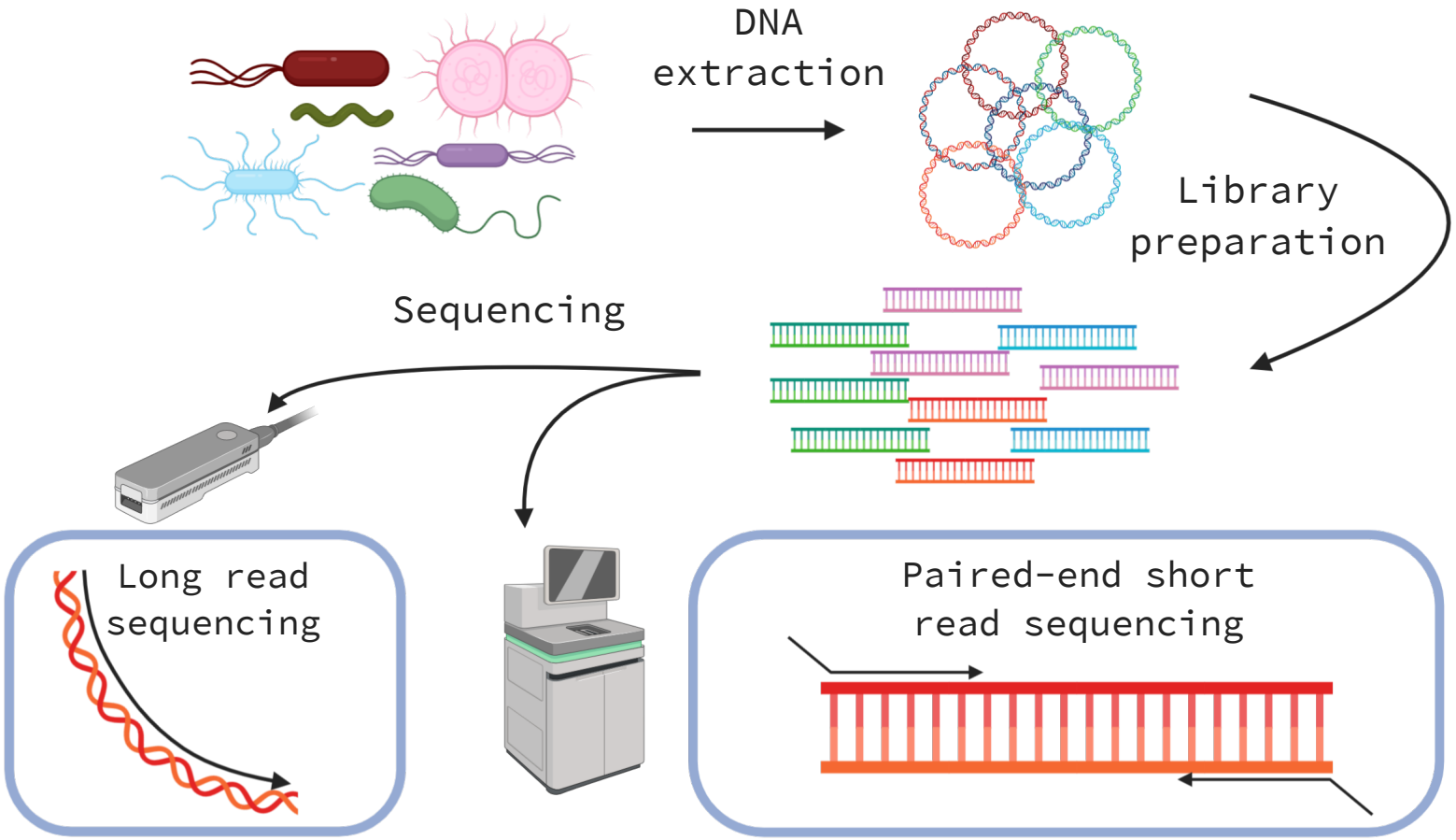
Environmental metagenomics

Metagenome assembly

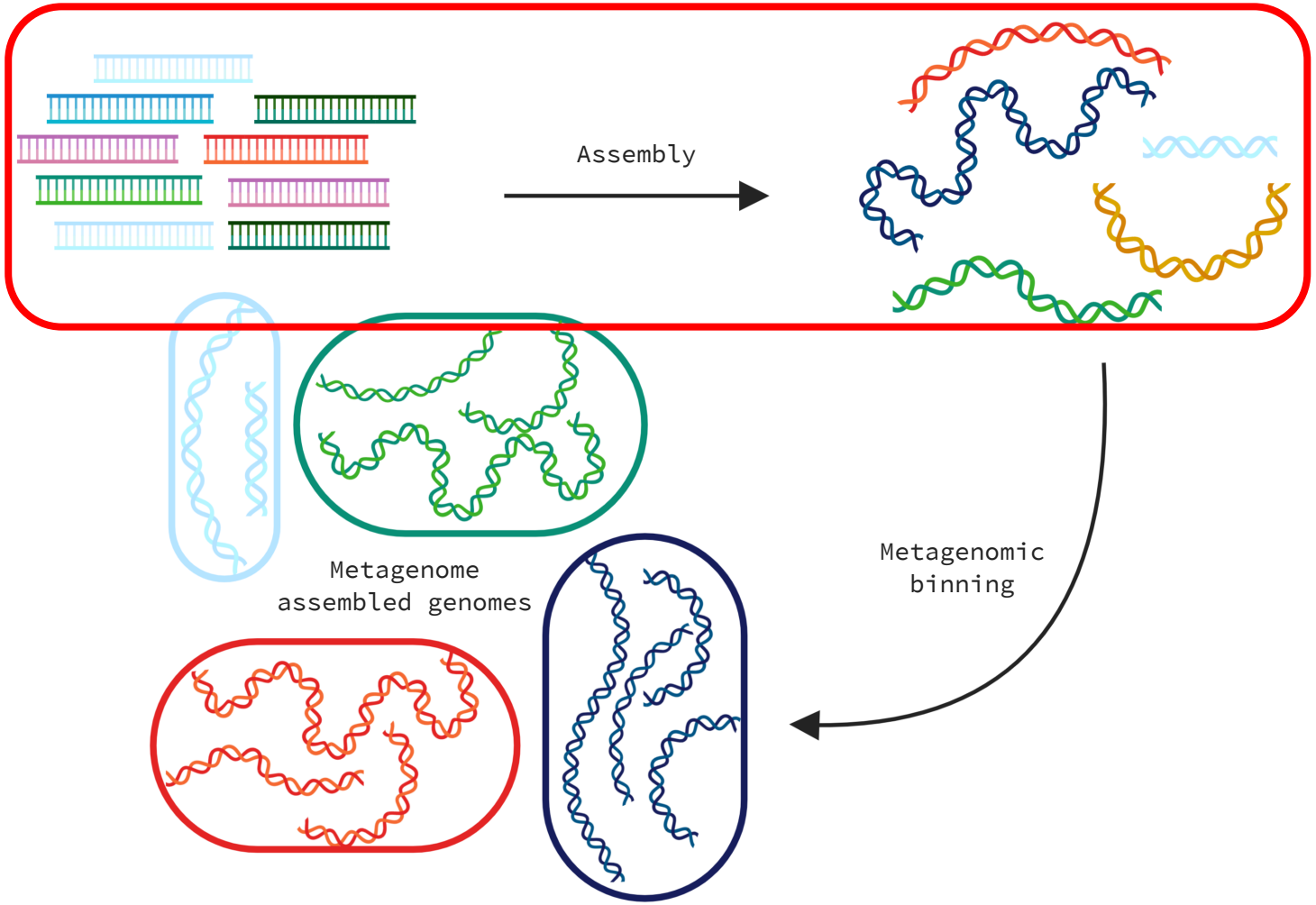
Igor S. Pessi & Antti Karkman, University of Helsinki

