Parasite Genomics: Day 2

Steve Doyle & Adam Reid

Wellcome Sanger Institute / LSHTM

LSHTM Pathogen Genomics 2021 (virtual)

Friday 23rd April 2021

Program

- Manual, individual presentations
  - https://stephenrdoyle.github.io/LSHTM_ParasiteGenomics/

- Day 1
  - Module 1: Artemis
  - Module 2: Short read mapping

- **Day 2**
  - **Module 3: Comparative genomics**
  - **Module 4: Genome assembly**

Module 3:
Comparative Genomics

Comparative Genomics

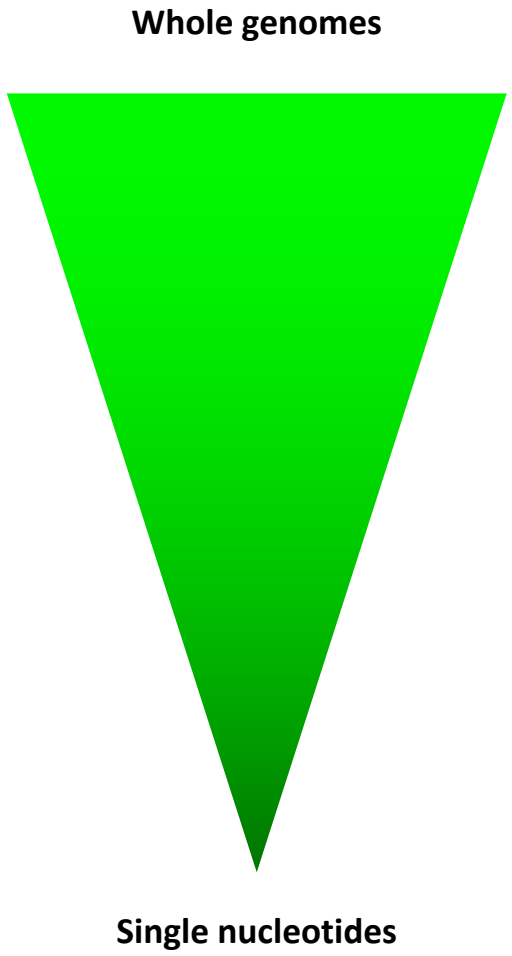
- Previously worked on individual genomes from a species - can teach us a lot about species biology

- We can learn more by analysing complete (or large parts of) genomes from within and/or between species

- Identification of similarities and differences between organisms can tell us about:
  - Core functions
  - Evolutionary history
  - Adaptation (for ex. Infectivity and virulence, alternative metabolic pathways, etc.)

- Particularly useful with increasing size of genomic datasets

Comparative Genomics

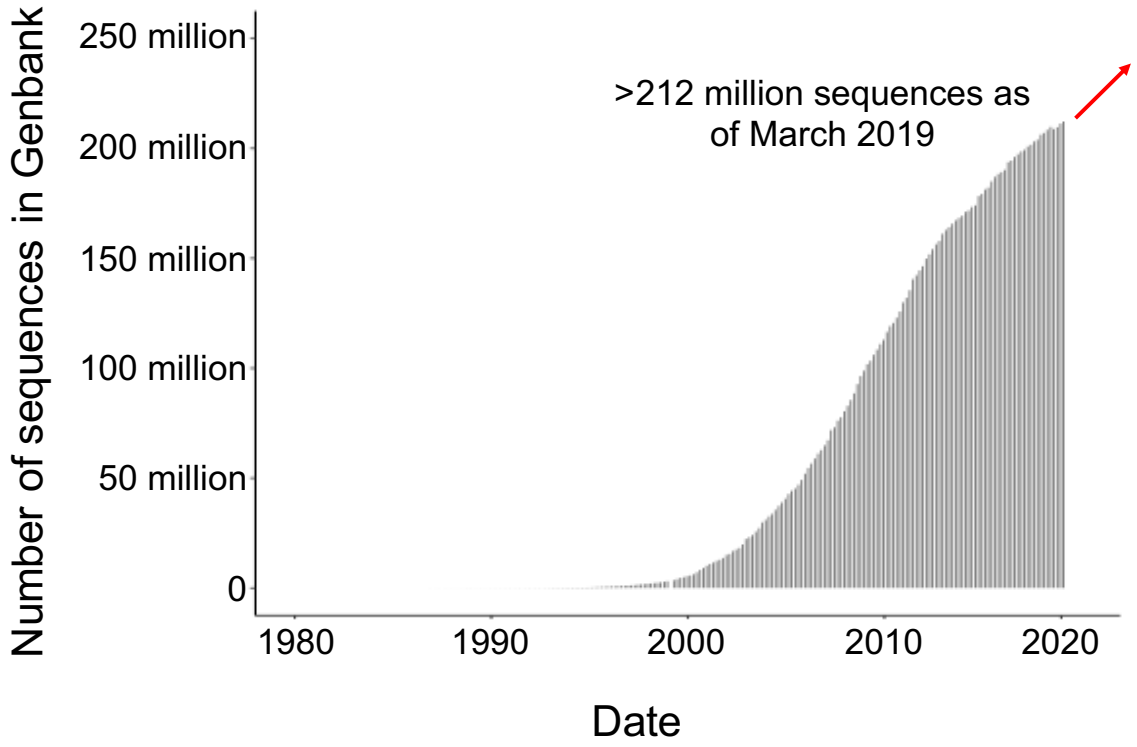


Useful over a wide variety of scales:

- Chromosomal structure & rearrangements
- Gene gain & loss
- Gene order & rearrangements
- Pseudogenes
- Coding and non-coding variation

Resources and tools

- Many resources exist for undertaking comparative genomic analyses
  - **Databases**: Genbank, Ensemble, MBGD…
  - **Datasets**: Metagenomic data, curated data, refseq...
  - **Dataset formats**: fasta, GFF, VCF…
  - **DNA similarity**: BLAST, Mummer…
  - **DNA composition**: GC%, codon usage, kmer…



Challenge: single gene to whole genome

- The output of these resources & tools is not easy to visualise / conceptualise, esp for large datasets / whole genomes

BLAST

```
Query: 1    MNPNQKIITIGSVCMTIGMANLILQIGNIISIWISHSIQLGNQNQIETCNQSVITYENNT 60
            MNPNQKIITIGSVCMTIGMANLILQIGNIISIWISHSIQLGNQNQIETCNQSVITYENNT
Sbjct: 1    MNPNQKIITIGSVCMTIGMANLILQIGNIISIWISHSIQLGNQNQIETCNQSVITYENNT 60

Query: 61   WVNQTYVNISNTNFAAGQSVVSVKLAGNSSLCPVSGWAIYSKDNSIRIGSKGDVFVIREP 120
            WVNQTYVNISNTNFAAGQSVVSVKLAGNSSLCPVSGWAIYSKDNSIRIGSKGDVFVIREP
Sbjct: 61   WVNQTYVNISNTNFAAGQSVVSVKLAGNSSLCPVSGWAIYSKDNSIRIGSKGDVFVIREP 120
```

