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

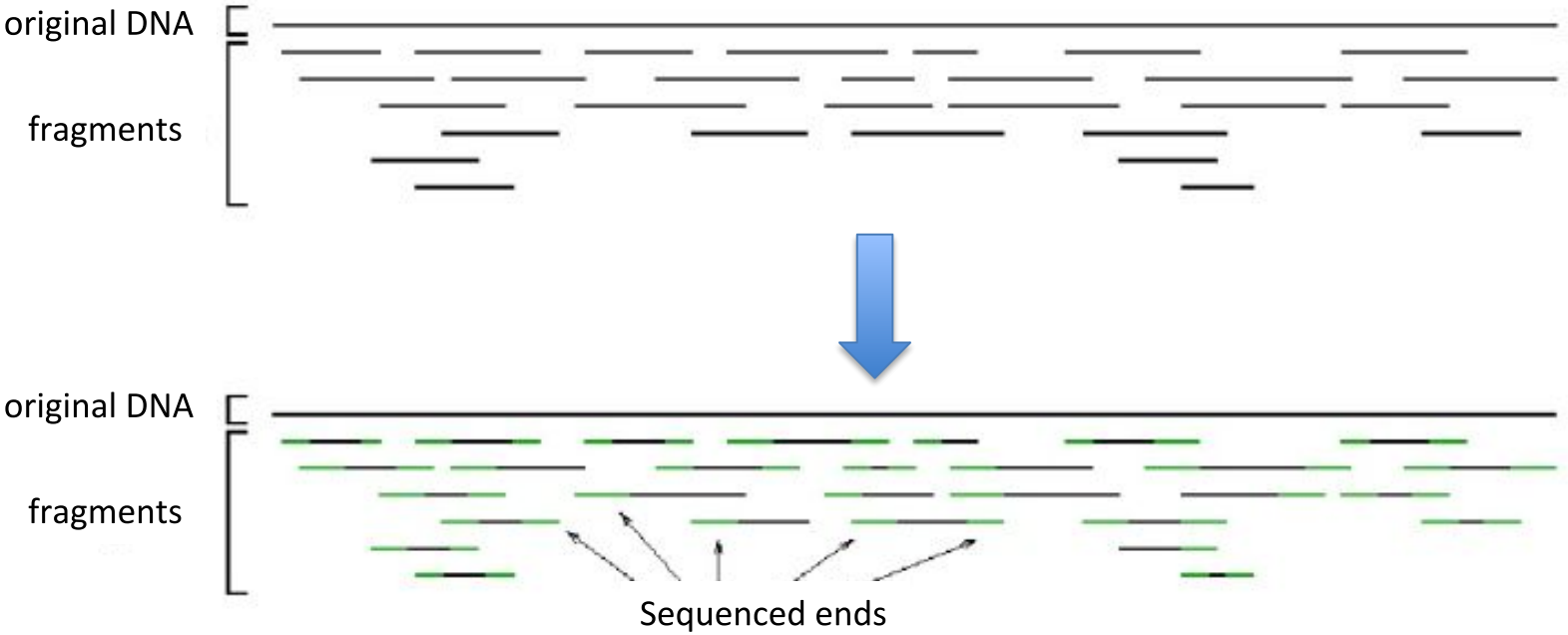
to a putative reconstruction of the target

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

Hierarchical structure

Sequence data

Reads



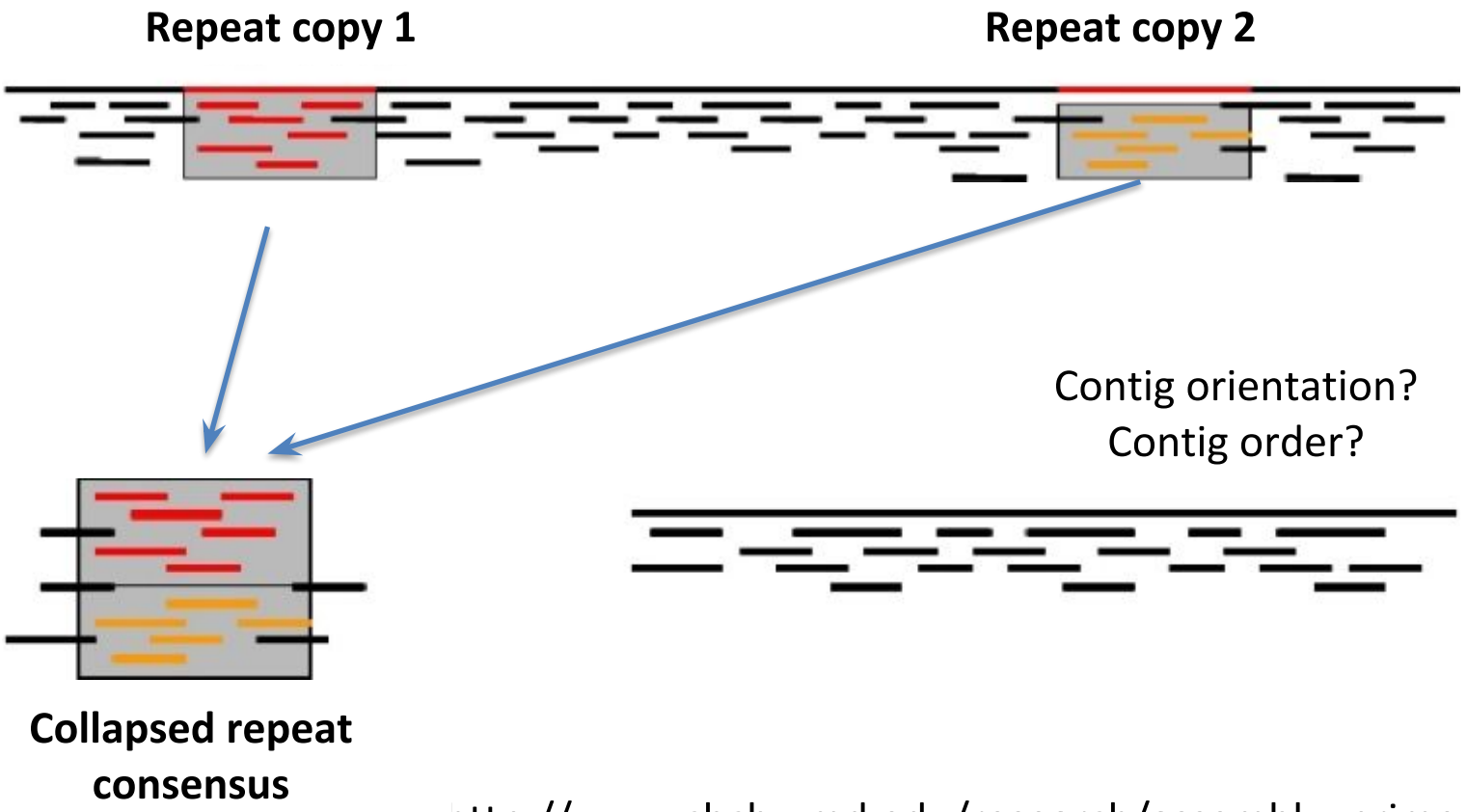
http://www.cbcb.umd.edu/research/assembly_primer.shtml

