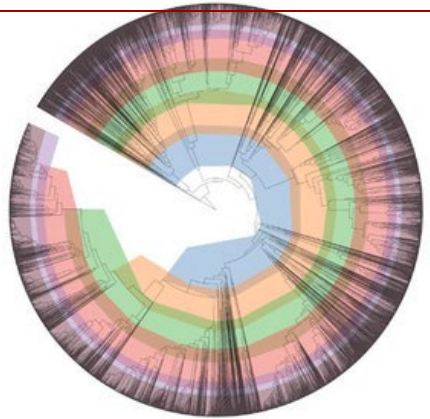
ENVIRONMENTAL METAGENOMICS

Physalia course, online, 11-15 November 2024

MAG binning

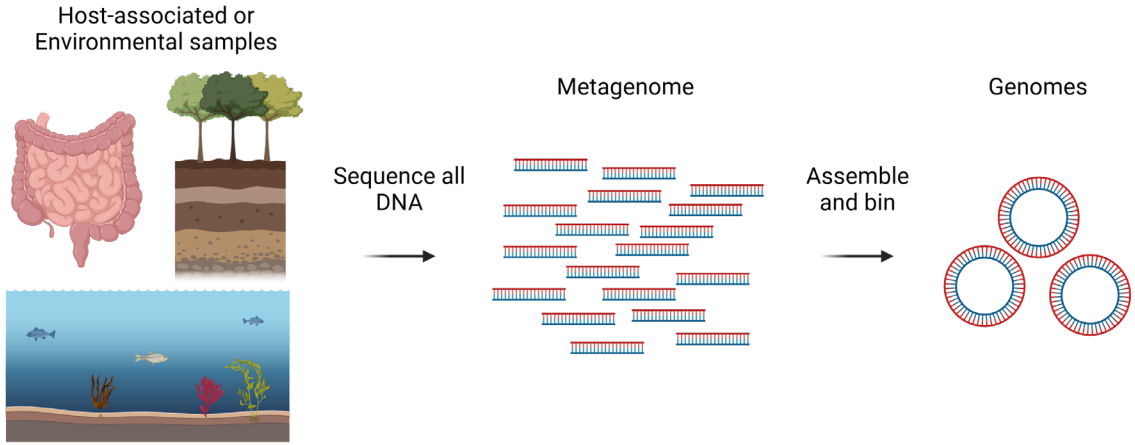
Nikolay Oskolkov, Lund University, NBIS SciLifeLab
Samuel Aroney, Queensland University of Technology

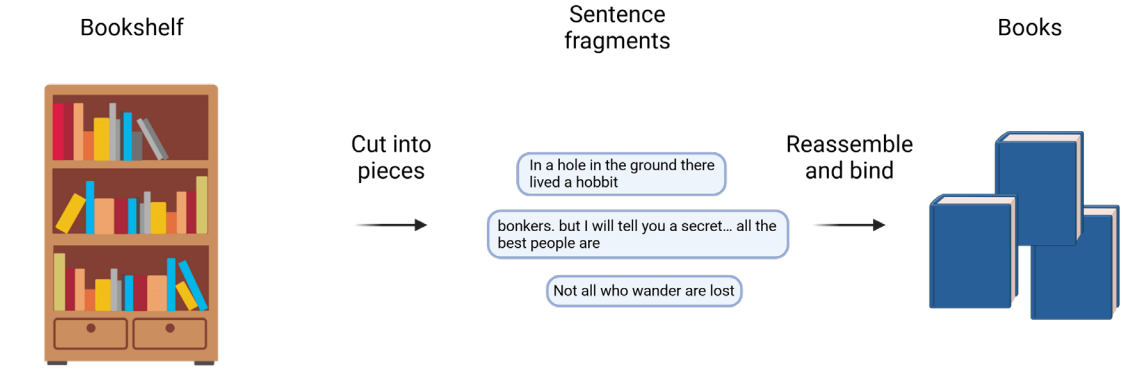


Physalia Courses

NB: original course material courtesy:
Dr. Antti Karkman, University of Helsinki
Dr. Igor Pessi, Finnish Environment Institute (SYKE)
As. Prof. Luis Pedro Coelho

