SciLifeLab

NBIS
NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

Illumina Assembly

elixir excelerate

Outline

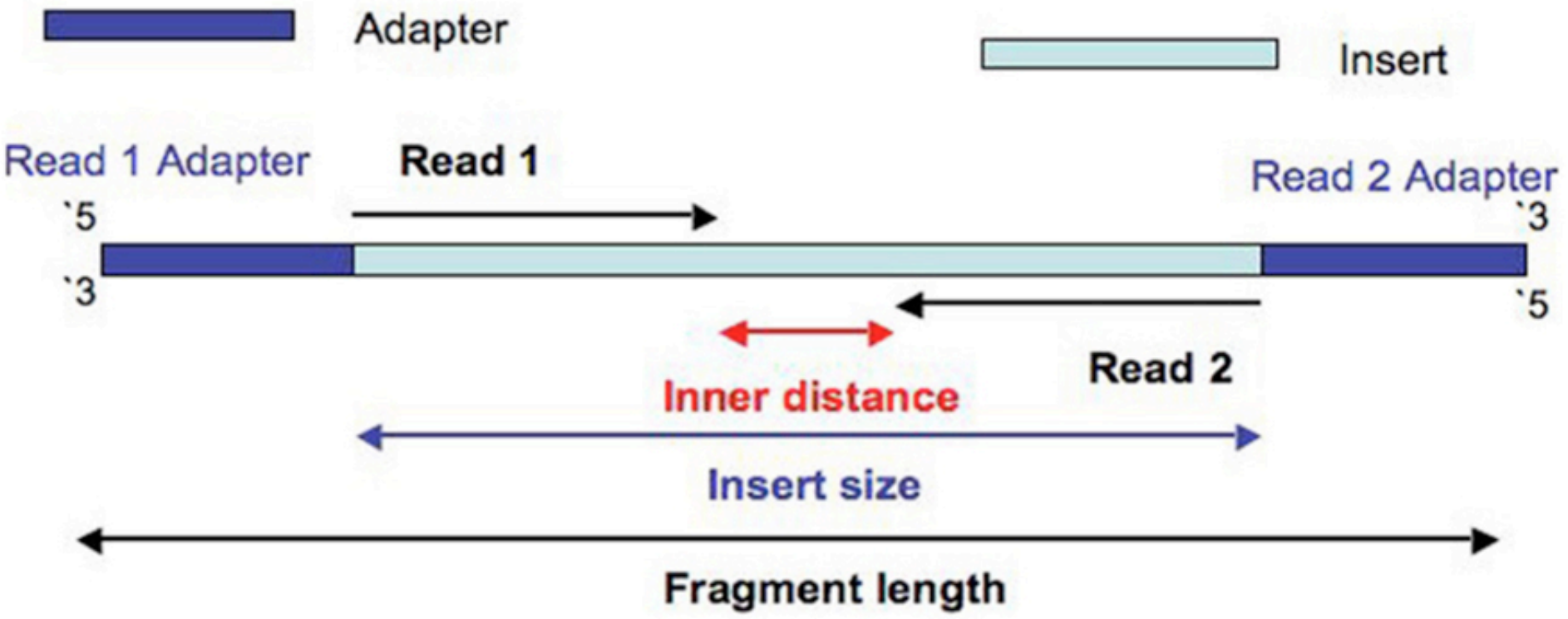
SciLifeLab

- The types of Illumina data
- Methods of assembly
  - Repeats
  - Selecting k-mer size
- Assembly Tools
- Assembly Diagnostics
- Assembly Polishing

