Bio 296: Microbial Bioinformatics Introduction to DNA extraction and Sequencing Technologies

Tim Roloff

Assemblies – 2 strategies: mapping and de novo

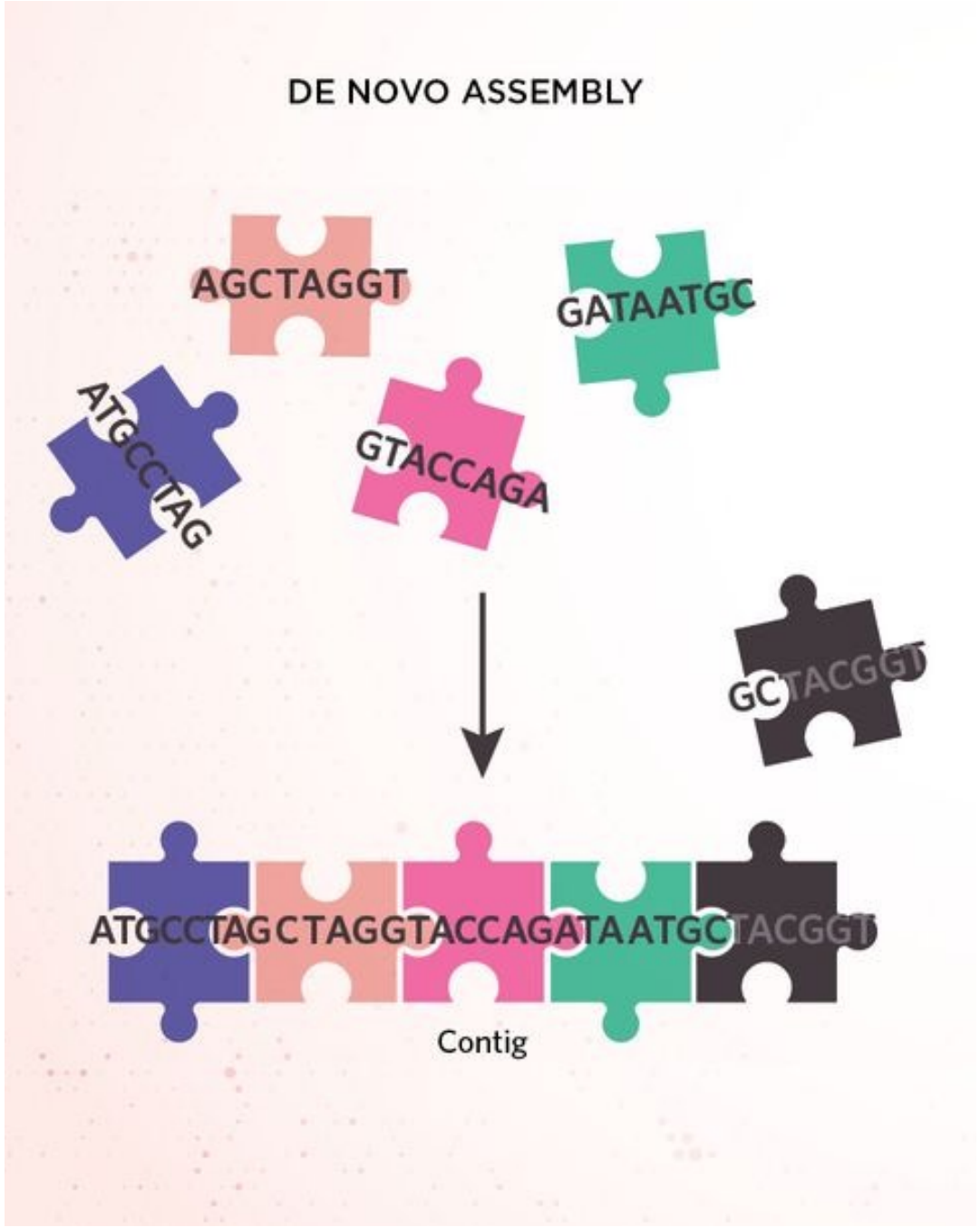
Mapping

- used for closely related genomes.
- Reference to map against has to be closesly related to genome to analyze
- Useful for close outbreaks and species with little variation (small accessory genome) e.g. *Mycobacteroum tuberculosis*

De novo assembly

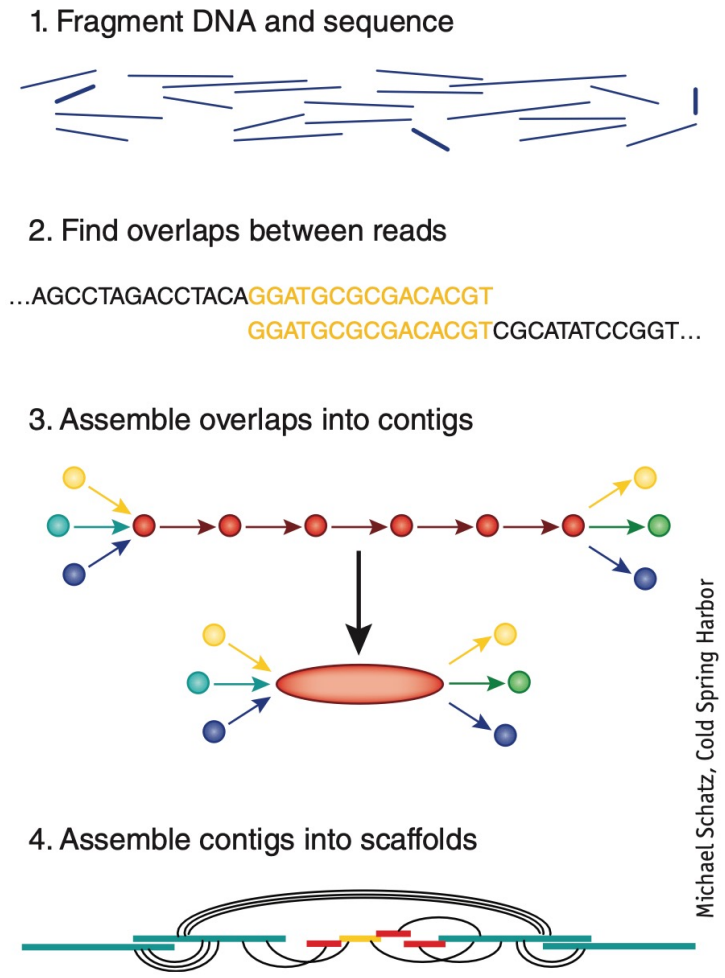
- Used for diverese genomes (large accessory genome) and new speices
- Long reads can help to get more complete assemblies

