Working with Pathogen Genomes 2022

Module 5: Genome Assembly

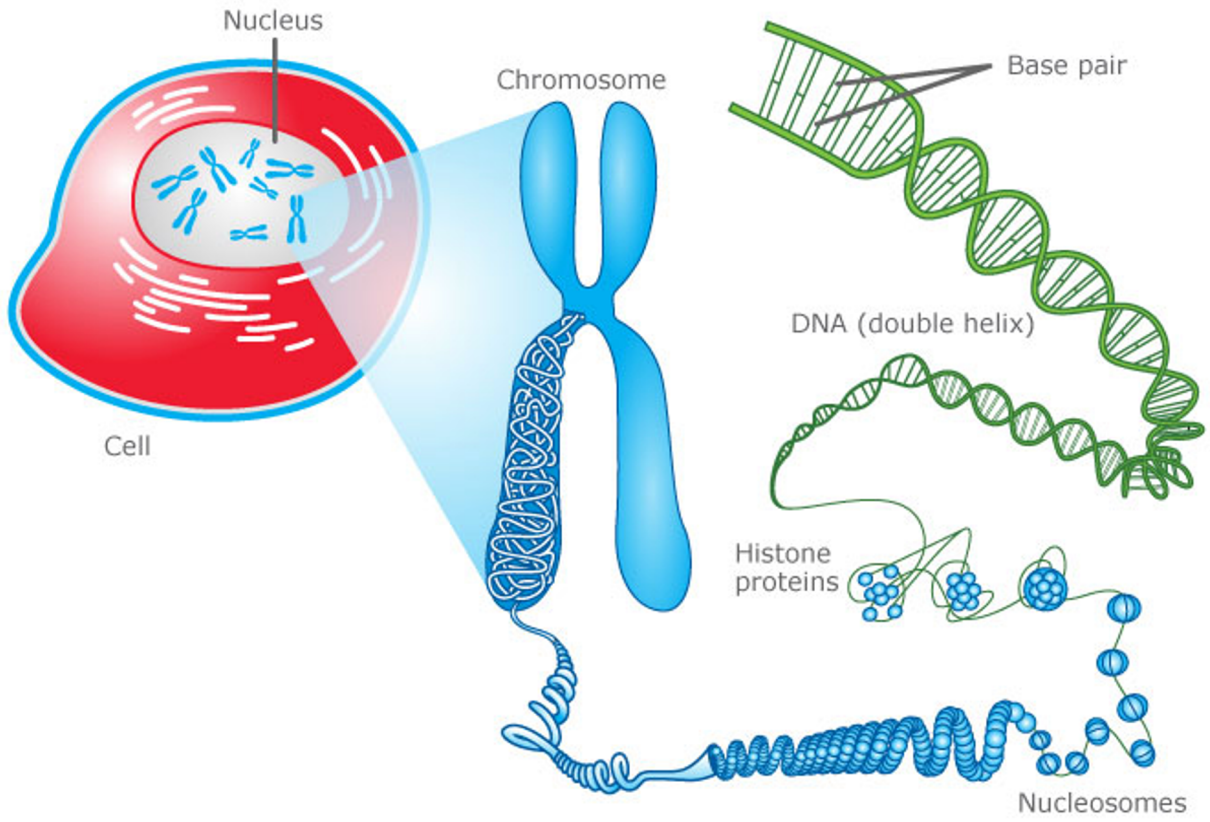
Steve Doyle: Stephen.doyle@sanger.ac.uk

Fernán Agüero: fernan@iib.unsam.edu.ar

wellcome sanger institute

wellcome connecting science

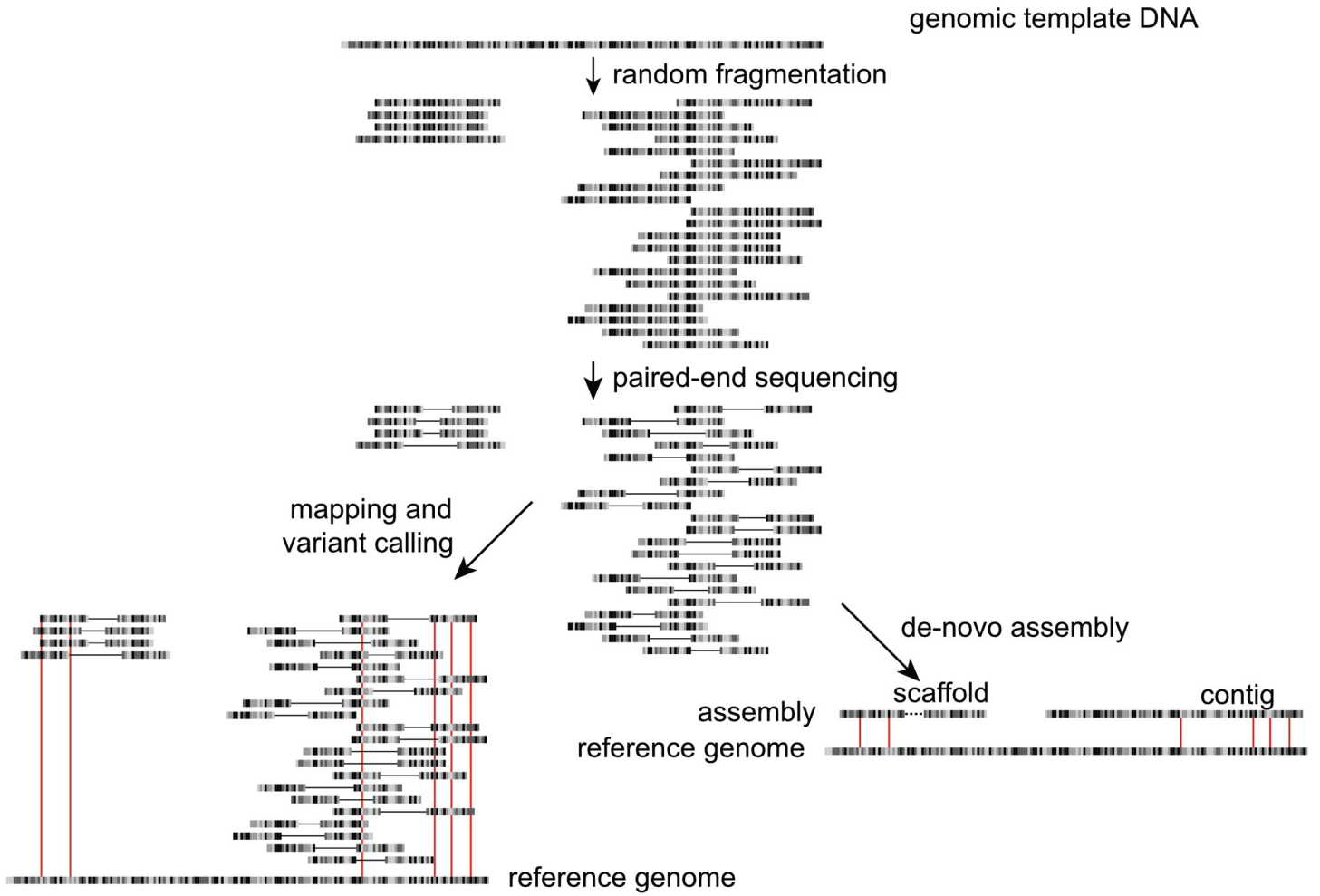
Genome sequencing is conceptually straight forward




```
CCACAAGTTCCTGACTGCCTGACTTCCCTTCACCGACTGGCACTTTCCACTCGGATCGCC
AGCGACCGTTACTAAAAAAACAAACATCGAATACTGTCTGCAAGACAGTGCAATAAAGCA
AATGAAAATAATTAGAATAAGAATAATGTTAATAATGATACCAAAAATTCTTCGGCTGGA
ACTGATGTGACTCTATGCATAATGTGAAATTTCCATGACGAACGAACACGCATCCTACAC
CAGATTTTGAGTAATGTTCCTCTATATATGCATCTAATTCCTAAGATAAATGGGTGTGAG
CAGCAACTAGAATTGGAAAGAACCACTGAACAGCTTGGTACCTTTTGCAGAGCTACACGC
