Metagenomic Binning Pipelines - the State of the Art

Theo Portlock

January 26, 2021

1 Abstract

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. Modular and user-friendly tools would greatly improve such 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. Oweing to the democritization of open source software... The following article reviews the ...

2 Background

- *General introduction to metagenomics history*
- *increase in popularity of the field of metagenomics*
- *Maybe some important discoveries found as a result of metagenomic analysis*
- *The key programs that constitute metagenomic pipelines*
- *Maybe some benchmarking? Not sure*

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...

The steps involved in metagenomic analysis Quality Control, filtering, and trimming A review of the efficiency of quality control algorythms can be found here (Zhou, Su, & Ning, 2014). Trimming is to remove adapters, primers, and over-represented sequences, and to trim poor quality basepairs Sequence alignment - Bowtie2, Tophat2, Hisat2 are used to map reads against a database Classifying taxonomy and Annotation Assembly Functional analyses Visualization

3 Binning

Kaiju, Kraken2, Braken, mOTU, fetchMG, MetaPhIAn Centrifuge, METEOR pipelines are chosen based on a number of factors 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. Galaxy EBI Metagenomics (MGnify) has doubled the number of publicly available anaysed datasets held within the resource in two years.

- *Section introduces the most popular pipelines*
- *Features that distinguish these pipelines*

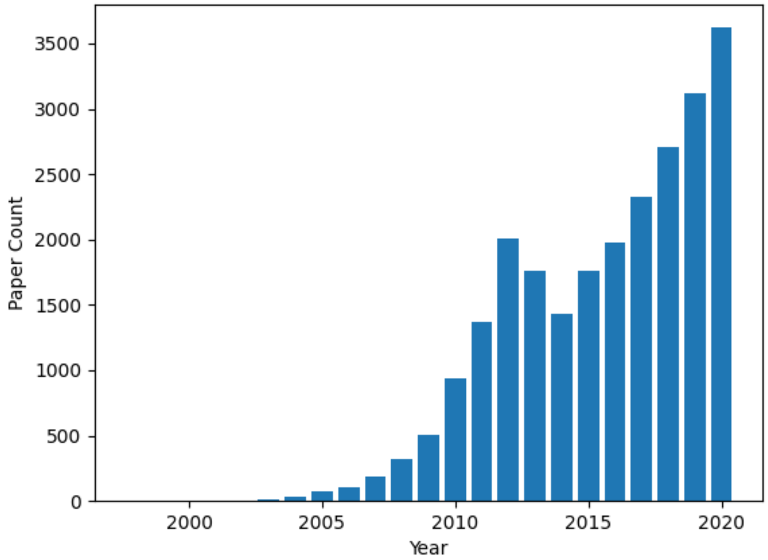

Figure 1: Popularity increase of the field of metagenomics. Measured by number of publications hosted on pubmed.gov

- *Some guidelines for chosing the correct pipeline appropriate for a given study*
- *Largest section*

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 that may have negative consequences for patient care.

4 Traditional binning pipelines

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

5 Machine learning assisted pipelines

Improved metagenome binning and assembly using deep variational autoencoders Nature biotechnology - 4th Jan 2021 the VAMB pipeline (Nissen et al., n.d.)

6 Conclusion

Future developments for metagenomic analysis

- *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*

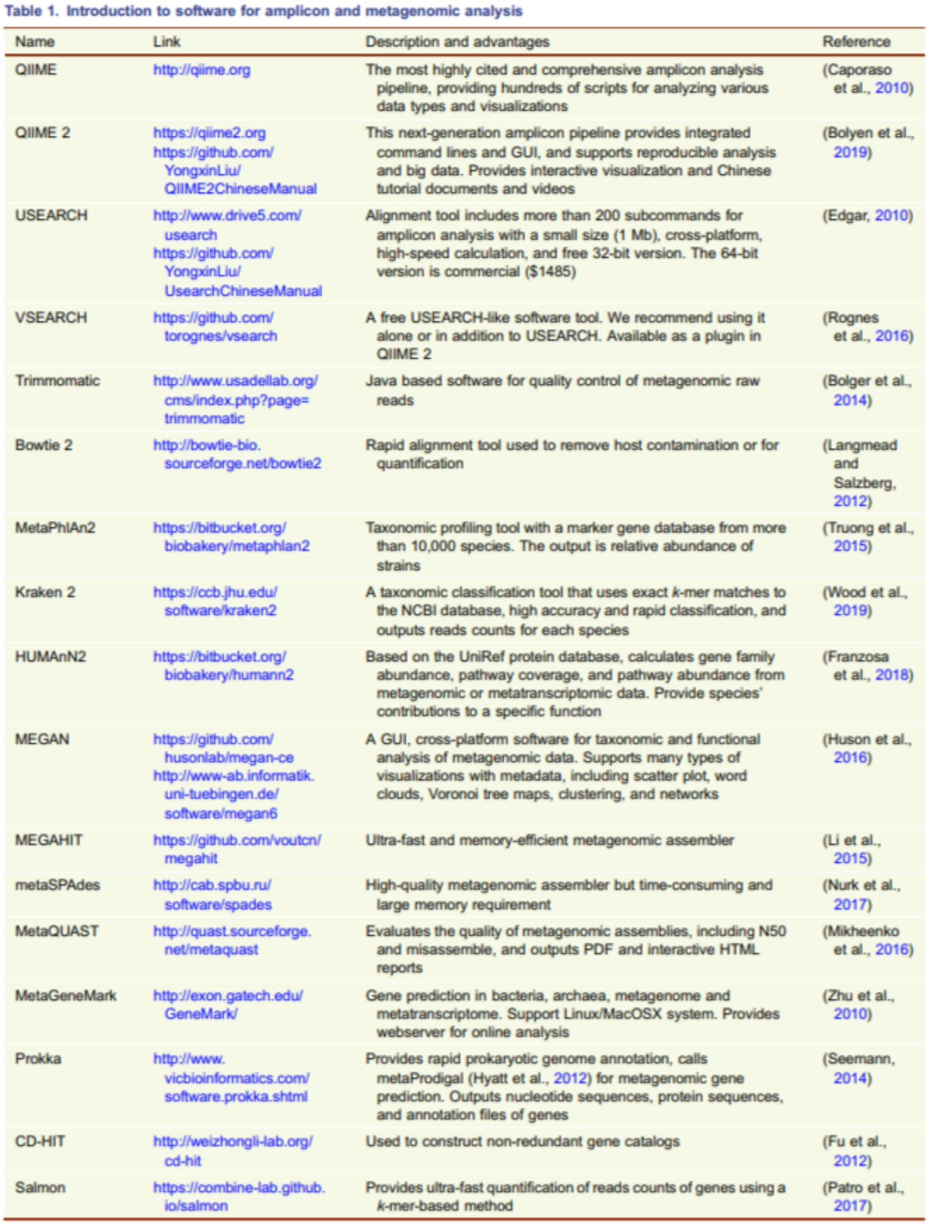

Table 1. Introduction to software for amplicon and metagenomic analysis

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

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.

Zhou, Q., Su, X., & Ning, K. (2014). Assessment of quality control approaches for metagenomic data analysis. *Scientific reports*, *4*(1), 1–11.