Principles and problems of de novo genome assembly

Karin Lagesen

Norwegian Veterinary Institute

Material adapted from slides provided by Lex Nederbragt

What is this thing called 'genome assembly'?

What is a genome assembly?

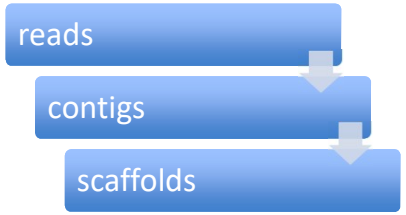
A hierarchical data structure

that maps the sequence data

to a putative reconstruction of the target

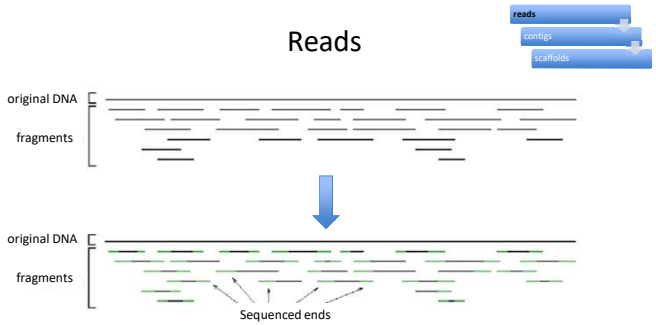
Miller et al 2010, Genomics 95 (6): 315-327

