# Approaches to metagenomic classification and assembly

J. Marić and M. Šikić
Faculty of Electrical Engineering and Computing,
Laboratory for Bioinformatics and Computational Biology, Unska 3, 10000 Zagreb
Email: josip.maric@fer.hr, mile.sikic@fer.hr

***Abstract* - Microbiome is an ecological community of commensal, symbiotic, and pathogenic microorganisms that share the same environment. The study of microbiome, i.e. genetic material sampled directly from environmental samples is called metagenomics. In recent years, genome sequencing methods have dramatically improved and the number and variety of sequenced genomes has rapidly increased. New technology has significantly increased the variety and complexity of the microbiome research and ever-larger datasets present new challenges in analysis of metagenomic data. Two main tasks in metagenomic analysis are classification of sequenced metagenomic data into taxonomic group of any rank, such as a species, family, or class, and assembly of the data into longer contiguous sequences. The final aim of both tasks is to correctly identify species presented in the metagenomic sample. This has various applications in medicine (e.g. infectious disease diagnosis), development of biofuels, biotechnology, agriculture, and many other areas. In this paper, we present a description of common procedures and methods for metagenomic data analysis and challenges facing these procedures. We give an overview of existing software tools and a review of public genome databases used for metagenomic analysis. Finally, we explore possible improvements to the existing methods for metagenomic classification and assembly.**

***Keywords* - *microbiome, metagenomics, genome sequencing, databases, assembly, classification***

## I. Introduction

DNA is a long molecule that consists of large number of simple molecules called nucleotides or bases. The process of resolving the structure of a DNA molecule is called DNA sequencing. The Human Genome Project was the first attempt to determine the sequence of bases that make up human DNA and it lasted for more than 10 years [1]. Since then, the technology has rapidly advanced and today we have various sequencing techniques. Next Generation Sequencing (NGS) technologies brought lower cost and higher throughput, but also produced short read lengths [2]. Most prominent second generation sequencing technology is Ilumina, which produces nucleotide base sequences called reads, ranging from 30 bases do 500 bases and having low error rates [4]. New technologies, called third-generation sequencing, such as PacBio [4] and ONT [5], produce much longer reads with average read length of several thousand base-pairs, but with higher error rates. While the error rates of reads produced by these technologies is higher than the ones produces by NGS, their high average read length enables sequencing through extended repetitive regions, detection of mutations, identification of gene isoforms and discovery of new genes. These properties make third generation of sequencers, and the data they produce, the central point around which new bioinformatic methods should be researched and developed. They are revolutionizing the assembly and structural variant analysis of single genomes and, as their throughput improves, these technologies have tremendous potential for metagenomic classification as well.

There are many computational challenges in the field of bioinformatics, such as reconstruction of the original DNA sequence from read fragments called assembly [6] or mapping DNA or RNA sequences onto a previously assembled reference DNA. DNA assembly methods have enabled reconstruction of many genomes ranging from simple bacteria genomes to large human genomes, while mappers have been used as a tool in solving many different problems, such as the analysis of expression of genes by mapping samples of RNA sequences, or DNA assembly itself.

The analysis of microbiota, the assemblage of microorganisms presented in a defined environment [7], is another complex computational challenge that uses DNA/RNA mappers as well as assembled genomes. This field of work is, in the broader sense, called metagenomics analysis. As shown in Figure 1, the main difference between genomic and metagenomics analysis is that genomic analysis is done with a culture of a single microbe, while metagenomics analysis uses the community of many, mainly unculturable, microbial species. There are many applications of metagenomics analysis, suchs as agriculture [8], pathogen identification

[9], [10], gut microbiota analysis [11], biofuels [12] and many more. All these applications rely on different metagenomic analysis methods and software tools that provide these analysis. There are two main goals of metagenomic analysis: (1) taxonomic composition (identification of species or other filogenetic entities) and (2) functional analysis (gene identification) of the environmental sample. In the next chapter we will present main terminology related to the metagenomic research and give an overview of different methods used in metagenomic analysis. The emphasis will be on the methods that provide taxonomic composition of the sample. In the third chapter we will give an overview of currently existing tools which are used in metagenomic analysis, and, in the last chapter, we will present existing genomic databases used for metagenomic analysis.