CACGATTCTAAAGCGCTCGATTCTGCTGGATACGGCTGGATGTCACTGGCCTCCTTCCGA
AGATGATCATACTTGAAGGATTTCTTCGACGCATCATCCACGTAGCGACCTCCTCCAATA
TCGTTGAATGGCGTCACAAGTGTCTAAGCGACATTTTTTGAAAGTGTAAGCAATACTTGA
TCAACTTCTTTTCCGAATCCTCGATGAATAATCATTAGAACAATACGTTTCCATTTTTTA
CTGCACCCTTTTCCATACAAATAGTTACAAGTTAGATTGGACAATGACAACAAATGAACA
CCGAGTTCTATGGTGAGAAAACATGCACTAGGGAATCGACCGCCTTGTGCAGCAGCATTC
ATGGTGAAAAGACAAGATCTGTCATAGATGAGTTTAAGTTTGAACAGTTCCCCTTCAAAT
TCCACACTATCACAGACCATTCCCCGAAGAATGTTCGACCTCTAGACCTGACCTCTACGA
```

https://www.sciencelearn.org.nz/images/198-cell-chromosomes-and-dna

Sequencing genomes is easy…!



wellcome sanger institute

wellcome connecting science

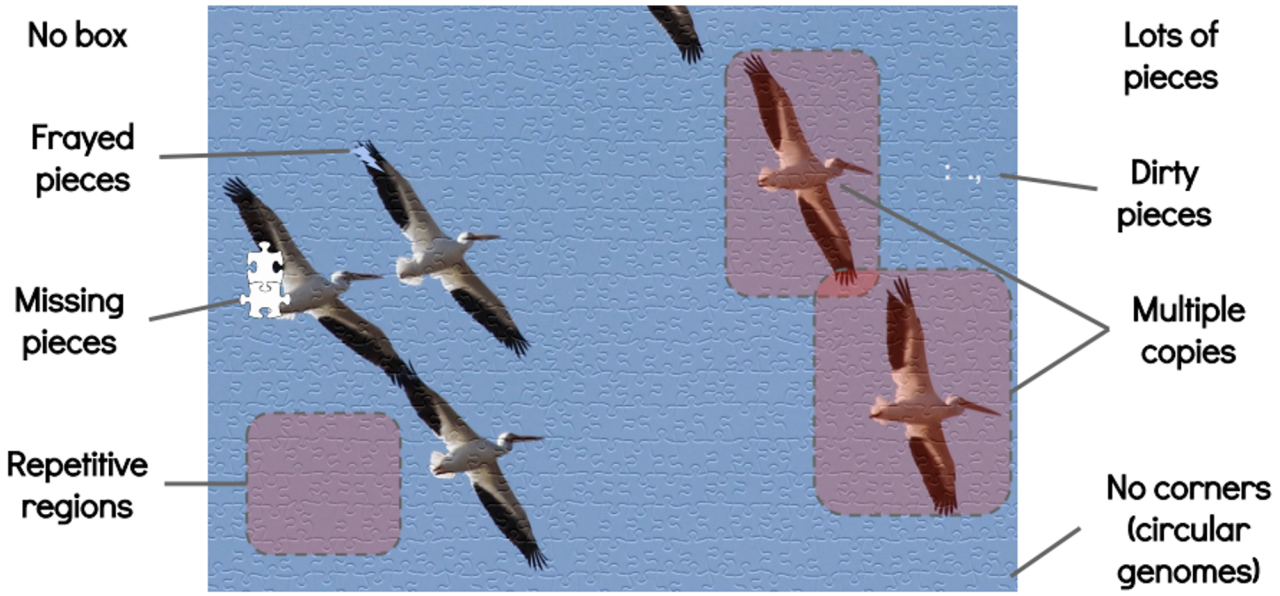
Sequencing genomes is easy, constructing good genomes is not

