NORWEGIAN SEQUENCING CENTRE

Principles and problems of de novo genome assembly

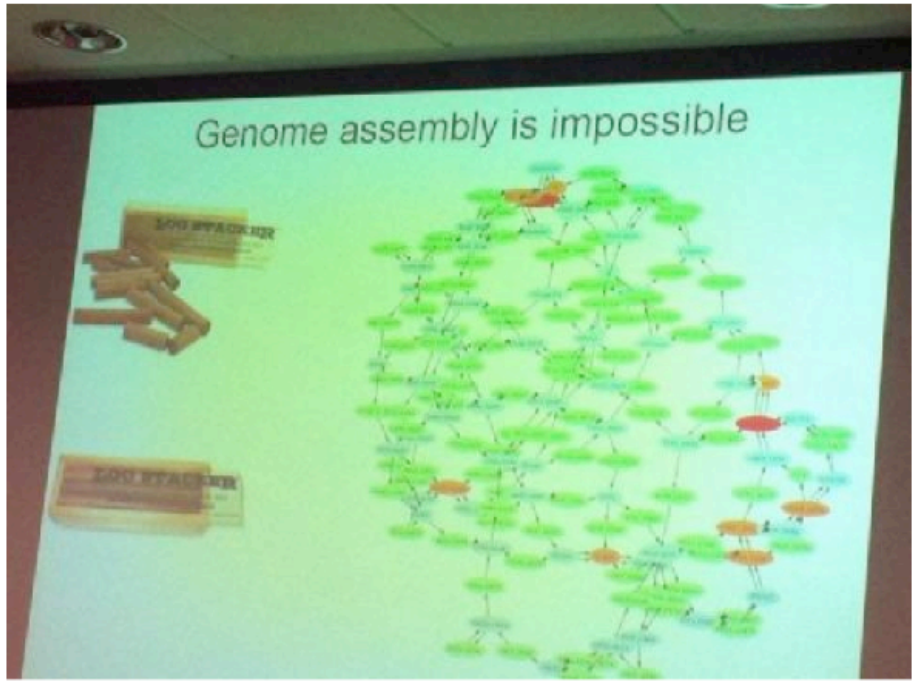
Lex Nederbragt
Norwegian High-Throughput Sequencing Centre (NSC)
and
Centre for Ecological and Evolutionary Synthesis (CEES)

CEES
Centre for Ecological and Evolutionary Synthesis

twitter Search Home Profile Messages Who To Follow lexnederbragt

@gilbertjacka
Jack A Gilbert

Finally someone (Mihai Pop) admits it!!
yfrog.com/nxen3ohj



YFrog Flag this media

5 hours ago via Twitter for iPhone Favorite Undo Retweet Reply

Retweeted by lexnederbragt

What is this thing called ‘genome assembly’?

What is a genome assembly?

A hierarchical data structure

that maps the sequence data

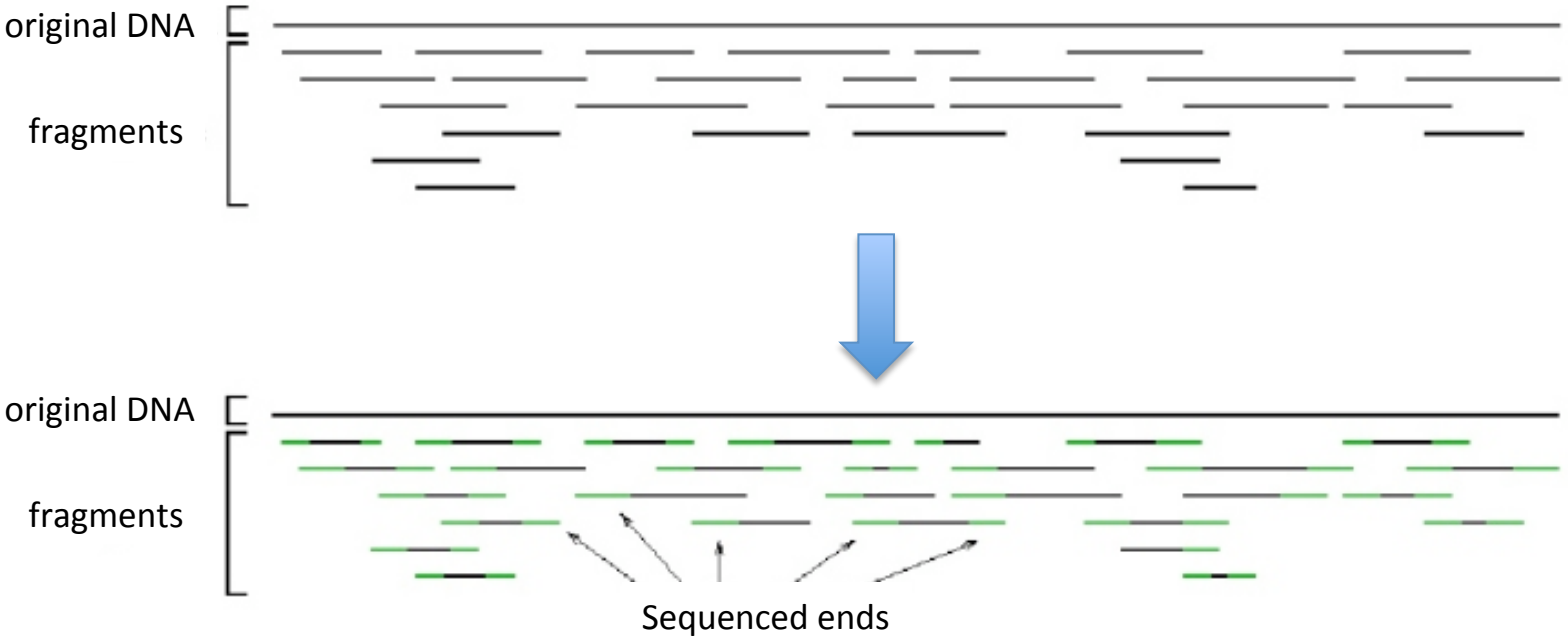
to a putative reconstruction of the target

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

Hierarchical structure

Sequence data

Reads



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

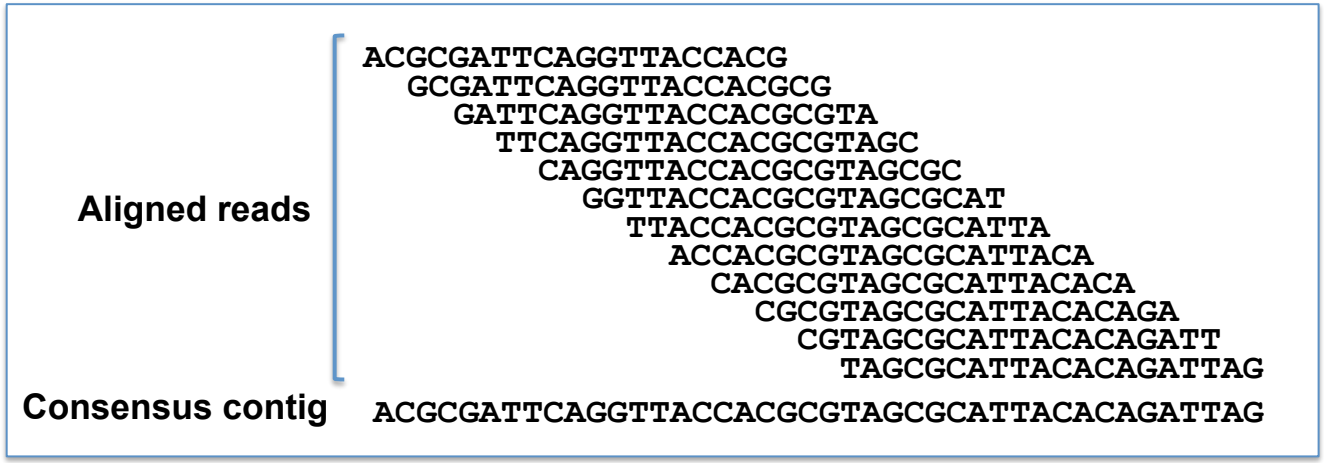
Reads!

