Principles and problems of de novo genome assembly

Karin Lagesen
Norwegian Veterinary Institute

Material adapted from slides provided by Lex Nederbragt

1

What is this thing called 'genome assembly'?

2

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

3

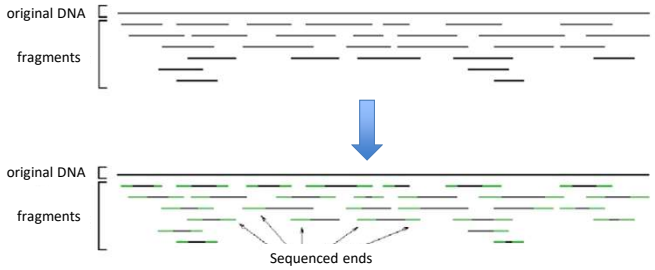
Hierarchical structure

4

Sequence data

reads
contigs
scaffolds

Reads



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

5

Reads!

reads
contigs
scaffolds

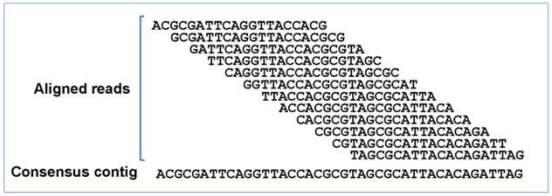
http://www.sciencephoto.com/media/210915/enlarge

6

Contigs

reads
contigs
scaffolds

Building contigs

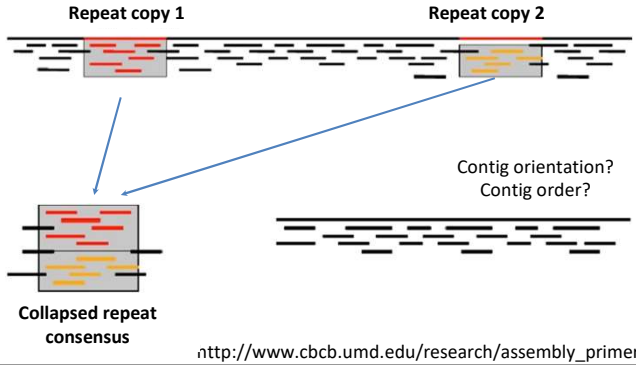


7

Contigs

reads
contigs
scaffolds

Building contigs

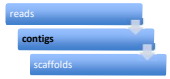


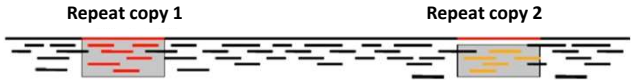
http://www.cbcb.umd.edu/research/assembly_primer.shtml

8

Contigs

Repeats: major problem



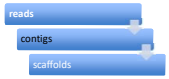


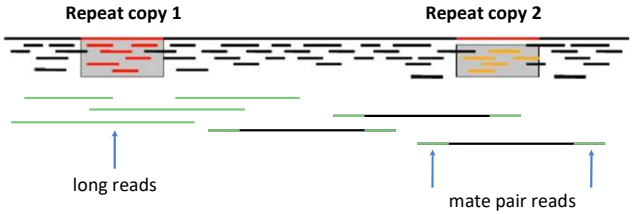
http://www.cbcb.umd.edu/research/assembly_primer.shtml

