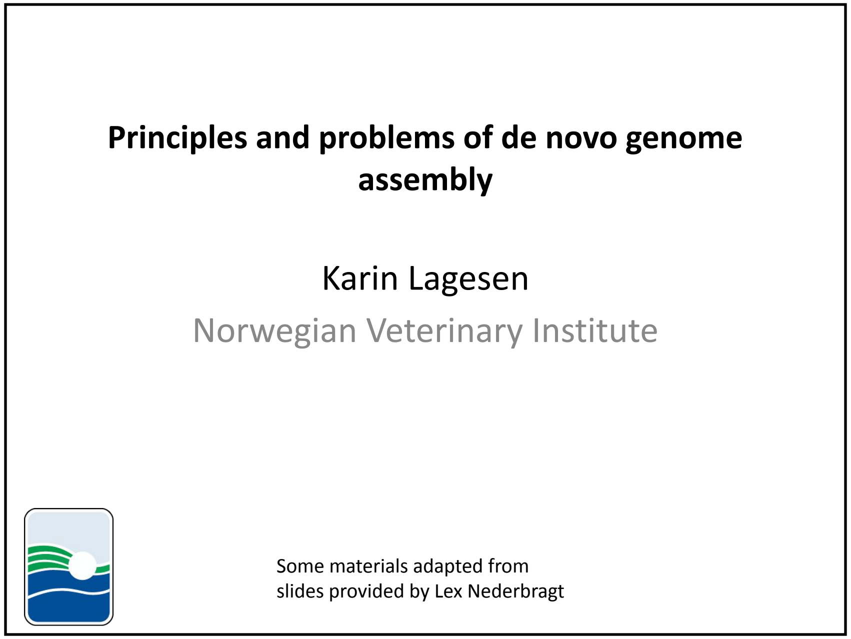
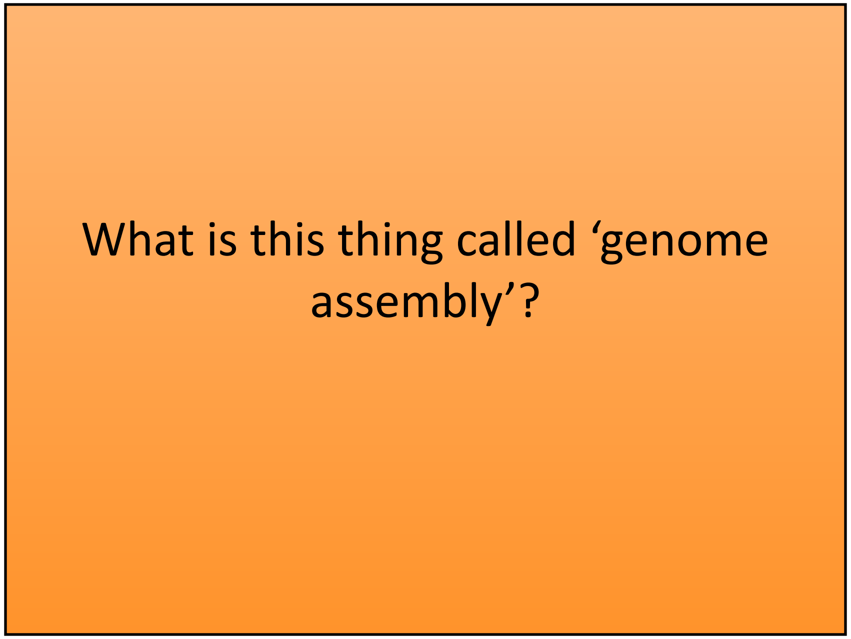


Principles and problems of de novo genome assembly

Karin Lagesen

Norwegian Veterinary Institute

Some materials adapted from
slides provided by Lex Nederbragt



What is this thing called 'genome assembly'?

What is a genome assembly?