In routine diagnostics: De-novo used by default, mapping used for closely related outbreaks and mycobaceria


DE NOVO ASSEMBLY
AGCTAGGT
GATAATGC
ATGCCTAG
GTACCAGA
GCTACGGT
ATGCCTAGCTAGGTACCAGATAATGCTACGGT
Contig

De novo assembly

- De novo is a Latin term that means "anew," "from the beginning," or "afresh."
- Arranging/Assembling the reads (jigsaw pieces) to complete genomes.
- Why do we have to assemble the reads?
- species without reference genomes, for characterizing the genetic variants present in the species (Chen et al., 2021). The end goal of a sequence assembler is to produce long contiguous pieces of sequence
- Greedy algorithm and **De Bruijn graph** are the most used algorithms by the assembly tools.
- Short read assembly - Illumina (Spades, Unicycler)
- Long read assembly - Nanopore, PacBio (Flye)
- Hybrid assembly - Best of two worlds (Unicycler, Trycycler)



Genome assembly stitches together a genome from short sequenced pieces of DNA.

https://doi.org/10.1038/nmeth.1935

Assembly software used over time

Various asssemly software available. Spades and Unicycler widely used.

	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
SPADES	0.0	0.0	0.0	0.5	3.1	11.3	30.8	45.4	49.4	49.1	54.8
CLC	1.5	9.6	8.8	4.9	9.9	22.8	17.0	19.6	10.6	7.3	6.6
VELVET	34.6	9.5	21.1	20.5	7.4	11.3	13.0	6.0	6.1	6.3	3.0
ALLPATH	0.0	1.1	7.9	35.4	49.6	4.4	0.2	1.3	0.8	0.4	0.0
HGAP	0.0	0.0	0.0	0.7	2.2	5.7	7.7	7.8	7.7	7.8	5.7
NEWBLER	55.2	64.0	25.2	13.3	8.9	9.8	7.6	2.7	2.8	2.4	2.9
SOAP	0.3	3.0	7.4	6.2	3.1	5.7	4.8	2.8	8.6	7.2	8.1
A5	0.0	0.0	0.0	1.0	0.6	6.0	2.0	2.1	5.2	4.1	0.5
ABYSS	0.0	0.2	0.7	1.3	3.2	14.2	5.3	1.8	2.6	0.5	2.7
CELERA	4.8	12.7	27.8	12.6	1.3	2.7	1.2	0.9	0.9	0.3	0.3
PLATANUS	0.0	0.0	0.0	0.0	0.0	0.4	0.1	1.5	0.5	4.3	0.1
UNICYCLER	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	2.3	4.1	10.3
MIRA	1.0	0.7	1.5	1.7	1.5	1.6	2.6	1.4	0.3	0.2	0.1
MASURCA	0.0	0.0	0.0	0.4	6.5	0.2	1.1	0.2	0.2	0.4	0.2
IDBA	0.0	0.0	0.1	0.1	2.3	4.2	0.8	1.0	0.2	0.1	1.0
CANU	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.7	1.5	1.6	2.1
PHRED/PHRAP/CONSED	1.3	1.6	1.4	1.6	1.9	1.7	0.2	0.1	0.1	0.0	0.0
GENEIOUS	0.0	0.1	0.0	0.2	0.2	0.5	0.6	0.6	0.4	0.6	0.1
RAY	0.0	0.0	0.1	0.5	0.2	2.0	0.2	0.9	0.1	0.0	0.7
DNASTAR	0.0	0.0	0.7	0.3	0.2	0.9	0.4	0.6	0.2	0.0	0.5
FALCON	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.2	0.3
SKESA	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0

https://doi.org/10.3389/fcimb.2020.592102

De Bruijn graph theory – A lecture on its own!!!

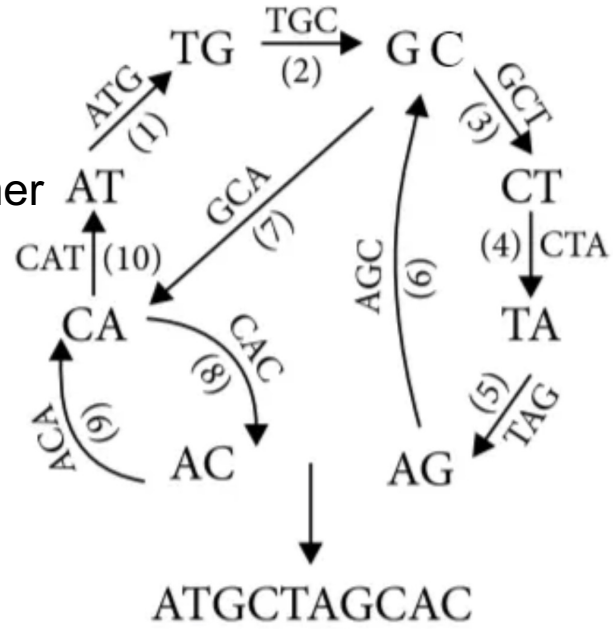
Genome: ATGCTAGCAC

Reads: ATGCTA
GCTAGC
TAGCAC
GCACAT
ACATGC

Kmer split eg kmer = 3

ATG
TGC
GCT
CTA

Multi graph with k-1 mer



Walk through the generated graph layout only once – find the shortest distance that connects the nodes and reconstruct the original genome sequence – Eulerian path

Kmer: represent the length of nucleotide bases

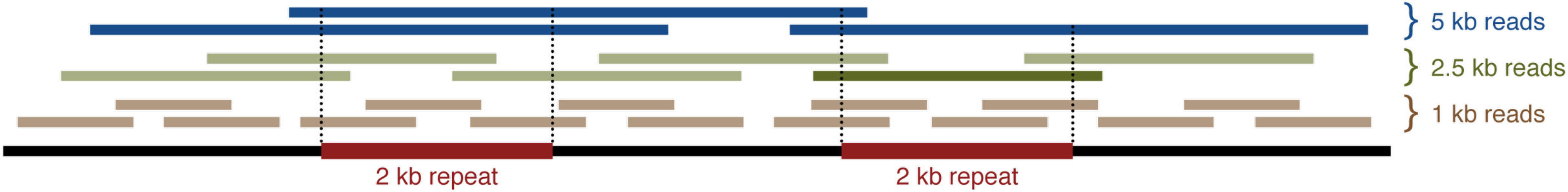
Node

Black arrows in the image - edges

https://towardsdatascience.com/genome-assembly-using-de-bruijn-graphs-69570efcc270

Challenges for good genome assemblies

- Sequencing errors - misassembly
- Uneven sequencing depth and coverage
- Computational resources, and cost
- repetitive regions, ribosomal regions
  - Optimal read length: longer than the longest repeat



Commonly used values

- **N50**: sequence length of the shortest contig at 50% of the total assembly length
- **N90**: sequence length of the shortest contig at 90% of the total assembly length
- **L50**: count of smallest number of contigs whose length sum makes up half of genome size
- **Assembly size**: sum of the length of all contigs
- **Coverage:** can mean 1 of 3 things. We typically use sequence coverage
  - **Sequence coverage** (or depth) is the number of unique reads that include a given nucleotide in the reconstructed sequence.
  - **Physical coverage**, the cumulative length of reads or read pairs expressed as a multiple of genome size.
  - **Genomic coverage**, the percentage of all base pairs or loci of the genome covered by sequencing.