Binning: collating contigs into collections

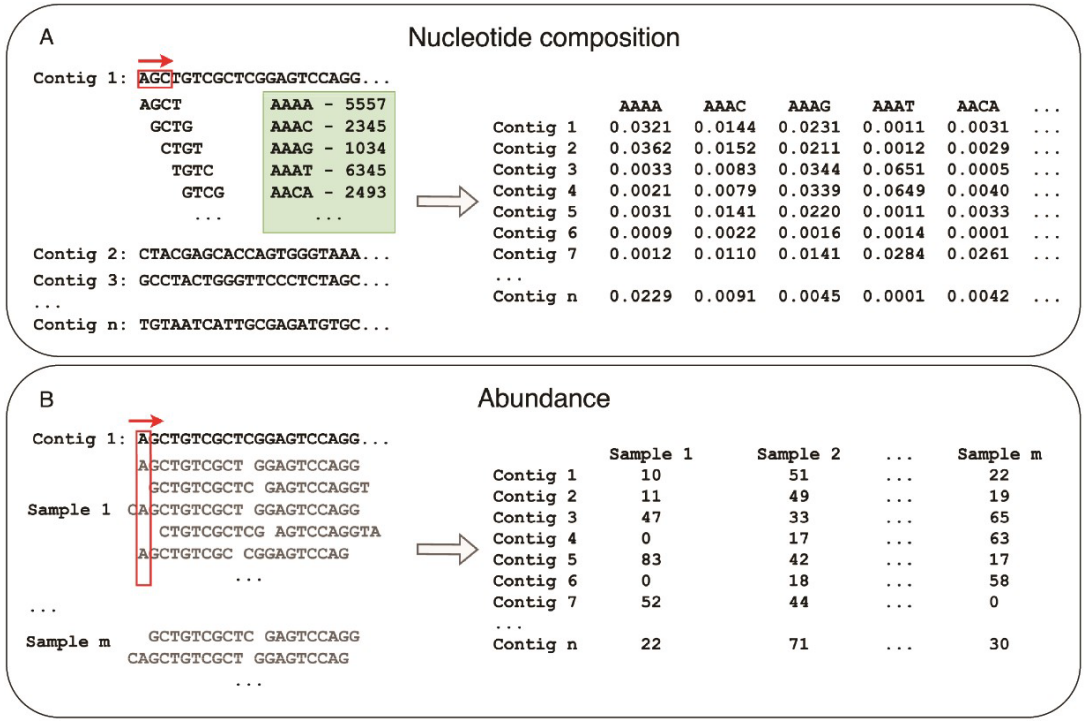




Physalia
Courses

Standard binning strategies

- Nucleotide composition
  - Microbes tend to have similar %GC and codon choice across genome
- Abundance
  - The whole genome should be present at similar copy number