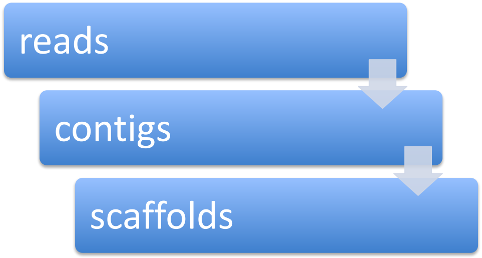
A hierarchical data structure

that maps the sequence data

to a putative reconstruction of the target

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

Hierarchical structure



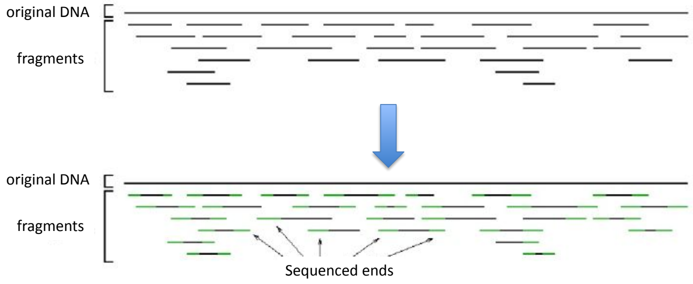
Sequence data

reads

contigs

scaffolds

Reads



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

Contigs

reads

contigs

scaffolds

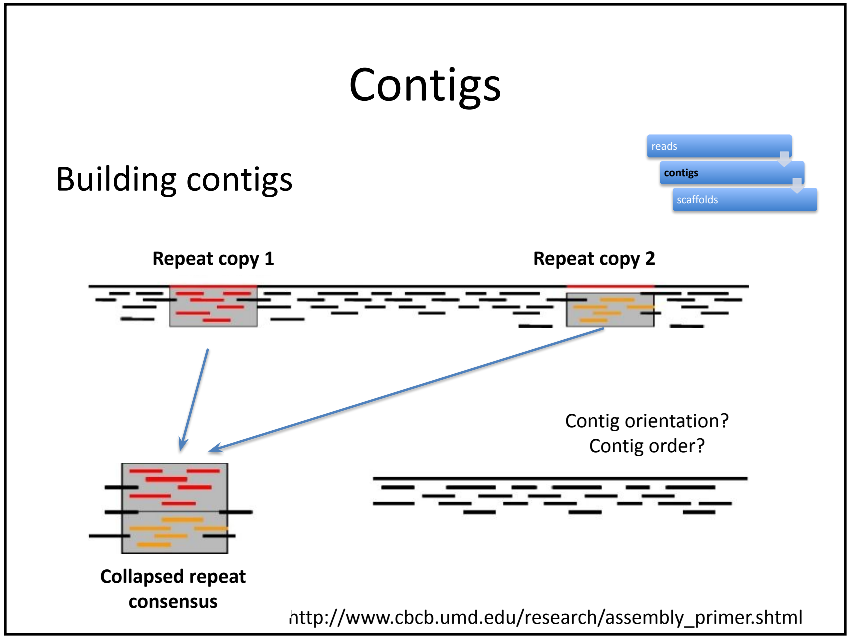
Building contigs

Aligned reads

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

Consensus contig

```
ACGCGATTCAGGTTACCACGCGTAGCGCATTACACAGATTAG
```



Contigs

Building contigs

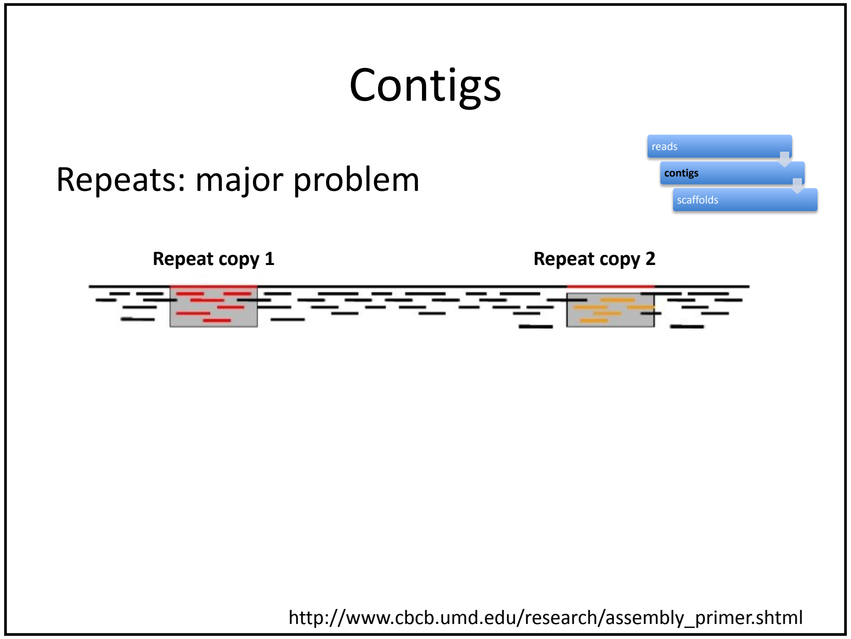
Repeat copy 1

