

1 Advances in Metagenomic Binning for the Reconstruction of  
2 Microbial Species

3 Jose Fernando Garcia Guevara<sup>1\*</sup>, Theo Portlock<sup>1\*</sup>, Adil Mardinoglu<sup>1,2</sup>, Mathias  
4 Uhlén<sup>1</sup>, and Saeed Shoaie<sup>1,2</sup>

5 <sup>1</sup>Science for Life Laboratory, Royal Institute of Technology (KTH), Stockholm,  
6 Sweden.

7 <sup>2</sup>Centre for HostMicrobiome Interactions, Faculty of Dentistry, Oral &  
8 Craniofacial Sciences, King's College London, London, UK.

9 **Contents**

10	<b>1 Abstract</b>	<b>2</b>
11	<b>2 Background</b>	<b>2</b>
12	<b>3 Overview of recent methods for metagenomic binning</b>	<b>4</b>
13	3.1 Progress in recent binning strategies . . . . .	4
14	3.1.1 Binning co-abundant genes . . . . .	5
15	3.1.2 Binning microbial genomes with deep learning . . . . .	5
16	3.2 Binning of viral genomes . . . . .	6
17	3.3 Binning Pipelines . . . . .	6
18	<b>4 Challenges in appropriate binning algorithm selection</b>	<b>7</b>
19	<b>5 Conclusion</b>	<b>9</b>
20	<b>References</b>	<b>9</b>

## 21 Glossary

22 **CPU** Central Processing Unit.

23 **HMP** Human Microbiome Project.

24 **MAG** Metagenome-Assembled Genome.

25 **MSP** Metagenomic Species Pan-genomes.

26 **VAE** Variational Auto Encoders.

## 27 1 Abstract

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

## 37 2 Background

38 The explosion in popularity and success in the field of metagenomics over the last 25 years can be  
39 largely attributed to the advances in computing technologies. An example of the outcomes of this  
40 can be found in the Human Microbiome Project (HMP); a project that has been greatly improved  
41 the understanding of the microbiome flora involved in human health and disease (Turnbaugh  
42 et al., 2007). These advances have brought with them greater demands for storage, Central  
43 Processing Unit (CPU) time, and consequently more efficient algorithms. The main function of  
44 binning tools is to reconstruct species/biological entities from metagenomic samples. Compared  
45 to amplicon, shotgun metagenome can provide functional gene profiles directly and reach a much  
46 higher resolution of taxonomic annotation. However, due to the high demands on computational