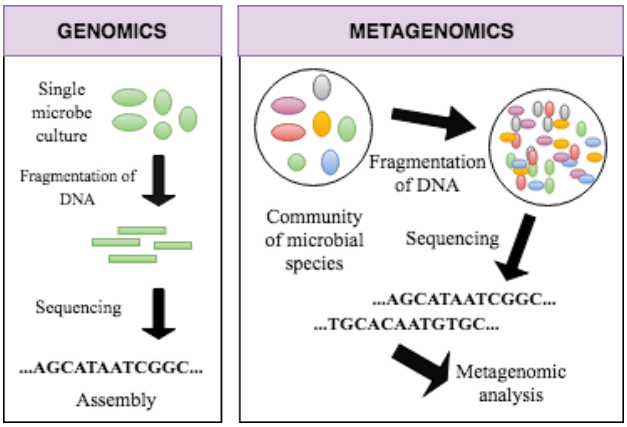


**Figure 1 -** The difference between genomics and metagenomics analysis

## II. Methodology

As the methods for genome sequencing rapidly improve so does the number and variety of sequenced genomes. This provides many opportunities as well as challenges in extracting information from environmental samples. Different approaches to researching microbiota can be divided into three categories: metataxonomics, metagenomics and metatranscriptomics.

Metataxonomics uses sequences of marker genes which are highly conserved across taxonomic groups, such as species, to identify organisms in an environmental sample. Those genes include 16S rRNA gene for bacteria, 18S rRNA gene for eucaryotes and ITS regions of fungal ribosomes for fungi. Since it requires only sequences from one gene, this approach is less computationally complex. However it has drawbacks since it does not capture viruses, and also many organisms evade detection with 16S genes [13]. Approach that uses reads from random shotgun sequencing of microbiota, without selecting a particular gene, is called metagenomics [7]. Metagenomic analysis has achieved unprecedented performance in microbial community profiling, allowing researchers to access the composition of many different microbial communities [14]. Metatranscriptomics, on the other hand, attempts to capture and sequence RNA in a sample in order to create a profile of genes that are currently being transcribed [15]. This analysis provides insight into gene expression and regulatory mechanisms that can contribute to discovery of new and personalised drugs which can significantly contribute to human health [14].

There are two main approaches in metagenomic analysis. First approach provides results by comparing microbiota reads from a sample, to some public genome database in order to determine taxonomic composition or gene expression. This approach is called direct metagenomic classification and it has advantages over metataxonomics since it alleviates biases from marker genes that metataxonomic approaches have and identifies organisms across all domains of life. However, when there are species in the sample with no known genomes in the database, then another approach has to be used - metagenomic assembly. While the assembly of single genome is already a challenging problem, assembly of a mixed sample with many species in different abundances is even more complicated and requires special-purpose assembly algorithms [16]. The biggest problem is highly uneven sequencing depth of different organisms in a metagenomic sample. Another issue is that metagenomic sample is not likely to contain deep coverage of more than one or two species. This means that the results of metagenomics assembly will never be as good as those from assembly of a single organism. However assembly of metagenomic sample often succeeds in merging many reads resulting in contigs that can be aligned to the reference genomes more easily or analysed without the reference genomes.

All these approaches in microbiome analysis have some common steps which are implemented in most of the existing tools. Common first step is to run various quality control methods that identify and remove low quality reads. If the sample was sequenced from a host, such as human, sometimes identification and removal of host reads is also done. Remaining reads can be directly used to classify species in the sample or assembled into larger contigs. Direct classifiers use DNA mapping tools to map reads from the sample to a genome database in order to place reads into certain taxons. These classifiers can be divided into two main categories based on the amount of reads the classifier uses, (1) marker gene classifiers and (2) all-reads classifiers.

Marker gene classifiers use reduced database, consisting only of specific marker genes that identify specific species. Since this approach involves comparing metagenomic reads to a relatively small database, marker gene analysis can be a rapid way to estimate diversity of a metagenomic sample. Since marker gene methods identify only a few genes per genome, most of the reads in a sample are not classified at all.

All-reads classifiers use the whole dataset of sequenced reads and map them to the database of whole sequenced genomes. These classifiers usually use much higher percentage of sequences in their analysis than marker gene classifiers and can provide abundances of organisms present in a metagenomic sample. Direct classifiers that use reference database sometime perform a step of assembly of metagenomic reads into longer contigs and then perform contig classification. In the following chapter we will give a review of software tools that implement these different approaches in metagenomic research. A breakdown of microbiota research approaches with the list of tools for each research is given in Figure 1.

| **Microbiota research approach** | Metatxonomics | Metagenomics | | | | Metatranscriptomics |
|---|---|---|---|---|---|---|
| **Research goal** | species/strain identification | species/strain identification | | | | gene identification |
| **Sequencing type** | specific gene sequencing | whole sample sequencing | | | | RNA sequencing |
| **Metagenomics research approach** | | direct classification | | | read assembly | |
| **Classification methods** | | all reads mapping | marker gene detection | strain identification | | |
| **Tools** | QIIME, UPARSE, Mothur, DADA2 | Kraken, CLARK | MetaPhlAn, FOCUS, MG-RAST | StrainPhlAn, PanPhlan, Constraint, Sigma | Meta Velvet, Ray meta, IDBA, MetaSPAdes | HUMAN2, Leimena-2013, SotMeRNA |