Repeat copy 2

Contig orientation?
Contig order?

Collapsed repeat consensus

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



Contigs

Repeats: major problem

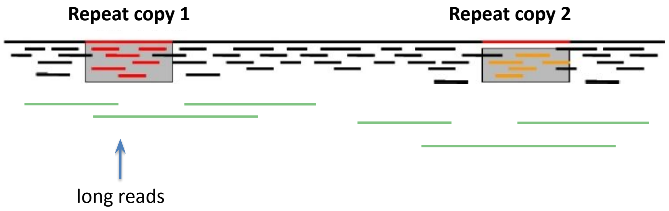
Repeat copy 1

Repeat copy 2

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

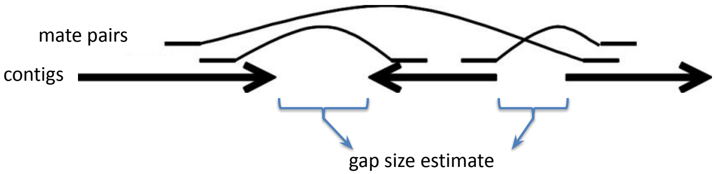
Long reads

Other read type

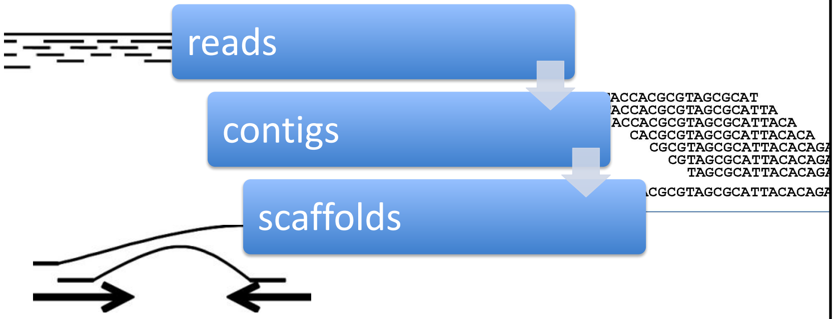


Scaffolds

Ordered, oriented contigs



Hierarchical structure



Assembly

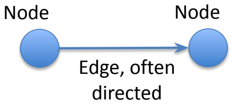
How to do this?

↓

Algorithms

Algorithms

All are graph-based



Node

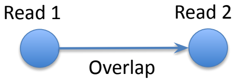
Node

Edge, often directed

Graph-theory!

