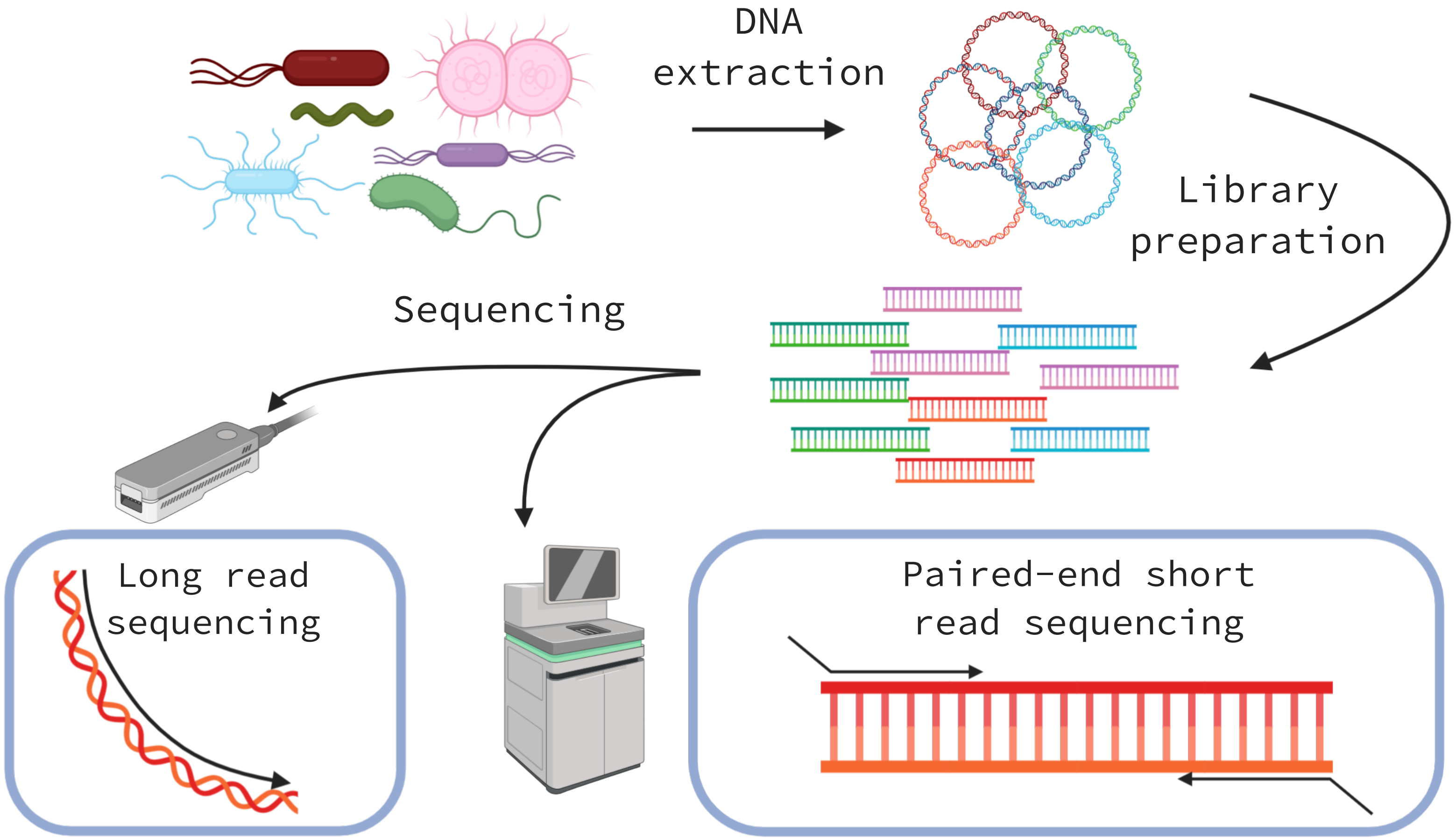
Environmental metagenomics

Metagenome assembly

Physalia Courses