**Figure 2 -** Breakdown of microbiota research approaches

## III. Tools

There are several bioinformatic tools that provide metataxonomic analysis of sequenced metagenomic samples by using 16s rRNA marker gene such as QIIME [17], UPARSE [18], Mothur [19], and DADA2 [20]. As mentioned earlier, these methods target only specific genes and can fail to detect many species, while not detecting viruses at all. A comprehensive analysis of metagenomic tools that use 16s tRNA gene is given in [14].

One of the first methods for assigning taxonomic labels to unknown reads was done with program BLAST [21], which classified sequences by aligning them to a large database of genomes. Certain tools used the results of BLAST and applied Bayesian rule to distribution of matches to identify species inside the sample, but these tools perform even slower than BLAST itself [21]. Since then, many metagenomic analysis tools have been developed.

Kraken [22] is a sequence classification tool which uses exact match database queries of k-mers. It is able to achieve genus-level sensitivity and precision. It uses database that connects a k-mer with the lowest common ancestor of all organisms whose genomes contain that k-mer. Using this database the Kraken is able to make a quick lookup of the most specific node in the taxonomic tree which is associated with the given k-mer. The Kraken database is built from the NCBI RefSeq database [23]. To classify the DNA sequence, Kraken collects all k-mers within the sequence into a set and forms taxonomy tree from that set. Each node in the tree is weighted with the number of k-mers that are mapped to the taxon associated to that node. Then every root-to-leaf path in the tree is scored by calculating the sum of all node weights along the path. The maximum path is called classification path and the sequence is assigned the label corresponding with the leaf of the path.

CLARK [24] (CLAssifier based on Reduced K-mers) is another sequence classification tool that matches reads with the database of genomes based on a set of k-mers.

Clark builds the reduced database of k-mers where all common k-mers between targets in the database (e.g. a collection of genomes) are removed which produces a set of genomic regions that uniquely describe each target. A read is classified as a target with which it shares the highest number of k-mers. Kraken and CLARK have been noted as most precise and fast tools that currently exist in [25].

MetaPhlAn2 [26] is a marker gene based direct sequence classification tool that mapps reads against a reduced set of clade-specific marker sequences. It infers the presence and read coverage of clade-specific markers to unequivocally detect the taxonomic clades present in microbiome sample and estimate their relative abundance. Clade-specific markers are coding sequences that satisfy the conditions of (1) being strongly conserved within the clade's genomes and (2) not possessing substantial local similarity with any sequence outside the clade. The marker set needs to be generated once for a set of database genomes. The set used in [26] contained around 1 million markers from over 7500 species. MetaPhlAn2 uses Bowtie2 [27], a fast mapper, to map reads to marker genes, assigns read counts to the microbial clades according to the alignment prediction and outputs a matrix containing relative abundances of the identified species, genus, family, order, class, phylum and kingdom. In the rare case of multiple matches with markers from different clades, only the best hit is considered. For each clade, read counts can be assigned directly using the clade-specific markers or indirectly by considering the reads assigned to all direct descendants.

Another marker gene based tool is FOCUS [28], which uses nonnegative least squares (NNLS) to report organisms profile. Focus calculates k-mer frequencies for sequences of genomes in database with length of k-mer from 6-8 bases. NNLS problem can be formalised as:

Given a matrix $A \in \Re^{mxn}$ and a vector $b \in \Re^{m}$, where $m \geq n$, find a non-negative vector $x \in \Re^{n}$ to minimize the function.

$$f(x) = 1/2 \,\|\, Ax - b \,\|, \; where\, x \geq 0 \; and \; \sum_{i=1}^{n} x_i = 1 \qquad (1)$$

In FOCUS, the reference matrix A is composed of m k-mer frequencies from n genomes, while a vector describing the user's metagenomic dataset is calculated from the k-mer frequencies of both strands from the dataset. NNLS is used to compute the set of k-mer frequencies x that gives optimal possible abundance of k-mers in the user's metagenome by selecting the optimal number of frequencies from the matrix A.

MG-RAST [29] is the tool that provides metagenomic and metatranscriptomic analysis of metagenomic data whose upload is offered to the users through the web page. The tool extracts multiple features for users to assess sequence quality and address some of the common issues like high error rates, contamination with adapters, duplicate reads. The tool maps metagenomic sequences with BLASTX [30] to the database sources of protein encoding genes [31]. In parallel the sequence data is compared to all accessed databases which include several rDNA databases, chloroplast database, mitochondrial database, and other.

