Metagenomic Binning Pipelines - the State of the Art

Theo Portlock

June 28, 2021

1 Abstract

- *Decision tree graphical abstract for the choice of binning algorithm*
- *Features that distinguish binning algorithms*
- *Some guidelines for chosing the correct binning techniques appropriate for a given study*

New generations of sequencing platforms coupled to numerous bioinformatics tools have led to rapid technological progress in metagenomics and metatranscriptomics 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', bigger data sets are available, and proportional costs of analysis have risen as a consequence. Binning is the grouping of assembled metagenomic contigs by their genome of origin. Algorithms for binning are a rapidly evolving field. The number of these algorithms are growing over time. Selecting the most appropriate binning algorithm can be a daunting task and is influenced on computational resources available and experimental variables relating to the sequencing. This review serves as a roadmap to direct the reserarcher to the binning algorithm that best suits their needs.

2 Background

- *General introduction to history of binning*
- *increase in popularity of the field of metagenomics*

2.1 History of binning

Compared to amplicon, shotgun metagenome can provide functional gene profiles directly and reach a much higher resolution of taxonomic annotation. However, due to the large amount of data, the fact that most software is only available for Linux systems, and the large amount of computing resources are needed to perform analysis... After collection, the steps involved in preparing the sequencing data for metagenomic analysis are quality control, filtering, and trimming. Sequence alignment - Bowtie2, Tophat2, Hisat2 are used to map reads against a database Classifying taxonomy and Annotation - Binning Binning pipelines: Kaiju, Kraken2, Braken, mOTU, fetchMG, metaphlan, Centrifuge, METEOR, MetaBAT Galaxy EBI Metagenomics (MGnify) has doubled the number of publicly available anaysed datasets held within the resource in two years. orgnaisms have similar tetranucleotide frequencies put contigs together into genome Spades tools for binning - Concoct, metabat, groupm, and crAss

3 Factors to concider before chosing a binning algorithm

Resource management Tradeoff between number of CPU's, memory, and time are important conciderations. Depends on the resources you have available and the required accuracy. Pipeline vs standalone? Alignement based or alignment free An analysis pipeline is defined as a program that combines several softwaare programs in a defined order to complete a complex analysis. Improperly developed, validated, and/or monitored pipelines may generate inaccurate results.

4 Methods for metagenomic binning

4.1 Metagenome Assembled Genomes

Viral, environmental, gut, long/short reading, computational,lab resourses etc - deep coverage, how did you recover the sequences, oxford nanopore vs illumina, shotgun vs 16s, number of samples? data preparation before binning, gene orientation, webserver vs local vs supercomuter, competency with the linux environment? sequence coverage, methylation signatures

4.2 Metagenomic Species Pan-genomes

4.3 Megagenome Assembled Genome

5 Binning microbial genomes with deep learning

- *Not sure if to focus on this or the appropriateness*
- *HMP and other*
- *The increased impact of machine learning in analysis*
- *Short section - just for past-present-future completeness*

supervised vs unsupervised

5.0.1 VAMB

The integration of deep learning techniques into the field of metagenomics has revolutionised the field of metagenomics. The VAMB pipeline was developed to take advantage of variational autoencoders; a generative machine learning model that uses a combination Improved metagenome binning and assembly using deep variational autoencoders Nature biotechnology - 4th Jan 2021 the VAMB pipeline (Nissen et al., n.d.)