From samples to sequences

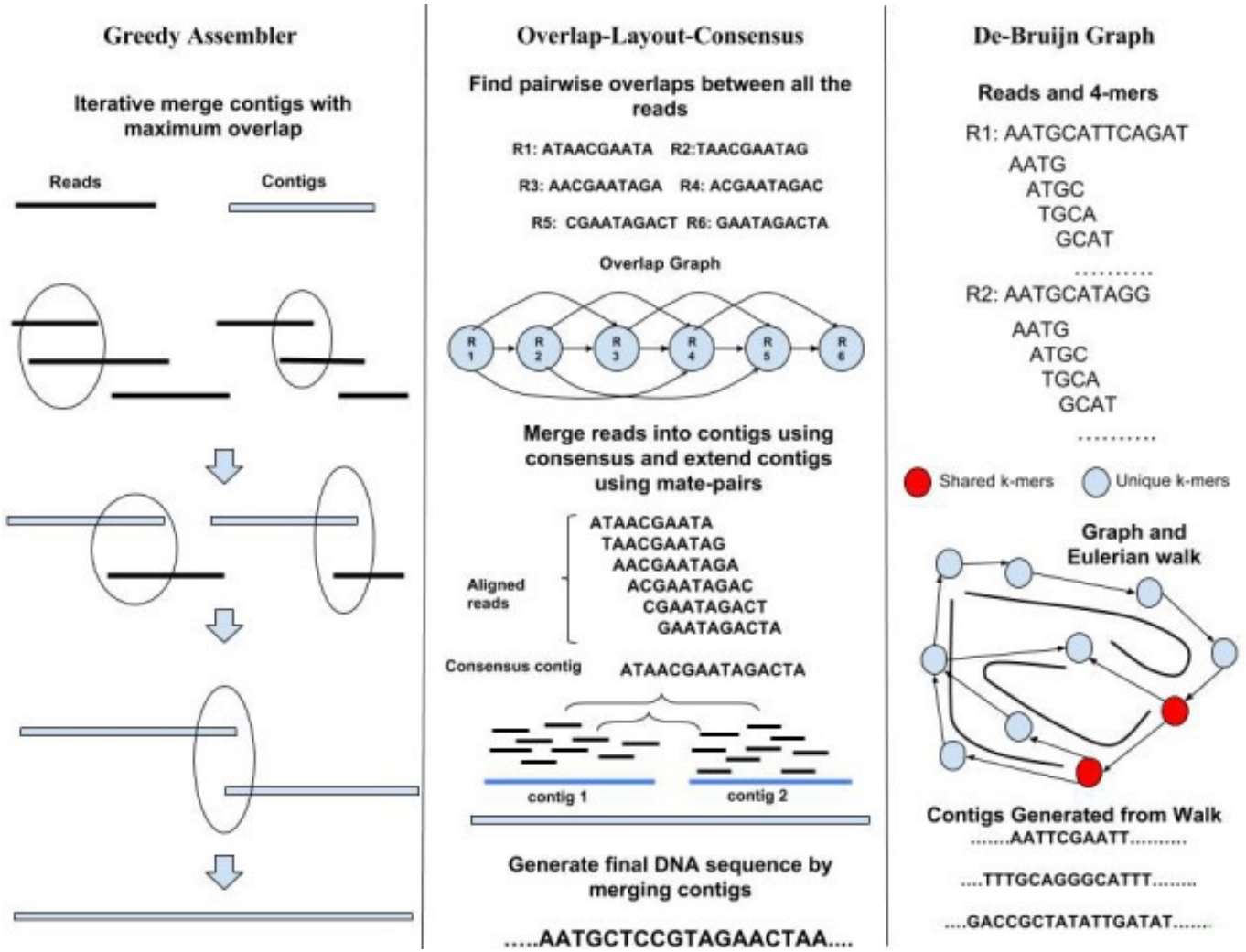


De novo assembly

- Uneven and complex communities
- No reference available

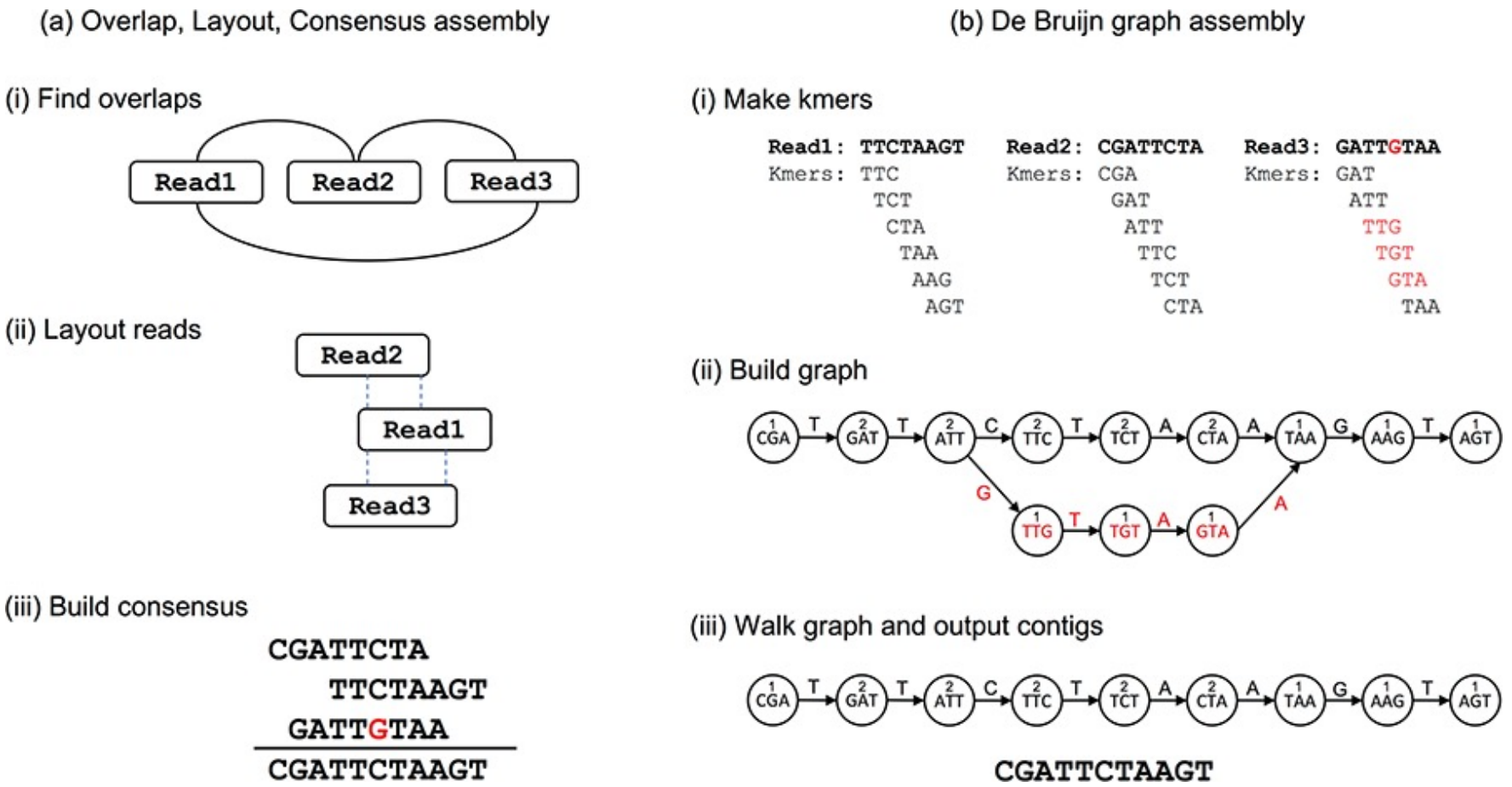


