Advances in Metagenomic Binning for the Reconstruction of Microbial Species

Jose Fernando Garcia Guevara[1*], Theo Portlock[1*], Adil Mardinoglu[1,2], Mathias Uhlén[1], and Saeed Shoaie[1,2]

[1]Science for Life Laboratory, Royal Institute of Technology (KTH), Stockholm, Sweden.

[2]Centre for Host Microbiome Interactions, Faculty of Dentistry, Oral & Craniofacial Sciences, King's College London, London, UK.

Contents

Glossary

CPU Central Processing Unit.

HMP Human Microbiome Project.

MAG Metagenome-Assembled Genome.

MSP Metagenomic Species Pan-genomes.

VAE Variational Auto Encoders.

1 Abstract

New generations of sequencing platforms coupled with numerous bioinformatics tools have led to rapid technological progress in metagenomics to investigate complex microorganism communities. Nevertheless, a combination of different bioinformatic tools remains necessary to draw conclusions out of microbiota studies. As sequencing costs have dropped at a rate above 'Moore's law', a greater number of large data sets are being produced than ever before. Newer algorithms that take advantage of the size of these datasets are continually being developed. Binning algorithms are defined as the grouping of assembled metagenomic contigs by their genome of origin (Figure 1). Selecting the most appropriate binning algorithm can be a daunting task and is influenced by many factors. This review details the current advances in the field of metagenomic binning.

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 (HMP); a project that has been greatly improved the understanding of the microbiome flora involved in human health and disease (Turnbaugh et al., 2007). These advances have brought with them greater demands for storage, Central Processing Unit (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