5.0.2 Phylopythia

5.0.3 Coconet

6 Chosing the most appropriate binning algorithm

6.1 Binning for viral genomes

New insights from uncultivated genomes of the global human gut microbiome Nature - 13th March 2019 (Nayfach, Shi, Seshadri, Pollard, & Kyrpides, 2019)

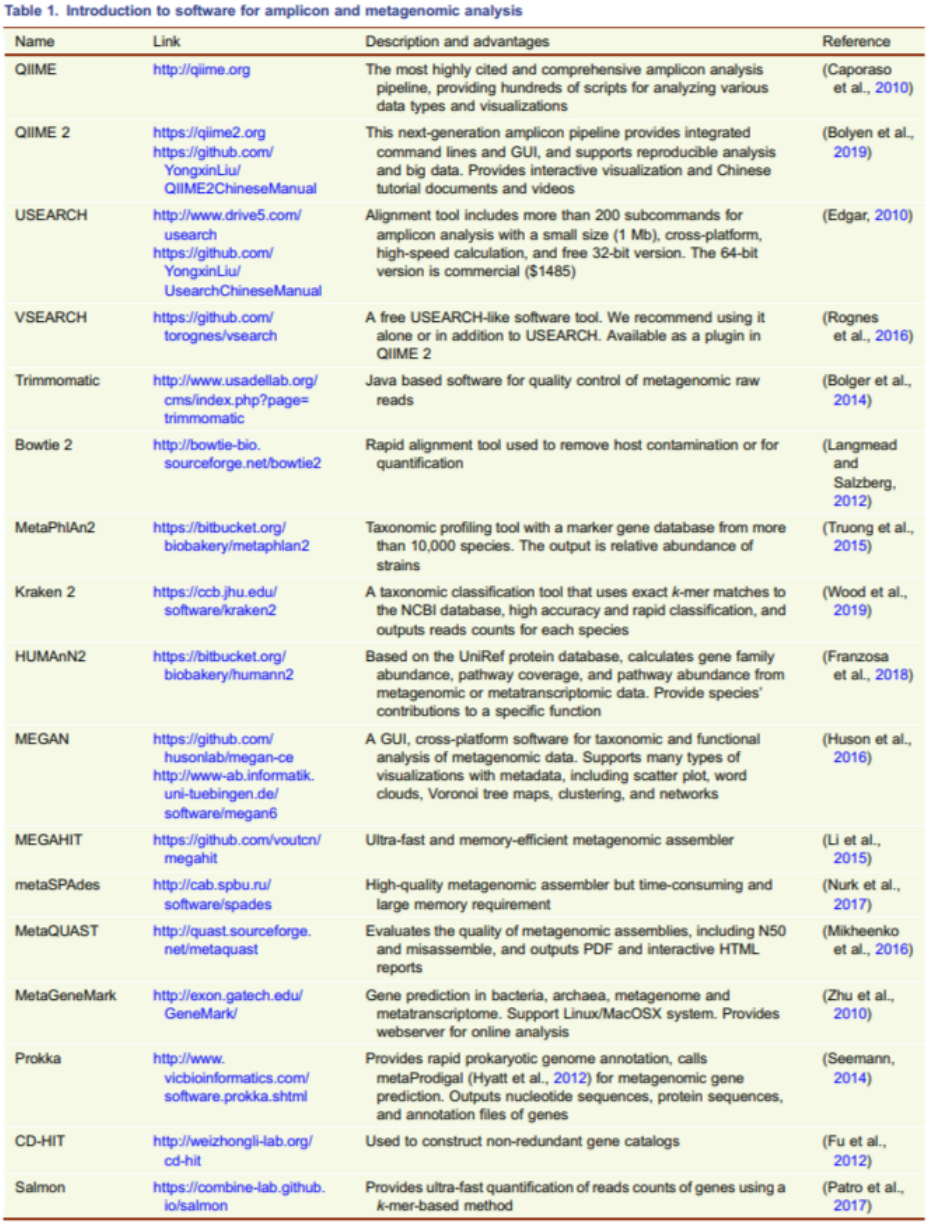

Table 1. Introduction to software for amplicon and metagenomic analysis

Name	Link	Description and advantages	Reference
QIIME	http://qiime.org	The most highly cited and comprehensive amplicon analysis pipeline, providing hundreds of scripts for analyzing various data types and visualizations	(Caporaso et al., 2010)
QIIME 2	https://qiime2.org https://github.com/YongxinLiu/QIIME2ChineseManual	This next-generation amplicon pipeline provides integrated command lines and GUI, and supports reproducible analysis and big data. Provides interactive visualization and Chinese tutorial documents and videos	(Bolyen et al., 2019)
USEARCH	http://www.drive5.com/usearch https://github.com/YongxinLiu/UsearchChineseManual	Alignment tool includes more than 200 subcommands for amplicon analysis with a small size (1 Mb), cross-platform, high-speed calculation, and free 32-bit version. The 64-bit version is commercial ($1485)	(Edgar, 2010)
VSEARCH	https://github.com/torognes/vsearch	A free USEARCH-like software tool. We recommend using it alone or in addition to USEARCH. Available as a plugin in QIIME 2	(Rognes et al., 2016)
Trimmomatic	http://www.usadellab.org/cms/index.php?page=trimmomatic	Java based software for quality control of metagenomic raw reads	(Bolger et al., 2014)
Bowtie 2	http://bowtie-bio.sourceforge.net/bowtie2	Rapid alignment tool used to remove host contamination or for quantification	(Langmead and Salzberg, 2012)
MetaPhlAn2	https://bitbucket.org/biobakery/metaphlan2	Taxonomic profiling tool with a marker gene database from more than 10,000 species. The output is relative abundance of strains	(Truong et al., 2015)
Kraken 2	https://ccb.jhu.edu/software/kraken2	A taxonomic classification tool that uses exact *k*-mer matches to the NCBI database, high accuracy and rapid classification, and outputs reads counts for each species	(Wood et al., 2019)