- **Genome: *biologically***
  - “the haploid set of chromosomes in a gamete or microorganism, or in each cell of a multicellular organism”
  - “the complete set of genes or genetic material present in a cell or organism”

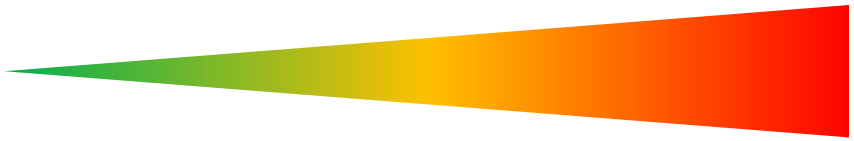
Adapted from Torsten Seemann presentation: “De novo genome assembly”
https://www.slideshare.net/torstenseemann/de-novo-genome-assembly-tseemann-imb-winter-school-2016-brisbane-au-4-july-2016

Sequencing genomes is easy, constructing good genomes is not

- **Genome: *bioinformatically***
  - Best guess, but often:
    - highly fragmented
    - misassembled to some degree
    - Haplotypic
    - contaminated
    - duplicated or missing



Draft genomes
"manageable"(?)



Chromosome-scale genomes
HARD

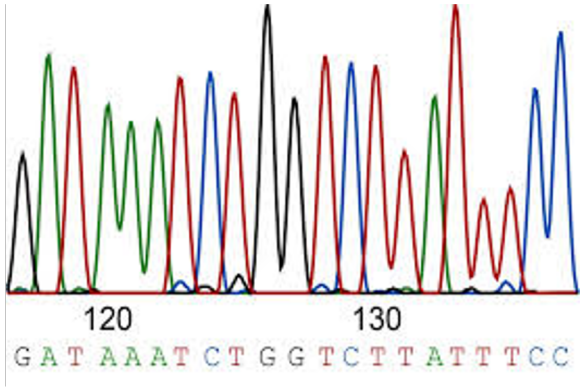
Time, money, expertise

wellcome sanger institute

wellcome connecting science

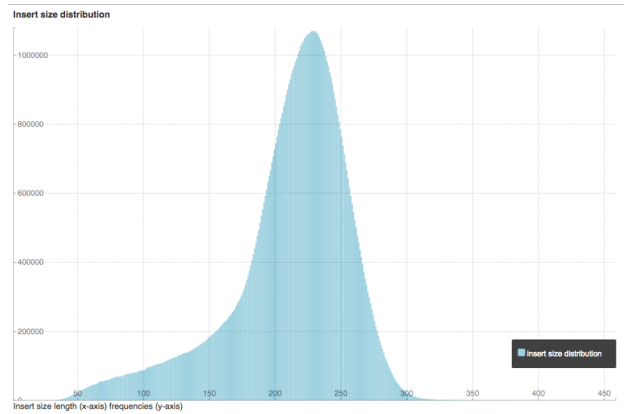
New technologies are making genome assembly easier