Illumina Sequencing

SciLifeLab

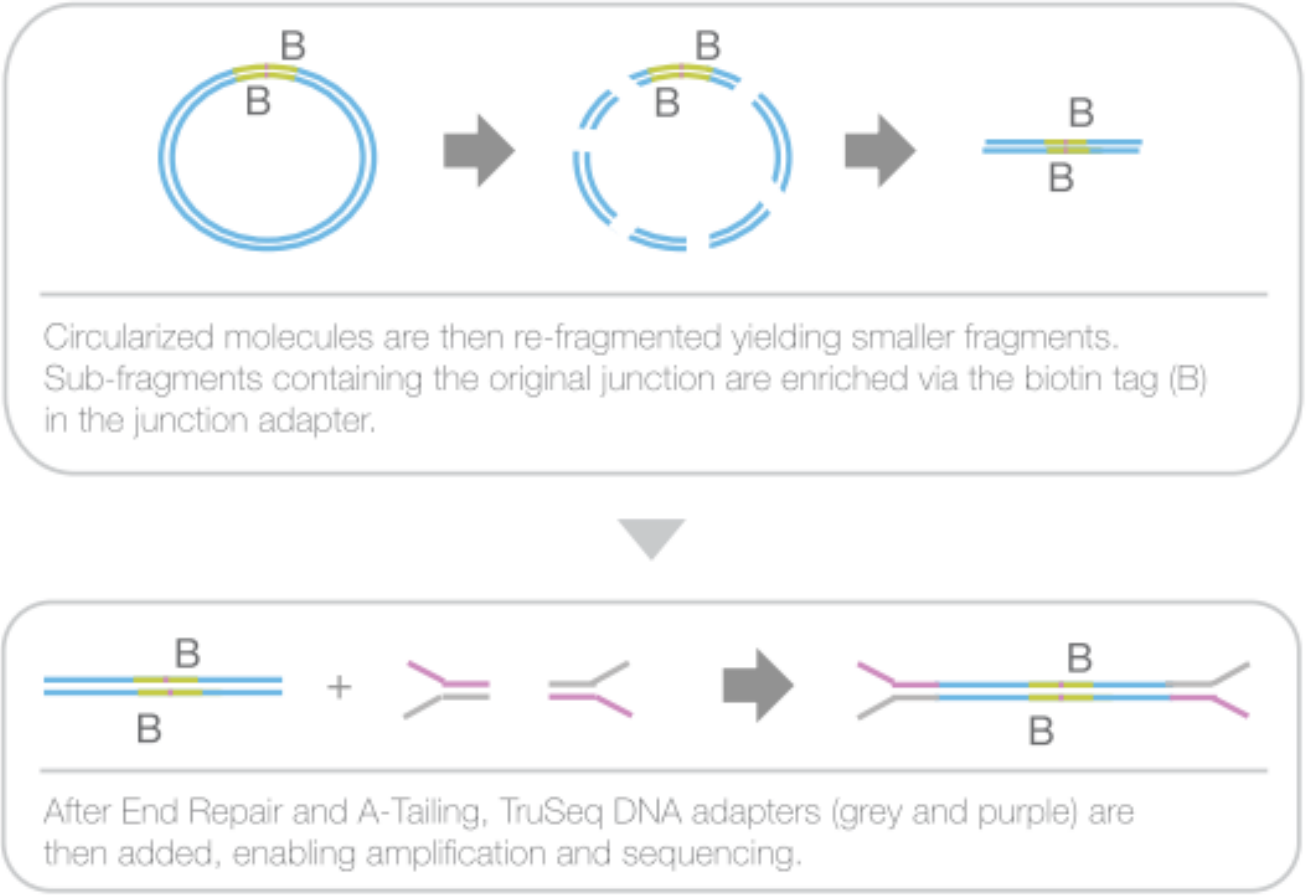
- Paired end Illumina library



Illumina Sequencing

SciLifeLab

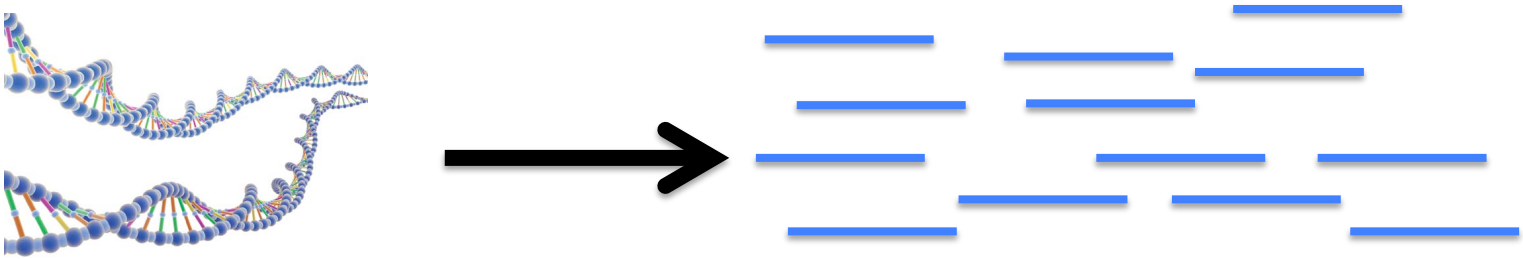
- Mate pair Illumina library



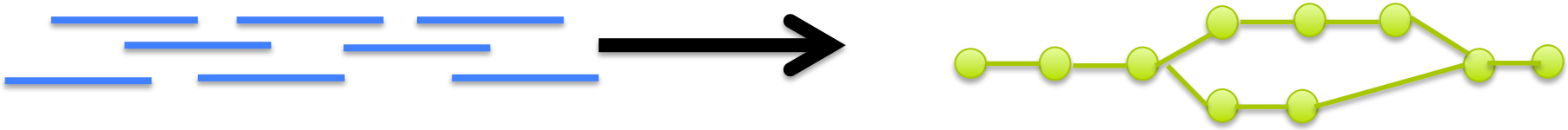
Methods of assembly

SciLifeLab

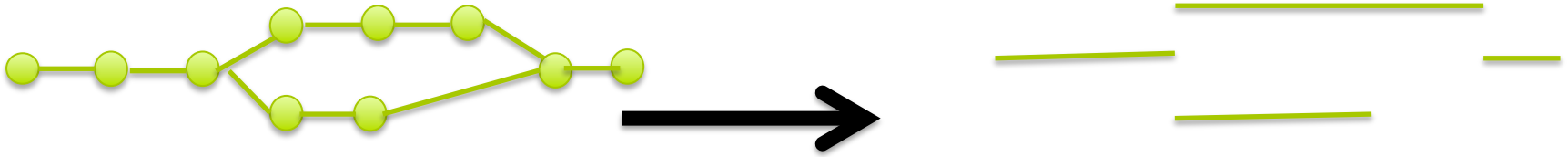
1. Shear and sequence the DNA.



2. Find overlaps in the reads and construct an assembly graph



3. Simply the assembly graph



Methods of assembly

SciLifeLab

- Used for first generation reads (e.g. Sanger)
- Overlap - Layout - Consensus
  - Overlap
    - All vs all pair-wise read comparison
    - Build graph; nodes = reads, edges = overlaps
  - Layout
    - Analyze, simplify, and clean the overlap graph
    - Determine a Hamiltonian path through the graph (visit each node only once, can ignore edges)
  - Consensus
    - Align reads to the assembly path
    - Call bases using weighted voting

