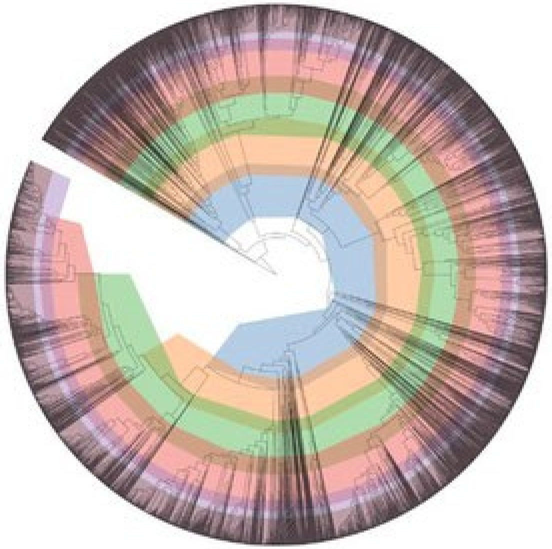
ENVIRONMENTAL METAGENOMICS

Physalia course, online, 13-17 October 2025

Metagenome de-novo assembly and quality control

Nikolay Oskolkov, Group Leader of Metabolic Research Group at LIOS, Riga, Latvia
Samuel Aroney, Postdoctoral Research Fellow, Queensland University of Technology



Physalia
Courses

NB: original course material courtesy:
Dr. Antti Karkman, University of Helsinki
Dr. Igor Pessi, Finnish Environment Institute (SYKE)

Typical analysis methods used in metagenomics

1) Alignment:

BOW TIE

BWA

stands for

Burrows Wheeler Aligner

ABBREVIATIONS

Abbreviations.com

HOPS

2) Classification:

Centrifuge

MetaPhlan

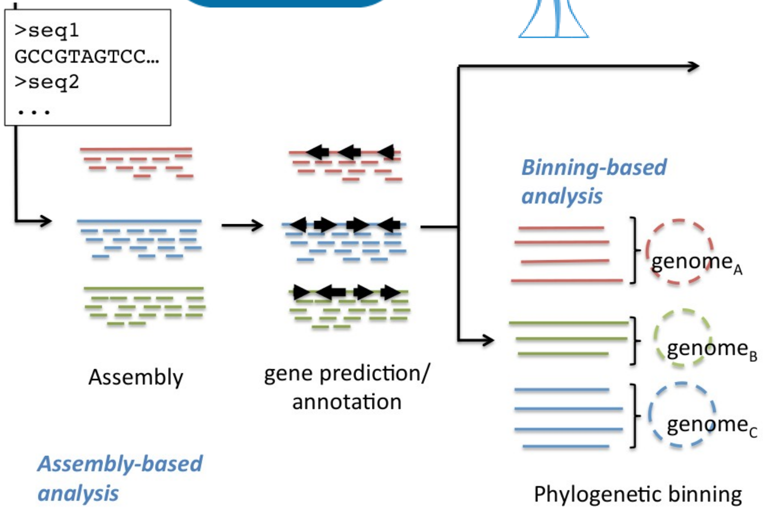
Clark

Reference based:
assume similarity to reference

3) De-novo assembly:

MEGA HIT

meta



```
>seq1
GCCGTAGTCC…
>seq2
...
```

Reference free:
unbiased but challenging

Strenghts and weaknesses of read-based metagenomics

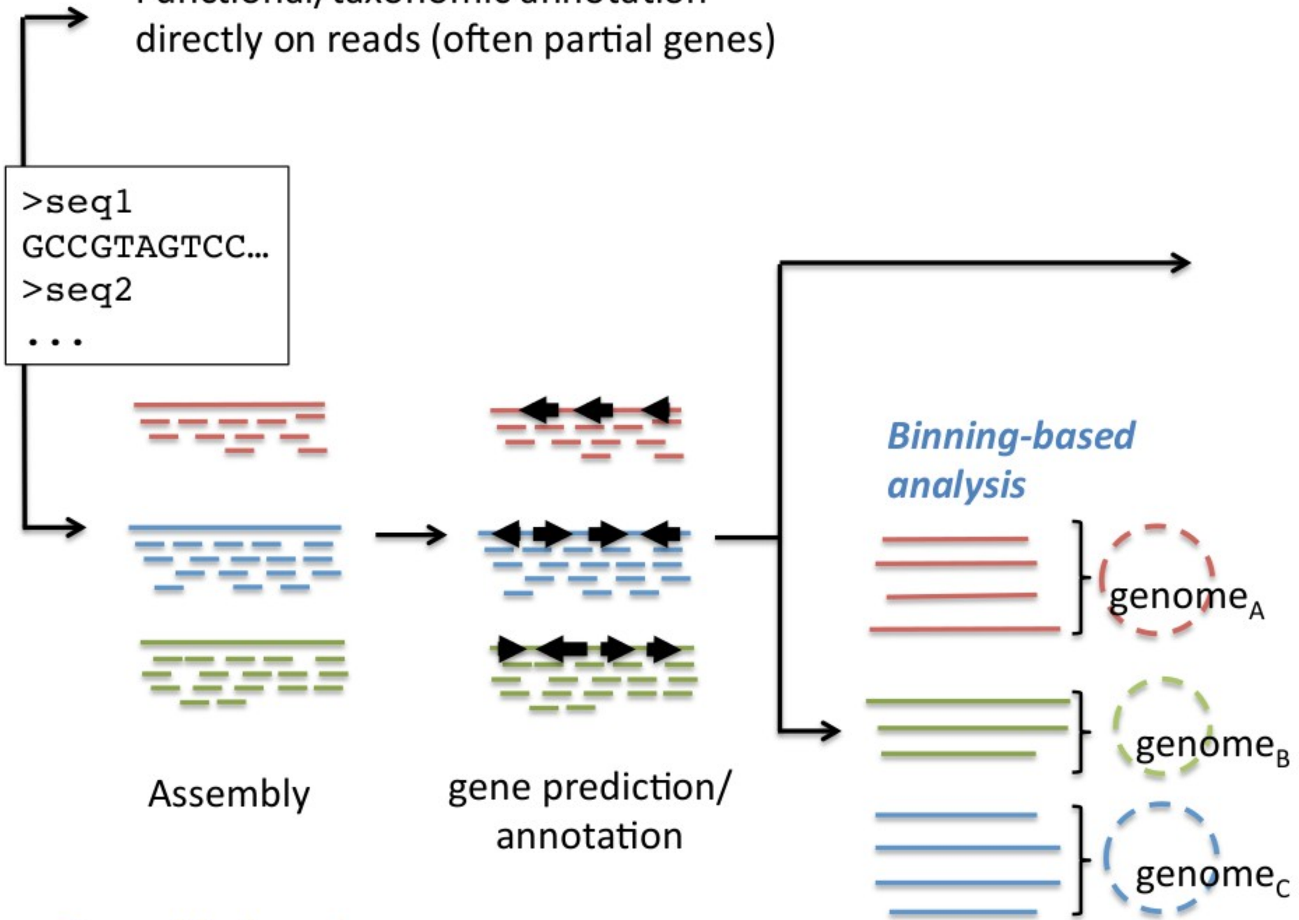
Comprehensiveness	Can provide an aggregate picture of community function or structure, but is based only on the fraction of reads that map effectively to reference dbs
Community complexity	Can deal with communities of arbitrary complexity given sufficient sequencing depth and satisfactory reference database coverage
Novelty	Cannot resolve organisms for which genomes of close relatives are unknown
Computational burden	Can be performed efficiently, enabling large meta-analyses
Genome-resolved metabolism	Can typically resolve only the aggregate metabolism of the community, and links with phylogeny are only possible in the context of known reference genomes