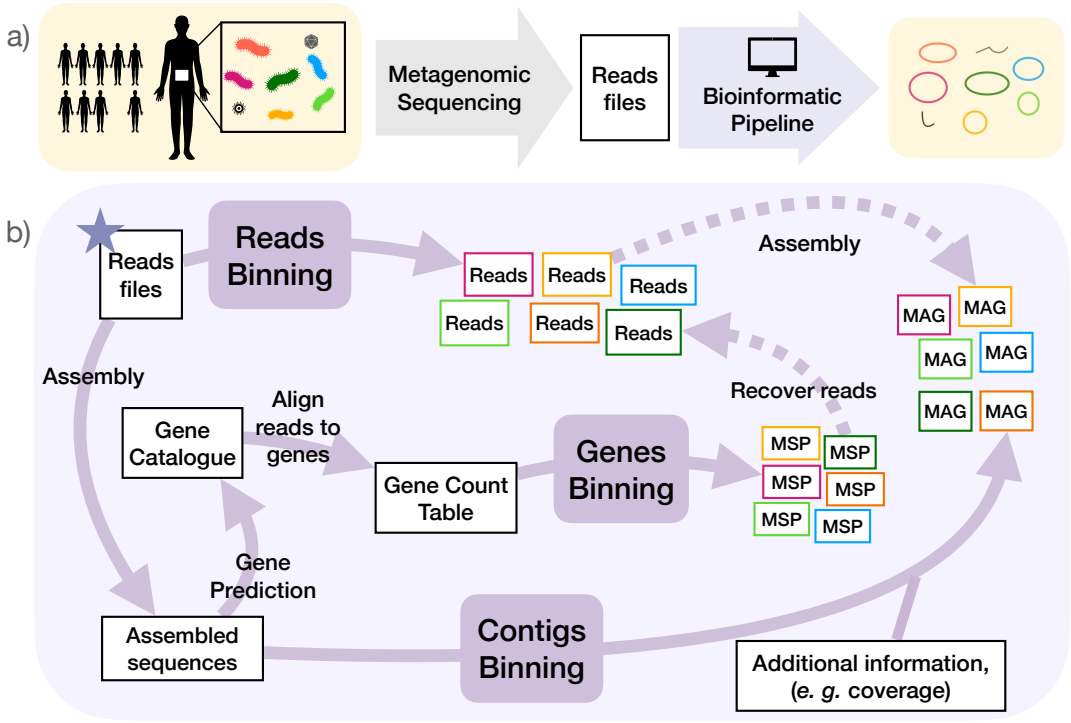

Figure 1: Overview of the existing binning strategies. a) Microbiome samples are collected and sequenced, the resulting sequences are processed, employing the existing binning strategies, which try to reconstruct the genome sequences from the original organisms. b) The scheme present a simplified workflow of the existing binning strategies, named contig-binning, reads-binning and genes-binning, represented as rounded rectangles in the middle of arrows. The star in the upper left indicates the entry point, simple arrows represent intermediate processing steps, black edge squares represent intermediate files, the color squares represent endpoint files. The binning strategies are not independent and can complement each other as shown by the dotted arrows.

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. Accurate, robust, and suitable binning methods are crucial for the proper interpretation of any metagenomic dataset. Here we will briefly recapitulate recent binning algorithms and highlight some of the developments in the field, among them, the use of new algorithms and strategies employed to achieve the goal of identifying the organisms composing microbiome communities.

3 Overview of recent methods for metagenomic binning

3.1 Progress in recent binning strategies

A metagenomic sample is comprised of many organisms and the goal of binning is to reconstruct the sequences from each organism present in the original sample. The majority of binning tools are oriented toward clustering contigs (contig-binning) into bins, which may represent the genome from a single biological entity/organism. A Metagenome-Assembled Genome (MAG) is a single-taxon assembly based on one or more binned metagenomes that has been asserted to be a close representation to an individual genome that could match an already existing isolate or represent a novel isolate. Current contig-binning tools are commonly reference free (i.e. they do not depend on reference sequences to perform clustering) and rely on coverage information and sequence composition (listed in Table 1). Progress in contig-binning algorithms can be seen in the proposals to integrate new sources of information (for example, from scaffold-graphs (Binnacle (Muralidharan, Shah, Meisel, & Pop, 2021)), paired-end reads (COCACOLA) (Lu, Chen, Fuhrman, & Sun, 2016), or 3D contact information (MetaTOR) (Baudry, Foutel-Rodier, Thierry, Koszul, & Marbouty, 2019)) and state of the art algorithms in machine learning (CoCoNet, VAMB) (Nissen et al., n.d.; Arisdakessian, Nigro, Steward, Poisson, & Belcaid, 2021). We also notice the development of Bin refinement tools (DAS-tool, Binning Refiner) that rely on the outputs from multiple contig-binning algorithms and combine them to produce better results (Song & Thomas, 2017; Sieber et al., 2018), and tools which allow visual inspection of bins (Broeksema et al., 2017; Laczny et al., 2017). Binning of contigs have played a central role in software development in the field, a review on the benchmarking of binning algorithms was done by Yue et al., 2020. Beside contig-binning tools we can also distinguish read-binning tools and co-abundant-gene-binning tools. The main purpose of read-binning tools is to pre-process reads into clusters for a posterior targeted assem-

bly. Here we find reference-free and non-reference-free tools, and tools designed for short-read or long-read sequencing technologies. Among the binning tools developed in recent years, a subset of them are dedicated to cluster reads (read-binning) (MetaBBC-LR, BioBloom Tools, CLAME, LVQ-KNN, Meta VW, HirBin, MEGAN-LR) (Wickramarachchi, Mallawaarachchi, Rajan, & Lin, 2020; Chu et al., 2014; Benavides, Isaza, Niño-García, Alzate, & Cabarcas, 2018; Belka, Fischer, Pohlmann, Beer, & Höper, 2018; Vervier, Mahé, Tournoud, Veyrieras, & Vert, 2016; Österlund, Jonsson, & Kristiansson, 2017; Huson et al., 2018).

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, Metagenomic Species Pan-genomess (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-occurrance. This tool groups co-abundant genes into 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 bioinforamtic 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 et al., 2021).