Contigs

Building contigs

Contigs

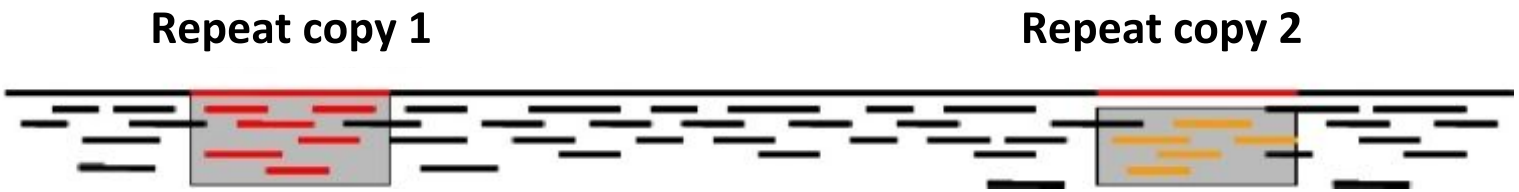
Building contigs



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

Contigs

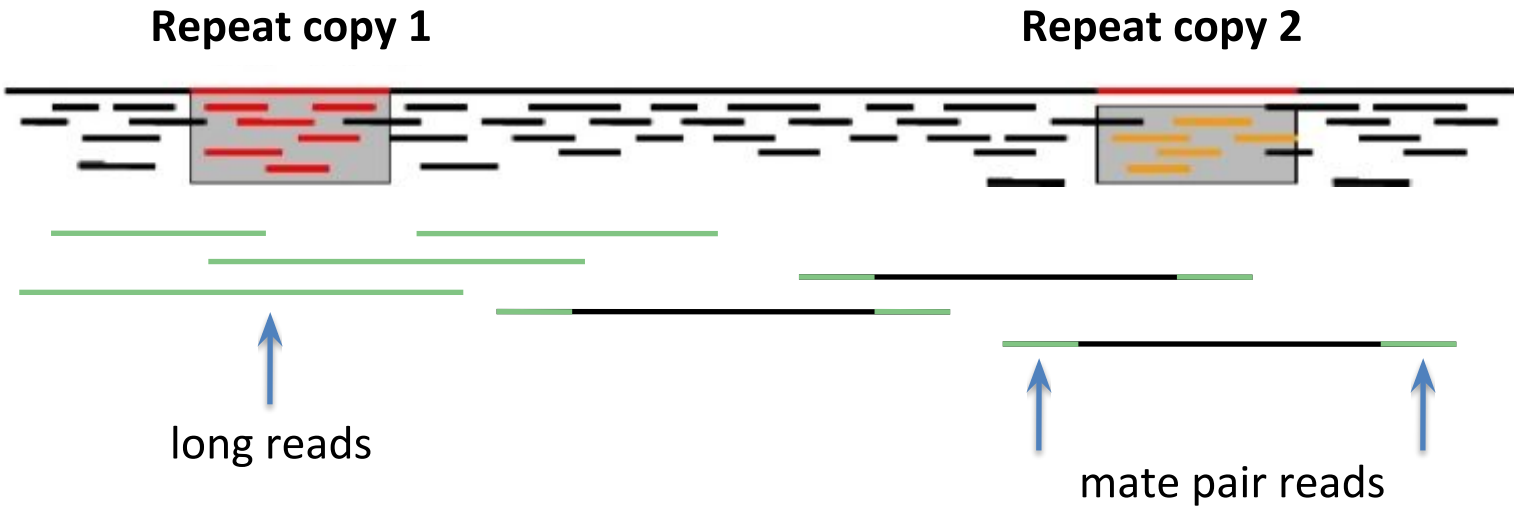
Repeats: major problem



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

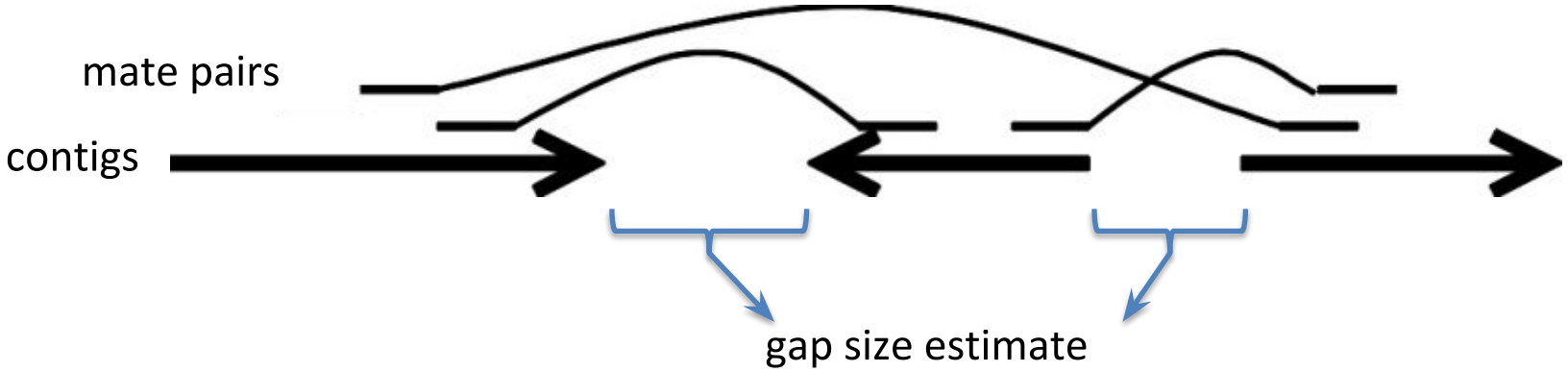
Mate pairs/long reads

Other read type



Scaffolds

Ordered, oriented contigs



Hierarchical structure

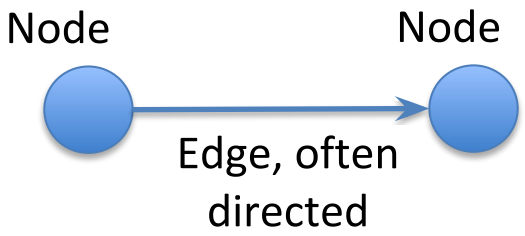
Assembly

How to do this?

Algorithms

Algorithms

All are graph-based



Graph-theory!

Algorithms

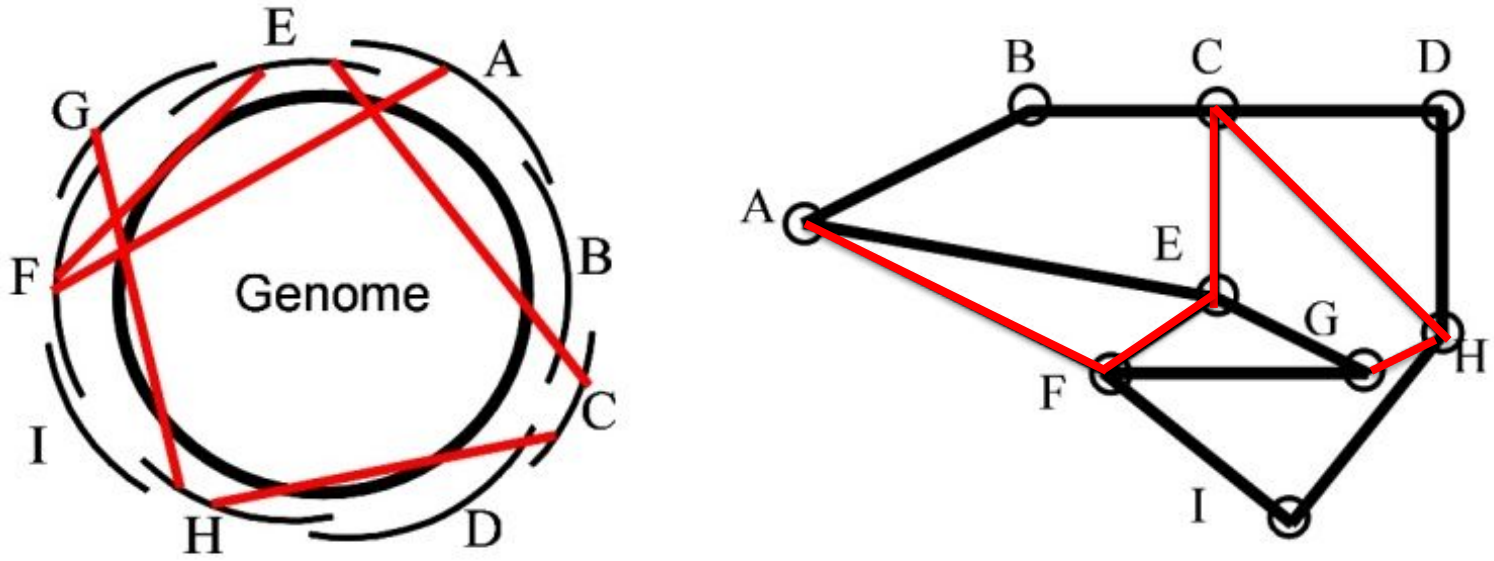
All are graph-based

Graph-theory!

Algorithms

Hamiltonian path

– a path that contains all the nodes



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

Algorithms

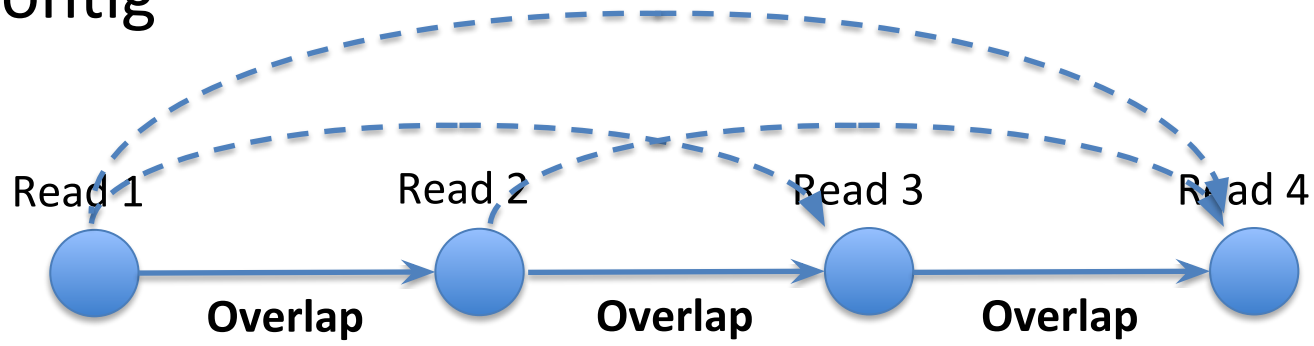
Overlap calculation (alignment)

