

BENCHMARKING METAGENOMICS BINNING TOOLS ON REAL DATASETS ACROSS SEQUENCING

PLATFORMS AND BINNING MODELS.

combination of data types and binning models
DATA-BINNING COMBINATIONS
long/short reads
single/multi sample, co-assembly

Metagenomes → sequences from an environmental sample which contains more than one organism's DNA

Settings → microbial metagenomics: some of their characteristics remain undiscovered because of impossibility to culture

- ↳ SOLUTION:
 - de novo assembly of DNA sequences from metagenomes,
 - reconstruction of DNA sequences without references
 - binning into metagenome-assembled genomes (MAGs), tries to group configs coming from the same organism

↳ GENOMIC SIGNATURES: GC content, k-mer frequencies, codon usage

from assembled configs to binned MAGs achieved through:

these "features" are used by binning tools to generate MAGs

- SEQUENCE COMPOSITION (genomic signatures)
- COVERAGE PROFILES

BINNING TOOLS OVERVIEW

- CONCOCT: sequence composition + coverage as config features, PCA, Gaussian Mixture Models for clustering configs
- MaxBin2: estimates likelihood of a given config belonging to a particular genome using tetranucleotide frequencies and coverage. This probability is used in an Expectation-Maximization algorithm
 - ↳ calculate probability of belonging to each cluster, update parameters, look for best fit.
- MetaBAT2: pairwise similarities between configs using tetranucleotide frequency and coverage, then clusters through Label Propagation Algorithm (LPA) → each config begins with a label that propagates in the similarity cluster. The label is updated based on similarity weights and statistical confidence

- Binny: k-mer composition and coverage for dimensionality reduction, then hierarchical density-based spatial clustering (HDBSCAN) \hookrightarrow t-SNE
- MetaDecoder: Dirichlet Process Gaussian Mixture Model (DPGMM) for preliminary clustering, semi-supervised k-mer frequency probabilistic model and GMM for further clustering.
- MetaBinner: stand alone ensemble algorithm employing partial seed k-means and multiple features
 \hookrightarrow combine models to improve guess
- VAMB: deep variational autoencoders (VAE) to encode tetranucleotide frequency and coverage for each config. The latent representation is processed with a clustering algorithm.
- CLMB: introduces simulated noise for each config and uses contrastive learning for config embeddings
- SemiBin1: deep siamese neural networks to build (or not build) links - semi supervised
- SemiBin2: self-supervised, ensemble based DBSCAN for long-read data
- CoMeBin: data augmentation to generate multiple views for each config, contrastive learning for high quality embeddings, Leiden-based method for clustering
- MetaWRAP, DAS Tool, MAGSciT: bin refinement tools combining multiple binning tools to reconstruct high quality MAGs

Metagenomic binning:

- 1) co-assembly: all sequencing samples into one assembly, configs binned with coverage information calculated across samples (co-abundance information, but chimeric configs, no sample specific validation)
- 2) single-sample: assembly and binning independently within each sample
- 3) multi-sample: calculates coverage information across samples