- **reads**
- contigs
- scaffolds

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

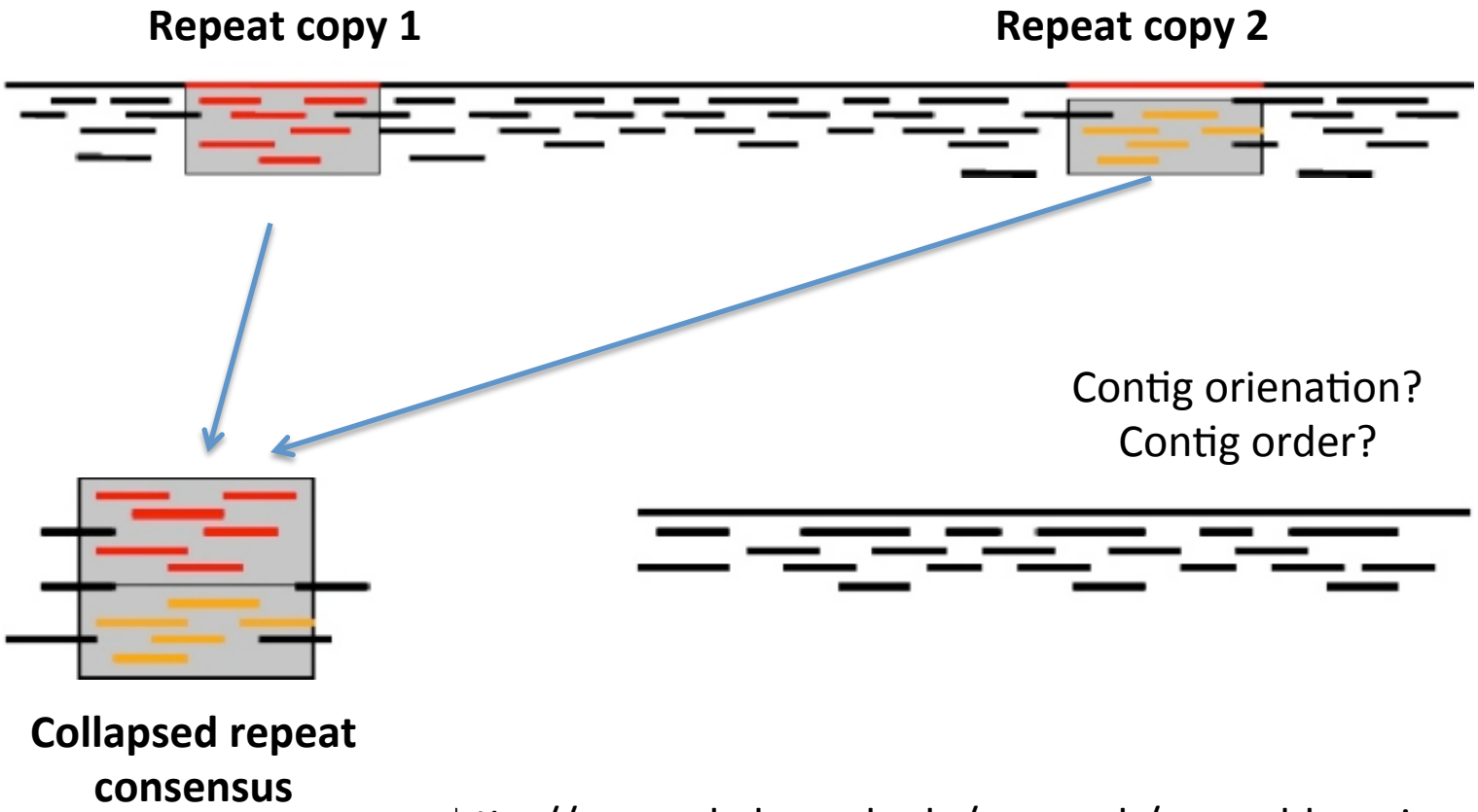
Contigs

Building contigs



Contigs

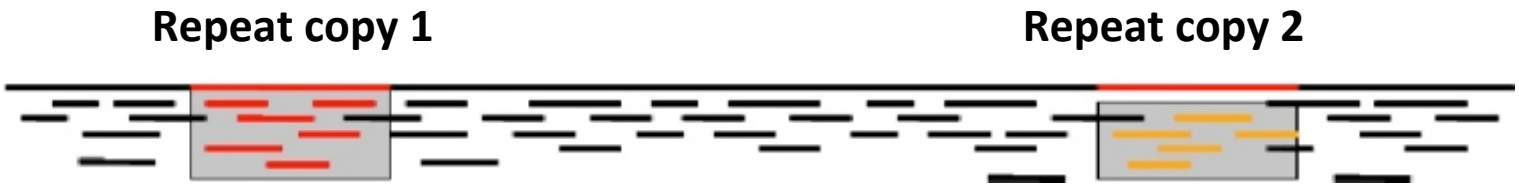
Building contigs



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

Contigs

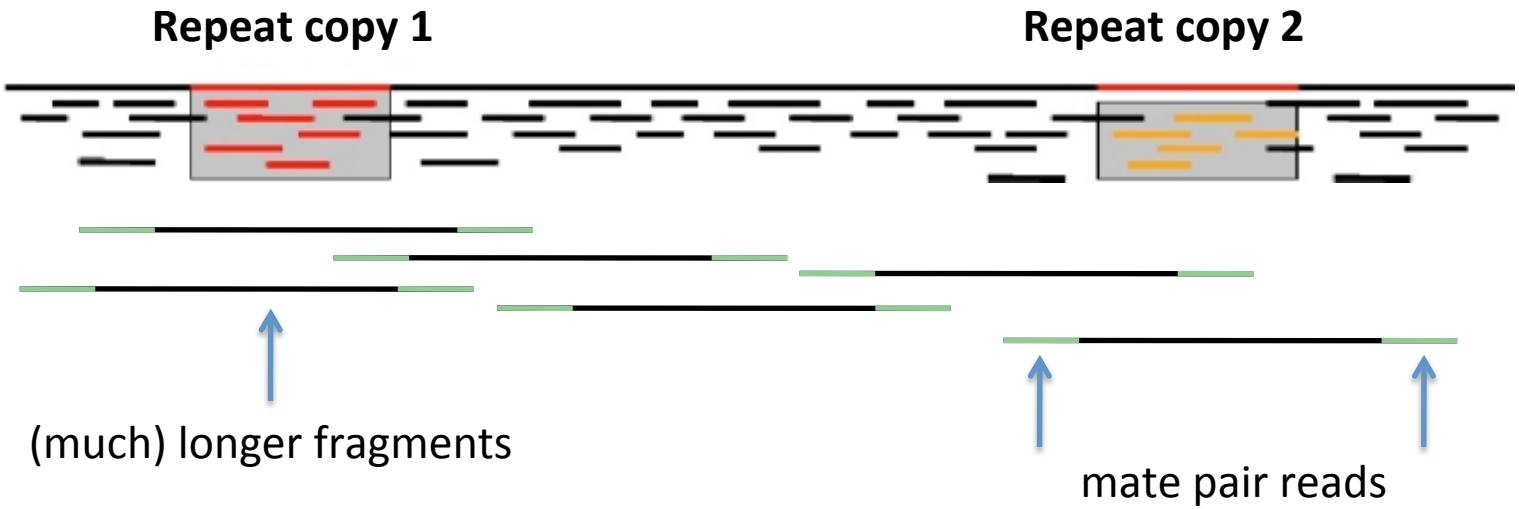
Repeats: major problem



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

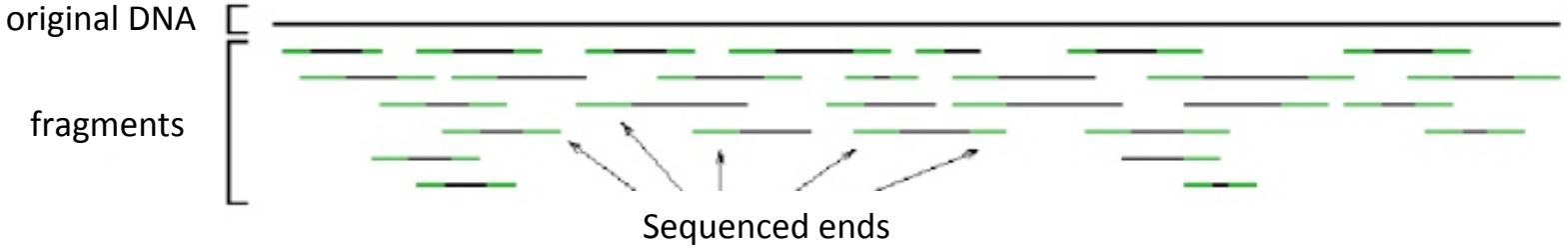
Mate pairs

Other read type