9

Long reads

Other read type

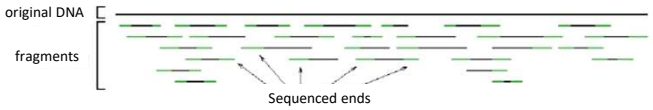




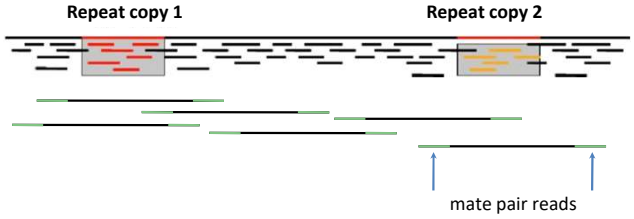
10

Mate pairs

Paired end reads → 100-500 bp insert



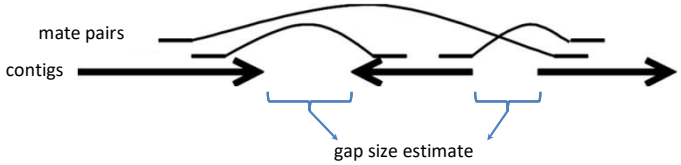
Mate pairs → 2-20 kb insert



11

Scaffolds

Ordered, oriented contigs



12

Hierarchical structure

13

Assembly

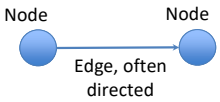
How to do this?

Algorithms

14

Algorithms

All are graph-based

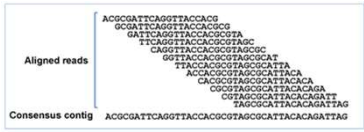


Graph-theory!

15

Algorithms

All are graph-based



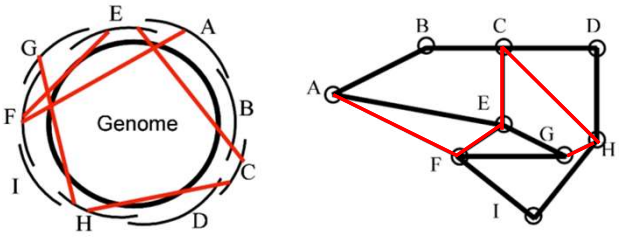
Graph-theory!

16

Algorithms

Hamiltonian path

– a path that contains all the nodes



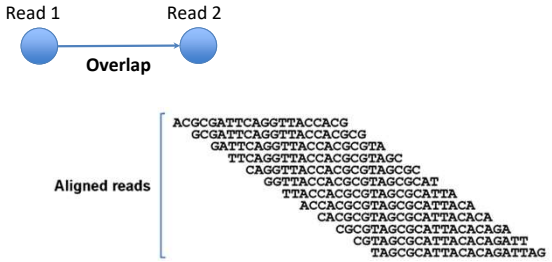
https://www.cbcb.umd.edu/research/assembly_primer

17

Algorithms

Overlap calculation (alignment)

– computationally intensive

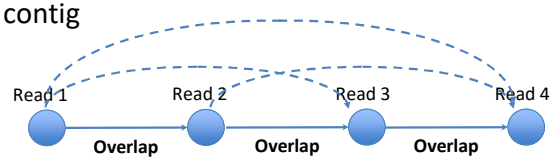


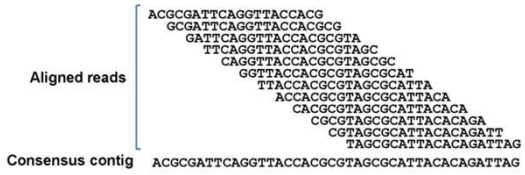
18

Algorithms

Path through the graph

→ contig





19

Algorithms

Many flavors

Abandoned

- Greedy extension

Two most used

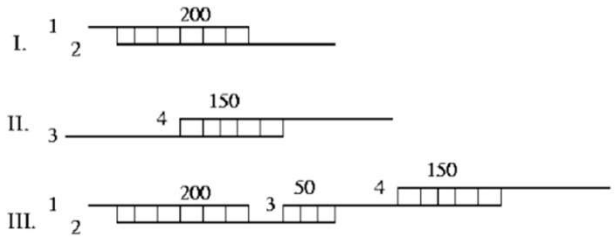
- Overlap Layout Consensus
- de Bruijn graph

