

1 Metagenomic Binning Pipelines - the State of the Art

2

3 **Contents**

4	1 Abstract	1
5	2 Background	2
6	3 Overview of recent methods for metagenomic binning	3
7	3.1 Progress in recent binning strategies	3
8	3.1.1 Binning co-abundant genes	4
9	3.1.2 Binning microbial genomes with deep learning	4
10	3.2 Binning of viral genomes	5
11	3.3 Binning Pipelines	5
12	4 Suggestions on choosing a binning algorithm	6
13	5 Conclusion	8
14	References	8

15 **1 Abstract**

16 New generations of sequencing platforms coupled with numerous bioinformatics tools have led to
17 rapid technological progress in metagenomics to investigate complex microorganism communities.
18 Nevertheless, a combination of different bioinformatic tools remains necessary to draw conclusions
19 out of microbiota studies. As sequencing costs have dropped at a rate above 'Moore's law', a
20 greater number of large data sets are being produced than ever before. Newer algorithms that
21 take advantage of the size of these datasets are continually being developed. Binning algorithms
22 are defined as the grouping of assembled metagenomic contigs by their genome of origin (Figure 1).

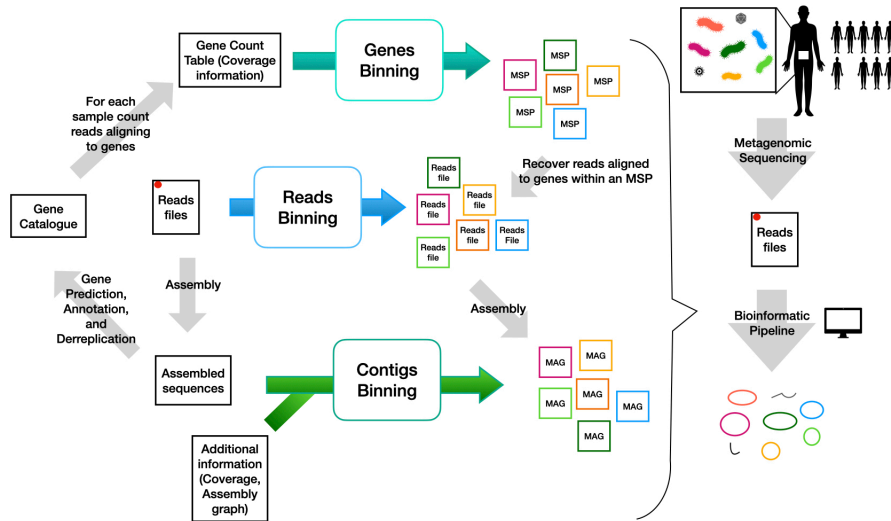


Figure 1: Summary of binning principles and techniques.

Selecting the most appropriate binning algorithm can be a daunting task and is influenced by many factors. This review serves as a guide to direct the researcher to the binning algorithm that best suits their needs.

2 Background

The explosion in popularity and success in the field of metagenomics over the last 25 years can be largely attributed to the advances in computing technologies. An example of the outcomes of 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 (Turnbaugh et al., 2007). 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. Whole Genome Shotgun sequences does not require cultivation. At the time of writing, shotgun metagenomic sequencing costs on average three times as much as 16S sequencing in comparison. Here we will briefly recapitulate recent binning algorithms and highlight

41 some of the developments in the field, among them, the use of new algorithms and strategies
42 employed to achieve the goal of identifying the organisms composing microbiome communities.
43 We hope this overview could aid the reader to choose a binning algorithm or a combination of
44 them based on their specific needs.