What do we use in the course?

Unicycler (Short reads)

https://github.com/rrwick/Unicycler

- Only individual bacterial isolates
- Short reads, long reads
- Circulization of the genomes
- Better handling of misassembly, repeat regions, and recovering plasmid sequences.

Flye (long reads)

https://github.com/fenderglass/Flye

- Individual and mixed community
- Long-reads only.
- Plasmid recovery.

Hybrid assembly – Unicycler

- Combining the accuracy (Illumina) with closing the gaps (long reads)

Output from assemblies:

- assembly.fasta - the assembled genome
- assembly.gfa - assembly graph
- Log file – Coverage, contig length, N50

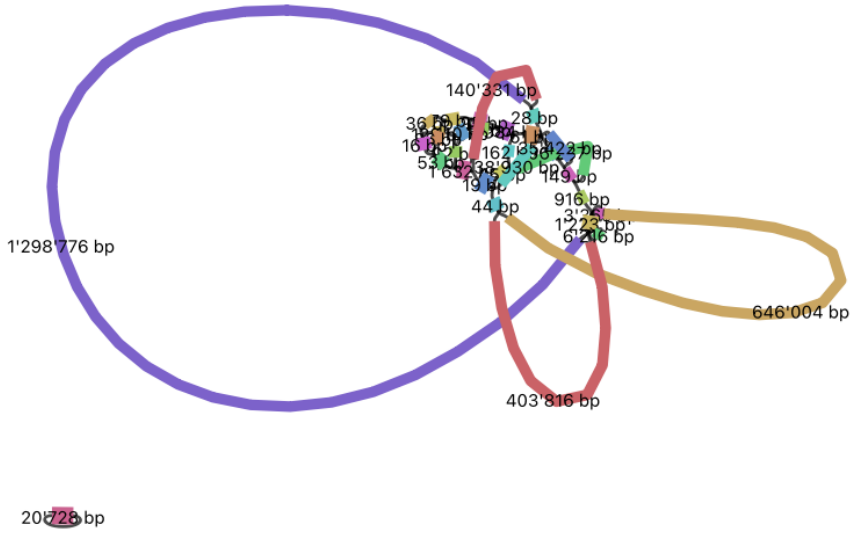
Recommended coverage for WGS - 30 to 50x

Coverage on reads C=L*N/G, where
C - Coverage
L - Read length
N - Number of reads sequenced
G - Genome Size of the bacteria of interest

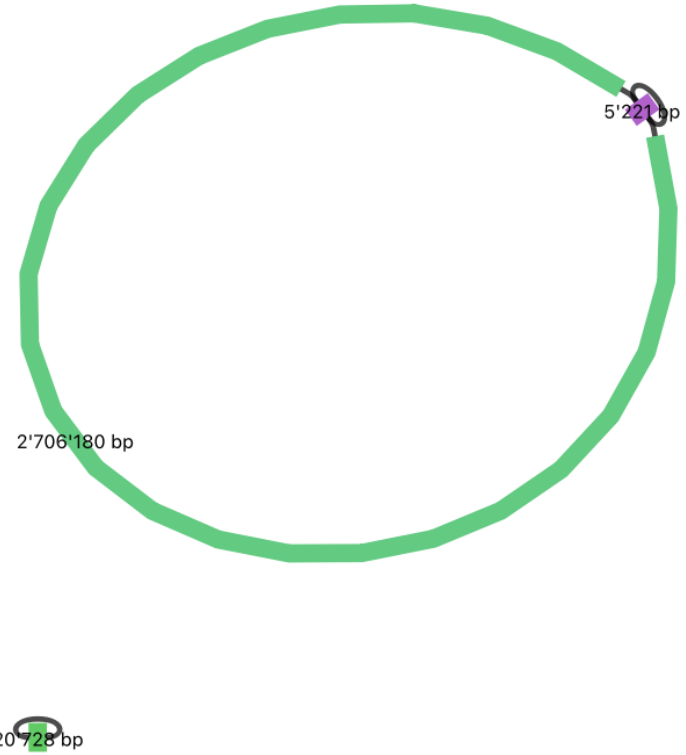
Visualizing the assembly

Bandage - https://rrwick.github.io/Bandage/ Sample SaH1P10622 – input graph file from assembly step – assembly.gfa

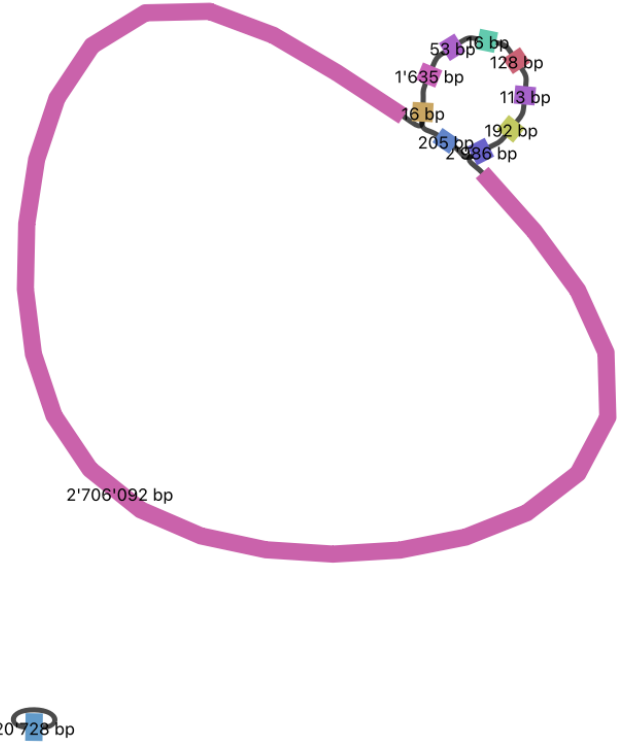
Illumina only (Unicycler)



Nanopore only (Flye)



