

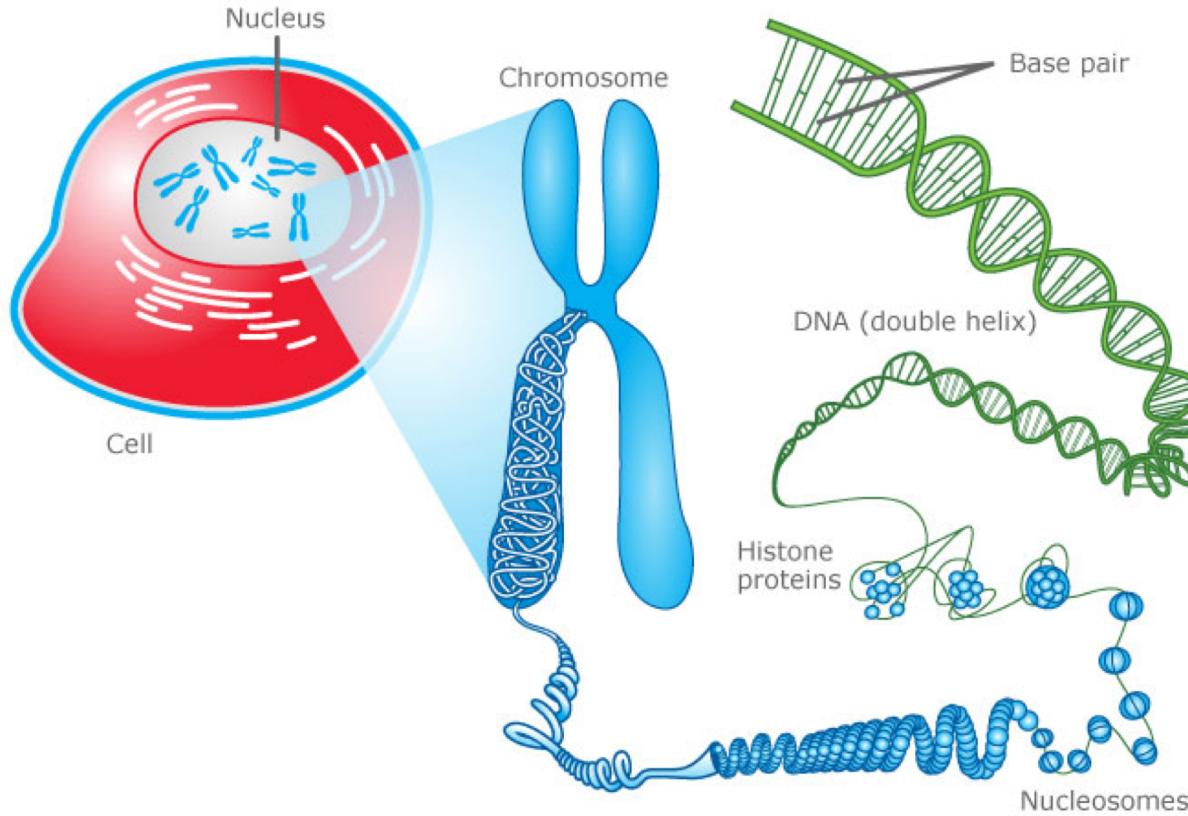
# **Module 5: Genome Assembly**

**Steve Doyle**

Wellcome Sanger Institute  
[stephen.doyle@sanger.ac.uk](mailto:stephen.doyle@sanger.ac.uk)

