Genome assembly strategies

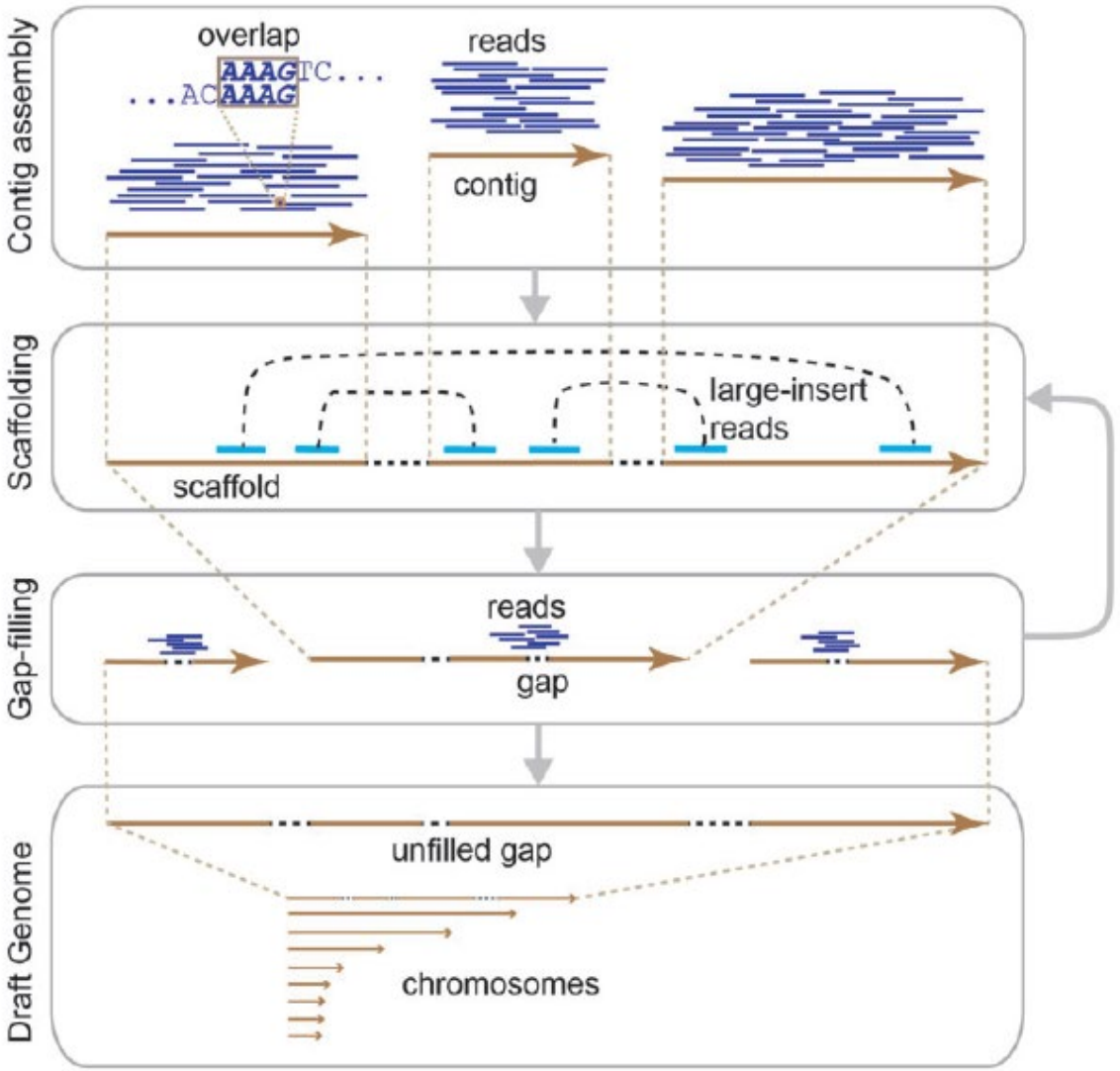
Arturo Vera Ponce de Leon

June 2021

arturo.vera.ponce.de.leon@nmbu.no

Genome assembly

Genome assembly



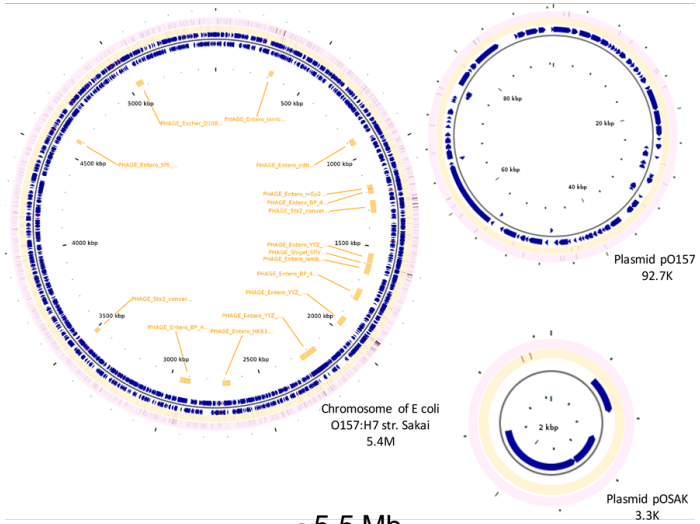
Reads
(fastq)

