GoSeqIt

Assembly and draft genome quality assessment

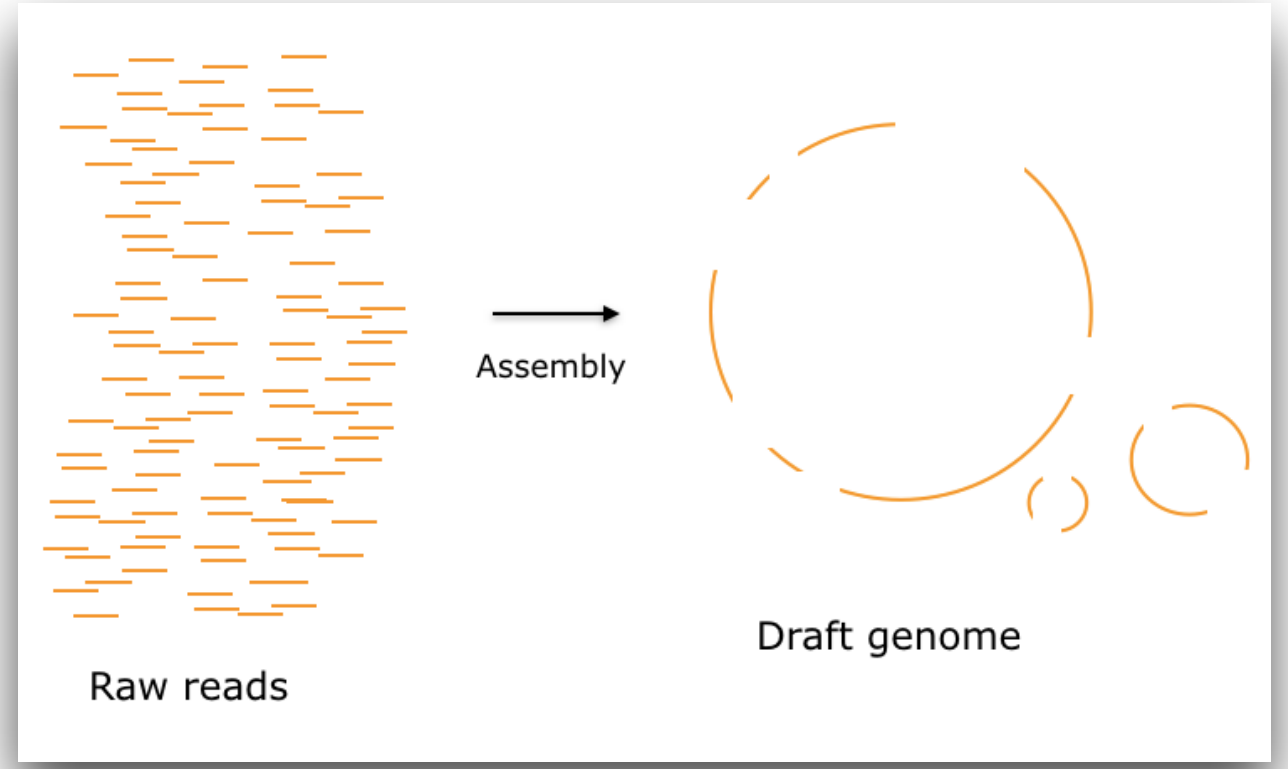
Learning objective:

After this lecture, you should be able to…

…in general terms describe different approaches for de novo genome assembly

..account for draft genome quality assessment by QUAST

Assembly



Two basic approaches

- Alignment: Use a reference genome and align your reads to the genome
- de novo assembly: Try to assemble the reads into a (draft) genome without any prior knowledge