All mentioned tools are used to report organisms present in metagenomic samples and profile their abundances. In some cases, such as diseases related to a single group of microorganisms, different strains in the same species may have different effect on human health This is why identification of species in a metagenomic sample is not sufficient for some applications. Identification and profile of different strains of the same species in the environment are crucial to revealing human-microbial interactions. We present bioinformatic tools that provide strain identification in a metagenomic sample.

StrainPhlAn [32] infers the strain-level phylogenetic structure of microbial species across metagenomic samples by reconstructing the consensus sequences of the dominant strain for each detected species in a sample and then comparing the consensus sequences in different samples. It uses MetaPhlAn2 to map the reads to MetaPhlAn2 marker database and uses the alignments to produce phylogenetic trees using the maximum-likelihood principle [33]. To identify the presence of multiple strains from the same species in a single sample, reads-to-markers mappings are analysed to find evidence of polymorphic sites [34] on the alignments suggestive of multiple alleles. For each position $s$ in the alignment $N_s$ is defined as the total number of reads covering it and $T_s$ as the number of reads supporting the dominant allele. With the sequencing error rate $E$, the nonpolymorphic null hypothesis is rejected if the probability that the number of reads $N_s - T_s$ that comes from the nondominant allele is $< \alpha = 0.05$. failing to reject the null hypothesis indicates the absence of alternative alleles.

PanPhlan [35] is a tool that uses metagenomic data and achieves strain-level microbial profiling. It identifies which genes are present or absent within different strains of species, based on the entire gene set of the species' pangenome [36]. The pangenome of the species includes genes present in all strains and genes present only in some strains of a species (variable genome). The core

genome represents the genes present in all strains of a species. PanPhlAn builds the pangenome of a species by extracting all genes from available reference genomes and merging them into gene clusters. It then maps the metagenomic sample reads against the reference genomes to obtain gene coverage levels and then reconstructs the unique gene set of a strain present in the sample. A strain is predicted to be present when the abundance of species-specific genes indicates a nearly uniform coverage depth across an expected number of N gene families representing the median number of gene families of all reference genomes of a species.

ConStrains [37] is another tool that identifies organisms in a sample at strain level. ConStrains compares raw metagenomic reads to reference genomes and identifies single nucleotide polymorphism (SNP) patterns as the basis in differentiation and quantification of different strains. The tool achieves this in two steps: (1) identifies species for which SNPs are detected and quantified, and (2) transforms individual SNPs into SNP profiles that represent individual strains. First step uses MetaPhlAn for species composition. For the species with sufficient sequencing depth, a custom database of marker genes is created from the PyloPhlan marker set [38], against which the raw reads are mapped using Bowtie2. The resulting alignments are used to generate a table of coverage by base position from which SNPs are identified. SNPs are counted across samples as those positions where the minos allele had at least two counts or more than 3% in relative abundance.

The last strain identification tool presented here is Sigma [39] which uses read mapping approach with a probabilistic model, maximum likelihood estimation (MLE) of the relative abundances of all genomes measured by the percentages of reads sampled from these genomes. Sigma's strain classification method does not use any marker genes, as methods described before, and it will be described in more detail here. Sigma calculates the P-value for identification of a genome using a likelihood ratio test of the null hypothesis that the genome is not present in the sample.

Let $Pr(g_j) \in [0,1]$ be the probability of sampling a random read from the reference genome $g_j$. Considering only reads sampled from reference genomes in the database, we have:

$$\sum_{\forall j} Pr(g_j) = 1 \tag{2}$$

$Pr(g_j)$ is determined by the relative abundance, size, and sequencing bias of the genome $g_j$. Sigma estimates $Pr(g_j)$ using MLE. Consider a read $r_i$ mapped to the genome $g_j$ with $z$ mismatches and a uniform probability of $\sigma$ assumed for any mismatch between a read and a genome. The probability of obtaining $r_i$ with $z$ mismatches and $(l-z)$ matches from the genome $g_i$ is:

$$Pr(r_i|g_j) = \sigma^z(1-\sigma)^{l-z};\ z \le U, \tag{3}$$

$$Pr(r_i|g_i) = 0\ ;z > U, \tag{4}$$

where $l$ is the length of $r_i$ and U is the maximum number of mismatches allowed in the alignment. $Pr(r_i|g_i)$ is calculated between each read and each genome based on their alignment and matrix $Q$ is populated:

$$Q_{i,j} = Pr(r_i|g_i). \tag{5}$$

The probability of sampling $r_i$ from $g_i$ is:

$$Pr(r_i, g_j) = Pr(r_i|g_j) \cdot Pr(g_j) = Q_{i,j} \cdot Pr(g_j), \tag{6}$$

And the probability of generating $r_i$, since it may originate from any of the reference genomes, is:

$$Pr(r_i) = \sum_{\forall j} Pr(r_i, g_i). \tag{7}$$

MLE finds the estimate of $Pr(g_j)$ by maximizing the probability of all reads, which is the joint probability of all reads:

$$max\ Pr(r_1, ..., r_n) = max \prod_{\forall i} Pr(r_i) \tag{8}$$

$$= max \prod_{\forall i} [\sum_{\forall j} Pr(r_i, g_j)] \tag{9}$$

$$= max \prod_{\forall i} [\sum_{\forall j} Q_{i,j} \cdot Pr(g_j)]. \tag{10}$$

This optimisation problem is solved using non-linear programming method implemented in the Ipopt library [40].

Here are some of the most notable metatranscriptomic tools that try to capture all of the RNA sequences in the sample and create profiles of all genes that are being transcribed. HUMAnN2 [41] is a tool designed for analysis in both metagenomics and metatranscriptomics. It identifies a community's known species, aligns reads to their pangenomes, performs translated search on unclassified reads, and finally quantifies gene families and pathways. A tool presented in [42] uses SotMeRNA [43] and BLASTN for the removal of rRNA and tRNA reads, and megaBLAST for mRNA reads alignment to reference genome database and classifies the alignments to protein and non protein encoding regions. More on metatranscriptomic tools is given in [14].

Lastly, metagenomic assembler tools are presented in [44]. MetaVelvet [45] and Ray meta [46] are single k-mer de Bruijn graph assemblers for metagenomic data. MetaVelvet extends single-genome assembler for short reads, known as Velvet [47], to metagenome assembly. It decomposes de Bruijn graph constructed from mixed short reads by Velvet, into individual sub-graphs and builds scaffolds based on each decomposed de Bruijn sub-graph as an isolated species genome. Ray Meta constructs contigs by a heuristics-guided graph traversal. IDBA assembler [48] is iterative De bruijn graph assembler that generates contigs from iteratively constructed and refined de bruijn graphs using multiple k-mer lengths from small k's to larger k's, replacing reads with reassembled contigs at each iteration. MetaSPAdes [49] is an extension of the SPAdes assembler which uses approach similar to IDBA with iterative de Bruijn graph refinement, but also implements various heuristics for graph simplification, filtering and storage.

## IV. Genomic resources

The microbial genome resources are of high importance for metagenomic research. The quantity and quality of genomes in the database greatly influences the precision of the classifiers in species identification. Most commonly used reference genome database is the database of complete and draft genomes at GenBank which has been the repository for genome sequences for more than a quarter of century [50]. GenBank genomes receive taxonomic identification by their uploaders and some genomes have incorrect species name. Another issue with GenBank data is contamination, where the vast majority of genomes in GenBank are 'draft' genomes. These are genomes for which the assembly was generated, but most of the chromosomes are fragmented into many pieces. Some of the contigs might be contaminants, i.e. they might not belong to sequenced species. The result of these contaminants is that reads from metagenomic sample will match some draft genomes extremely well because the metagenomic sample contains same contamination.

The RefSeq project uses GenBank sequences and filters them to get more curated genome resources and provides alternative database for metagenomic research [23]. RefSeq database provides, for each individual species, a complete, non-redundant and richly annotated genome sequences. Currently it contains more than 88 thousand organisms, a number that will surely grow in the future.

Since 2014 only novel species are assigned new taxonomy ID in GenBank database while assigned strain taxonomy ID stays in the database [51]. This means that single species can have genomes at species and strain level which can be challenging for algorithms that try to characterize metagenomic samples at strain level.

Clade specific markers are another important genomic resource in analysing microbiota. MetaPhlAn2 is the most prominent tool that utilises such resource. The process of identification and extraction of clade-specific core genes is given in [52]

## V. Future prospects

With all the presented metagenomics research approaches and tools that analyse metagenomic samples there are certain ways in which the future research can progress beyond the state-of-the-art. In order to utilise good properties of reads created by third generations sequencers, a new metagenomic analysis tool that classifies these reads to different strains of the same species needs to be developed. A good start would be to place reads into certain classes by using methods with k-mers, similar to Kraken or CLARK. With MetaPhlAn2 marker database available, the fast mapping of the reads to the database of marker genes should be researched and if possible used alongside k-mer approach. These methods can be improved by utilising currently best mappers evaluated in [53] which have shown good results in mapping DNA or RNA third generation sequences to a reference genome. Finally, strain classification could be done with the classification algorithm used in SIGMA which would use the results provided by previous steps of algorithm. Before tackling these challenges, a thorough benchmark of currently existing tools for metagenomic classification should be done, with exact and clearly defined measures of quality of the tools.