Algorithms

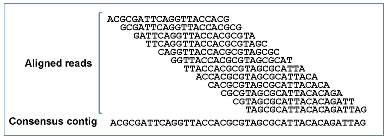
All are graph-based



Read 1

Read 2

Overlap



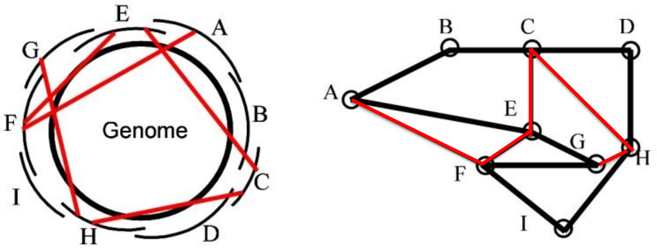
```
                 ACGCGATTCAGGTTACCACG
                   GCGATTCAGGTTACCACGCG
                     GATTCAGGTTACCACGCGTA
                       TTCAGGTTACCACGCGTAGC
                         CAGGTTACCACGCGTAGCGC
Aligned reads              GGTTACCACGCGTAGCGCAT
                             TTACCACGCGTAGCGCATTA
                               ACCACGCGTAGCGCATTACA
                                 CACGCGTAGCGCATTACACA
                                   CGCGTAGCGCATTACACAGA
                                     CGTAGCGCATTACACAGATT
                                       TAGCGCATTACACAGATTAG
Consensus contig ACGCGATTCAGGTTACCACGCGTAGCGCATTACACAGATTAG
```

Graph-theory!

Algorithms

Hamiltonian path

- a path that contains all the nodes

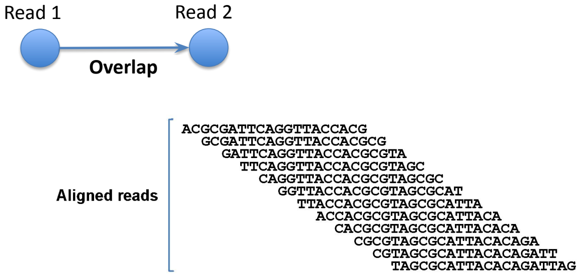


https://www.cbcb.umd.edu/research/assembly_primer

Algorithms

Overlap calculation (alignment)

- computationally intensive

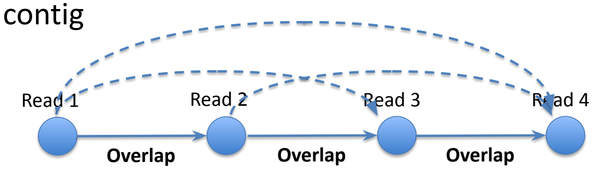


```
ACGCGATTCAGGTTACCACG
  GCGATTCAGGTTACCACGCG
    GATTCAGGTTACCACGCGTA
      TTCAGGTTACCACGCGTAGC
        CAGGTTACCACGCGTAGCGC
          GGTTACCACGCGTAGCGCAT
            TTACCACGCGTAGCGCATTA
              ACCACGCGTAGCGCATTACA
                CACGCGTAGCGCATTACACA
                  CGCGTAGCGCATTACACAGA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

Algorithms

Path through the graph

□contig



```
                 ACGCGATTCAGGTTACCACG
                   GCGATTCAGGTTACCACGCG
                     GATTCAGGTTACCACGCGTA
                       TTCAGGTTACCACGCGTAGC
                         CAGGTTACCACGCGTAGCGC
Aligned reads              GGTTACCACGCGTAGCGCAT
                             TTACCACGCGTAGCGCATTA
                               ACCACGCGTAGCGCATTACA
                                 CACGCGTAGCGCATTACACA
                                   CGCGTAGCGCATTACACAGA
                                     CGTAGCGCATTACACAGATT
                                       TAGCGCATTACACAGATTAG
Consensus contig  ACGCGATTCAGGTTACCACGCGTAGCGCATTACACAGATTAG
```

Algorithms

Many flavors

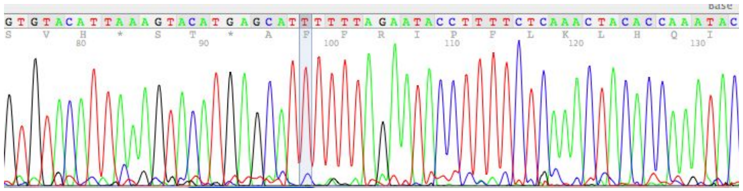
Two most used

- Overlap Layout Consensus
- de Bruijn graph

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

Overlap-Layout-Consensus

Developed for Sanger-type reads (longer reads)



Overlap-Layout-Consensus

Steps

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

Overlap-Layout-Consensus

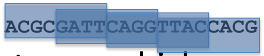
Overlap phase: find “similar enough” reads

Comparing all against all: expensive

Trick for finding “similar enough” reads:

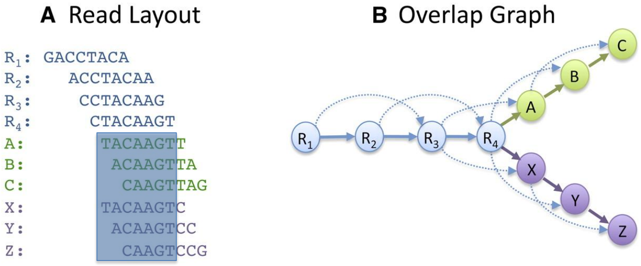
- Split reads into k-mers

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



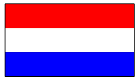
- Make list over which read has which k-mers
- If two reads share k-mers, test for similarity

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

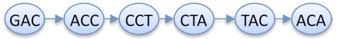
Read GACCTACA

K-mers (K=3)

