

# Metagenomic Binning Pipelines - the State of the Art

## Outline

- *Abstract*
- *Background/Introduction*
  - *Binning problem definition (recover biological entities from metagenomic sequencing)*
  - *problem relevance (Explosion in metagenomics, reduction in sequencing cost, increased computer capacity)*
  - *Review objectives (Brief summary on popular tools, innovations overview of recent tools)*
- *Popular/Previous Binning software*
  - *Proposed solutions (bin contigs into bins(MAG if good quality) based on their kmer composition and abundance/coabundance)*
  - *Tools available (Cite recent benchmark)*
- *Overview of recent metagenomic binning tools*
  - *Innovations in binning tools*
    - \* *Innovations in proposed solutions/ strategy innovations* Read binning gene-abundance binning (CAG, MGS, MSPi) Integrate new experimental data
    - \* *Software/algorithms innovations* machine-learning/deep-learning implementation
  - *Innovations in specific biological questions* Viral genomes and viral strains; Endosymbionts
- *Choosing a binning algorithm*
  - *Identify start point variables*

- 23           \* *Sample origin (Host contamination, diversity)*
- 24           \* *Number of samples (some tools require many samples to perform well)*
- 25           \* *Sequencing technology (Most tools employ illumina, LongReads are increasing)*
- 26           \* *Computational resources available*
- 27        – *Identify endpoint*
- 28           \* *organism of interest viral(ref viral catalogue), bacteria, all*
- 29        – *Tools are complementary MSP/Metabat*
- 30        • *Conclusions*
- 31           – *Current limitations and future directions* Do not perform well on multiple strains, on
- 32           the same sample

33       Figure. Binning software historical citations barplot Figure. Decision tree, overview of  
 34       metagenomic binning Table. List of binning software since 2017

## 35 Abstract

36 New generations of sequencing platforms coupled with numerous bioinformatics tools have led to  
 37 rapid technological progress in metagenomics to investigate complex microorganism communities.  
 38 Nevertheless, a combination of different bioinformatic tools remains necessary to draw conclusions  
 39 out of microbiota studies. As sequencing costs have dropped at a rate above 'Moore's law', a  
 40 greater number of large data sets are being produced than ever before. Newer algorithms that  
 41 take advantage of the size of these datasets are continually being developed. Binning algorithms  
 42 are defined as the grouping of assembled metagenomic contigs by their genome of origin. Selecting  
 43 the most appropriate binning algorithm can be a daunting task and is influenced by many factors.  
 44 This review serves as a guide to direct the researcher to the binning algorithm that best suits their  
 45 needs.

## 46 Background

47 The explosion in popularity and success in the field of metagenomics over the last 25 years can  
 48 be largely attributed to the advances in computing technologies. An example of the outcomes of  
 49 this can be found in the Human Microbiome Project; a project that has been greatly improved

the understanding of the microbial flora involved in human health and disease. These advances have brought with them greater demands for storage, CPU time, and consequently more efficient algorithms. The main function of binning tools is to reconstruct species/biological entities from metagenomic samples. Compared to amplicon, shotgun metagenome can provide functional gene profiles directly and reach a much higher resolution of taxonomic annotation. However, due to the high demands on computational resources, cost, and expertise necessary to perform this analysis, researchers have historically been limited in their capacity to collect and analyse sequencing data. As the cost of sequencing is rapidly falling, this burden has been significantly lessened. At the time of writing, shotgun metagenomic sequencing costs on average three times as much as 16S sequencing in comparison. The objectives of this review is for the reader to be better informed about the latest algorithms (since 2017) for binning metagenomic samples. The second part of this review is for the reader to be informed about distinguishing factors between the methods. The last part is for the reader to make an informed decision based on those factors for their needs. This review will be broken down into the following sections:

## **Recent methods for metagenomic binning**

A metagenomic sample is comprised of many organisms and the standard procedure is to retrieve the sequences from the mixture of organisms. The final goal of binning is to reconstruct the sequences from each organism present in the original sample. Among the binning tools developed in recent years we can distinguish a subset dedicated to cluster reads (read-binning) (MetaBBC-LR, BioBloom Tools, CLAME, LVQ-KKN, Meta VW, HirBin, MEGAN-LR). The main purpose of read-binning tools is to pre-process reads into clusters for a posterior targeted assembly, here we find reference-free and non-reference-free tools, and tools designed for short-read or long-read sequencing technologies. The majority of binning tools we can find are oriented toward clustering contigs (contig-binning) into bins, which may represent the genome from a single biological entity/organism. Contig-binning tools normally rely on coverage information and sequence composition. Progress in contig-binning algorithms can be seen in the proposals to integrate new sources of information (for example, from scaffold-graphs(Binnacle), paired-end reads(COCACOLA), or 3D contact information(MetaTOR)) and state of the art algorithms in machine learning (CoCoNet, Variational Autoencoders for Metagenomic Binning (VAMB)).

## 79 Metagenome Assembled Genomes

80 A Metagenome-Assembled Genome (MAG) is a single-taxon assembly based on one or more binned  
81 metagenomes that has been asserted to be a close representation to an actual individual genome  
82 (that could match an already existing isolate or represent a novel isolate).

## 83 Binning microbial genomes with deep learning

84 The integration of deep learning techniques into the field of metagenomics has revolutionised  
85 the field of metagenomics. The VAMB pipeline was developed to take advantage of variational  
86 autoencoders; a generative machine learning model that uses a deep variational autoencoders  
87 (Nissen et al., n.d.)... COCONET (Arisdakessian, Nigro, Steward, Poisson, & Belcaid, 2021)...

## 88 Binning for viral genomes

89 2021 viral catalog (Nayfach et al., 2021)... New insights from uncultivated genomes of the global  
90 human gut microbiome (Nayfach, Shi, Seshadri, Pollard, & Kyrpides, 2019)... Also mention  
91 coconet suitability for viral genomes...

## 92 Choosing the most appropriate binning algorithm (Classifi- 93 cation by output)

94 A review on the benchmarking binning algorithms was done by Yue et al., 2020. Resource man-  
95 agement is an important factor in the choice of binning algorithm. The tradeoff between number  
96 of Central Processing Units (CPUs), memory, and time are important considerations. Newer ad-  
97 vances in pipeline technologies have ameliorated these costs. An analysis pipeline is defined as  
98 a program that combines several programs in a defined order to complete a complex analysis.  
99 Improperly developed, validated, and/or monitored pipelines may generate inaccurate results.

## 100 MSPs, binning co-abundant genes

101 Binning of co-abundant genes represents an alternative proposal to reconstruct species/biological  
102 entities from a set of metagenomic samples. Co-abundant gene binning methods assume each gene  
103 coming from a shared chromosome will display proportional abundances across samples, if you have  
104 enough samples from a common environment you can identify the sets of genes from a common