nucmer

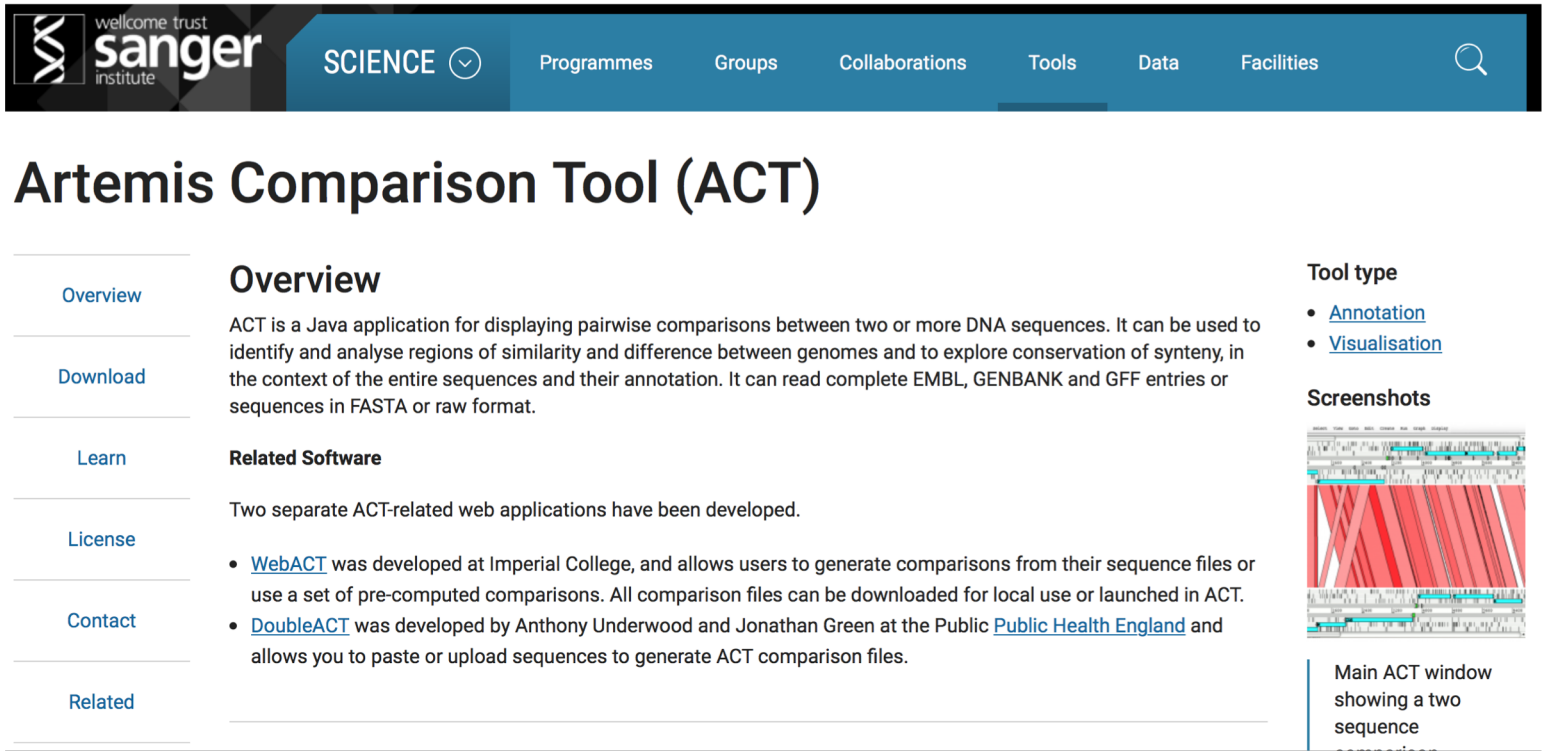
[S1]	[E1]	[S2]	[E2]	[LEN 1]	[LEN 2]	[% IDY]	[TAGS]	
1	364851	595755	960570	364851	364816	99.99	Wolbachia_onchocerca_volvulus|SC|contig000001	OVOC.WOLBACHIA
1811	1974	6260008	6260171	164	164	99.39	Wolbachia_onchocerca_volvulus|SC|contig000001	OVOC.OM2
1955	2338	1503891	1503509	384	383	97.14	Wolbachia_onchocerca_volvulus|SC|contig000001	OVOC.OM1b
2077	2222	6220019	6219877	146	143	92.47	Wolbachia_onchocerca_volvulus|SC|contig000001	OVOC.OM1b
4475	4686	779617	779824	212	208	89.62	Wolbachia_onchocerca_volvulus|SC|contig000001	OVOC.OM2
5510	5620	13325546	13325656	111	111	98.20	Wolbachia_onchocerca_volvulus|SC|contig000001	OVOC.OM4
13658	13753	6260606	6260697	96	92	94.79	Wolbachia_onchocerca_volvulus|SC|contig000001	OVOC.OM2
14169	14309	7196434	7196575	141	142	91.61	Wolbachia_onchocerca_volvulus|SC|contig000001	OVOC.OM4

Artemis Comparison Tool (ACT)

- Interactive tools to visualise and compare two or more sequences and their annotations
  - Useful for whole genomes down to single genes

- Based on, and builds upon, Artemis
  - Many of the functions learnt in Artemis are the same

Where to find ACT?