De novo assembly

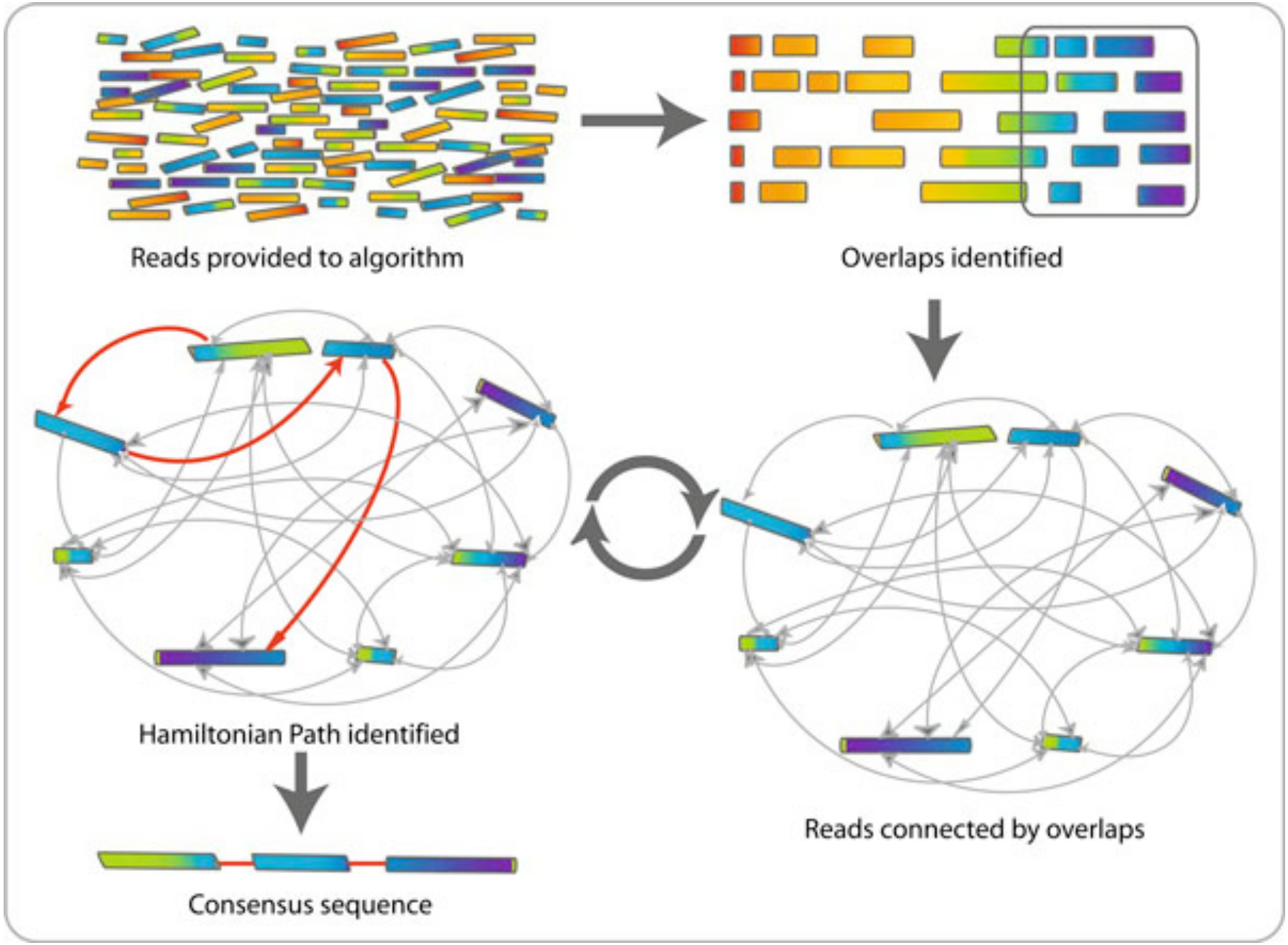
1. Greedy (simple approach)
2. Overlap, layout and consensus (OLC) approaches
3. de Bruijn Graphs

Greedy

1. Pairwise align all reads
2. Identify fragments that have the largest overlap
3. Merge these
4. Repeat until all overlaps are used

- *Can only resolve repeats that are smaller than the read length*
- ***High*** *computational cost with increasing no. of reads*