Methods of assembly

SciLifeLab

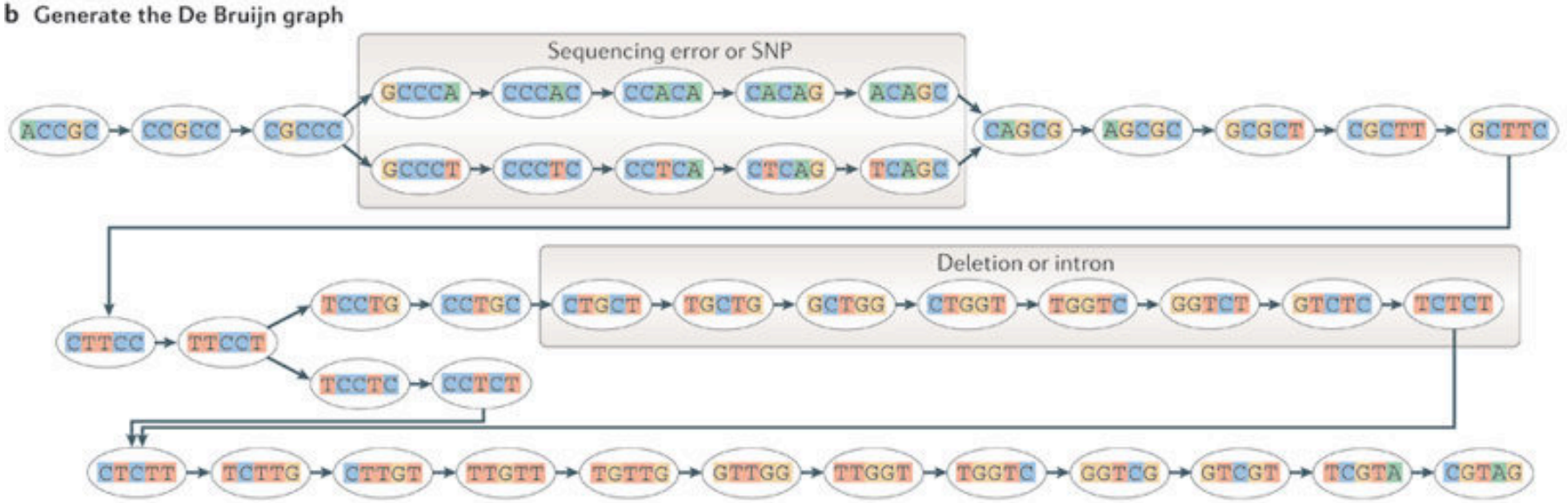
- Short reads too numerous for OLC
- De Bruijn graph
  - Break all reads into k-mers. A read has L-k+1 k-mers.
  - Construct graph
    - Nodes = distinct k-mer
    - Edges = k-1 exact overlap between two nodes.

ACGTTAGT

ACGTTA -> CGTTAG -> GTTAGT

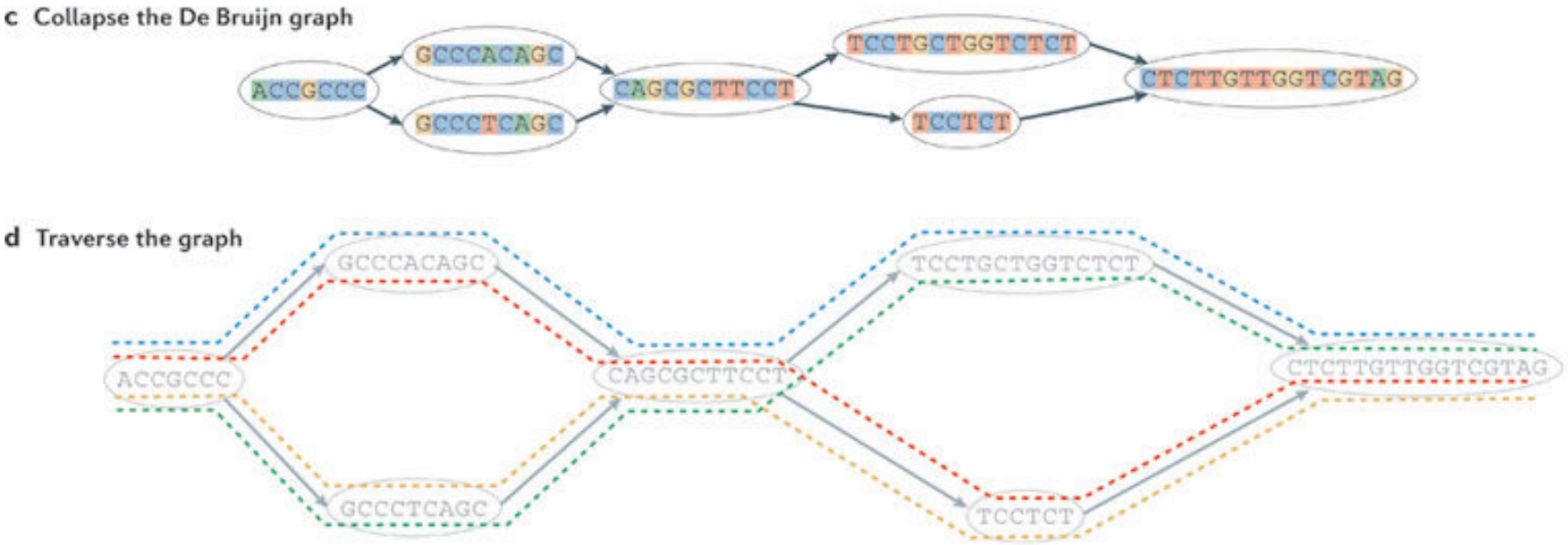
  - Simplify the graph
    - Remove bubbles, tips, and poorly supported edges.
  - Find Eulerian path through the graph (Every edge visited once - faster than finding Hamiltonian path)