Assembly strategies



Ghurye et al. 2016. Yale J Biol Med.

Assembly strategies

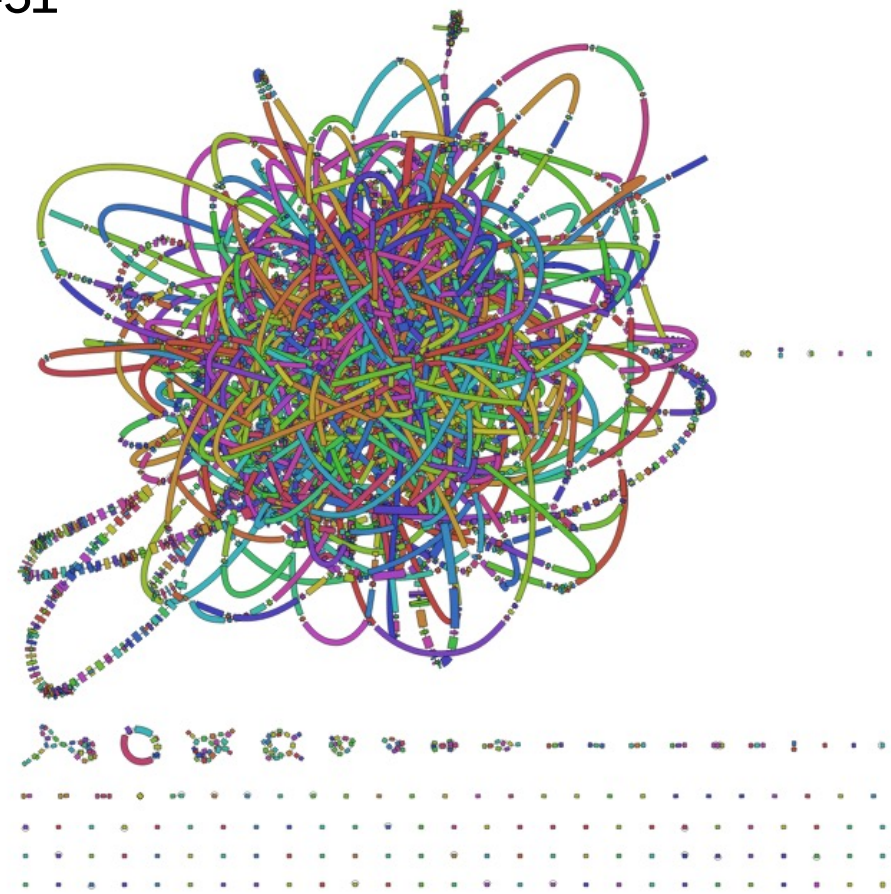


Ayling et al. 2020. Brief Bioinform,

Choice of kmer(s)?