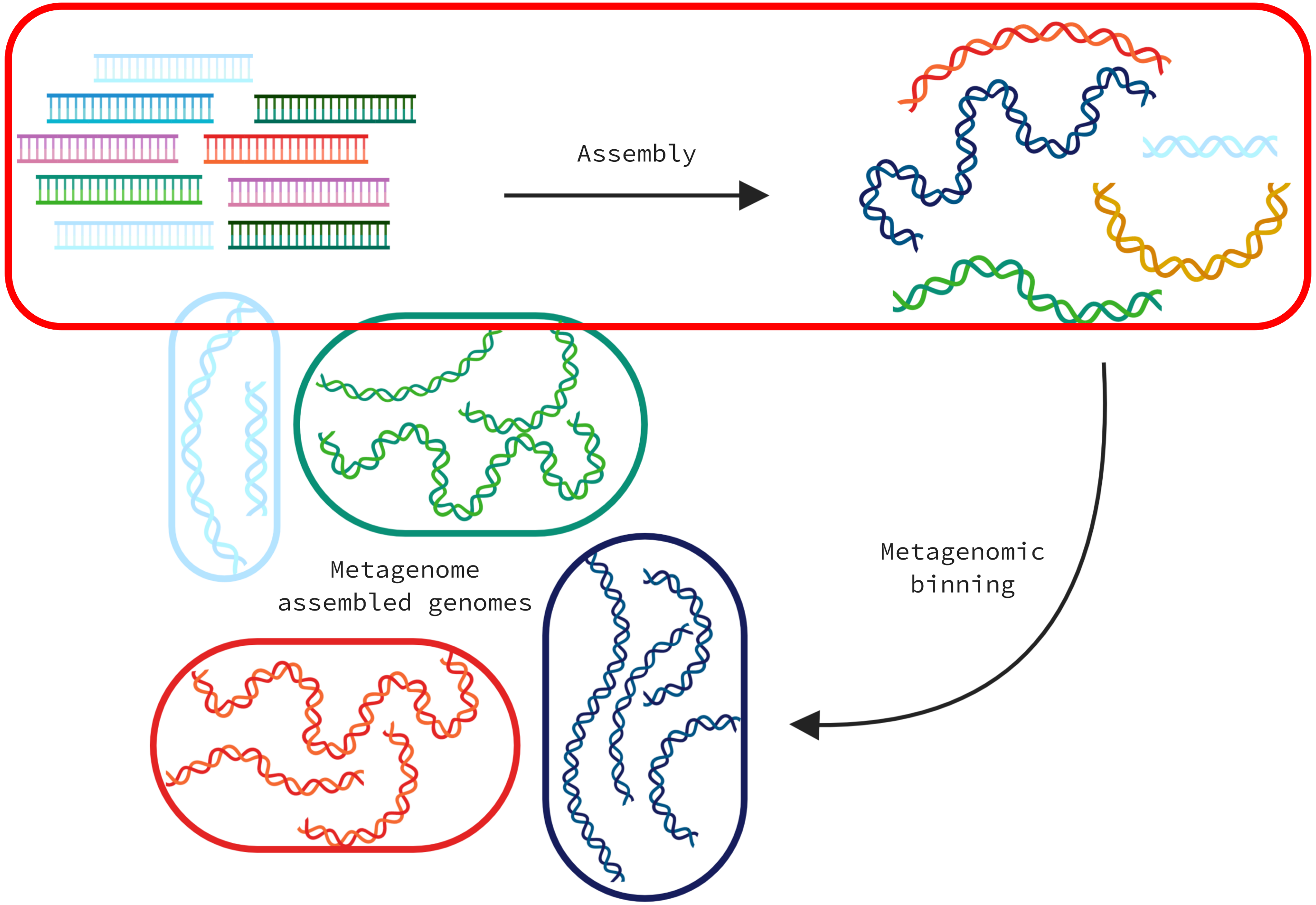
Igor S. Pessi & Antti Karkman

From samples to sequences

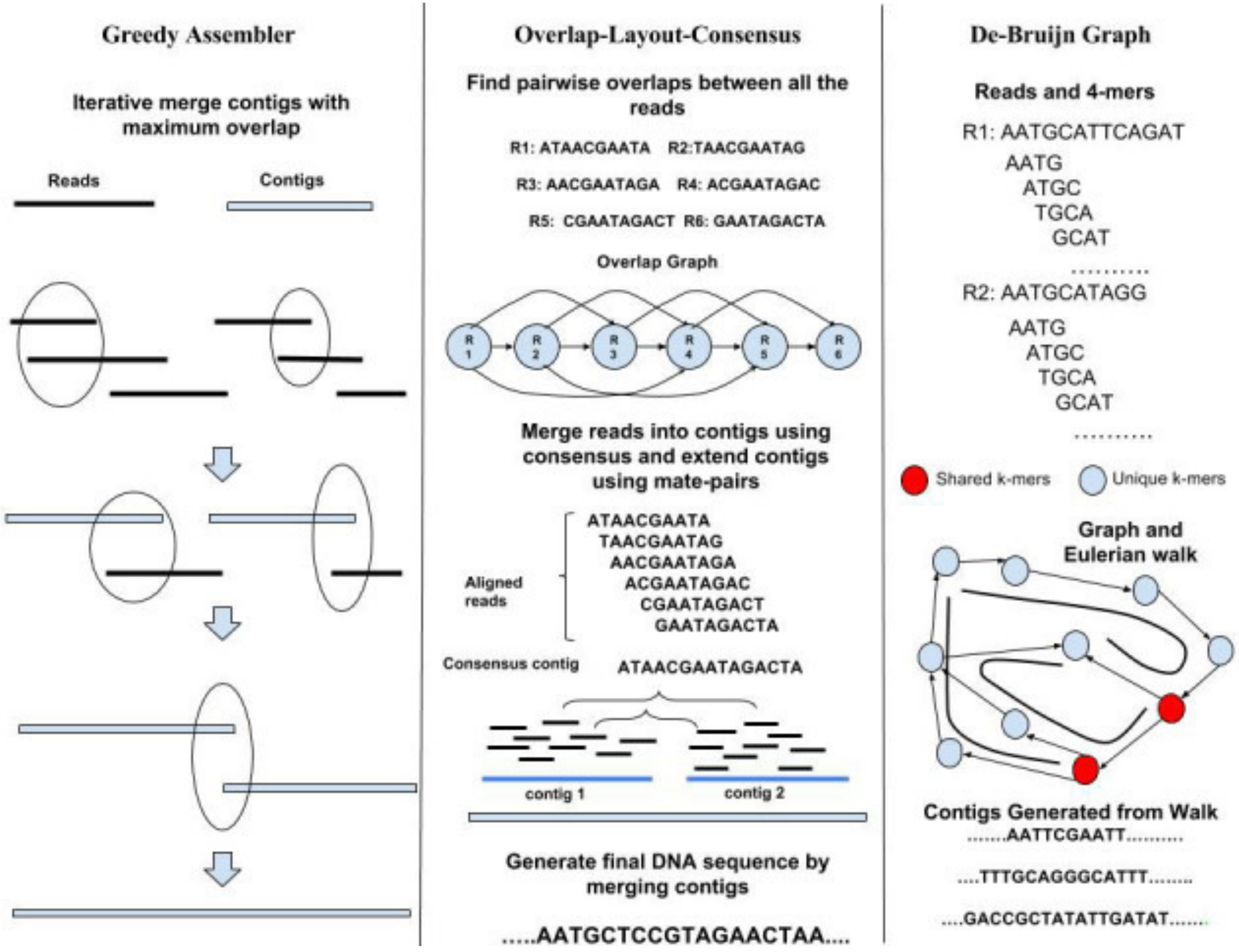