Methods of assembly



Methods of assembly

SciLifeLab



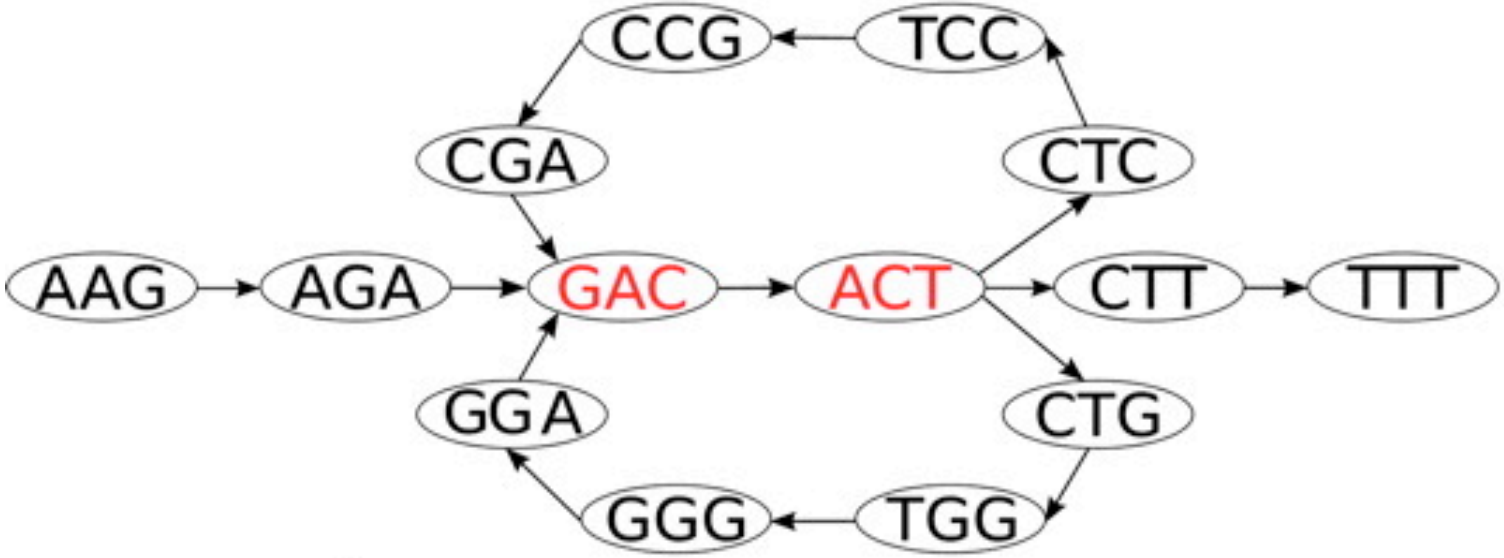
Martin and Wang 2011

Repeats

SciLifeLab

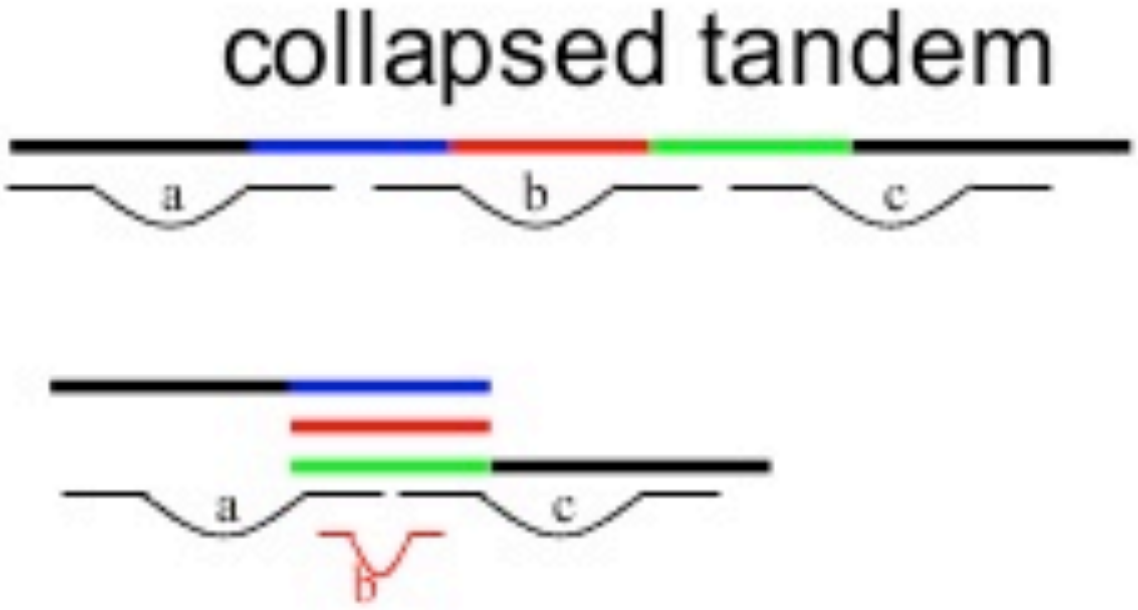
- The biggest problem for short read assembly are repeats.