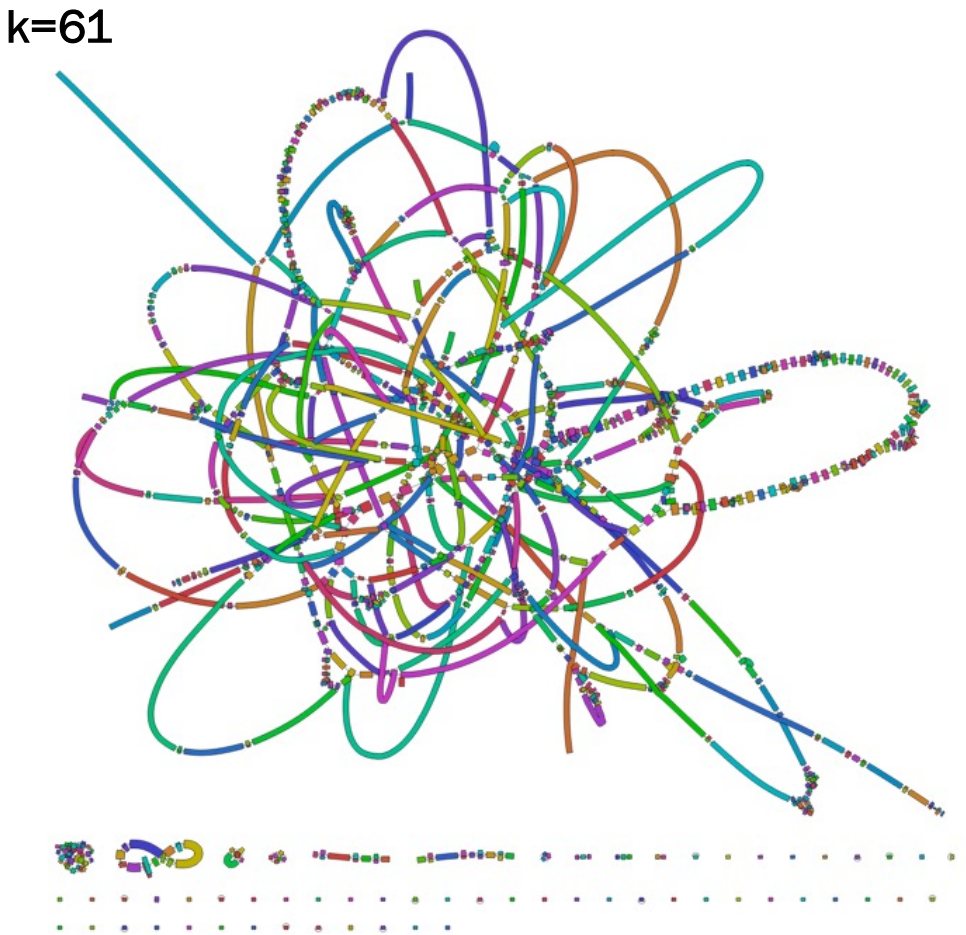
Choice of kmer(s)?

k=51



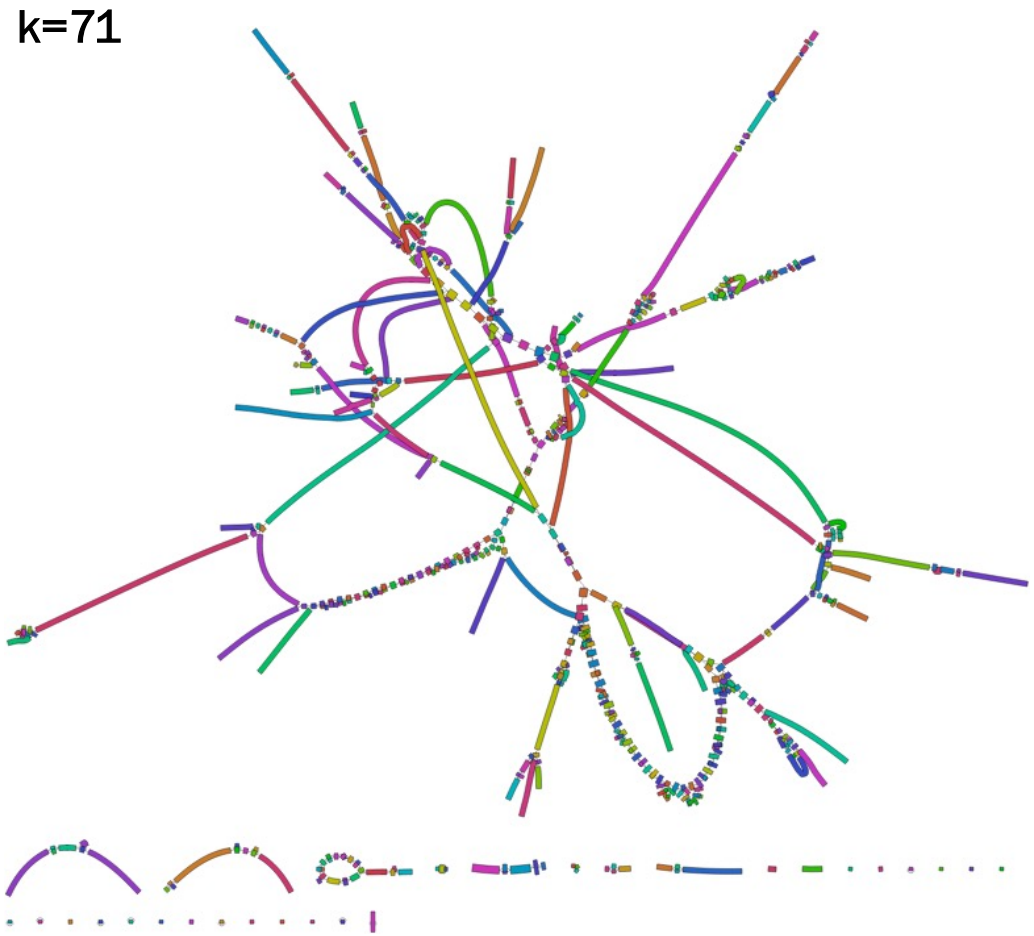
https://github.com/rrwick/Bandage/wiki/Effect-of-kmer-size