best quality genome kept
remove redundant MAGs

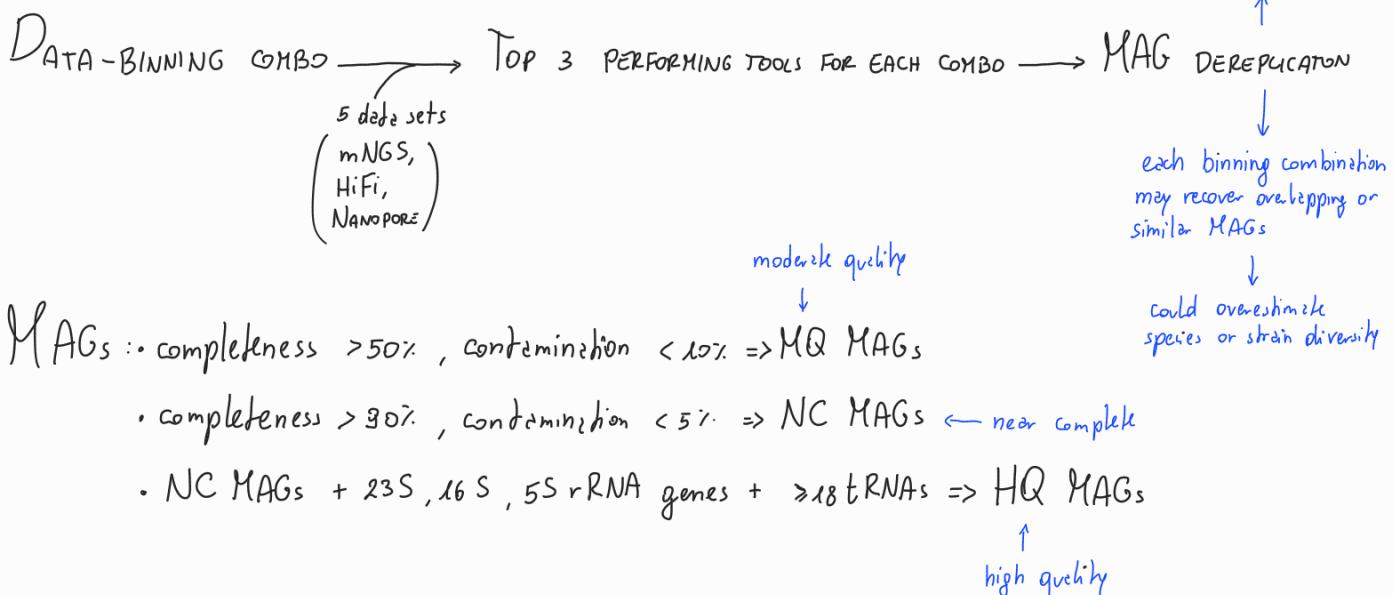


Fig. 2 - Number of MAGs retrieved from the data-binning combinations by the binning tools were different



SHORT READ DATA: - co-assembly , fewest MC, NC, HQ MAGs

- multi-sample compared to single sample:

- 22% more MQ and comparable NC, HQ MAGs (human gut I);
- 44% more MQ, 82% more NC, 233% more HQ (human gut II);
- 100% more MQ, 186% more NC, 82% more HQ (marine);
- better (cheese, sludge).

LONG READ DATA: - multi-sample compared to single sample:

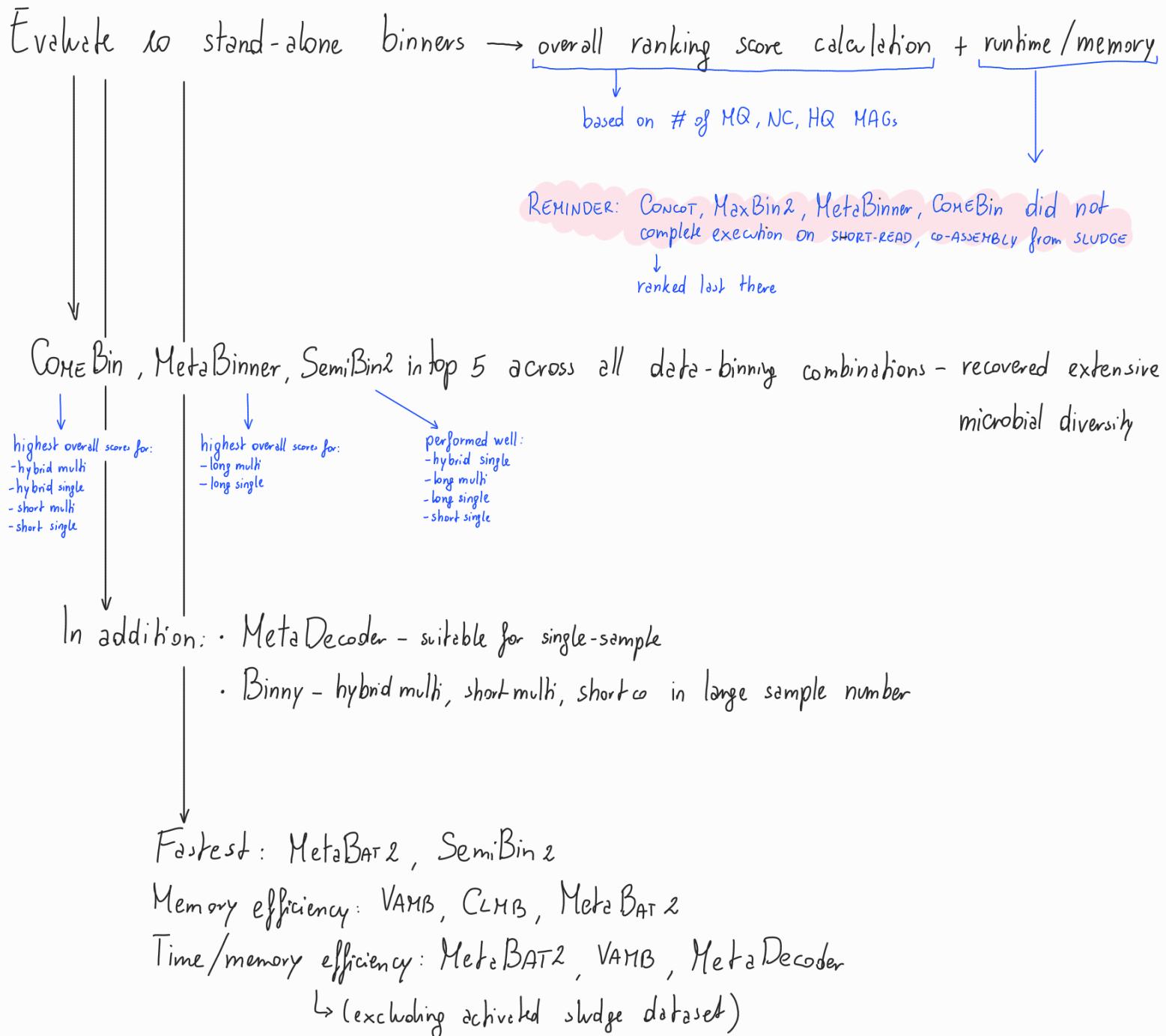
- comparable (cheese, human gut I);
- 50% more MQ, 55% more NC, 57% more HQ (marine);
- better (human gut II, sludge)
- conclusion: requirement of larger number of samples to demonstrate improvement

HYBRID DATA: - multi-sample compared to single sample:

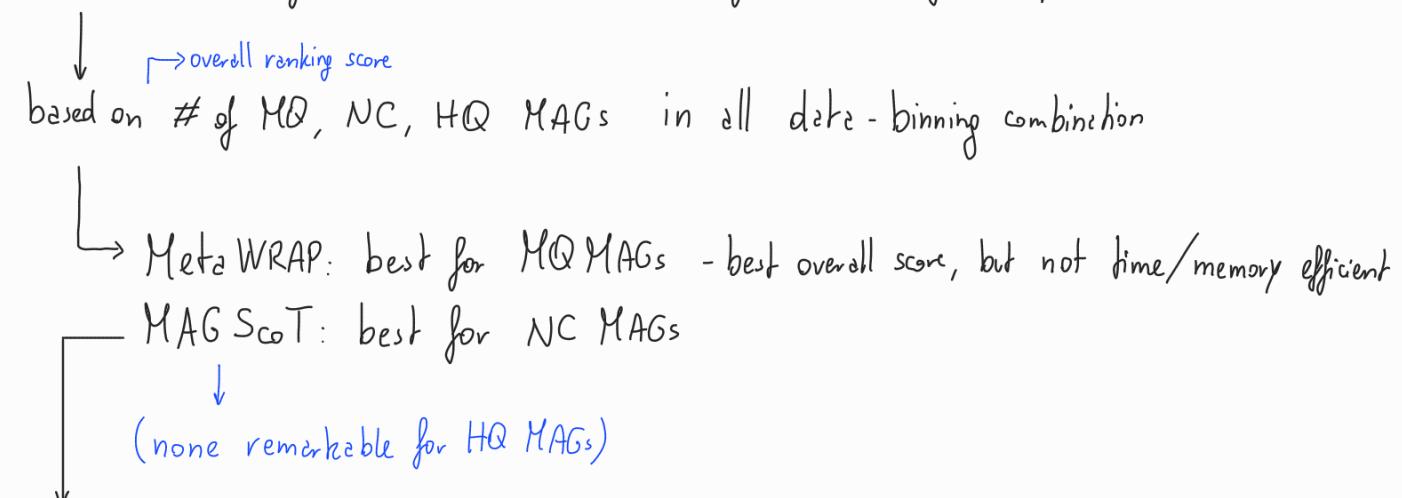
- slight outperformance (human gut I)
- 46% more MQ, 70% more NC, 60% more HQ (marine)
- better (cheese, human gut II)
- CONCOT, Binny, MaxBin2, MetaBinner, COMEBin ; excellent performance (sludge)

Conclusion: multi-sample outperformed single sample binning

- short-read data exhibited poor performance in retrieving HQ MAGs.
- hybrid data (short+long reads) demonstrated best overall performance in both single and multi-sample binning



Evaluate 3 bin-refinement tools - used to refine MAGs from top-three binners



compared to MetaBAT2, MAGSciT showed improvements in # MAGs retrieved

NC and HQ MAGs dereplication on marine set



51% for NC and 71% for HQ MAGs more strains in short multi compared to short single
↳ similar improvement regarding multisample in long and hybrid data

-
- HQ MAGs from hybrid data contained the highest number of species and strains
 - coassembly exhibited lowest diversity of species and strains
 - multi sample demonstrated better diversity

↓
The top performing binning tool recovered more species compared to the the 2nd and 3rd ranked

Antibiotic Resistance Genes (ARGs) → potential hosts identified: NC strains containing at least one ARG, length of config associated with it $\geq 10\text{kb}$

↓
multi-sample outperformed single-sample and the identified hosts exhibited greater number of multi-resistance features

BGCs for secondary metabolites (discovery and development)



Again multi-sample out performed → novelty assessed by comparison with BigFAM database

To assess the quality of MAGs - CheckM2

5 DATASETS: → both long and short read , diverse environments , public

↳ MARINE: 30 mNGS + 30 PacBio

CHEESE: 15 mNGS + 15 PacBio

HUMAN GUT I: 3 mNGS + 3 PacBio

HUMAN GUT II: 30 mNGS + 30 NANOPORE

ACTIVATED SLUDGE: 23 mNGS + 23 NANOPORE

→ FASTQC , MultQC (mNGS , PacBio)
NANOPLT (NANOPORE)

PREPROCESSING

mNGS preprocessing : Fastp -q 20 --length-required 100 --low-complexity-filter

NANOPORE trimming: qcadapt --trim --detect middle (adaptor + barcode)

processing: Filterlong --min-length 4000 --min-mean-q80 (eliminate low quality and short reads)

residual adaptors / barcodes: Porechop --min-split-read-size 4000

Filterlong again using above options

For human gut datasets were mapped on hg38 using BOWTIE2

ASSEMBLY AND ALIGNMENT

short-read co-assembly and single-sample assembly : MEGAHIT



Bowtie2

long-read single-sample : Flye



minimap2

hybrid single-sample : OPERA-MS → short-read sample and corresponding long-read assembled together



Pilon (if long-read is Nanopore) → select contigs longer 1kb.