**Working with Pathogen Genomes 10-14<sup>th</sup> May 2021**

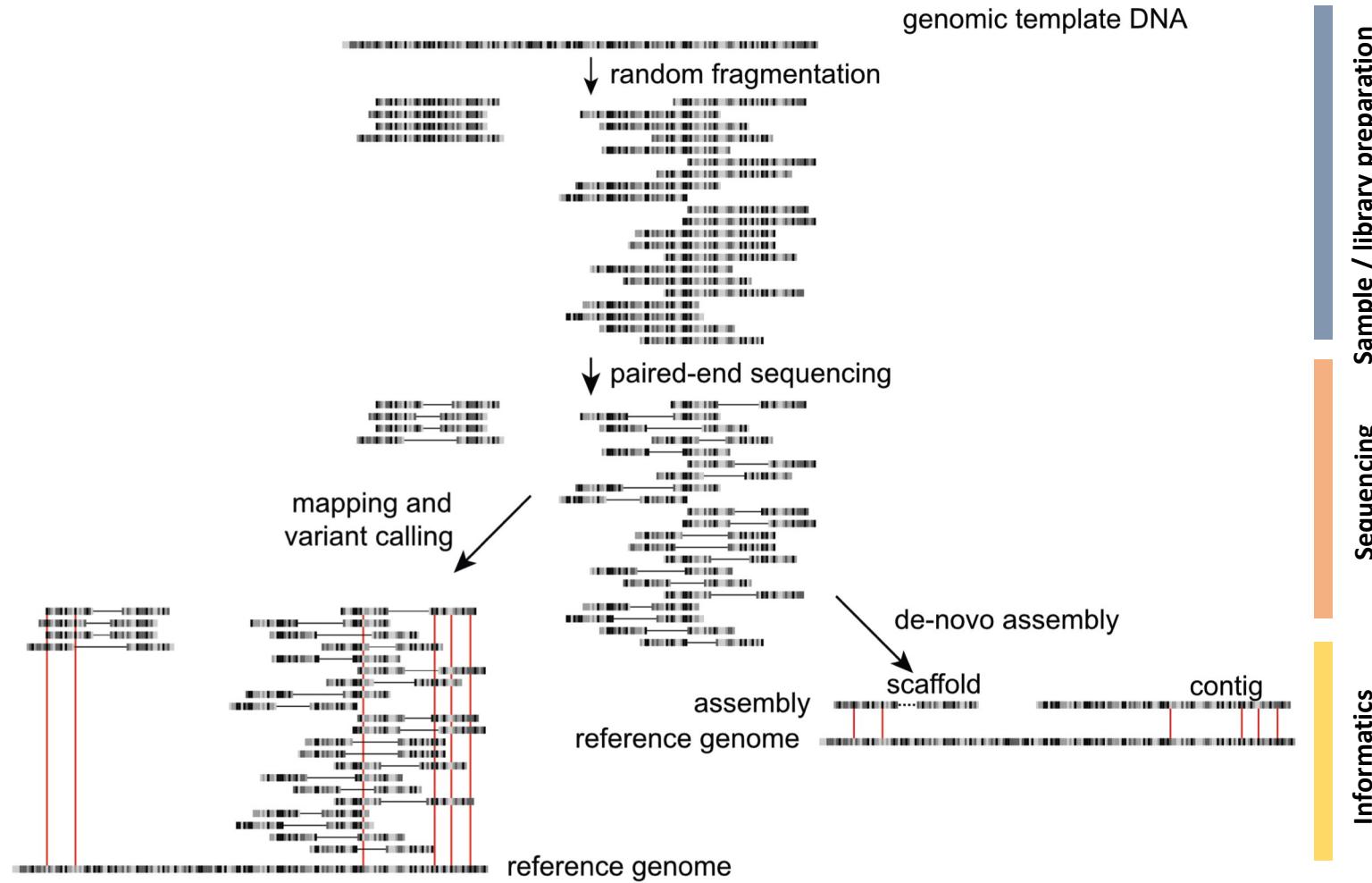
# Genome sequencing is conceptually straight forward



© 2007-2011 The University of Waikato | www.sciencelearn.org.nz

```
CCACAAGTCCTGACTGCCTGACTTCCCTCACCGACTGGCACTTCCACTCGGATGCC  
AGCAGCCGTTACTAAAAAAACAAACATCGAATACTGTCTGCAAGACAGTCGAATAAAGCA  
AATGAAAATAATTAGAATAAGAATAATGTTAATAATGATAACAAAAATTCTCGGCTGGA  
ACTGATGTGACTCTATGCATAATGTGAAATTCCATGACGAACAGCATCCTACAC  
CAGATTTGAGTAATGTTCTCTATATGCATCTAATTCTAAGATAAATGGGTGTGAG  
CAGCAACTAGAATTGGAAGAACACTGAACAGCTGGTACCTTGAGAGCTACACGC  
CACGATTCTAACCGCTCGATTCTGCTGGATACGGCTGGATGTCACTGGCCTCCTCCGA  
AGATGATCATACTGAAGGAGTTCTTGACGCATCCACGTAGCGACCTCCTCCAATA  
TCGTTGAATGGCGTCACAAGTGTCAAGCGACATTGGAAAGTGTAAAGCAATACTTGA  
TCAACTTCTTTCCGAATCTCGATGAATAATCATTAGAACATACTGTTCCATTNTTA  
CTGCACCCTTCCATACAAATAGTTACAAGTTAGATTGGACAATGACAACAAATGAACA  
CCGAGTTCTATGGTAGAGAAACATGCACTAGGGAATCGACCGCCTGTGCAGCAGCATT  
ATGGTAAAAGACAAGATCTGTCAAGATGAGTTAAGTTAACAGTCCCCCTCAAAT  
TCCACACTATCACAGACCATTCCCCGAAGAATGTTGACCTCTAGACCTGACCTCTACGA
```