Overlap, layout and consensus approaches



Overlap, layout and consensus approaches

- Not good with many short reads -> lots of alignments!
- With short read lengths it is difficult to resolve repeats
- Good for long reads (PacBio, ONT)
- Assemblers include Celera (PMID: 10731133) and MIRA (https://sourceforge.net/p/mira-assembler/wiki/Home/), Miniasm (PMID: 27153593)

de Bruijn graphs

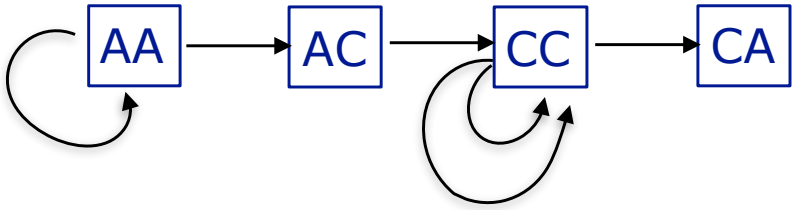
Instead of comparing reads, reads are initially chopped into kmers (nucleotides of length *k*)

Very simple example:

Original genome: A A A C C C C A

Kmers (3-mers): AAA AAC ACC CCC CCC CCA

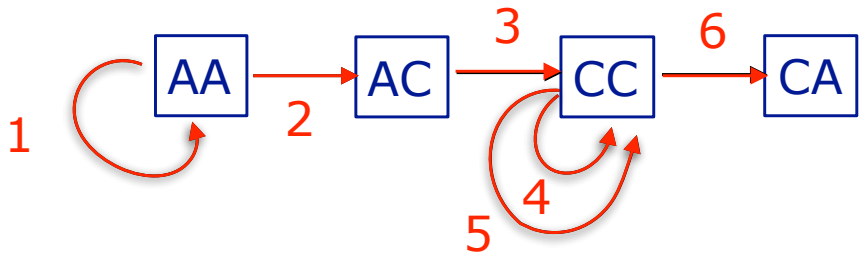
Extract all L/R 2-mers: AA,AA AA,AC AC,CC CC,CC CC,CC CC,CA



