**MetaSPAdes**

[10.1101/gr.213959.116](https://dx.doi.org/10.1101%2Fgr.213959.116)

* Versatile metagenomic assembler
* The assembly of metagenomic data remains challenging
* Results in high-quality assemblies across diverse data sets.
* High fragmentation of metagenomic assemblies negatively affects both the accuracy of binning and the contiguity of genomes attributed to specific bins
* enormous microdiversity of related strains within various microbial communities.
* Challenges
  + widely different abundance levels of various species in a microbial sample result in a highly nonuniform read coverage across different genomes.
  + various species within a microbial community often share highly conserved genomic regions
  + many bacterial species in a microbial sample are represented by strain mixtures, that is, multiple related strains with varying abundances
* SPAdes was initially developed to assemble data sets with nonuniform coverage, one of the key challenges of single-cell assembly
* Pipeline
  + constructs the de Bruijn graph of all reads with SPAdes
  + transforms it into the assembly graph using various graph simplification procedures
  + reconstructs paths in the assembly graph that correspond to long genomic fragments within a metagenome
* Responding to the microdiversity challenge, metaSPAdes focuses on reconstructing a consensus backbone of a strain mixture, thus ignoring some strain-specific features corresponding to rare strains.

**BWA-MEM**

* BWA is a software package for mapping low-divergent sequences against a large reference genome
* RPKM stands for Reads Per Kilobase of transcript per Million mapped reads. FPKM stands for Fragments Per Kilobase of transcript per Million mapped reads. In RNA-Seq, the relative expression of a transcript is proportional to the number of cDNA fragments that originate from it. Normalized for sequence length and read depth.

**Prodigal**

[10.1186/1471-2105-11-119](https://dx.doi.org/10.1186%2F1471-2105-11-119)

* Microbial gene prediction is a well studied, and some would say solved, problem
* we constructed a novel gene-finding algorithm called Prodigal
* This pipeline consisted of a combination of Critica [11] and Glimmer [1], BLAST [9] to locate missing genes and correct errors, and a final round of manual expert curation.
* Prodigal runs very quickly, analyzing a 4 MB genome in about 20 seconds on a typical workstation. It is also extremely easy to use relative to other methods, consisting of only a single executable that can be run without the user needing to supply any organism-specific parameters.

**CD-HIT**

<https://doi.org/10.1093/bioinformatics/bts565>

* CD-HIT is a very widely used program for clustering and comparing protein or nucleotide sequences
* CD-HIT is very fast and can handle extremely large databases. CD-HIT helps to significantly reduce the computational and manual efforts in many sequence analysis tasks and aids in understanding the data structure and correct the bias within a dataset
* Removing redundancy from such data by clustering could be crucial for reducing storage space, computational time and noise interference in some analysis methods, etc.
* CD-HIT is a greedy incremental algorithm that starts with the longest input sequence as the first cluster representative, and then process the remaining sequences from long to short to classify each sequence as a redundant or representative sequence based on its similarities to the existing representatives.
* The algorithm behind cd-hit is short word filtering, which can determine that the similarity between two sequences is below a certain value without performing an actual sequence alignment.
* sequence identity threshold, default 0.9 this is the default cd-hit's "global sequence identity" calculated as: number of identical amino acids in alignment divided by the full length of the shorter sequence
* alignment coverage for the shorter sequence, default 0.0 if set to 0.9, the alignment must covers 90% of the sequence