Table 1: Comparison of binning algorithms

Software/Algorithm	Year	Description/purpose	Comment/Highlight	Doi	PubmedID
CoNet	2021	Deep learning tool for Viral Metagenome Binning	Reconstructs viral genomes	10.1093/bioinformatics/btab213	33822891
Biatic	2021	Using scaffolds to improve Metagenomic bin quality	Incorporates scaffold information	10.3389/fmicb.2021.685614	33717083
VAMB	2021	Metagenomic binning and MAG assembly	Autocoder algorithm, fast processing	10.1186/s13057-020-00777-4	33398153
phyloFlash	2020	mRNA profiling and MAG assembly	Incorporates asRNA profiling info into MAG as...	10.1093/bioinformatics/btaz441	33109753
hyBRCC-LR	2020	Metagenomic binning for Long-Reads	Suitable for Long Reads sequencing technology	10.1093/bioinformatics/btaz441	32657364
BinBam Tools	2020	Refined binning of metagenomic contigs using as...	Data preparation for targeted assembly, using s...	10.1093/bioinformatics/btaz441	32641514
MetaBin	2020	Metagenomic binning using assembly graphs	Incorporates assembly graphs	10.1093/bioinformatics/btaz441	32167328
MetaSPSim	2020	Simulating metagenomic stable isotope probing d...	Augment binning resolution with extra experimen...	10.1186/s12859-020-3372-6	32000876
MetaCon	2019	Unsupervised binning k-mers and coverage, focus...	Augment binning resolution with extra experimen...	10.1186/s12859-019-2904-4	31757198
VireBin	2019	Binning viral haplotypes from assembled contigs	Viral haplotypes MAGs	10.1186/s12859-019-3138-1	31634576
MAGO (*only tool pipeline)	2019	Framework for Proton and analysis of MAGs	Identification of endosymbiont	10.1093/bioinformatics/btaz441	31633780
SeqDox	2019	Genome separation of Endosymbionts from mixed s...	Incorporates 2D contact information	10.3389/fgene.2019.00853	31608107
MetaTOR	2019	High quality MAGs from mammalian gits using met...	Eliminates misassembly, post binning from previou...	10.7717/peerj.7359	31481973
MetaBAT 2	2019	Adaptive binning algorithm for genome recons...	Employs sample X mappings of mapped read counts	10.1093/bioinformatics/btaz441	31388474
MetaBMF	2019	Scalable binning algorithm for large-scale meta...	Haplotypes for polyploid genomes	10.1186/s12859-019-5828-5	31299888
PolyCRACKER	2019	Method for partitioning polyploid genomes b...	NaN	10.1093/bioinformatics/btaz441	30977806
SolidBin	2019	Improving metagenomic binning with individual...	Handles eukaryotic contamination	10.1093/bioinformatics/btaz441	30838416
Autmeta	2019	Signal processing method for alignat free met...	Alternative description of sequences designed f...	10.1038/s41598-018-38197-9	30770850
MLBP MrGBP (Algorithm)	2019	Alignat based algorithm for alignat free met...	Alignment based for reads	10.1186/s12859-018-5191-y	30537931
CLAME	2018	Fuzzy binning of metagenomic sequence fragmen...	Horizontal gene transfer and regions of uncerta...	10.1109/EMBC.2018.8512529	30447633
3D BH SNE (Algorithm)	2018	Fuzzy binning of metagenomic sequence fragmen...	Classify into DNA or RNA sequence	10.1016/j.virus.2018.10.002	30291874
LVQ-KNN	2018	Abundance based reconstitution of microbial pa...	Paq genome reconstitution	10.1093/bioinformatics/btaz441	30252023
MSPinner	2018	Flexible pipeline for genome resolved metageno...	Hybrid bin extraction algorithm	10.1186/s40168-018-0541-1	30219103
MetaWRAP*	2018	Large scale Machine Learning Sequence classific...	Machine learning for reads based on Kmer profile	10.1007/978-1-4939-8561-6_2	30030800
BM3C	2018	Metagenomic binning through low density binning	Improvement at higher taxonomic levels, discove...	10.1093/bioinformatics/btaz441	30010790
BM3C	2018	Binning contigs using codon usage sequence comp...	Add codon usage information	10.1093/bioinformatics/btaz441	29947757
Assessment of Metagenome Binners	2018	Derreplication aggregation and scoring strategy	NaN	10.1093/bioinformatics/btaz441	29893851
DAS Tool	2018	Long Read/ contigs taxonomic binning	Combines several binning algorithm results	10.1038/s41564-018-0171-1	29807988
MEGAN-LR	2018	Binning workflow using contain coverage and com...	Alignment of long reads against reference sequences	10.1186/s13062-018-0208-7	29678199
CoMet	2017	Metagenomic binning and association of plasmids...	Single sample, include gc content and 4mer fre...	10.1186/s12859-017-1967-3	29297295
?	2017	Reference-free learning with multiple metagenom...	Plasmid binning at strain level using methylati...	10.1038/nbt.4037	29227468
MetaGen	2017	Improved formula for calculate oligonucleotide...	Requires multiple samples	10.1186/s13059-017-1323-y	28974263
BinBam add onn	2017	Boosttrapped supervises binning and annotation	Math formula to calculate oligo sequence dissim...	10.1186/s12859-017-1835-1	28931373
BusyBee Web	2017	Interactive visualisation tool for verification...	2d interactive scatterplots supervised binning	10.1093/nar/akx348	28472498
ICoVer	2017	High resolution identification of differential...	Interactive visualisation tool	10.1186/s12859-017-1653-5"	28464793
BinSanity	2017	Unsupervised clustering using coverage and affi...	Supervised annotation, unsupervised clustering ...	10.1186/s12859-017-3686-6	28431529
BinSanity	2017	Unsupervised clustering using coverage and affi...	Reduce bias for high/low abundance	10.7717/peerj.3035	28289564
IFCM add on	2016	Improved binning using Fuzzy C-Means Method	Combination of different binning algorithms	10.1093/bioinformatics/btaz441	28186226
COCACOLA	2016	Binning contigs using composition, read coverage...	Add estimated distribution of real genome lengths	10.1109/TCBB.2016.2576452	27295684
CoCoM (2)	2014	Tool for automatic recovery of population genom...	Adds paired end read and coaligner information	10.1093/bioinformatics/btaz441	27256312
			Adds differential coverage to complement compos...	10.7717/peerj.603	25289188