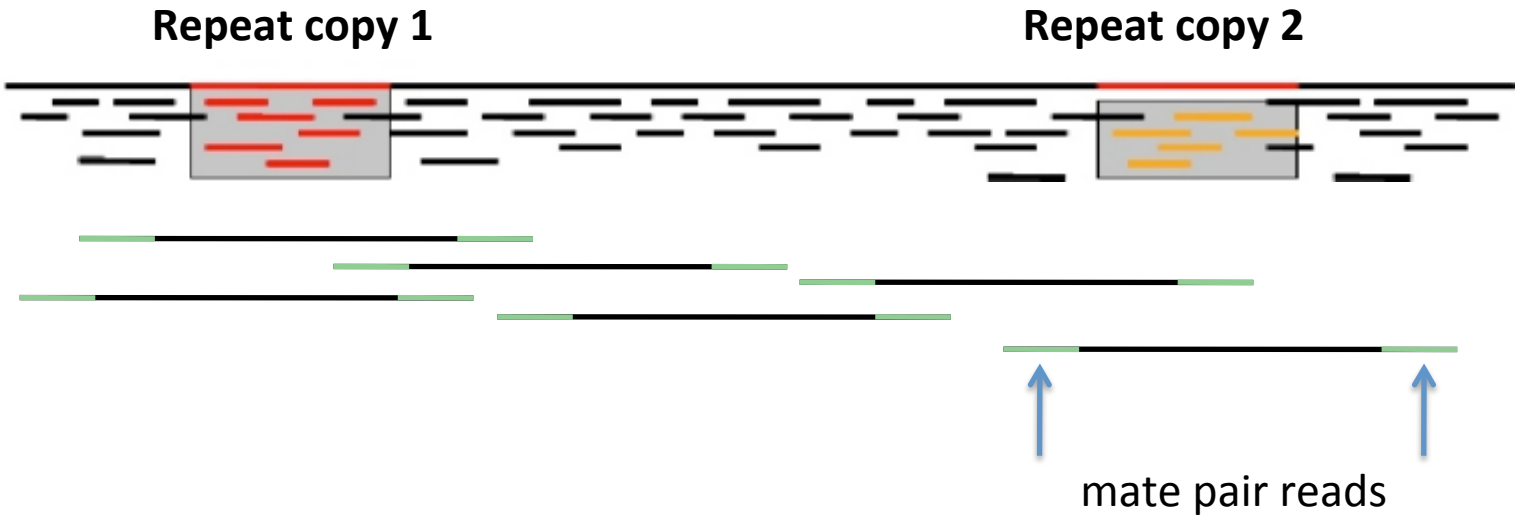


Mate pairs

Paired end reads → 100-500 bp insert

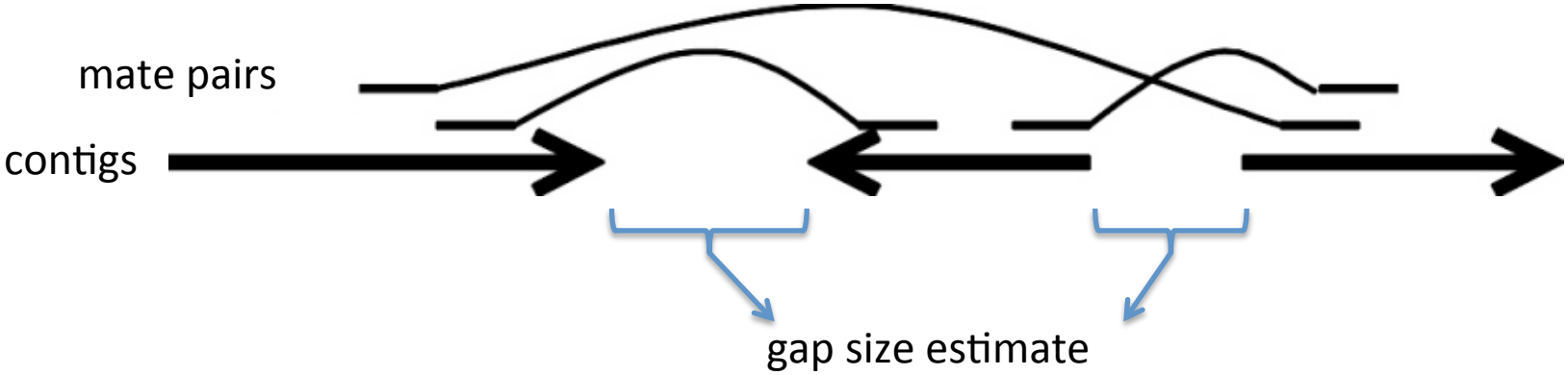


Mate pairs → 2-20 kb insert

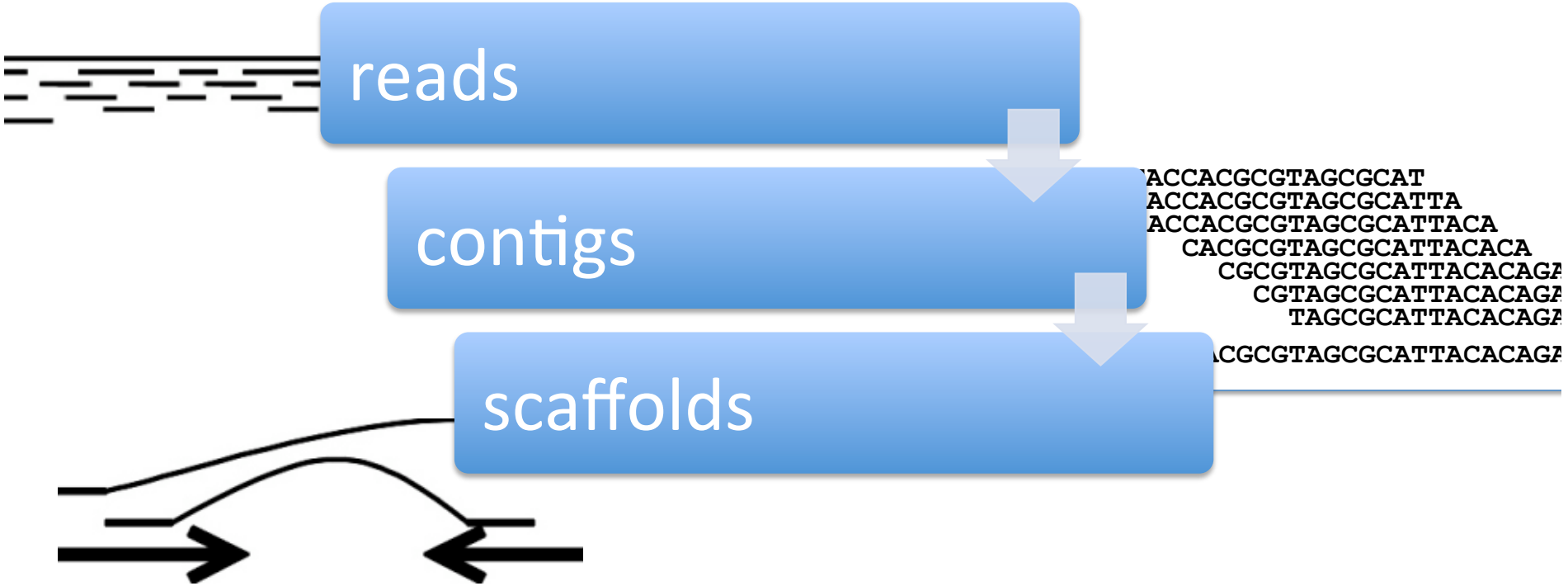


Scaffolds

Ordered, oriented contigs



Hierarchical structure



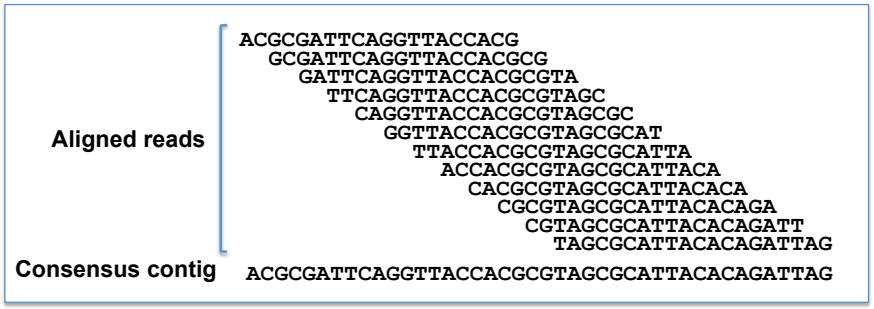
Assembly

How to do this?

Algorithms

Algorithms

All are graph-based

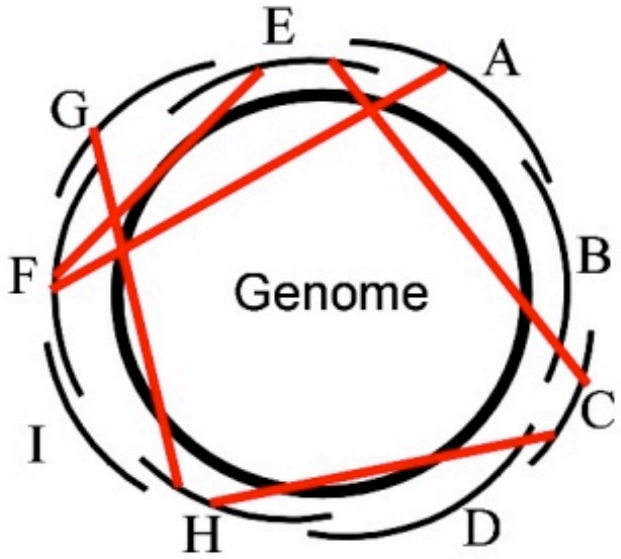


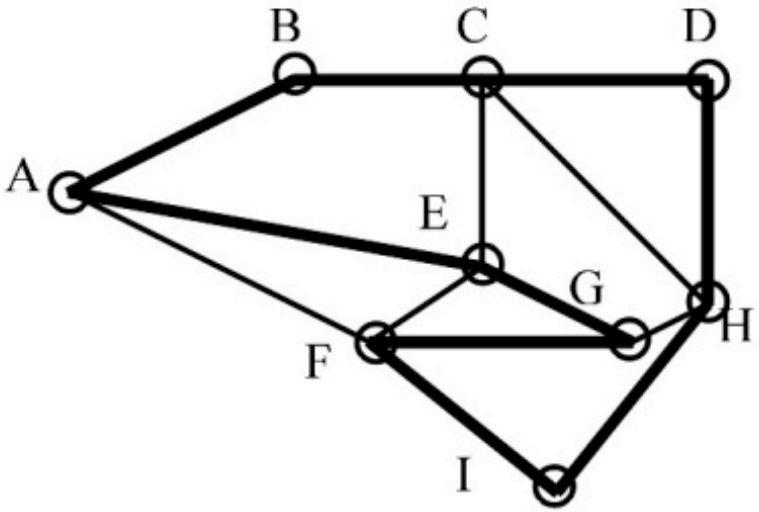
Graph-theory!