```
GAC
 ACC
  CCT
   CTA
    TAC
     ACA
```

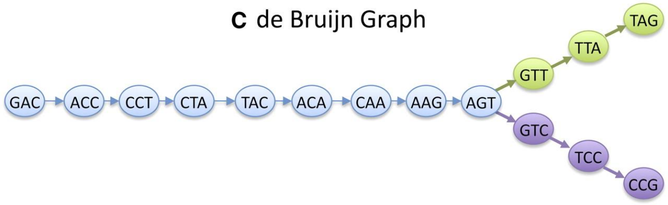
K-1 bases overlap

de Bruijn graph



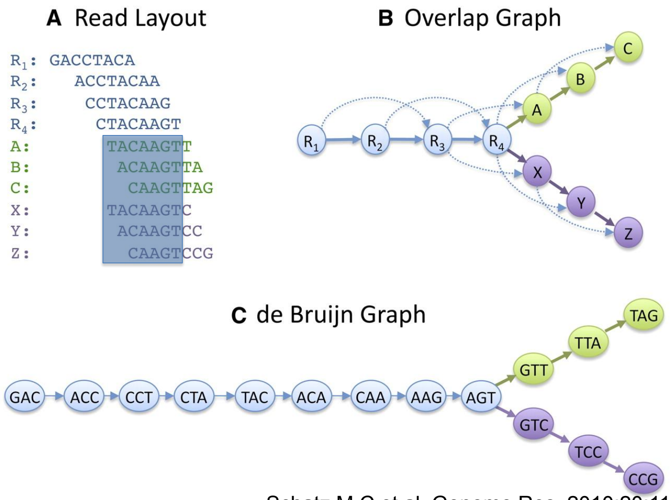
GAC
 ACC

Graphs



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

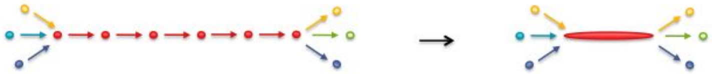
Graphs



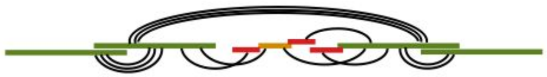
Schatz M C et al. Genome Res. 2010;20:1165-1173

Graphs

Simplify the graph



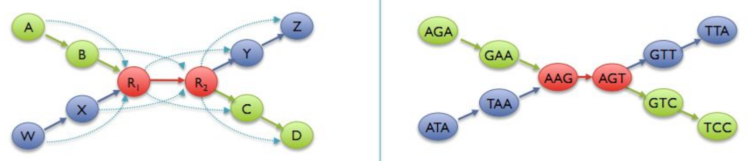
Add scaffolding information



de Bruijn Graphs

Overlap Graph

de Bruijn Graph



Long read assemblers

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

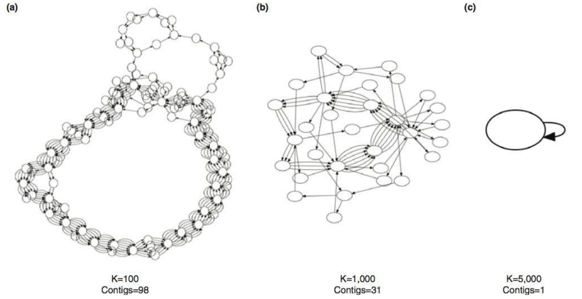
Short read assemblers

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

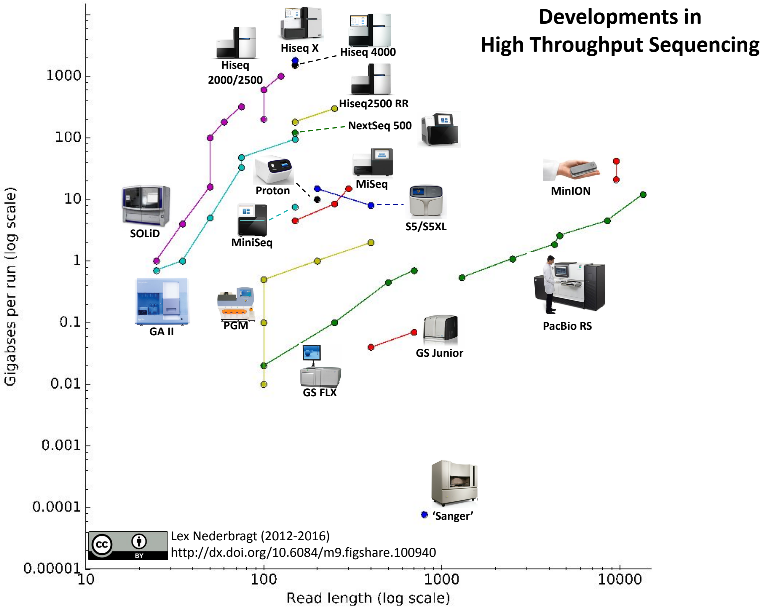
Mike Schatz

Read length matters

5.2 Mb circular genome, infinite error-free reads

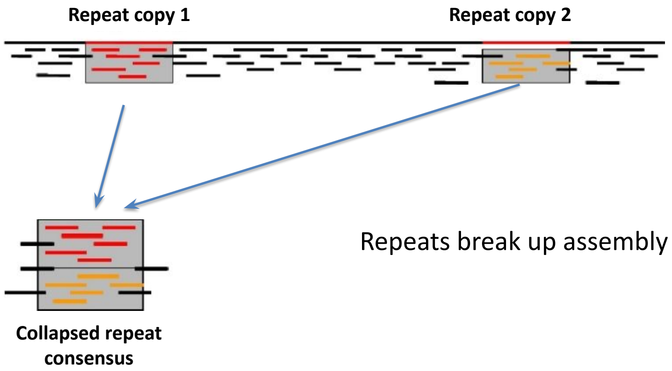


Roberts et al (2013) doi:10.1186/gb-2013-14-6-405