## VI. Conclusion

NGS sequencers provide powerful tool to analyse microbiota. There is a variety of tools developed to help analyze metagenomic datasets. In this paper, we presented an overview of different types of analysis of metagenomic samples. Metataxonomic methods do not capture viruses and usually don't classify many present species. Metataxonomic methods analyse metagenomic samples of RNA and provide gene expression analysis of the sample. Metagenomic methods classify read in the metagenomic sample to the reference database, or assemble them into larger contigs, both with goal to identify species present in the sample. With many applications in different areas of interest like medicine, agriculture and other, the analysis of metagenomic

samples is of high importance for future of human species and will continue at an even faster pace.

## VII. Acknowledgements

This work has been supported in part by the Croatian Science Foundation under the project Single genome and metagenome assembly (IP-2018-01-5886), in part by the European Regional Development Fund under the grant KK.01.1.1.01.0009 (DATACROSS) and in part by the Croatian Academy of Arts and Sciences under the project Classification of DNA and RNA reads sequenced by third generation sequencing tools.

## References

[1] Hood, L., & Rowen, L. (2013). The Human Genome Project: big science transforms biology and medicine. *Genome Medicine*, *5*(9), 79. https://doi.org/10.1186/gm483

[2] I. Sović, K. Skala and M. Šikić, "Approaches to DNA de novo assembly," 2013 36th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, 2013, pp. 351-359.

[3] van Dijk, E. L., Auger, H., Jaszczyszyn, Y., & Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends in Genetics*, *30*(9), 418–426. https://doi.org/10.1016/j.tig.2014.07.001

[4] Rhoads, A., & Au, K. F. (2015). PacBio Sequencing and Its Applications. *Genomics, Proteomics & Bioinformatics*, *13*(5), 278–289. https://doi.org/10.1016/J.GPB.2015.08.002

[5] Jain, M., Olsen, H. E., Paten, B., & Akeson, M. (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology*, *17*(1), 239. https://doi.org/10.1186/s13059-016-1103-0

[6] Vaser R.(2016). Approaches to haplotype assembly. *Laboratory for Bioinformatics and Computational Biology*

[7] Marchesi, J. R., & Ravel, J. (2015). The vocabulary of microbiome research: a proposal. *Microbiome*, *3*(1), 31. https://doi.org/10.1186/s40168-015-0094-5

[8] Piškur, J., Ling, Z., Marcet-Houben, M., Ishchuk, O. P., Aerts, A., LaButti, K., … Phister, T. (2012). The genome of wine yeast Dekkera bruxellensis provides a tool to explore its food-related properties. *International Journal of Food Microbiology*, *157*(2), 202–209. https://doi.org/10.1016/j.ijfoodmicro.2012.05.008

[9] Cao, M. D., Ganesamoorthy, D., Elliott, A. G., Zhang, H., Cooper, M. A., & Coin, L. J. M. (2016). Streaming algorithms for identification of pathogens and antibiotic resistance potential from real-time MinIONTM sequencing. *GigaScience*, *5*(1), 32. https://doi.org/10.1186/s13742-016-0137-2

[10] Greninger, A. L., Naccache, S. N., Federman, S., Yu, G., Mbala, P., Bres, V., … Chiu, C. Y. (2015). Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. *Genome Medicine*, *7*(1), 99. https://doi.org/10.1186/s13073-015-0220-9

[11] Zmora, N., Suez, J., & Elinav, E. (2019). You are what you eat: diet, health and the gut microbiota. *Nature Reviews Gastroenterology & Hepatology*, *16*(1), 35–56. https://doi.org/10.1038/s41575-018-0061-2

[12] Hess, M., Sczyrba, A., Egan, R., Kim, T.-W., Chokhawala, H., Schroth, G., … Rubin, E. M. (2011). Metagenomic Discovery of Biomass-Degrading Genes and Genomes from Cow Rumen. *Science*, *331*(6016), 463–467. https://doi.org/10.1126/science.1200387

[13] Eloe-Fadrosh, E. A., Ivanova, N. N., Woyke, T., & Kyrpides, N. C. (2016). Metagenomics uncovers gaps in amplicon-based detection of microbial diversity. *Nature Microbiology*, *1*(4), 15032. https://doi.org/10.1038/nmicrobiol.2015.32

