Melbourne Bioinformatics
BIOINFORMATICS + DATA SERVICES + INFRASTRUCTURE, FOR LIFE SCIENCES TODAY

THE UNIVERSITY OF MELBOURNE

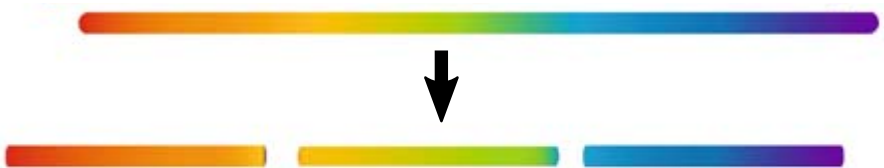
De novo genome assembly

Computational Genomics | Lecture 14

Tom Harrop

melbournebioinformatics.org.au

De novo genome assembly



- regions that can't be assembled result in ***gaps*** in the assembly
- the assembled fragments are called ***contigs*** (contiguous sequence)
- contigs can be joined into ***scaffolds*** with the gaps filled by Ns

Imagine trying to reassemble a book from strings of letters…

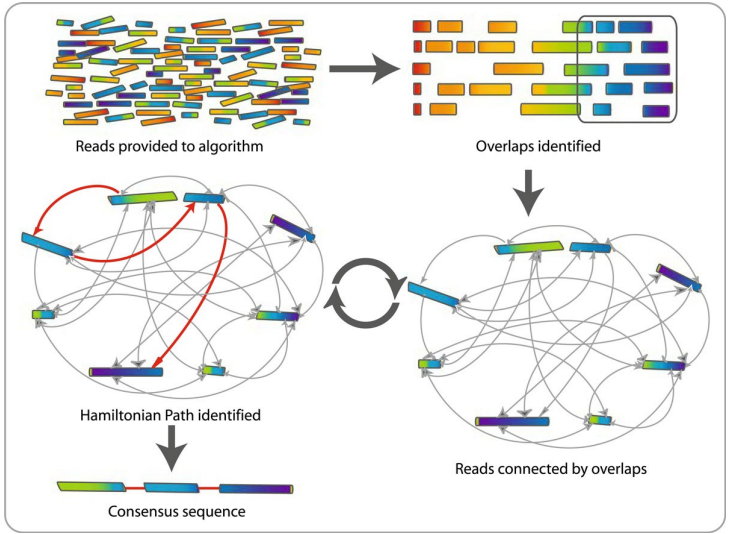
It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity

```
# "genome"
itwasthebestoftimesitwastheworstoftimesitwastheageofw…

# short "reads" (8 b)
itwasthe
 twastheb
  wasthebe
   asthebes
     ...

itwasthe           itwasthe            itwasthe

# longer "reads" (19 b)
         estoftimesitwasthew
```

```
itwasthebest
itwasthewors
itwastheageo
itwastheepoc
```

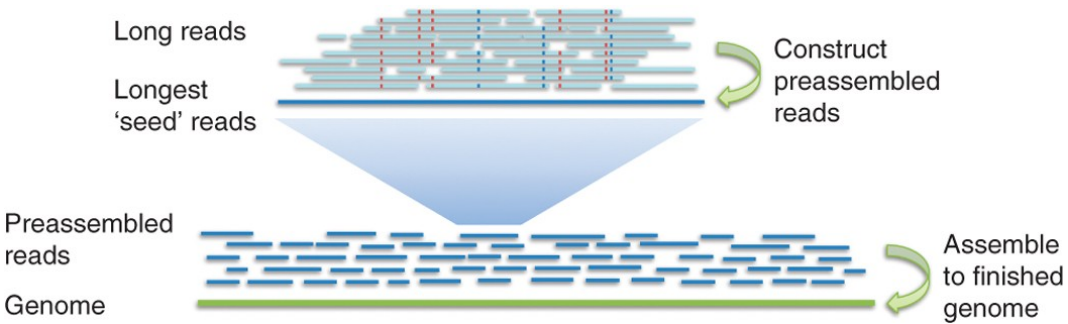
Charles Dickens, *A Tale of Two Cities*. Example: Jeremy Leipzig

Overlap-layout-consensus assembly (OLC)