Sanger Sequecing: ABi



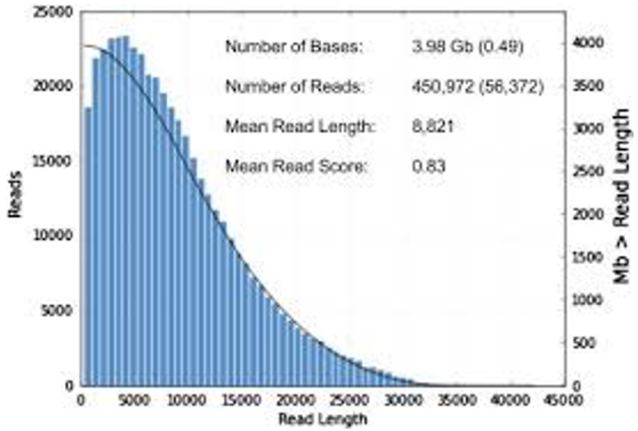
Read length:500-1000 bp

High Throughput Sequencing: Illumina



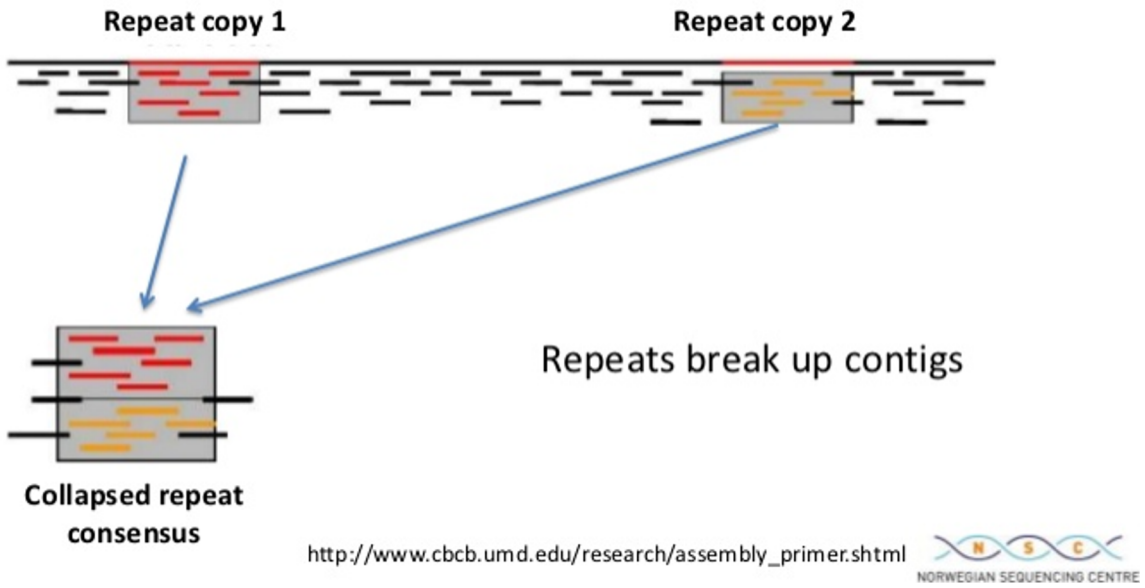
Read lengths: 100-300 bp
Insert lengths: ave 300-500 bp

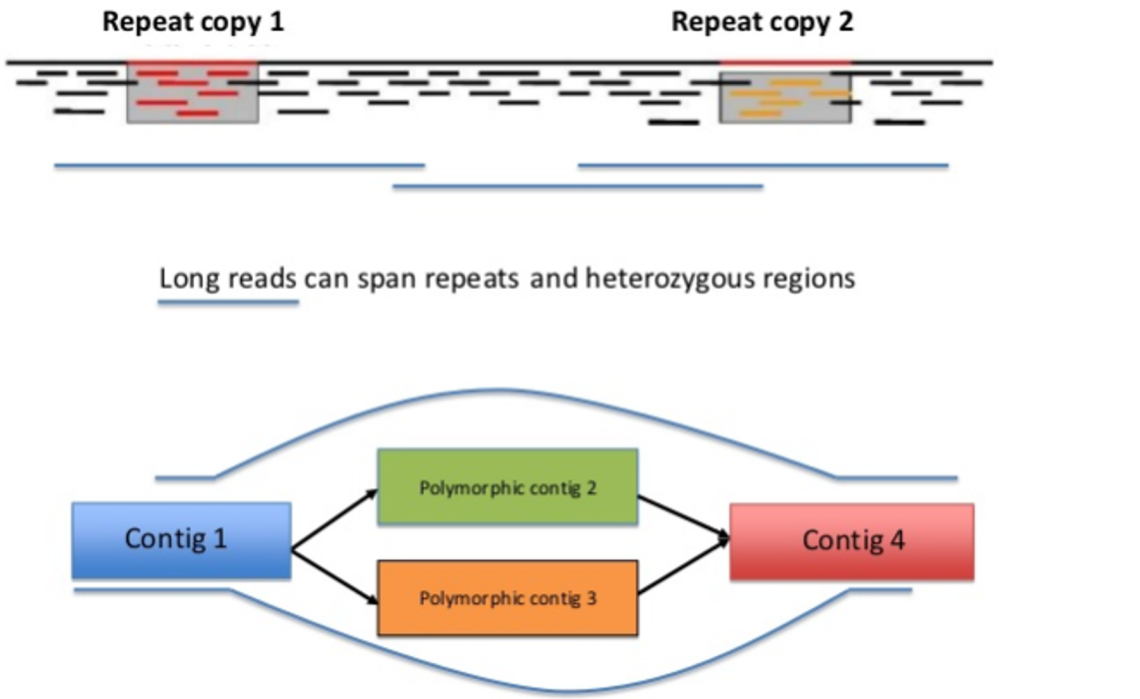
Long read sequencing: Pacbio & Nanopore



Read lengths: 5-10 kb
- Pacbio: up to 60 kb
- Nanopore: up to 1Mb

Repeats / polymorphic loci can break genomes

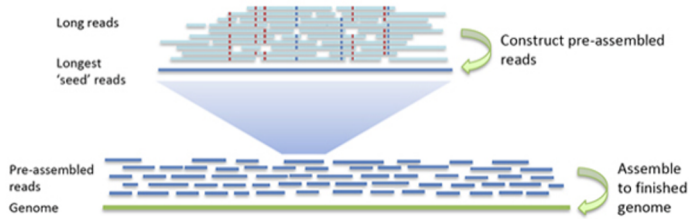




https://www.slideshare.net/flxlex/combining-pacbio-with-short-read-technology-for-improved-de-novo-genome-assembly