[14] Niu, S.-Y., Yang, J., McDermaid, A., Zhao, J., Kang, Y., & Ma, Q. (2017). Bioinformatics tools for quantitative and functional metagenome and metatranscriptome data analysis in microbes. *Briefings in Bioinformatics*, *19*(6), 1415–1429. https://doi.org/10.1093/bib/bbx051

[15] Moran, M. A., Satinsky, B., Gifford, S. M., Luo, H., Rivers, A., Chan, L.-K., … Hopkinson, B. M. (2013). Sizing up metatranscriptomics. *The ISME Journal*, *7*(2), 237–243. https://doi.org/10.1038/ismej.2012.94

[16] Cao, M. D., Ganesamoorthy, D., Elliott, A. G., Zhang, H., Cooper, M. A., & Coin, L. J. M. (2016). Streaming algorithms for identification of pathogens and antibiotic resistance potential from real-time MinIONTM sequencing. *GigaScience*, *5*(1), 32. https://doi.org/10.1186/s13742-016-0137-2

[17] Kuczynski, J., Stombaugh, J., Walters, W. A., González, A., Caporaso, J. G., & Knight, R. (2011). Using QIIME to Analyze 16S rRNA Gene Sequences from Microbial Communities. In *Current Protocols in Bioinformatics* (Vol. Chapter 10, p. Unit 10.7.). Hoboken, NJ, USA: John Wiley & Sons, Inc. https://doi.org/10.1002/0471250953.bi1007s36

[18] Edgar, R. C. (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nature Methods*, *10*(10), 996–998. https://doi.org/10.1038/nmeth.2604

[19] Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., … Weber, C. F. (2009). Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. *Applied and Environmental Microbiology*, *75*(23), 7537–7541. https://doi.org/10.1128/AEM.01541-09

[20] Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, *13*(7), 581–583. https://doi.org/10.1038/nmeth.3869

[21] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

[22] Wood, D. E., & Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, *15*(3), R46. https://doi.org/10.1186/gb-2014-15-3-r46

[23] Pruitt, K. D., Tatusova, T., Brown, G. R., & Maglott, D. R. (2012). NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Research*, *40*(D1), D130–D135. https://doi.org/10.1093/nar/gkr1079

[24] Ounit, R., Wanamaker, S., Close, T. J., & Lonardi, S. (2015). CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC Genomics*, *16*(1), 236. https://doi.org/10.1186/s12864-015-1419-2

[25] Lindgreen, S., Adair, K. L., & Gardner, P. P. (2016). An evaluation of the accuracy and speed of metagenome analysis tools. *Scientific Reports*, *6*(1), 19233. https://doi.org/10.1038/srep19233

[26] Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., & Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods*, *9*(8), 811–814. https://doi.org/10.1038/nmeth.2066

[27] Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

[28] Silva, G. G. Z., Cuevas, D. A., Dutilh, B. E., & Edwards, R. A. (2014). FOCUS: an alignment-free model to identify organisms in metagenomes using non-negative least squares. *PeerJ*, *2*, e425. https://doi.org/10.7717/peerj.425

[29] Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E., Kubal, M., … Edwards, R. (2008). The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, *9*(1), 386. https://doi.org/10.1186/1471-2105-9-386

[30] Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, *25*(17), 3389–3402. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9254694

[31] Overbeek, R., Begley, T., Butler, R. M., Choudhuri, J. V, Chuang, H.-Y., Cohoon, M., … Vonstein, V. (2005). The Subsystems Approach to Genome Annotation and its Use in the Project to Annotate 1000 Genomes. *Nucleic Acids Research*, *33*(17), 5691–5702. https://doi.org/10.1093/nar/gki866

[32] Truong, D. T., Tett, A., Pasolli, E., Huttenhower, C., & Segata, N. (2017). Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Research*, *27*(4), 626–638. https://doi.org/10.1101/gr.216242.116

[33] Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312–1313. https://doi.org/10.1093/bioinformatics/btu033

[34] Cann, R. L., Brown, W. M., & Wilson, A. C. (1984). Polymorphic sites and the mechanism of evolution in human mitochondrial DNA. *Genetics*, *106*(3), 479–499. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/6323246

[35] Scholz, M., Ward, D. V, Pasolli, E., Tolio, T., Zolfo, M., Asnicar, F., … Segata, N. (2016). Strain-level microbial epidemiology and population genomics from shotgun metagenomics. *Nature Methods*, *13*(5), 435–438. https://doi.org/10.1038/nmeth.3802