Algorithms

Hamiltonian path

– a path that contains all the nodes





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

Algorithms

Overlap calculation (alignment)

– computationally intensive

Aligned reads

```
ACGCGATTCAGGTTACCACG
  GCGATTCAGGTTACCACGCG
    GATTCAGGTTACCACGCGTA
      TTCAGGTTACCACGCGTAGC
        CAGGTTACCACGCGTAGCGC
          GGTTACCACGCGTAGCGCAT
            TTACCACGCGTAGCGCATTA
              ACCACGCGTAGCGCATTACA
                CACGCGTAGCGCATTACACA
                  CGCGTAGCGCATTACACAGA
                    CGTAGCGCATTACACAGATT
                      TAGCGCATTACACAGATTAG
```

Algorithms

Path through the graph

→contig

Algorithms

Many flavors

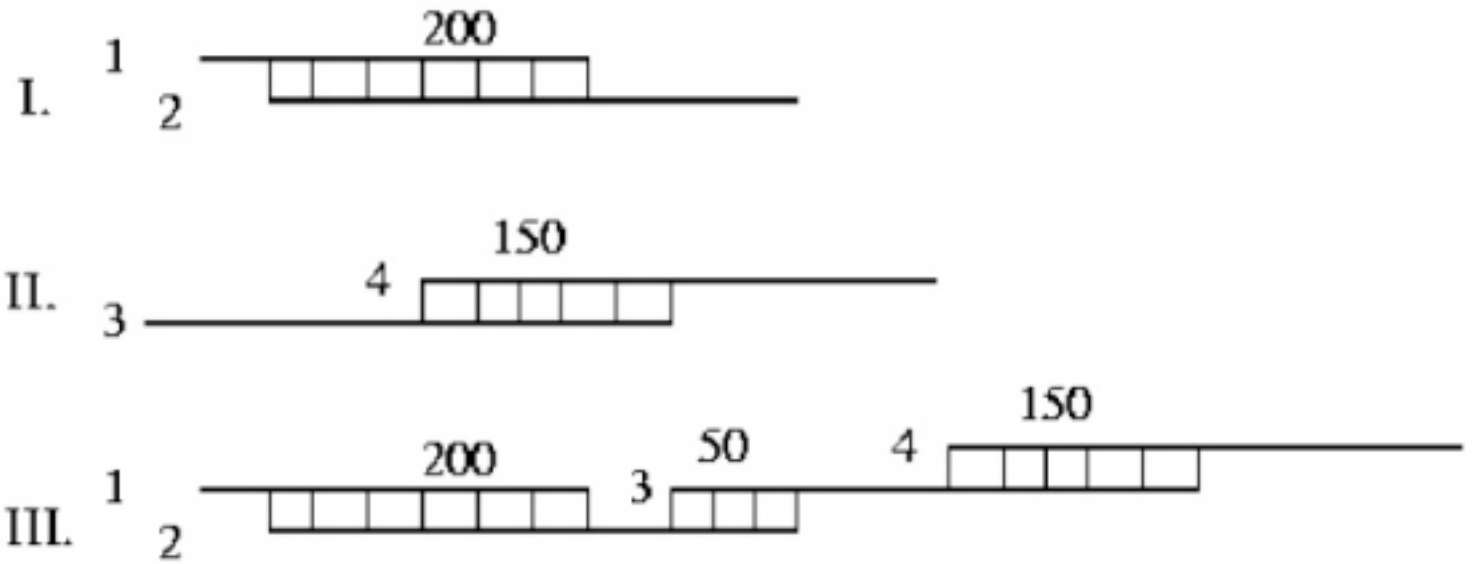
Abandoned
→ Greedy extension

Two most used
→ Overlap Layout Consensus
→ de Bruijn graph

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

Greedy extension

Oldest

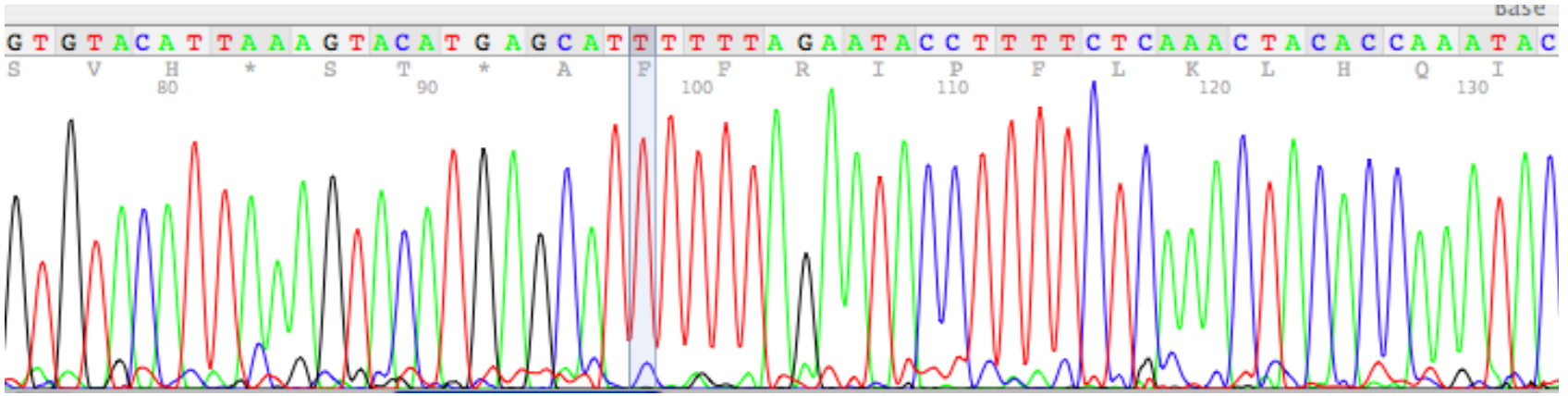


http://www.cbcb.umd.edu/research/assembly_primer.shtml

Overlap-Layout-Consensus

Typical for Sanger-type reads

– also used by newbler from 454 Life Sciences



Overlap-Layout-Consensus

Steps

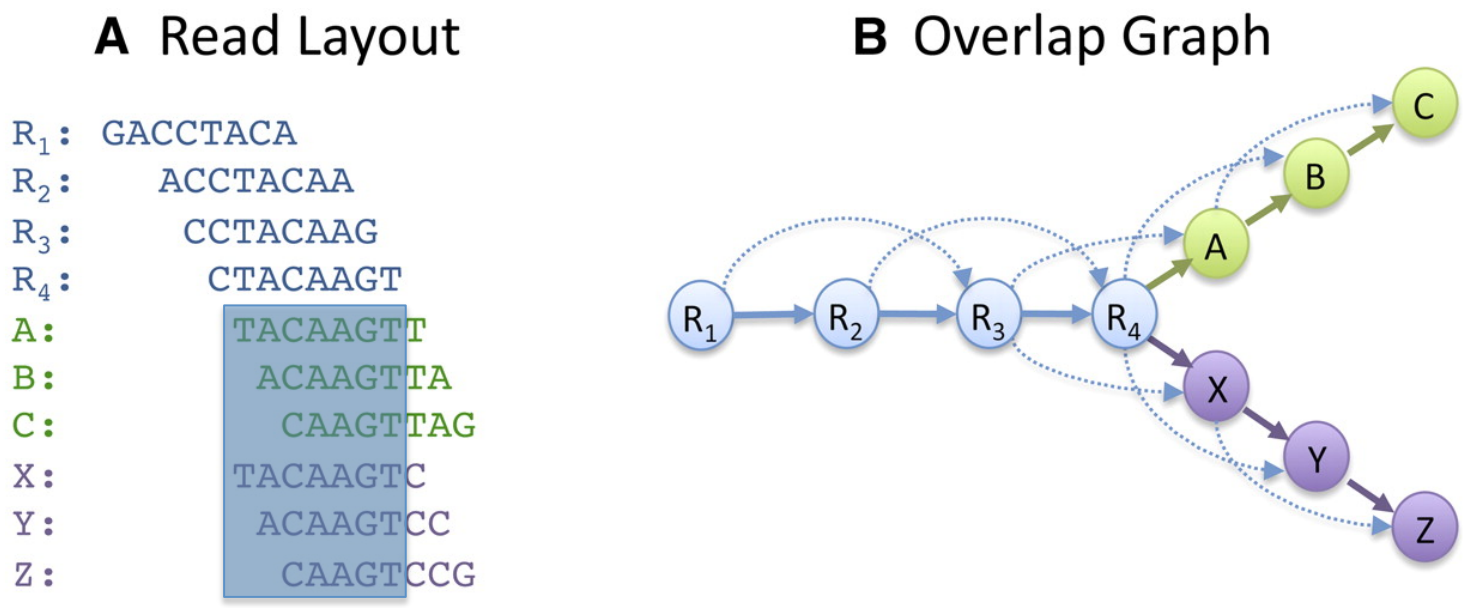
- Overlap computation
- Layout: graph simplification
- Consensus: sequence

Overlap-Layout-Consensus

Overlap phase:

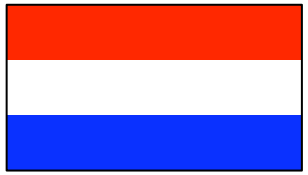
- K-mer seeds initiate overlap

Overlap-Layout-Consensus



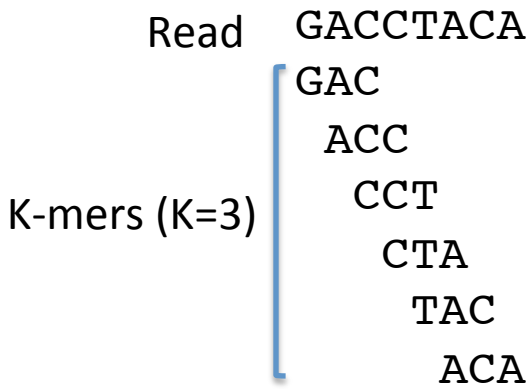
Schatz M C et al. Genome Res. 2010;20:1165-1173