AAGACTCCGACTGGGACTTT

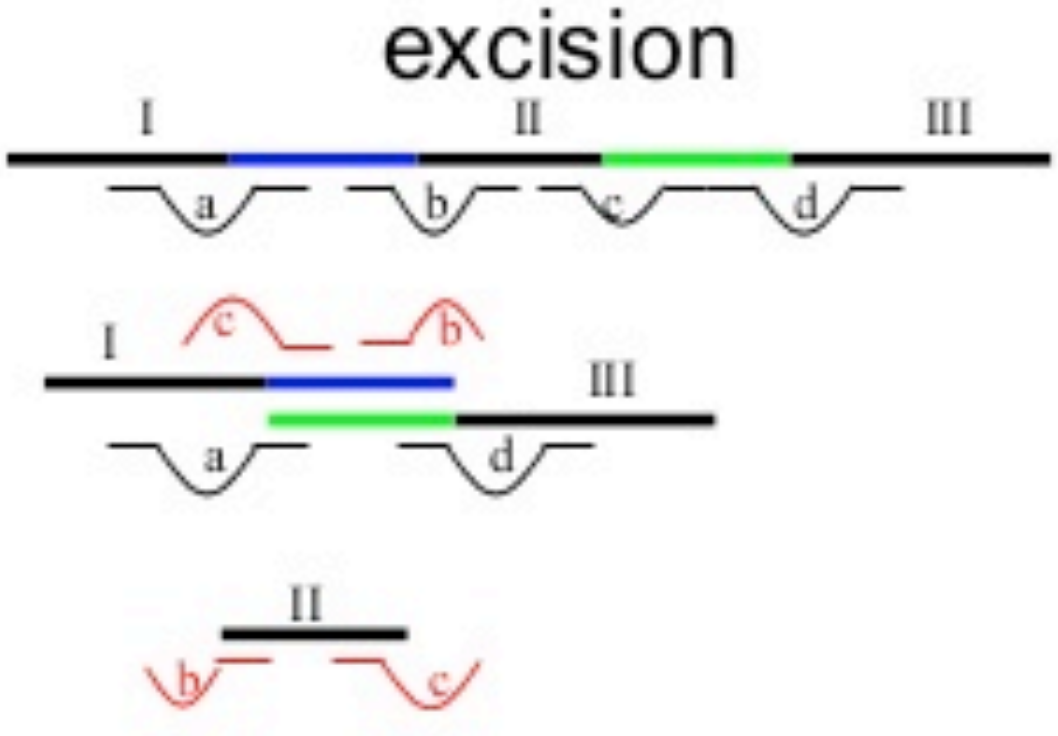


A de Bruijn graph of a sequence

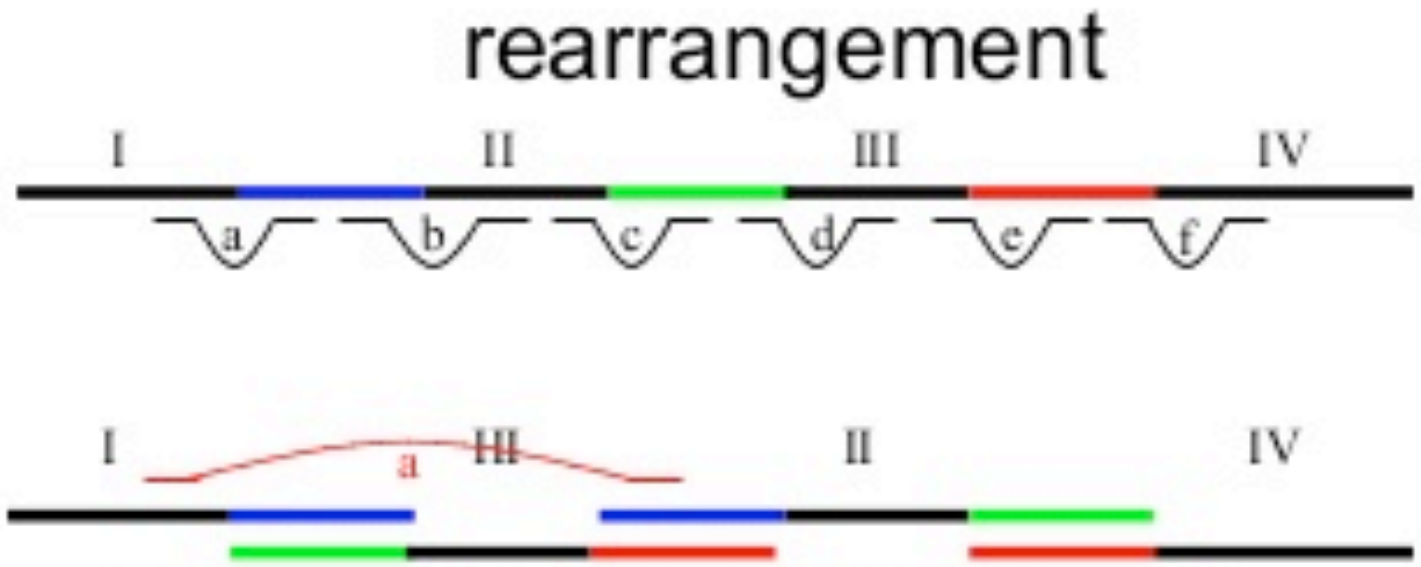
Repeats



Repeats



Repeats



Selecting k-mer size

SciLifeLab

- De Bruijn Graph simplification is highly affected by your choice of k-mer size.
  - Some tools explore multiple k-mer sizes
  - Others only use one k-mer size.

- Small k-mer sizes result in lots of connections.

- Large k-mer sizes help resolve complex regions
  - If the k-mer size is larger than the repeat, the repeat can be anchored in unique sequence on both sides.
  - Errors mean connections between k-mers may be lost.

- Size selection can by guided by tools like K-mer Genie

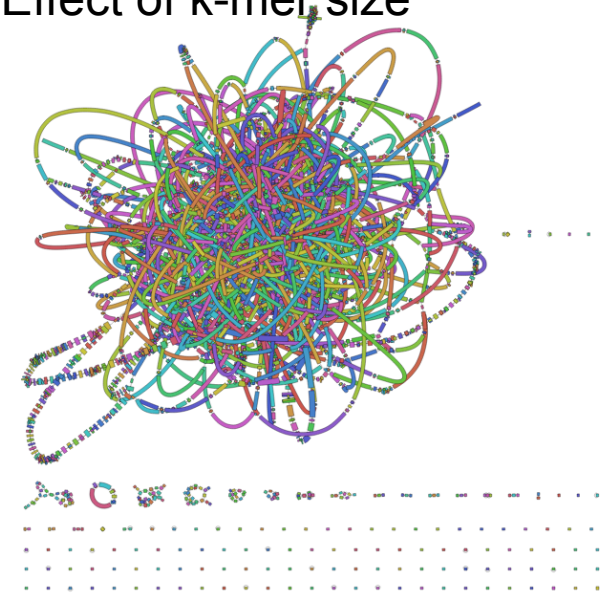
- Best k depends entirely on the properties of your genome and the error rate in the reads.