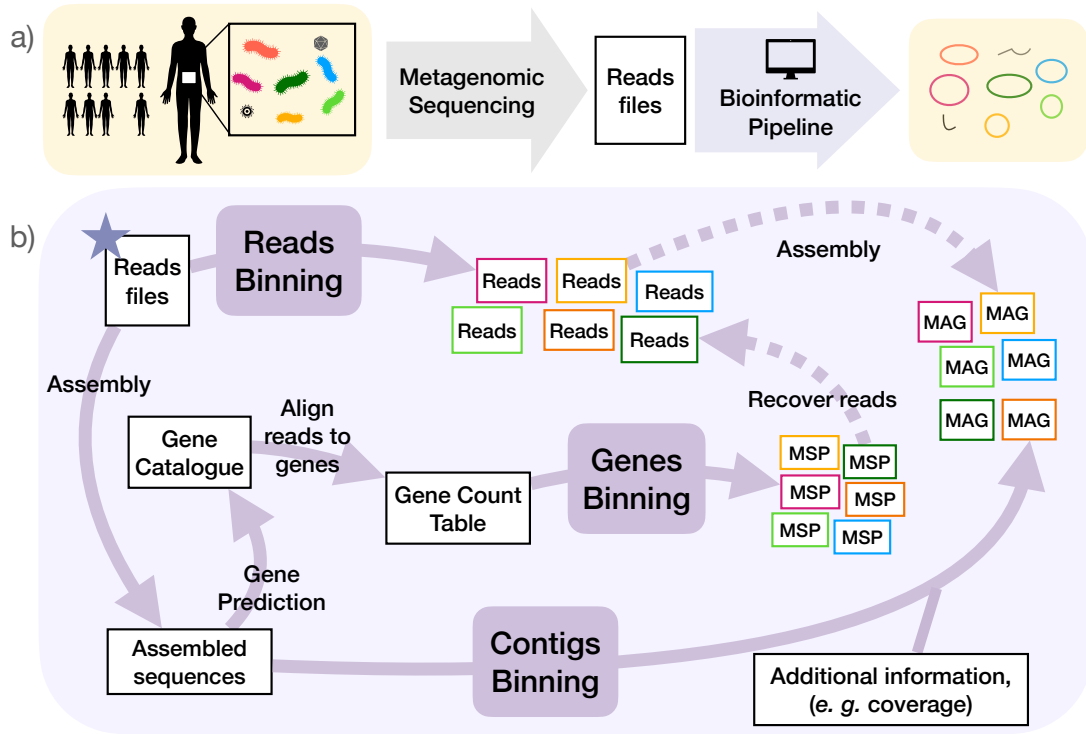


Figure 1: Overview of the existing binning strategies. a) Microbiome samples are collected and sequenced, the resulting sequences are processed, employing the existing binning strategies, which try to reconstruct the genome sequences from the original organisms. b) The scheme present a simplified workflow 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. 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 (CO-CACOLA), or 3D contact information (MetaTOR)) and state of the art algorithms in machine learning (CoCoNet, VAMB). 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 (Sieber et al., 2018). 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 assembly. 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)

77 (Wickramarachchi, Mallawaarachchi, Rajan, & Lin, 2020; Chu et al., 2014; Benavides, Isaza,  
78 Niño-García, Alzate, & Cabarcas, 2018; Belka, Fischer, Pohlmann, Beer, & Höper, 2018; Vervier,  
79 Mahé, Tournoud, Veyrieras, & Vert, 2016; Österlund, Jonsson, & Kristiansson, 2017; Huson et  
80 al., 2018).

### 81 **3.1.1 Binning co-abundant genes**

82 Binning of co-abundant genes represents an alternative proposal to reconstruct species/biological  
83 entities from a set of metagenomic samples. Co-abundant gene binning methods assumes that  
84 each gene coming from a shared chromosome will display proportional abundances across samples.  
85 Therefore, if there are enough samples from a similar environment you can identify the sets of  
86 genes from a common organism of origin (MLGs Chameleon-clust 2012, CAGs and MGSs Canopy  
87 2014, Markovclust-MGCs Karlsson 2013, Metagenomic Species Pan-genomess (MSPs) MSPminner  
88 2018) (Karypis, Han, & Kumar, 1999; Plaza Oñate et al., 2019). The MSPminer software was  
89 developed to exploit this approach. MSPminer introduced a robust proportionality measure to  
90 detect co-abundance but no necessarily co-occurrence. This tool groups co-abundant genes into  
91 MSPs and classifies genes within an MSP as core, accessory and shared. Core genes are present  
92 in all strains, accessory are present only in some, and the shared category applies for those genes  
93 which may be present in more than one MSP due to horizontal transfer (Tettelin et al., 2005).  
94 Factors that impact directly on MSP quality include sample composition, sequencing depth, and  
95 previous bioinformatic steps to build the reference gene dataset and map the reads. MSPs can be  
96 used for taxonomic profiling of new samples from similar ecosystems at the species level, and also  
97 to compare strains between samples by building a presence/absence table of accessory genes and  
98 for biomarker discovery. By binning contigs carrying genes from the same MSP it is also possible  
99 to build a MAG.

### 100 **3.1.2 Binning microbial genomes with deep learning**

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

105 The main novelty of VAMB is the application of the Deep Learning technique known as Vari-  
106 ational Auto Encoders (VAE). In this case, VAEs learn how to integrate two data types, coabun-

107 dance and k-mer composition. The resulting latent representation is able to cluster better than  
108 either of the inputs alone. In principle this technology is not limited by only two input data  
109 types. VAMB also applies a "multisplit" approach whereby each cluster should correspond to an  
110 organism representation across samples and each bin in a cluster to a per-sample representation  
111 of the genome of that organism.