# Sequencing genomes is easy...!



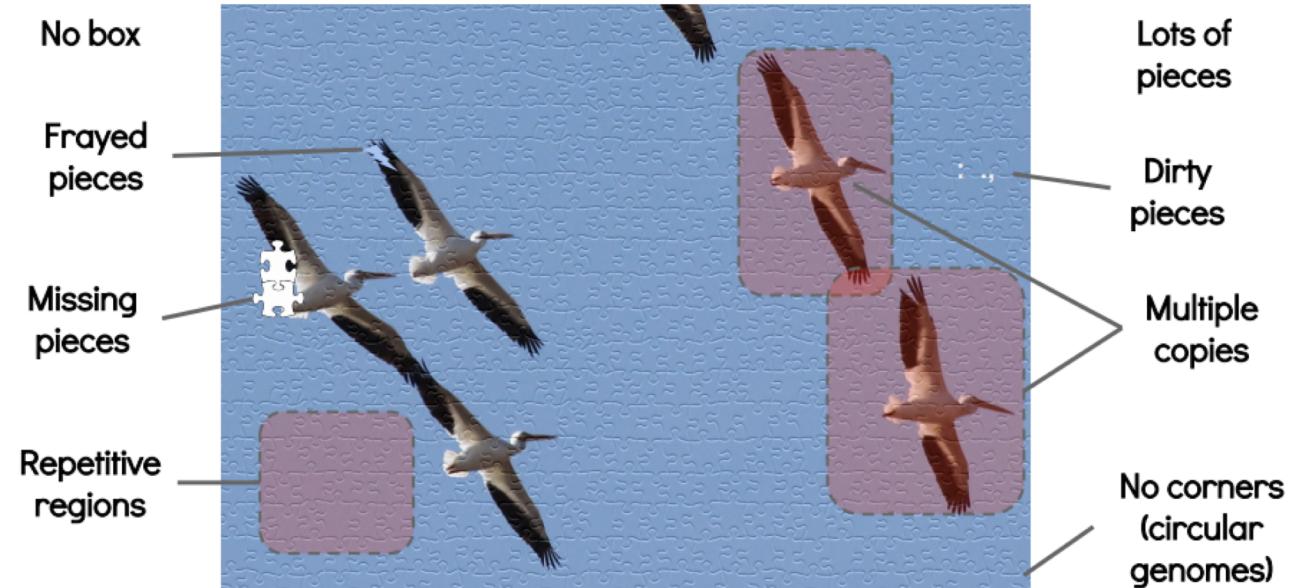
# Sequencing genomes is easy, constructing good genomes is not

- **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”

# 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**

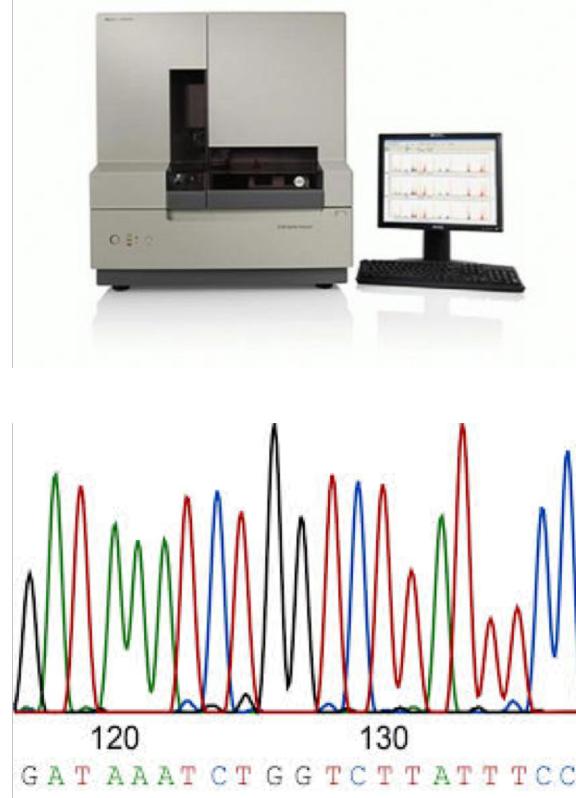
Time, money, expertise

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>

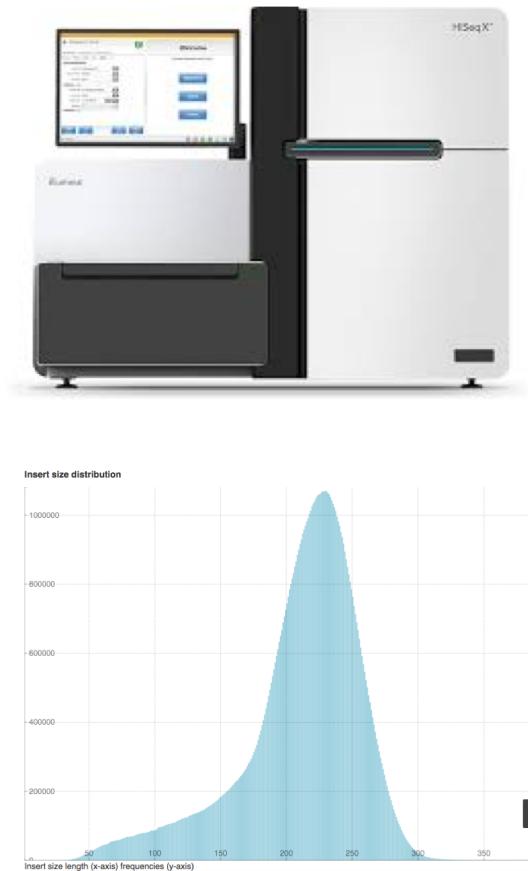
# New technologies are making genome assembly easier