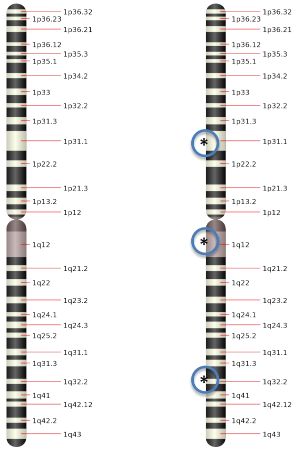
Why is genome assembly such a difficult problem?

1) Repeats



Repeats break up assembly

2) Diploidy



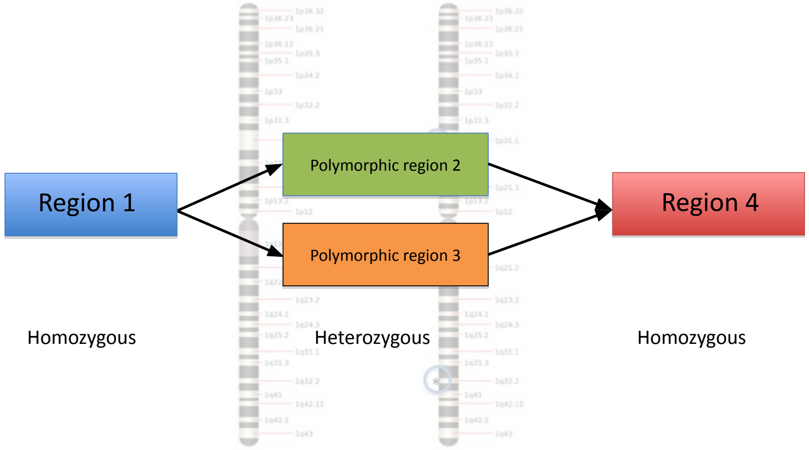
Differences between sister chromosomes

↓

'heterozygosity'

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

2) Diploidy



Region 1

Polymorphic region 2

Polymorphic region 3

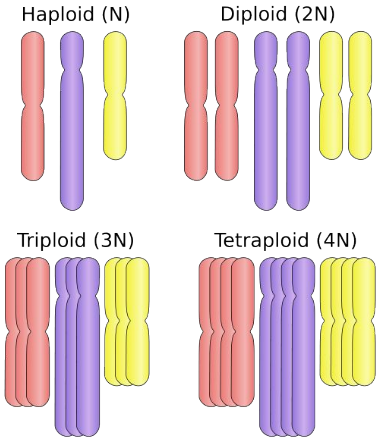
Region 4

Homozygous

Heterozygous

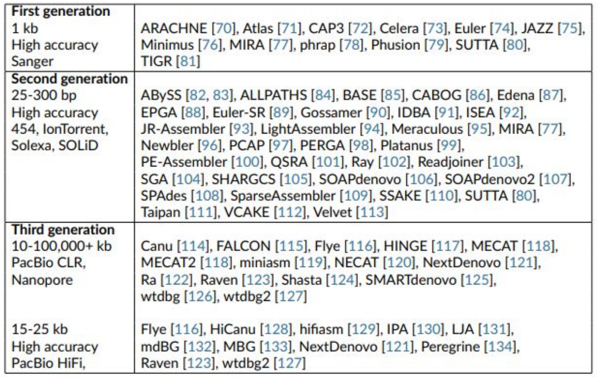
Homozygous

3) Polyploidy



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

4) Lots of tools to choose from

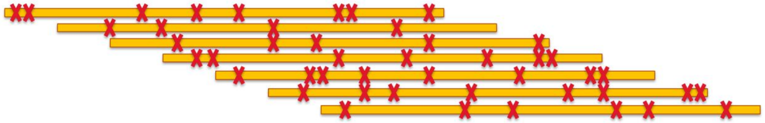


First generation 1 kb High accuracy Sanger	ARACHNE [70], Atlas [71], CAP3 [72], Celera [73], Euler [74], JAZZ [75], Minimus [76], MIRA [77], phrap [78], Phusion [79], SUTTA [80], TIGR [81]