Hierarchical structure



Sequence data

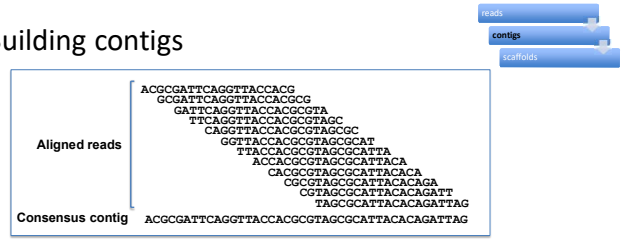
Reads



http://www.cbcb.umd.edu/research/assembly_primer.shtml

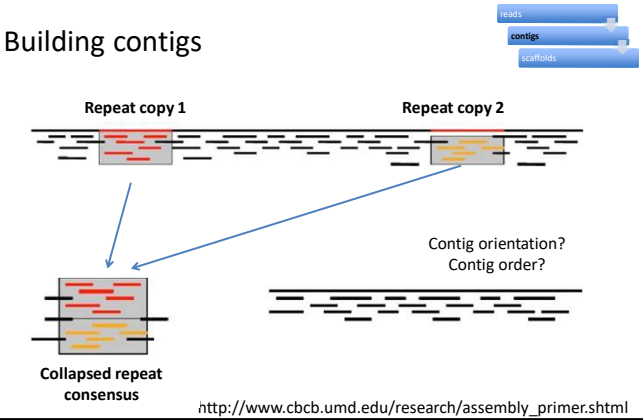
Contigs

Building contigs



Contigs

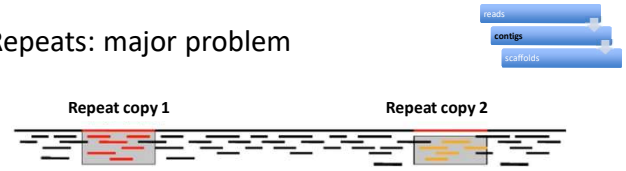
Building contigs



http://www.cbcb.umd.edu/research/assembly_primer.shtml

Contigs

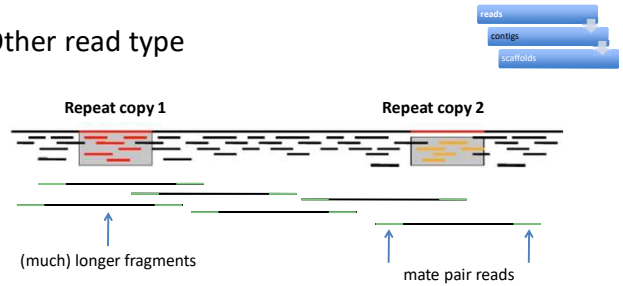
Repeats: major problem



http://www.cbcb.umd.edu/research/assembly_primer.shtml

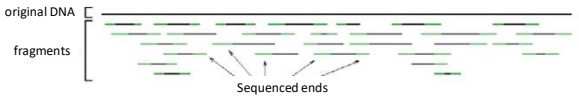
Mate pairs

Other read type

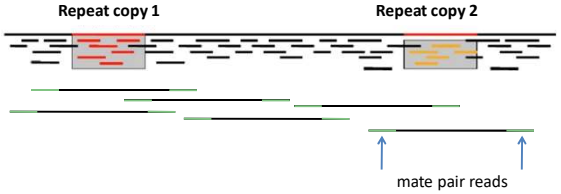


Mate pairs

Paired end reads → 100-500 bp insert

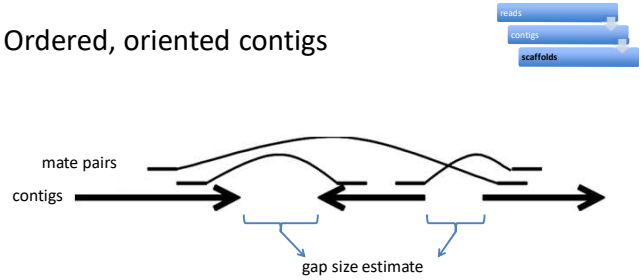


Mate pairs → 2-20 kb insert

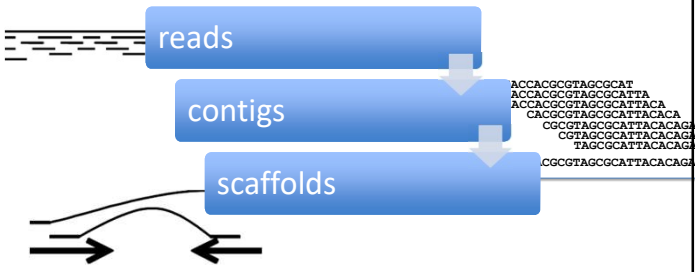


Scaffolds

Ordered, oriented contigs



Hierarchical structure



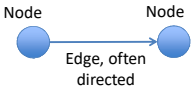
Assembly

How to do this?

Algorithms

Algorithms

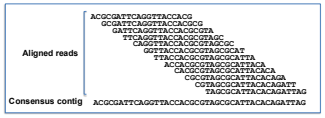
All are graph-based



Graph-theory!

Algorithms

All are graph-based

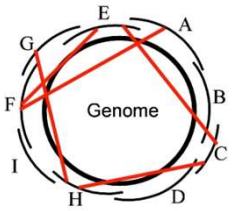


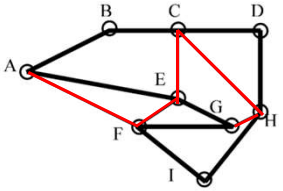
Graph-theory!

Algorithms

Hamiltonian path

– a path that contains all the nodes





https://www.cbcb.umd.edu/research/assembly_primer

Algorithms

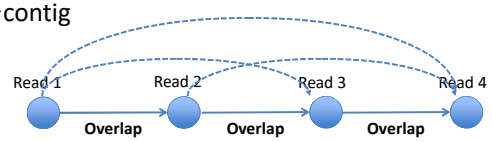
Overlap calculation (alignment)

– computationally intensive

Algorithms

Path through the graph

→contig



Algorithms

Many flavors

Abandoned

→ Greedy extension

Two most used

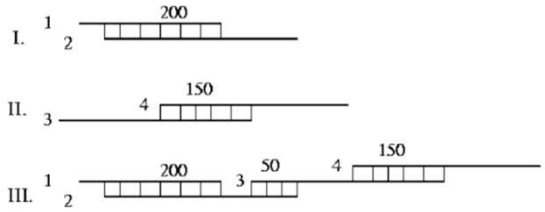
→ Overlap Layout Consensus

→ de Bruijn graph

http://www.waialuasodaworks.com/images/flavors2009.jpg

Greedy extension

Oldest

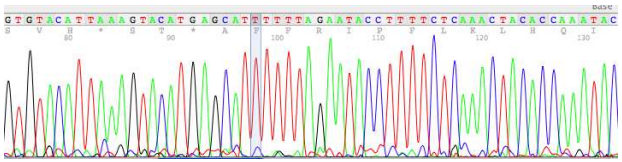


https://www.cbcb.umd.edu/research/assembly_primer

Overlap-Layout-Consensus

Typical for Sanger-type reads (longer reads)