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

## 119 **3.2 Binning of viral genomes**

120 Most binning algorithms are designed for prokaryotic organisms leaving viruses out of the soft-  
121 ware scope. In recent years the virome and its importance in health and disease has recognised.  
122 CoCoNet uses deep learning to model co-occurrence of contigs from the same viral genome. The  
123 network was optimized for diverse viral metagenomes, the network learns to model coverage vari-  
124 ability within samples, a critical feature in viral metagenomes where DNA amplification methods  
125 are needed to increase input genetic material. VirBin clusters contigs for genome reconstruction of  
126 viral strains, different strains within viral species may show different biological properties such as  
127 transmissibility or virulence. Composition based features are usually not enough to separate  
128 haplotypes, VirBin receives contigs as inputs and outputs the estimated number of haplotypes via  
129 contig alignment and returns the contigs for each haplotype based on relative abundance distribu-  
130 tion, when the contigs are long enough VirBin produce better results. Newer strategies have been  
131 proposed and employed to reconstruct viral genomes from metagenomic samples, in a recent work  
132 (Nayfach et al., 2021) a new compendium of 189680 DNA viruses from the human gut microbiome  
133 was produced. In this work they use viral informative features including presence of viral protein  
134 families (Paez-Espino et al., 2016), absence of non-viral families (El-Gebali et al., 2019), gene  
135 strand switch rate (Roux, Beloin, & Ghigo, 2005), and the score produced from the VirFinder  
136 (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). Autometata 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. 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 (Herath, Tang, Tandon, Ackland,

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

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...	Handles eukaryotic contamination	10.1093/bioinformatics/btz253	30977806
Autometa	2019	extraction of microbial genomes from individua...	Alternative description of sequences designed ...	10.1093/nar/gkz148	30838416
MLBP MrGBP (Algorithm)	2019	Signal processing method for alignment free me...	Alignement based for reads	10.1038/s41598-018-38197-9	30770850
CLAME	2018	Alignment based algorithm allowed description o...	Horizontal gene transfer and regions of uncert...	10.1186/s12864-018-5191-y	30537931
3D BH SNE (Algorithm)	2018	Fuzzy binning of metagenomic sequence fragments	Classify into DNA or RNA sequence	10.1109/EMBC.2018.8512529	30440633
LVQ-KNN	2018	Composition based RNA or DNA binning of short ...	Pan genome reconstitution	10.1016/j.virusres.2018.10.002	30291874
MSPminer	2018	Abundance based reconstitution of microbial pa...	Hybrid bin extraction algorithm	10.1093/bioinformatics/bty830	30252023
MetaWRAP*	2018	Flexible pipeline for genome resolved metageno...	Machine learning for reads based on Khmer profile	10.1186/s40168-018-0541-1	30219103
MetaVW	2018	Large scale Machine Learning Sequence classifi...	Improvement at higher taxonomic levels, discov...	10.1007/978-1-4939-8561-6_2	30030800
Opal (algorithm*)	2018	Metagenomic binning through low density binning	Add codon usage information	10.1093/bioinformatics/bty611	30010790
BMC3C	2018	Binning contigs using codon usage sequence com...	NaN	10.1093/bioinformatics/bty519	29947757
AMBER tool	2018	Assessment of Metagenome Binner	Combines several binning algorithm results	10.1093/gigascience/giy069	29893851
DAS Tool	2018	Derreplication aggregation and scoring strategy	Alignement of long reads against reference seque...	10.1038/s41564-018-0171-1	29807988
MEGAN-LR	2018	Long Read/ contigs taxonomic binning	Single sample, include gc content and 4mer fr...	10.1186/s13062-018-0208-7	29678199
CoMet	2017	Binning workflow using contain coverage and co...	Requires multiple samples	10.1186/s12859-017-1967-3	29297295
d2sBin add onn	2017	Improved formula for calculate oligonucleotide...	Math formula to calculate oligo sequence dissi...	10.1186/s13059-017-1323-y	28974263
BusyBee Web	2017	Bootstrapped supervises binning and annotation	2d interactive scatterplots supervised binning	10.1186/s12859-017-1835-1	28931373
ICoVer	2017	High resolution identification tool for verificatio...	Interactive visualisation tool	10.1093/nar/gkx348	28472498
HirBin*	2017	Unsupervised clustering tool for differential...	Supervised annotation, unsupervised clustering...	10.1186/s12859-017-1653-5	28464793
BinSanity	2017	Improve genome bins through the combination of...	Reduce bias for high/low abundance	10.1186/s12864-017-3686-6	28431529
Binning_refinner	2017	Improved binning using Fuzzy C-Means Method	Combination of different binning algorithms	10.1093/bioinformatics/btx086	28186226
IFCM add on	2016	Binning contigs 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



167 & Halgamuge, 2017). Long read sequence technology is gaining momentum and some tools also  
 168 integrate the characteristic features generated with this technology. The environment under study  
 169 also play an important role for binning. Sometimes there exists host organisms whose genome  
 170 sequences would be removed before starting the analysis. The environment also has a profound ef-  
 171 fect on the sample’s diversity with samples that have greater diversity requiring greater sequencing  
 172 depth making binning more difficult. It is also difficult for binning tools to discern between similar  
 173 strains within the same sample. It is also worth mentioning that there is no mutual exclusivity  
 174 between the currently available tools and it is possible to benefit from the relative advantages each  
 175 has to offer and merge the results depending of the aim of the study. Besides binning, other types  
 176 of metagenomic analysis can be performed on microbiomes. Recent reviews provide an overview of  
 177 the complete process and practical guides to apply available software (Breitwieser, Lu, & Salzberg,  
 178 2019).