Expert manual supervision	Usually does not require manual curation, but selection of reference genomes to use could involve human supervision
Integration with microbial genomics	Obtained profiles cannot be directly put into the context of genomes derived from pure cultured isolates

Physalia
Courses

Summary metagenome analysis strategies

Read-based analysis

Functional/taxonomic annotation directly on reads (often partial genes)



```
>seq1
GCCGTAGTCC…
>seq2
...
```

Assembly

gene prediction/ annotation

Binning-based analysis

Assembly-based analysis

Phylogenetic binning

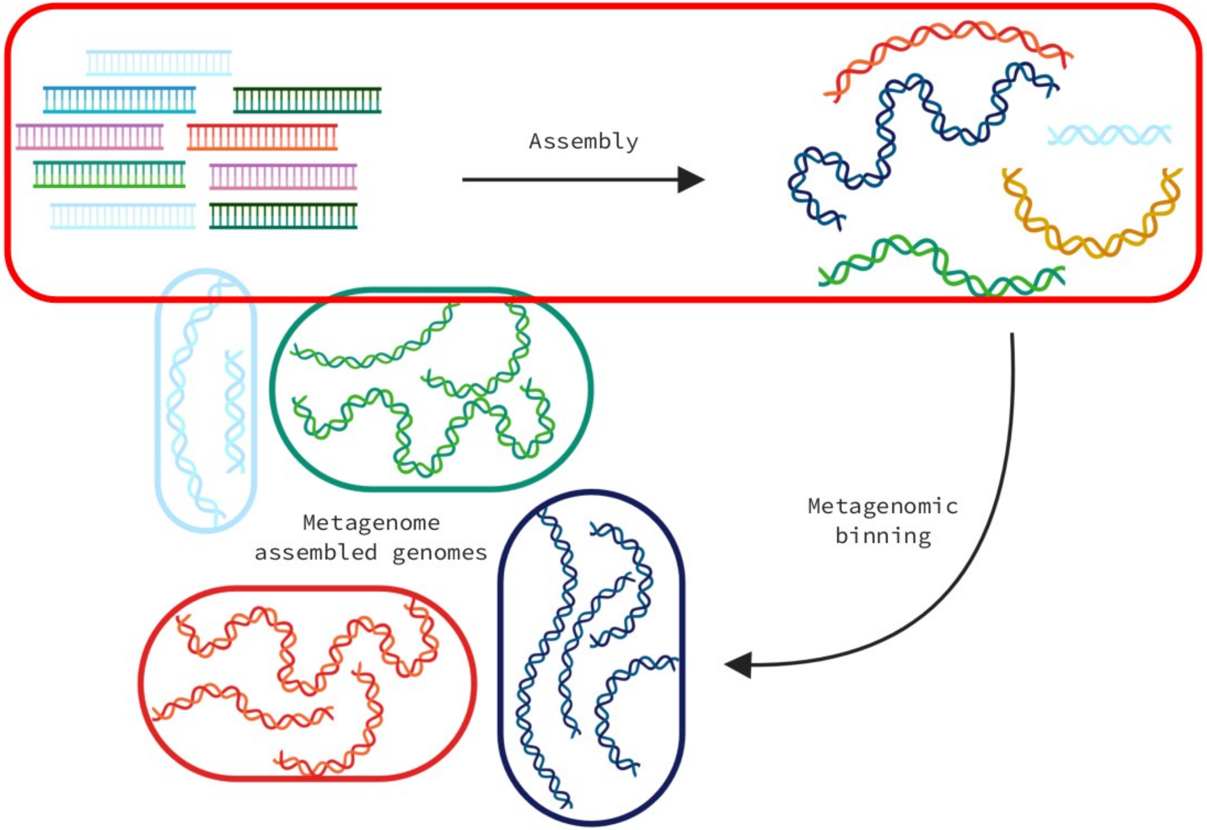
Strengths and weaknesses of assembly-based metagenomics