– also used by canu



Overlap-Layout-Consensus

Steps

– Overlap computation

– Layout: graph simplification

– Consensus: sequence

Overlap-Layout-Consensus

Overlap phase: find “similar enough” reads

Comparing all against all: expensive

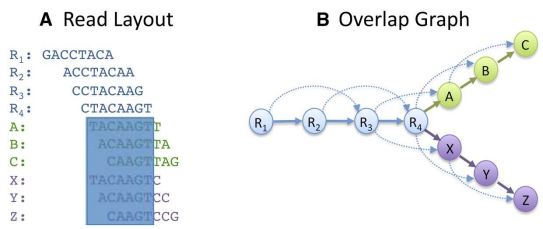
Trick for finding “similar enough” reads:

- Split reads into k-mers

K-mer: substring of length *k* from a longer string

- Make list over which read has which k-mers
- If two reads share k-mers, test for similarity

Overlap-Layout-Consensus

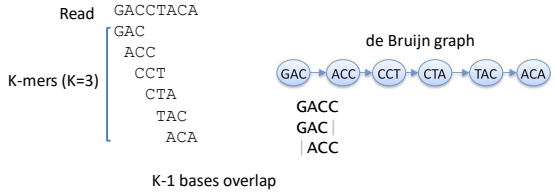


Schatz M C et al. Genome Res. 2010;20:1165-1173

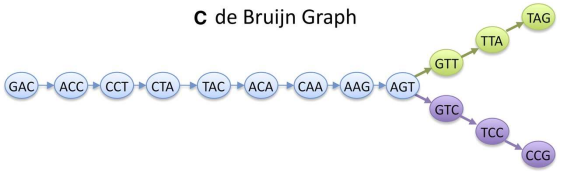
de Bruijn graphs

Developed outside of DNA-related work

– Best solution for short(er) reads

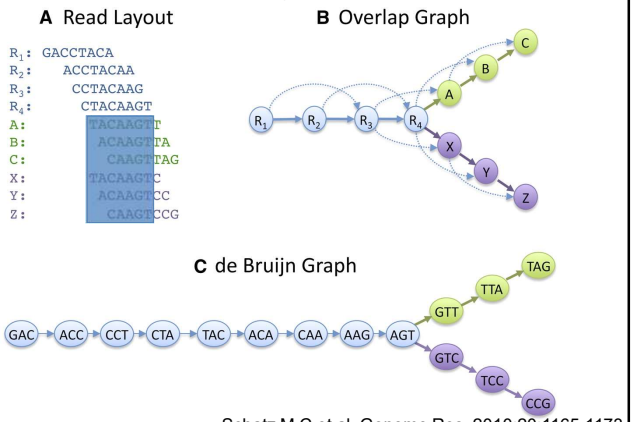


Graphs



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

Graphs

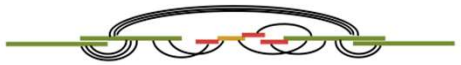


Schatz M C et al. Genome Res. 2010;20:1165-1173

Graphs

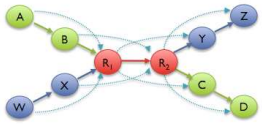
Simplify the graph

Add scaffolding information



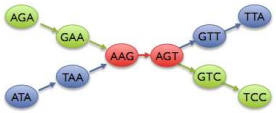
de Bruijn Graphs

Overlap Graph



Long read assemblers
- Repeats depends on read length
- Read coherency, placements kept
- Tangled by high coverage

de Bruijn Graph

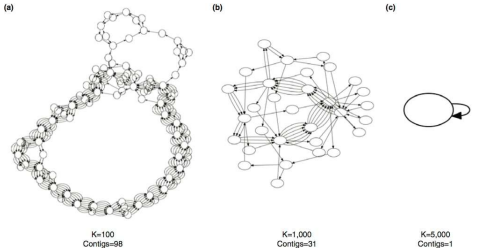


Short read assemblers
- Repeats depends on word length
- Read coherency, placements lost
- Robust to high coverage