De novo assembly

- No reference available
- Uneven and complex communities

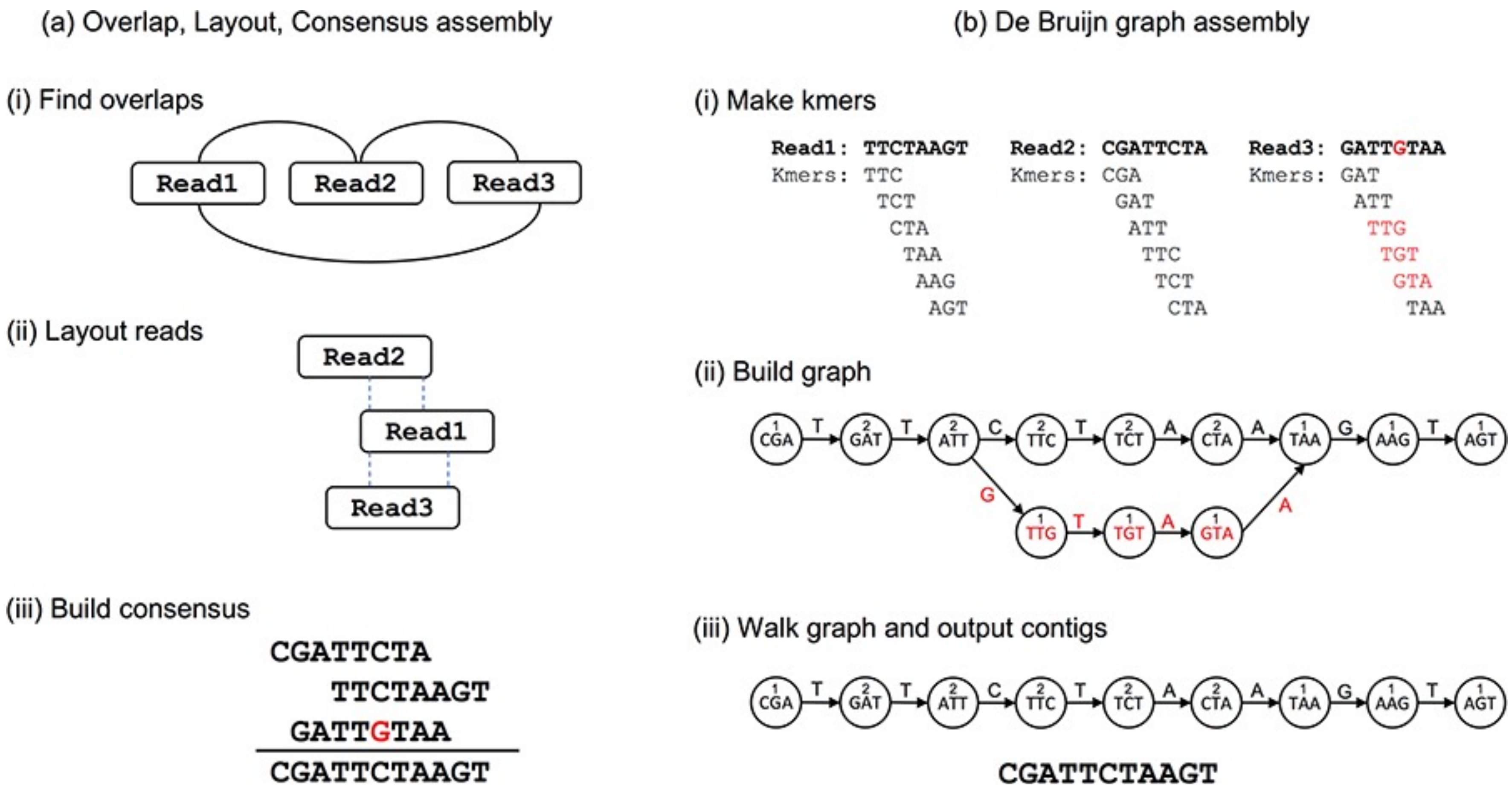


Assembly strategies


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.......

Ghurye et al. 2016. Yale J Biol Med.