Summarizing

Escherichia coli

Circular chromosome

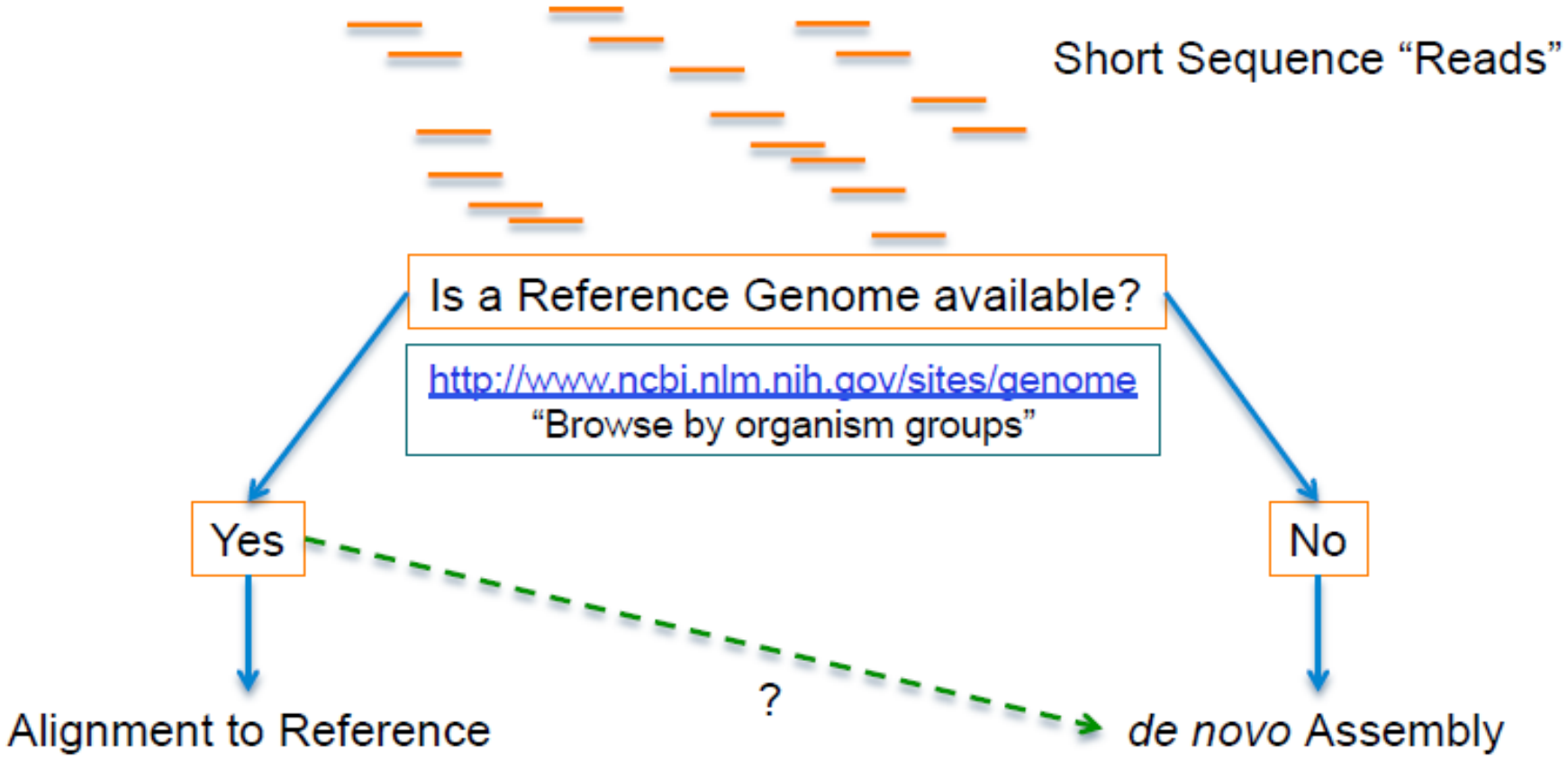
Plasmids



~5.5 Mb

Briefings in Bioinformatics, 2016, 1–18

Genome *de-novo* assembly or reference mapping



Genome Assembly

Work flow to classic genome assembly strategies

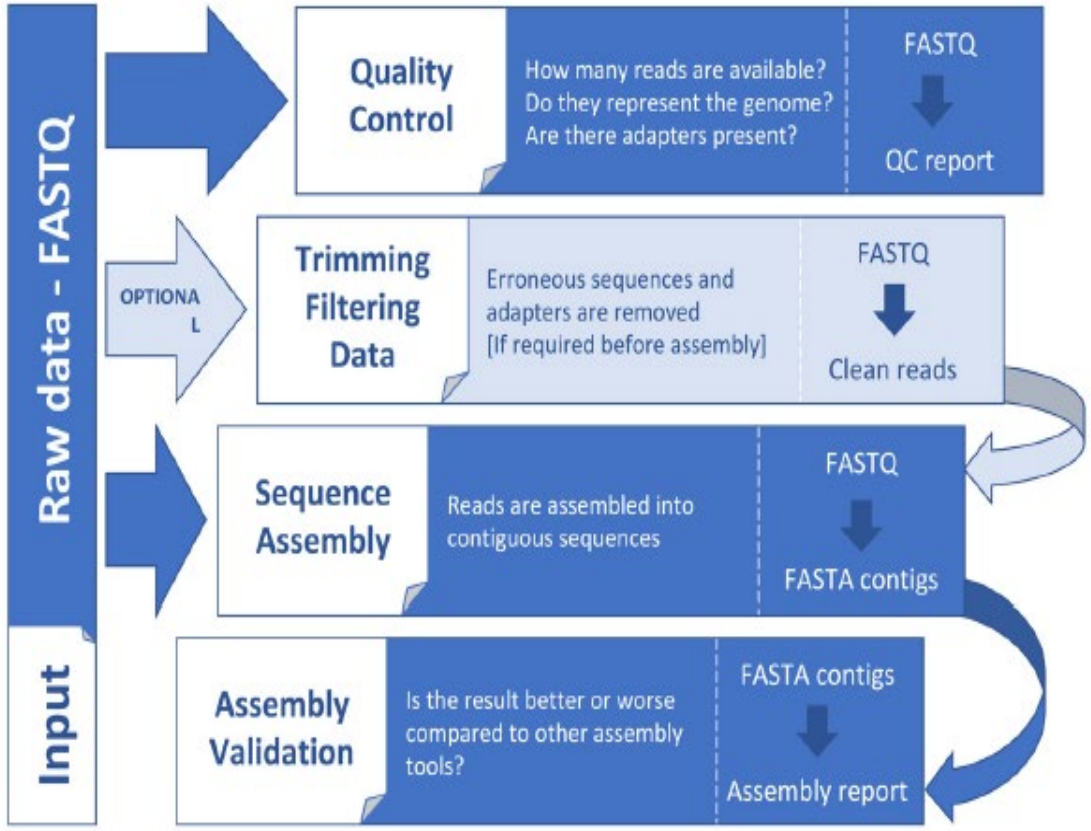


Figure 2. General steps in a genome assembly workflow. Input and output data are indicated for each step.