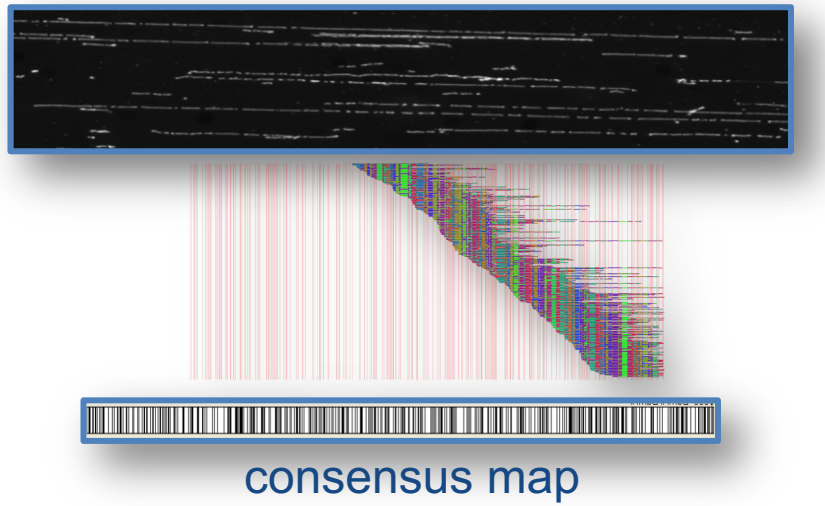
Long read / range sequencing is key to good genomes

Pacific Biosciences (PacBio)



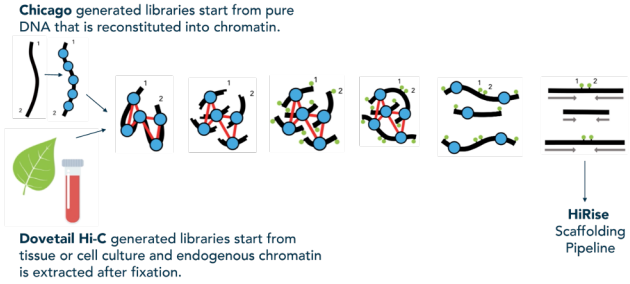
Oxford Nanopore

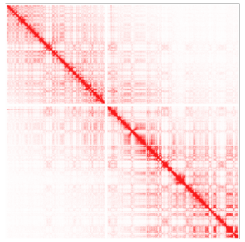
Optical Mapping (OpGen, Bionano genomics)



Linked reads (10X Genomics)

Chromosome confirmation capture, ie Hi-C (Dovetail Genomics)