if mNGS + PacBio: short and long read alignments used for binning

SAM tools - sort alignment files

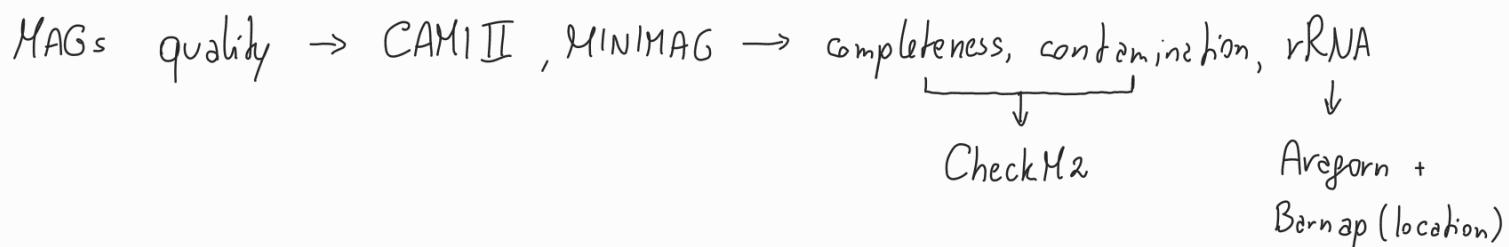
CONTIG BINNING:

- single-sample : · independent assembly for each sequencing sample.
 - coverage profiles calculation from a single sample
- multi-sample : · independent assembly for each sequencing sample.
 - coverage profiles calculation from all samples
- co-assembly : · all sequencing samples are initially assembled together
 - coverage profiles derived from all samples

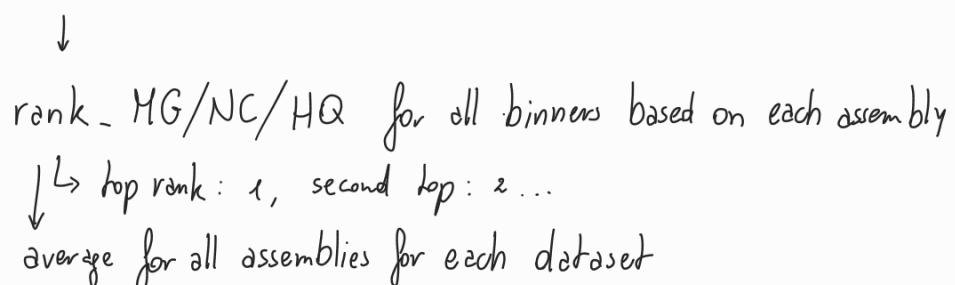
COMBOS:

- hybrid-multi
- hybrid-single
- long-multi
- long-single
- short-multi
- short-single
- short-co

EVALUATION AND RANKING



Ranking score: performance across datasets based on MG, NC, HQ :



$$\text{ranking-score-dataset} = \frac{1}{3} \left(\frac{1}{n} \sum \text{rank_MG}_i + \frac{1}{n} \sum \text{rank_NC}_i + \frac{1}{n} \sum \text{rank_HQ}_i \right)$$

#samples in dataset

$$\text{overall ranking score} = \frac{1}{N} \sum \text{ranking-score-dataset}$$

↑
average ranking scores for all datasets

DEREPLICATION

↳ dRep -p 32 -nc 0.6 -sa 0.95 or -sa 0.99
|
| species ↓ strain
| 95% ani 88% ANI

→ annotation: Genome Taxonomy Database

phylogenetic trees: IQ-TREE -m LG+RG , visualization iTOL

ARGs: RGI (default parameters), rely on CARD (antibiotic database)

↳ host of ARGs: non-redundant NC MAGs at strain level containing ≥ 1 ARG length ARG associated config $\geq 10\text{kb}$

BGCs: annotation: antiSMASH

novelty: BiG-SLICE (query mode, comparison against BiG-FAM)
(database, threshold 800)

COMPUTATIONAL RESOURCES:

↳ runtime and memory on activated sludge