Second generation 25-300 bp High accuracy 454, IonTorrent, Solexa, SOLiD	ABySS [82, 83], ALLPATHS [84], BASE [85], CABOG [86], Edena [87], EPGA [88], Euler-SR [89], Gossamer [90], IDBA [91], ISEA [92], JR-Assembler [93], LightAssembler [94], Meraculous [95], MIRA [77], Newbler [96], PCAP [97], PERGA [98], Platanus [99], PE-Assembler [100], QSRA [101], Ray [102], Readjoiner [103], SGA [104], SHARGCS [105], SOAPdenovo [106], SOAPdenovo2 [107], SPAdes [108], SparseAssembler [109], SSAKE [110], SUTTA [80], Taipan [111], VCAKE [112], Velvet [113]
Third generation 10-100,000+ kb PacBio CLR, Nanopore	Canu [114], FALCON [115], Flye [116], HINGE [117], MECAT [118], MECAT2 [118], miniasm [119], NECAT [120], NextDenovo [121], Ra [122], Raven [123], Shasta [124], SMARTdenovo [125], wtdbg [126], wtdbg2 [127]
15-25 kb High accuracy PacBio HiFi,	Flye [116], HiCanu [128], hifiasm [129], IPA [130], LJA [131], mdBG [132], MBG [133], NextDenovo [121], Peregrine [134], Raven [123], wtdbg2 [127]

Guiglielmoni, *et. al.* Peer Comm. J. 2022