45 **3 Overview of recent methods for metagenomic binning**

46 **3.1 Progress in recent binning strategies**

47 A metagenomic sample is comprised of many organisms and the goal of binning is to reconstruct
48 the sequences from each organism present in the original sample. The majority of binning tools
49 are oriented toward clustering contigs (contig-binning) into bins, which may represent the genome
50 from a single biological entity/organism. A Metagenome-Assembled Genome (MAG) is a single-
51 taxon assembly based on one or more binned metagenomes that has been asserted to be a close
52 representation to an individual genome that could match an already existing isolate or represent
53 a novel isolate. Current contig-binning tools are commonly reference free (i.e. they do not de-
54 pend on reference sequences to perform clustering) and rely on coverage information and sequence
55 composition. Progress in contig-binning algorithms can be seen in the proposals to integrate
56 new sources of information (for example, from scaffold-graphs (Binnacle), paired-end reads (CO-
57 CACOLA), or 3D contact information (MetaTOR)) and state of the art algorithms in machine
58 learning (CoCoNet, VAMB). We also notice the development of Bin refinement tools (DAS-tool,
59 Binning Refiner) that rely on the outputs from multiple contig-binning algorithms and combine
60 them to produce better results (Sieber et al., 2018). Binning of contigs have played a central role
61 in software development in the field, a review on the benchmarking of binning algorithms was done
62 by Yue et al., 2020. Beside contig-binning tools we can also distinguish read-binning tools and
63 co-abundant-gene-binning tools. The main purpose of read-binning tools is to pre-process reads
64 into clusters for a posterior targeted assembly. Here we find reference-free and non-reference-free
65 tools, and tools designed for short-read or long-read sequencing technologies. Among the binning
66 tools developed in recent years, a subset of them are dedicated to cluster reads (read-binning)
67 (MetaBBC-LR, BioBloom Tools, CLAME, LVQ-KKN, Meta VW, HirBin, MEGAN-LR).

3.1.1 Binning co-abundant genes

Binning of co-abundant genes represents an alternative proposal to reconstruct species/biological entities from a set of metagenomic samples. Co-abundant gene binning methods assumes that each gene coming from a shared chromosome will display proportional abundances across samples. Therefore, if there are enough samples from a similar environment you can identify the sets of genes from a common organism of origin (MLGs Chameleon-clust 2012, CAGs and MGSs Canopy 2014, Markovclust-MGCs Karlsson 2013, MSPs MSPminner 2018) (Karypis, Han, & Kumar, 1999; Plaza Oñate et al., 2019). The MSPminer software was developed to exploit this approach. MSPminer introduced a robust proportionality measure to detect co-abundance but no necessarily co-occurrence. This tool groups co-abundant genes into Metagenomic Species Pan-genomess (MSPs) and classifies genes within an MSP as core, accessory and shared. Core genes are present in all strains, accessory are present only in some, and the shared category applies for those genes which may be present in more than one MSP due to horizontal transfer (Tettelin et al., 2005). Factors that impact directly on MSP quality include sample composition, sequencing depth, and previous bioinformatic steps to build the reference gene dataset and map the reads. MSPs can be used for taxonomic profiling of new samples from similar ecosystems at the species level, and also to compare strains between samples by 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.

3.1.2 Binning microbial genomes with deep learning

The integration of deep learning techniques has revolutionised the field of metagenomics. Deep learning approaches have benefitted from the rapid acceleration in GPU efficiency over the past few years. The Software VAMB and CoCoNet constitute two such examples that employ deep learning for binning (Nissen et al., n.d.; Arisdakessian, Nigro, Steward, Poisson, & Belcaid, 2021).

The main novelty of VAMB is the application of the Deep Learning technique known as Variational Auto Encoders (VAE). In this case, variational autoencoders learn how to integrate two data types, coabundance and k-mer composition. The resulting latent representation is able to cluster better than either of the inputs alone. In principle this technology is not limited by only two input data types. VAMB also applies a "multisplit" approach whereby each cluster should correspond to an organism representation across samples and each bin in a cluster to a per-sample representation of the genome of that organism.

99 The CoCoNet software uses deep learning and clustering to bin contigs into clusters repre-
 100 senting species present in the samples. The algorithm consists of two phases. During the first
 101 phase, a neural network is trained to estimate the probability that two contigs arise from the same
 102 genome given their composition and coverage information. The second use a heuristic to bin the
 103 contigs using the probabilities inferred in the first stage. An interesting feature in CoCoNet is it
 104 was trained on viral genomes. In the following section we discuss more about binning on viral
 105 genomes.

106 3.2 Binning of viral genomes

107 Most binning algorithms are designed for prokaryotic organisms leaving viruses out of the soft-
 108 ware scope. In recent years the virome and its importance in health and disease has recognised.
 109 CoCoNet uses deep learning to model co-occurrence of contigs from the same viral genome. The
 110 network was optimized for diverse viral metagenomes, the network learns to model coverage vari-
 111 ability within samples, a critical feature in viral metagenomes where DNA amplification methods
 112 are needed to increase input genetic material. VirBin clusters contigs for genome reconstruction of
 113 viral strains, different strains within viral species may show different biological properties such as
 114 transmissibility or virulence. Composition based features are usually not enough to separate
 115 haplotypes, VirBin receives contigs as inputs and outputs the estimated number of haplotypes via
 116 contig alignment and returns the contigs for each haplotype based on relative abundance distribu-
 117 tion, when the contigs are long enough VirBin produce better results. Newer strategies have been
 118 proposed and employed to reconstruct viral genomes from metagenomic samples, in a recent work
 119 (Nayfach et al., 2021) a new compendium of 189680 DNA viruses from the human gut microbiome
 120 was produced. In this work they use viral informative features including presence of viral protein
 121 families (Paez-Espino et al., 2016), absence of non-viral families (El-Gebali et al., 2019), gene
 122 strand switch rate (Roux 2015), and the score produced from the VirFinder (Ren, Ahlgren, Lu,
 123 Fuhrman, & Sun, 2017) software.