Comprehensiveness	Can construct multiple whole genomes, but only for organisms with enough coverage to be assembled and binned
Community complexity	In complex communities, only a fraction of the genomes can be resolved by assembly
Novelty	Can resolve genomes of entirely novel organisms with no sequenced relatives
Computational burden	Requires computationally costly assembly, mapping and binning
Genome-resolved metabolism	Can link metabolism to phylogeny through completely assembled genomes, even for novel diversity
Expert manual supervision	Manual curation required for accurate binning and scaffolding and for misassembly detection
Integration with microbial genomics	Assemblies can be fed into microbial genomic pipelines designed for analysis of genomes from pure cultured isolates

De novo assembly

Assemble short nucleotide sequences into longer sequences by finding their overlap / consensus without reference genome

- No reference available
- Uneven and complex communities



Physalia
Courses

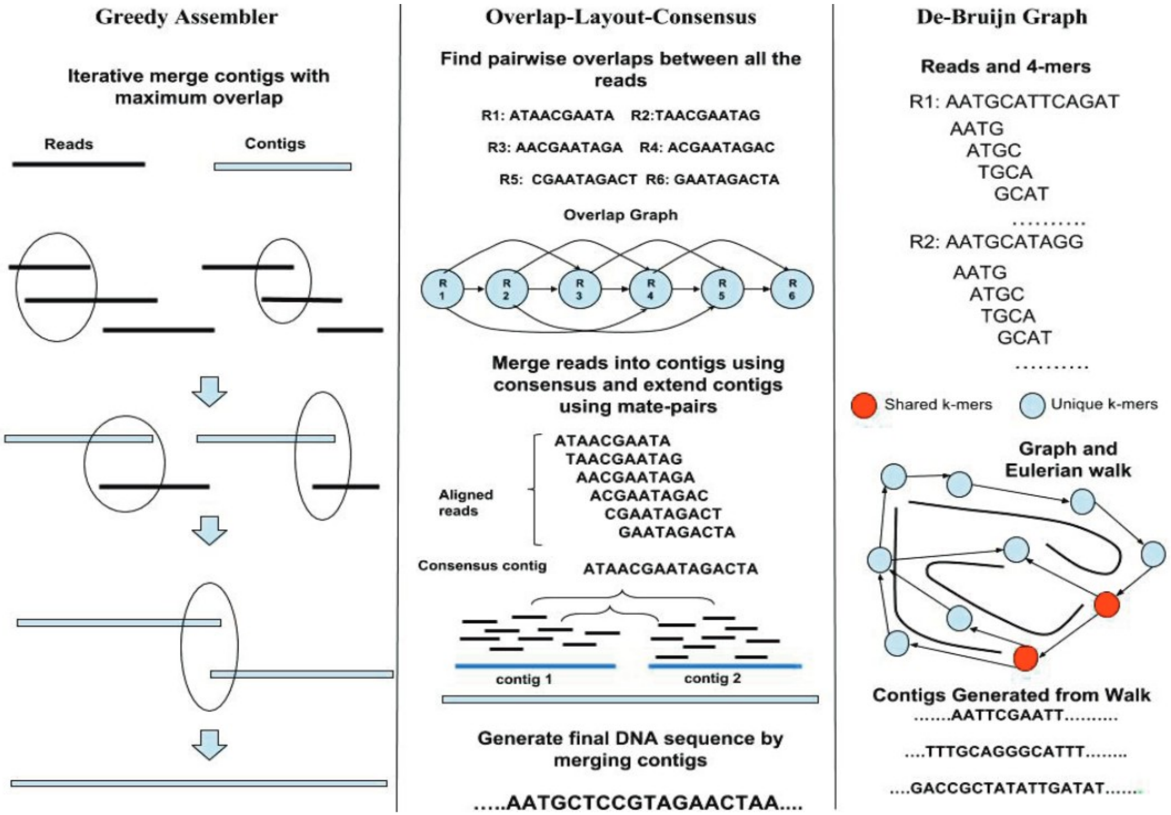

Greedy Assembler
Iterative merge contigs with maximum overlap
Reads
Contigs
Overlap-Layout-Consensus
Find pairwise overlaps between all the reads
R1: ATAACGAATA R2:TAACGAATAG
R3: AACGAATAGA R4: ACGAATAGAC
R5: CGAATAGACT R6: GAATAGACTA
Overlap Graph
R 1
R 2
R 3
R 4
R 5
R 6
Merge reads into contigs using consensus and extend contigs using mate-pairs
Aligned reads
ATAACGAATA
TAACGAATAG
AACGAATAGA
ACGAATAGAC
CGAATAGACT
GAATAGACTA
Consensus contig
ATAACGAATAGACTA
contig 1
contig 2
Generate final DNA sequence by merging contigs
.....AATGCTCCGTAGAACTAA....
De-Bruijn Graph
Reads and 4-mers
R1: AATGCATTCAGAT
AATG
ATGC
TGCA
GCAT
....
R2: AATGCATAGG
AATG
ATGC
TGCA
GCAT
....
Shared k-mers
Unique k-mers
Graph and Eulerian walk
Contigs Generated from Walk
.......AATTCGAATT..........
....TTTGCAGGGCATTT........
....GACCGCTATATTGATAT......

Figure 1: Overview of different de novo assembly paradigms. Schematic representation of the three main paradigms for genome assembly – Greedy, Overlap-Layout-Consensus, and de Bruijn. In Greedy assembler, reads with maximum overlaps are iteratively merged into contigs. In Overlap-Layout-Consensus approach, a graph is constructed by finding overlaps between all pairs of reads. This graph is further simplified and contigs are constructed by finding branch-less paths in the graph, and taking the consensus sequence of the overlapping reads implied by the corresponding paths. Contigs are further organized and extended using mate pair information. In de Bruijn graph assemblers, reads are chopped into short overlapping segments (k-mers) which are organized in a de Bruijn graph structure based on their co-occurrence across reads. The graph is simplified to remove artifacts due to sequencing errors, and branch-less paths are reported as contigs.

