BENCHMARKING METAGENOMICS BINNING TOOLS ON REAL DATASETS ACROSS SEQUENCING PLATFORMS AND BINNING MODELS.

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

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), tryies to group contigs coming from the same organism
↳ GENOMIC SIGNATURES: GC content, k-mer frequencies, codon usage

from assembled contigs 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 contig features, PCA, Gaussian Mixture Models for clustering contigs
  EM algorithm ↑ data as mixture of Gaussian distributions ↑
- MaxBin2: estimates likelyhood of a given contig belonging to a particular genome using tetranucleotides frequencies and coverage. This probability is used in an Expectation-Maximization algorithm to recover genomes from metagenomics
  ↳ calculate probability of belonging to each cluster, update parameters, look for best fits.
- MetaBAT2: pairwise similarities between contigs using tetranucleotide frequency and coverage, then clusters through Label Propagation Algorithm (LPA) → each contig begins with a label that propagate in the similarity cluster. The label is updated based on similarity weights and statistical confidence

- Binny: k-mer composition and coverage for dimentionality reduction, then hierarchical density-based spatial clustering (HDBSCAN) ↳ 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
  ↳ combine models to improve guess
- VAMB: deep variational autoencoders (VAE) to encode tetranucleotide frequency and coverage for each contig. The latent representation is processed with a clustering algorithm.
- CLMB: introduces simulated noise for each contig and uses contrastive learning for contig 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 contig, contrastive learning for high quality embeddings, Leiden-based method for clustering
- MetaWRAP, DAS Tool, MAGScoT: bin refinement tools combining multiple binning tools to reconstruct high quality MAGs

Metagenomic binning:

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

DATA-BINNING COMBO → TOP 3 PERFORMING TOOLS FOR EACH COMBO → MAG DEREPLICATION

5 data sets (mNGS, HiFi, Nanopore)

best-quality genome kept
remove redundant MAGs

each binning combination may recover overlapping or similar MAGs → could overestimate species or strain diversity

MAGs: • completeness >50%, contamination <10% => MQ MAGs (medium quality)
• completeness >90%, contamination <5% => NC MAGs ← near complete
• NC MAGs + 23S, 16S, 5S rRNA genes + ≥18 tRNAs => HQ MAGs (high quality)

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 comparible NC, HQ MAGs (human gut I);
  - 66% more MQ, 82% more NC, 233% more HQ (human gut II);
  - 100% more MQ, 156% 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, 66% more HQ (marine)
  - better (cheese, human gut II)
  - CONCOT, Binny, MaxBin 2, MetaBinner, COMEBin: excellent performance (sludge)

Conclusion: - multisample 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 10 stand-alone binners → overall ranking score calculation + runtime/memory

↓ based on # of MQ, NC, HQ MAGs

↓ REMINDER: CONCOT, MaxBin2, MetaBinner, COMEBin did not complete execution on SHORT-READ, CO-ASSEMBLY from SLUDGE

↓ ranked last there

COMEBin, MetaBinner, SemiBin2 in top 5 across all data-binning combinations - recovered extensive microbial diversity

↓ highest overall scores for:
- hybrid multi
- hybrid single
- short multi
- short single

↓ highest overall scores for:
- long multi
- long single

↘ performed well:
- hybrid single
- long multi
- long single
- short single

In addition:
- MetaDecoder - suitable for single-sample
- Binny - hybrid multi, short multi, short co in large sample number

Fastest: MetaBAT2, SemiBin2

Memory efficiency: VAMB, CLMB, MetaBAT2

Time/memory efficiency: MetaBAT2, VAMB, MetaDecoder

↳ (excluding activated sludge dataset)

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

↓ based on # of MQ, NC, HQ MAGs in all data-binning combination → overall ranking score

↳ MetaWRAP: best for MQ MAGs - best overall score, but not time/memory efficient

MAGScoT: best for NC MAGs

↓ (none remarkable for HQ MAGs)

compared to MetaBAT2, MAGScoT 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 contig associated with it ≥ 10kb

↓

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 data base

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, MultiQC (mNGS, PacBio)
NanoPlot (NANOPORE)

PREPROCESSING

mNGS preprocessing: Fastp -q 20 --lenght_required 100 --low_complexity_filter
NANOPORE trimming: qcat --trim --detect middle (adaptor + barcode)
processing: Filtlong --min_lenght 6000 --min_mean_q80 (eliminate low quality and short reads)
residual adapters / barcodes: Porechop --min_split_read_size 6000
Filtlong 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
↓
Bowtie 2

long-read single-sample: Flye
↓
minimap 2

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

MAGs quality → CAMI II, MINIMAG → completeness, contamination, rRNA

completeness, contamination → CheckM2

rRNA → Aragorn + Barnap (location)

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

↓

rank_MG/NC/HQ for all binners based on each assembly

↓ ↳ top rank: 1, second top: 2 ...

average for all assemblies for each dataset

$$\text{ranking\_score\_dataset} = \frac{1}{3}\left(\frac{1}{n}\sum \text{rank\_MG}_i + \frac{1}{n}\sum \text{rank\_NC}_i + \frac{1}{n}\sum \text{rank\_HQ}_i\right)$$

n ↑ # samples in dataset

$$\text{overall ranking score} = \frac{1}{N}\sum \text{ranking\_score\_dataset}$$

N ↑ average ranking scores for all datasets

DEREPLICATION

↳ dRep -p 32 -nc 0.6 -sa 0.95 or -sa 0.99

-sa 0.95 → species 95% ani

-sa 0.99 → strain 99% ANI

↳ annotation: Genome Taxonomy Database

phylogenetic trees: IQ-TREE -m LG+R4 , visualization iTOL

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

↳ host of ARGs: non-redundant NC MAGs at strain level containing ≥ 1 ARG

lenght ARG associated contig ≥ 10kb

BGCs: annotation: antiSMASH

novelty: BIG-SLICE (query mode, comparison against BIG-FAM database, threshold 900

COMPUTATIONAL RESOURCES:

↳ runtime and memory on activated sludge