Selecting k-mer size

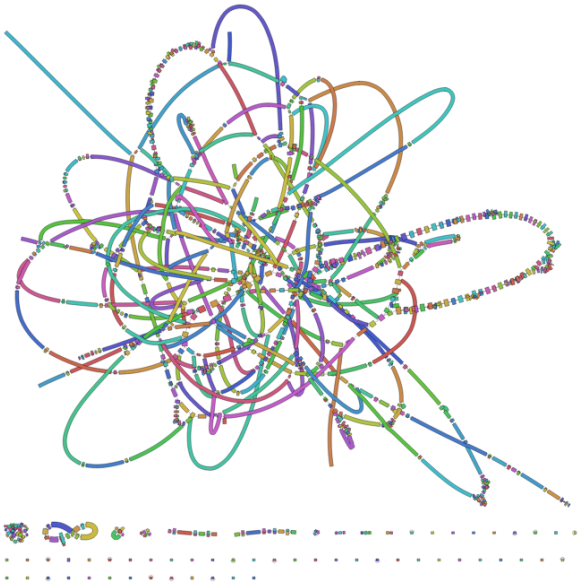
SciLifeLab

R. Wick. Effect of k-mer size

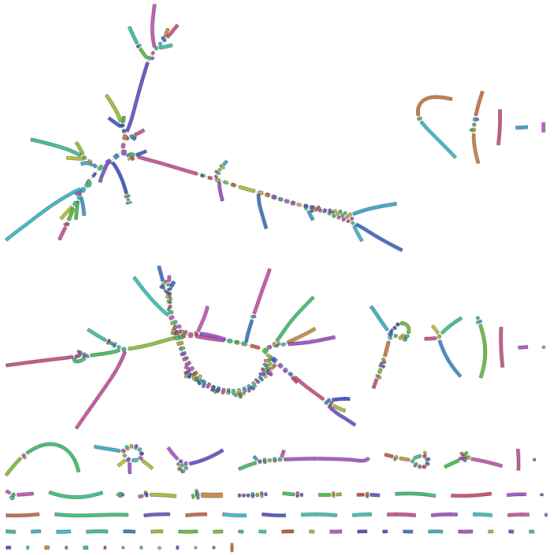
k = 51



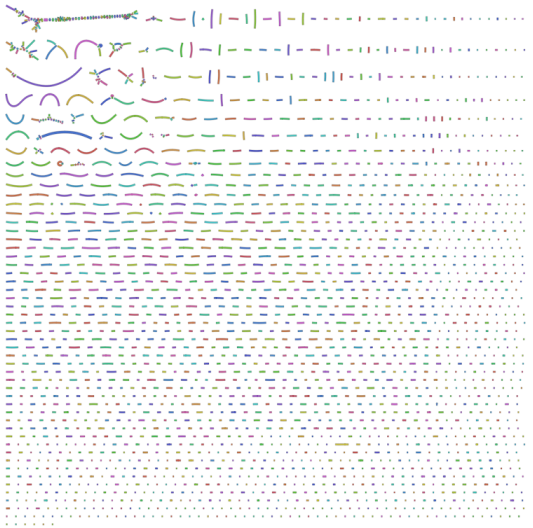
k = 61



k = 81



k = 91



Assembly Tools

SciLifeLab

- Many short read assemblers available
  - Spades
    - Often produces the best assembly for small genomes.
    - Explores multiple k-mer sizes and merges results.
    - Handles single cell, plasmid, and meta-genome assembly too.
    - Can use long reads in scaffolding.
  - AllPaths-LG
    - Suitable for large genomes
    - Requires specific Illumina libraries
  - MaSuRCA
    - Suitable for very large genomes
    - Sometimes over-assembles
  - Abyss
    - Suitable for very large genomes
  - Soap de novo
    - Suitable for large genomes

Assembly Tools

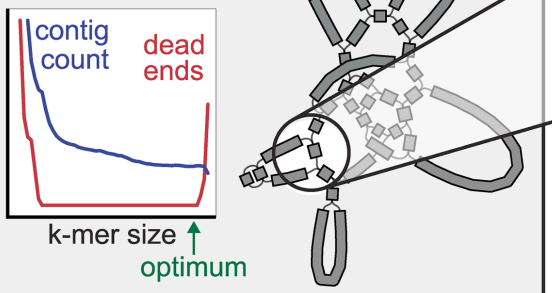
SciLifeLab

- Honorable mentions
  - Discovar - requires specialized overlapping PCR-free Illumina libraries
  - Velvet - Formerly popular bacterial assembler
  - SGA - Experimental string graph assembler
  - Mira - Flexible assembler for many different platforms
  - Platanus - Assembler that claims to work with high heterozygosity
  - CLC Workbench (not free like some others) - often performs well in comparative assessments
  - And many more …

Assembly Tools - Unicycler

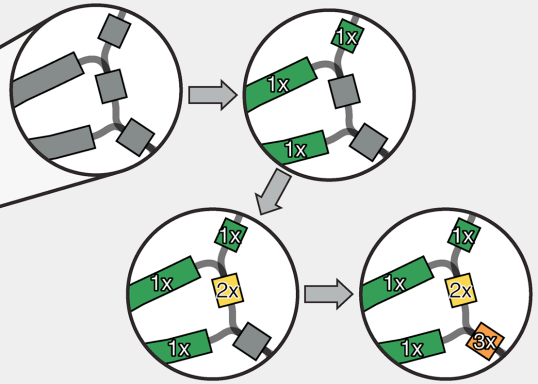
SciLifeLab

A. Short read assembly with SPAdes



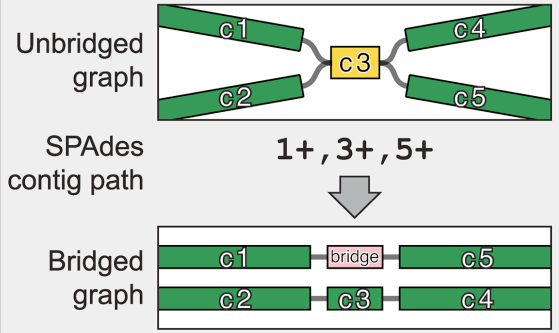
A thorough sweep of k-mer sizes finds an optimal assembly graph with few dead ends.

B. Multiplicity



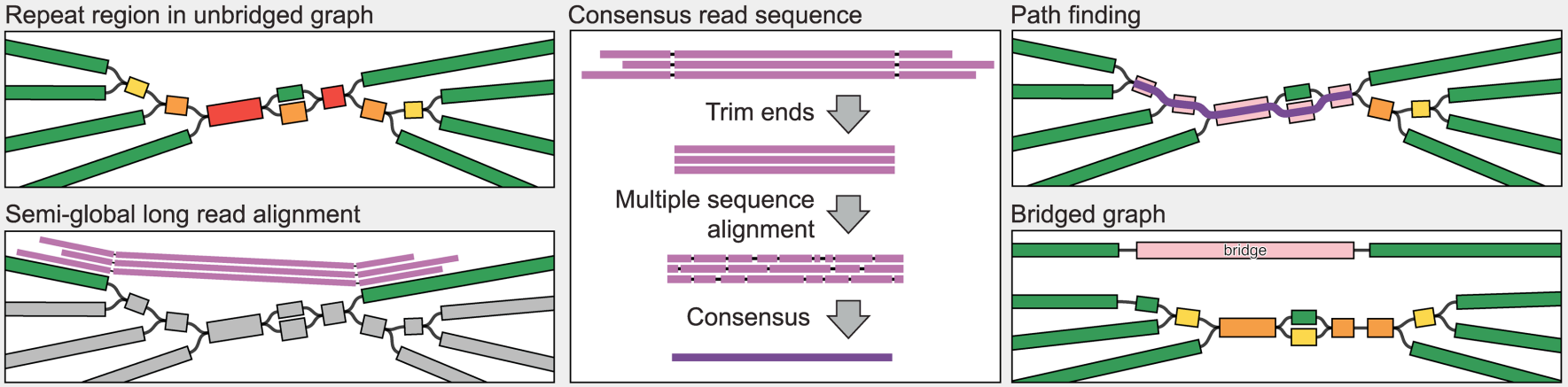
A greedy algorithm assigns copy numbers to contigs using depth and graph connections.