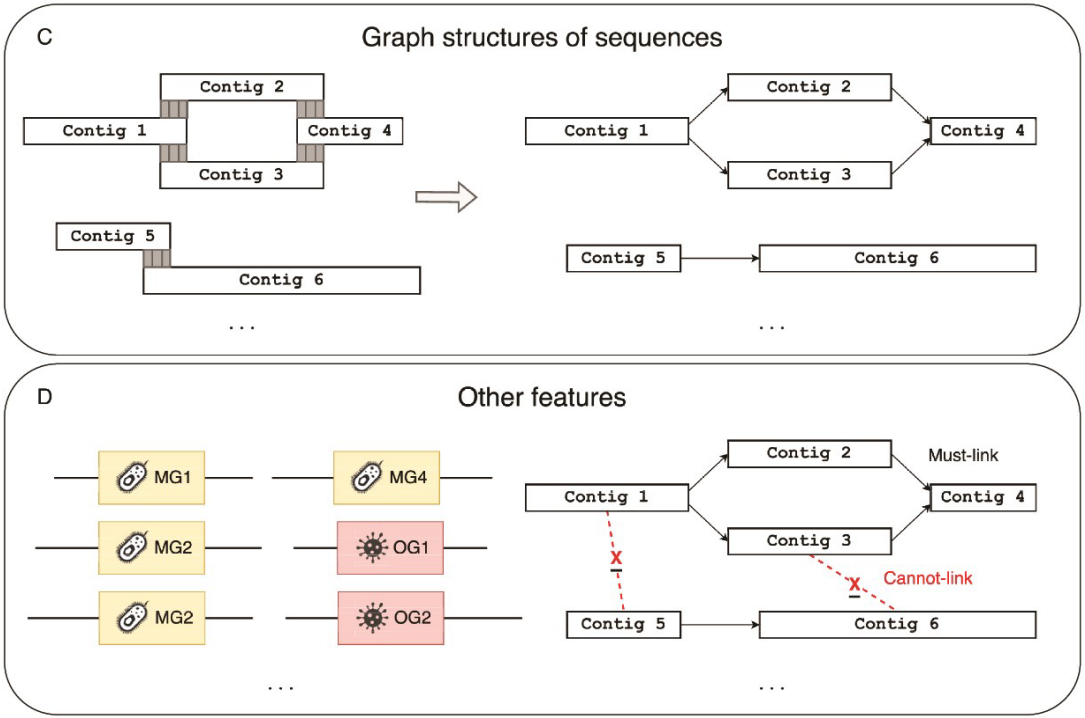
Physalia Courses

Mallawaarachchi et al. *Briefings in Bioinformatics* 2024

More binning strategies

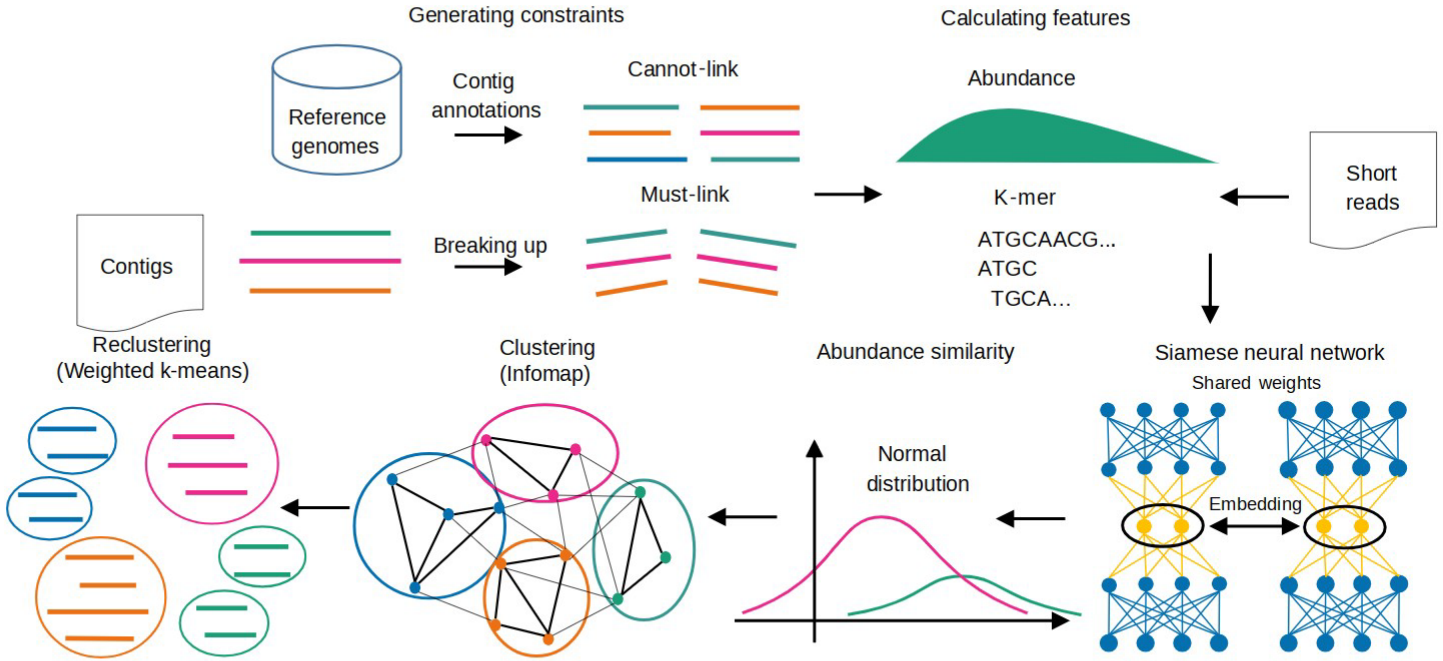
- Graph structures
  - Assembly graph structures can be reused