Ghurye et al., Yale Journal of Biology and Medicine, 2016

SOAPdenovo
Velvet
Spades

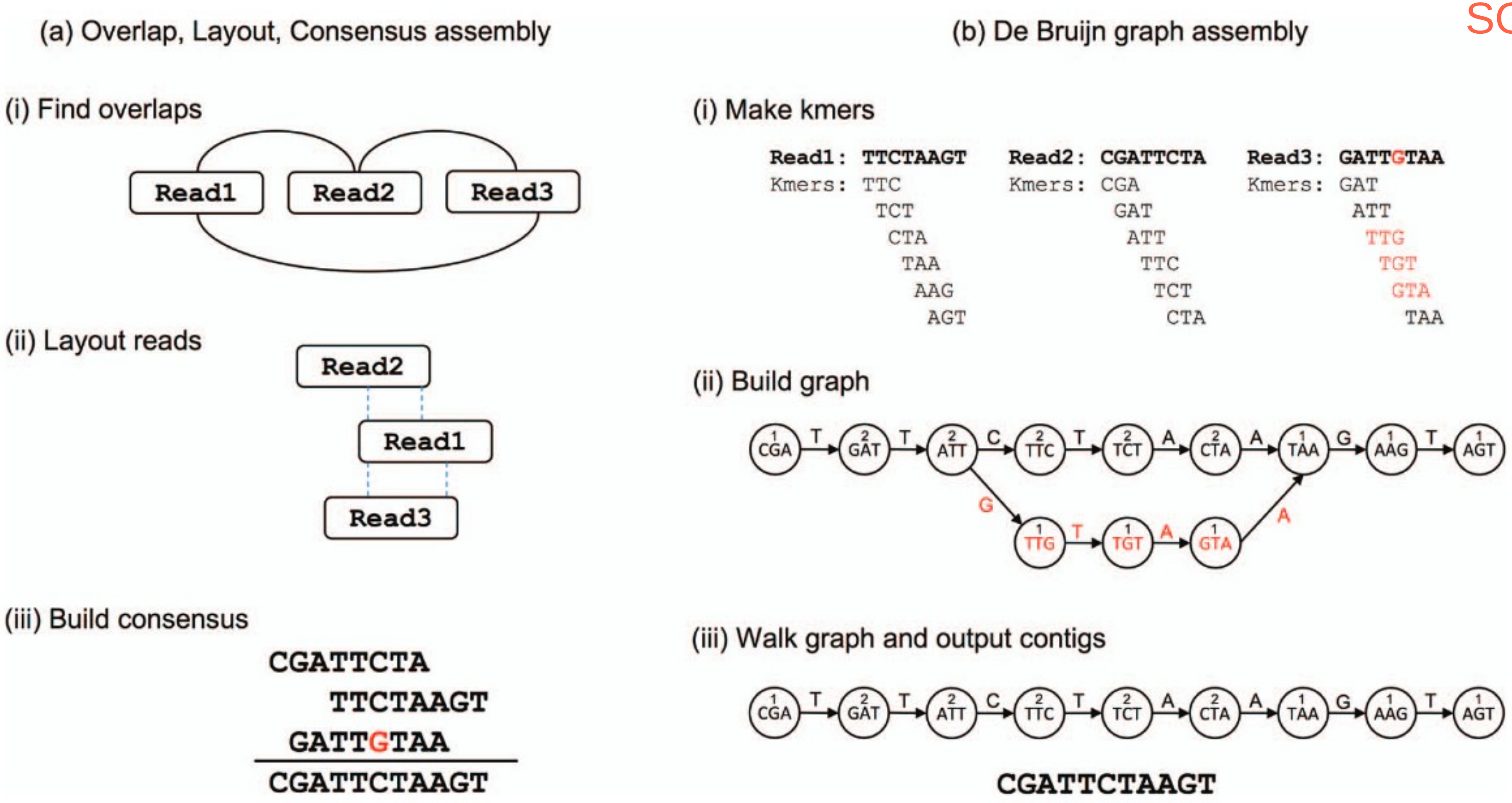
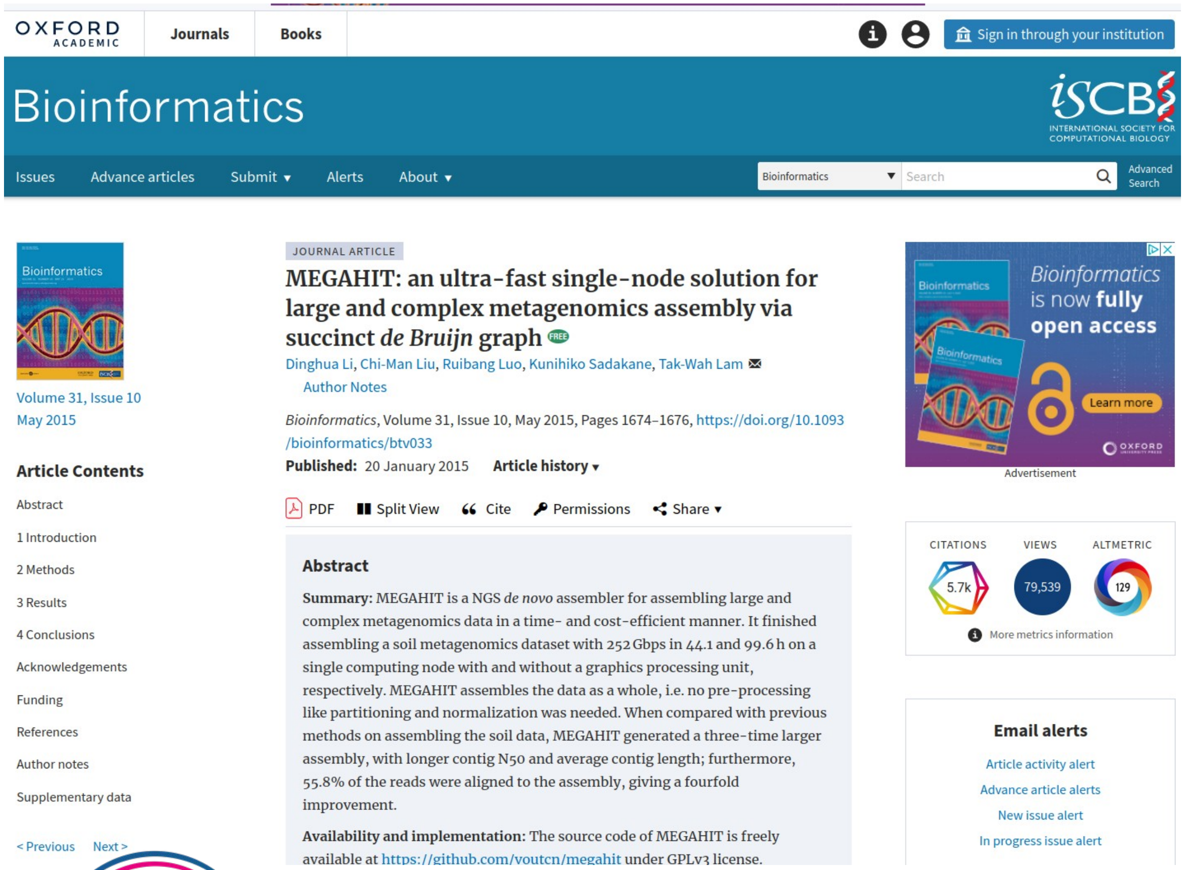


Figure 1. Two different approaches to genome assembly: (**a**) in Overlap, Layout, Consensus assembly, (i) overlaps are found between reads and an overlap graph constructed (edges indicate overlapping reads). (ii) Reads are laid out into contigs based on the overlaps (dashed lines indicate overlapping portions). (iii) The most likely sequence is chosen to construct consensus sequence. (**b**) In dBg assembly, (i) reads are decomposed into kmers by sliding a window of size *k* across the reads. (ii) The kmers become vertices in the dBg, with edges connecting overlapping kmers. Polymorphisms (red) form branches in the graph. A count is kept of how many times a kmer is seen, shown here as numbers above kmers. (iii) Contigs are built by walking the graph from edge nodes. A variety of heuristics handle branches in the graphs—for example, low coverage paths, as shown here, may be ignored.