C. Short read bridging



Bridges simplify the graph by resolving repeats between single-copy contigs. Short read bridges are made from SPAdes paths.

D. Long read bridging

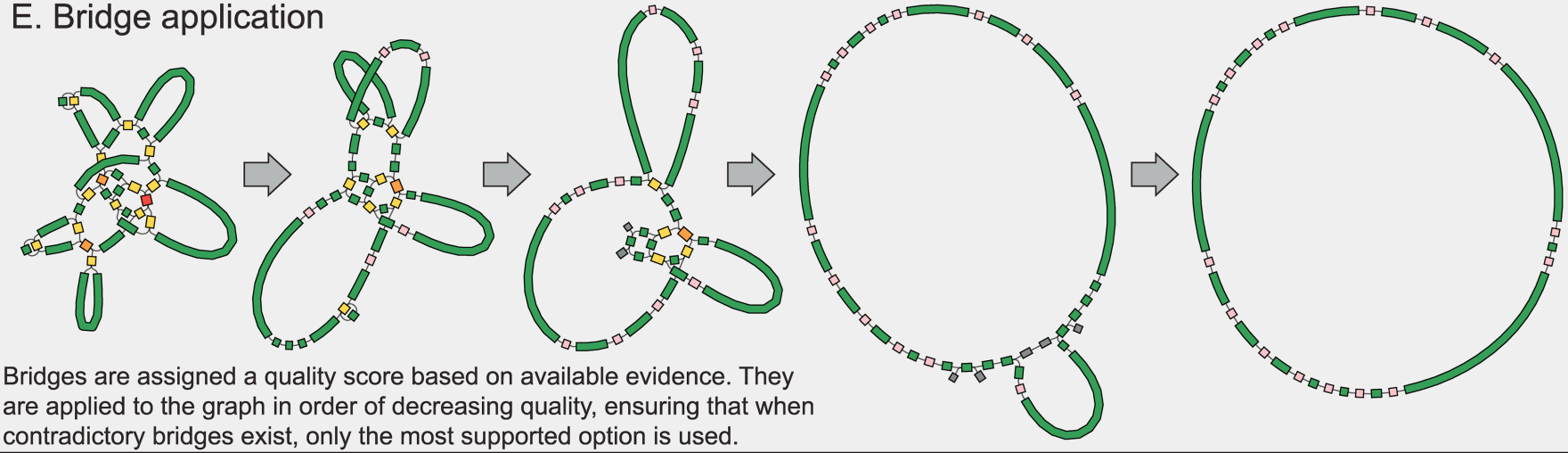


Bridges made using long reads can resolve larger repeats than short-read bridges. They are made from long reads which align to two or more single-copy contigs. The bridge sequence comes from the graph path between the two contigs, not the long reads, providing greater accuracy. When multiple possible bridge paths exist, the best path is chosen based on agreement with the long-read consensus sequence.

Assembly Tools - Unicycler

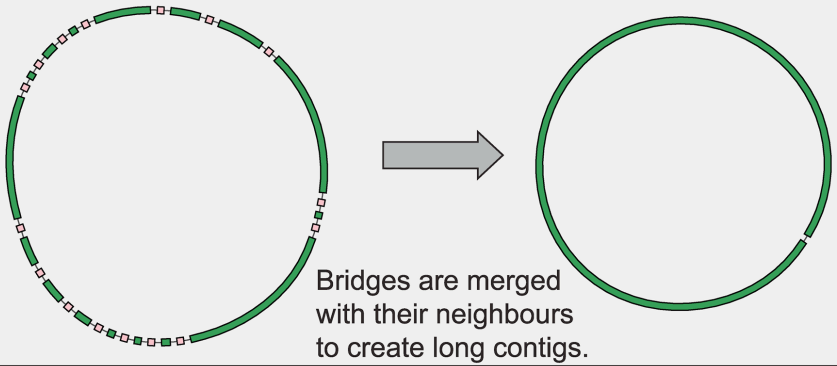
SciLifeLab

E. Bridge application



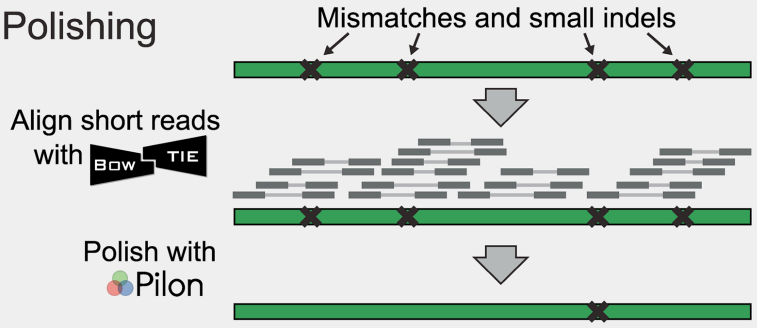
Bridges are assigned a quality score based on available evidence. They are applied to the graph in order of decreasing quality, ensuring that when contradictory bridges exist, only the most supported option is used.

F. Contig merging



Bridges are merged with their neighbours to create long contigs.