Sanger Sequencing: ABI

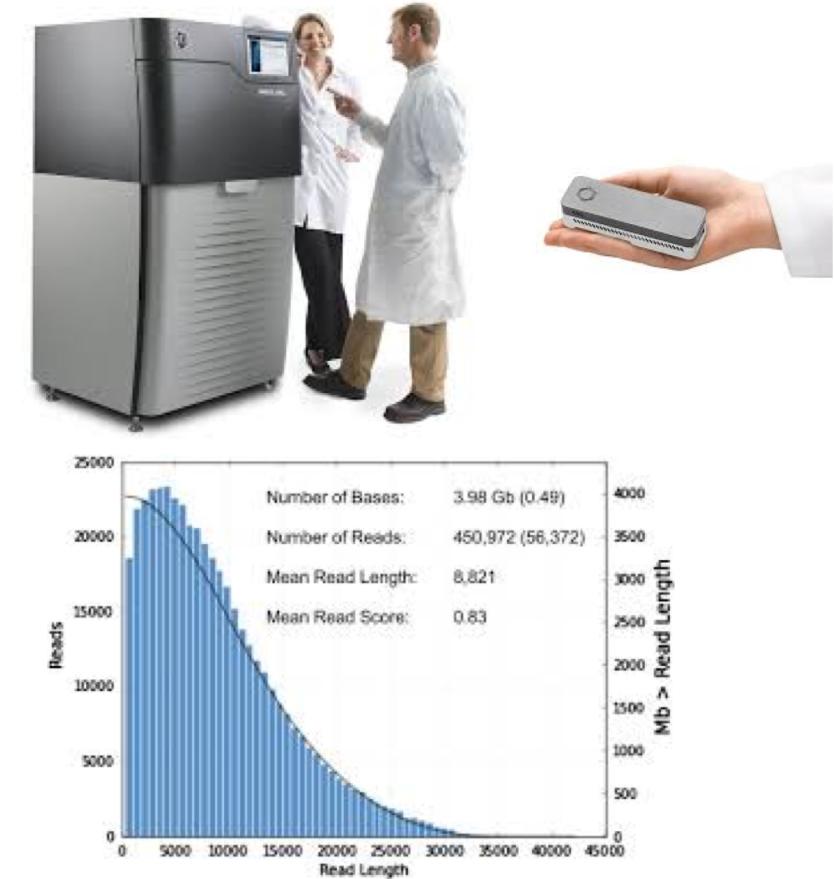


Read length: 500-1000 bp

High Throughput Sequencing: Illumina

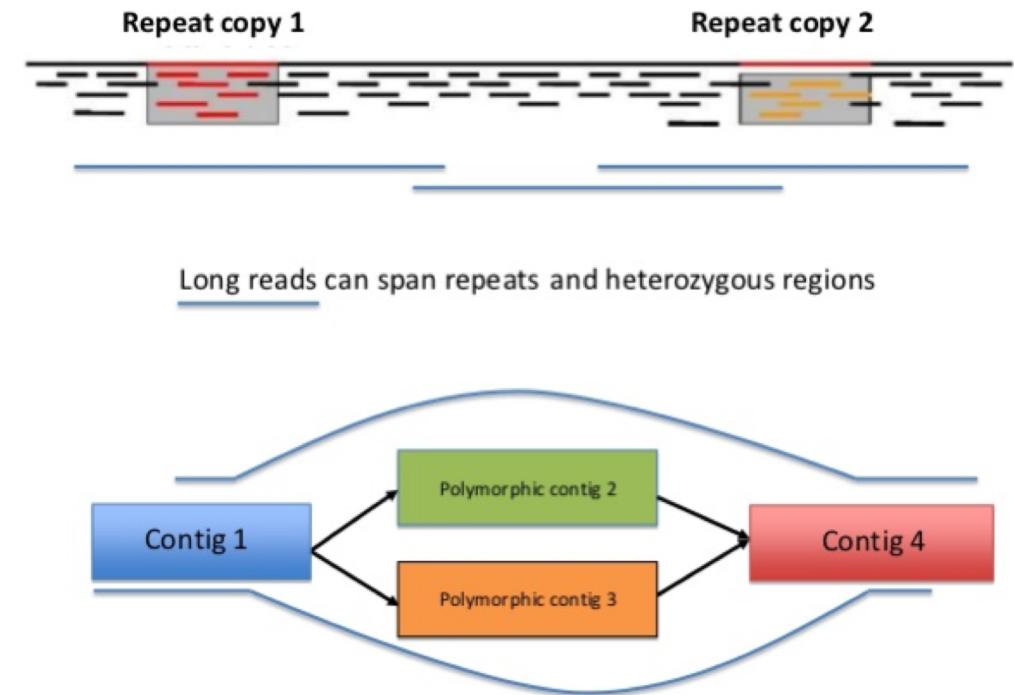
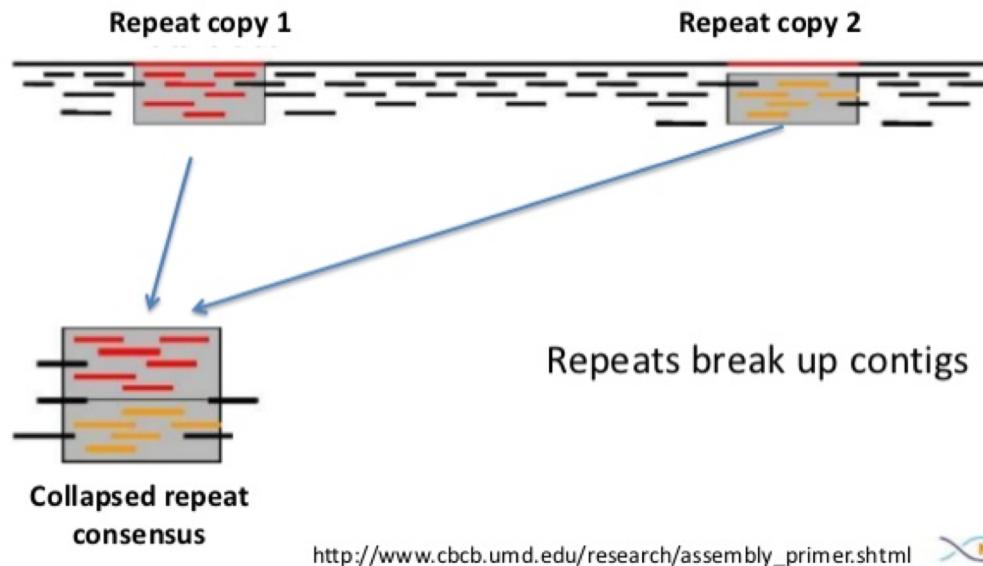