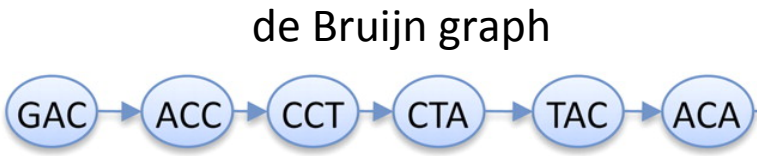
de Bruijn graphs



Developed outside of DNA-related work

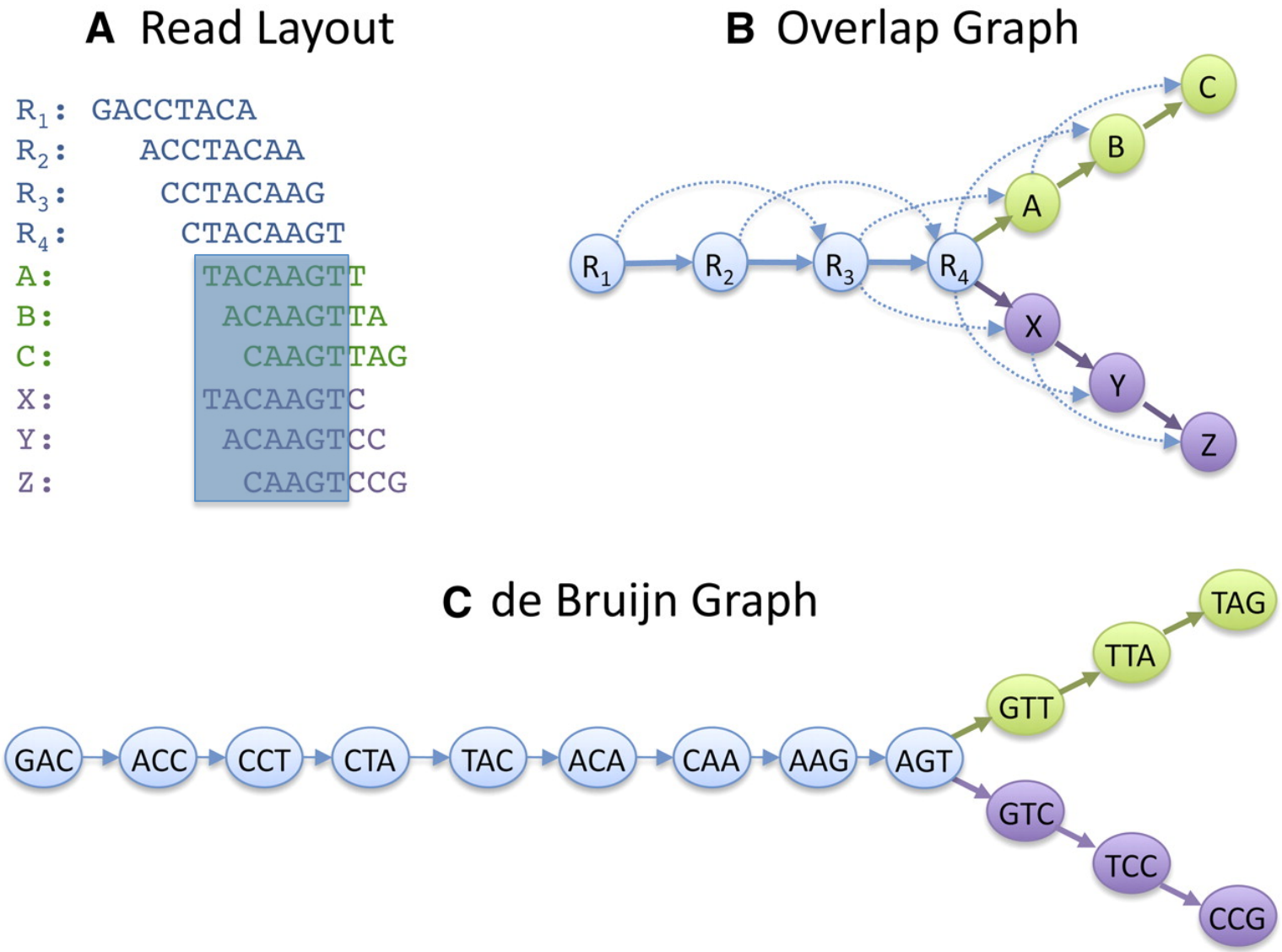
– Best solution for very short reads ≤100 nt