Ayling et al., Briefings in Bioinformatics, 2020

Popular de Bruijn graph *de-novo* metagenomic assemblers for short Illumina reads



Volume 31, Issue 10
May 2015

Article Contents

Abstract
1 Introduction
2 Methods
3 Results
4 Conclusions
Acknowledgements
Funding
References
Author notes
Supplementary data

< Previous Next >

JOURNAL ARTICLE

MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct *de Bruijn* graph FREE

Dinghua Li, Chi-Man Liu, Ruibang Luo, Kunihiko Sadakane, Tak-Wah Lam ✉
Author Notes

Bioinformatics, Volume 31, Issue 10, May 2015, Pages 1674–1676, https://doi.org/10.1093/bioinformatics/btv033

Published: 20 January 2015 **Article history** ▾

Abstract

Summary: MEGAHIT is a NGS *de novo* assembler for assembling large and complex metagenomics data in a time- and cost-efficient manner. It finished assembling a soil metagenomics dataset with 252 Gbps in 44.1 and 99.6 h on a single computing node with and without a graphics processing unit, respectively. MEGAHIT assembles the data as a whole, i.e. no pre-processing like partitioning and normalization was needed. When compared with previous methods on assembling the soil data, MEGAHIT generated a three-time larger assembly, with longer contig N50 and average contig length; furthermore, 55.8% of the reads were aligned to the assembly, giving a fourfold improvement.

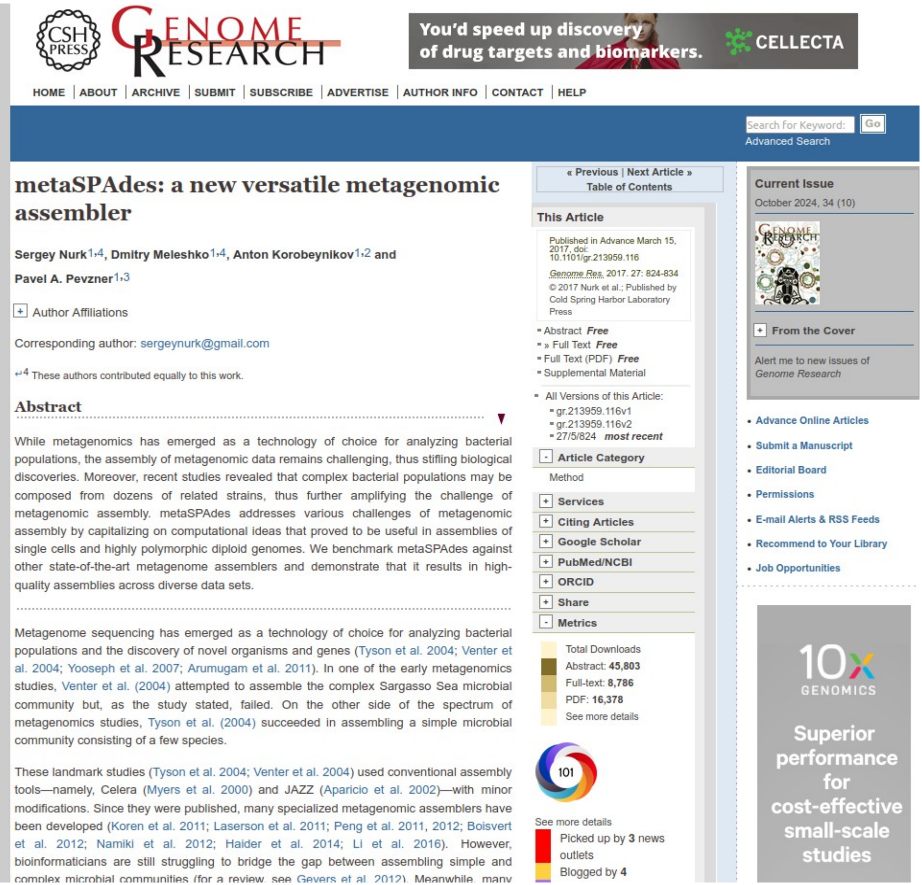
Availability and implementation: The source code of MEGAHIT is freely available at https://github.com/voutcn/megahit under GPLv3 license.

CITATIONS	VIEWS	ALTMETRIC
5.7k	79,539	129

More metrics information

Email alerts

Article activity alert
Advance article alerts
New issue alert
In progress issue alert



CSH PRESS
GENOME RESEARCH
HOME | ABOUT | ARCHIVE | SUBMIT | SUBSCRIBE | ADVERTISE | AUTHOR INFO | CONTACT | HELP
Search for Keyword: Go
Advanced Search

metaSPAdes: a new versatile metagenomic assembler

Sergey Nurk[1,4], Dmitry Meleshko[1,4], Anton Korobeynikov[1,2] and Pavel A. Pevzner[1,3]

+ Author Affiliations

Corresponding author: sergeynurk@gmail.com

[4] These authors contributed equally to this work.

Abstract

While metagenomics has emerged as a technology of choice for analyzing bacterial populations, the assembly of metagenomic data remains challenging, thus stifling biological discoveries. Moreover, recent studies revealed that complex bacterial populations may be composed from dozens of related strains, thus further amplifying the challenge of metagenomic assembly. metaSPAdes addresses various challenges of metagenomic assembly by capitalizing on computational ideas that proved to be useful in assemblies of single cells and highly polymorphic diploid genomes. We benchmark metaSPAdes against other state-of-the-art metagenome assemblers and demonstrate that it results in high-quality assemblies across diverse data sets.

Metagenome sequencing has emerged as a technology of choice for analyzing bacterial populations and the discovery of novel organisms and genes (Tyson et al. 2004; Venter et al. 2004; Yooseph et al. 2007; Arumugam et al. 2011). In one of the early metagenomics studies, Venter et al. (2004) attempted to assemble the complex Sargasso Sea microbial community but, as the study stated, failed. On the other side of the spectrum of metagenomics studies, Tyson et al. (2004) succeeded in assembling a simple microbial community consisting of a few species.