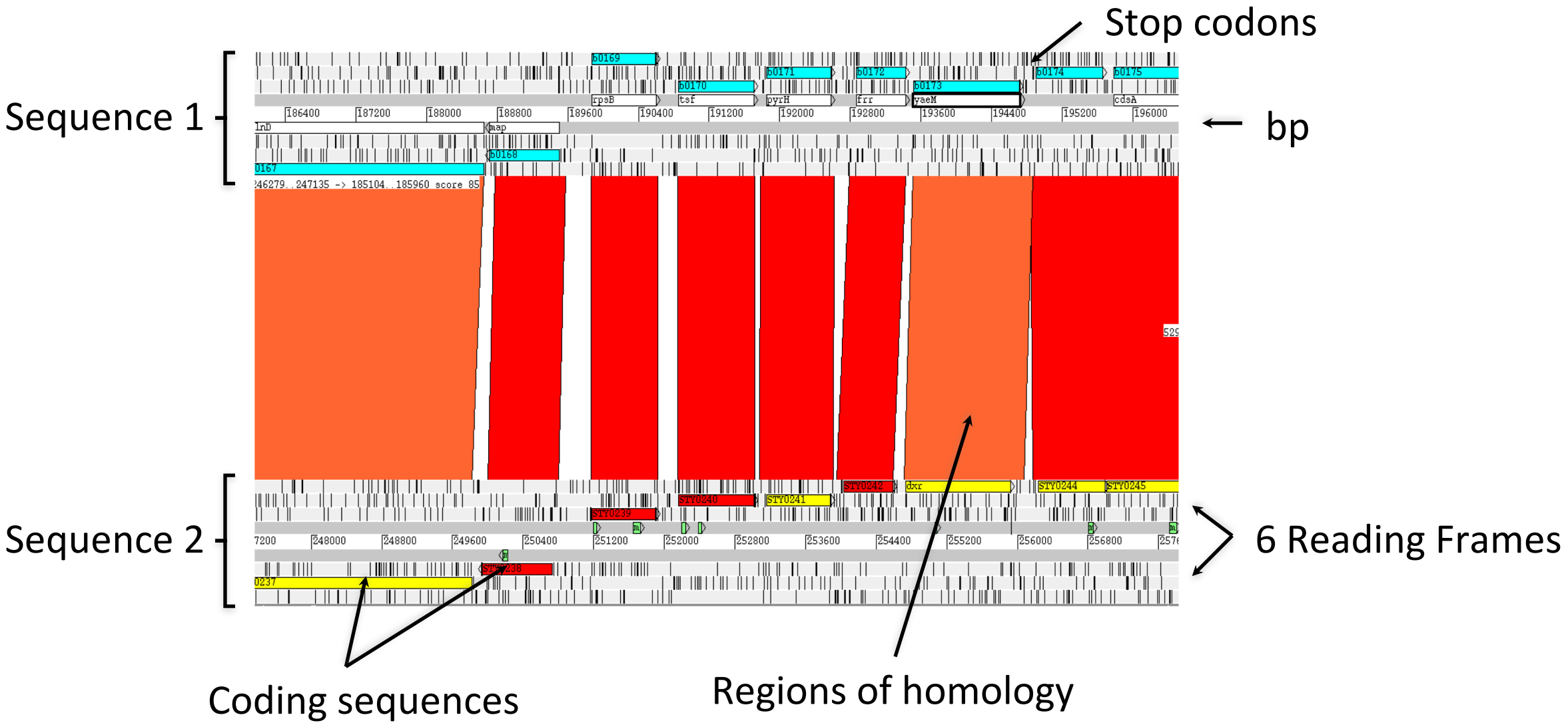


- Freely available (unix, mac, win)
  - WEB: http://www.sanger.ac.uk/science/tools/artemis-comparison-tool-act
  - MANUAL: ftp://ftp.sanger.ac.uk/pub/resources/software/act/act.pdf
  - PAPER: Carver et al. (2005) Bioinformatics. 21;16;3422-3

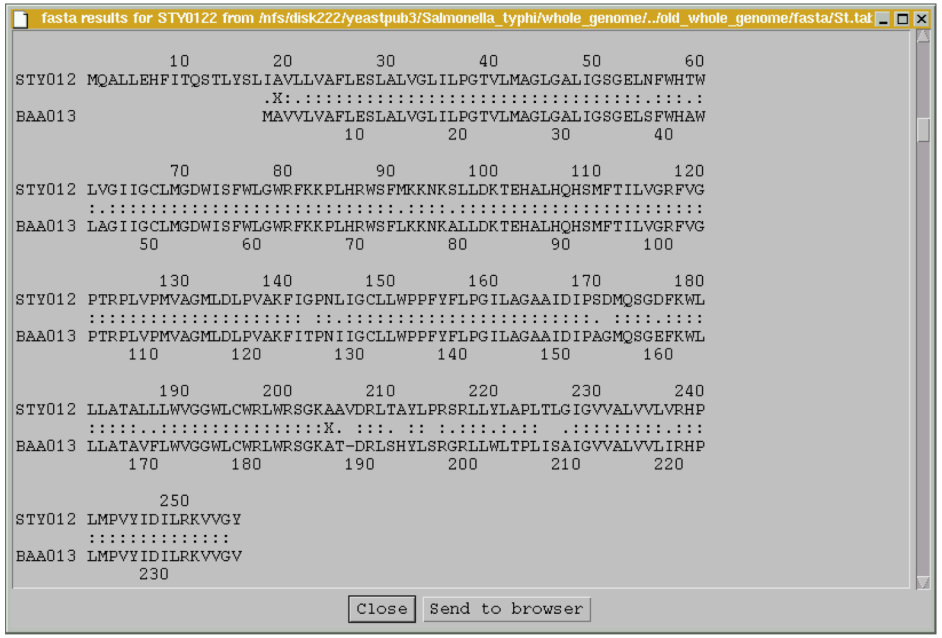
What do you need to run ACT?

- The program
  - Local install (currently on your VM desktop) or on the web (as described in your manual)
- Two or more files containing sequence information
  - Format: Fasta, EMBL, Genbank
- **Output from a comparative analysis of those sequence files**
  - Format: BLAST, mummer, VCF
- Optional: Additional metadata
  - Format: GFF, EMBL