124 3.3 Binning Pipelines

125 Other advances in binning can be found in the integration of existing tools and software into
 126 bioinformatic pipelines. These innovations allow the automatic complete processing of read sam-
 127 ples into bins or the addition of extra processing steps to address specific biological questions or
 128 problems related to the sample of origin. MetaWRAP is a modular pipeline ready to perform

common tasks in metagenomic analysis, starting from read quality checks up to bin creation, refinement, reassembly quantification, taxonomic annotation and functional annotation. MAGO pipeline integrates metagenome assembly, binning, bin improvement, bin quality check, bin functional annotation, and bin taxonomic annotation. SqueezeMeta also integrates external software to perform the complete analysis of metagenomic samples from sequences reading to MAG construction and annotation. `nf-mag` supports both short and long reads, performs quality and adapter trimming, quality check, performs assembly, binning, checks bin quality and assigns taxonomy (Ewels et al., 2020).

Autometa was developed to deal with non-model Eukariotic host contamination and complex single metagenomes, the application integrate sequence homology, nucleotide composition, coverage and single-copy marker genes to separate microbial genomes from non model host genomes. Seqdex is a tool written in R which separates endosymbionts from their host sequences. Their approach uses specific features in endosymbiotic systems to better solve this problem. This tool combines partial taxonomic annotations obtained through homology searches and sequence composition to predict the contig's organism of origin from host and its endosymbionts and helps the user to understand how effective is the classification. Reproducibility, scalability, and ease of use from people with little computational experience are attractive features that pipelines for metagenomic analysis provide.

4 Suggestions on choosing a binning algorithm

A number of aspects should be considered when performing binning analysis on metagenomic samples. Computational resources available, sequencing technology, number of samples, and the sample's source are important factors to consider. Some tools employ more resources than others, and some perform better under specific circumstances (as reviewed by Yue et al., 2020). If you are dealing with a large number of samples, a gene-binning strategy could be taken into consideration. Tools such as CoMet were built around single sample binning. Long read sequence technology is gaining momentum and some tools also integrate the characteristic features generated with this technology. The environment under study also play an important role for binning. Sometimes there exists host organisms whose genome sequences would be removed before starting the analysis. The environment also has a profound effect on the sample's diversity with samples that have greater diversity requiring greater sequencing depth making binning more difficult. It is also difficult for binning tools to discern between similar strains within the same sample. It is also worth mentioning