The main novelty of VAMB is the application of the Deep Learning technique known as Variational Auto Encoders (VAE). In this case, VAEs 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 "mulitsplit" 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 orgamisn.

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

3.2 Binning of viral genomes

Most binning algorithms are designed for prokariotic organisms leaving viruses out of the software scope. In recent years the virome and its importance in health and disease has recognised. CoCoNet uses deep leaning to model co-ocurrence of contigs from the same viral genome. The network was optimized for diverse viral metagenomes, the network learns to model coverage variability within samples, a critical feature in viral metagenomes where DNA amplification methods are needed to increase input genetic material. VirBin clusters contigs for genome reconstruction of viral strains, different strains within viral species may show different biological properties such as transmissibility or virulence. Composition based features are usually are not enough to separate haplotypes, VirBin receives contigs as inputs and outputs the estimated number of haplotypes via contig aligment and returns the contigs for each haplotype based on relative abundance distribution, when the contigs are long enough VirBin produce better results. Newer strategies has been proposed and employed to reconstruct viral genomes from metagenomic samples, in a recent work (Nayfach et al., 2021) a new compendium of 189680 DNA viruses from the human gut microbiome was produced. In this work they use viral informative features including presence of viral protein families (Paez-Espino et al., 2016), absence of non-viral families (El-Gebali et al., 2019), gene

strand switch rate (Roux, Beloin, & Ghigo, 2005), and the score produced from the VirFinder (Ren, Ahlgren, Lu, Fuhrman, & Sun, 2017) software.

3.3 Binning Pipelines

Other advances in binning can be found in the integration of existing tools and software into bioinformatic pipelines. These innovations allow the automatic complete processing of read samples into bins or the addition of extra processing steps to address specific biological questions or 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 (Tamames & Puente-Sánchez, 2019) 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 (Miller et al., 2019). Seqdex is a tool written in R which separates endosymbionts from their host sequences (Chiodi et al., 2019). Their approach uses specific features in endosymbiotic systems to better solve this problem. This tool combines partial taxonomic annotations obtained trough 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 Challenges in appropriate binning algorithm selection

A number of aspects should be considered when performing binning analysis on metagenomic samples. A list of binning algorithms, including some that we have not mentioned above, are listed here Table 1. Computational resources available, sequencing technology, number of samples, and

Table 1: Comparison of popular binning algorithms updated since 2017.