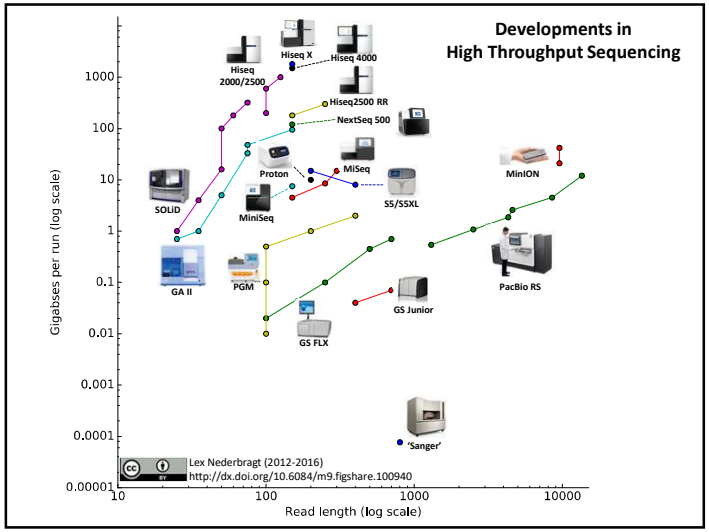
Mike Schatz

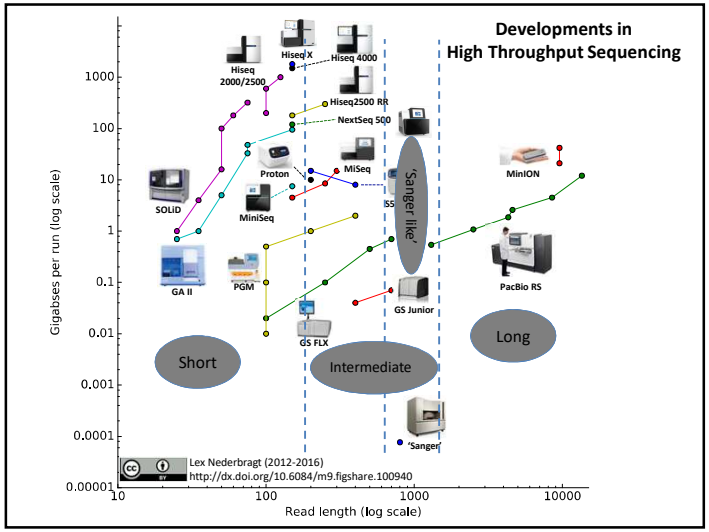
Read length matters

5.2 Mb circular genome, infinite error-free reads



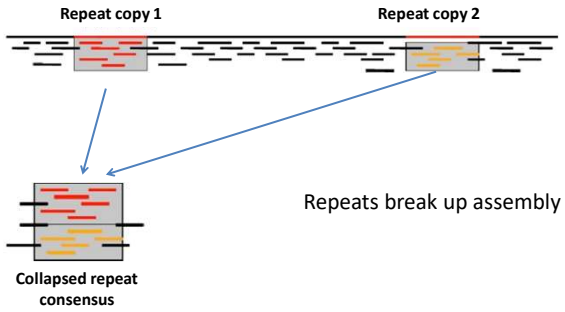
Roberts et al (2013) doi:10.1186/gb-2013-14-6-405





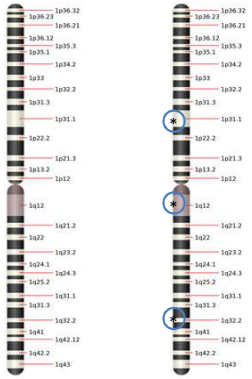
Why is genome assembly such a difficult problem?

1) Repeats



Repeats break up assembly

2) Diploidy



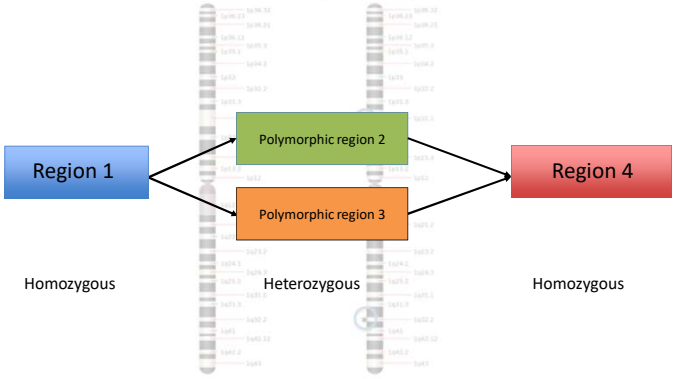
Differences between sister chromosomes

↓

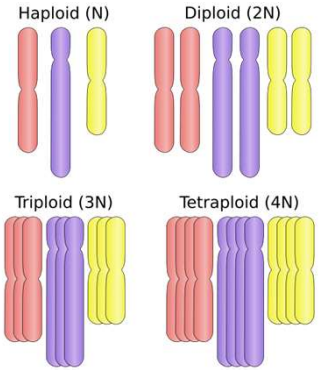
'heterozygosity'