Table 1: Comparison of binning algorithms

Software/Algorithm	Year	Description/purpose	Comment/Highlight	Doi	PubmedID
CoCoNet	2021	Deep learning tool for Viral Metagenome Binning	Reconstructs viral genomes	10.1093/bioinformatics/btab213	33822891
Binnacle	2021	Using scaffolds to improve Metagenomic bin qua...	Incorporates scaffold information	10.3389/fmicb.2021.638561	33717033
VAMB	2021	Metagenome binning using deep variational auto...	Autoencoder algorithm, fast processing	10.1038/s41587-020-00777-4	33398153
phyloFlash	2020	ssrRNA profiling and MAG assembly	Incorporates ssrRNA profiling info into MAG as...	10.1128/mSystems.00920-20	33109753
MetaBCC-LR	2020	Metagenomic binning for Long-Reads	Suitable for Long Reads sequencing technology	10.1093/bioinformatics/btaa441	32657364
BioBloom Tools	2020	Reads binning for targeted assembly, alignment...	Data preparation for targeted assembly, using ...	10.1073/pnas.1903436117	32641514
GraphBin	2020	Refined binning of metagenomic contigs using a...	Incorporates assembly graphs information	10.1093/bioinformatics/btaa180	32167528
MetaSPSim	2020	Simulating metagenomic stable isotope probing ...	Augment binning resolution with extra experime...	10.1186/s12859-020-3372-6	32000676
MetaCon	2019	Unsupervised binning k-mers and coverage, focu...	Focus different lengths contigs	10.1186/s12859-019-2904-4	31757198
VirBin	2019	Binning viral haplotypes from assembled contigs	Viral haplotypes MAGs	10.1186/s12859-019-3138-1	31684876
MAGO (*only tool pipeline)	2019	Framework for Production and analysis of MAGs	Identification of endosymbiont	10.1093/molbev/msz237	31633780
SeqDex	2019	Genome separation of Endosymbionts from mixed ...	Incorporates 3D contact information	10.3389/fgene.2019.00853	31608107
MetaTOR	2019	High quality MAGs from mammalian guts using me...	Eliminates manual parameter tuning from previo...	10.3389/fgene.2019.00753	31481973
MetaBAT (v2)	2019	Adaptive binning algorithm for genome recon...	Employs sample X contigs cf mapped read counts	10.7717/peerj.7359	31388474
MetaBMF	2019	Scalable binning algorithm for large scale met...	Haplotypes for polyploid genomes	10.1093/bioinformatics/btz577	31347687
PolyCRACKER	2019	Method for partitioning polyploid sub genomes ...	NaN	10.1186/s12864-019-5828-5	31299888
SolidBin	2019	Improving metagenome binning with semi-supervi...	NaN	10.1093/bioinformatics/btz253	30977806
Autometa	2019	Improvement of microbial genomes from individua...	Handles eukaryotic contamination	10.1093/nar/gkz148	30838416
MLBP MrGBP (Algorithm)	2019	Signal processing method for alignment free me...	Alternative description of sequences designed ...	10.1038/s41598-018-38197-9	30770850
CLAME	2018	Alignement based algorithm allowed description o...	Alignement based for reads	10.1186/s12864-018-5191-y	30537931
3D BH SNE (Algorithm)	2018	Fuzzy binning of metagenomic sequence fragments	Horizontal gene transfer and regions of uncert...	10.1109/EMBC.2018.8512529	30440633
LVQ-KNN	2018	Composition based RNA or DNA binning of short ...	Classify into DNA or RNA sequence	10.1016/j.virusres.2018.10.002	30291874
MSPminer	2018	Abundance based reconstitution of microbial pa...	Pan genome reconstitution	10.1093/bioinformatics/bty830	30252023
MetaWRAP*	2018	Flexible pipeline for genome resolved metageno...	Hybrid bin extraction algorithm	10.1186/s40168-018-0541-1	30219103
MetaVW	2018	Large scale Machine Learning Sequence classifi...	Machine learning for reads based on Khmer profile	10.1007/978-1-4939-8561-6_2	30030800
Opal (algorithm*)	2018	Metagenomic binning through low density binning	Improvement at higher taxonomic levels, discov...	10.1093/bioinformatics/bty611	30010790
BMC3C	2018	Binning contigs using codon usage sequence com...	Add codon usage information	10.1093/bioinformatics/bty519	29947757
AMBER tool	2018	Assessment of Metagenome Bimmers	NaN	10.1093/gigascience/giy069	29893851
DAS Tool	2018	Derreplication aggregation and scoring strategy	Combines several binning algorithm results	10.1038/s41564-018-0171-1	29807988
MEGAN-LR	2018	Long Read/ contigs taxonomic binning	Alignement of long reads against reference seque...	10.1186/s13062-018-0208-7	29678199
CoMet	2017	Binning workflow using contain coverage and co...	Single sample, include gc content and 4mer fr...	10.1186/s12859-017-1967-3	29297295
MetaGen	2017	reference-free learning with multiple metageno...	Requires multiple samples	10.1186/s13059-017-1323-y	28974263
d2sBin add onn	2017	Improved formula for calculate oligonucleotide...	Math formula to calculate oligo sequence dissi...	10.1186/s12859-017-1835-1	28931373
BusyBee Web	2017	Bootstrapped supervises binning and annotation	2d interactive scatterplots supervised binning	10.1093/nar/gkx348	28472498
ICoVer	2017	High resolution identification tool for verificatio...	Interactive visualisation tool	10.1186/s12859-017-1653-5	28464793
HirBin*	2017	Unsupervised clustering tool for differential...	Supervised annotation, unsupervised clustering...	10.1186/s12864-017-3686-6	28431529
BinSanity	2017	Improve genome bins through the combination of...	Reduce bias for high/low abundance	10.7717/peerj.3035	28289564
Binning_refinner	2017	Improve genome bins through the combination of...	Combination of different binning algorithms	10.1093/bioinformatics/btx086	28186226
IFCM add on	2016	Improved binning using Fuzzy C-Means Method	Add estimated distribution of real genome lengths	10.1109/TCBB.2016.2576452	27295684
COCACOLA	2016	binning contigs using composition, read covera...	Adds paired end read and coalignment information	10.1093/bioinformatics/btw290	27256312
GroopM (v2)	2014	Tool for automatic recovery of population geno...	Adds differential coverage to complement compo...	10.7717/peerj.603	25289188