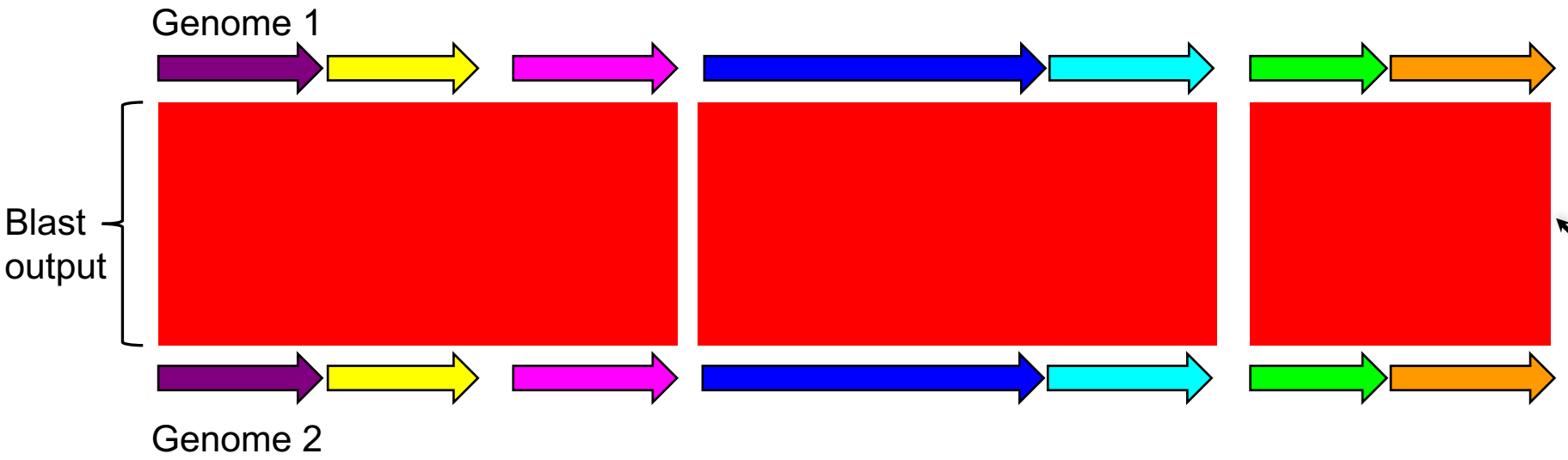
Basic setup of an ACT session



Visualising sequence similarity



Visualising genome rearrangements

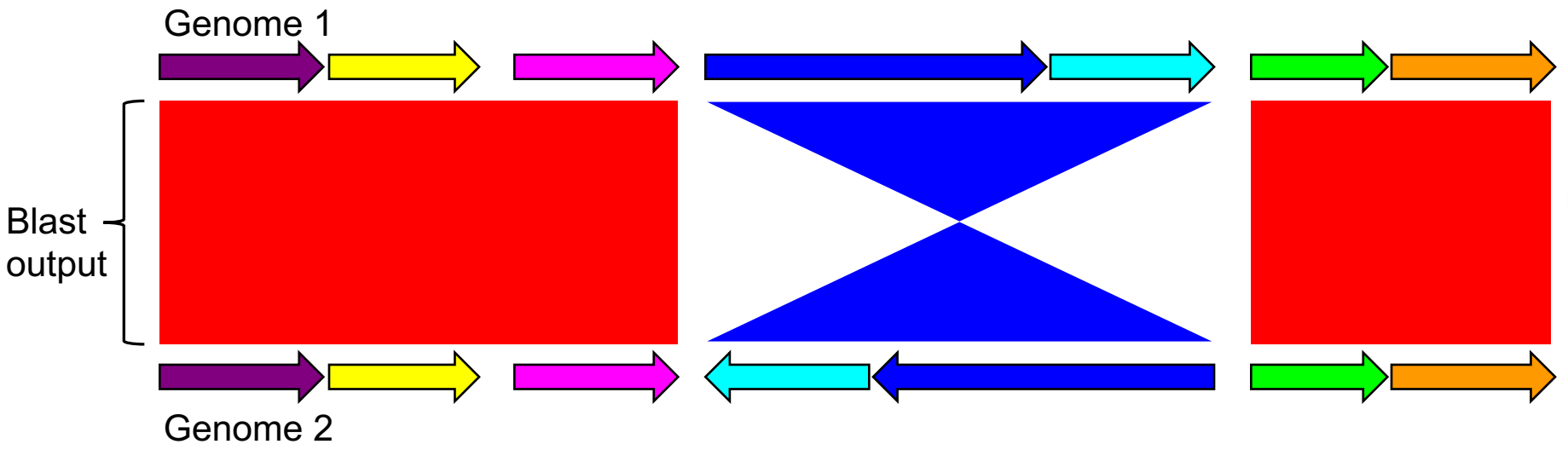


Region of homology in:

- Same orientation
- Reverse orientation

Same genes present in the same orientation

Visualising genome rearrangements

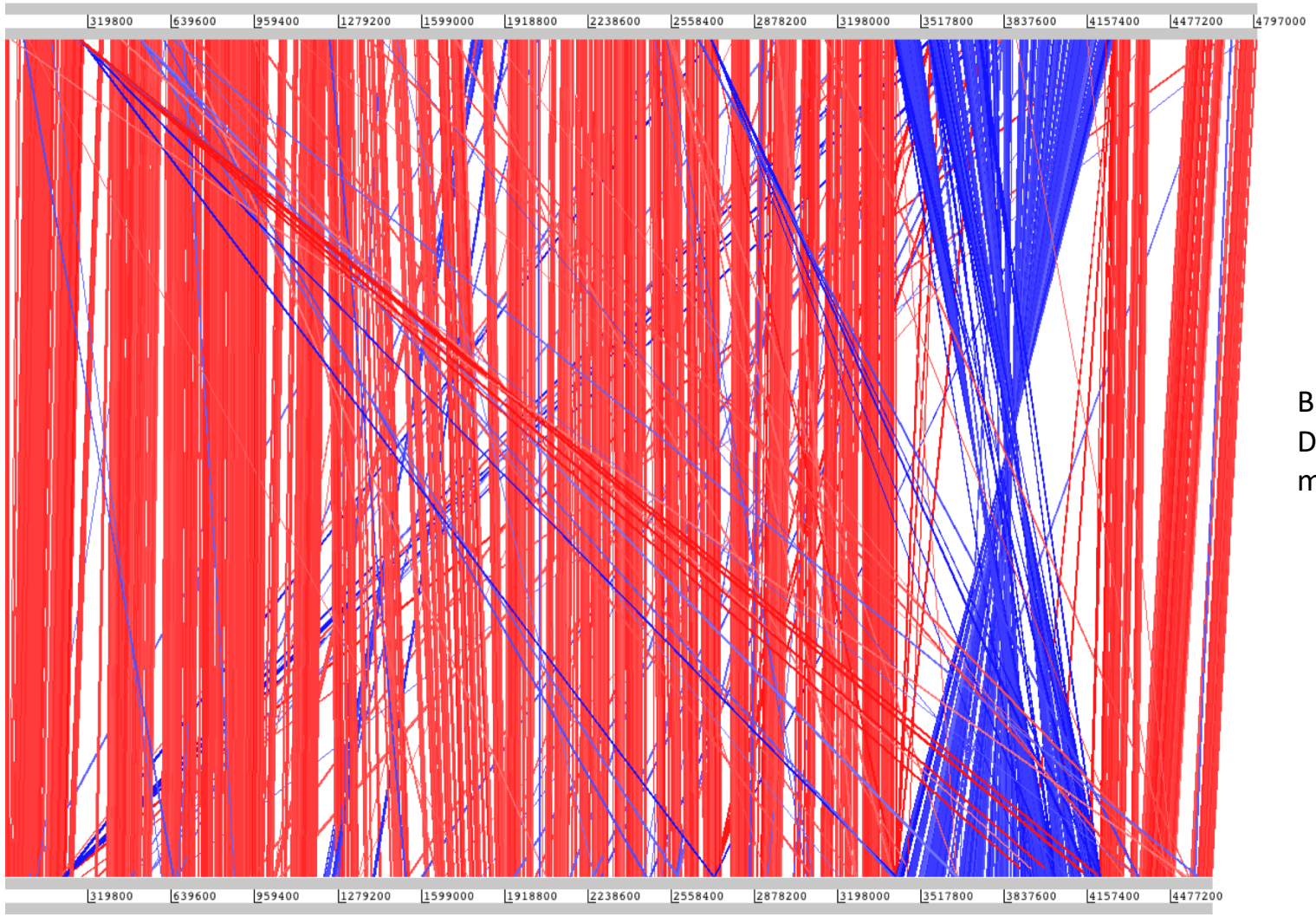


Region of homology in:

- Same orientation
- Reverse orientation

The same genes are present but the blue genes have undergone a rearrangement

Visualising genome rearrangements

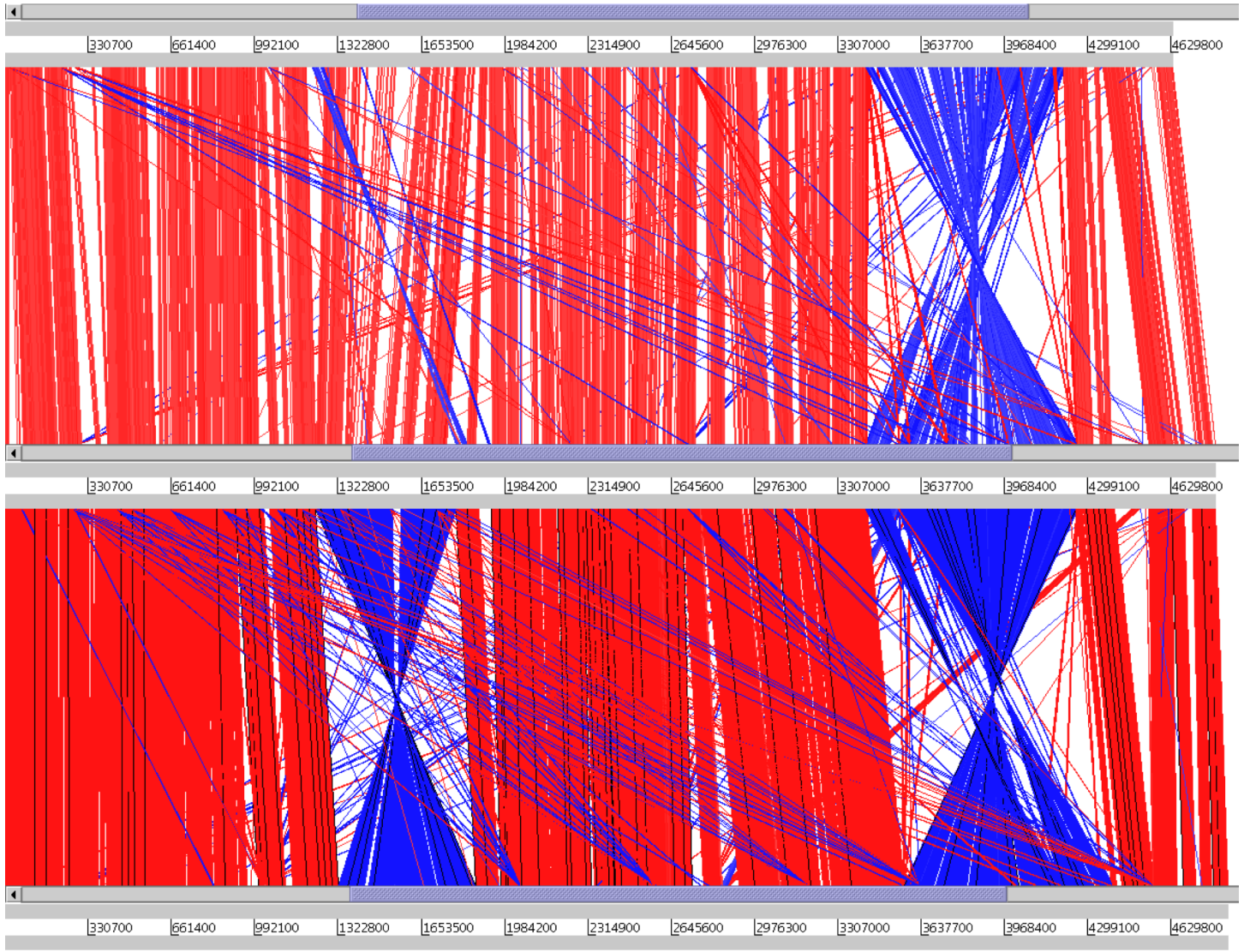


S. typhi

BLASTN
DNA-DNA
matches

E. coli

Visualising genome rearrangements

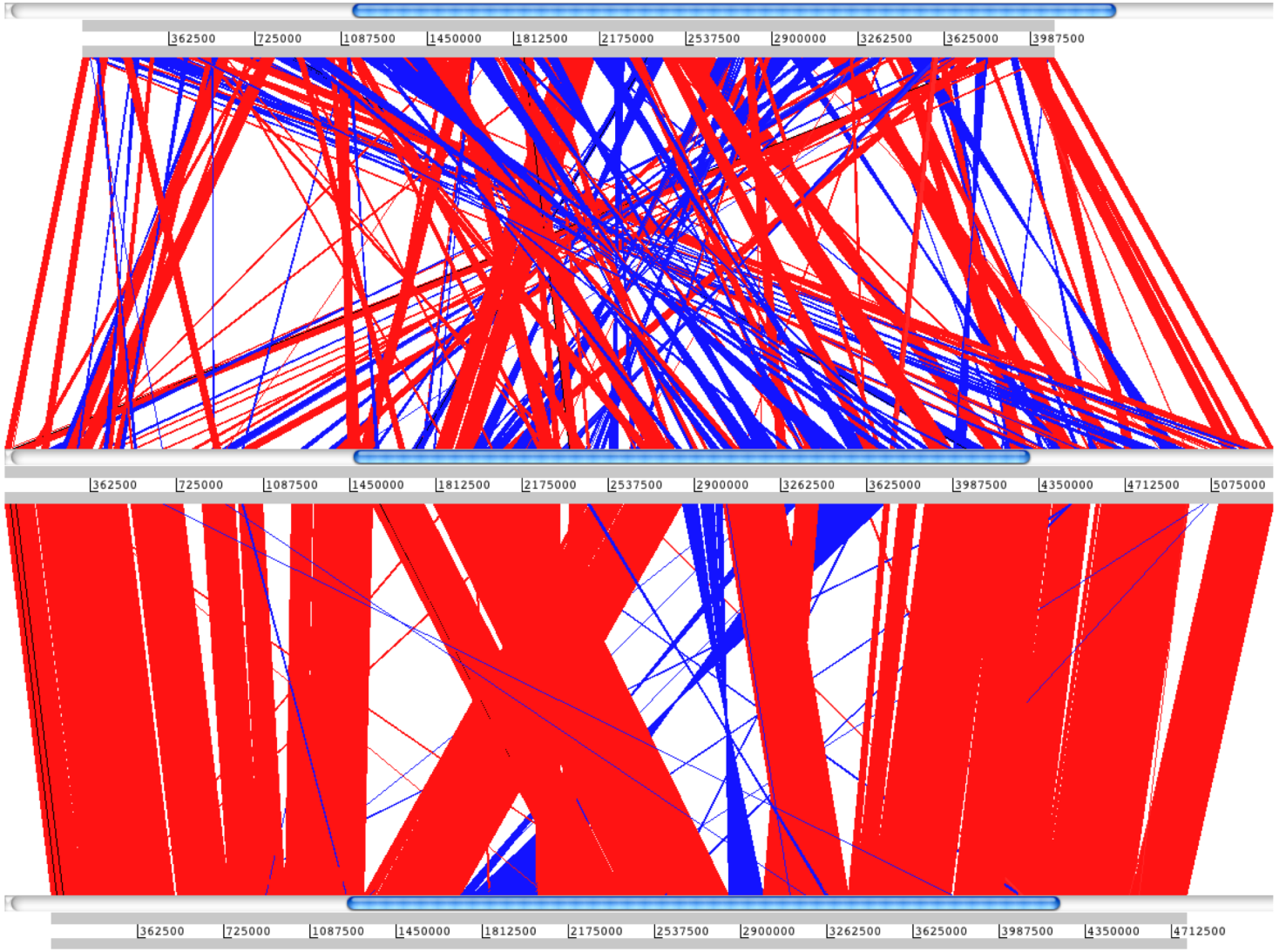