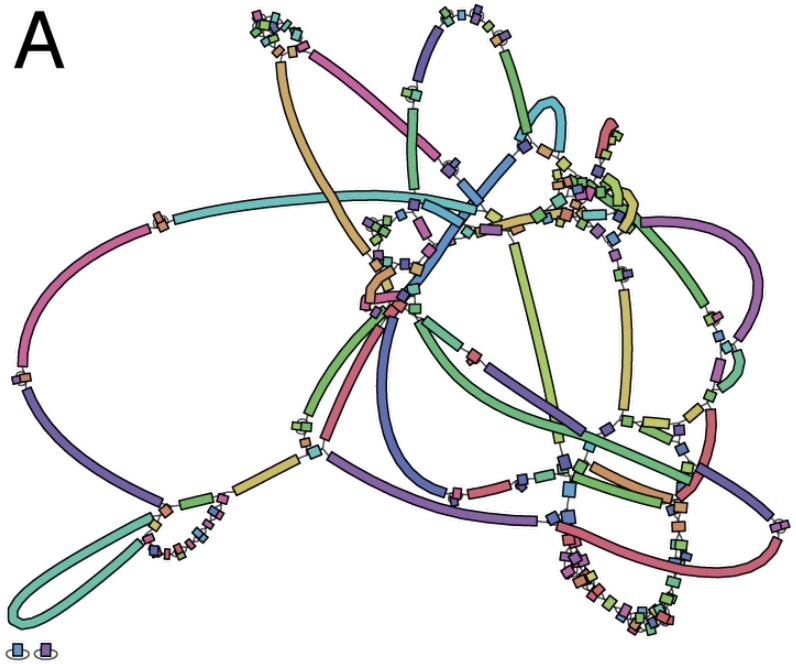
Hybrid (Unicycler)