## 179 5 Conclusion

180 Popularity and successes of metagenomic binning have accelerated in the last ten years. Current  
 181 limitations that still remain include the difficulty in classifying similar strains within samples.  
 182 They additionally do not perform well assigning 16S sequences to bins likely due to the high copy  
 183 number of these sequences within a genome. As binning has been focused mainly in prokariotic  
 184 organisms, binning of organisms outside prokariotes need more development. Although there  
 185 have been signifiante advances in the characterisation of viral genomes as of late (Nayfach et al.,  
 186 2021), the huge diversity in viral genomes still poses a challenge for current methodologies. The  
 187 continuously increasing number of sequences available require more efficient/faster algorithms and  
 188 new strategies to reconstruct single organisms from environmental samples. However, with the  
 189 breakneck pace of technological advancements in computing resources, this requirement is sure to  
 190 be met and will pave the way for greater insights into the microbial world. With the integration  
 191 of Machine learning algorithms into binning, we expect to see significant developments in the near  
 192 future.

## 193 References

194 Arisdakessian, C. G., Nigro, O. D., Steward, G. F., Poisson, G., & Belcaid, M. (2021). Coconet:  
 195 an efficient deep learning tool for viral metagenome binning. *Bioinformatics*.

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

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

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

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

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

210 El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., ... others (2019).  
211 The pfam protein families database in 2019. *Nucleic acids research*, 47(D1), D427–D432.

212 Ewels, P. A., Peltzer, A., Fillinger, S., Patel, H., Alneberg, J., Wilm, A., ... Nahnsen, S. (2020).  
213 The nf-core framework for community-curated bioinformatics pipelines. *Nature biotechnol-  
214 ogy*, 38(3), 276–278.

215 Herath, D., Tang, S.-L., Tandon, K., Ackland, D., & Halgamuge, S. K. (2017). Comet: a workflow  
216 using contig coverage and composition for binning a metagenomic sample with high precision.  
217 *BMC bioinformatics*, 18(16), 161–172.

218 Huson, D. H., Albrecht, B., Bağcı, C., Bessarab, I., Gorska, A., Jolic, D., & Williams, R. B.  
219 (2018). Megan-lr: new algorithms allow accurate binning and easy interactive exploration  
220 of metagenomic long reads and contigs. *Biology direct*, 13(1), 1–17.

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

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

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

228 *Microbiology*, 1–11.

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

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

234 Paez-Espino, D., Eloë-Fadrosch, E. A., Pavlopoulos, G. A., Thomas, A. D., Huntemann, M.,  
235 Mikhailova, N., ... Kyrpides, N. C. (2016). Uncovering earth’s virome. *Nature*, 536(7617),  
236 425–430.

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

240 Ren, J., Ahlgren, N. A., Lu, Y. Y., Fuhrman, J. A., & Sun, F. (2017). Virfinder: a novel k-mer  
241 based tool for identifying viral sequences from assembled metagenomic data. *Microbiome*,  
242 5(1), 1–20.

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

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

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

251 Tettelin, H., Massignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., ... others  
252 (2005). Genome analysis of multiple pathogenic isolates of streptococcus agalactiae: impli-  
253 cations for the microbial “pan-genome”. *Proceedings of the National Academy of Sciences*,  
254 102(39), 13950–13955.

255 Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., & Gordon, J. I.  
256 (2007). The human microbiome project. *Nature*, 449(7164), 804–810.

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

259 Wickramarachchi, A., Mallawaarachchi, V., Rajan, V., & Lin, Y. (2020). Metabcc-lr: meta

260       genomics binning by coverage and composition for long reads. *Bioinformatics*,  
261       36(Supplement\_1), i3–i11.  
262   Yue, Y., Huang, H., Qi, Z., Dou, H.-M., Liu, X.-Y., Han, T.-F., ... Tu, J. (2020). Evaluating  
263       metagenomics tools for genome binning with real metagenomic datasets and caml datasets.  
264       *BMC bioinformatics*, 21(1), 1–15.

## 265   **Acknowledgements**

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