G. Polishing



The final assembly is polished using the accurate short reads to reduce the rate of mismatches and small insertions/deletions.

Assembly Tools

SciLifeLab

- Error correction is often a primary stage in many assemblers
  - Assemblers:
    - Spades
    - Allpaths
    - MaSuRCA
  - Datasets for these assemblers should not be filtered (except adapter trimmed)

- Stand alone short read error correction tools
  - Quake
  - Reptile
  - QuorUM
  - LoRMA
  - Hammer
  - …

Preliminary Assembly Diagnostics

SciLifeLab

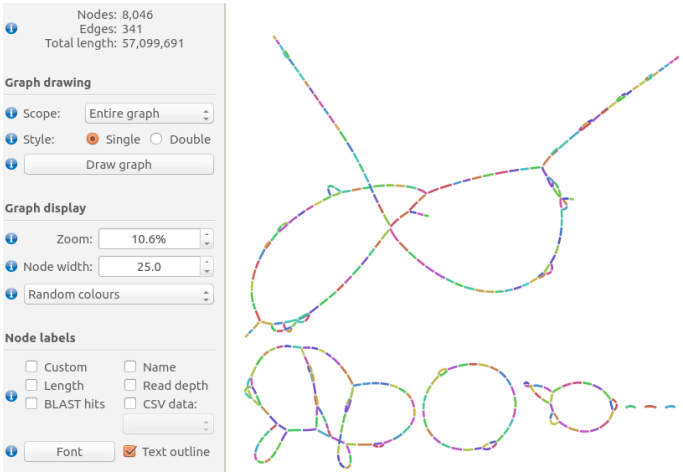
- Assembly Size
  - Assemblathon Script (https://github.com/KorfLab/Assemblathon)
  - Quast

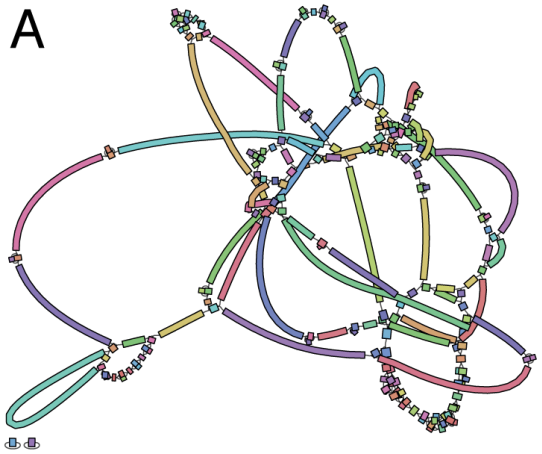
Number of scaffolds	556	
Total size of scaffolds	31318563	← Expected genome size
Longest scaffold	447934	
Shortest scaffold	8580	
Number of scaffolds > 1K nt	556	100.0%
Number of scaffolds > 10K nt	555	99.8%
Number of scaffolds > 100K nt	38	6.8%
Number of scaffolds > 1M nt	0	0.0%
Number of scaffolds > 10M nt	0	0.0%
Mean scaffold size	56328	
Median scaffold size	43995	
N50 scaffold length	60037	
L50 scaffold count	152	

Preliminary Assembly Diagnostics

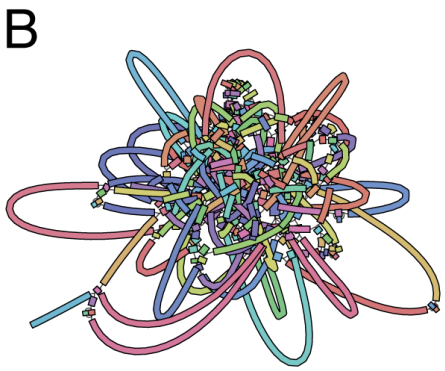
SciLifeLab

- Assembly Graph
  - Check connectedness of contigs
    - Bandage
    - Created from GFA output
    - Not supported by all assemblers

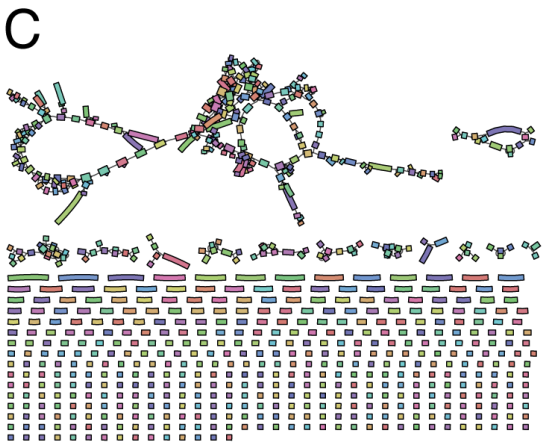




Good



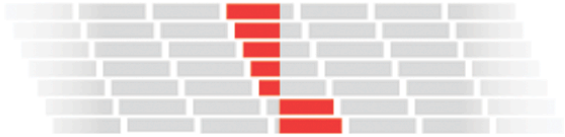
OK



Bad

Assembly Polishing

SciLifeLab

PROCESS	RESULT	
Pilon protocol	Assembly improvement (Fasta)	Variation detection (VCF)
Evaluate alignment pileups TAATGGGGGCGGTGCCATATCATGAGA TAATGGGGGCGGTGCCATATCATGAGA TAATGGGG*CGGTGCCATATCTAGAGA TAATGGGGGCGGTGCCATATCATGAGA	Identify and fix base errors	Identify SNPs and small indels
Scan read coverage and alignment discrepancies 	Identify potential local misassemblies	Identify larger insertions and deletions
Reassemble across gaps and discrepant regions 	Attempt to fill gaps and fix local misassemblies	Attempt to build out the full sequence of larger insertions

Summary

SciLifeLab

- Check whether your reads need pre-processing or if the assembler does it.
- Use a variety of assemblers (and settings).
  - Some assemblers handle certain properties in genomes better than others.
  - Explore multiple values of k-mer size for de Bruijn graph assemblers.
- Check the basic assembly statistics.
- Check assembly graph properties if you can.
- Select your best assemblies.
- Polish your assemblies
- Assess correctness.