- Other features
  - Methylation patterns
  - Single-copy marker genes
  - Machine learning inputs



Mallawaarachchi et al. *Briefings in Bioinformatics* 2024

Semibin

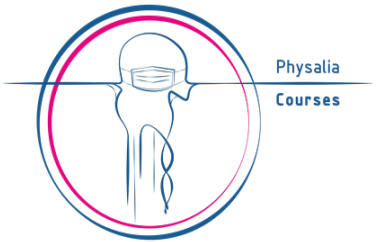


Physalia
Courses

Learning a new model is resource intensive

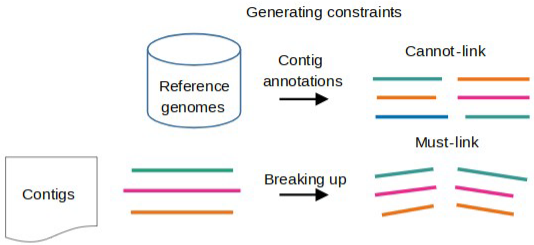
- SemiBin uses a learned model
- Generating the data for training is expensive (mmseqs2 is slow)
- Learning is slow (a GPU helps, but few people have a GPU cluster)

		Single-sample		
		Human gut	Dog gut	Ocean
Computing features	Time(min)	5	9	12
	Memory(MB)	923	736	1,675
Generating cannot-link	Time(min)	88	81	114
	Memory(MB)	39,070	37,904	46,091
Training(CPU)	Time(min)	181	209	222
	Memory(MB)	2,497	2,373	3,211
Training(GPU)	Time(min)	34	36	45
	Memory(MB)	4,487	4,355	5,222
Binning	Time(min)	2	2	3
	Memory(MB)	4,501	3,641	7,622



From Semi- to self-supervised learning

- In SemiBin1, must-links were self-supervised
- In SemiBin1, cannot-links were from a reference taxonomy



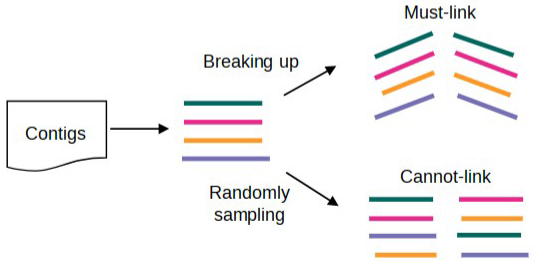
Use random sampling to generate cannot-links