that most of the available tools are not mutually exclusive and it is possible to benefit from the advantages each one has to offer and merge results depending of the aim of the study. Besides binning, other types of metagenomic analysis can be performed on microbiomes. Recent reviews provide an overview of the complete process and practical guides to apply available software.

5 Conclusion

Until now binning methods perform poorly in samples that contain similar strains. Also do not perform great assigning 16S sequences to bins maybe due to high copy number of these sequences within a genome. Binning has been focused mainly in prokariotic organisms. Binning of organisms outside prokariotes need more development, lately some advances have been observed in viral genomes (cite viral catalogue and viral binning organims) but the huge diversity in viral genomes still poses a challenge for current methodologies. Eukariotic microscopic organisms does not appear in the current picture.

The continuously increasing number of sequences available require more efficient/faster algorithms and new strategies to reconstruct single organisms from environmental samples. New sources of experimental information might add up into solving the binning central problem. Development of Machine learning algorithms have started in the field and we expect to see more development soon

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*.
- El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., ... others (2019). The pfam protein families database in 2019. *Nucleic acids research*, 47(D1), D427–D432.
- Ewels, P. A., Peltzer, A., Fillinger, S., Patel, H., Alneberg, J., Wilm, A., ... Nahnsen, S. (2020). The nf-core framework for community-curated bioinformatics pipelines. *Nature biotechnology*, 38(3), 276–278.
- Karypis, G., Han, E.-H., & Kumar, V. (1999). Chameleon: Hierarchical clustering using dynamic modeling. *Computer*, 32(8), 68–75.

187 Nayfach, S., Páez-Espino, D., Call, L., Low, S. J., Sberro, H., Ivanova, N. N., ... others (2021).
188 Metagenomic compendium of 189,680 dna viruses from the human gut microbiome. *Nature*
189 *Microbiology*, 1–11.

190 Nissen, J. N., Johansen, J., Allesøe, R. L., Sønderby, C. K., Armenteros, J. J. A., Grønbech,
191 C. H., ... others (n.d.). Improved metagenome binning and assembly using deep variational
192 autoencoders. *Nature Biotechnology*, 1–6.

193 Paez-Espino, D., Eloë-Fadrosh, E. A., Pavlopoulos, G. A., Thomas, A. D., Huntemann, M.,
194 Mikhailova, N., ... Kyrpides, N. C. (2016). Uncovering earth’s virome. *Nature*, 536(7617),
195 425–430.

196 Plaza Oñate, F., Le Chatelier, E., Almeida, M., Cervino, A. C., Gauthier, F., Magoulès, F., ...
197 Pichaud, M. (2019). Mspminer: abundance-based reconstitution of microbial pan-genomes
198 from shotgun metagenomic data. *Bioinformatics*, 35(9), 1544–1552.

199 Ren, J., Ahlgren, N. A., Lu, Y. Y., Fuhrman, J. A., & Sun, F. (2017). Virfinder: a novel k-mer
200 based tool for identifying viral sequences from assembled metagenomic data. *Microbiome*,
201 5(1), 1–20.

202 Sieber, C. M., Probst, A. J., Sharrar, A., Thomas, B. C., Hess, M., Tringe, S. G., & Banfield,
203 J. F. (2018). Recovery of genomes from metagenomes via a dereplication, aggregation and
204 scoring strategy. *Nature microbiology*, 3(7), 836–843.

205 Tettelin, H., Massignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., ... others
206 (2005). Genome analysis of multiple pathogenic isolates of streptococcus agalactiae: impli-
207 cations for the microbial “pan-genome”. *Proceedings of the National Academy of Sciences*,
208 102(39), 13950–13955.

209 Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., & Gordon, J. I.
210 (2007). The human microbiome project. *Nature*, 449(7164), 804–810.

211 Yue, Y., Huang, H., Qi, Z., Dou, H.-M., Liu, X.-Y., Han, T.-F., ... Tu, J. (2020). Evaluating
212 metagenomics tools for genome binning with real metagenomic datasets and caml datasets.
213 *BMC bioinformatics*, 21(1), 1–15.