Choice of kmer(s)?



https://github.com/rrwick/Bandage/wiki/Effect-of-kmer-size

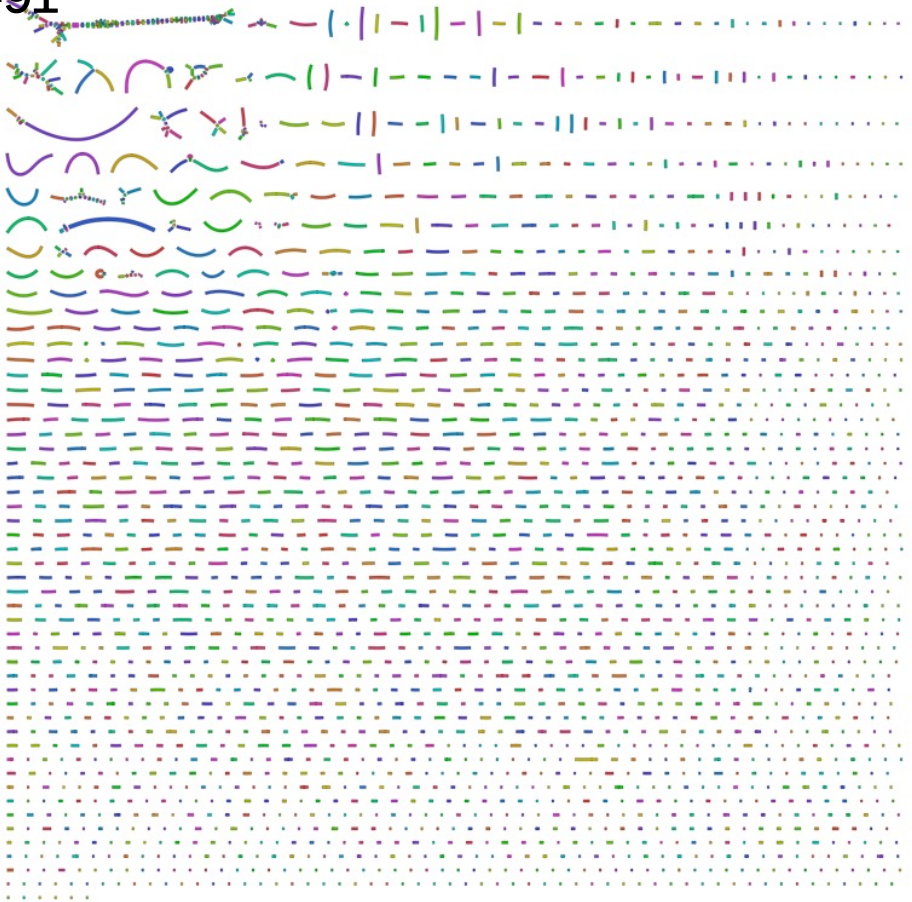
Choice of kmer(s)?



https://github.com/rrwick/Bandage/wiki/Effect-of-kmer-size

Choice of kmer(s)?