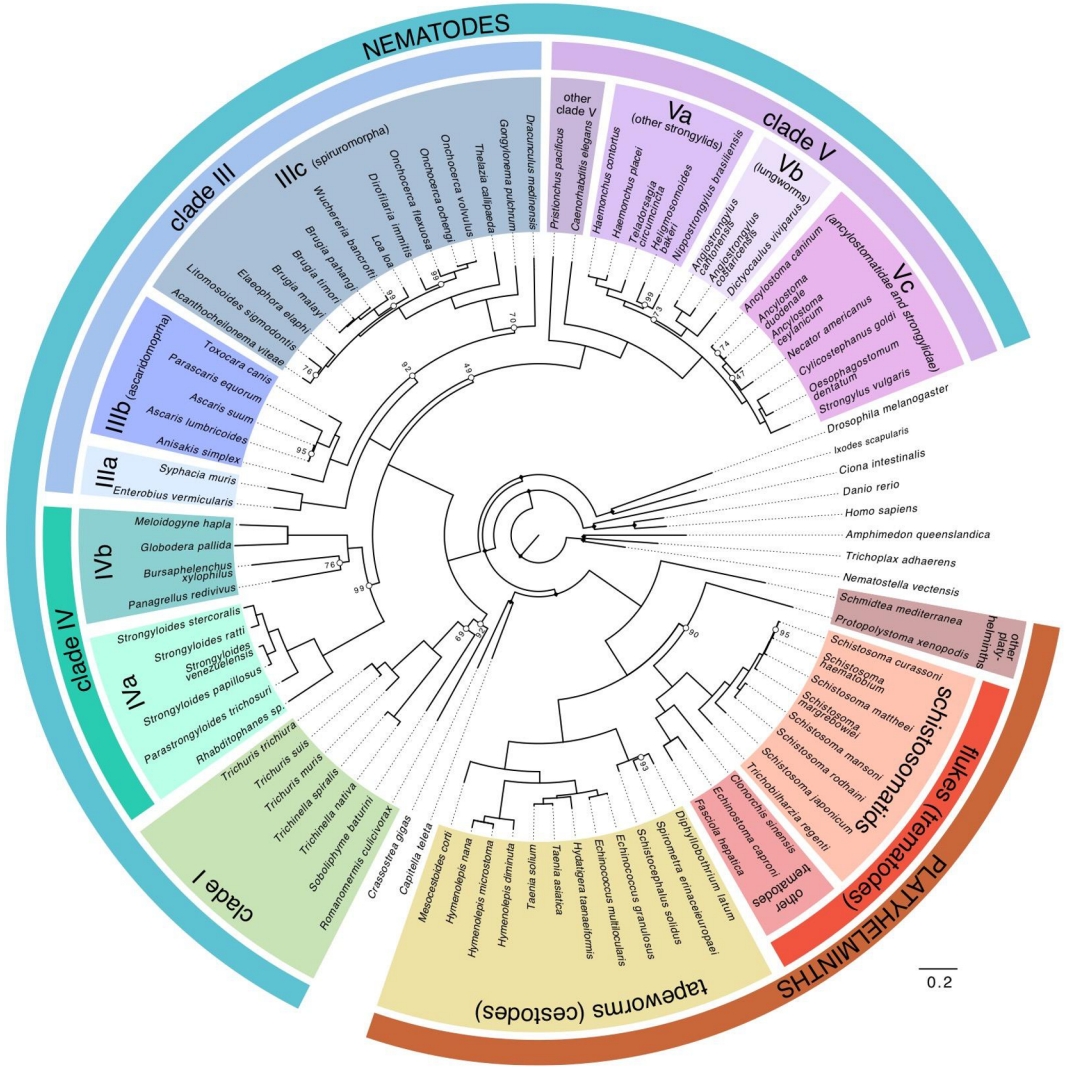




wellcome sanger institute

wellcome connecting science

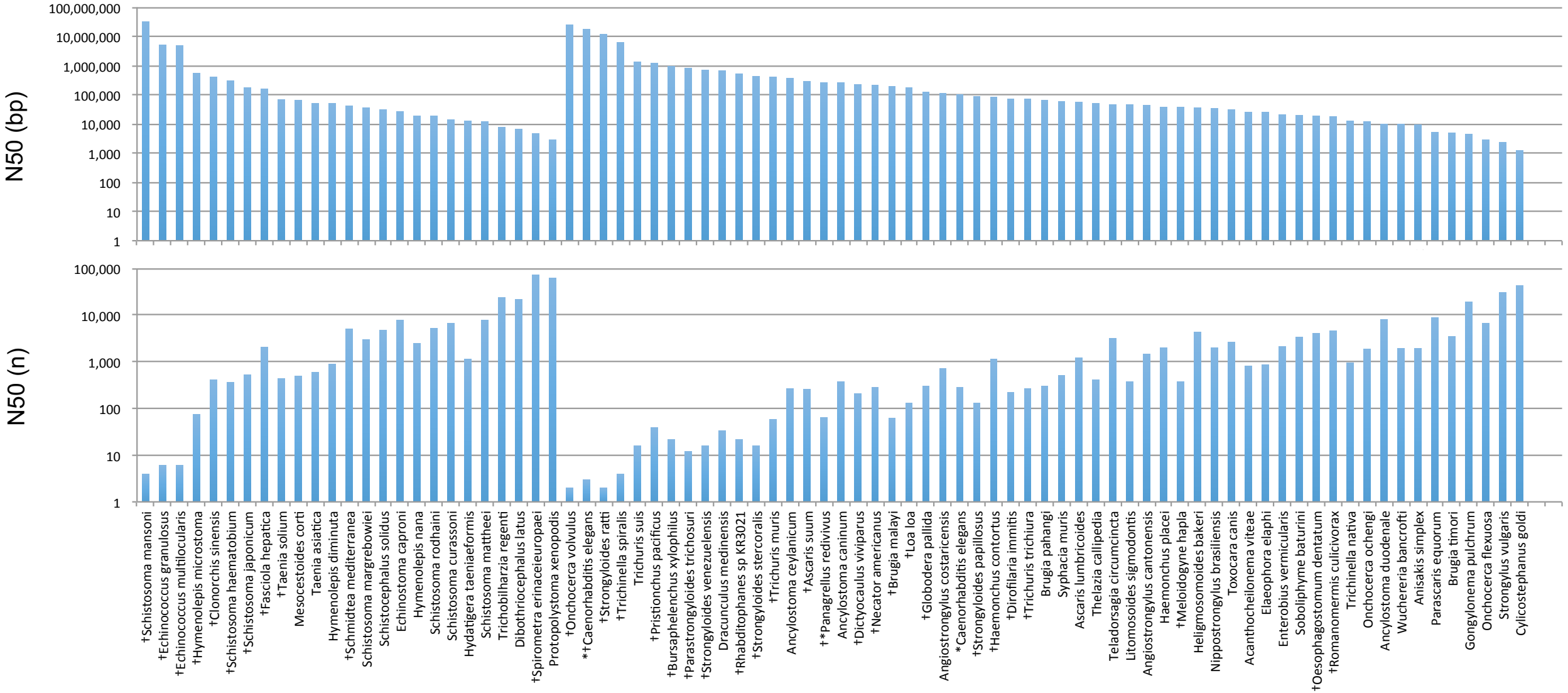
Parasite Genomics @ Sanger: 50 Helminth Genome Project



- Aims:
  - generate ***draft*** genome assemblies for (a) clinically and veterinary important organisms and (b) parasitic groups lacking exemplars in current genome projects and (c) comparators to ‘reference’ species
  - (Try to) ensure similar sequencing, assembly and annotation approaches for each genome so they are truly **comparable**

wellcome connecting science

Not all helminths, nor their assemblies, are created equal



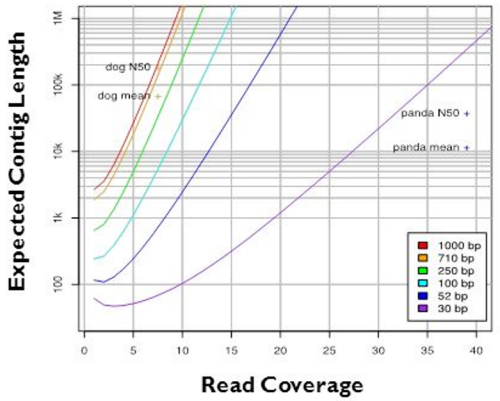
N50: measure of contiguity. The minimum sequence length for which 50% of the genome is in sequences at least this length