Assembly strategies


(a) Overlap, Layout, Consensus assembly
(i) Find overlaps
Read1
Read2
Read3
(ii) Layout reads
Read2
Read1
Read3
(iii) Build consensus
CGATTCTA
TTCTAAGT
GATTGTAA
CGATTCTAAGT
(b) De Bruijn graph assembly
(i) Make kmers
Read1: TTCTAAGT
Kmers: TTC TCT CTA TAA AAG AGT
Read2: CGATTCTA
Kmers: CGA GAT ATT TTC TCT CTA
Read3: GATTGTAA
Kmers: GAT ATT TTG TGT GTA TAA
(ii) Build graph
1 CGA, T, 2 GAT, T, 2 ATT, C, 2 TTC, T, 2 TCT, A, 2 CTA, A, 1 TAA, G, 1 AAG, T, 1 AGT
G, 1 TTG, T, 1 TGT, A, 1 GTA, A
(iii) Walk graph and output contigs
1 CGA, T, 2 GAT, T, 2 ATT, C, 2 TTC, T, 2 TCT, A, 2 CTA, A, 1 TAA, G, 1 AAG, T, 1 AGT
CGATTCTAAGT

Ayling et al. 2020. Brief Bioinform,

Which assembler to choose?

Which assembler to choose?

- Type of data
  - Long-read
  - Short-read

Which assembler to choose?

- Type of data
  - Long-read
  - Short-read
- Sequencing instrument

Which assembler to choose?

- **Availability**: The tool should be freely available either as download or webserver.
- **Usability:** The tool should have a proper manual, readme file or help function describing how to use it. In case of problems, the respective authors were contacted.
- **Adoption:** The tool should be widely used, or show potential of being widely adopted in the future.
- Reference: Lindgreen et al. 2016. Sci Rep.

Assembly polishing

- Assemblies from noisy reads can contain errors

Assembly polishing

- Assemblies from noisy reads can contain errors
- Several methods for polishing:
  - Better quality sequence data from same samples
    - `pilon, racon`
  - Using the same raw noisy reads to polish the assembly
    - `medaka, racon`
  - Reference-based frame correction
    - `Proovframe`

Assembly polishing

- Assemblies from noisy reads can contain errors
- Several methods for polishing:
    - Better quality sequence data from same samples
        - `pilon`
    - Using the same raw noisy reads to polish the assembly
        - `medaka, racon`
    - Reference-based frame correction
        - `Proovframe`
- Also assemblers might do polishing
    - `metaflye`