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

20

Greedy extension

Oldest

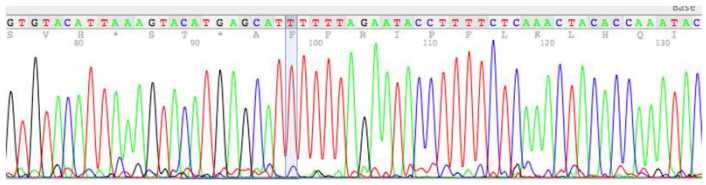


https://www.cbcb.umd.edu/research/assembly_primer

21

Overlap-Layout-Consensus

Developed for Sanger-type reads (longer reads)



22

Overlap-Layout-Consensus

Steps

- Overlap computation
- Layout: graph simplification
- Consensus: sequence

23

Overlap-Layout-Consensus

Overlap phase: find "similar enough" reads

Comparing all against all: expensive

Trick for finding "similar enough" reads:

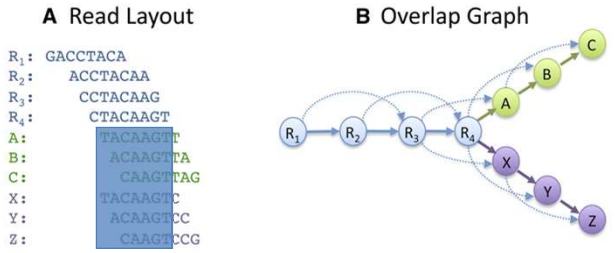
- Split reads into k-mers

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

24

Overlap-Layout-Consensus



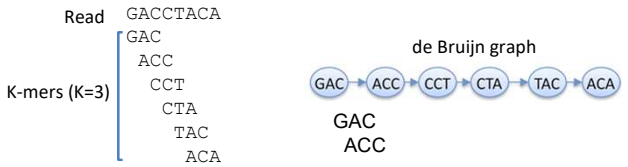
Schatz M C et al. Genome Res. 2010;20:1165-1173

25

de Bruijn graphs

Developed outside of DNA-related work

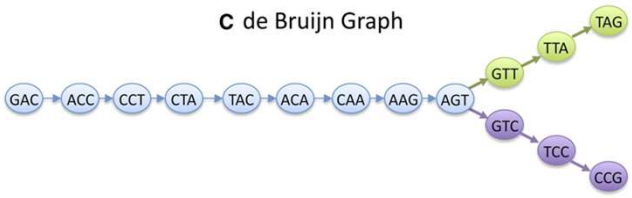
– Best solution for short(er) reads



K-1 bases overlap

26

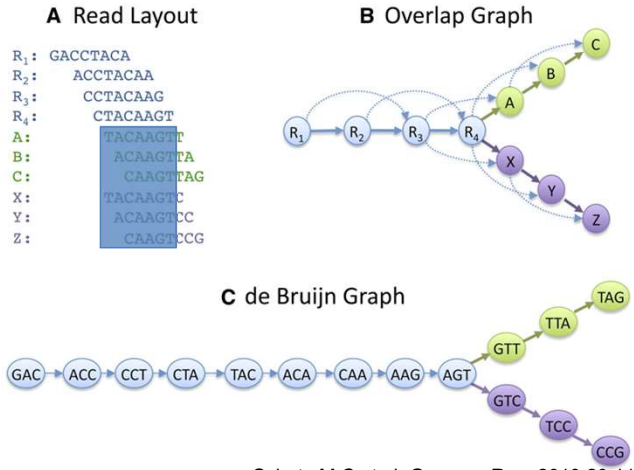
Graphs



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

27

Graphs



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

28

Graphs

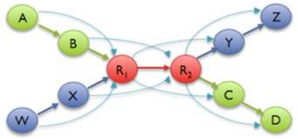
Simplify the graph

Add scaffolding information

29

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

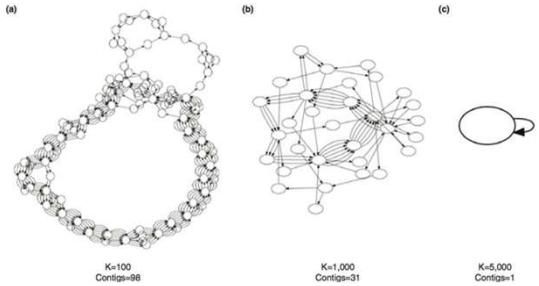
- Repeats depends on word length
- Read coherency, placements lost
- Robust to high coverage