– computationally intensive

Algorithms

Path through the graph

- contig



Algorithms

Many flavors

Abandoned

- Greedy extension

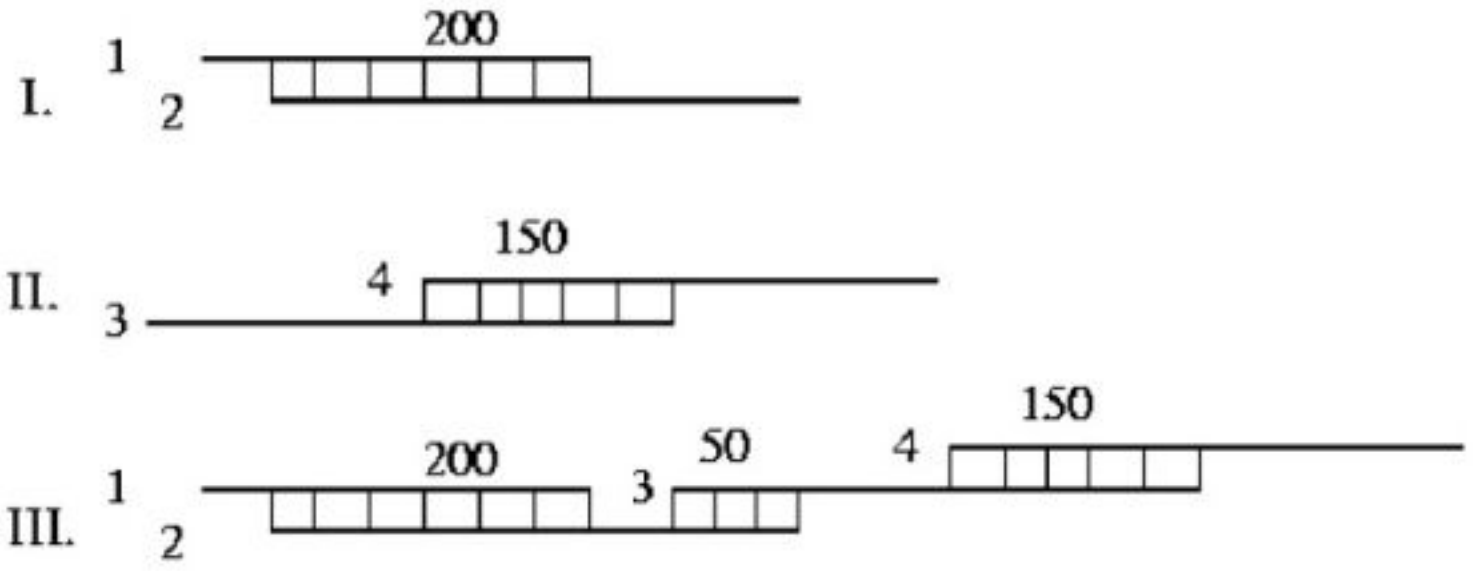
Two most used

- Overlap Layout Consensus
- de Bruijn graph

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

Greedy extension

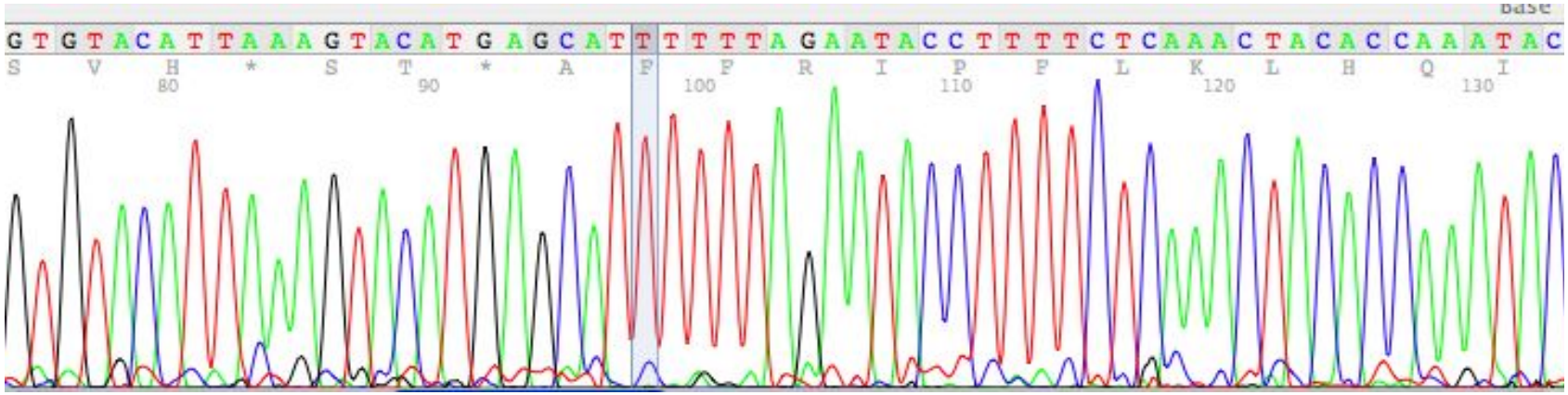
Oldest



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

Overlap-Layout-Consensus

Developed for Sanger-type reads (longer reads)



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

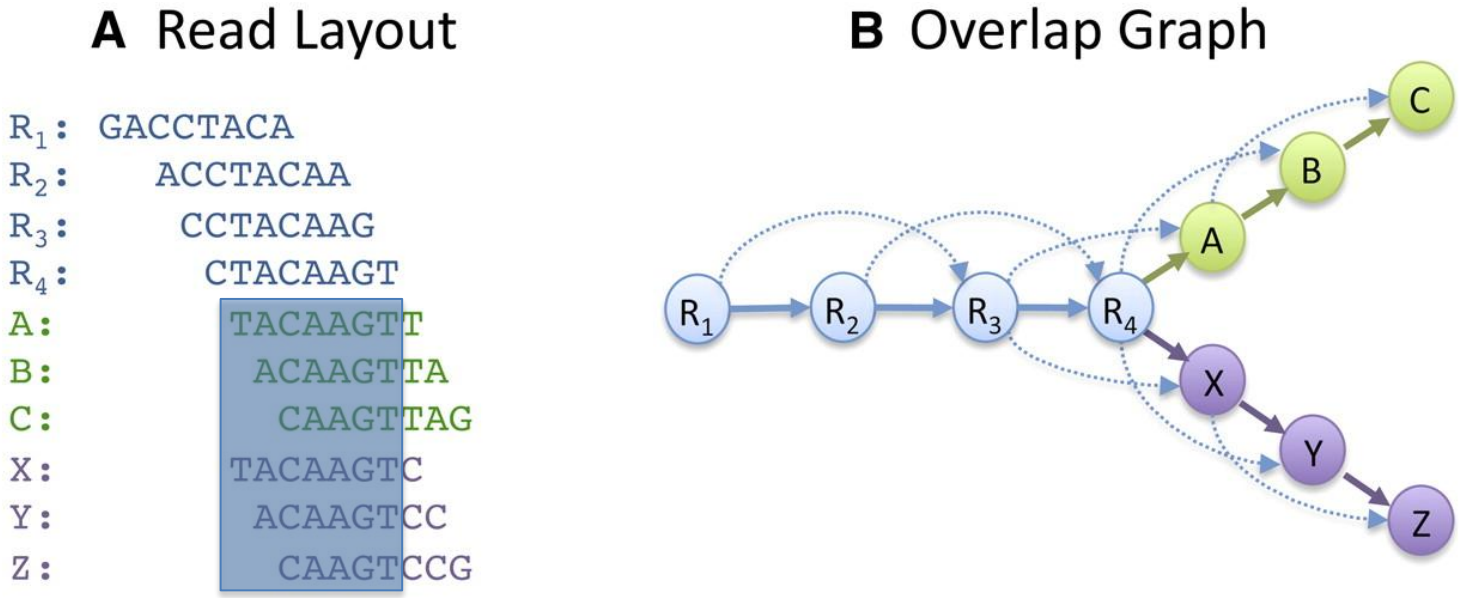
Trick for finding “similar enough” reads:

- Split reads into k-mers

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

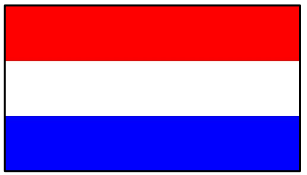
- Make list over which read has which k-mers
- If two reads share k-mers, test for similarity

Overlap-Layout-Consensus



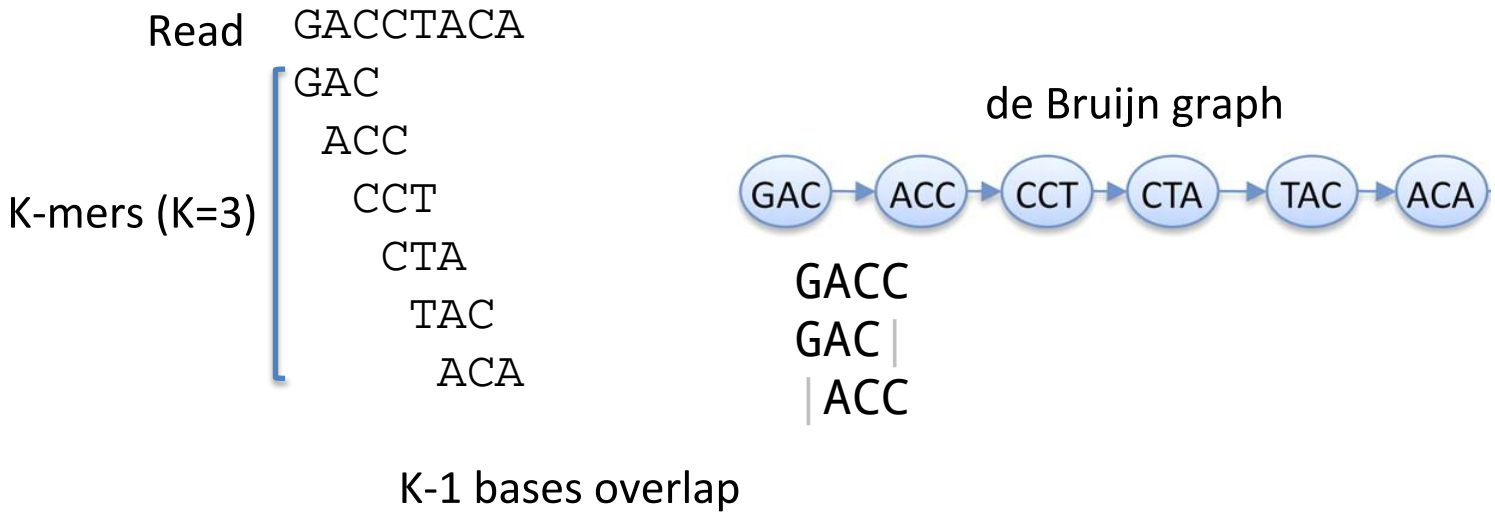
Schatz M C et al. Genome Res. 2010;20:1165-1173