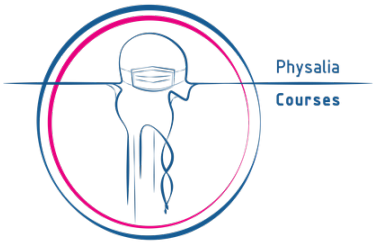
Random sampling advantages

1. Random sampling is faster than MMSeqs2
2. It has no database biases

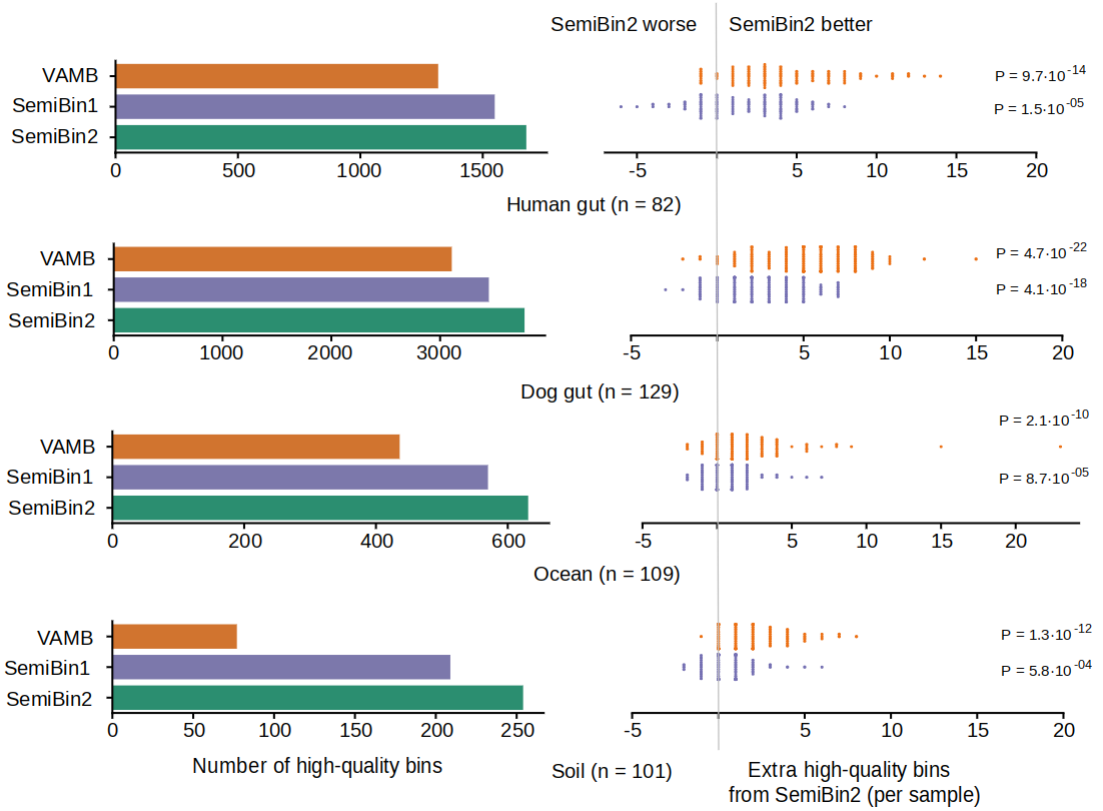
Random sampling disadvantages

1. Random sampling introduces errors (naturally)





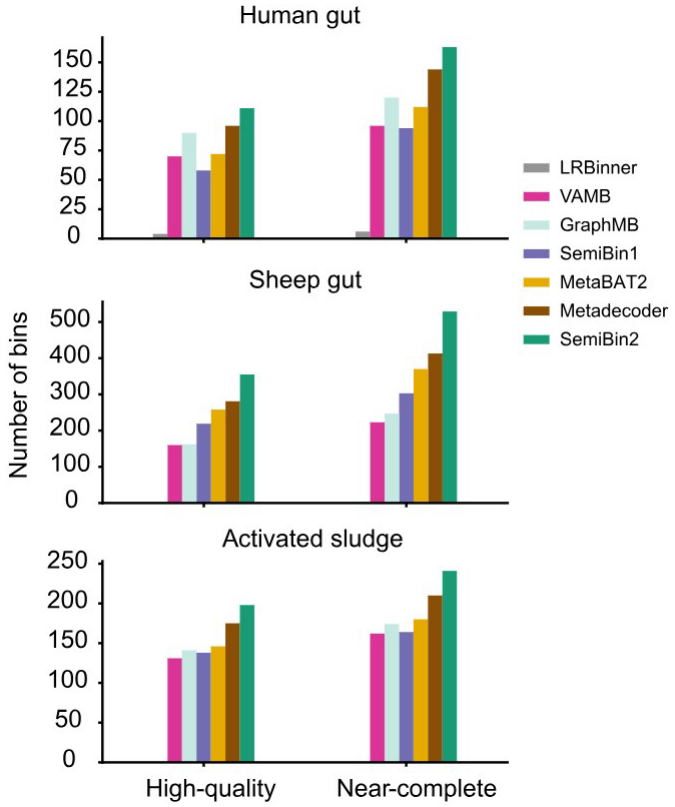
Semibin vs Semibin2



Physalia
Courses

SemiBin2 surpasses alternatives for long-read binning

- Three datasets: human gut, sheep gut, and activated sludge
- Some of the alternatives (like SemiBin1) are not designed for long-reads
- We used CheckM2 & GUNC to evaluated the results
- SemiBin2 is also best using checkM1, but the two tools share some of the same marker genes (so evaluation is partly circular)



Physalia
Courses

All the binners

Binning metagenomic contigs by coverage and composition

Johannes Alneberg[1,8], Brynjar Smári Bjarnason[1,8], Ino de Bruijn[1,2], Melanie Schirmer[3], Joshua Quick[4,5], Umer Z Ijaz[3], Leo Lahti[6,7], Nicholas J Loman[4], Anders F Andersson[1,9] & Christopher Quince[3,9]

Effective binning of metagenomic contigs using contrastive multi-view representation learning

Received: 28 June 2023
Accepted: 7 December 2023