Chromosomal scale helminth assemblies

	Samples / biology	Sequencing	Stage	Genome size (Mb)	Fragments	N50 (Mb)	N50(n)
Onchocerca volvulus	Single worm	SR, optical map	final	97	715	25.5	2
Trichuris muris	Pooled worms, inbred lab strain	SR, Pacbio, optical map	PB only	123	3708	0.140	193
			final	112	803	28.9	2
Trichuris trichura	Single worm (SR), 3 males (pacbio)	SR, Pacbio	PB only	97	1344	0.257	98
			final	80	113	11.3	2
Hymenolepis microstoma	Inbred, clonal	Pacbio, optical map	PB only	161	288	4.6	10
			final	164	27	21	3
Schistosoma mansoni	Clonal (Pacbio), single worms (SR)	SR, Pacbio, genetic map	PB only	409	1598	1.05	97
			final	409	320	50.4	3
Haemonchus contortus	Pooled (PB), single worm (SR), semi-inbred, haplotypic	SR, PacBio, optical map	PB only	487	5284	0.184	563
			final	279	8	47	3

Recipe for a good genome assembly

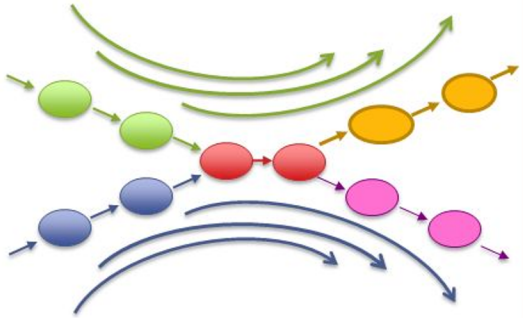
Coverage



High coverage is required

- Oversample the genome to ensure every base is sequenced with long overlaps between reads
- Biased coverage will also fragment assembly

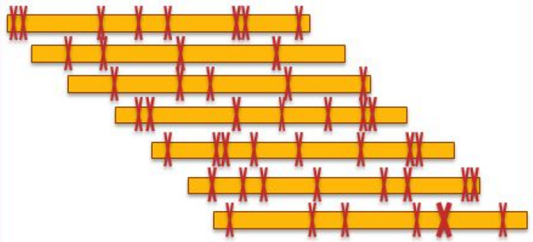
Read Length



Reads & mates must be longer than the repeats

- Short reads will have ***false overlaps*** forming hairball assembly graphs
- With long enough reads, assemble entire chromosomes into contigs

Quality



Errors obscure overlaps

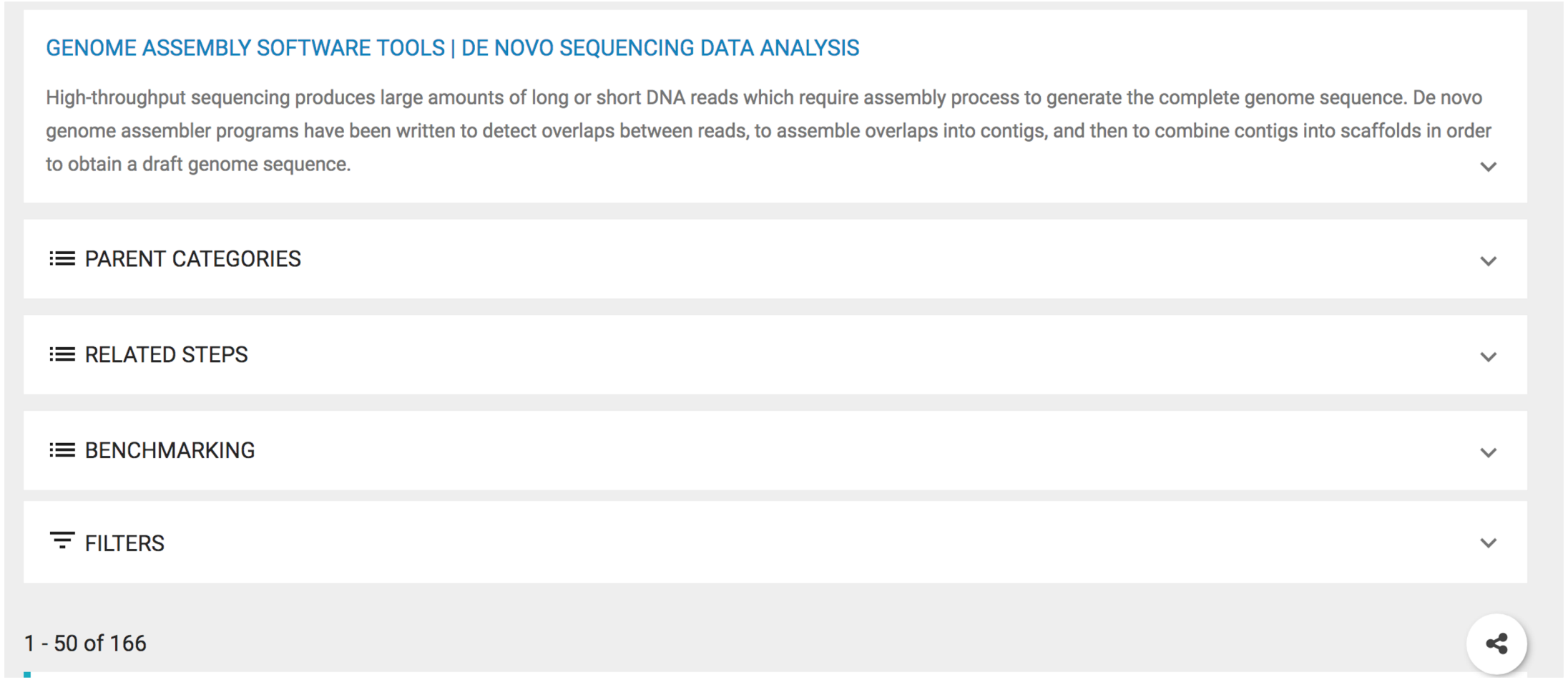
- Reads are assembled by finding kmers shared in pair of reads
- High error rate requires very short seeds, increasing complexity and forming assembly hairballs