de Bruijn graphs

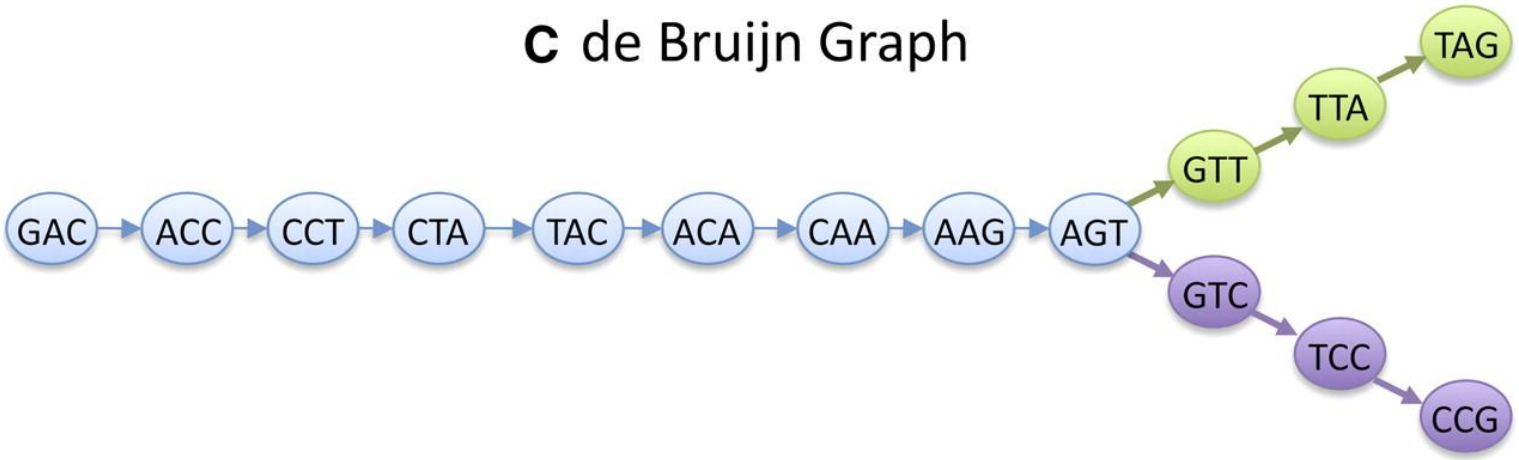


Developed outside of DNA-related work

- Best solution for short(er) reads

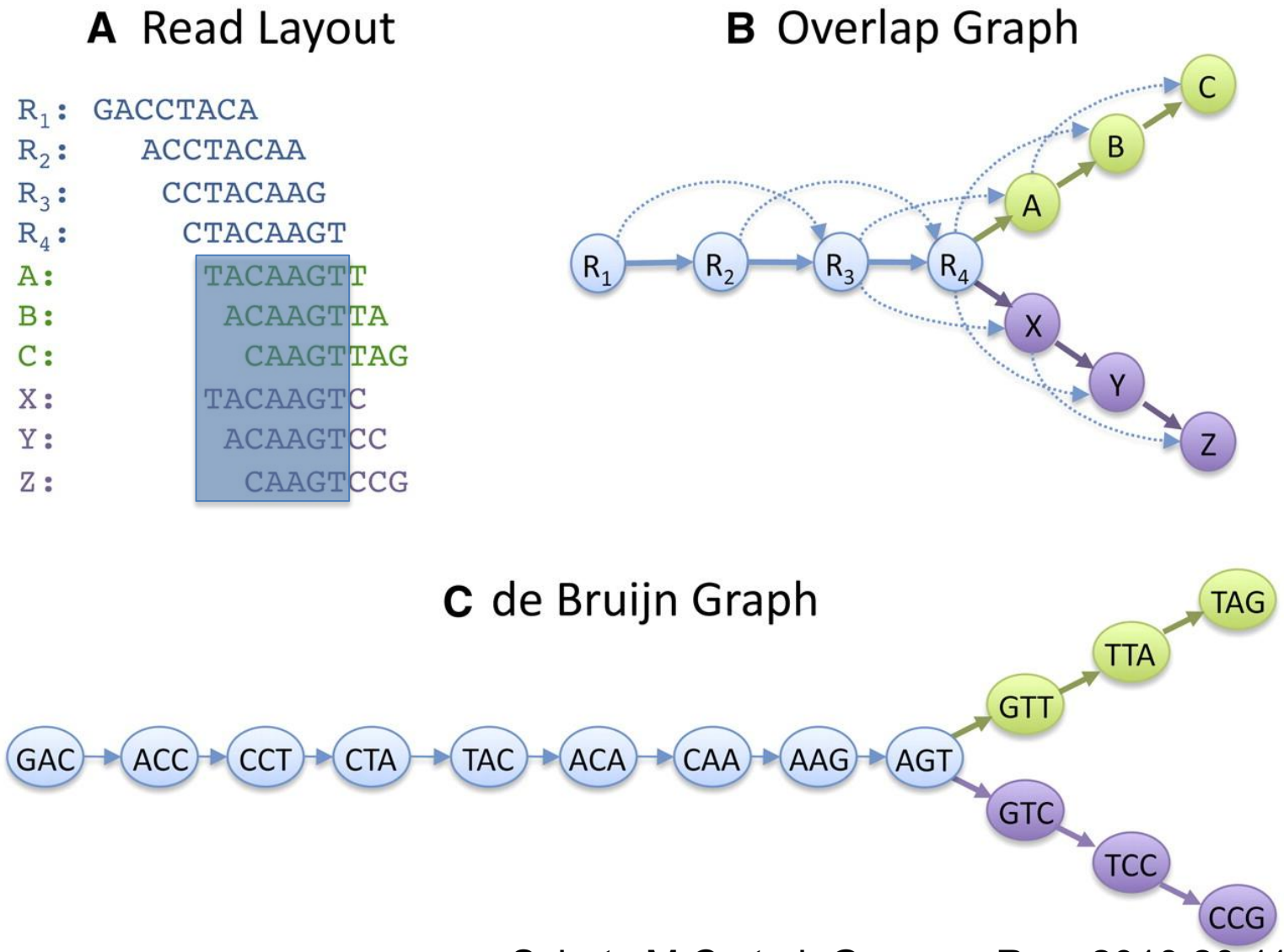


Graphs



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

Graphs



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

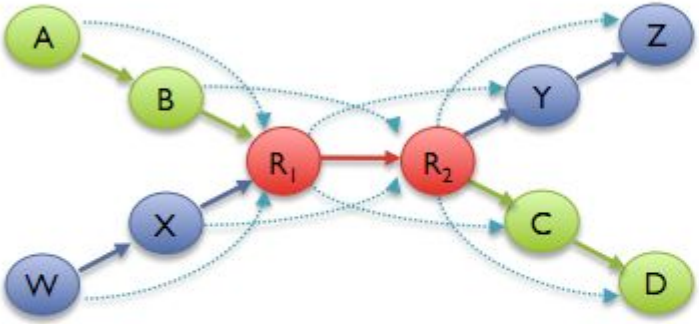
Graphs

Simplify the graph

Add scaffolding information

de Bruijn Graphs

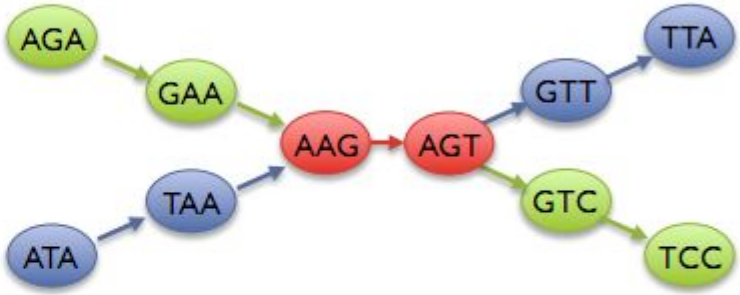
Overlap Graph



Long read assemblers

- Repeats depends on read length
- Read coherency, placements kept
- Tangled by high coherage