k=91

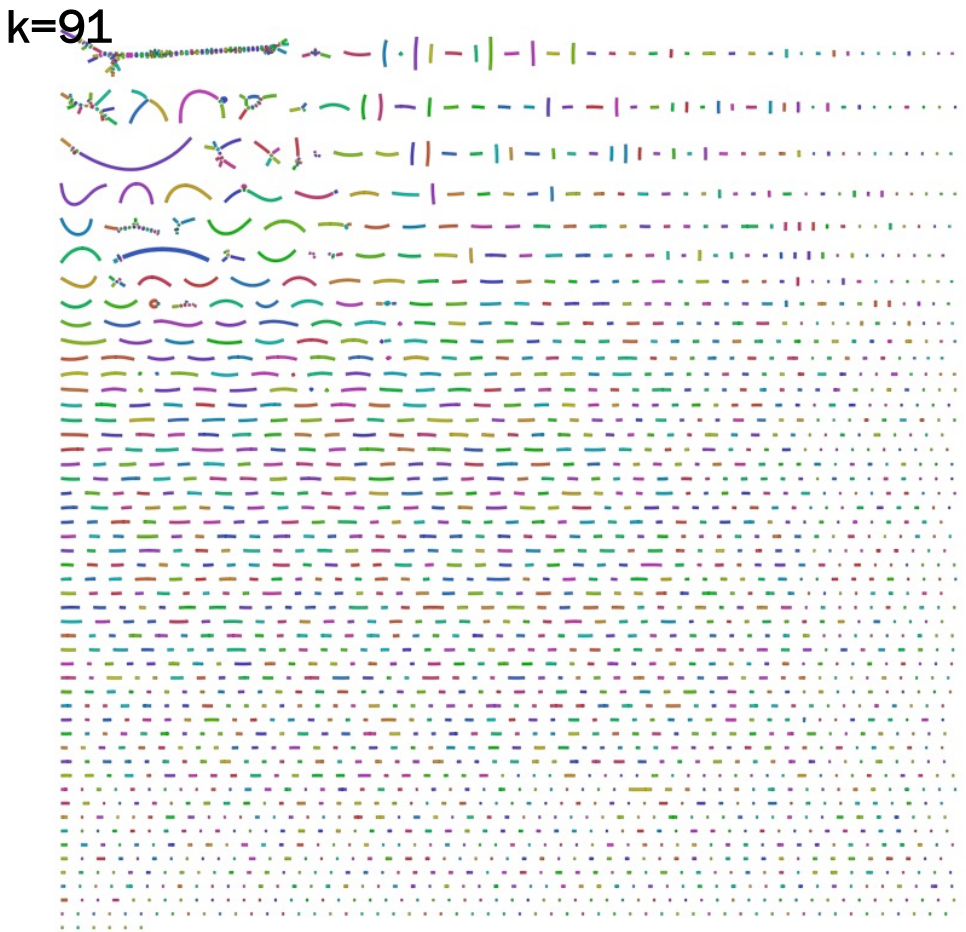


https://github.com/rrwick/Bandage/wiki/Effect-of-kmer-size

Choice of kmer(s)?

- Short kmers → short contigs & many connections
- Longer kmers → longer contigs & fewer connections
- Things affecting optimal kmers:
  - Read length
  - Read depth
  - Sequence complexity

- **Multiple kmers**



https://github.com/rrwick/Bandage/wiki/Effect-of-kmer-size

Which assembler to choose?

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.