http://commons.wikimedia.org/wiki/File:Chromosome_1.svg

2) Diploidy

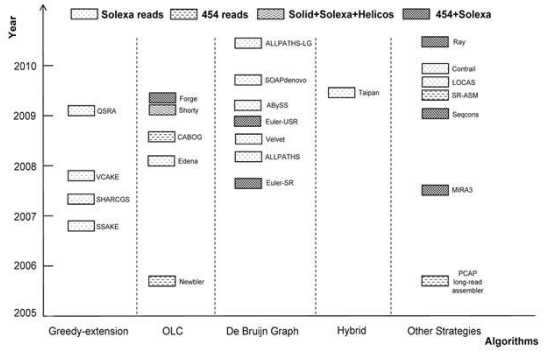


3) Polyploidy



http://en.wikipedia.org/wiki/Polyploidy

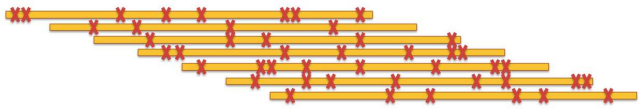
4) Many programs to choose from



Zhang et al. PLoSOne 2011

Assembly with noisy single molecule sequencing data

Usage of long reads

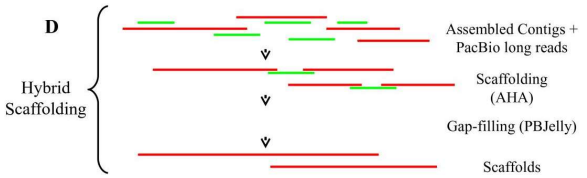


- Problem: higher error rates
- Overlaps more difficult/expensive to find
- OLC more commonly used than for 2nd generation data

Long read assembly strategies

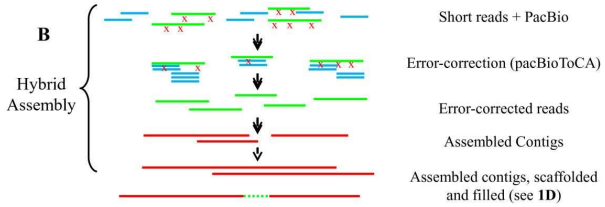
- Scaffolding and gap closing (hybrid)
  - Make short read assembly, scaffold/gap close with long reads (SPAdes)
- Mapping and error correcting (hybrid)
  - Map short reads to long reads, error correct, assemble (MaSuRCA)
- Hierarchical approach (self-correcting)
  - Map shorter long reads to longer long reads, error correct, assemble (HGAP, canu)
- Direct assembly (miniasm)

Scaffolding and gap closing (hybrid)



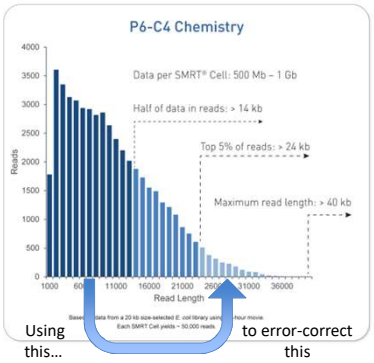
Powers *et.al.*, BMC genomics 2013

Mapping and error correcting (hybrid)



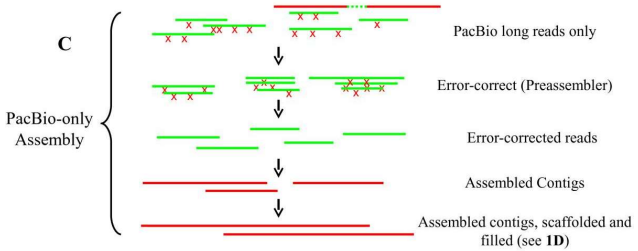
Powers *et.al.*, BMC genomics 2013

Hierarchical approach (self-correcting)



https://genome.duke.edu/cores-and-services/sequencing-and-genomic-technologies/pacbio

Short read error correction



Powers *et.al.*, BMC genomics 2013

Questions?