de Bruijn Graph



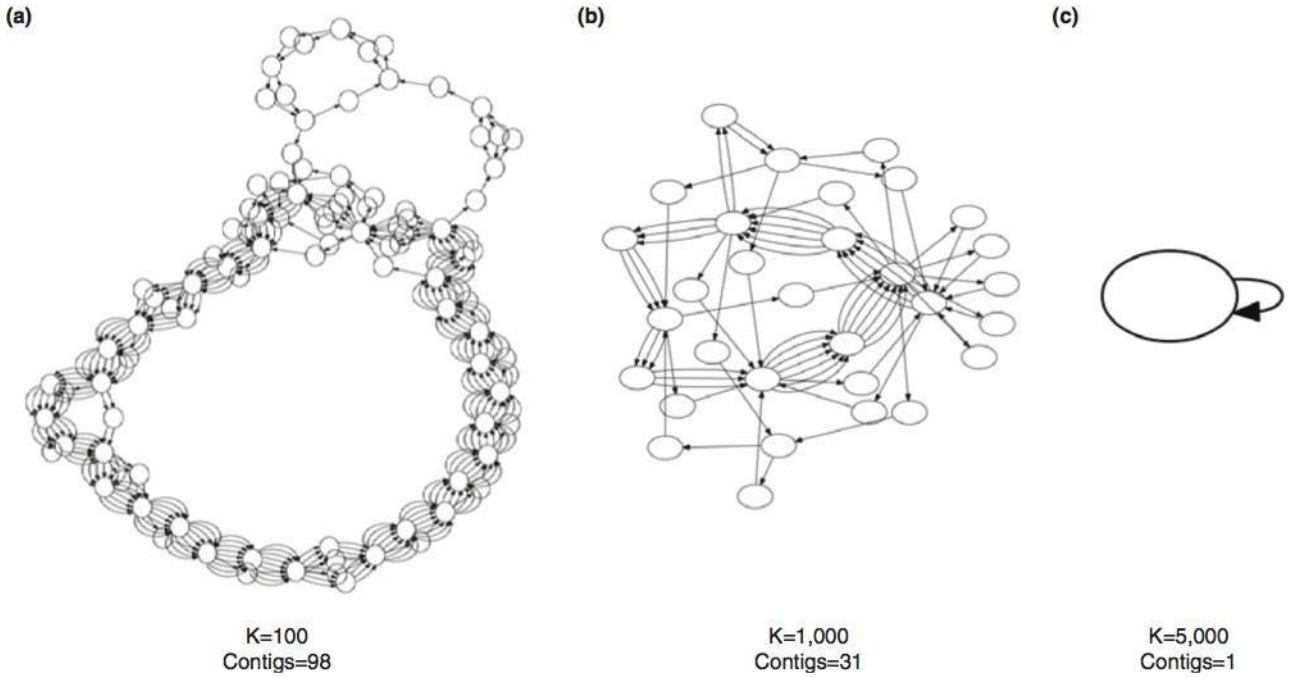
Short read assemblers

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

Mike Schatz

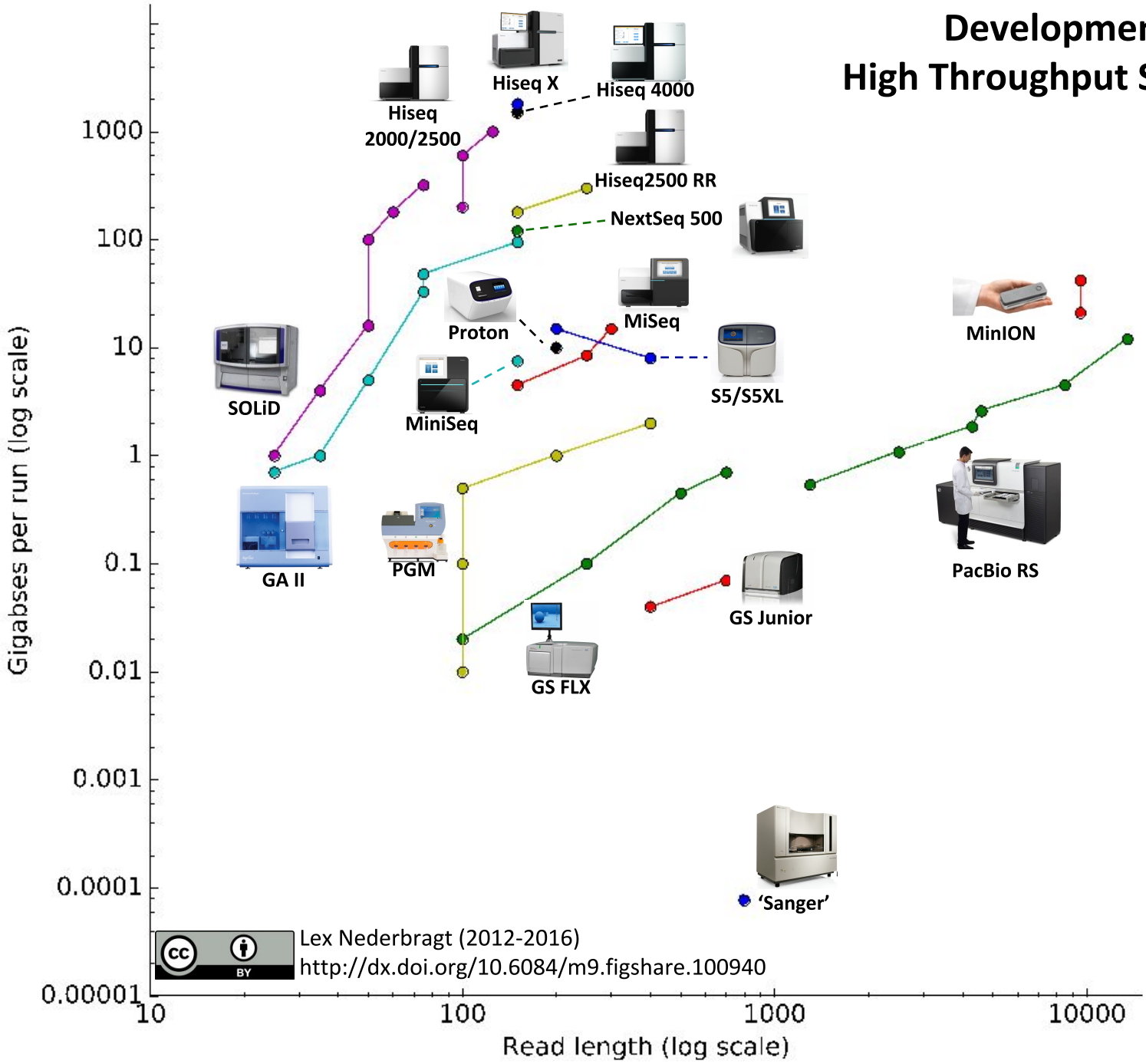
Read length matters

5.2 Mb circular genome, infinite error-free reads

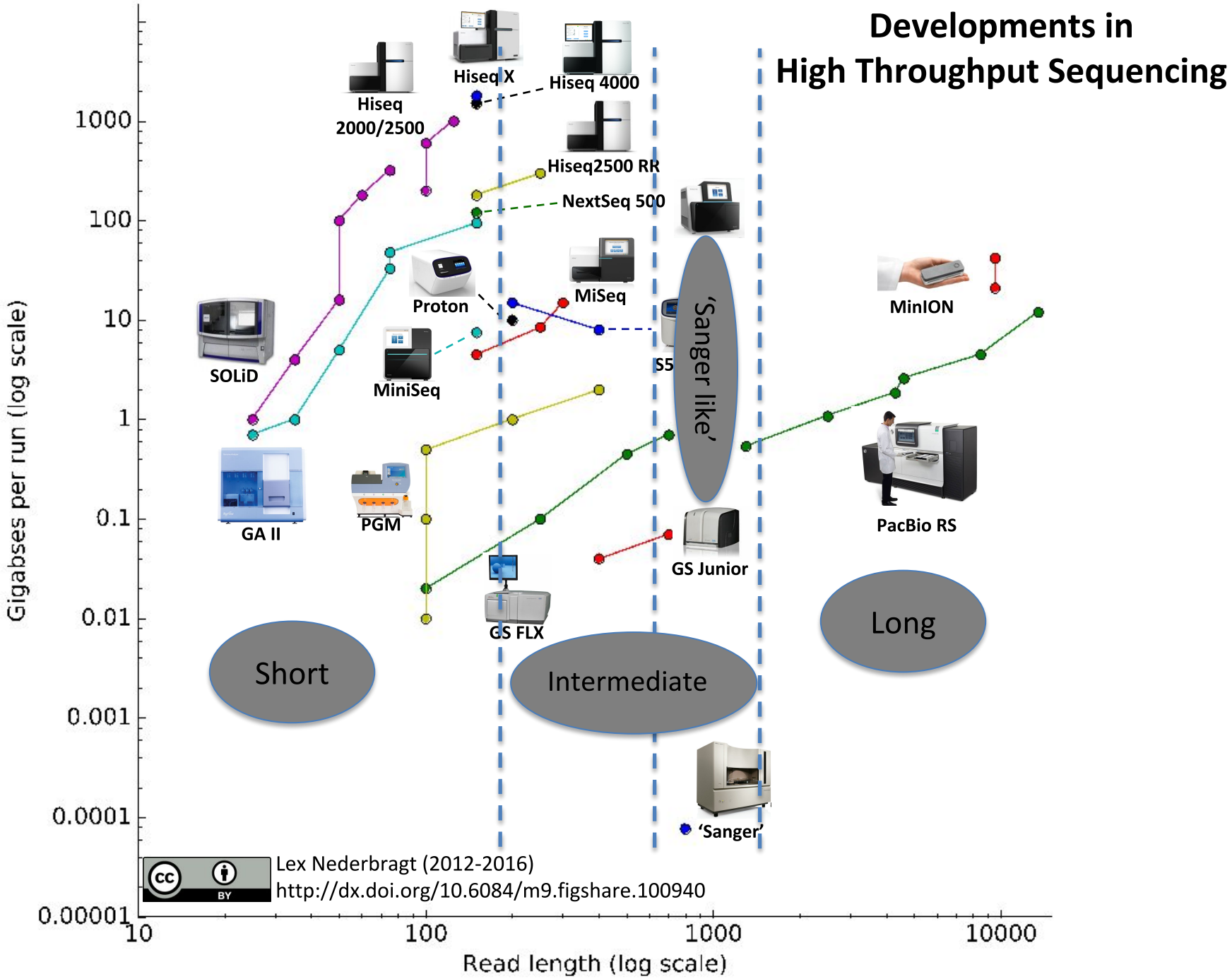


Roberts et al (2013) doi:10.1186/gb-2013-14-6-405

Developments in High Throughput Sequencing

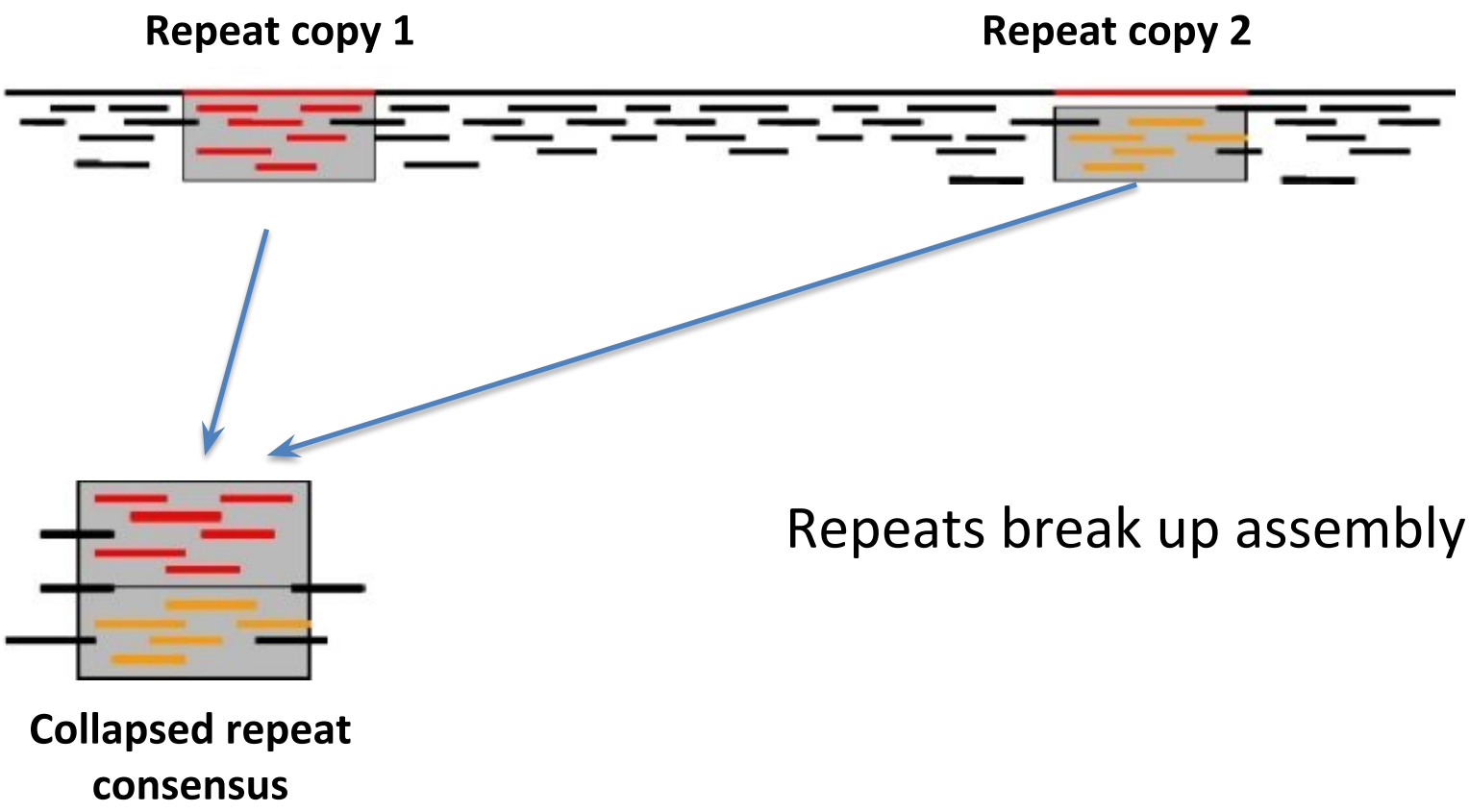


http://dx.doi.org/10.6084/m9.figshare.100940