organism of origin (MLGs Chameleon-clust 2012, CAGs and MGSs Canopy 2014, Markovclust-  
MGCs Karlsson 2013, MSPs MSPminner 2018). To the extent of our knowledge, in the past few  
years MSPminer is the only available Software exploiting this approach. MSPminer introduced a  
robust proportionality measure detecting co abundant but no necessarily co-occurring. MSPminer  
introduced a robust proportionality measure detecting co abundant but no necessarily co-occurring.  
This tools groups co-abundant genes into Metagenomic Species Pan-genomes or Metagenomic  
Species Pan-genomess (MSPs) and classify genes within an MSP as core, accessory and shared.  
The factors that impact directly on MSP quality include the sample composition, the sequencing  
depth, the previous bioinformatic steps to build the reference gene dataset and to map the reads.  
A high number of samples with varying phenotypes improve the quality of MSPs. MSPs can  
be employed for taxonomic profiles of new samples from similar ecosystems, to compare strains  
between samples building a presence/absence table of accessory genes and for biomarker discovery.  
By binning contigs carrying genes from the same MSP it is also possible to build a MAG. Co-  
abundant gene binning methods perform better in large sample datasets.

## **Metagenomic Species Pan-genomes**

Microbial pan-genomes are gene repertoires composed of core genes present in all strains and  
accessory genes present in only some of them (Medini et al., 2005). In a shotgun metagenomic  
sequencing context, we define as shared the genes detected in some samples where the species is  
not present. A strain found in a sample is an instance of the species pan-genome: it is made of all  
the species (shared) core genes and of a subset of (shared) accessory genes. Core genes are suitable  
for taxonomic profiling at species-level while accessory genes can be used to compare strains across  
samples. Genes tagged as shared should be used carefully as they contain false positives counts  
or are subject to horizontal transfer. Core genes are suitable for taxonomic profiling at species-  
level while accessory genes can be used to compare strains across samples. Genes tagged as shared  
should be used carefully as they contain false positives counts or are subject to horizontal transfer.

## **Weaknesses and future developments**

Until now binning methods perform poorly in samples that contain similar strains...

## Conclusion

New and open areas of research in which the application of metagenomic pipelines are relevant  
The increased impact of machine learning in analysis Short section - just for past-present-future  
completeness Future developments for metagenomic analysis

## References

- Arisdakessian, C. G., Nigro, O. D., Steward, G. F., Poisson, G., & Belcaid, M. (2021). Coconet: an efficient deep learning tool for viral metagenome binning. *Bioinformatics*.
- Nayfach, S., Páez-Espino, D., Call, L., Low, S. J., Sberro, H., Ivanova, N. N., ... others (2021). Metagenomic compendium of 189,680 dna viruses from the human gut microbiome. *Nature Microbiology*, 1–11.
- Nayfach, S., Shi, Z. J., Seshadri, R., Pollard, K. S., & Kyrpides, N. C. (2019). New insights from uncultivated genomes of the global human gut microbiome. *Nature*, 568(7753), 505–510.
- Nissen, J. N., Johansen, J., Allesøe, R. L., Sønderby, C. K., Armenteros, J. J. A., Grønbech, C. H., ... others (n.d.). Improved metagenome binning and assembly using deep variational autoencoders. *Nature Biotechnology*, 1–6.
- Yue, Y., Huang, H., Qi, Z., Dou, H.-M., Liu, X.-Y., Han, T.-F., ... Tu, J. (2020). Evaluating metagenomics tools for genome binning with real metagenomic datasets and caml datasets. *BMC bioinformatics*, 21(1), 1–15.