Ziye Wang[1], Ronghui You[1], Haitao Han[1], Wei Liu[1], Fengzhu Sun[2] & Shanfeng Zhu[1,3,4,5,6]

A de
met
biome
datasets across different environments

Shaojun Pan[1,2], Chengkai Zhu[1,2,3], Xing-Ming Zhao[1,2,4,5] & Luis Pedro Coelho[1,2]

MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets

Yu-Wei Wu[1,2,*], Blake A

Improved metagenome binning and assembly using deep variational autoencoders

Jakob Nybo Nissen[1,2], Joachim Johansen[2], Rosa Lundbye Allesøe[2], Casper Kaae Sønderby[3], Jose Juan Almagro Armenteros[1], Christopher Heje Grønbech[3,4], Lars Juhl Jensen[2], Henrik Bjørn Nielsen[5], Thomas Nordahl Petersen[6], Ole Winther[3,4,7] and Simon Rasmussen[2]

MetaBAT, an efficient tool for accurately reconstructing single genomes from
c
D
[1] De
[2] Ge
[3] Sch

MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies

Dongwan D. Kang[1], Feng Li[2], Edward Kirton[1], Ashleigh Thomas[1], Rob Egan[1], Hong An[2] and Zhong Wang[1,3,4]

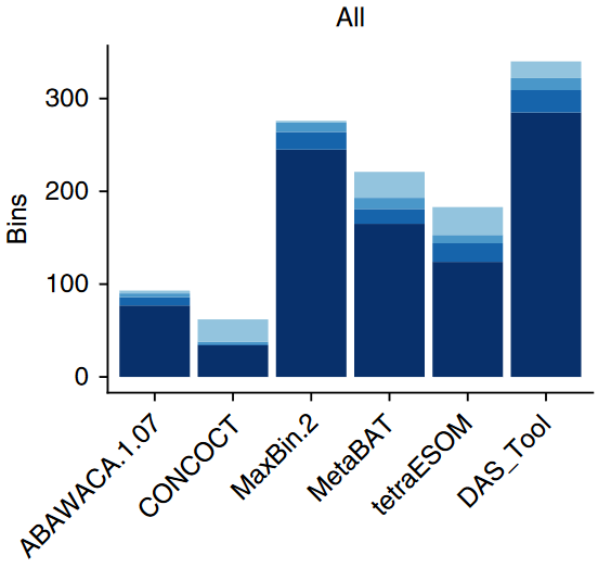
SemiBin2: self-supervised contrastive learning leads to better MAGs for short- and long-read sequencing

Shaojun Pan[1,2], Xing-Ming Zhao[1,2,3,4,*], Luis Pedro Coelho[1,2,*]