One edge per kmer

one node per unique k-1-mer

Reconstructing the original genome from the graph - walking the graph



AAACCCCA

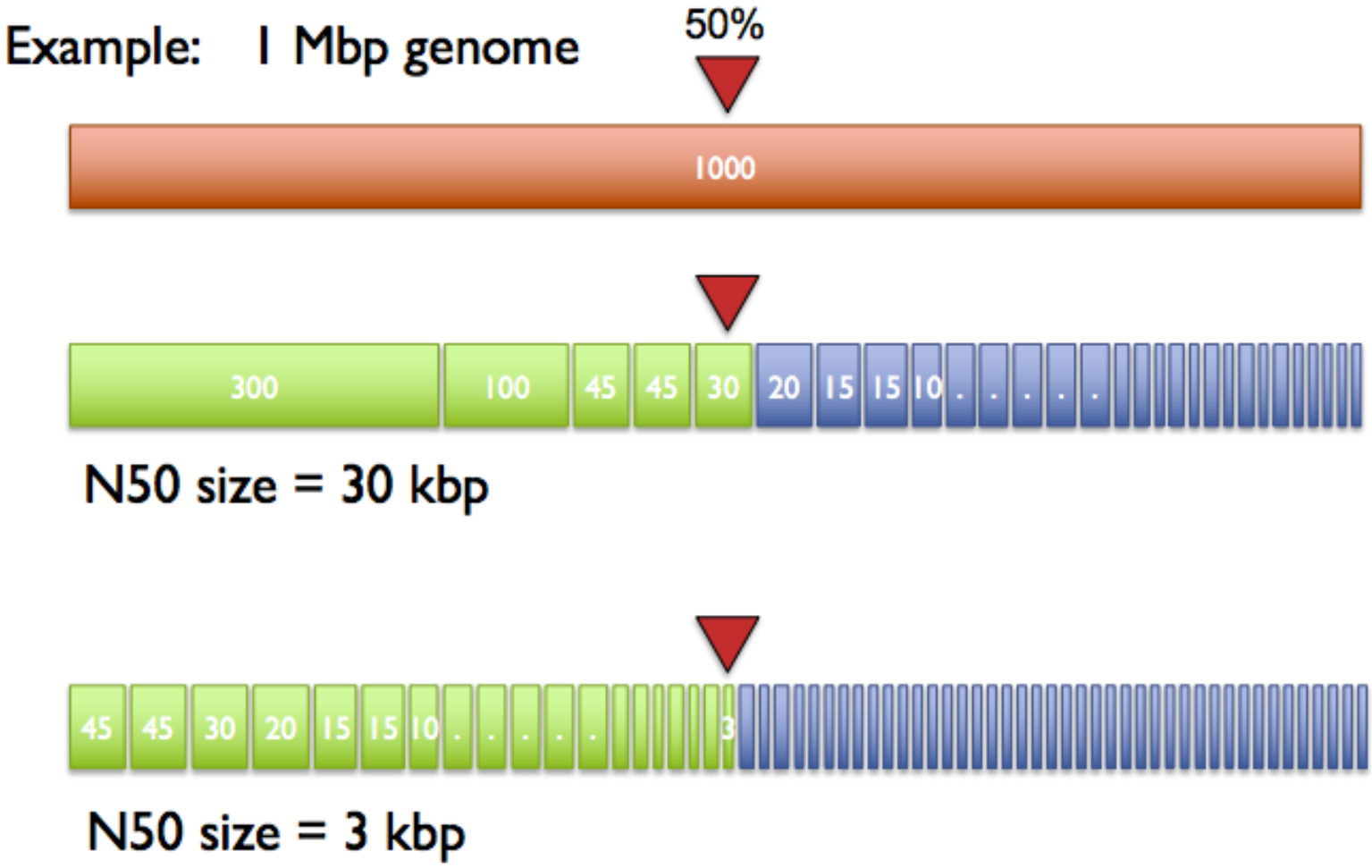
Original genome: A A A C C C C A

de Bruijn graphs

- Optimal kmer length cannot be identified a priori, must be experimentally tested for each dataset
- Requires a lot of RAM
- Assemblers include Velvet (PMID: 18349386) and SPAdes (PMID: 22506599)
    - Velvet uses only one length of the kmer, while SPAdes uses multiple

Assessing the quality of draft genomes

N50 value:
The N50 value for draft genomes is defined as the length of the shortest contig, in the set of largest contigs that represents at least 50% of the assembly



QUAST

Quality Assessment Tool for Genome Assemblies

- Evaluates genome assemblies
- Works both with and without a reference genome
- Accepts multiple assemblies, thus is suitable for comparison
- Web-service available: http://quast.bioinf.spbau.ru/