E. coli

S. typhi

S. typhimurium

Visualising genome rearrangements: *Bordetella spp.*

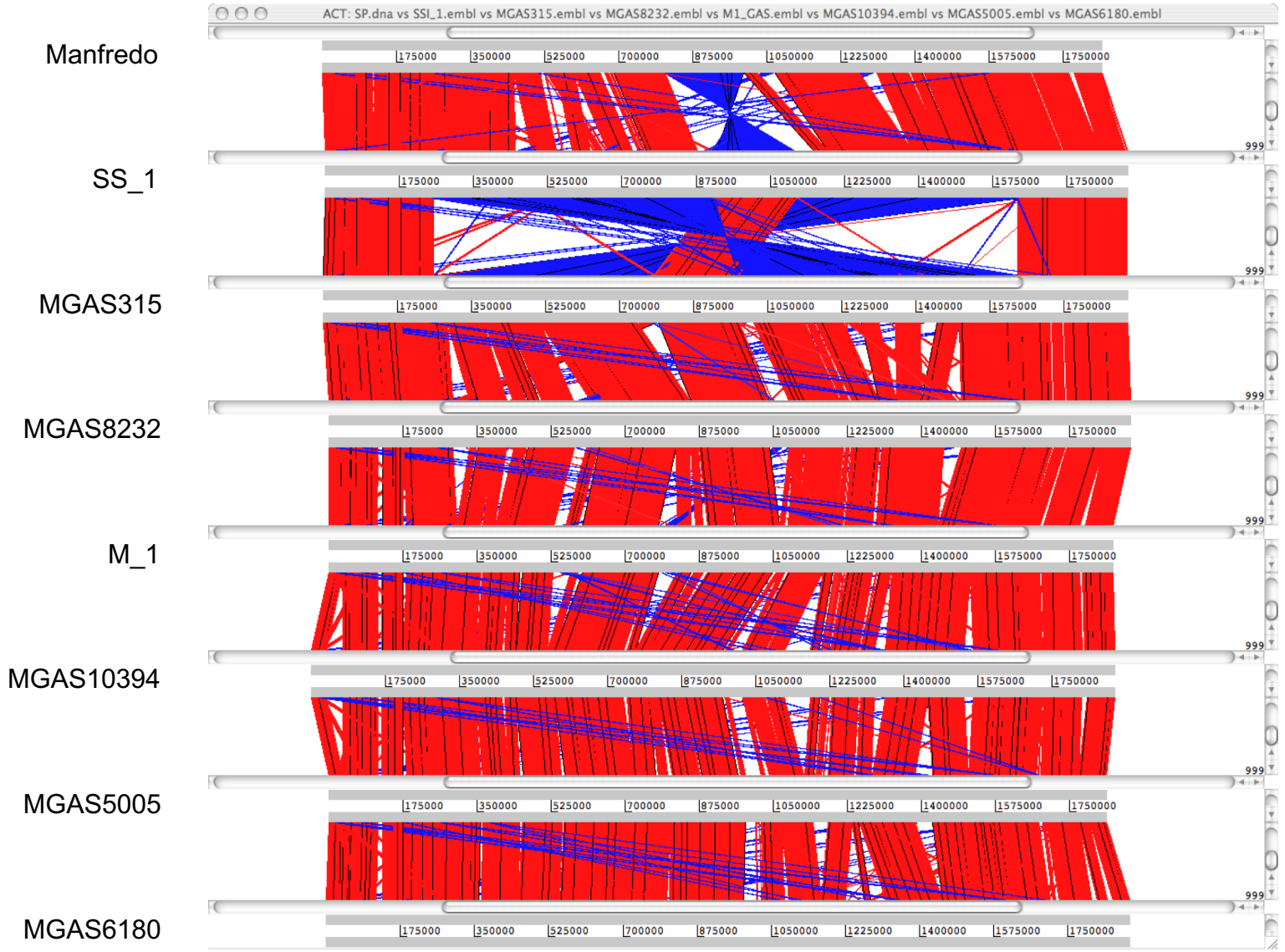


Bordetella pertussis

B. bronchiseptica

B. parapertussis

Comparison of strain diversity: *Streptococcus pyogenes*



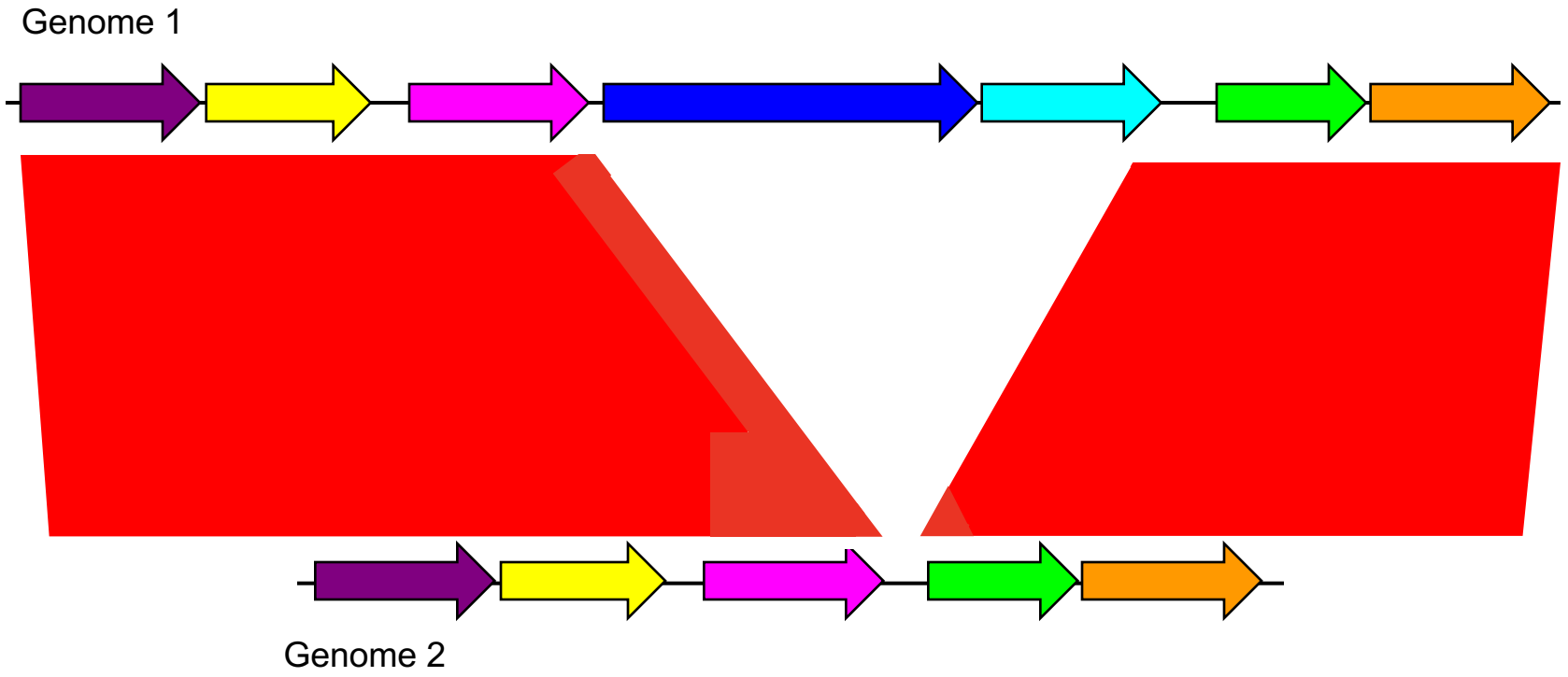