Current challenges in *de novo* plant genome sequencing and assembly
Schatz MC, Witkowski, McCombie, WR (2012) *Genome Biology*. 12:243

wellcome sanger institute

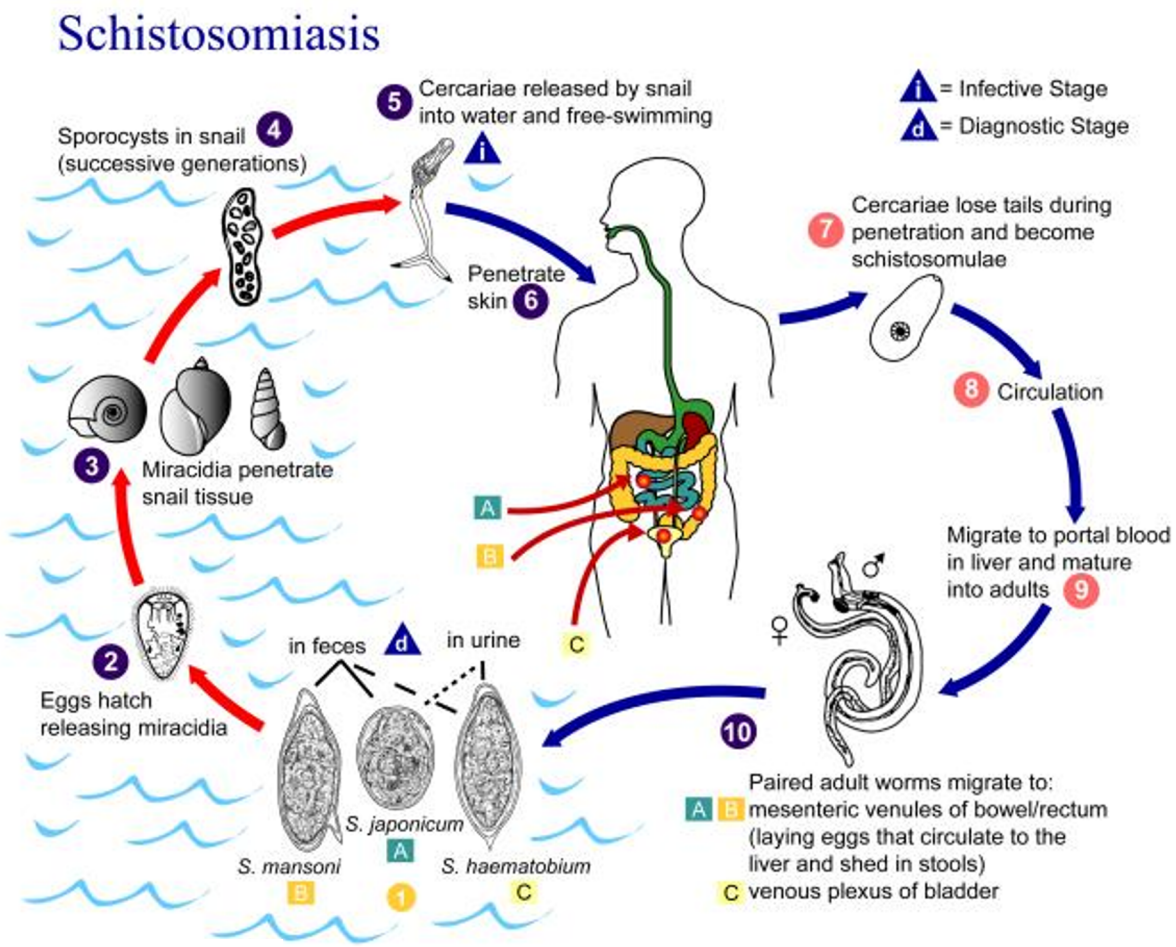
wellcome connecting science

What tool(s) should I use?



https://omictools.com/genome-assembly-category

Today: Assembly of a *Schistosoma mansoni* chromosome



- Genome
  - 7 autosomes + Z/W sex chromosomes
  - approximately 380 Mb
- We will work with chromosome IV
  - ~40 Mb

wellcome sanger institute

wellcome connecting science

Aims and workflow

Step 1: Checking raw sequencing data before assembly

Step 2: Estimating your genome size from raw sequence data

Step 3: Exploring different genome assembly using either Illumina short read or Pacbio long read data

Step 4: Comparison of your assemblies against a known reference sequence

Step 5: Further exploration of your genome assemblies

NOTE: Genome assembly is memory intensive!!!

- Page 12: Unfortunately, the computers were are working on are unlikely to finish the minimap assembly.
- You can **skip this step**, and move on to using assembly-stats to compare the pre-prepared assemblies.