Assembly with noisy single molecule sequencing data

Usage of long reads

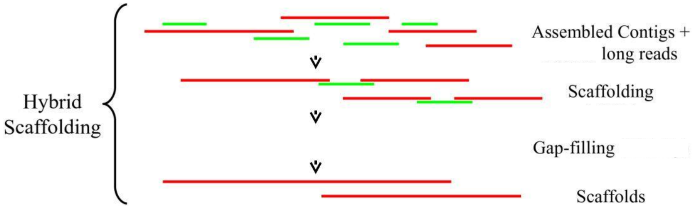


- Problem: higher error rates
- Overlaps more difficult/expensive to find
- OLC more commonly used than for 2nd generation data

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)

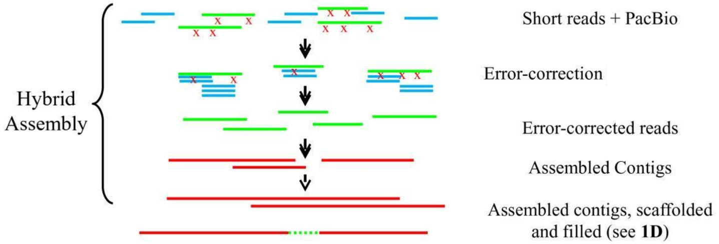


Powers *et.al.*, BMC genomics 2013

Correct, assemble

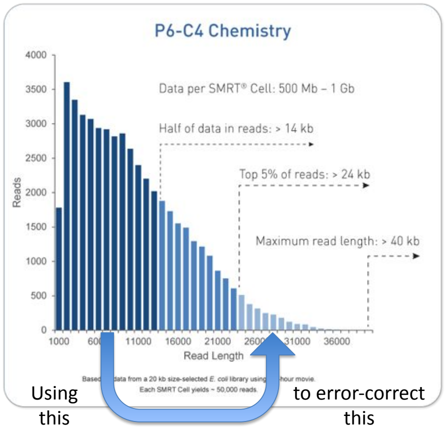
- Do pairwise comparison, find shorter reads that support the longer
- Align supporting reads, correct longer reads
- Overlap-Layout-Consensus on corrected reads
- Polish assembly

Mapping and error correcting (hybrid)



Powers *et.al.*, BMC genomics 2013

Hierarchical approach (self-correcting)



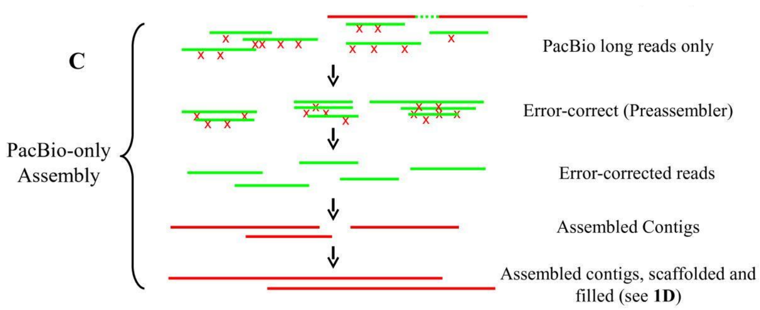
Based data from a 20 kb size-selected *E. coli* library using hour movie.
Each SMRT Cell yields ~ 50,000 reads.

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?