25th - 29th March 2019

Genome Assembly

THE UNIVERSITY OF MELBOURNE

Melbourne Bioinformatics

Activities

- A paper genome assembly
- *de novo* Assembly of a bacterial genome sequence from an Illumina readset
- Annotation of the draft genome sequence

The first analysis step is either:

- ***de novo* assembly**
  - reconstruct the original sequences from reads alone
  - like a jigsaw puzzle but ambiguous
- **Align to reference (Read Mapping)**
  - find where reads fit on a known sequence
  - can not always be uniquely placed

de novo Assembly

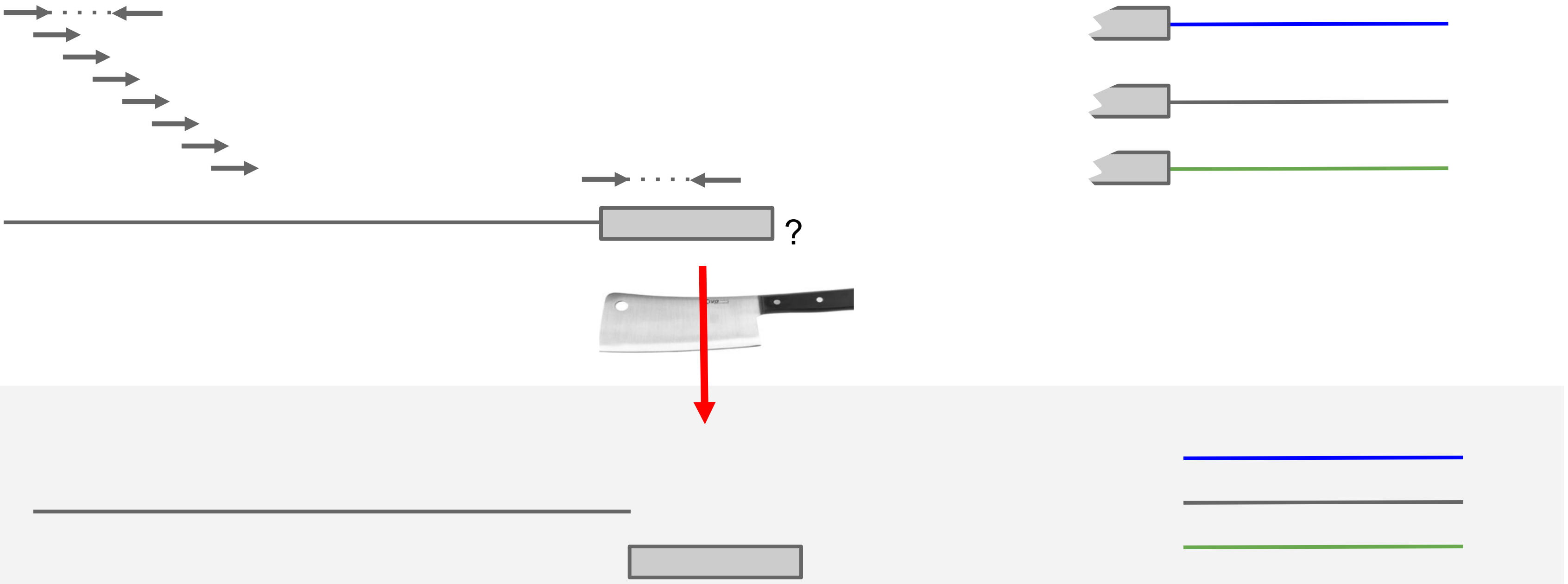
de novo Assembly

- *Reconstruct the original DNA sequences using the sequence reads alone*

- When is *de novo* assembly required?
  - new "non-model" organisms
    - no sufficiently related reference genome
  - novel DNA segments
  - novel RNA transcripts and splice variants
  - discover fusion genes
  - identify contamination

de novo Assembly

- *One contig per chromosome/plasmid*
    - not with current Illumina sequencing technology



Long Reads

- One contig per Chromosome?
  - When Reads are longer than Longest Repeated in a genome
  - Bacteria: rRNA operons, transposons: longest ~7 kb
  - PacBio, Oxford Nanopore - Reads longer than 7 kb
- Still expensive
  - extra information does not always justify the extra cost

de novo Assembly

What Purpose?