Example QUAST report

Quality Assessment Tool for Genome Assemblies by CAB

16 August 2018, Thursday, 09:36:10

View in Icarus contig browser

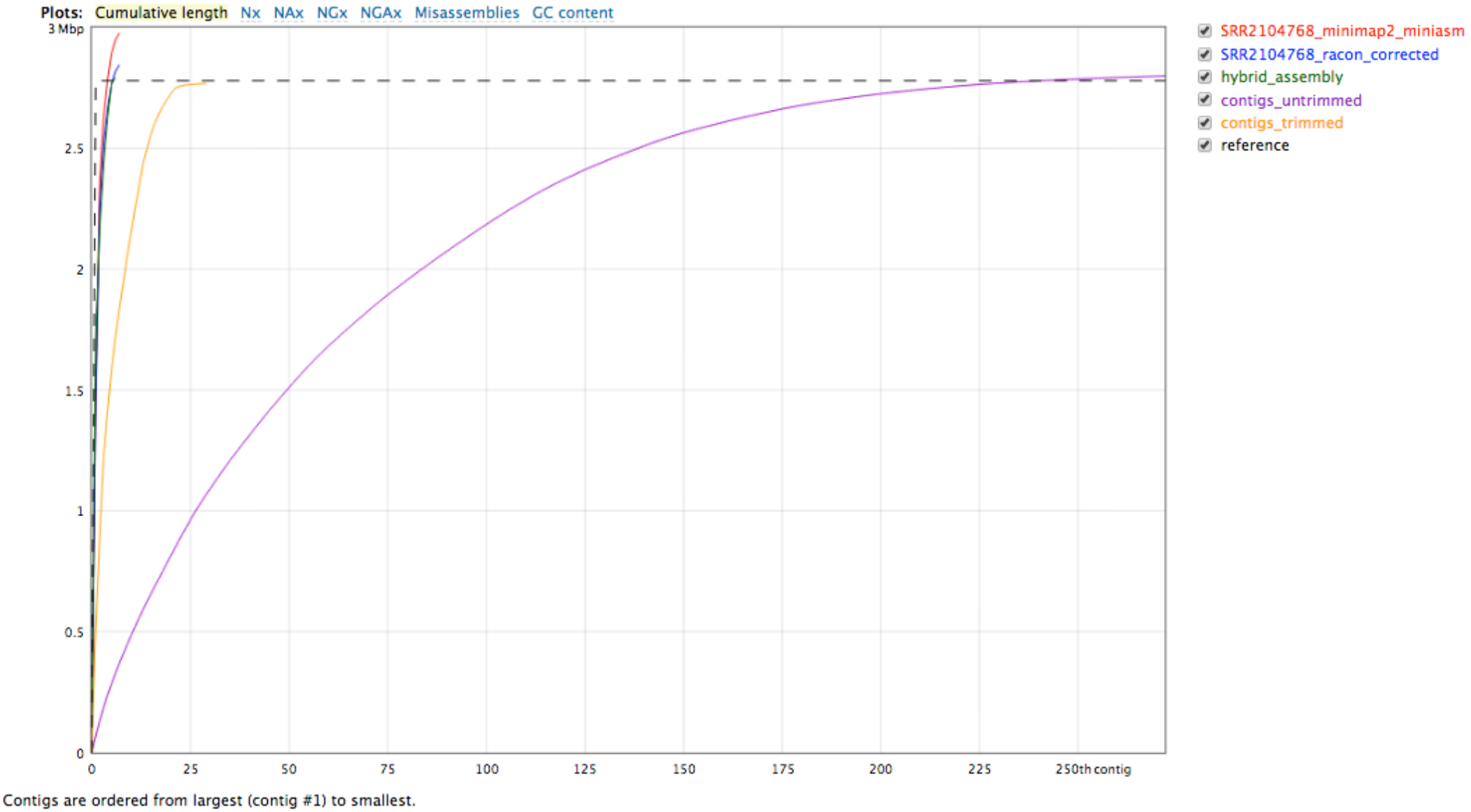
All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g., "# contigs (>= 0 bp)" and "Total length (>= 0 bp)" include all contigs).

Aligned to "NZ_CP009361" | 2 778 854 bp | 1 fragment | 32.88% G+C

Worst Median Best — Show heatmap

Genome statistics	SRR2104768_minimap2_miniasm	SRR2104768_racon_corrected	hybrid_assembly	contigs_untrimmed	contigs_trimmed
Genome fraction (%)	0.013	97.836	99.555	98.569	98.617
Duplication ratio	0.989	1.01	1	1.011	1
Largest alignment	360	562 778	1 727 799	67 407	508 890
Total aligned length	360	2 746 006	2 766 945	2 768 258	2 741 485
NGA50	-	433 044	1 727 799	19 579	189 171
LGA50	-	3	1	44	4
Misassemblies					
# misassemblies	0	10	0	1	0
Misassembled contigs length	0	2 766 259	0	681	0
Mismatches					
# mismatches per 100 kbp	274.73	23.5	0.61	1.13	0.18
# indels per 100 kbp	3296.7	280.09	0.43	0.07	0.07
# N's per 100 kbp	0	0	0	3.57	0
Statistics without reference					
# contigs	7	7	6	272	29
Largest contig	1 509 913	1 446 627	1 727 799	67 407	508 890
Total length	2 976 675	2 845 188	2 792 688	2 798 824	2 767 228
Total length (>= 1000 bp)	2 976 675	2 845 188	2 792 688	2 771 320	2 765 885
Total length (>= 10000 bp)	2 976 675	2 845 188	2 792 688	2 237 142	2 747 270
Total length (>= 50000 bp)	2 948 288	2 817 865	2 766 691	237 532	2 558 562

Example QUAST report



Example QUAST report