Unicycler graphs – Illumina only

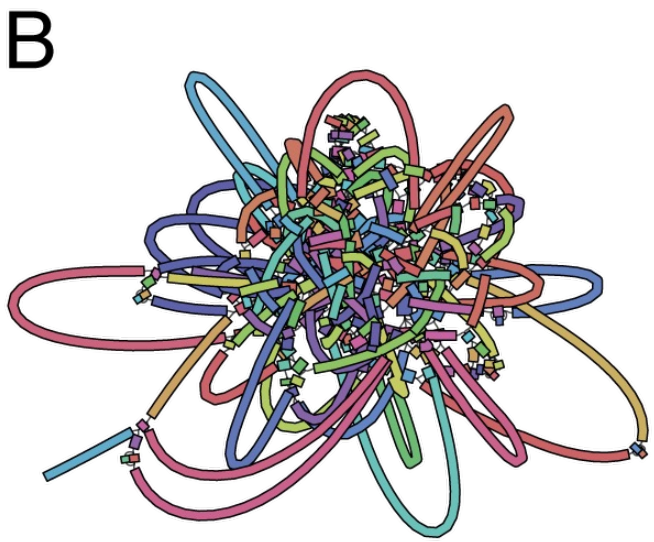
The good

Very good quality – good graph



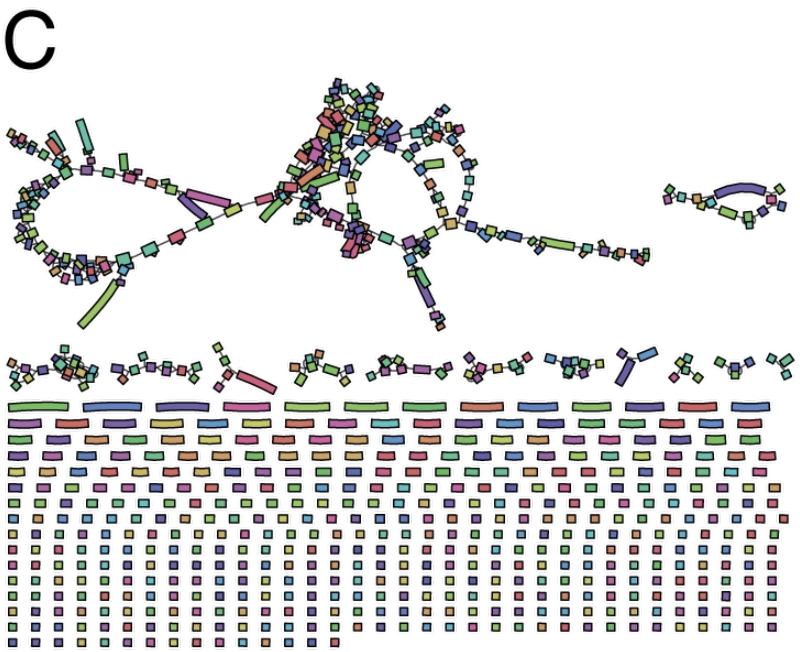
The bad

Good quality – more complex graph

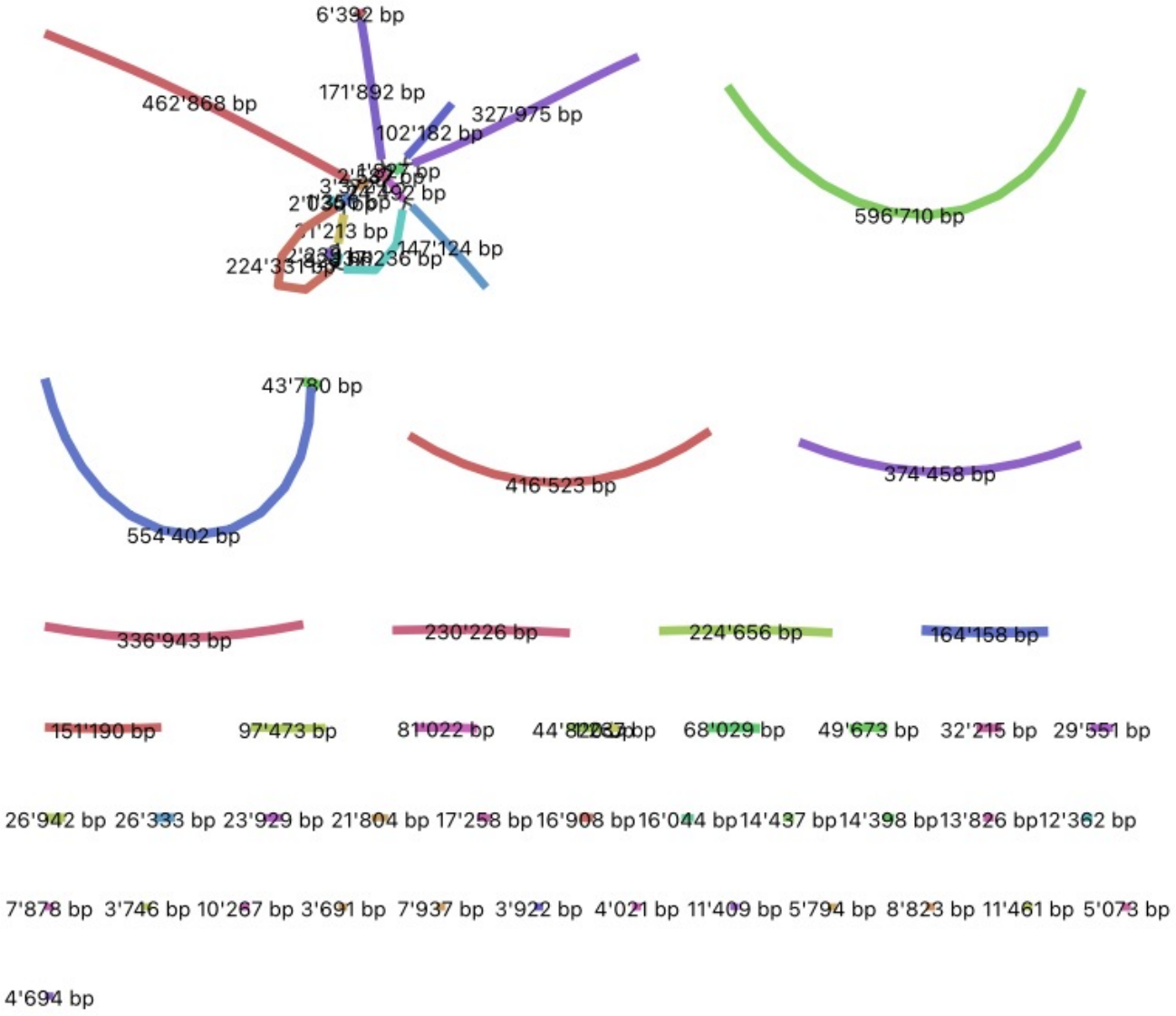


The ugly

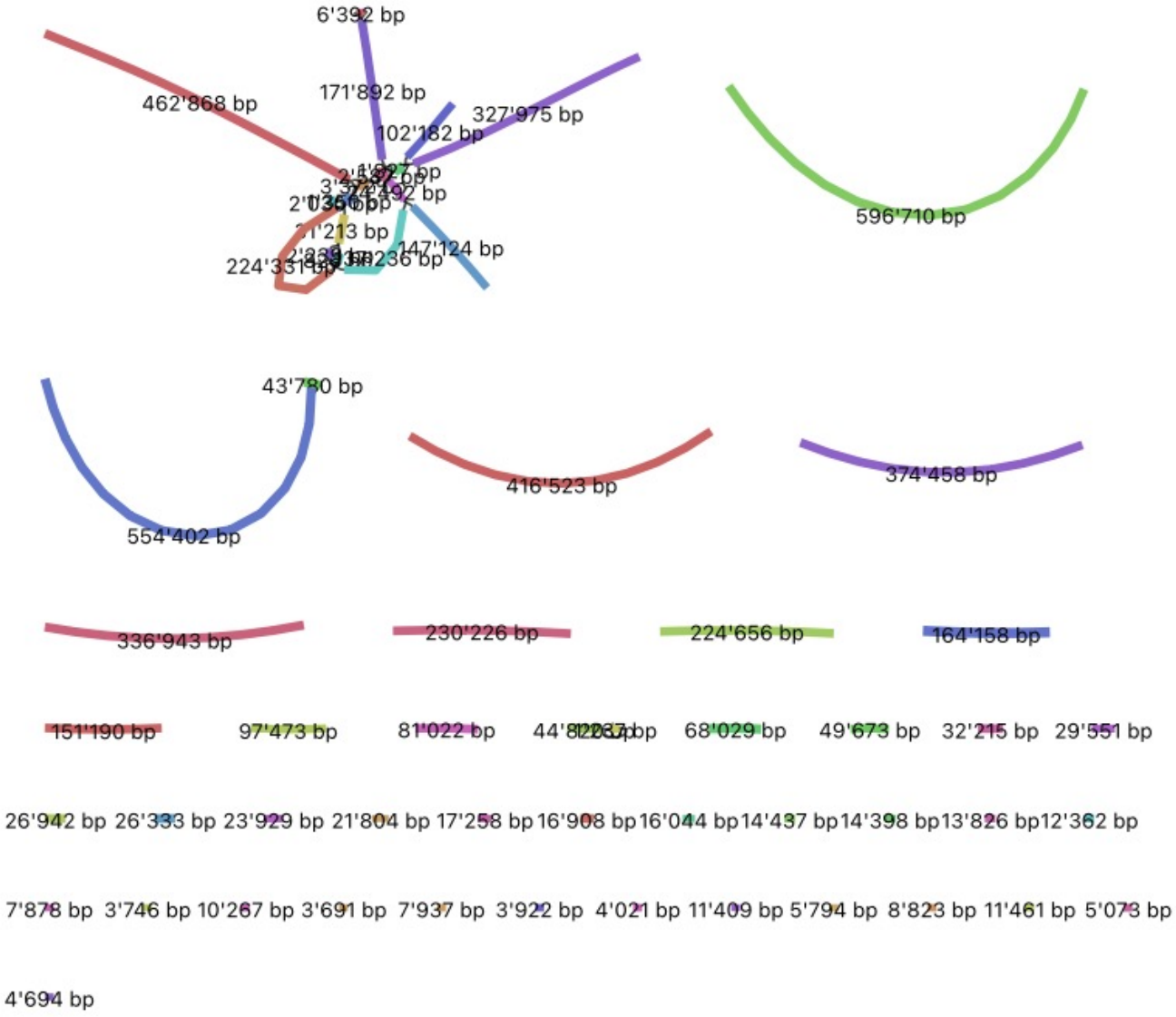
Poor quality – highly fragmented graph



QUIZ – What type of assembly is this?