[36] Medini, D., Donati, C., Tettelin, H., Masignani, V., & Rappuoli, R. (2005). The microbial pan-genome. *Current Opinion in Genetics & Development*, *15*(6), 589–594. https://doi.org/10.1016/j.gde.2005.09.006

[37] Luo, C., Knight, R., Siljander, H., Knip, M., Xavier, R. J., & Gevers, D. (2015). ConStrains identifies microbial strains in metagenomic datasets. *Nature Biotechnology*, *33*(10), 1045–1052. https://doi.org/10.1038/nbt.3319

[38] Segata, N., Börnigen, D., Morgan, X. C., & Huttenhower, C. (2013). PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. *Nature Communications*, *4*(1), 2304. https://doi.org/10.1038/ncomms3304

[39] Ahn, T.-H., Chai, J., & Pan, C. (2015). Sigma: Strain-level inference of genomes from metagenomic analysis for biosurveillance. *Bioinformatics*, *31*(2), 170–177. https://doi.org/10.1093/bioinformatics/btu641

[40] Wächter, A., & Biegler, L. T. (2006). On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical Programming*, *106*(1), 25–57. https://doi.org/10.1007/s10107-004-0559-y

[41] Franzosa, E. A., McIver, L. J., Rahnavard, G., Thompson, L. R., Schirmer, M., Weingart, G., … Huttenhower, C. (2018). Species-level functional profiling of metagenomes and metatranscriptomes. *Nature Methods*, *15*(11), 962–968. https://doi.org/10.1038/s41592-018-0176-y

[42] Leimena, M. M., Ramiro-Garcia, J., Davids, M., van den Bogert, B., Smidt, H., Smid, E. J., … Kleerebezem, M. (2013). A comprehensive metatranscriptome analysis pipeline and its validation using human small intestine microbiota datasets. *BMC Genomics*, *14*(1), 530. https://doi.org/10.1186/1471-2164-14-530

[43] Kopylova, E., Noé, L., & Touzet, H. (2012). SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics*, *28*(24), 3211–3217. https://doi.org/10.1093/bioinformatics/bts611

[44] Breitwieser, F. P., Lu, J., & Salzberg, S. L. (2017). A review of methods and databases for metagenomic classification and assembly. *Briefings in Bioinformatics*. https://doi.org/10.1093/bib/bbx120

[45] Namiki, T., Hachiya, T., Tanaka, H., & Sakakibara, Y. (2012). MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Research*, *40*(20), e155–e155. https://doi.org/10.1093/nar/gks678

[46] Boisvert, S., Raymond, F., Godzaridis, É., Laviolette, F., & Corbeil, J. (2012). Ray Meta: scalable de novo

metagenome assembly and profiling. *Genome Biology*, *13*(12), R122. https://doi.org/10.1186/gb-2012-13-12-r122

[47] Zerbino, D. R., & Birney, E. (2008). Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, *18*(5), 821–829. https://doi.org/10.1101/gr.074492.107

[48] Peng, Y., Leung, H. C. M., Yiu, S.-M., Lv, M.-J., Zhu, X.-G., & Chin, F. Y. L. (2013). IDBA-tran: a more robust de novo de Bruijn graph assembler for transcriptomes with uneven expression levels. *Bioinformatics*, *29*(13), i326–i334. https://doi.org/10.1093/bioinformatics/btt219

[49] Nurk, S., Meleshko, D., Korobeynikov, A., & Pevzner, P. A. (2017). metaSPAdes: a new versatile metagenomic assembler. *Genome Research*, *27*(5), 824–834. https://doi.org/10.1101/gr.213959.116

[50] Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2013). GenBank. *Nucleic Acids Research*, *41*(Database issue), D36-42. https://doi.org/10.1093/nar/gks1195

[51] Federhen, S., Clark, K., Barrett, T., Parkinson, H., Ostell, J., Kodama, Y., … Karsch-Mizrachi, I. (2014). Toward richer metadata for microbial sequences: replacing strain-level NCBI taxonomy taxids with BioProject, BioSample and Assembly records. *Standards in Genomic Sciences*, *9*(3), 1275–1277. https://doi.org/10.4056/sigs.4851102

[52] Segata, N., & Huttenhower, C. (2011). Toward an Efficient Method of Identifying Core Genes for Evolutionary and Functional Microbial Phylogenies. *PLoS ONE*, *6*(9), e24704. https://doi.org/10.1371/journal.pone.0024704

[53] Križanović, K., Echchiki, A., Roux, J., & Šikić, M. (2018). Evaluation of tools for long read RNA-seq splice-aware alignment. *Bioinformatics*, *34*(5), 748–754. https://doi.org/10.1093/bioinformatics/btx668