Assembly QC

Assembly QC

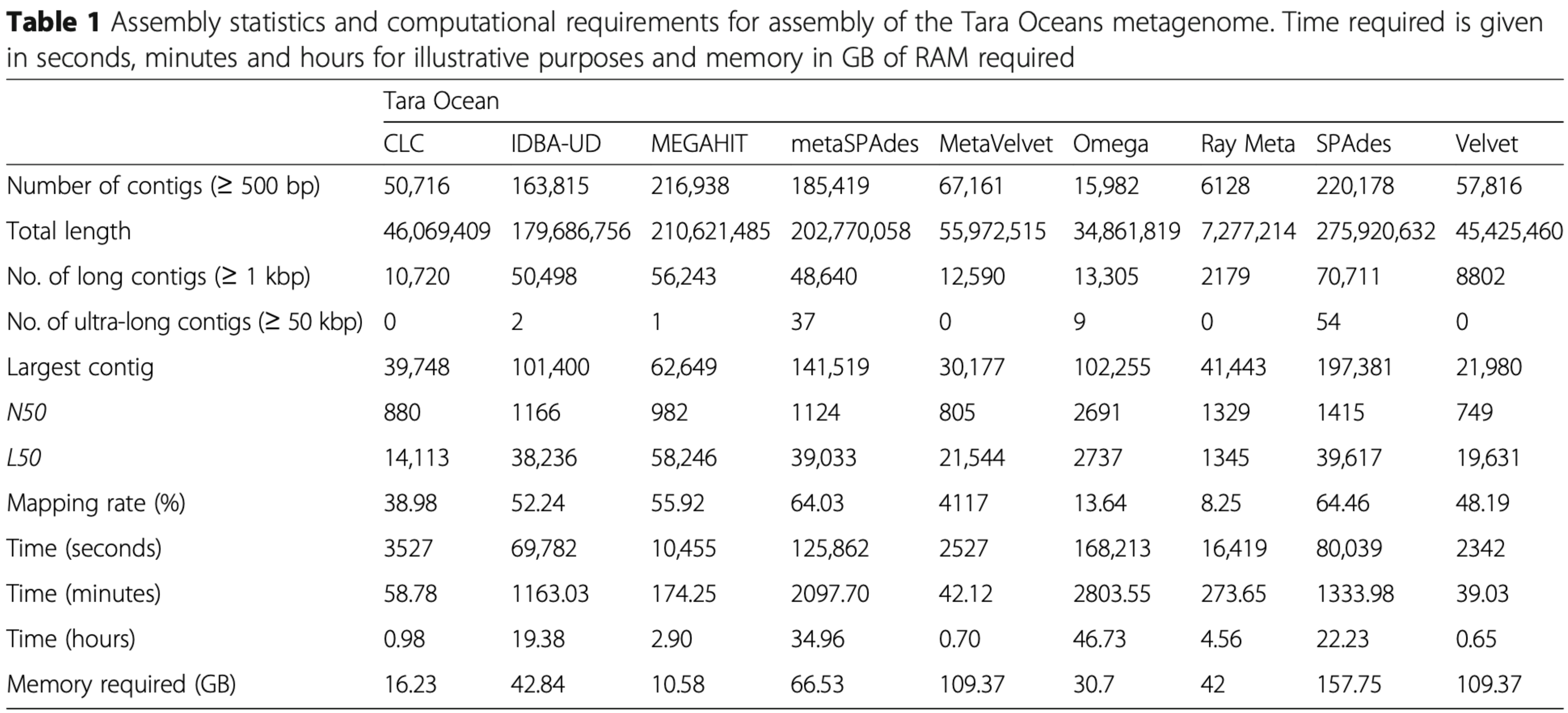


Table 1 Assembly statistics and computational requirements for assembly of the Tara Oceans metagenome. Time required is given in seconds, minutes and hours for illustrative purposes and memory in GB of RAM required

	Tara Ocean								
	CLC	IDBA-UD	MEGAHIT	metaSPAdes	MetaVelvet	Omega	Ray Meta	SPAdes	Velvet
Number of contigs (≥ 500 bp)	50,716	163,815	216,938	185,419	67,161	15,982	6128	220,178	57,816
Total length	46,069,409	179,686,756	210,621,485	202,770,058	55,972,515	34,861,819	7,277,214	275,920,632	45,425,460
No. of long contigs (≥ 1 kbp)	10,720	50,498	56,243	48,640	12,590	13,305	2179	70,711	8802
No. of ultra-long contigs (≥ 50 kbp)	0	2	1	37	0	9	0	54	0
Largest contig	39,748	101,400	62,649	141,519	30,177	102,255	41,443	197,381	21,980
N50	880	1166	982	1124	805	2691	1329	1415	749
L50	14,113	38,236	58,246	39,033	21,544	2737	1345	39,617	19,631
Mapping rate (%)	38.98	52.24	55.92	64.03	4117	13.64	8.25	64.46	48.19
Time (seconds)	3527	69,782	10,455	125,862	2527	168,213	16,419	80,039	2342
Time (minutes)	58.78	1163.03	174.25	2097.70	42.12	2803.55	273.65	1333.98	39.03
Time (hours)	0.98	19.38	2.90	34.96	0.70	46.73	4.56	22.23	0.65
Memory required (GB)	16.23	42.84	10.58	66.53	109.37	30.7	42	157.75	109.37

Open questions in metagenomic assembly

- To co-assemble or not?
- No reference – How to define a good assembly?
- Challenging elements for assembly
  - Repeat regions
  - Horizontally transferred genes
  - Extrachromosomal elements
- Others?