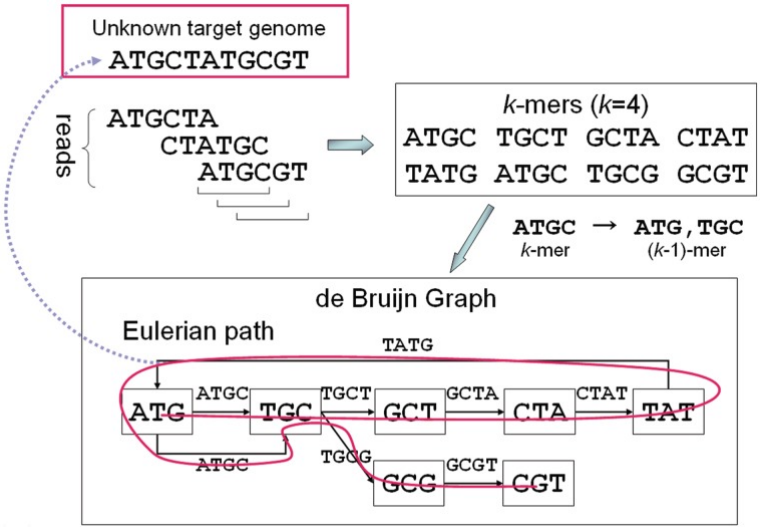
Commins *et al*., 2009

Long-read OLC assembly



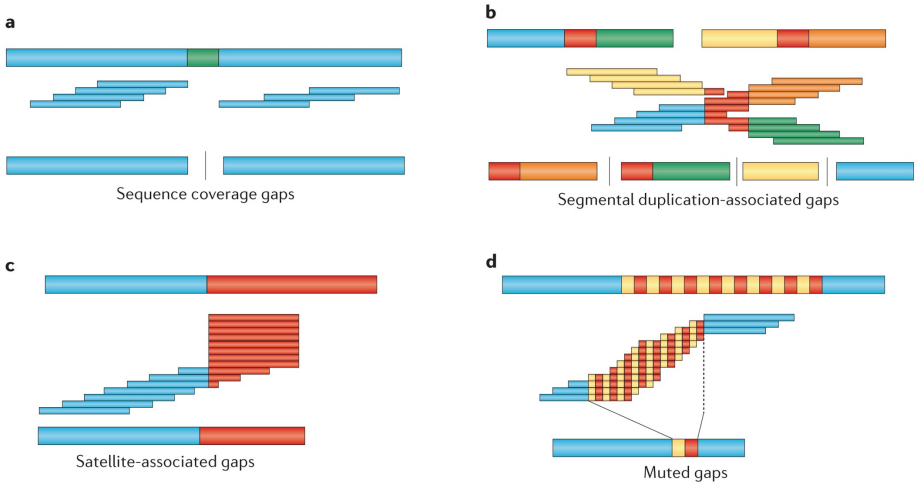
Chin *et al.*, 2013

de Bruijn graph assembly



Namiki *et al*., 2012

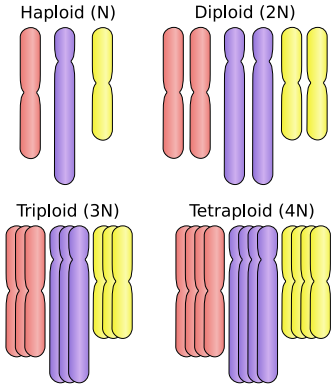
Challenges



Nature Reviews | Genetics

Chaisson *et al.*, 2015

Challenges - ploidy

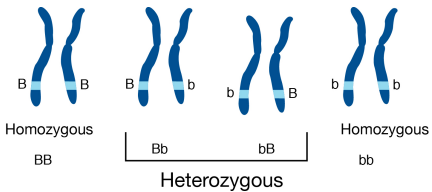


Ehamberg via Wikimedia Commons

Ploidy:

- diploid organisms have two ***homologous copies*** of each chromosome
- ***heterozygous*** individuals have different alleles on homologous chromosomes
- a genome sequence is intended to be a ***haploid representation*** of the organism's genetic material

Challenges - ploidy

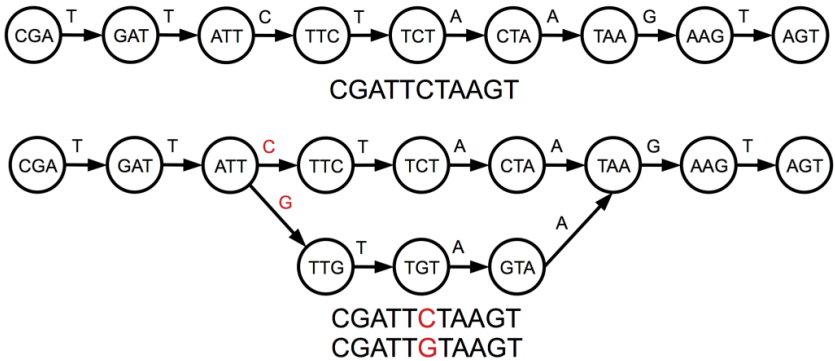


Ploidy:

- diploid organisms have two ***homologous copies*** of each chromosome
- ***heterozygous*** individuals have different alleles on homologous chromosomes
- a genome sequence is intended to be a ***haploid representation*** of the organism's genetic material

Darryl Leja via Wikimedia Commons

Challenges - ploidy



Leggett *et al.*, 2013

Some commonly-used assemblers

Short-read based

de Bruijn graph:

- `SPAdes`
- `Velvet`
- `AbySS`
- `DISCOVAR` / `ALLPATHS`
- `Meraculous`
- `SOAPdenovo`
- many more: see *De novo sequence assemblers* on Wikipedia

OLC algorithm:

- `wgs-assembler` (`celera`)

Long read (mostly OLC)

- `Canu`
- `FALCON`
- `Flye`
- `Shasta`

Hybrid

- `MaSuRCA`
- `Unicycler`

Special cases

- metagenomes, transcriptomes ...