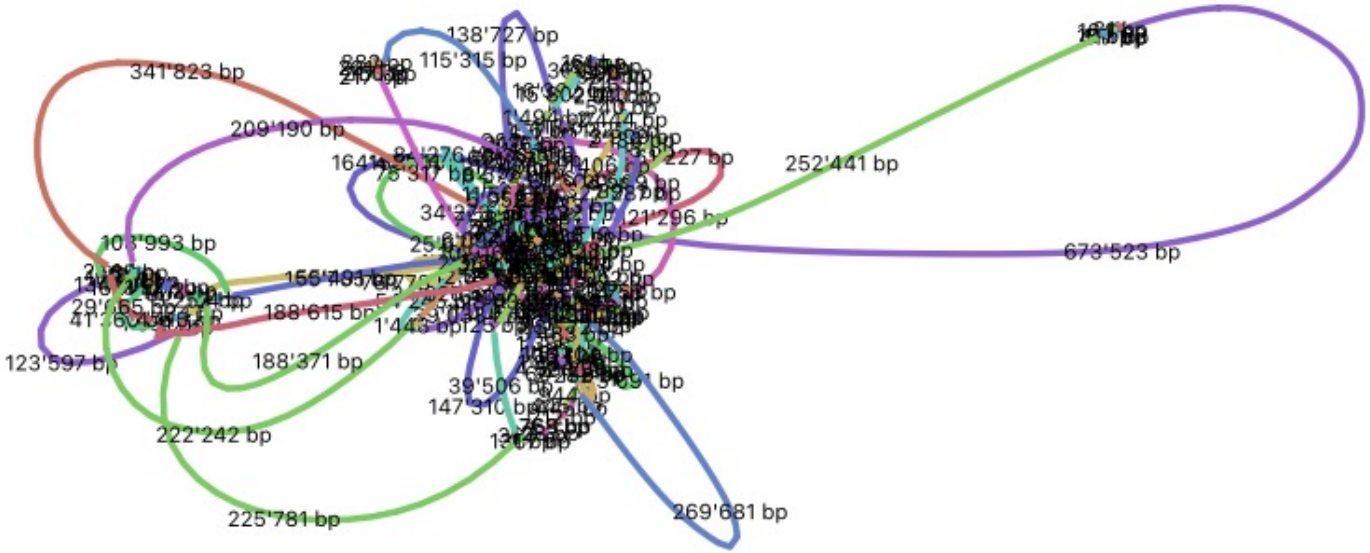


QUIZ – What type of assembly is this?

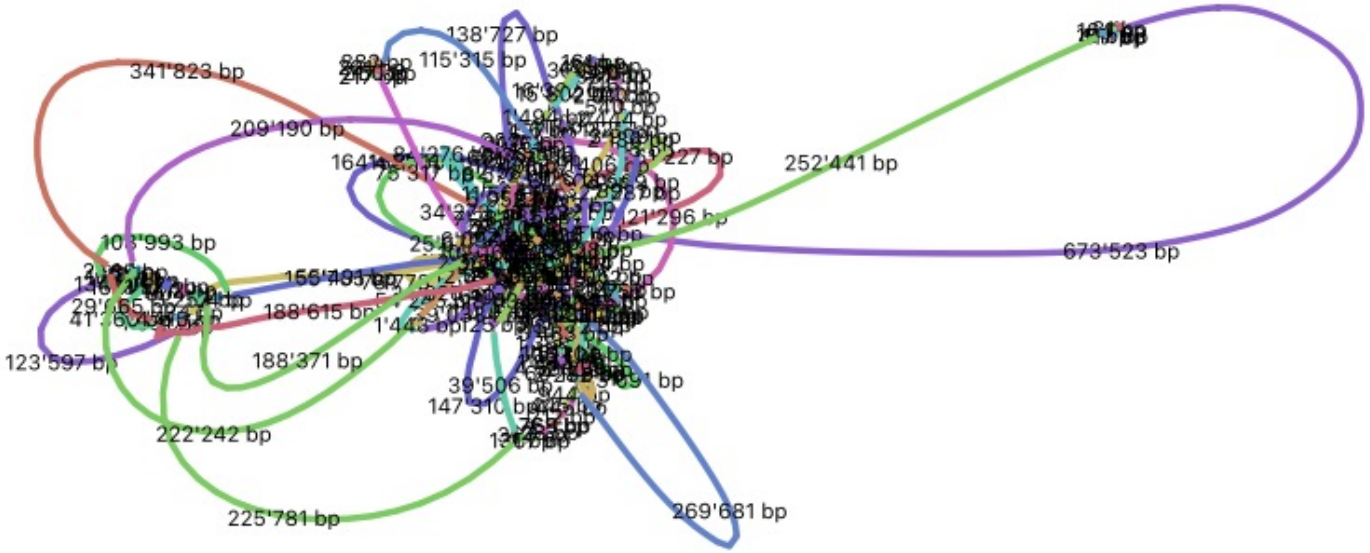


Nanopore

QUIZ – What type of assembly is this?