Mike Schatz

30

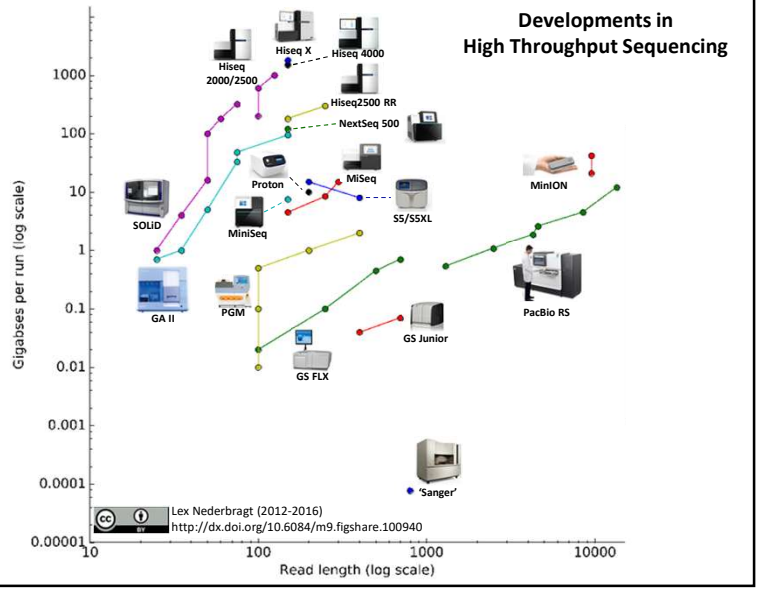
Read length matters

5.2 Mb circular genome, infinite error-free reads

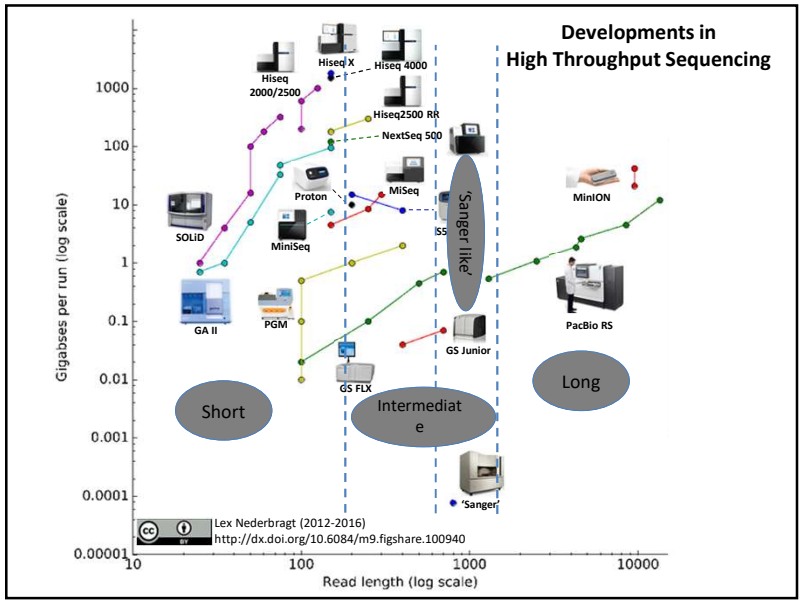


Roberts et al (2013) doi:10.1186/gb-2013-14-6-405

31



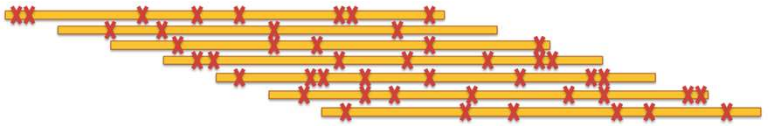
32



33

Quality matters

Too many errors → hard to find overlaps



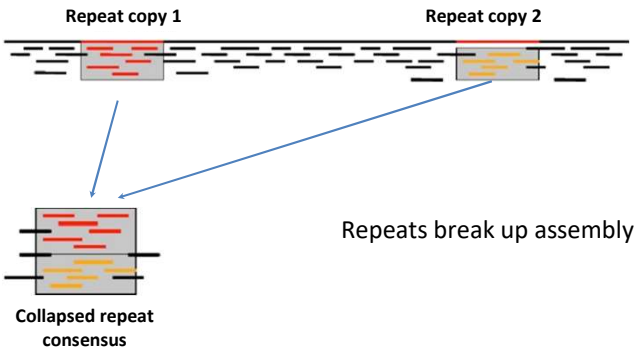
http://schatzlab.cshl.edu/presentations/2012-01-17.PAG.SMRTassembly.pdf

34