F1000Research 2018, 7(ELIXIR):148

Software used in this lecture

fastQC
NanoPlot

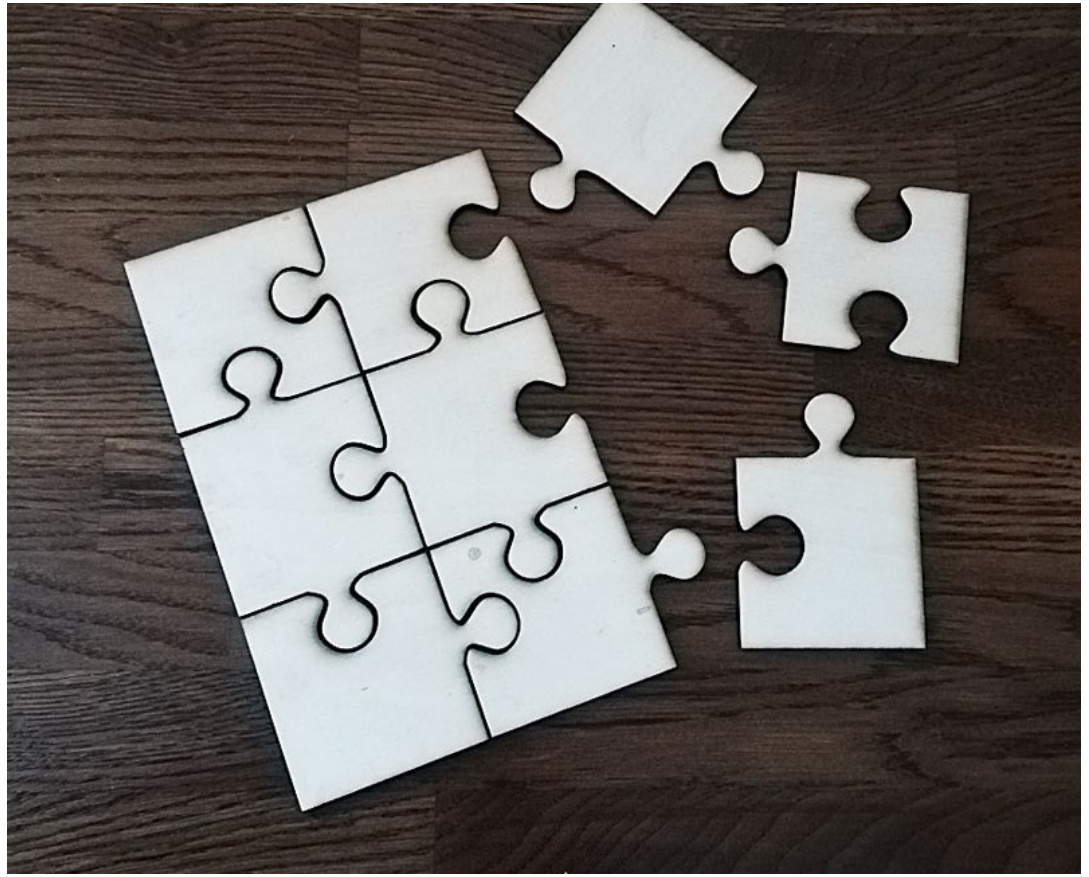
TrimGalore

SPADES
Unicycler

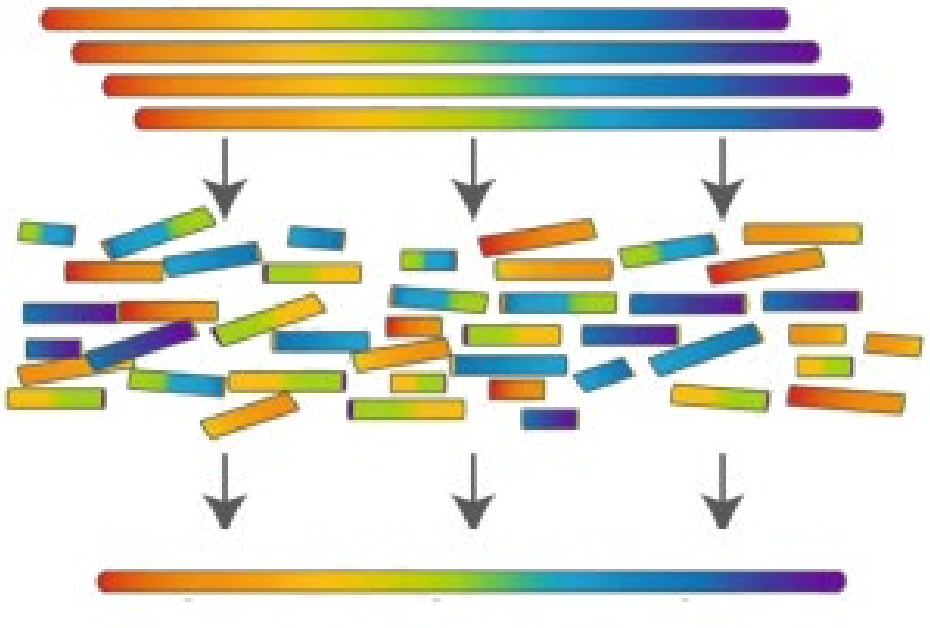
BUSCO

Steps of the Assembly (recap):