Different assemblers for short-reads

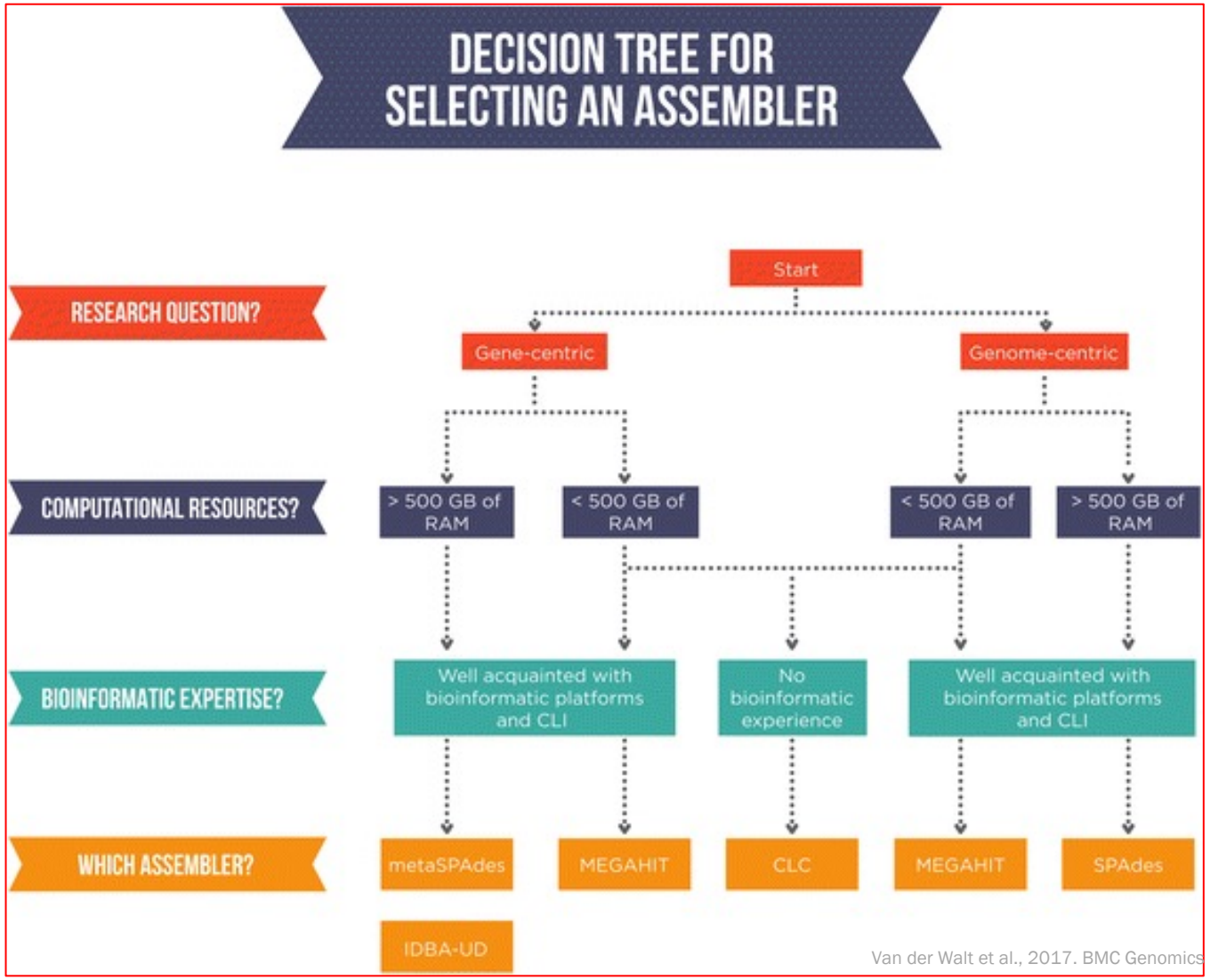
Different assemblers for short-reads

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

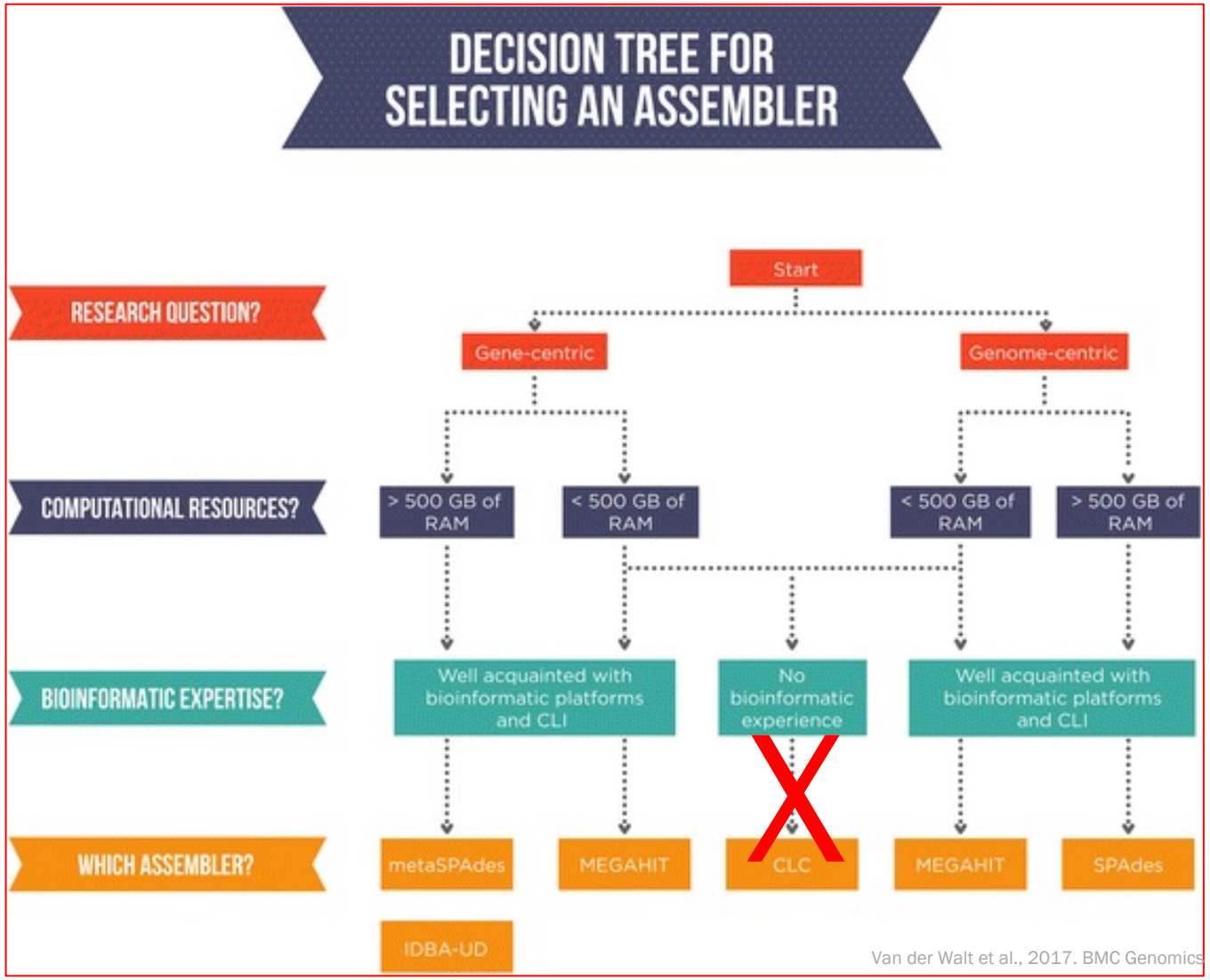
Van der Walt et al., 2017. BMC Genomics

Which assembler to choose?



Van der Walt et al., 2017. BMC Genomics

Which assembler to choose?



Van der Walt et al., 2017. BMC Genomics

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

- Choice of kmers
- 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?