These landmark studies (Tyson et al. 2004; Venter et al. 2004) used conventional assembly tools—namely, Celera (Myers et al. 2000) and JAZZ (Aparicio et al. 2002)—with minor modifications. Since they were published, many specialized metagenomic assemblers have been developed (Koren et al. 2011; Laserson et al. 2011; Peng et al. 2011, 2012; Boisvert et al. 2012; Namiki et al. 2012; Haider et al. 2014; Li et al. 2016). However, bioinformaticians are still struggling to bridge the gap between assembling simple and complex microbial communities (for a review, see Gevers et al. 2012). Meanwhile, many

« Previous | Next Article »
Table of Contents

This Article

Published in Advance March 15, 2017, doi: 10.1101/gr.213959.116

Genome Res. 2017. 27: 824-834

© 2017 Nurk et al.; Published by Cold Spring Harbor Laboratory Press

- Abstract *Free*
- » Full Text *Free*
- Full Text (PDF) *Free*
- Supplemental Material

- All Versions of this Article:
  - gr.213959.116v1
  - gr.213959.116v2
  - 27/5/824 *most recent*

- Article Category

Method

+ Services
+ Citing Articles
+ Google Scholar
+ PubMed/NCBI
+ ORCID
+ Share
- Metrics

Total Downloads	
Abstract:	45,803
Full-text:	8,786
PDF:	16,378
See more details	

101

See more details

Picked up by 3 news outlets
Blogged by 4

Current Issue

October 2024, 34 (10)

+ From the Cover

Alert me to new issues of *Genome Research*

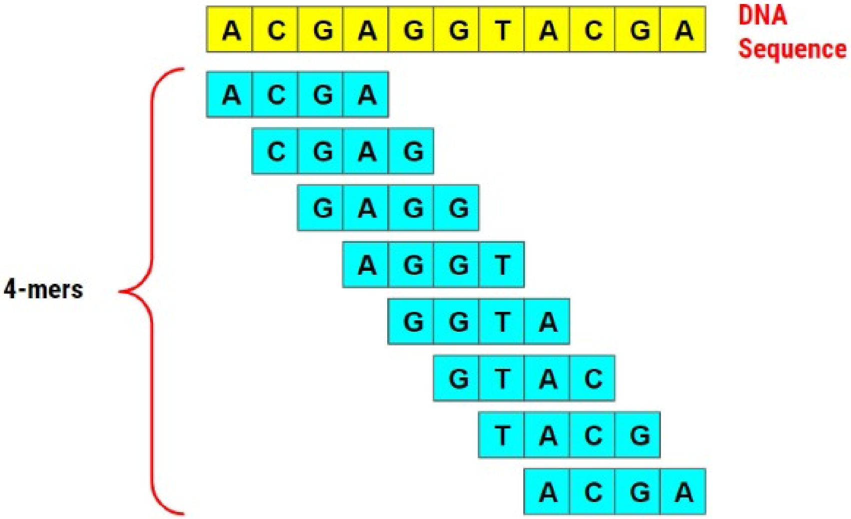
- Advance Online Articles
- Submit a Manuscript
- Editorial Board
- Permissions
- E-mail Alerts & RSS Feeds
- Recommend to Your Library
- Job Opportunities

Physalia Courses

We are going to use Megahit in the exercises

De novo assembly using MEGAHIT

MEGAHIT: de Bruijn-graph assembler using a distribution of different k-mer lengths inferred from the length of the sequencing data



A C G A G G T A C G A

A C G A
C G A G
G A G G
A G G T
G G T A
G T A C
T A C G
A C G A

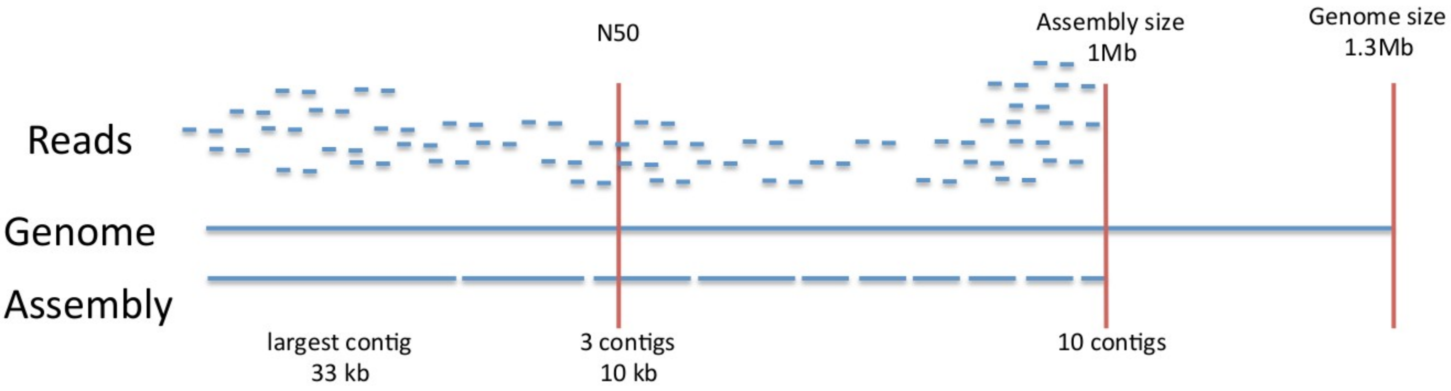
reasons for using MEGAHIT:

- **low-memory** footprint
- has little issues with the **presence of ancient DNA damage**
- works with **single-end data**

BUT: lower assembly quality than other assemblers for modern sequencing data (see CAMI II challenge; DOI: 10.1038/s41592-022-01431-4)

Assembly metrics

- assembly size
- number of contigs, largest contig
- N50

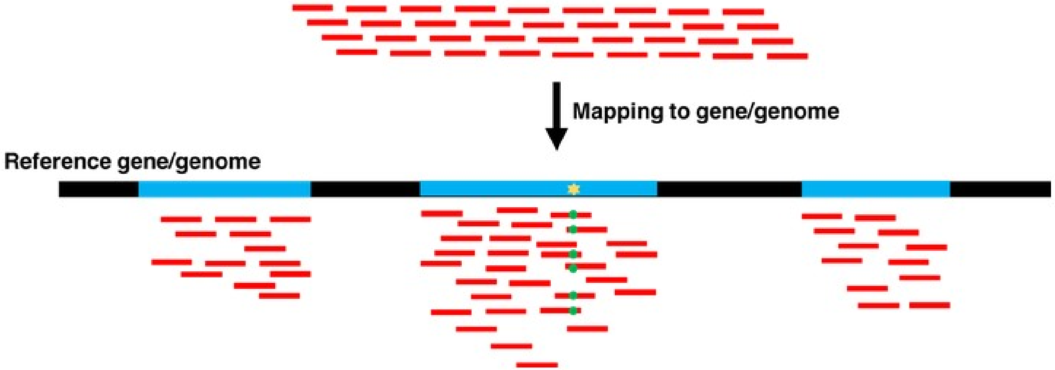


Physalia Courses

Alignment against the contigs

Many of the following steps require the alignment of the short-read data against the de novo assembled contigs, e.g.

- correction of the contig sequences
- binning of the contigs into MAGs (coverage along the contigs)
- quantification of the presence of ancient DNA damage



Taxonomic classification - on contig level

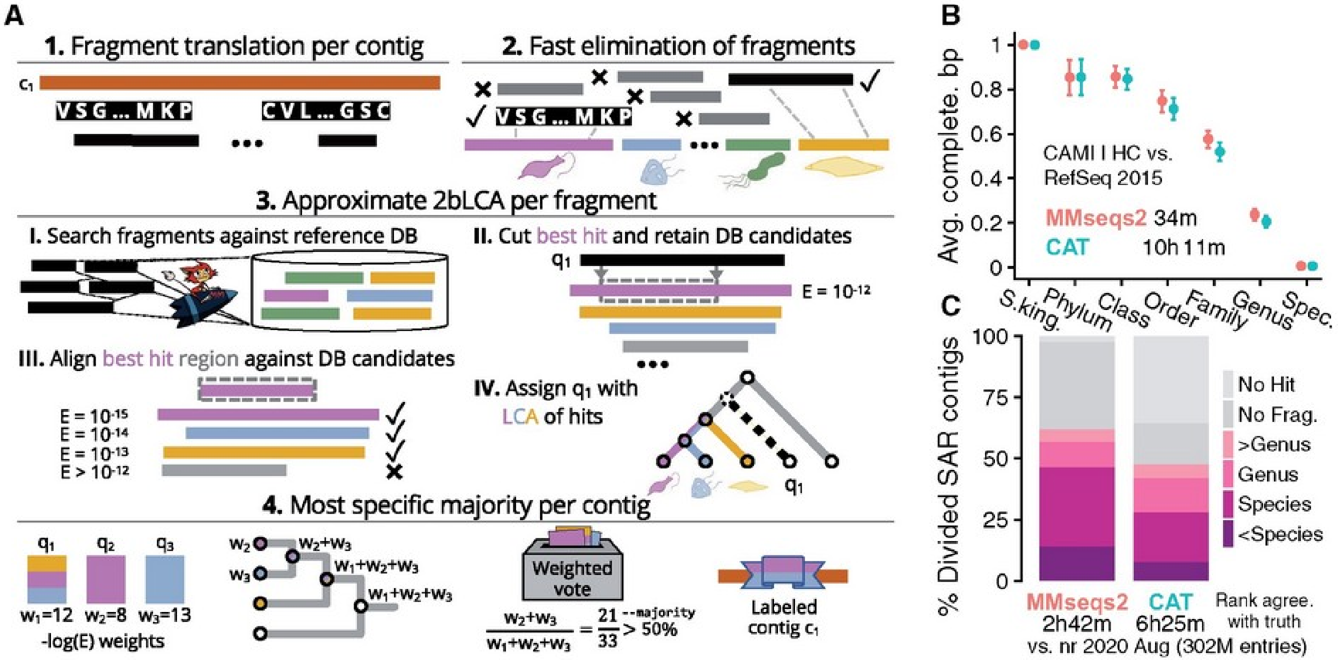
The likely taxonomic origin of contigs can be determined by aligning them against a reference database.

available aligners:

- BLAST/DIAMOND
- Kraken2
- Centrifuge
- MMSeqs2

available databases:

- NCBI NT/RefSeq
- GTDB



Mirdita *et al.* (Bioinformatics, 2021; doi: bioinformatics/btab184) Fig. 1

Physalia Courses