Why is genome assembly such a difficult problem?

35

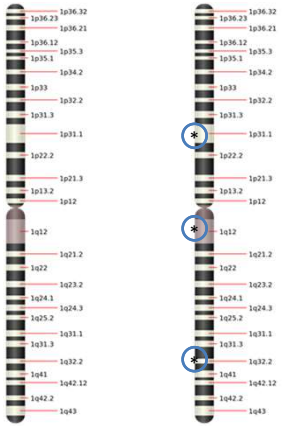
1) Repeats



Repeats break up assembly

36

2) Diploidy



Differences between sister chromosomes

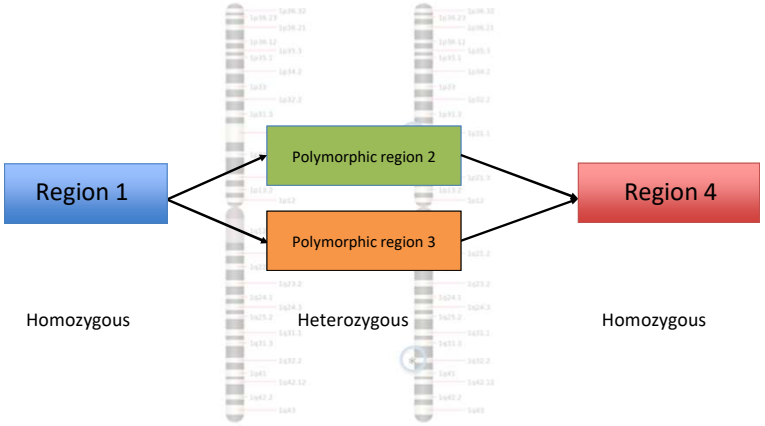
↓

'heterozygosity'

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

37

2) Diploidy



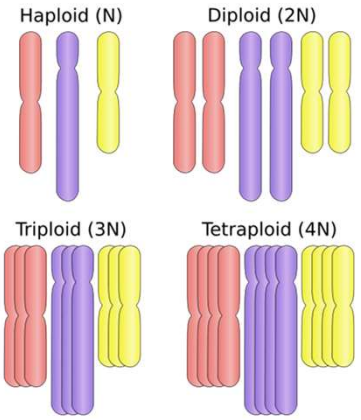
38

2) Diploidy

http://www.astraean.com/borderwars/wp-content/uploads/2012/04/heterozygoats.jpg
and many other sites