K-1 bases overlap

Graphs

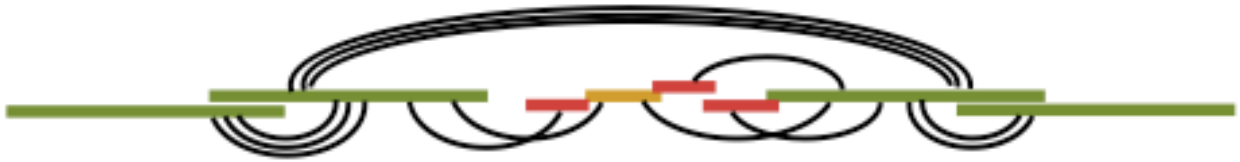


Schatz M C et al. Genome Res. 2010;20:1165-1173

Graphs

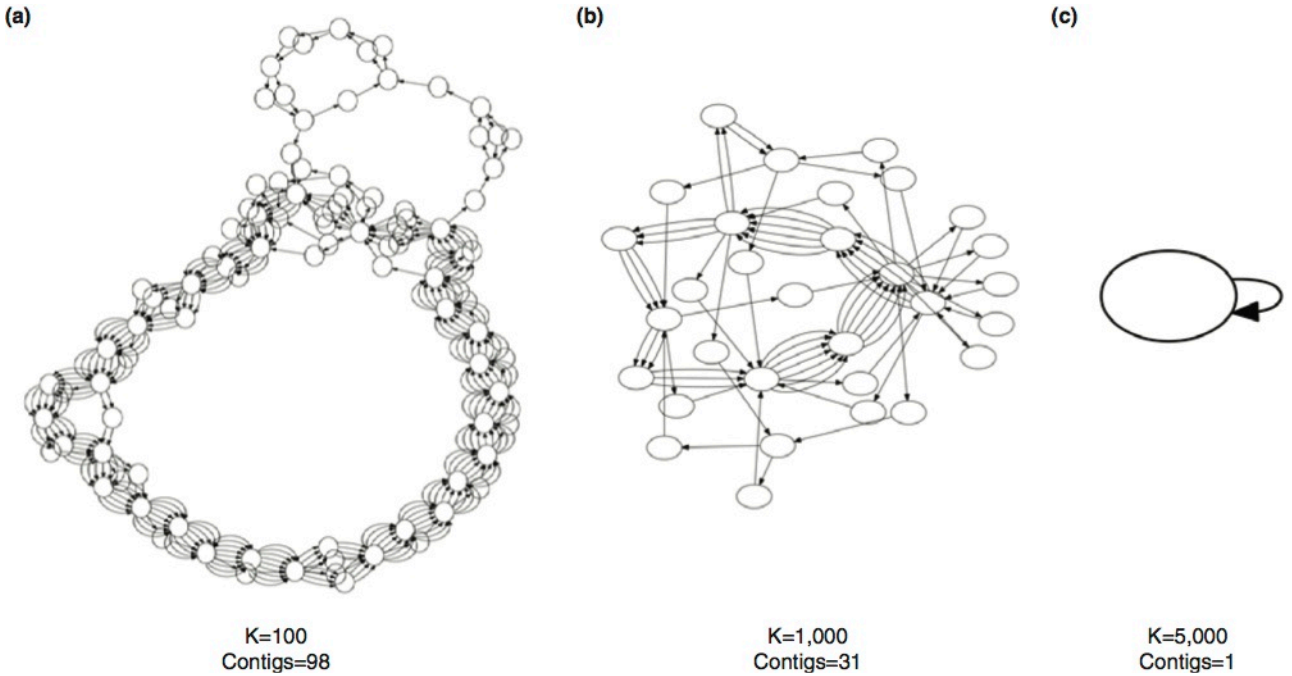
Simplify the graph

Add scaffolding information

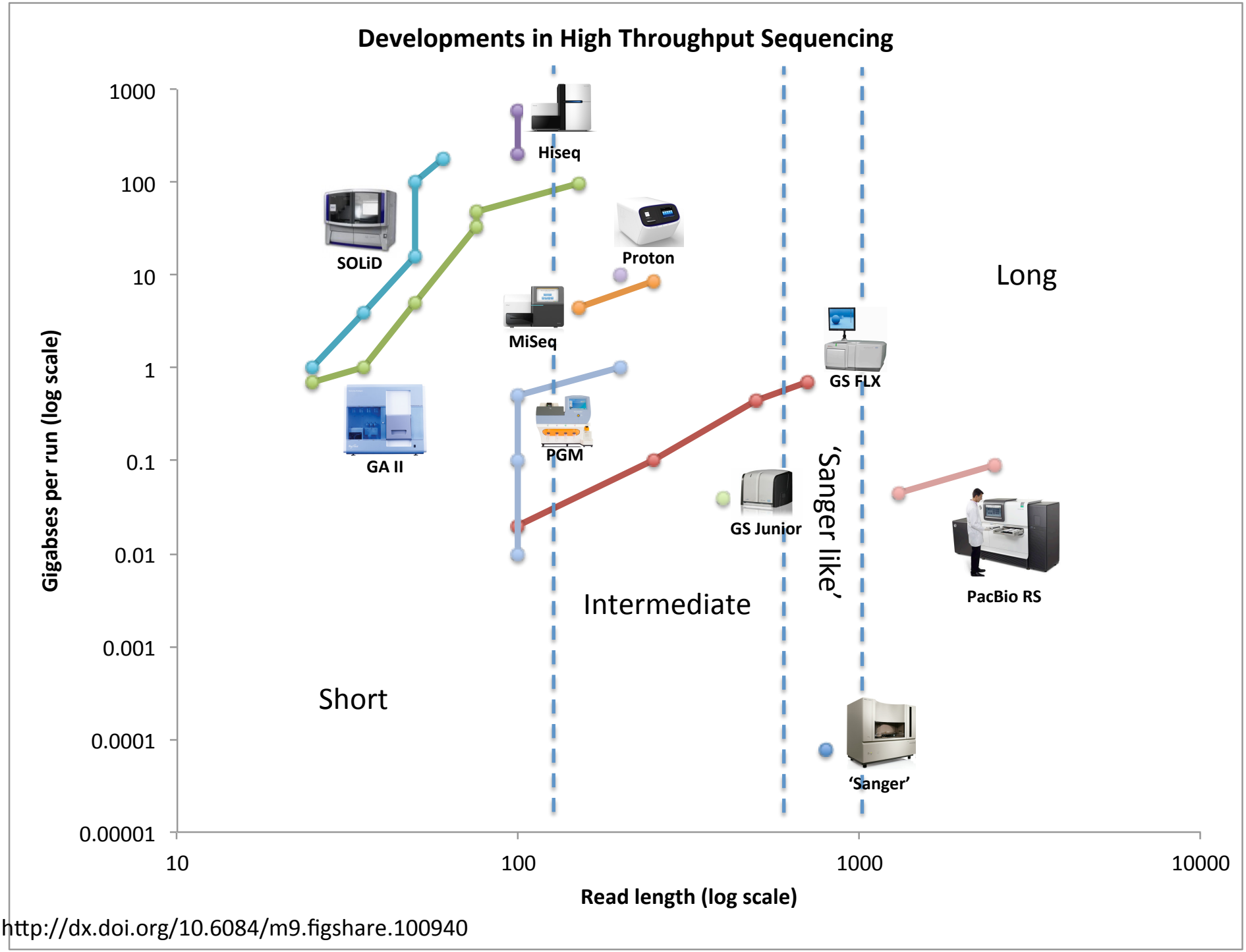


Read length matters

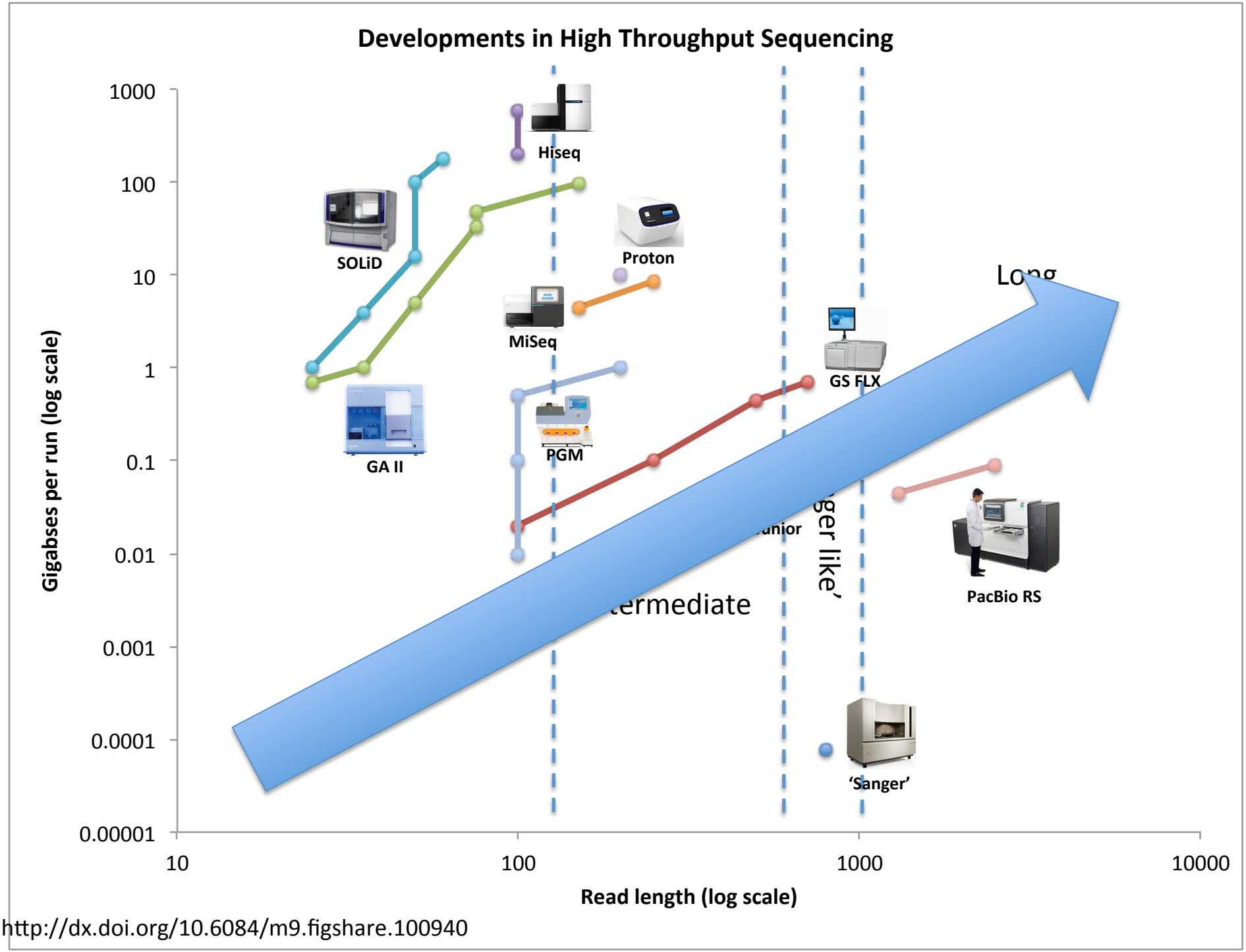
5.2 Mb circular genome, infinite error-free reads