Software/Algorithm	Year	Description/purpose	Comment/Highlight	Doi
CoCoNet	2021	Deep learning tool for Viral Metagenome Binning	Reconstuction of viral genomes	10.1093/bioinformatics/btab213
Binnacle	2021	Using scaffolds to improve Metagenomic bin quality	Incorporates scaffold information	10.3389/fmicb.2021.638561
VAMB	2021	Metagenome binning using deep variational autoencoders	Autoencoder algorithm, fast processing	10.1038/s41587-020-00777-4
phyloFlash	2020	ssrRNA profiling and MAG assembly	Incorporates ssrRNA profiling information into MAG assembly	10.1128/mSystems.00920-20
MetaBCC-LR	2020	Metagenomic binning for Long-Reads	Suitable for long reads sequencing technology	10.1093/bioinformatics/btaa441
BioBloom Tools	2020	Binning of reads for alignment free targeted assembly	Data preparation for targeted assembly using space seeds	10.1073/pnas.1903436117
GraphBin	2020	Refined binning of metagenomic contigs using assembly graphs	Incorporates assembly graph information	10.1093/bioinformatics/btaa180
MetaCon	2019	Unsupervised binning k-mers and coverage focussing on contigs of different lenghts	Focusses different contig lengths	10.1186/s12859-019-2904-4
VirBin	2019	Binning viral haplotypes from assembled contigs	Viral haplotypes MAGs	10.1186/s12859-019-3138-1
MAGO	2019	Framework for Production and analysis of MAGs	Integrated pipeline	10.1093/molbev/msz237
SeqDex	2019	Genome separation of Endosymbionts from mixed sequencing samples	Identification of endosymbiont	10.3389/fgene.2019.00853
SqueezeMeta	2019	Pipeline covering the complete metagenomic/metatranscriptomic analysis	Internal checks that provides information about consistency of contigs and bins	10.3389/fmicb.2018.03349
MetaTOR	2019	High quality MAG binning from mammalian guts using meta3C libraries	Incorporates 3D contact information	10.3389/fgene.2019.00753
MetaBAT (v2)	2019	Adatptative binning algorithm for genome reconstruction from metagenome assemblies	Eliminates manual parameter tuning from previous version	10.7717/peerj.7359
MetaBMF	2019	Scalable binning algorithm for species and strain level large scale metagenomic studies	Applicable to large scale studies	10.1093/bioinformatics/btz577
SolidBin	2019	Improving metagenome binning with semi-supervised normalized cut	Improved clustering performance	10.1093/bioinformatics/btz253
Autometa	2019	Extraction of microbial genomes from individual shotgun metagenomes	Handles eukaryotic contamination	10.1093/nar/gkz148
CLAME	2018	Aligment based algorithm allowed description of novel species	Aligment based read binning	10.1186/s12864-018-5191-y
LVQ-KNN	2018	Composition based RNA or DNA binning of short sequences	Classification into DNA or RNA sequence	10.1016/j.virusres.2018.10.002
MSPminer	2018	Abundance based reconstitution of microbial pan-genomes from metagenomic data	Pan genome reconstitution	10.1093/bioinformatics/bty830
MetaWRAP	2018	Flexible pipeline for genome resolved metagenomic data analysis	Hybrid bin extraction algorithm	10.1186/s40168-018-0541-1
Meta VW	2018	Large scale Machine Learning Sequence classification	Machine learning for reads based on Kmer profile	10.1007/978-1-4939-8561-6_2
BMC3C	2018	Binning contigs using codon usage sequence composition and read coverage	Adds codon usage information	10.1093/bioinformatics/bty519
DAS Tool	2018	Dereplication aggregation and scoring strategy	Combines several binning algorithm results	10.1038/s41564-018-0171-1
MEGAN-LR	2018	Long read/contig taxonomic binning	Aligment of long reads against reference sequences	10.1186/s13062-018-0208-7
CoMet	2018	Binning workflow using coverage and composition	Single sample, includes GC content and 4-mer frequency	10.1186/s12859-017-1967-3
MetaGen	2017	Reference-free learning with multiple metagenomic samples	Requires multiple samples	10.1186/s13059-017-1323-y
BusyBee Web	2017	Bootstrapped supervised binning and annotation	Supervised binning	10.1093/nar/gkx348
ICoVer	2017	Interactive visualisation tool for verification and refinement of metagenomic bins	Interactive visualisation tool	10.1186/s12859-017-1653-5
BinSanity	2017	Unsupervised clustering using coverage and affinity propagation	Reduce bias for high/low abundance	10.7717/peerj.3035
Binning refiner	2017	Improved genome bins through the combination of different binning algorithms	Combination of different binning algorithms	10.1093/bioinformatics/btx086
COCACOLA	2016	Binning contigs using composition, read coverage, co-aligment and paired end read linkage	Adds paired end read and coaligment information	10.1093/bioinformatics/btw290
GroopM (v2:2017)	2014	Tool for automatic recovery of population genomes from related metagenomes	Adds differential coverage to complement composition based binning	10.7717/peerj.603

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 tool like MetaBMF (Xing, Liu, & Zhong, 2017; Ma, Xiao, & Xing, 2019) or a gene-binning strategy could be taken into consideration. Tools such as CoMet were built around single sample binning (Herath, Tang, Tandon, Ackland, & Halgamuge, 2017). 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 that there is no mutual exclusivity between the currently available tools and it is possible to benefit from the relative advantages each has to offer and merge the 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 (Breitwieser, Lu, & Salzberg, 2019).