Read lengths: 100-300 bp  
Insert lengths: ave 300-500 bp



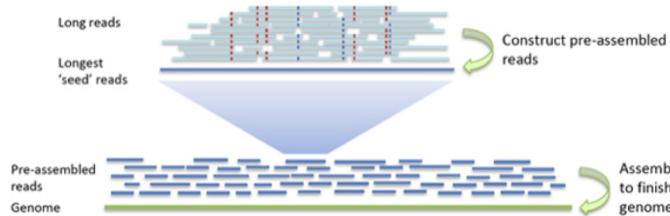
Read lengths: 5-10 kb  
- Pacbio: up to 60 kb  
- Nanopore: up to 1Mb

# Repeats / polymorphic loci can break genomes



# Long read / range sequencing is key to good genomes

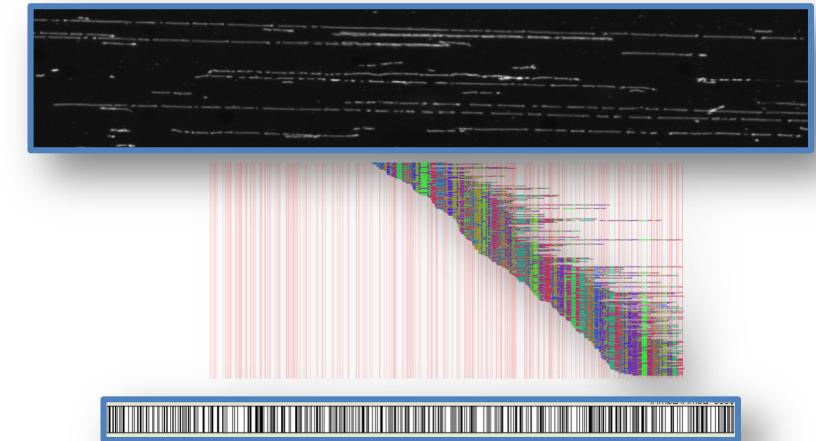
Pacific Biosciences (PacBio)



Oxford Nanopore



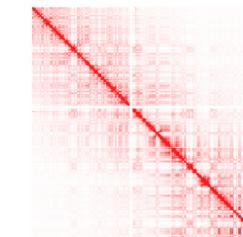
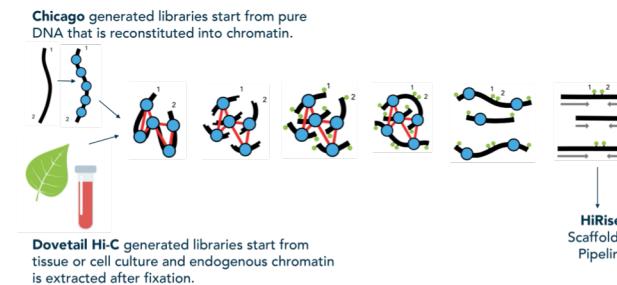
Optical Mapping (OpGen, Bionano genomics)