Comparison of genic diversity

- Gene gain and/or loss

  Duplication, horizontal gene transfer, indels

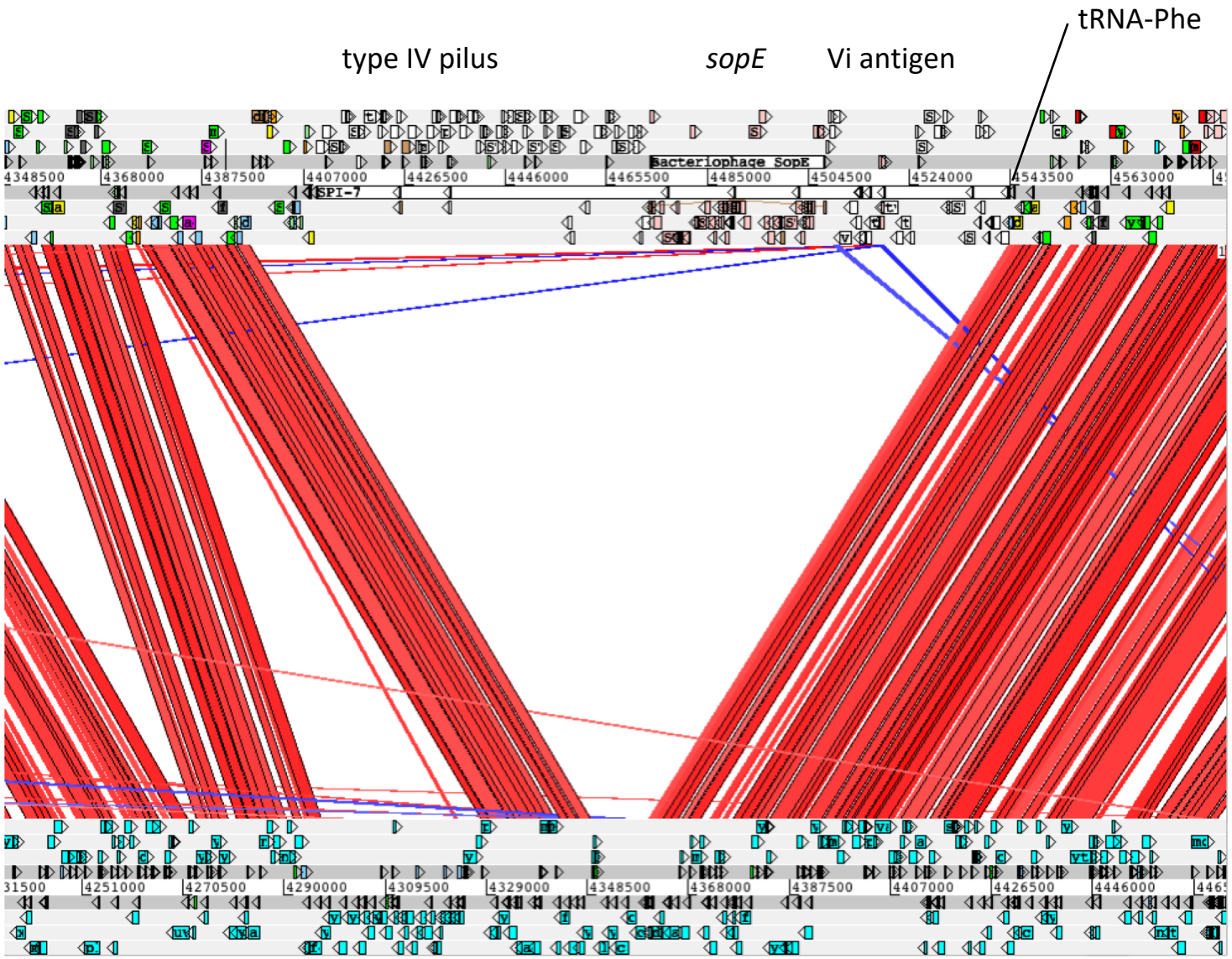
- Gene function modification
  - Loss of function (pseudogenes, gene fission)
  - Gain of function (gene fusion)
  - Mutation accumulation (SNPs)