Developments in
High Throughput Sequencing
Hiseq X
Hiseq 4000
Hiseq
2000/2500
Hiseq2500 RR
NextSeq 500
Proton
MiSeq
MinION
SOLiD
MiniSeq
S5
'Sanger like'
GA II
PGM
PacBio RS
GS Junior
GS FLX
Short
Intermediate
Long
'Sanger'
Lex Nederbragt (2012-2016)
http://dx.doi.org/10.6084/m9.figshare.100940
Gigabses per run (log scale)
Read length (log scale)
1000
100
10
1
0.1
0.01
0.001
0.0001
0.00001
10
100
1000
10000

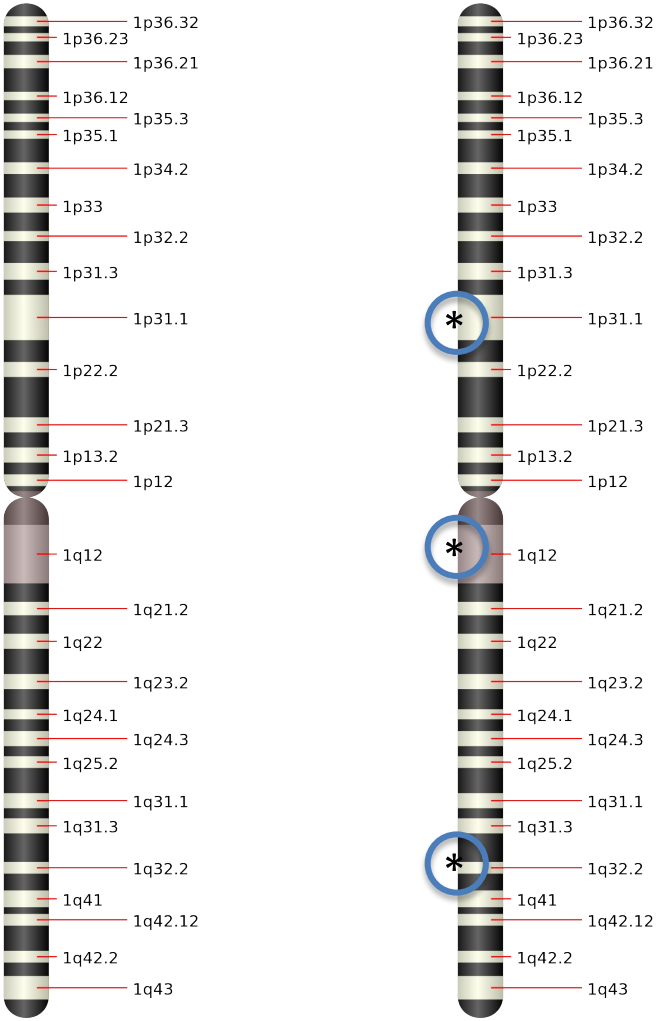
Why is genome assembly such a difficult problem?