1. Find all **overlaps** between reads
2. Build a **graph** (read connections)
3. **Simplify** the graph
4. Find a sensible path in the graph to generate a **consensus**



Assembly - expectation



Genome copies

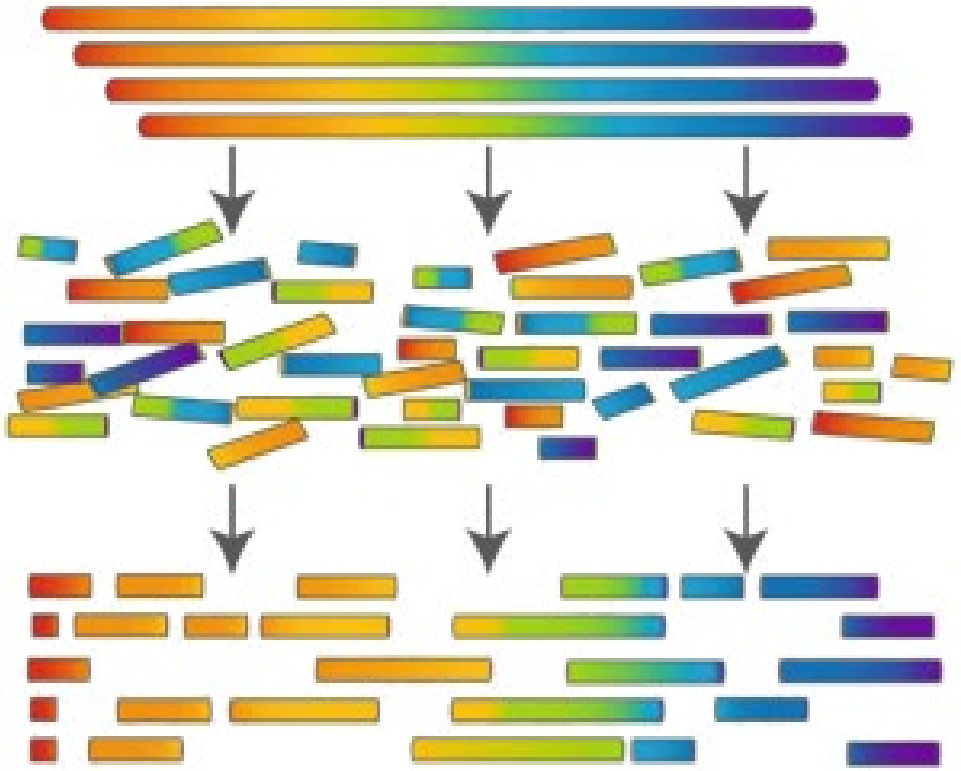
Reads

Reconstructed genome

Picture adapted from:
Commins, et al. (2009) Biological Procedures Online

Assembly - Reality

Especially true for short reads, such as those from Ilumina sequencing (150 - 300 bp)