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



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



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

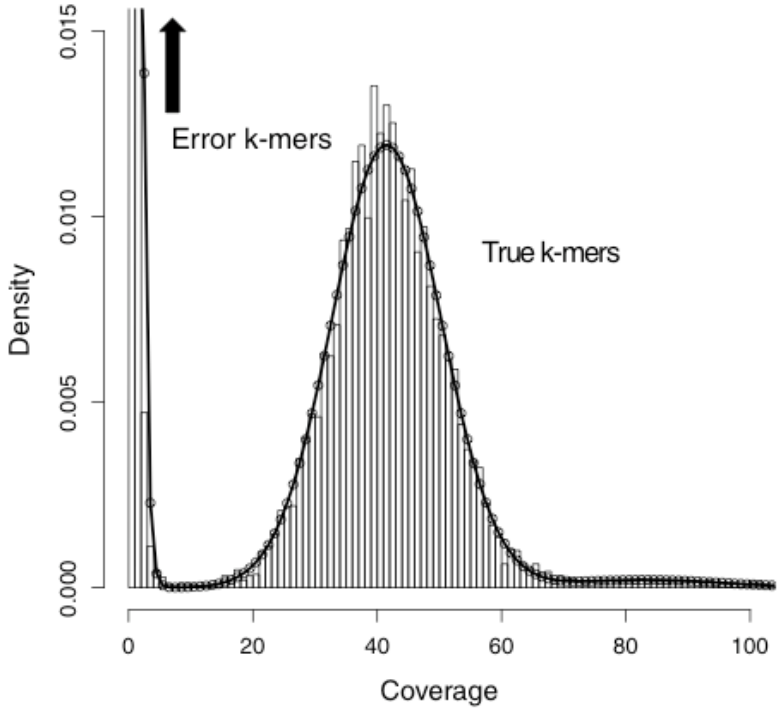
Quality matters

Sequencing errors

- add complexity to graph
- create new k-mers

Correction of errors

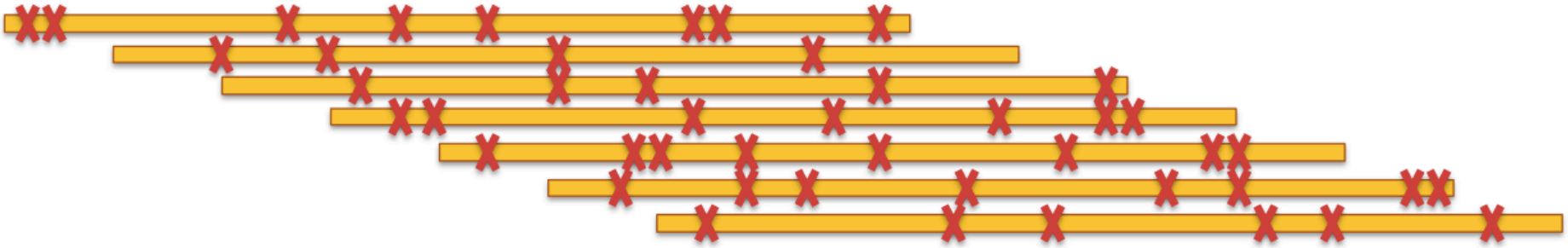
- k-mer frequency



Kelley *et al. Genome Biology* 2010 **11**:R116

Quality matters

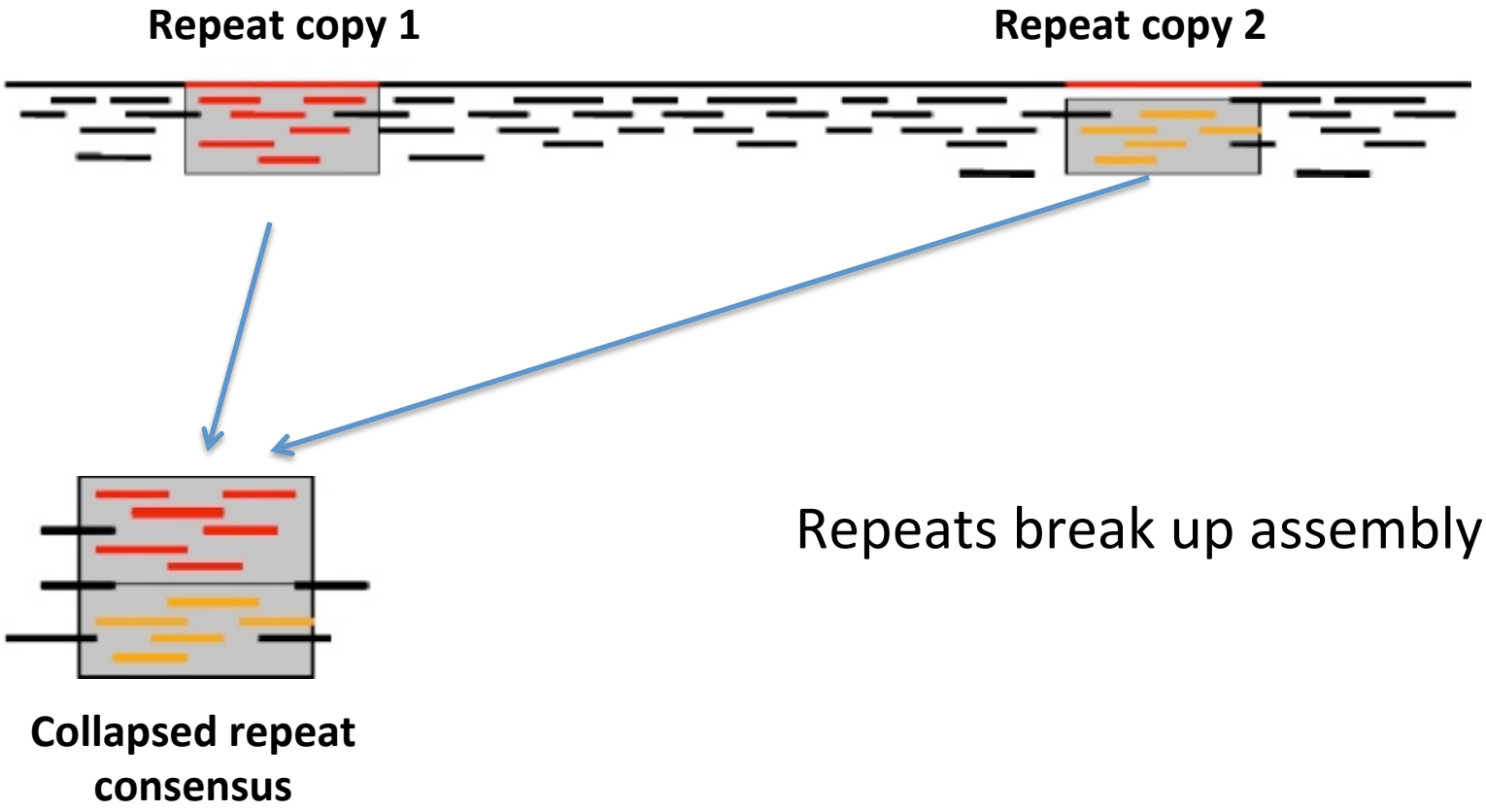
Too many errors → hard to find overlaps



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

Why is genome assembly such a difficult problem?