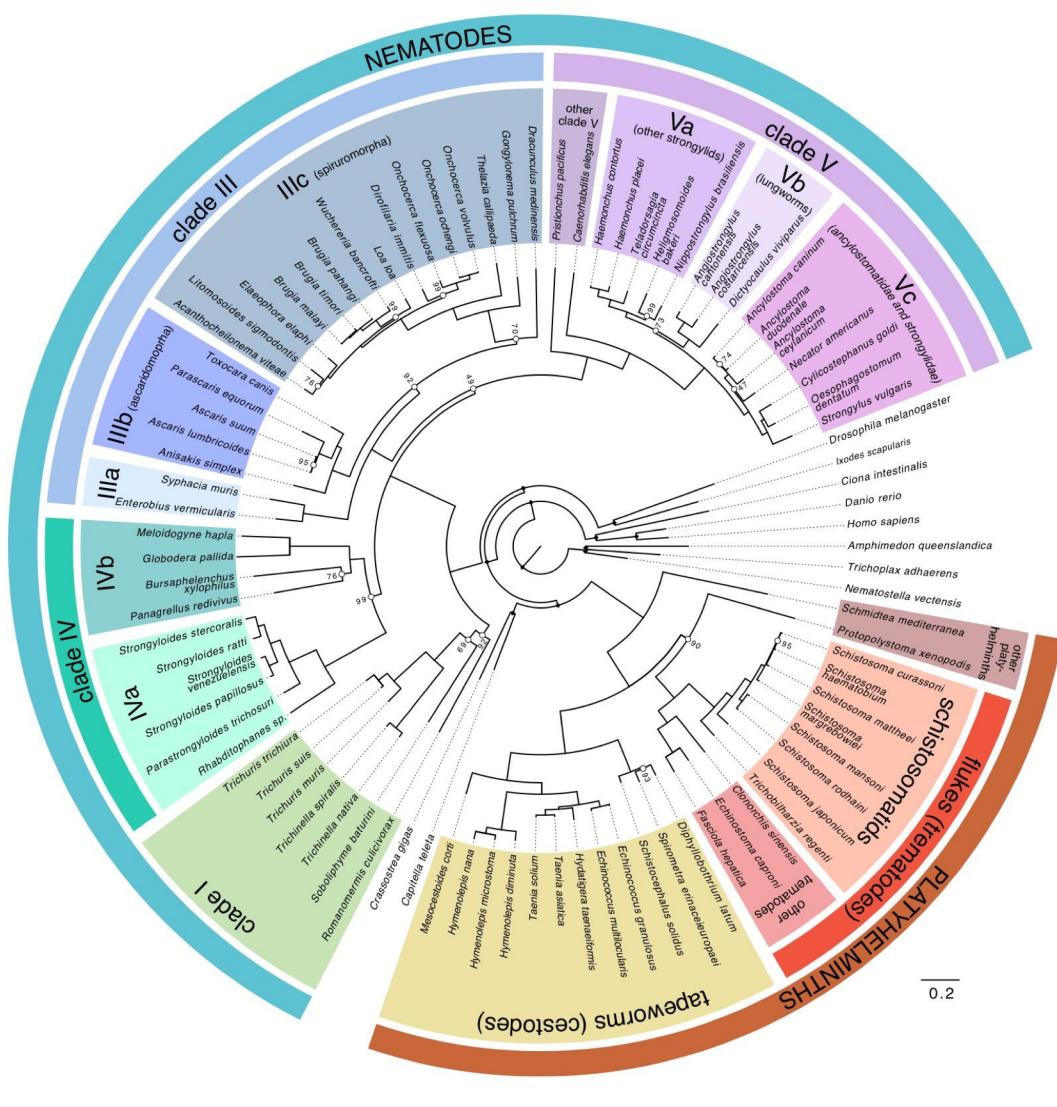
Linked reads (10X Genomics)



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



# Parasite Genomics @ Sanger: 50 Helminth Genome Project

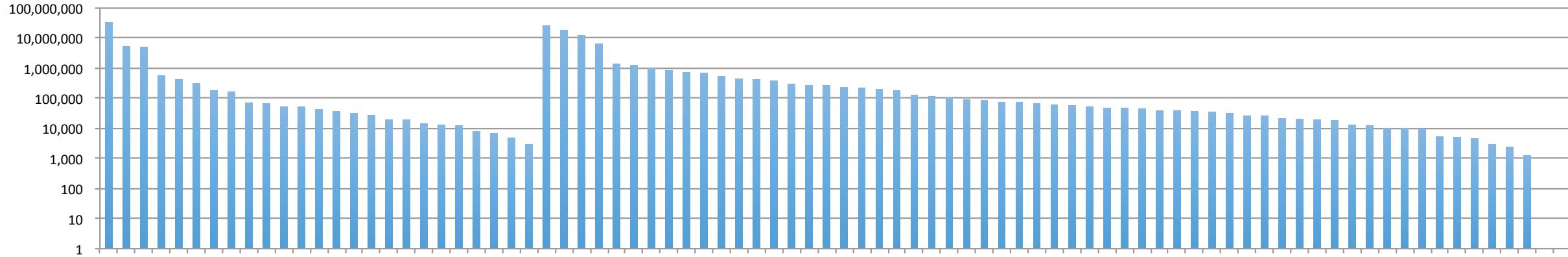


- Aims:

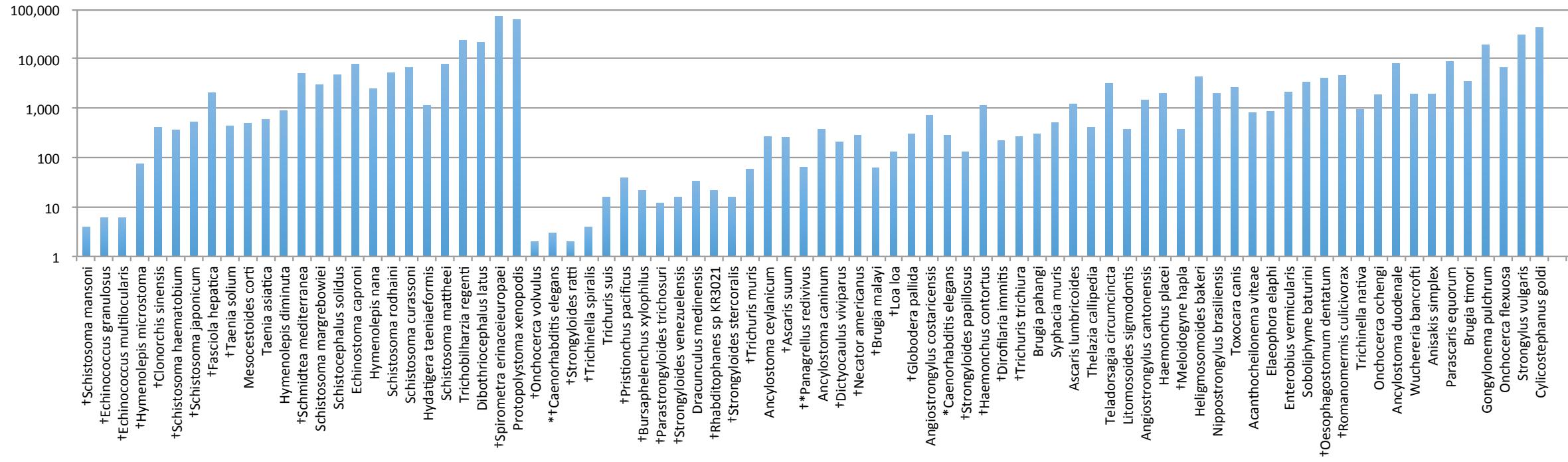
- generate **draft** genome assemblies for (a) clinically and veterinary important organisms and (b) parasitic groups lacking exemplars in current genome projects and (c) comparators to 'reference' species
- (Try to) ensure similar sequencing, assembly and annotation approaches for each genome so they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