1) Repeats



Repeats break up assembly

2) Diploidy



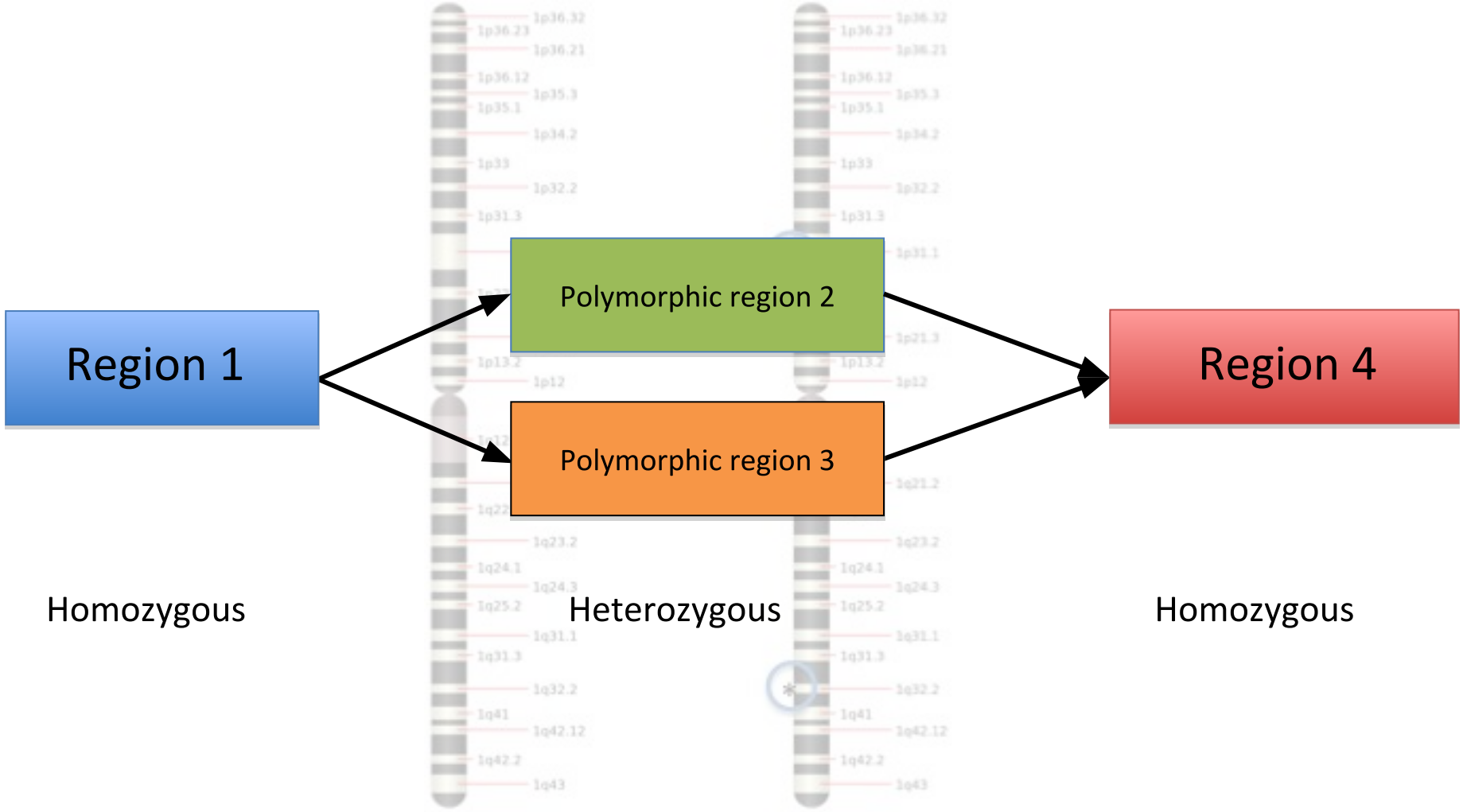
Differences between sister chromosomes

↓

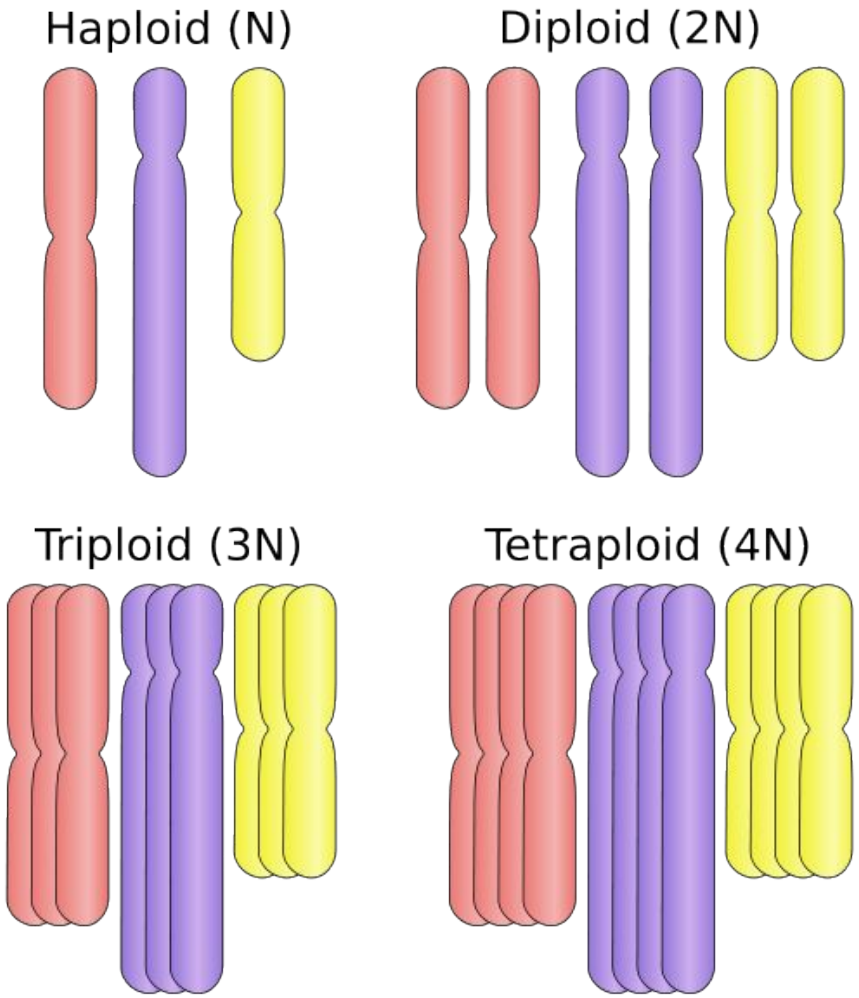
'heterozygosity'

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

2) Diploidy

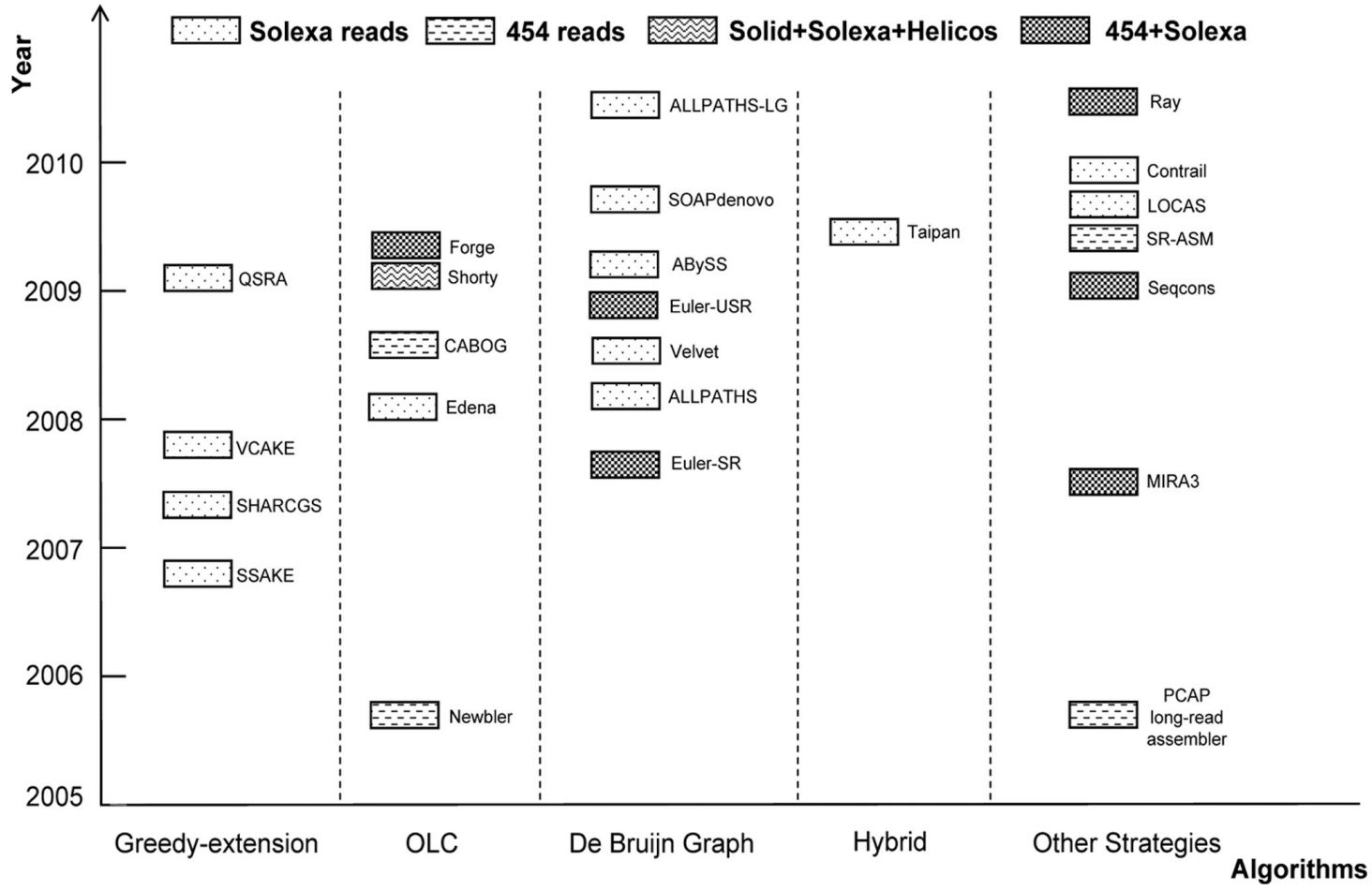

Region 1
Polymorphic region 2
Polymorphic region 3
Region 4
Homozygous
Heterozygous
Homozygous

3) Polyploidy



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

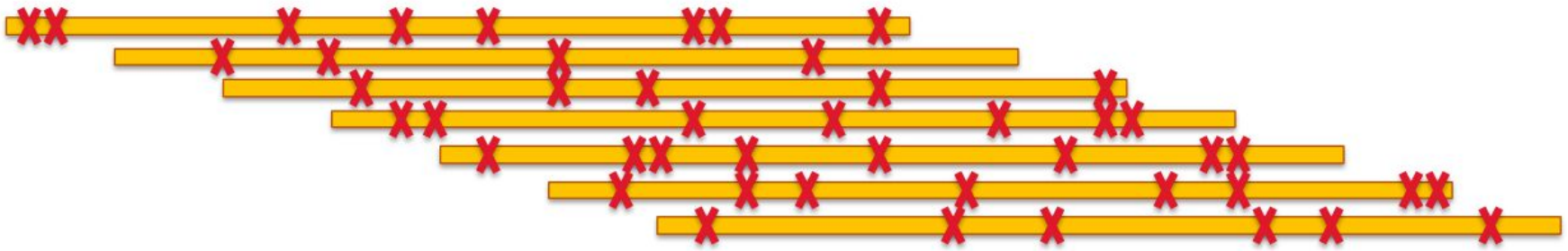
4) Many programs to choose from



Zhang et al. PLoSOne 2011

Assembly with noisy single molecule sequencing data

Usage of long reads

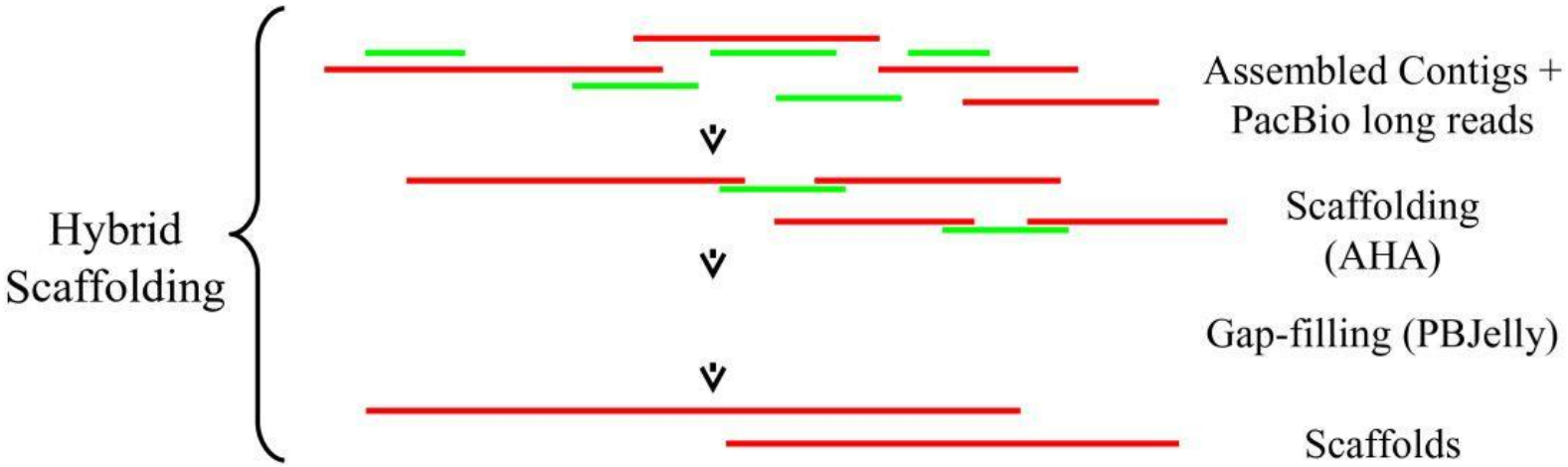


- Problem: higher error rates
- Overlaps more difficult/expensive to find
- OLC more commonly used than for 2nd generation data