39

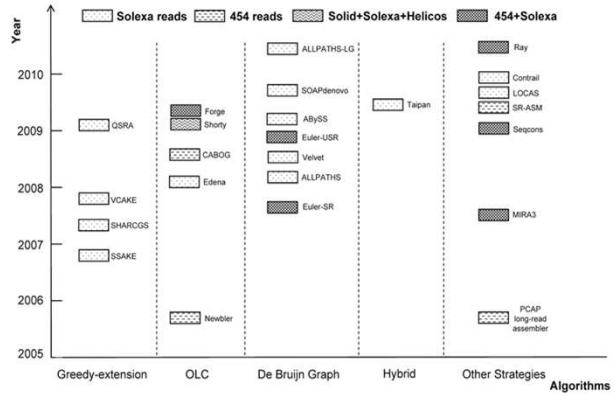
3) Polyploidy



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

40

4) Many programs to choose from



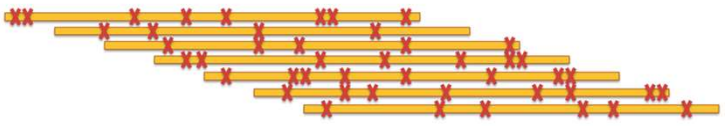
Zhang et al. PLoSOne 2011

41

Assembly with noisy single molecule sequencing data

42

Usage of long reads



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

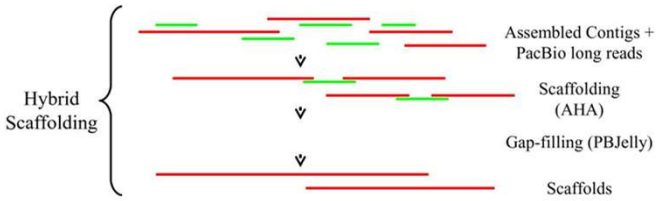
43

Long read assembly strategies

- Alt 0: Scaffolding short read asms
- Alt 1: Correct reads, then assemble
- Alt 2: Assemble reads, then correct

44

Scaffolding and gap closing (hybrid)



Powers *et.al.*, BMC genomics 2013

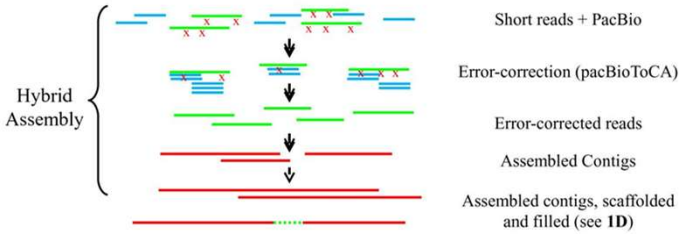
45

Correct, assemble

- Do pairwise comparison, find shorter reads that support the longer
- Align supporting reads, correct longer reads
- Overlap-Layout-Consensus on corrected reads
- Polish assembly

46

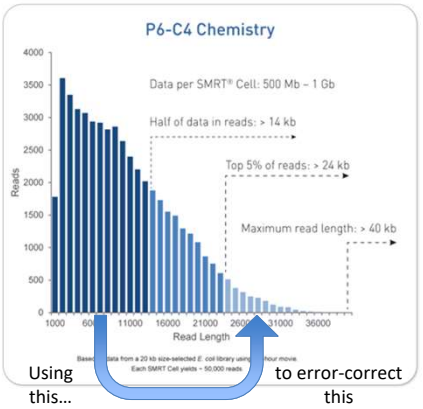
Mapping and error correcting (hybrid)



Powers *et.al.*, BMC genomics 2013

47

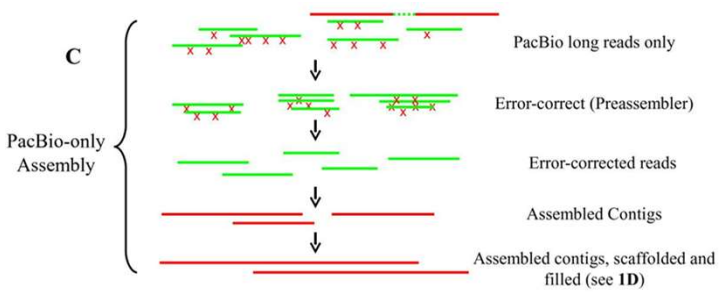
Hierarchical approach (self-correcting)



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

48

Short read error correction



Powers *et.al.*, BMC genomics 2013

49

Assemble, correct

- Compare reads, find overlaps
- Assemble reads, knowing things will be wrong
- Align reads to assembly
- Correct assembly

50

Questions?

51