# Not all helminths, nor their assemblies, are created equal

N50 (bp)



N50 (n)

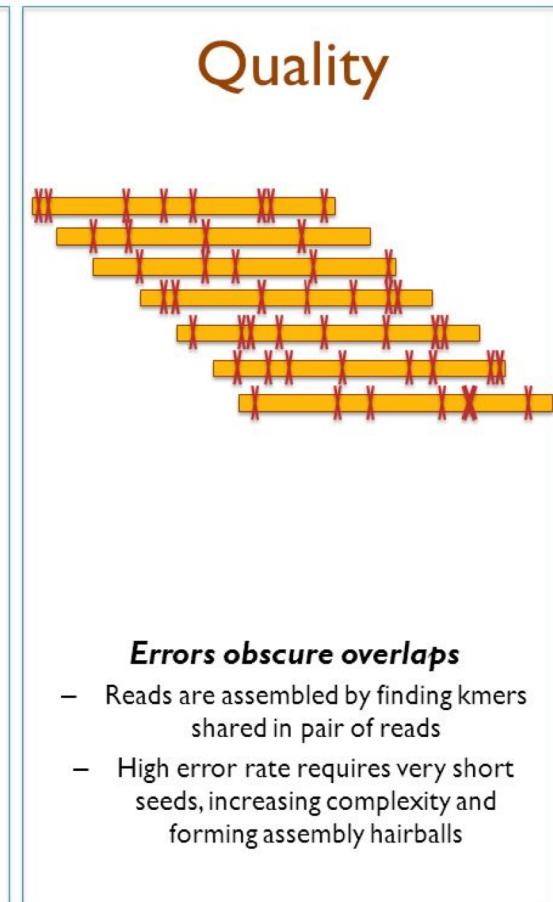
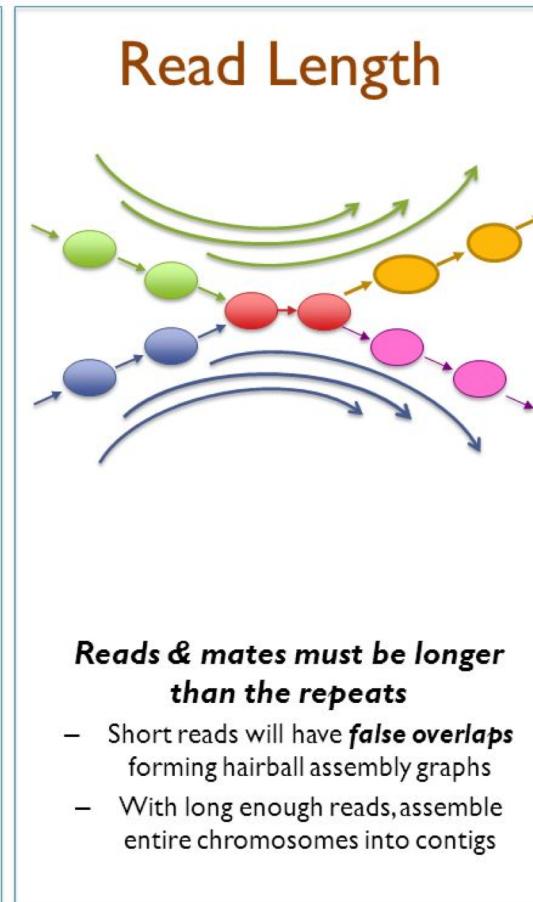
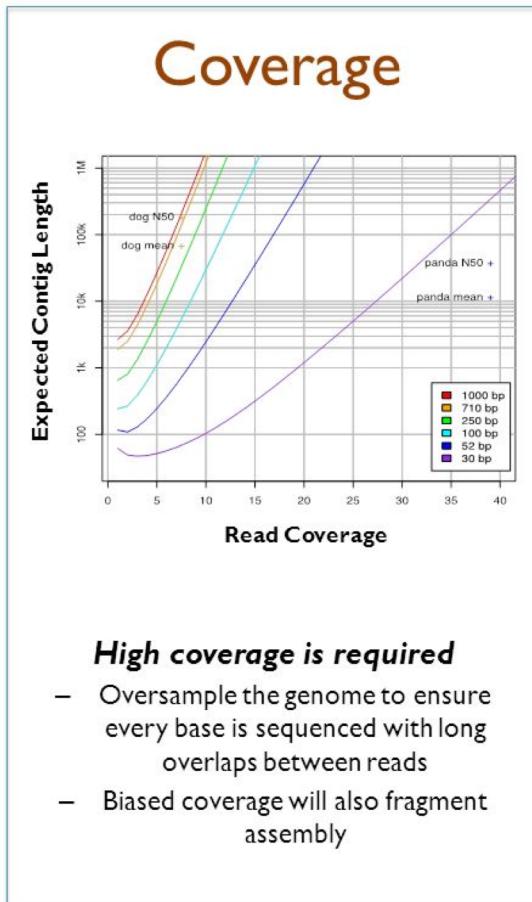