Long read assembly strategies

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

Scaffolding and gap closing (hybrid)

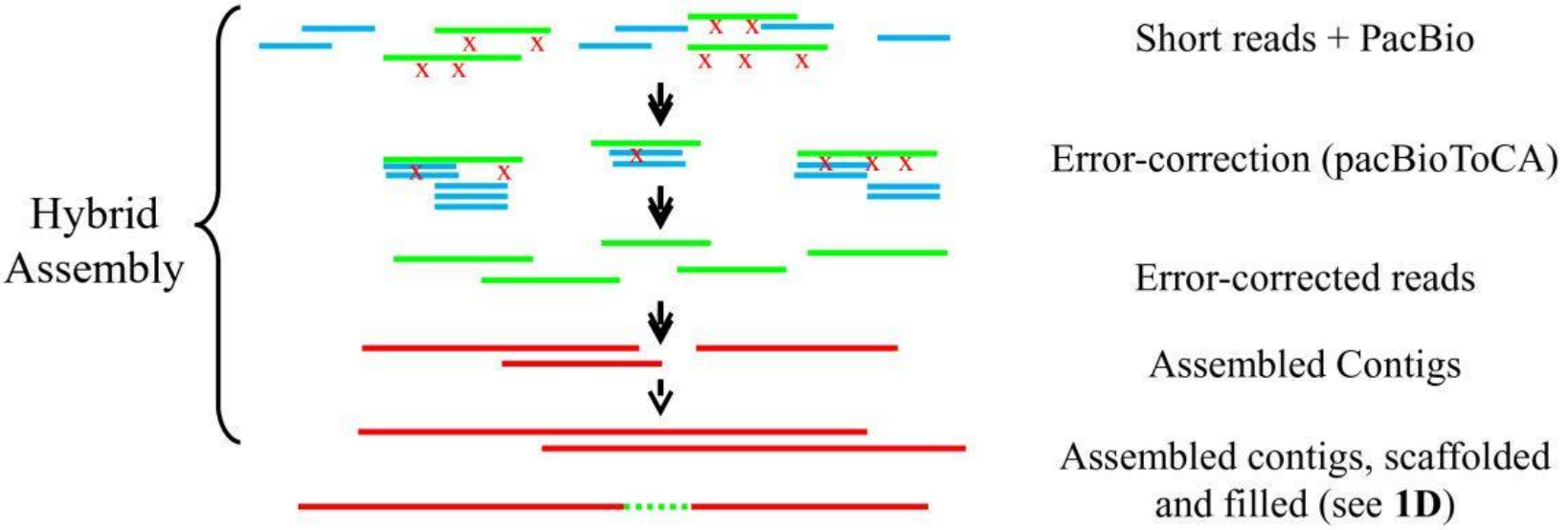


Powers *et.al.*, BMC genomics 2013

Correct, assemble

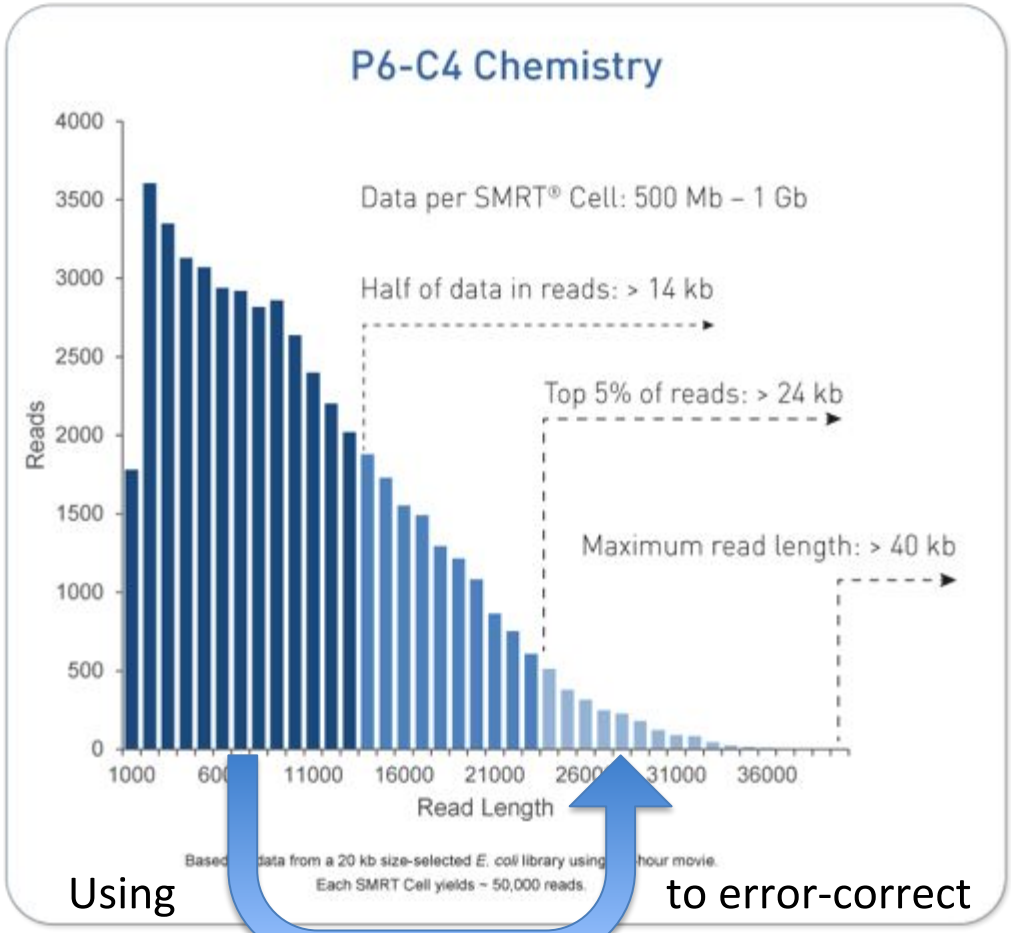
- Do pairwise comparison, find shorter reads that support the longer
- Align supporting reads, correct longer reads
- Overlap-Layout-Consensus on corrected reads
- Polish assembly

Mapping and error correcting (hybrid)



Powers *et.al.*, BMC genomics 2013

Hierarchical approach (self-correcting)

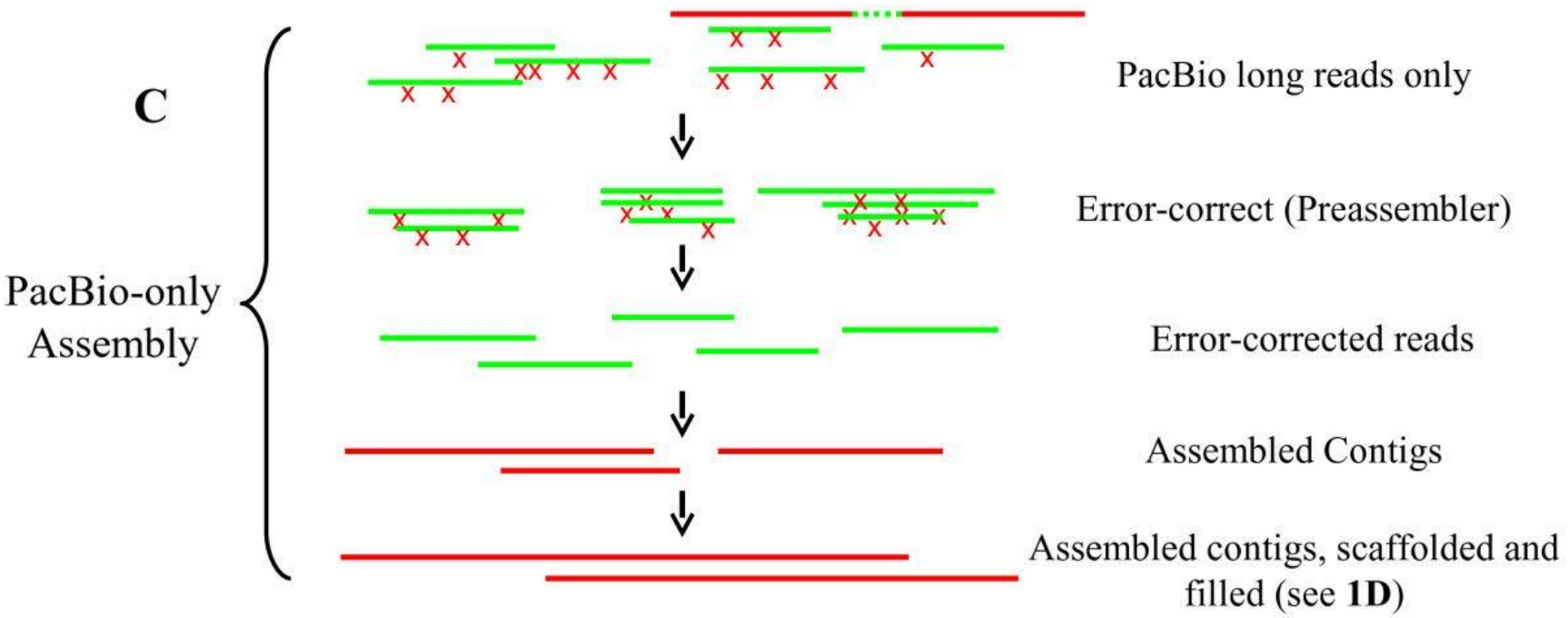


Using this …

to error-correct this

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

Short read error correction



Powers *et.al.*, BMC genomics 2013

Assemble, correct

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

Questions?