Ensemble binning

- Combining bins from multiple binners

  - DAS Tool
  - MetaWRAP
  - MetaBinner
  - BASALT



Sieber et al. *Nature Microbiology* 2018

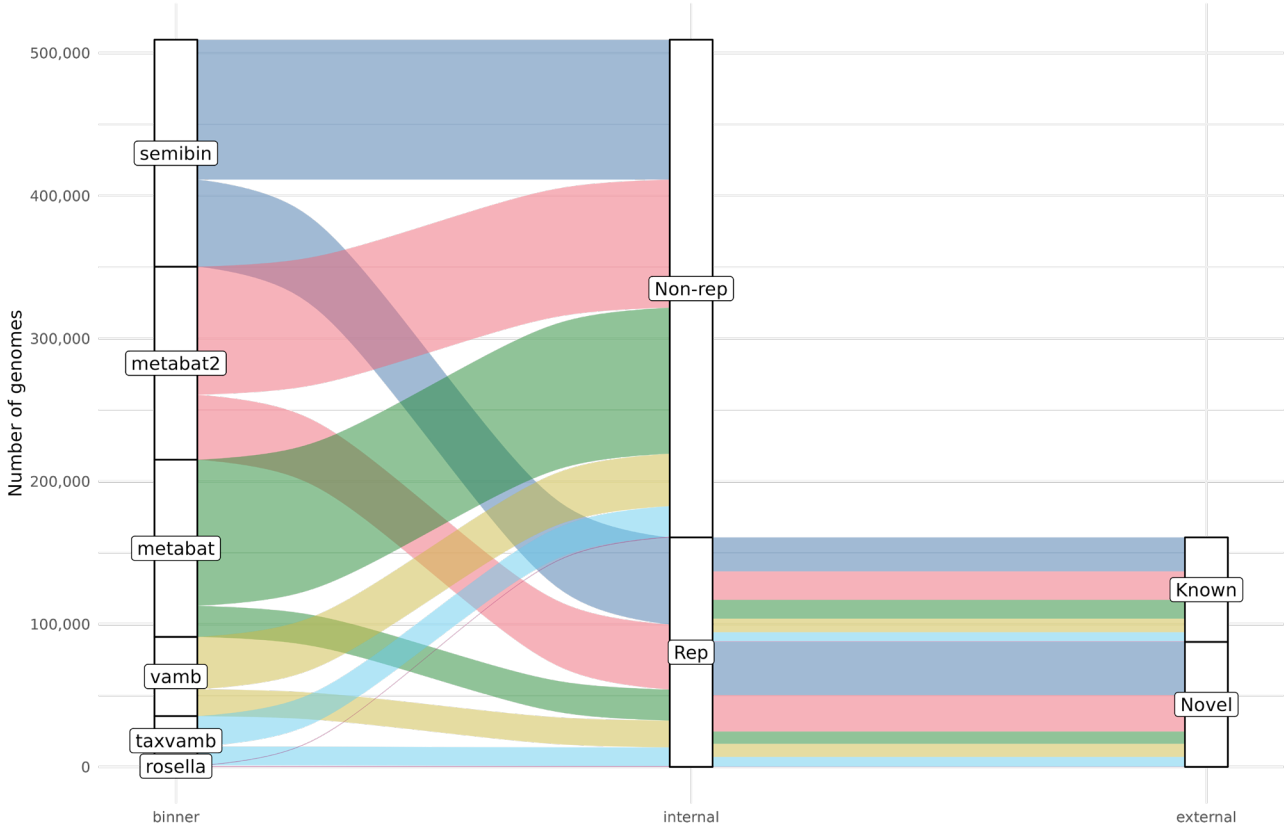
Aviary: Our Assembly + Binning pipeline

- Snakemake pipeline (automatically submit jobs to a cluster)
- Assembly: metaSPAdes or MegaHit
  - Particularly good at hybrid short+long read
- Binning
  - MetaBAT1/2, VAMB, SemiBin1/2, CONCOCT, MaxBin2, Rosella
  - DAS Tool

Physalia
Courses

All the binners - comparison

semibin = Semibin2



Physalia Courses