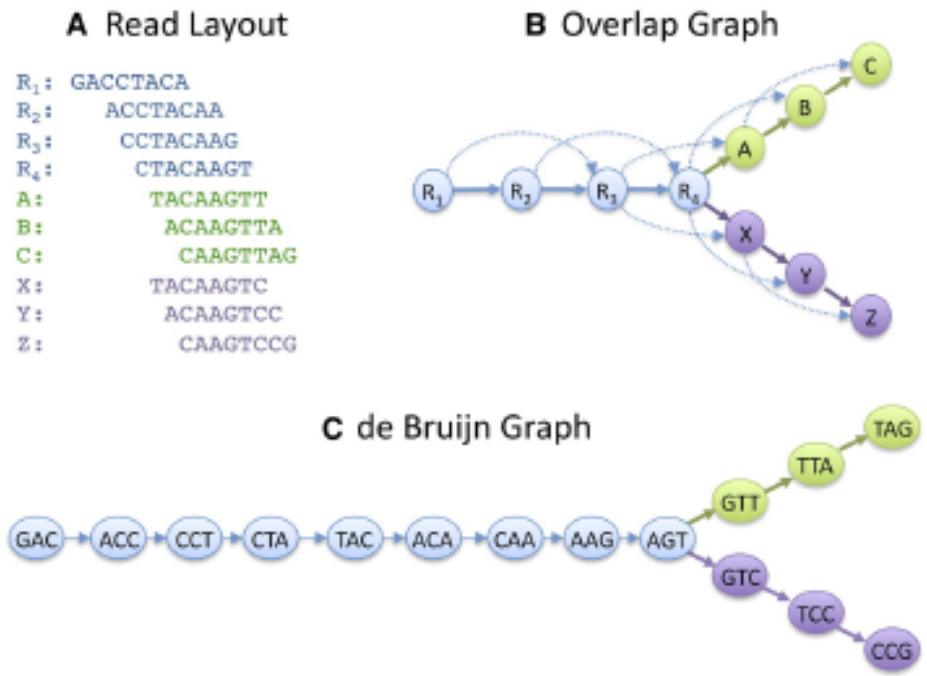


Picture adapted from:
Commins, et al. (2009) Biological Procedures Online

Strategies to genome assembly

- Algorithms

1. Greedy
2. Overlap-layout-consensus (OLC)
3. De Bruijn Graph

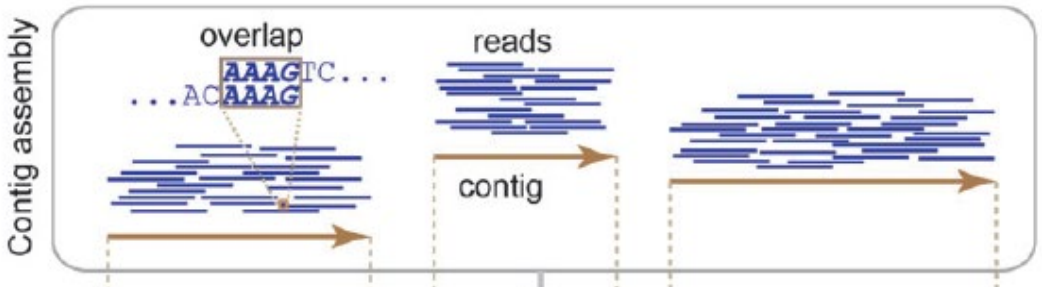


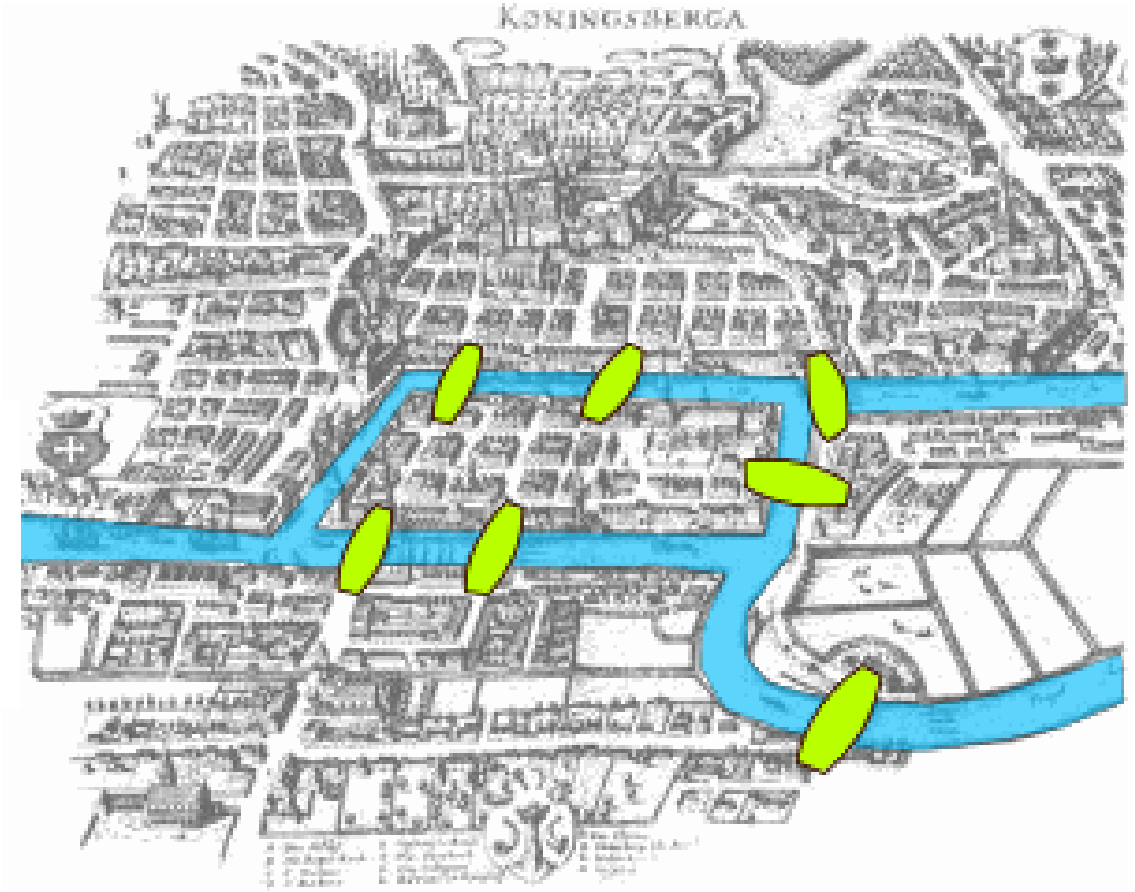
Schatz M C et al. Genome Res. 2010;20:1165-1173

CSH PRESS

Steeps to genome assembly using De-Bruijn graph

The basic strategy for de novo assembly for short NGS reads comprises three steps: (i) contig assembly, (ii) scaffolding and (iii) gap filling.





Seven Bridges of Königsberg

The K-mers: divide and conquer

- It breaks reads into successive k-mers and the graph maps the k-mers
- Each k-mer is a node and edges are drawn between each k-mer in a read.
- Repeat sequences create a fork in the graph; alternative sequences create a bubble.
- The k-mer size can only be determined by “trial and error”.
- A small value of K will create a complex graph but a large value of K may miss small overlaps. A good starting point would be a k-mer size that is 2/3 the size of the read
- Good for short reads or small genomes. With long reads and/or large genomes, may require lots of RAM (e.g., ~0.5 TB for human)

K-mers

- k-mers are subsequences of length ***k***

For sequence ATCG;