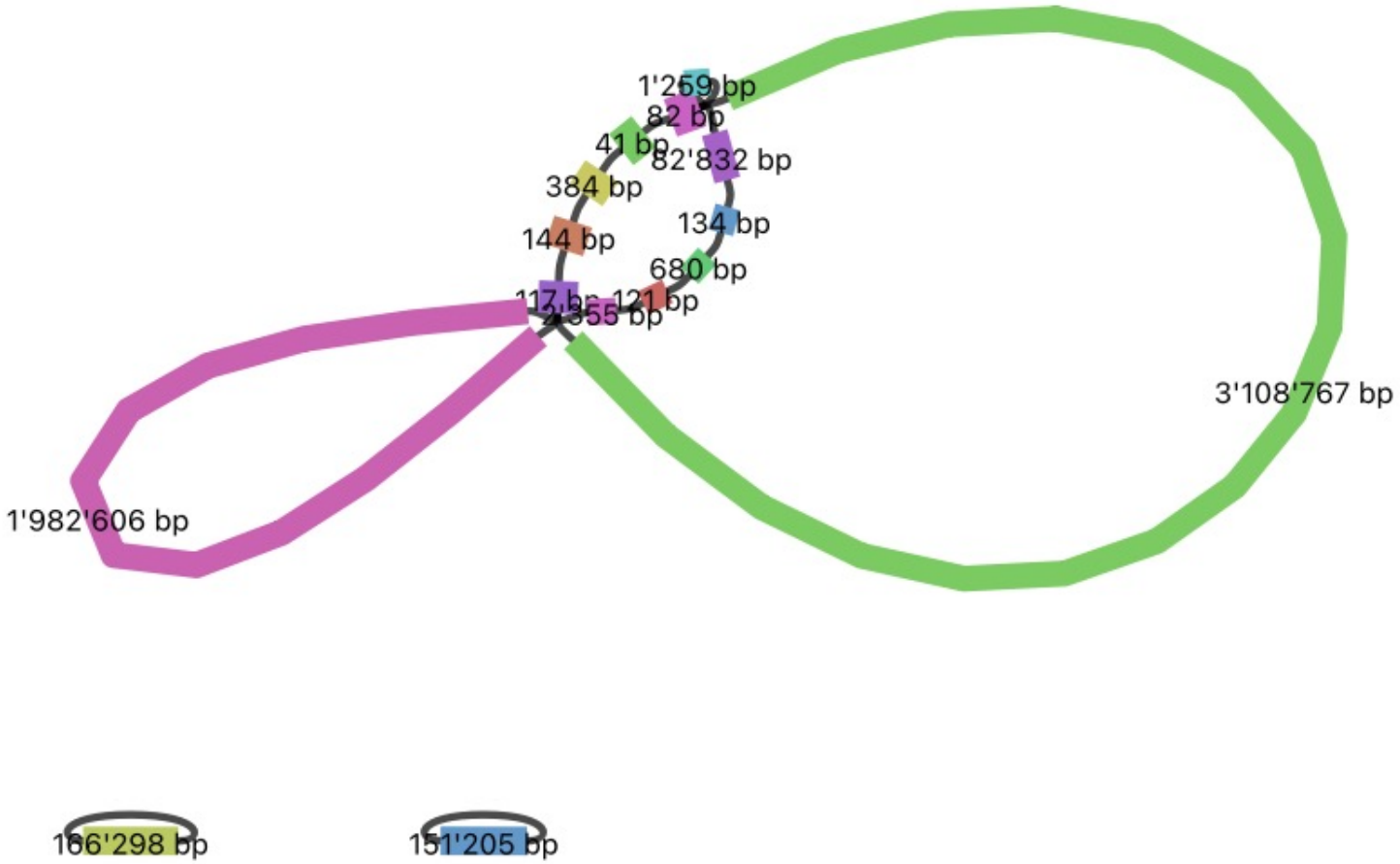


QUIZ – What type of assembly is this?

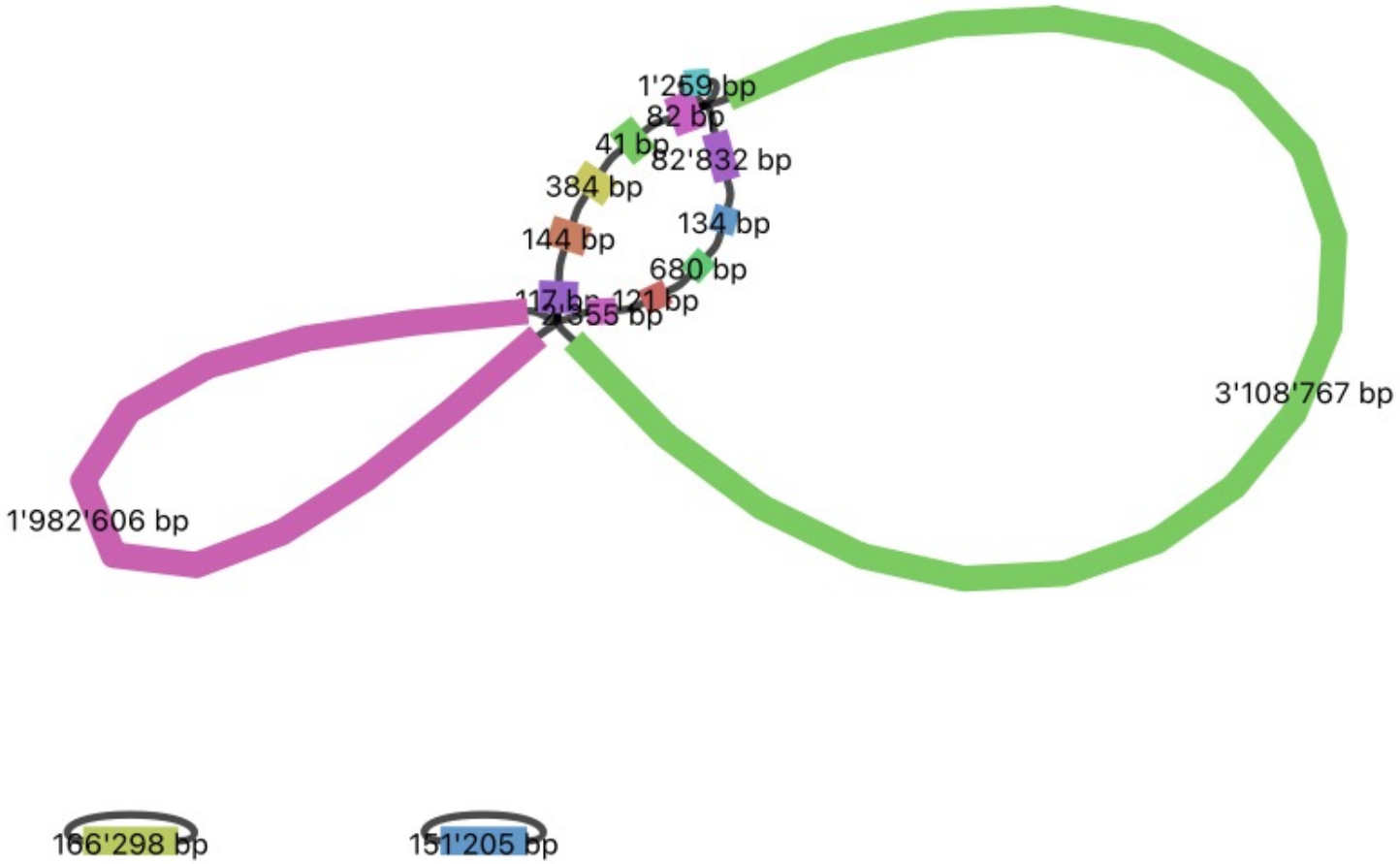


Illumina

QUIZ – What type of assembly is this?



QUIZ – What type of assembly is this?