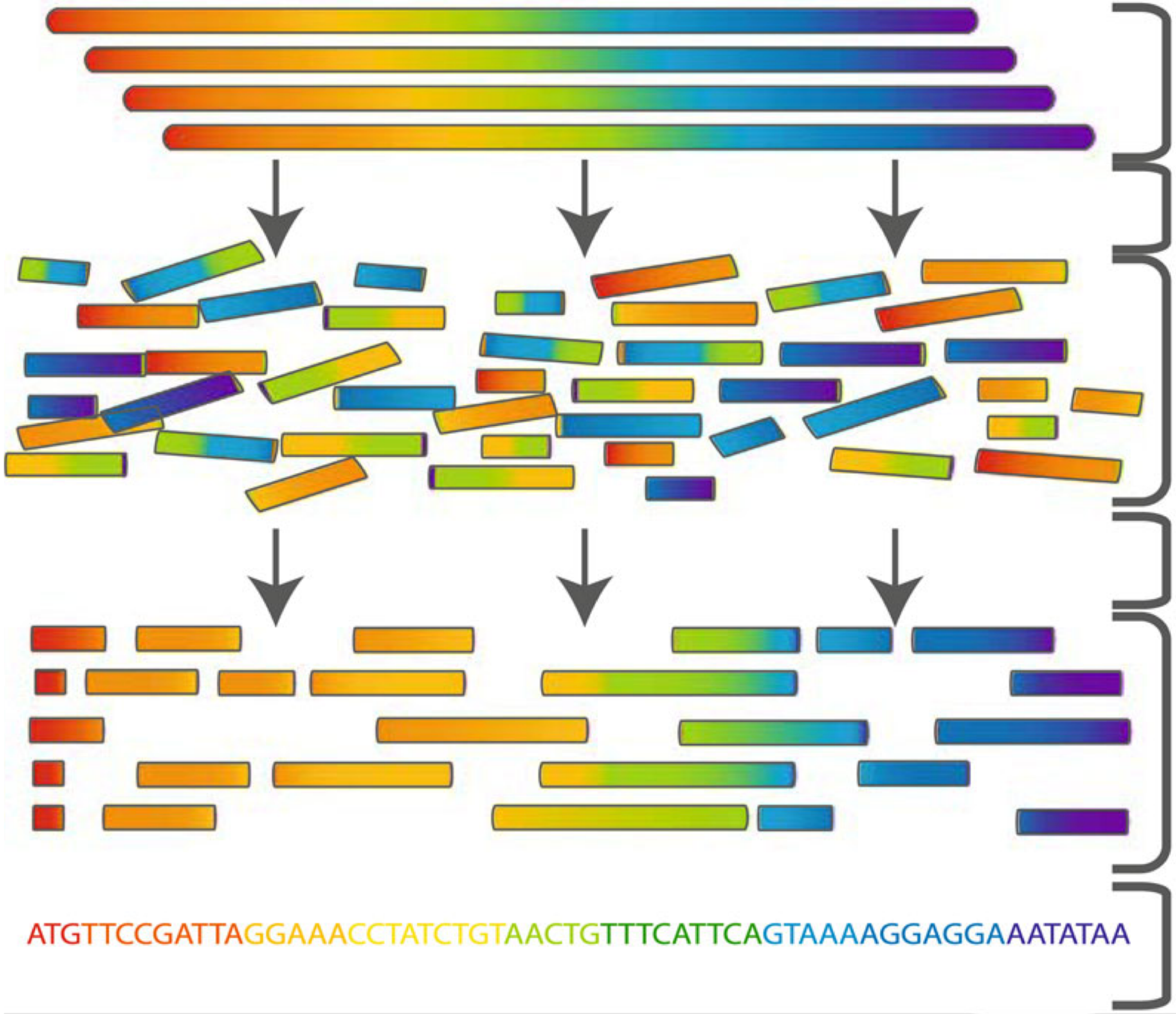
- Reconstruction of the genome sequence
  - conservative (no false joins!)