N50: measure of contiguity. The minimum sequence length for which 50% of the genome is in sequences at least this length

# Chromosomal scale helminth assemblies

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

# Recipe for a good genome assembly



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

# What tool(s) should I use?

## GENOME ASSEMBLY SOFTWARE TOOLS | DE NOVO SEQUENCING DATA ANALYSIS

High-throughput sequencing produces large amounts of long or short DNA reads which require assembly process to generate the complete genome sequence. De novo genome assembler programs have been written to detect overlaps between reads, to assemble overlaps into contigs, and then to combine contigs into scaffolds in order to obtain a draft genome sequence.

### ≡ PARENT CATEGORIES

### ≡ RELATED STEPS

### ≡ BENCHMARKING

### ≡ FILTERS

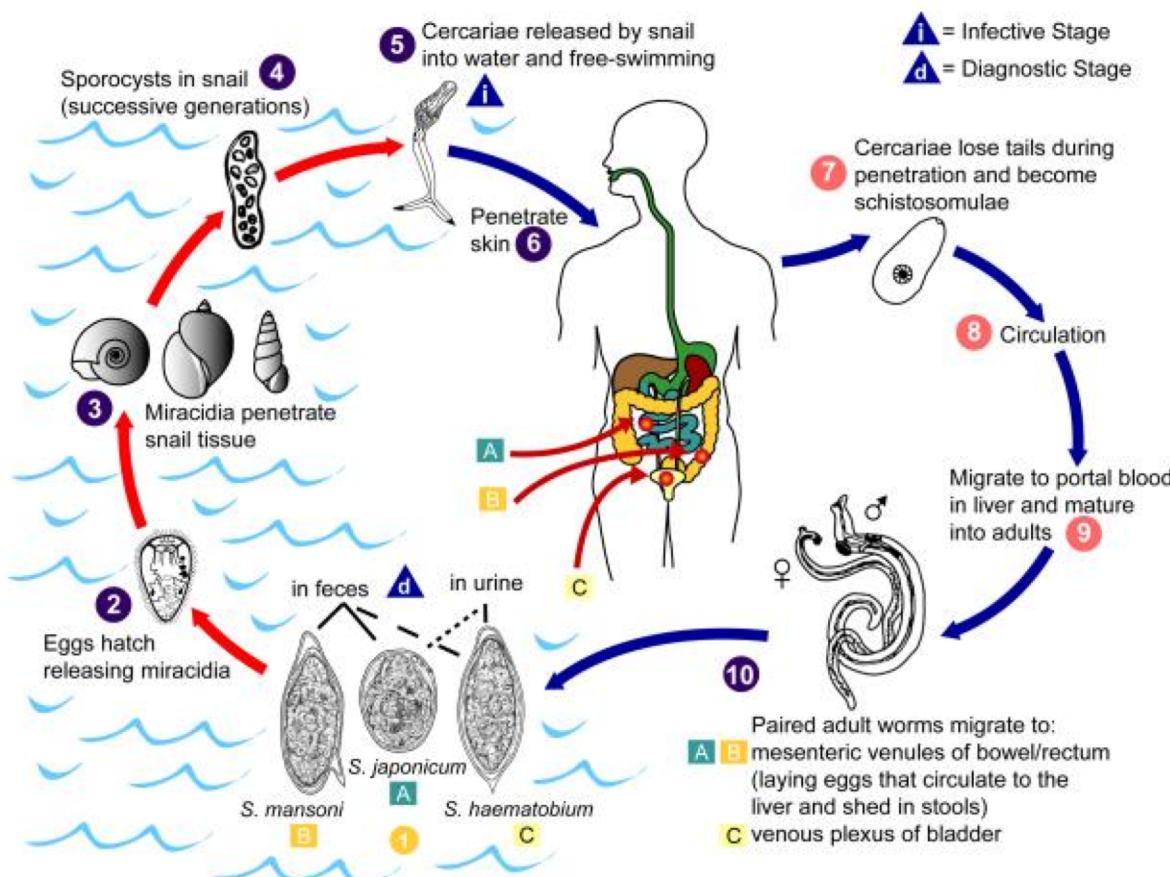
1 - 50 of 166



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

# Today: Assembly of a *Schistosoma mansoni* chromosome

## Schistosomiasis



## • Genome

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

# 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 we 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.