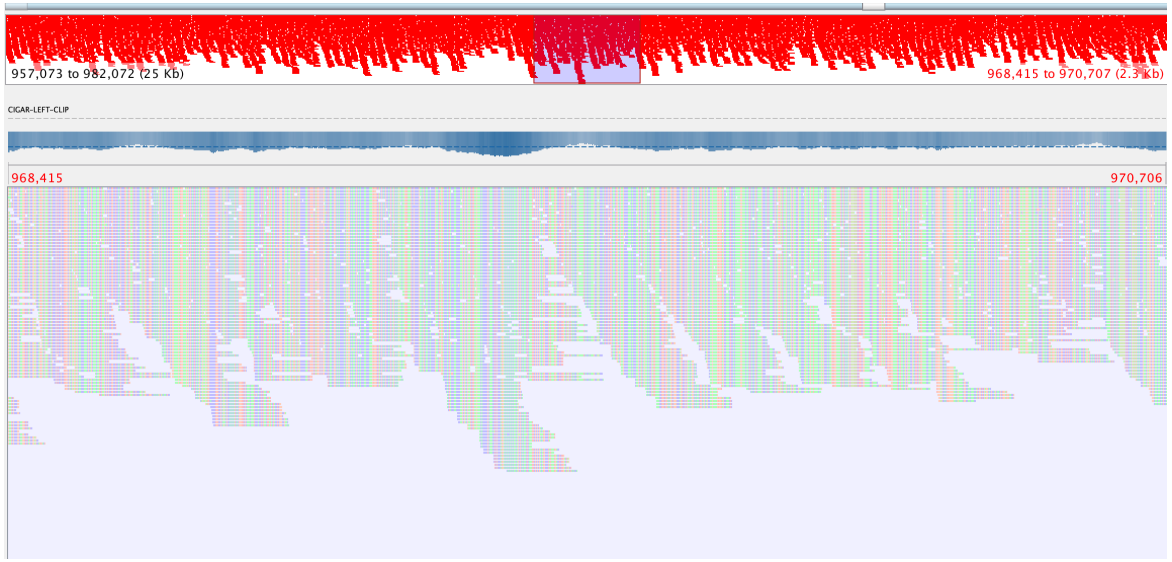
Hybrid assembly

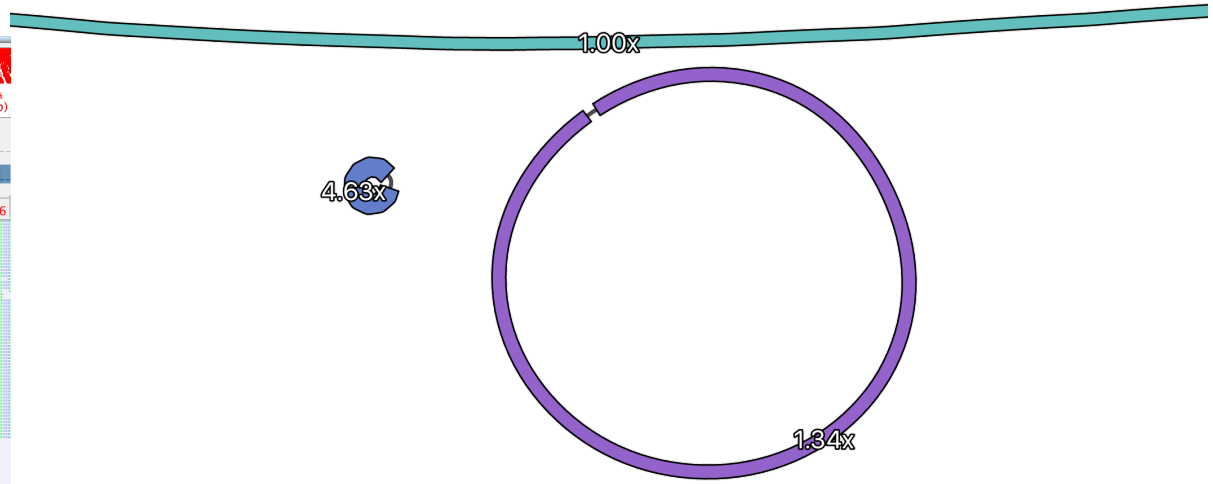
Unicycler graphs – data quality and sequencing depth matter

Coverage plot

- Tablet genome browser
- Input: bam / bai files generated by Pilon
- Output: graph with coverage

Coverage of chromosome set to 1
Coverage of plasmids higher





Polishing the assembly

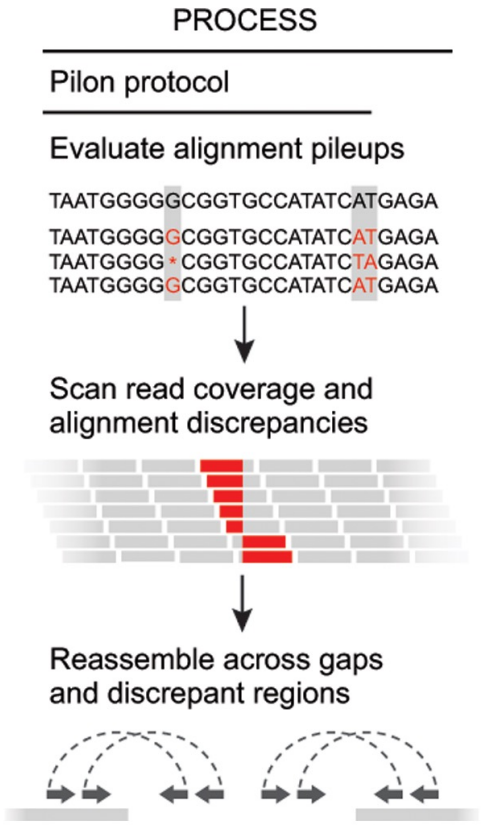
Polishing used for:

– Trying to correct the errors/improve draft genome assembly

Software for Polishing

– Pilon for short reads (Alignment method)
– Medaka for long reads (neural networks)

Polishing using Pilon



https://doi.org/10.1371/journal.pone.0112963

Assembly quality - Metaquast

- To assess the assembled genome quality from the de novo assemblers.
- Parameters like - Contig length, N50, GC content.
- Contig length - For WGS it translates to the expected genome size of the bacteria of interest.
- N50 - used to assess the contiguity of an assembly.
- Metaquast - input - assembly.fasta and raw reads

Statistics without reference	IDBA_UD	Ray	SOAPdenovo2	SPAdes
# contigs	31 224	10 327	36 468	40 546
Largest contig	305 144	99 107	40 707	189 063
Total length	80 325 286	30 411 921	46 741 224	92 397 329
Total length (>= 1000 bp)	69 223 529	27 080 646	30 720 336	77 823 828
Total length (>= 10000 bp)	34 930 908	13 755 677	2 800 864	33 477 263
Total length (>= 50000 bp)	16 008 349	2 346 322	0	11 409 912
Misassemblies				
# misassemblies	1132	407	831	1240
Misassembled contigs length	10 448 260	4 115 772	911 826	10 780 557
Mismatches				
# mismatches per 100 kbp	904.95	1054.68	888.21	1401.84
# indels per 100 kbp	31.88	27.7	17.09	51.64
# N's per 100 kbp	238.48	2087.27	3730.51	1425.14
Genome statistics				
Genome fraction (%)	12.796	4.386	8.055	11.585
Akkermansia_muciniphila_ATCC	0.003	-	-	0.011
Alistipes_putredinis	1.366	0.595	0.61	1.117