5 Conclusion

Popularity and successes of metagenomic binning have accelerated in the last ten years. Current limitations that still remain include the difficulty in classifying similar strains within samples. They additionally do not perform well assigning 16S sequences to bins likely due to the high copy number of these sequences within a genome. As binning has been focused mainly in prokariotic organisms, binning of organisms outside prokariotes need more development. Although there have been signifiance advances in the characterisation of viral genomes as of late (Nayfach et al., 2021), the huge diversity in viral genomes still poses a challenge for current methodologies. The continuously increasing number of sequences available require more efficient/faster algorithms and new strategies to reconstruct single organisms from environmental samples. However, with the breakneck pace of technological advancements in computing resources, this requirement is sure to be met and will pave the way for greater insights into the microbial world. With the integration of Machine learning algorithms into binning, we expect to see significant developments in the near future.

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*.

Baudry, L., Foutel-Rodier, T., Thierry, A., Koszul, R., & Marbouty, M. (2019). Metator: A computational pipeline to recover high-quality metagenomic bins from mammalian gut proximity-ligation (meta3c) libraries. *Frontiers in Genetics*, *10*, 753. Retrieved from `https://www.frontiersin.org/article/10.3389/fgene.2019.00753` doi: 10.3389/fgene.2019.00753

Belka, A., Fischer, M., Pohlmann, A., Beer, M., & Höper, D. (2018). Lvq-knn: Composition-based dna/rna binning of short nucleotide sequences utilizing a prototype-based k-nearest neighbor approach. *Virus research*, *258*, 55–63.

Benavides, A., Isaza, J. P., Niño-García, J. P., Alzate, J. F., & Cabarcas, F. (2018). Clame: a new alignment-based binning algorithm allows the genomic description of a novel xanthomonadaceae from the colombian andes. *BMC genomics*, *19*(8), 9–30.

Breitwieser, F. P., Lu, J., & Salzberg, S. L. (2019). A review of methods and databases for metagenomic classification and assembly. *Briefings in bioinformatics*, *20*(4), 1125–1136.

Broeksema, B., Calusinska, M., McGee, F., Winter, K., Bongiovanni, F., Goux, X., ... Ghoniem, M. (2017, May). Icover – an interactive visualization tool for verification and refinement of metagenomic bins. *BMC Bioinformatics*, *18*(1), 233. Retrieved from `https://doi.org/10.1186/s12859-017-1653-5` doi: 10.1186/s12859-017-1653-5

Chiodi, A., Comandatore, F., Sassera, D., Petroni, G., Bandi, C., & Brilli, M. (2019). Seqdeχ: a sequence deconvolution tool for genome separation of endosymbionts from mixed sequencing samples. *Frontiers in genetics*, *10*, 853.

Chu, J., Sadeghi, S., Raymond, A., Jackman, S. D., Nip, K. M., Mar, R., ... Birol, I. (2014). Biobloom tools: fast, accurate and memory-efficient host species sequence screening using bloom filters. *Bioinformatics*, *30*(23), 3402–3404.

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.

Herath, D., Tang, S.-L., Tandon, K., Ackland, D., & Halgamuge, S. K. (2017). Comet: a workflow using contig coverage and composition for binning a metagenomic sample with high precision.

BMC bioinformatics, *18*(16), 161–172.

Huson, D. H., Albrecht, B., Bağcı, C., Bessarab, I., Gorska, A., Jolic, D., & Williams, R. B. (2018). Megan-lr: new algorithms allow accurate binning and easy interactive exploration of metagenomic long reads and contigs. *Biology direct*, *13*(1), 1–17.

Karypis, G., Han, E.-H., & Kumar, V. (1999). Chameleon: Hierarchical clustering using dynamic modeling. *Computer*, *32*(8), 68–75.

