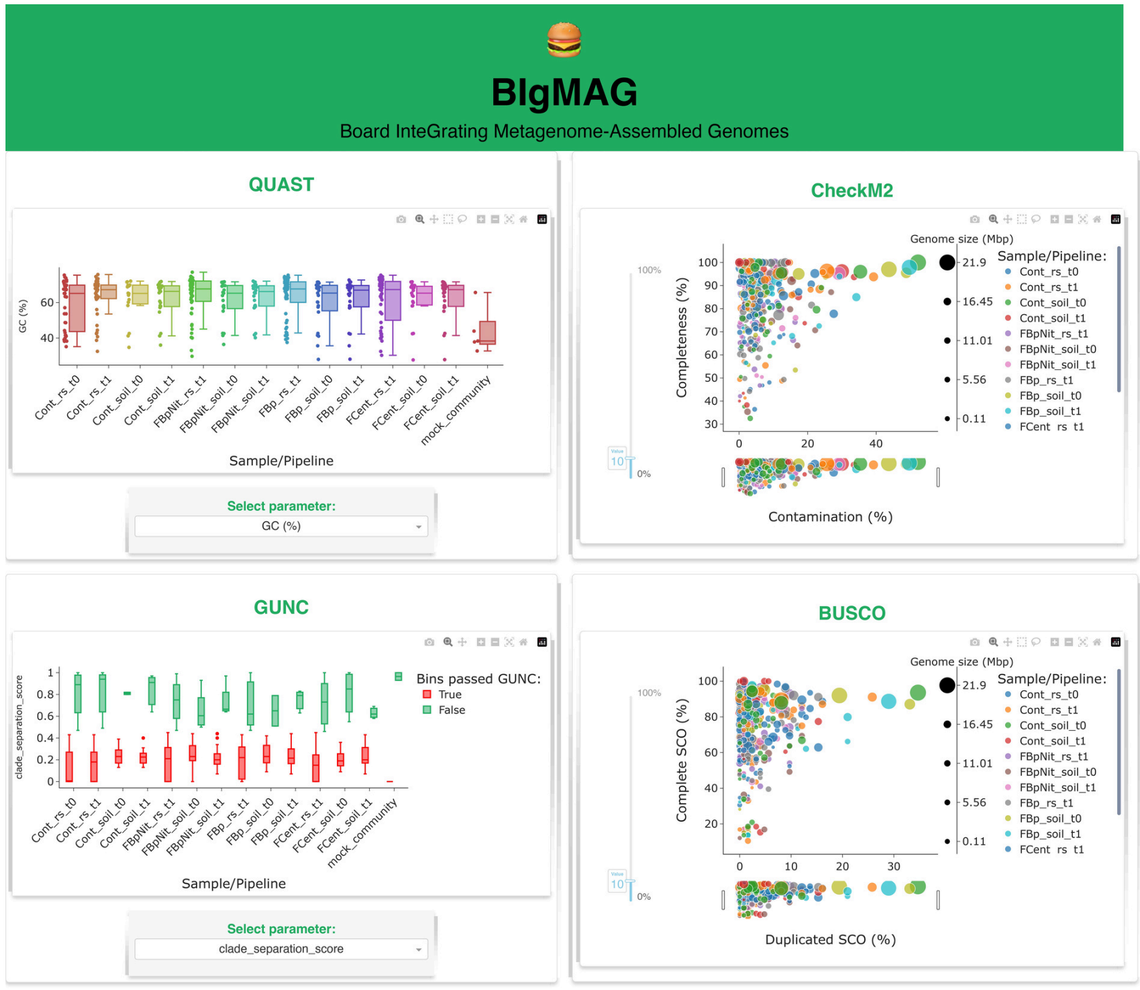
MAGFlow/BIgMAG

Pipeline:



MAGFlow/BIgMAG

Dashboard:



Take-home messages

- Assembling and binning are the core steps for MAG recovery, and therefore the most **computational demanding**.

- Selecting the correct pipeline requires careful analysis of **multiples variables** and **requirements** highly user-dependent.

- Tools/pipeline can have **several parameters to fine-tune**, try to test their effect on the final results do not trust them blindly.

- Try to use as many tools to measure MAG quality in order to assure confidence on your analysis: **Time for Divide-and-Rule Tactics**.