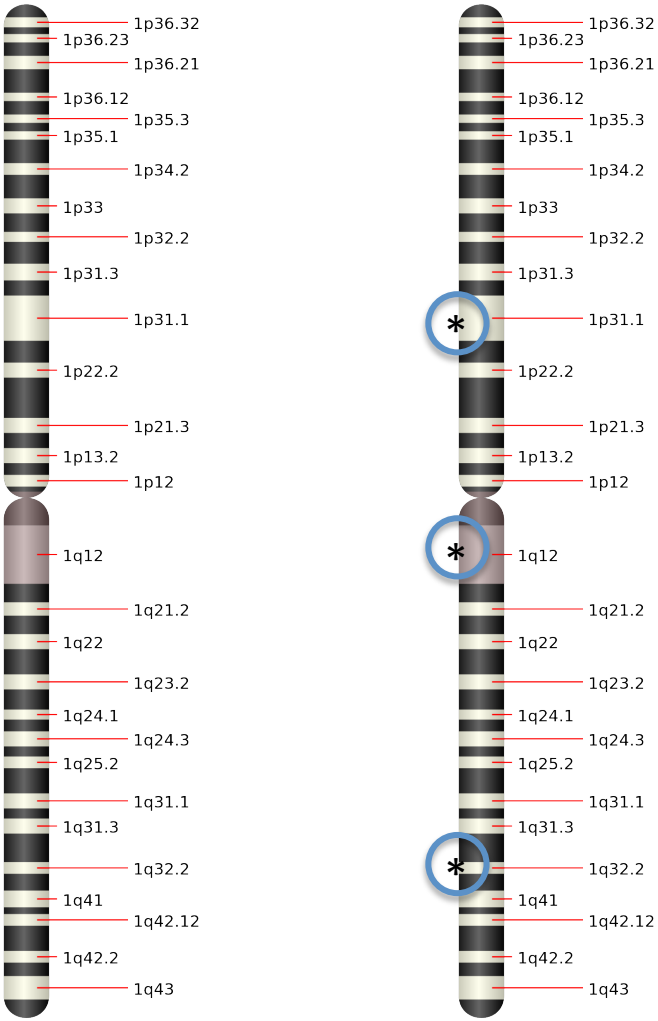
1) Repeats



Repeats break up assembly

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

2) Diploidy



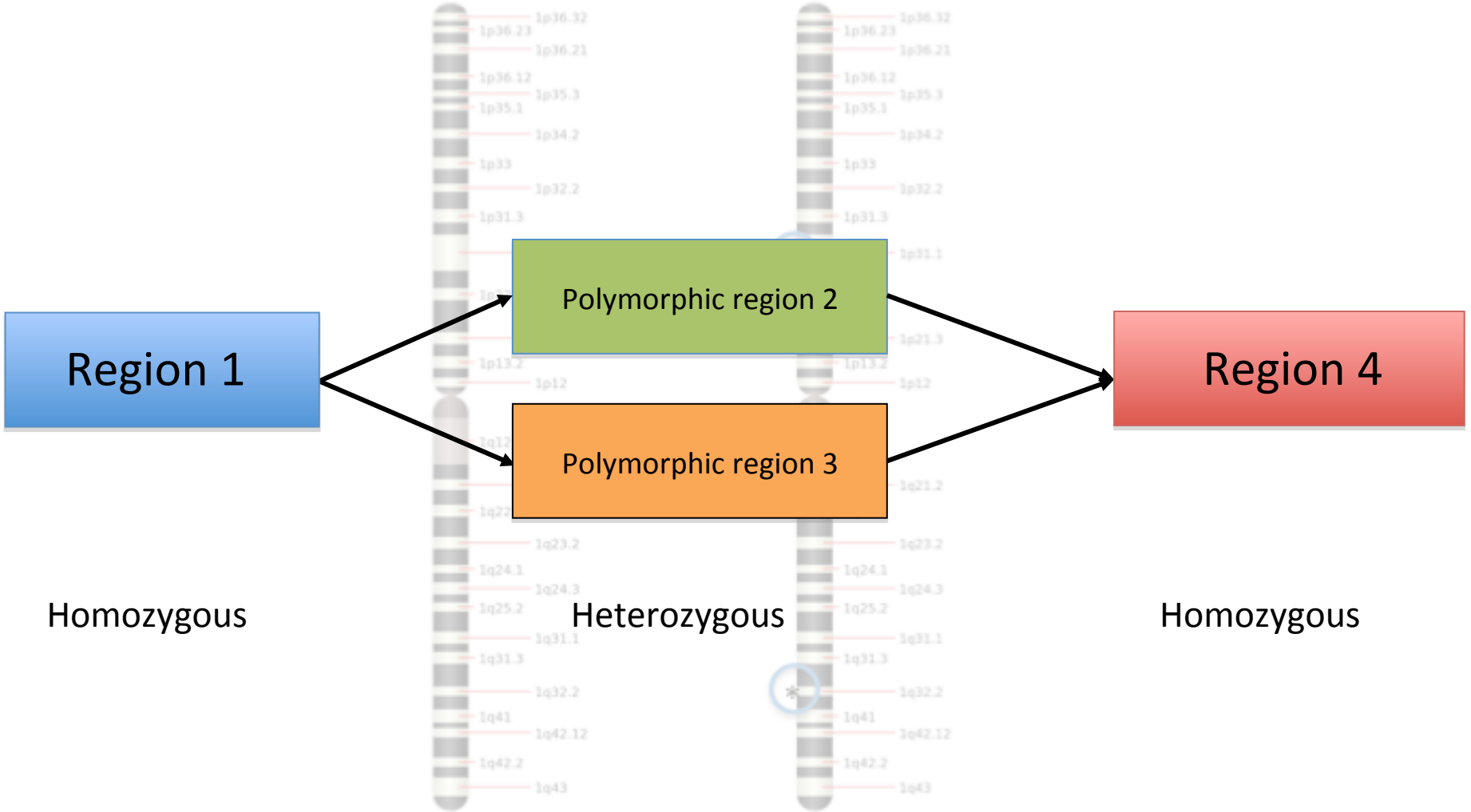
Differences between sister chromosomes

↓

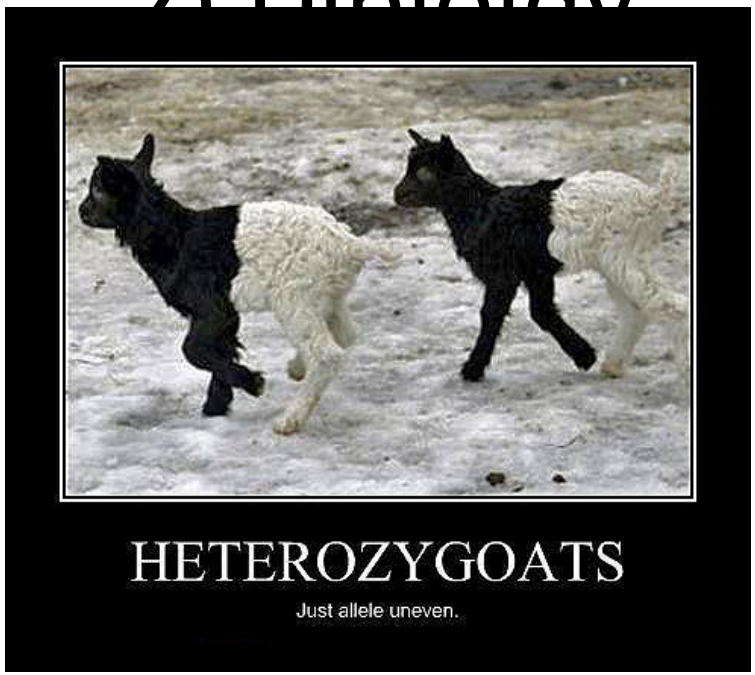
'heterozygosity'

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

2) Diploidy

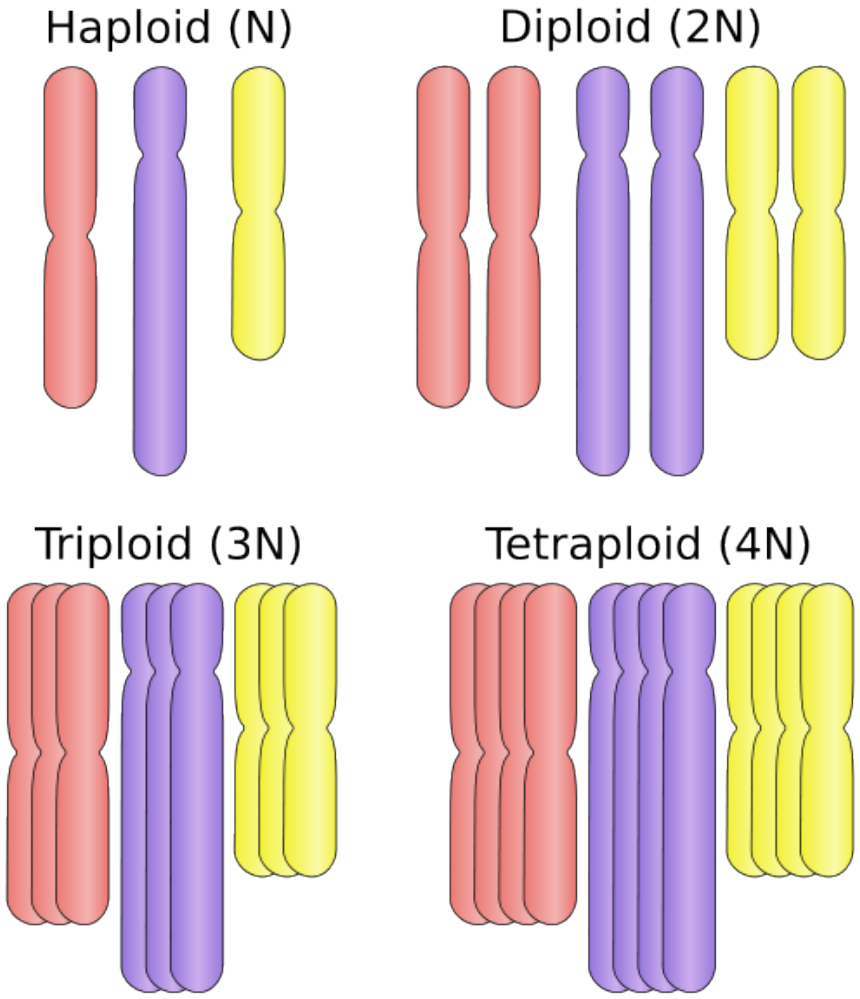

Region 1
Polymorphic region 2
Polymorphic region 3
Region 4
Homozygous
Heterozygous
Homozygous

2\ Diploidy



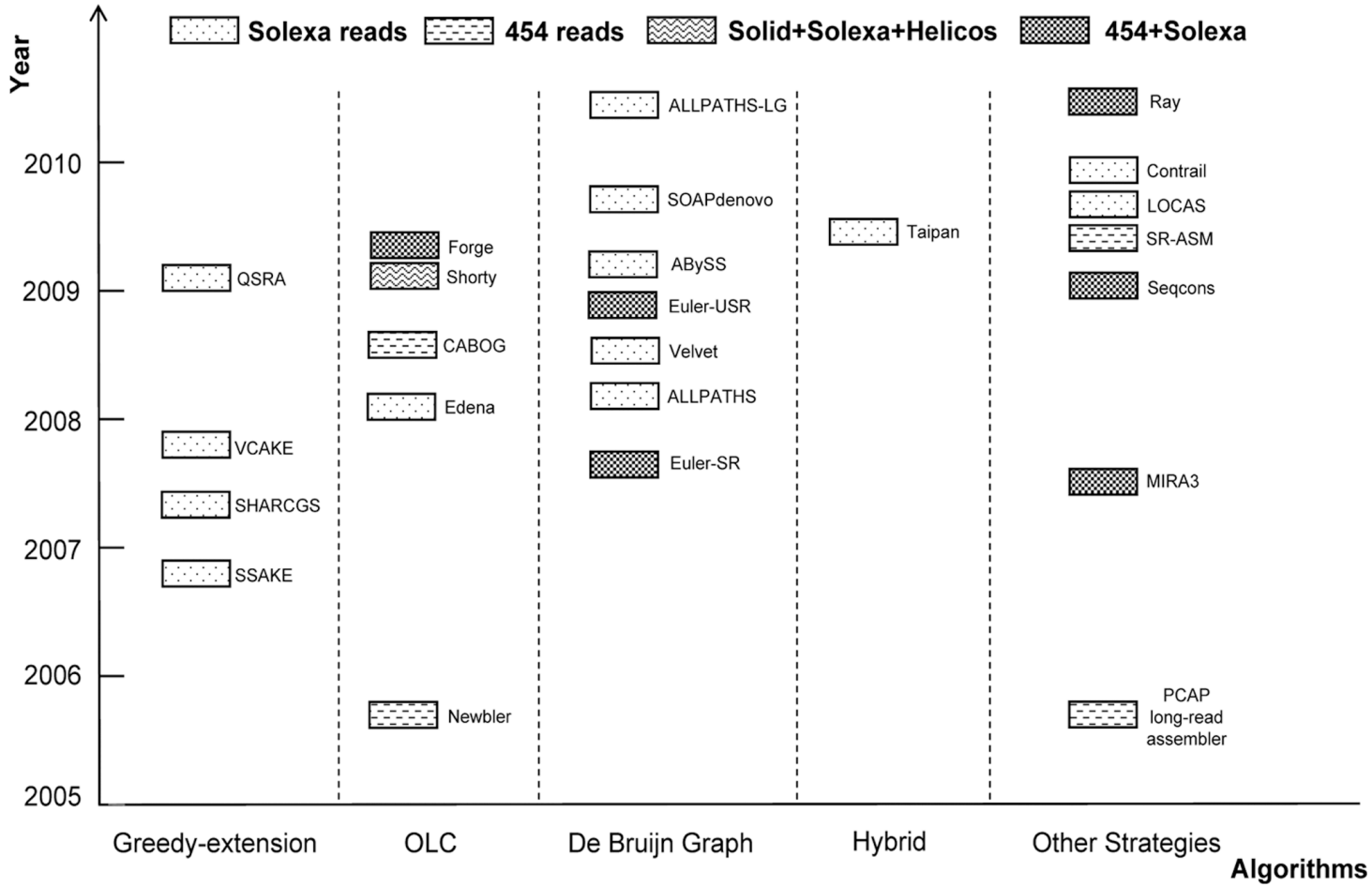
http://www.astraean.com/borderwars/wp-content/uploads/2012/04/heterozygoats.jpg
and many other sites

3) Polyploidy



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

4) Many programs to choose from



Zhang et al. PLoSOne 2011

Shameless self-promotion