HUMAnN2	https://bitbucket.org/biobakery/humann2	Based on the UniRef protein database, calculates gene family abundance, pathway coverage, and pathway abundance from metagenomic or metatranscriptomic data. Provide species' contributions to a specific function	(Franzosa et al., 2018)
MEGAN	https://github.com/husonlab/megan-ce http://www-ab.informatik.uni-tuebingen.de/software/megan6	A GUI, cross-platform software for taxonomic and functional analysis of metagenomic data. Supports many types of visualizations with metadata, including scatter plot, word clouds, Voronoi tree maps, clustering, and networks	(Huson et al., 2016)
MEGAHIT	https://github.com/voutcn/megahit	Ultra-fast and memory-efficient metagenomic assembler	(Li et al., 2015)
metaSPAdes	http://cab.spbu.ru/software/spades	High-quality metagenomic assembler but time-consuming and large memory requirement	(Nurk et al., 2017)
MetaQUAST	http://quast.sourceforge.net/metaquast	Evaluates the quality of metagenomic assemblies, including N50 and misassemble, and outputs PDF and interactive HTML reports	(Mikheenko et al., 2016)
MetaGeneMark	http://exon.gatech.edu/GeneMark/	Gene prediction in bacteria, archaea, metagenome and metatranscriptome. Support Linux/MacOSX system. Provides webserver for online analysis	(Zhu et al., 2010)
Prokka	http://www.vicbioinformatics.com/software.prokka.shtml	Provides rapid prokaryotic genome annotation, calls metaProdigal (Hyatt et al., 2012) for metagenomic gene prediction. Outputs nucleotide sequences, protein sequences, and annotation files of genes	(Seemann, 2014)
CD-HIT	http://weizhongli-lab.org/cd-hit	Used to construct non-redundant gene catalogs	(Fu et al., 2012)
Salmon	https://combine-lab.github.io/salmon	Provides ultra-fast quantification of reads counts of genes using a *k*-mer-based method	(Patro et al., 2017)

Figure 1: Current pipelines available for metagenomic analysis - Something like this? from a 2017 review

6.2 Binning for viral genomes

7 Conclusion

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

8 Notes

Make a table with the following software:

- *MSPminer*
- *METEOR*
- *VAMB*
- *Metaphlan*
- *MGnify*
- *metabat*
- *cocacola*
- *NF-core-mag*

A review on the benchmarking binning algorithms was done by Yue et al., 2020.

References

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