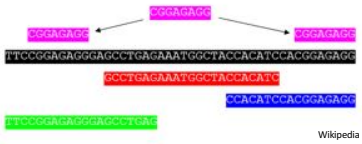
- *De novo* genome assembly
- De Bruijn graphs
- Genome assembly software
- Exercise: assemble a genome

Assembling Short Reads

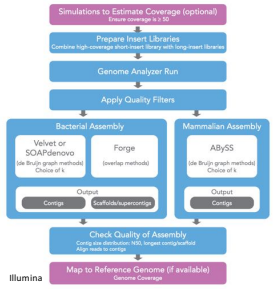


Wikipedia

De-novo vs. mapping assembly

- **de-novo**: assembling short reads to generate a near full-length sequence.
- **mapping:** assembling reads against a reference sequence. The sequence may be similar but not necessarily identical to the reference sequence.

Genome Assembly Workflow



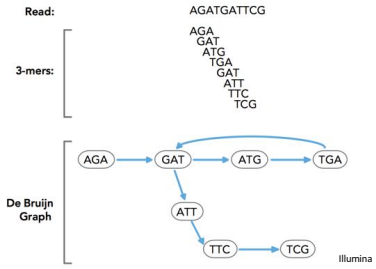
Illumina

Genome Assembly Software

Algorithm	Description	Strength	Genomes Assembled
Velvet	De Bruijn graph based Error corrections after graph is built	Fast (~30 mins) Easy to use Larger supercontig N50	Bacterial (Ref. 1; this technical note)
SOAPdenovo	De Bruijn graph based Error correction before graph is built	Easy to use Multi-threaded mode	Panda, Bacterial (Ref. 11; this technical note)
ABySS	De Bruijn graph based Can be run in parallel Distributed memory model (efficient)	Easy to use Largest contigs/scaffolds Best suited for large genomes	Human (Ref. 3; this technical note)
Forge	Overlap-layout-consensus method Modifications to accommodate Illumina reads	Largest contigs/supercontigs Good "long read" assembler	Bacterial (this technical note)

Illumina

De Bruijn Graph

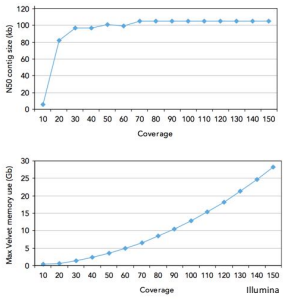


Illumina

Stats

- **N50:** The contig length at which 50% of the assembled genome is in contigs of N50 or greater
- **Coverage:** The percentage of the genome covered by contigs. Requires a reference genome to calculate.
- **Max contig length:** longest contig
- **Min contig length:** shortest contig. Often determined by assigning a threshold.
- **Average contig length:** average contig size. Often very small contigs are removed before calculating;

Coverage



Exercise

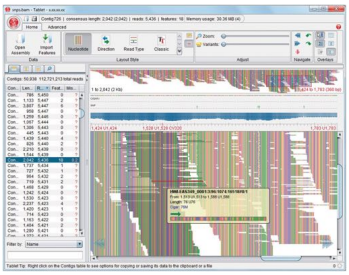
Assemble contigs using Velvet for the *E. coli* genome using paired-end high-throughput sequencing data generated with Illumina.

What we can except from our data:

Coverage	N50 contig size	Largest contig	Genome coverage
320×	95,313 bp	215,645 bp	99.47%
160×	95,368 bp	209,234 bp	99.72%
50×	97,333 bp	223,793 bp	99.72%
21×	35,828 bp	119,071 bp	99.38%

Illumina

Tablet for Viewing NextGen Sequencing Data



Milne et al (2013)