Laczny, C. C., Kiefer, C., Galata, V., Fehlmann, T., Backes, C., & Keller, A. (2017, 05). BusyBee Web: metagenomic data analysis by bootstrapped supervised binning and annotation. *Nucleic Acids Research*, *45*(W1), W171-W179. Retrieved from `https://doi.org/10.1093/nar/gkx348` doi: 10.1093/nar/gkx348

Lu, Y. Y., Chen, T., Fuhrman, J. A., & Sun, F. (2016, 06). COCACOLA: binning metagenomic contigs using sequence COmposition, read CoverAge, CO-alignment and paired-end read LinkAge. *Bioinformatics*, *33*(6), 791-798. Retrieved from `https://doi.org/10.1093/bioinformatics/btw290` doi: 10.1093/bioinformatics/btw290

Ma, T., Xiao, D., & Xing, X. (2019, 07). MetaBMF: a scalable binning algorithm for large-scale reference-free metagenomic studies. *Bioinformatics*, *36*(2), 356-363. Retrieved from `https://doi.org/10.1093/bioinformatics/btz577` doi: 10.1093/bioinformatics/btz577

Miller, I. J., Rees, E. R., Ross, J., Miller, I., Baxa, J., Lopera, J., ... Kwan, J. C. (2019). Autometa: automated extraction of microbial genomes from individual shotgun metagenomes. *Nucleic acids research*, *47*(10), e57–e57.

Muralidharan, H. S., Shah, N., Meisel, J. S., & Pop, M. (2021). Binnacle: Using scaffolds to improve the contiguity and quality of metagenomic bins. *Frontiers in Microbiology*, *12*, 346. Retrieved from `https://www.frontiersin.org/article/10.3389/fmicb.2021.638561` doi: 10.3389/fmicb.2021.638561

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.

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.

Österlund, T., Jonsson, V., & Kristiansson, E. (2017). Hirbin: high-resolution identification of differentially abundant functions in metagenomes. *BMC genomics*, *18*(1), 1–11.

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

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

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

Roux, A., Beloin, C., & Ghigo, J.-M. (2005). Combined inactivation and expression strategy to study gene function under physiological conditions: application to identification of new escherichia coli adhesins. *Journal of bacteriology*, *187*(3), 1001–1013.

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

Song, W.-Z., & Thomas, T. (2017, 02). Binning_refiner: improving genome bins through the combination of different binning programs. *Bioinformatics*, *33*(12), 1873-1875. Retrieved from `https://doi.org/10.1093/bioinformatics/btx086` doi: 10.1093/bioinformatics/btx086

Tamames, J., & Puente-Sánchez, F. (2019). Squeezemeta, a highly portable, fully automatic metagenomic analysis pipeline. *Frontiers in microbiology*, *9*, 3349.

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., ... others (2005). Genome analysis of multiple pathogenic isolates of streptococcus agalactiae: implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences*, *102*(39), 13950–13955.

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

Vervier, K., Mahé, P., Tournoud, M., Veyrieras, J.-B., & Vert, J.-P. (2016). Large-scale machine learning for metagenomics sequence classification. *Bioinformatics*, *32*(7), 1023–1032.

Wickramarachchi, A., Mallawaarachchi, V., Rajan, V., & Lin, Y. (2020). Metabcc-lr: meta genomics b inning by c overage and c omposition for l ong r eads. *Bioinformatics*, *36*(Supplement_1), i3–i11.

Xing, X., Liu, J. S., & Zhong, W. (2017, Oct). Metagen: reference-free learning with multiple metagenomic samples. *Genome Biology*, *18*(1), 187. Retrieved from `https://doi.org/10.1186/s13059-017-1323-y` doi: 10.1186/s13059-017-1323-y

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 cami datasets. *BMC bioinformatics*, *21*(1), 1–15.

Acknowledgements

We thank members of Sysbiomelab and Sysmedicine (Mardinoglu lab) for their invaluable support during this work. This work was funded by KTH Royal Institute of Technology and School of Engineering Sciences in Chemistry, Biotechnology and Health (CBH).