Output?

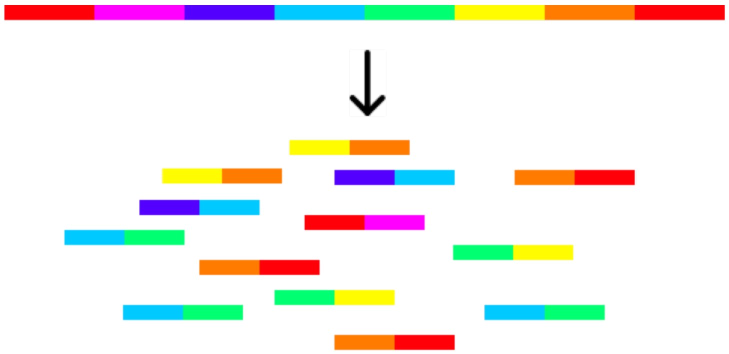
- A set of Contigs - multi-fasta file

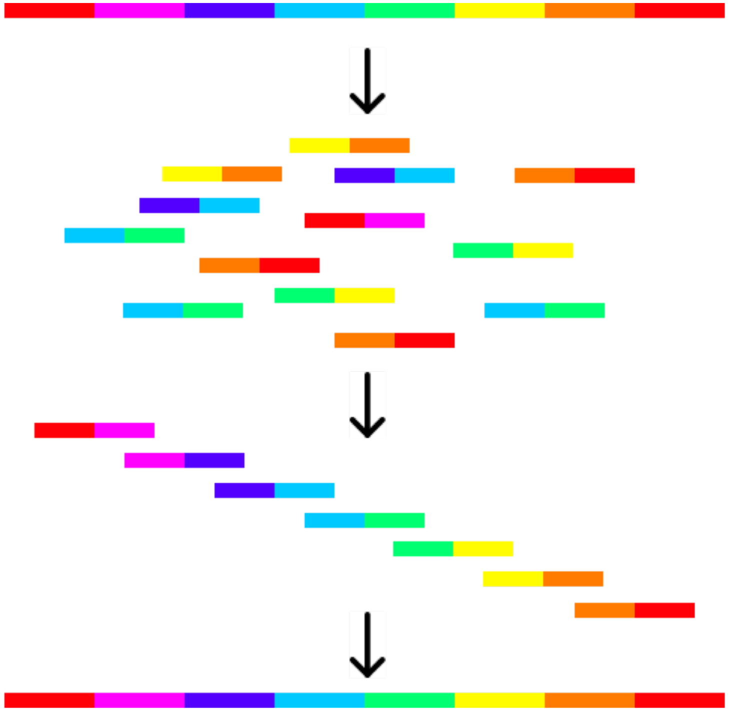
Software?

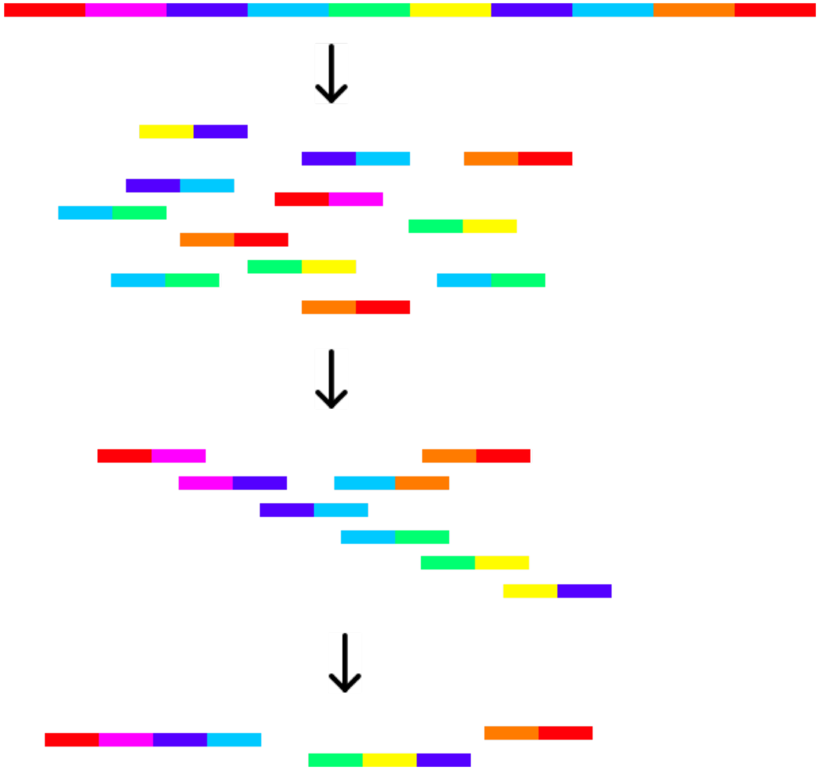
- Illumina reads: *Velvet, MegaHit* or *SPAdes Shovill Unicycler*