k=1 is A – T – C – G

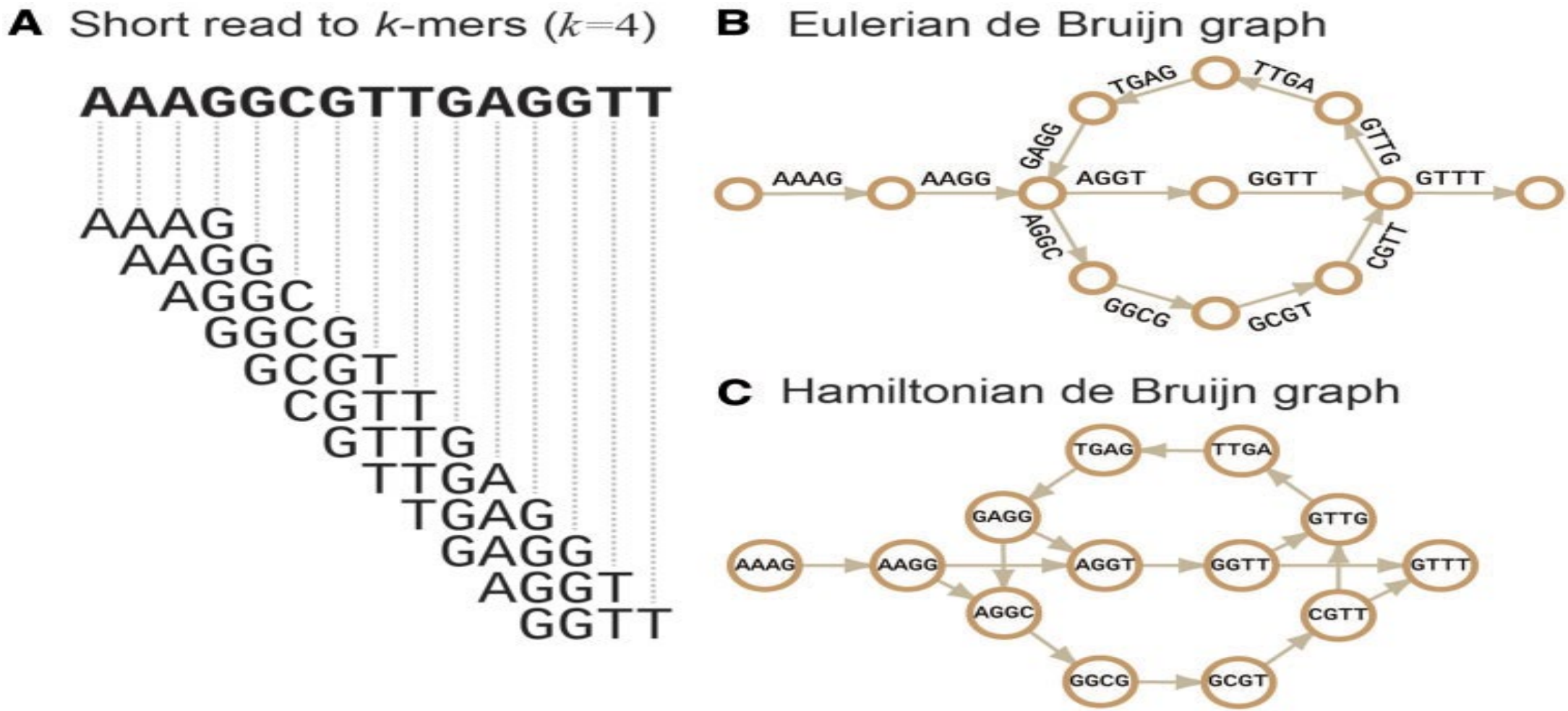
k=2 is AT – TC – CG

Assemblers (and other bioinformatic tools) are often based on k-mers

***k*-mers for GTAGAGCTGT**

k	***k*-mers**
1	G, T, A, G, A, G, C, T, G, T
2	GT, TA, AG, GA, AG, GC, CT, TG, GT
3	GTA, TAG, AGA, GAG, AGC, GCT, CTG, TGT
4	GTAG, TAGA, AGAG, GAGC, AGCT, GCTG, CTGT
5	GTAGA, TAGAG, AGAGC, GAGCT, AGCTG, GCTGT
6	GTAGAG, TAGAGC, AGAGCT, GAGCTG, AGCTGT
7	GTAGAGC, TAGAGCT, AGAGCTG, GAGCTGT
8	GTAGAGCT, TAGAGCTG, AGAGCTGT
9	GTAGAGCTG, TAGAGCTGT
10	GTAGAGCTGT

The K-mers: divide and conquer

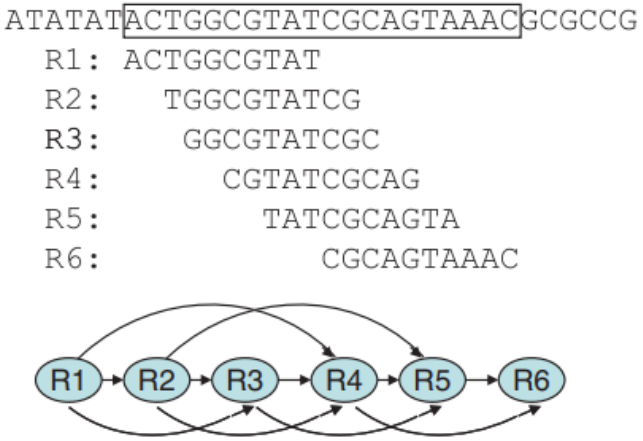


De novo assembly – graphs

De novo assembly is assembly without any prior knowledge (such as reference genomes)

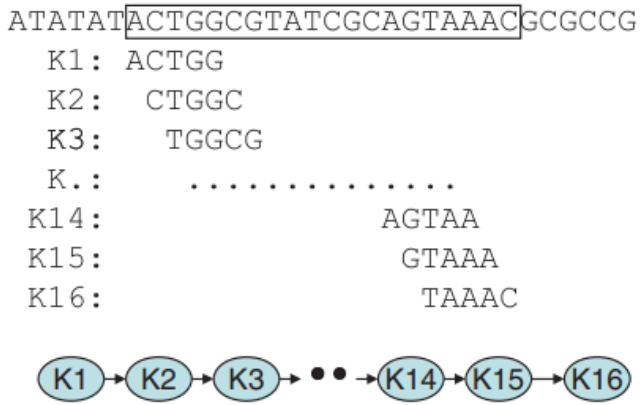
Overlap graphs (or Overlap-Layout-consensus, OLC)

- All-reads vs all-reads alignments
  - Provides connections between reads
- Need long reads → Nanopore data



***De Bruijn* graphs**

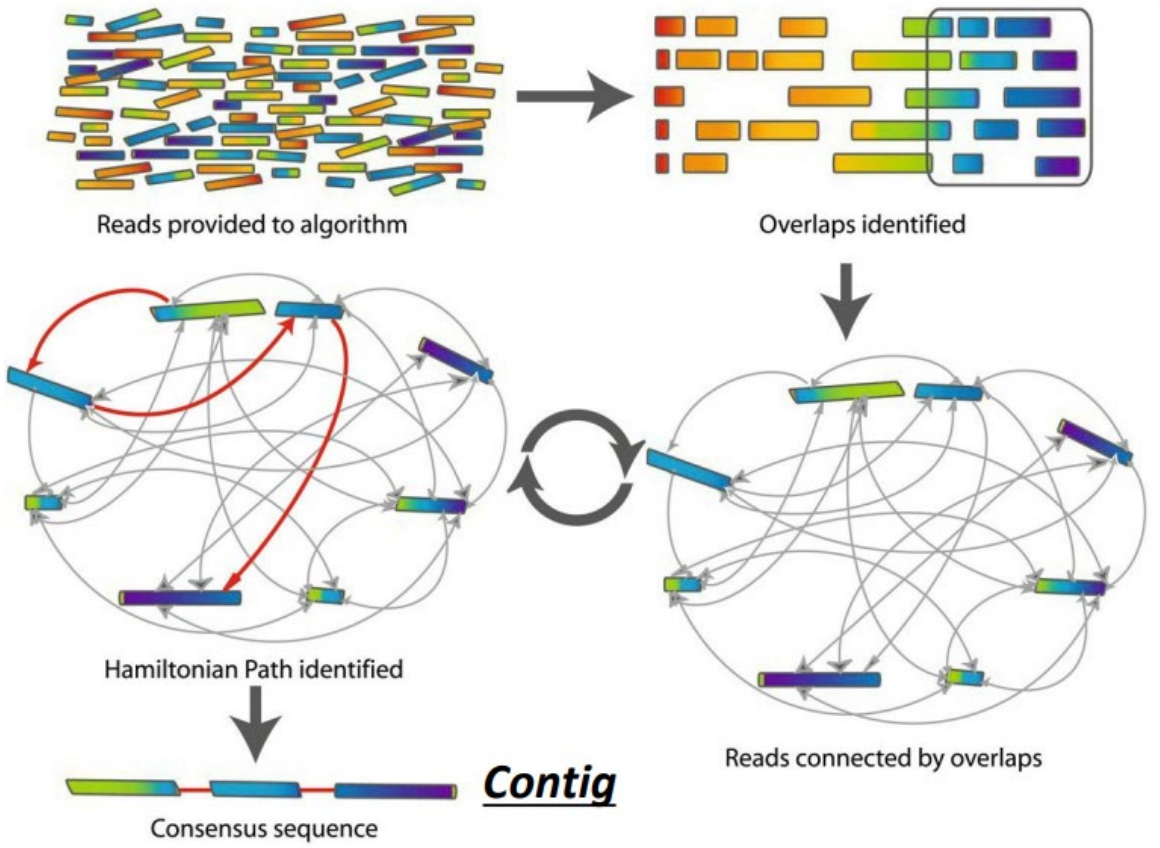
- The dominating approach for many years
- Optimal for short reads with high coverage → Illumina data
- Some long-reads assemblers use a «fuzzy» version of the *de Bruijn* graph



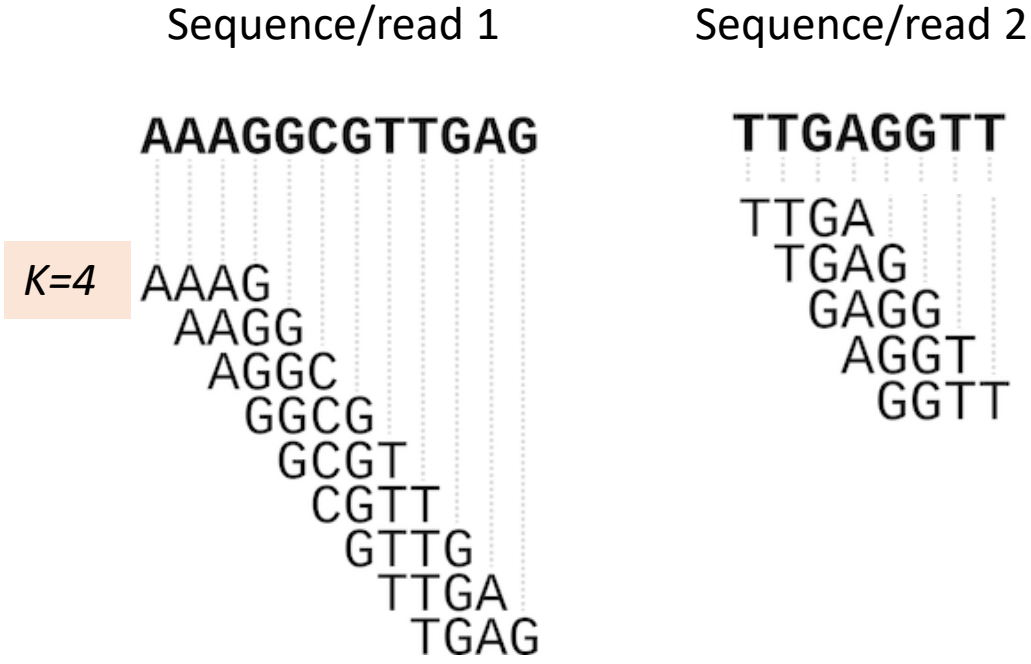
Overlap graphs (or Overlap-Layout-consensus, OLC)

OLC is composed of three main steps:

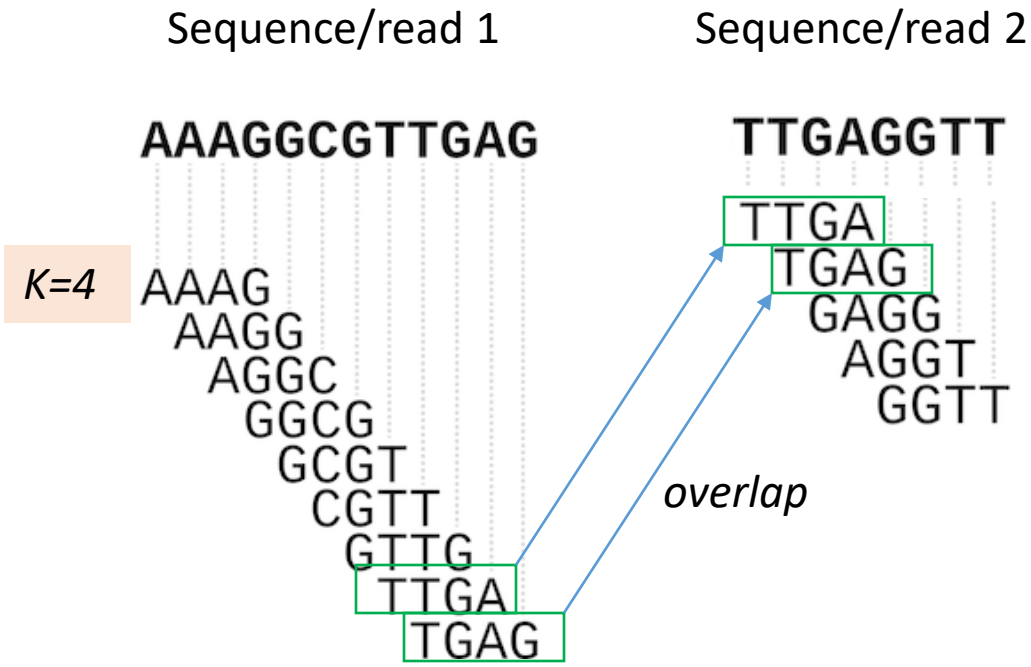
1. Computing the overlaps between the reads (blast-like alignment or k-mer)
2. Laying out the overlap information on a graph – try to simplify
3. Generate a consensus from the alignment



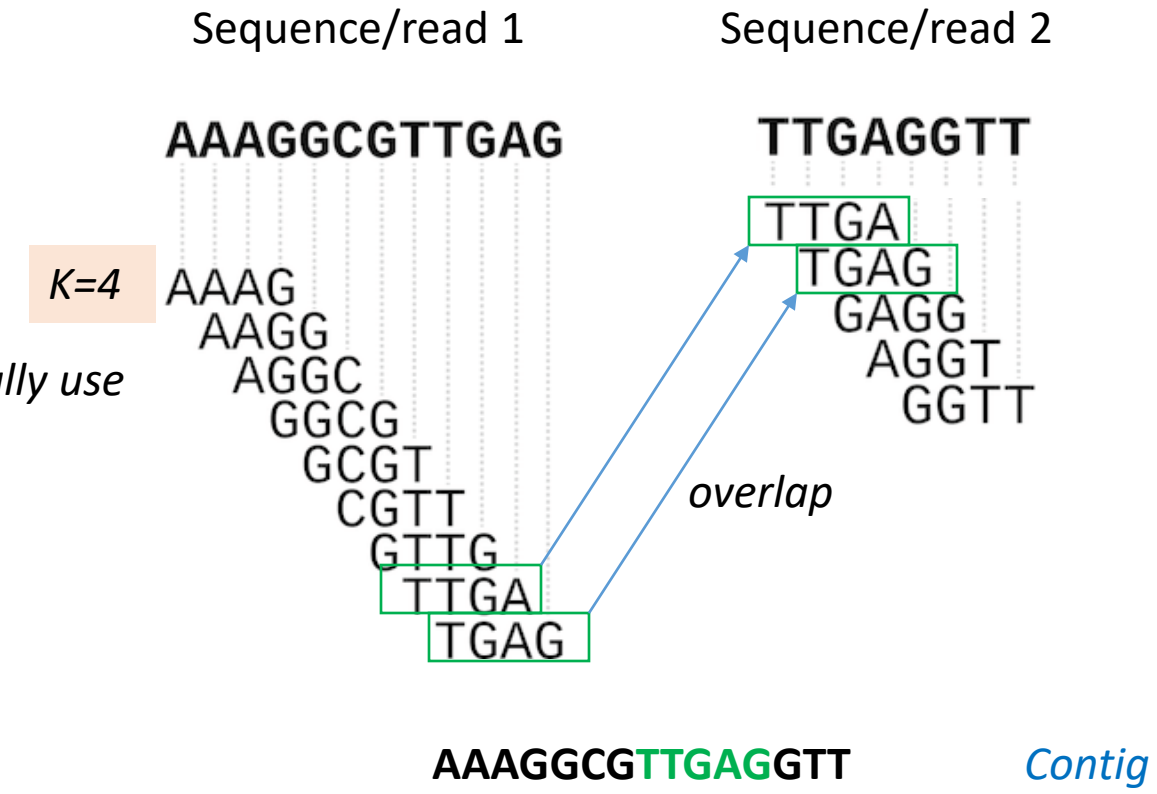
De Bruijn graphs – **simplified**



De Bruijn graphs – **simplified**



De Bruijn graphs – **simplified**



*NB: Assemblers usually use **k** greater than 31*

Evaluating the assembly

- **Genome assembly results:**
  - **contig size and number of contigs produced**
  - **scaffold size and number**
  - **N50 and N90**
- **Coverage**
- **GC Content**
- **Genome annotation**
  - repeats analysis and annotation
  - protein-coding gene annotation (including gene structure prediction and gene function annotation)
  - non-coding RNA gene annotation (including annotation of microRNA, tRNA, rRNA, and other ncRNA)
  - transposon and tandem repeats annotation
- Comparative genomics and evolution (chromosome structure, conserved gene families)

Basic stats

Basic statistics

N50 the length of the shortest contig such that the sum of contigs of equal length or longer is at least 50% of the total length of all contigs.

Contig size (bp)
3000
2000 N50
1200
800
600 N90
400
Total: **8000**

N90 = the length of the shortest contig such that the sum of contigs of equal length or longer is at least 90% of the total length of all contigs.