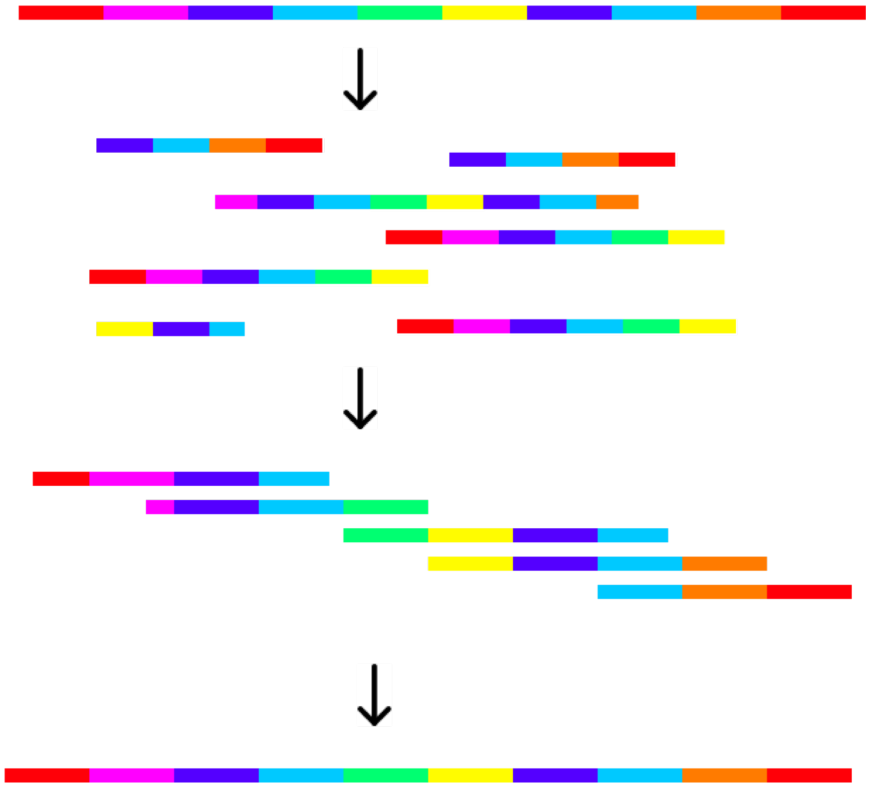


ATGTTCCGATTAGGAAACCTATCTGTAACTGTTTCATTCAGTAAAAGGAGGAAATATAA









Exercise:
Paper genome assembly

Assemble these 45 sequences

```
_ass _gen _gen  mbly acte al_g al_g al_g al_g al_g asse

asse asse bact cter cter cter cter e_as e_as e_as enom

enom eria   gejo geno ial_ ial_ ial_ me_a me_a me_a mial

ome_ ome_ ome_ ome_ me_a rial rial semb semb semb teri
```

Answer

```
bacterial_genome_assembly
```

Assessing assembly quality

- How many contigs?
- How long are the contigs?
- How correct are the contigs?