Visualising gene gain or loss



Example of an insertion or deletion

Visualising gene gain or loss

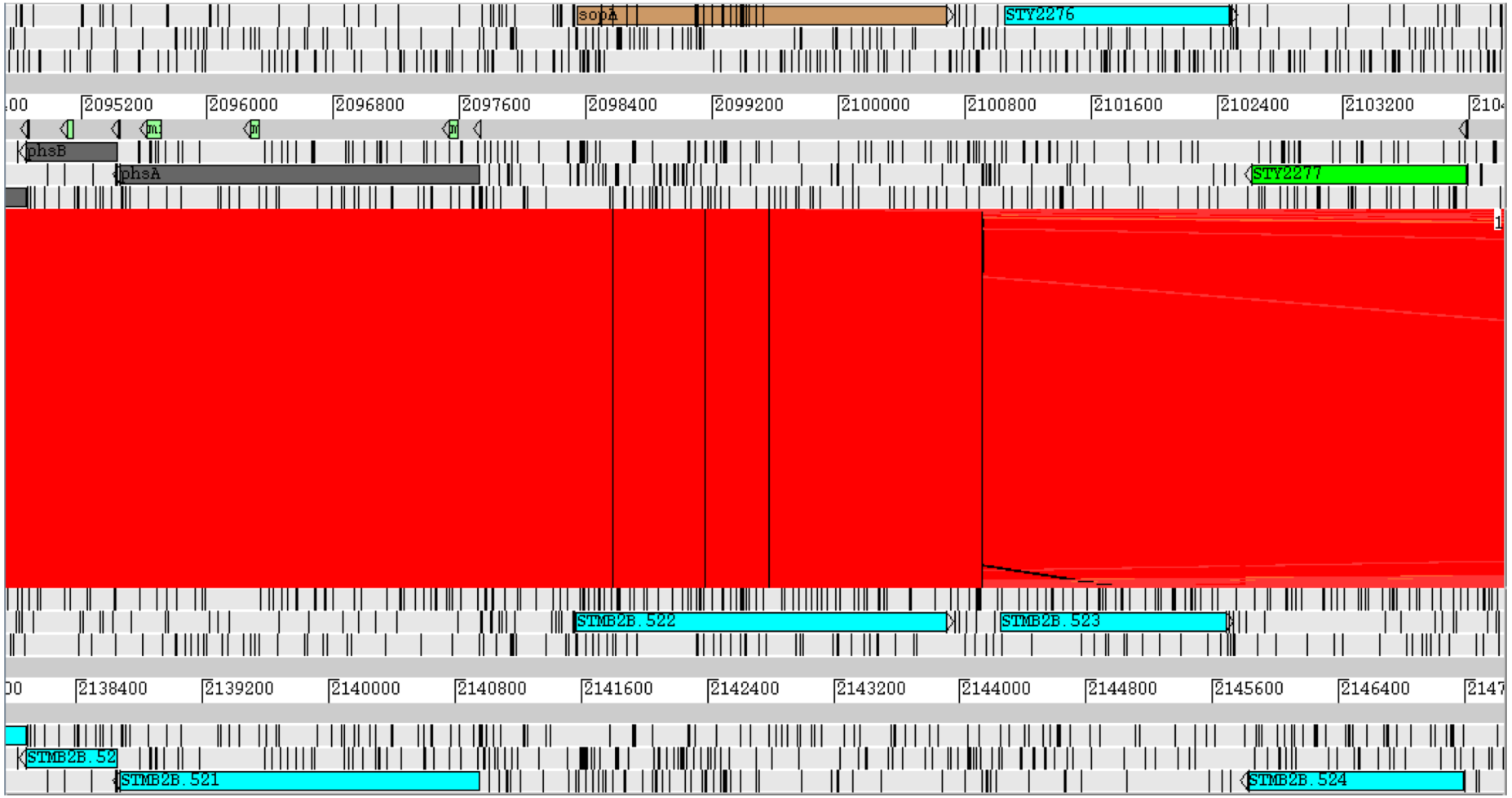


S. typhi

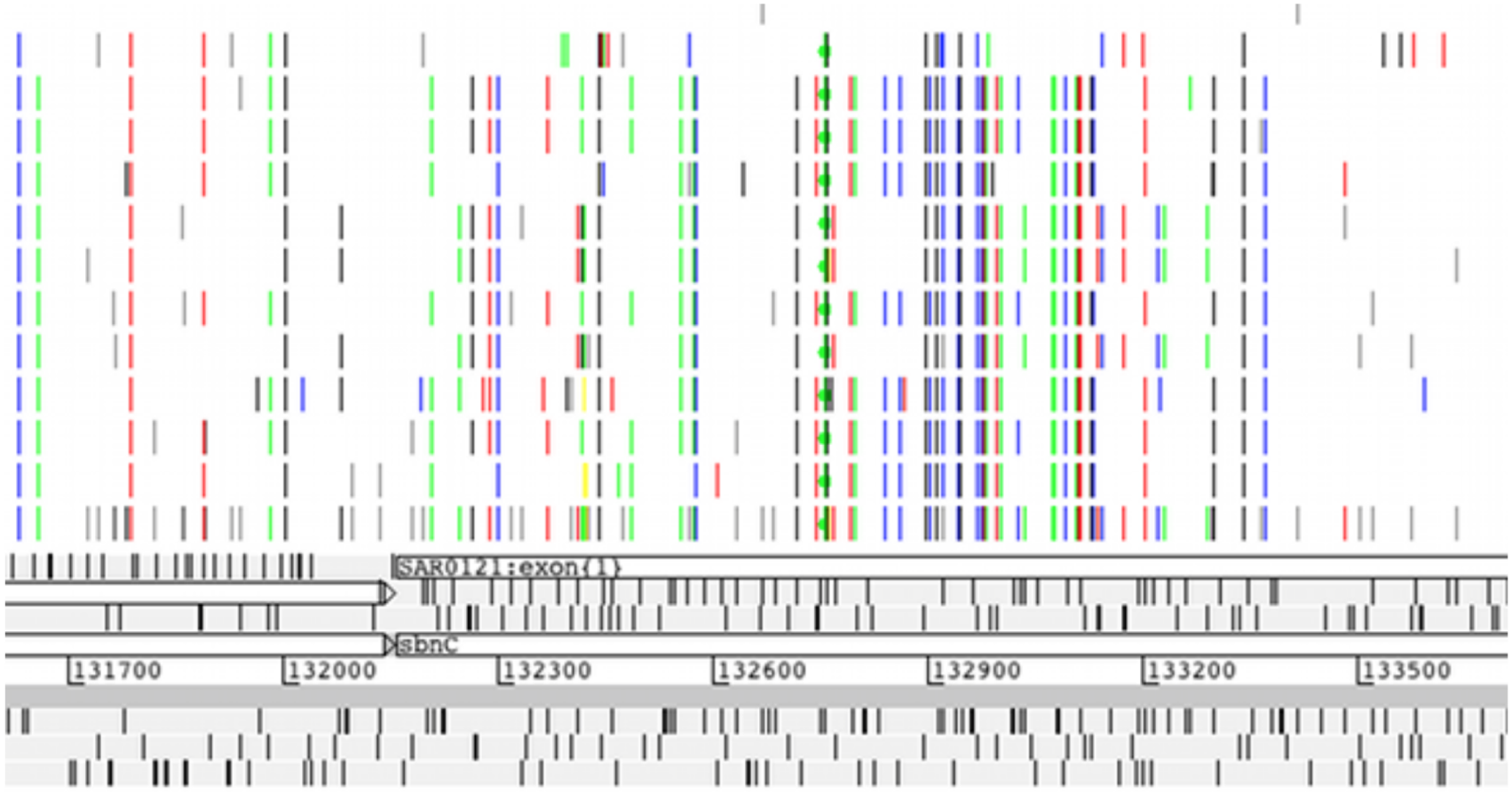
E. coli

Loss of function: pseudogenes

secreted effector protein (type-III)



Genetic variation among populations



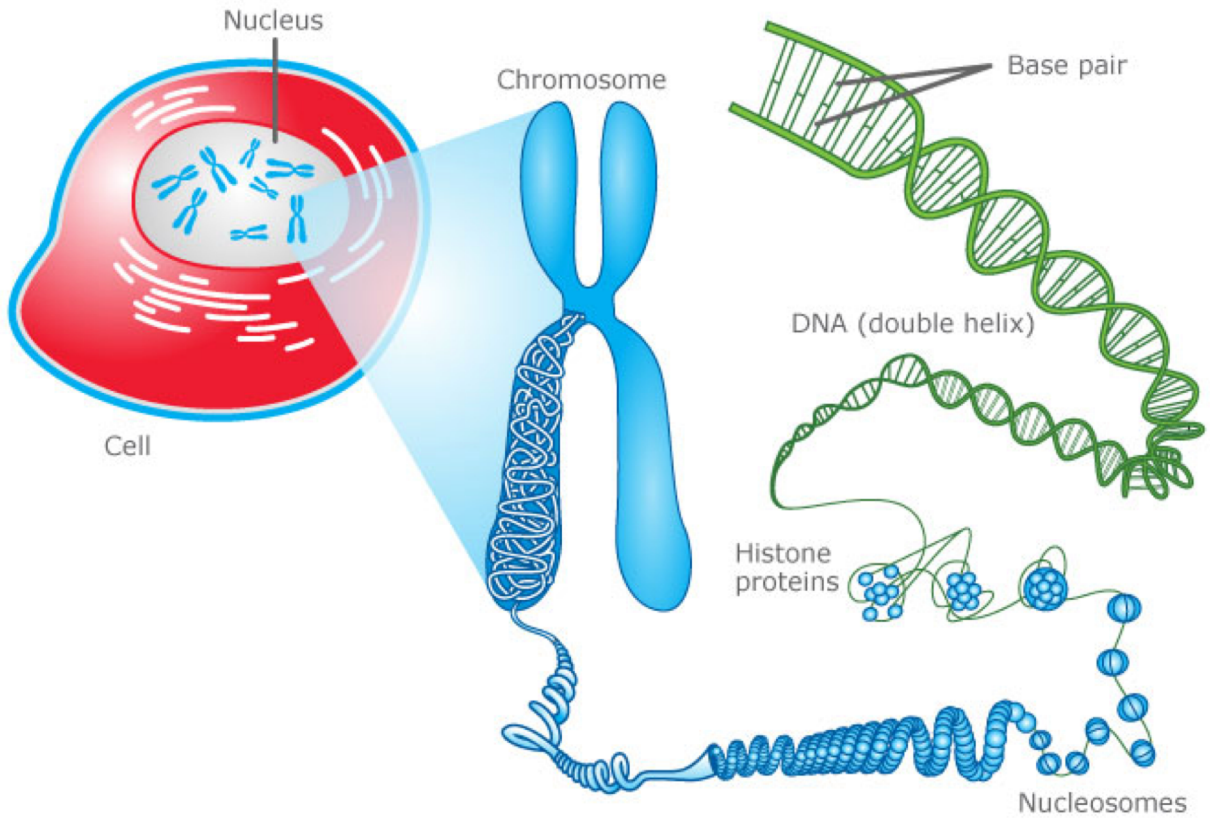
VCF input: row represents an individual, line represents a SNP

Aims of this module

- Demonstrate some of the the basic functions of ACT
- Enable you to perform basic comparative analyses

Module 4:
De novo genome assembly

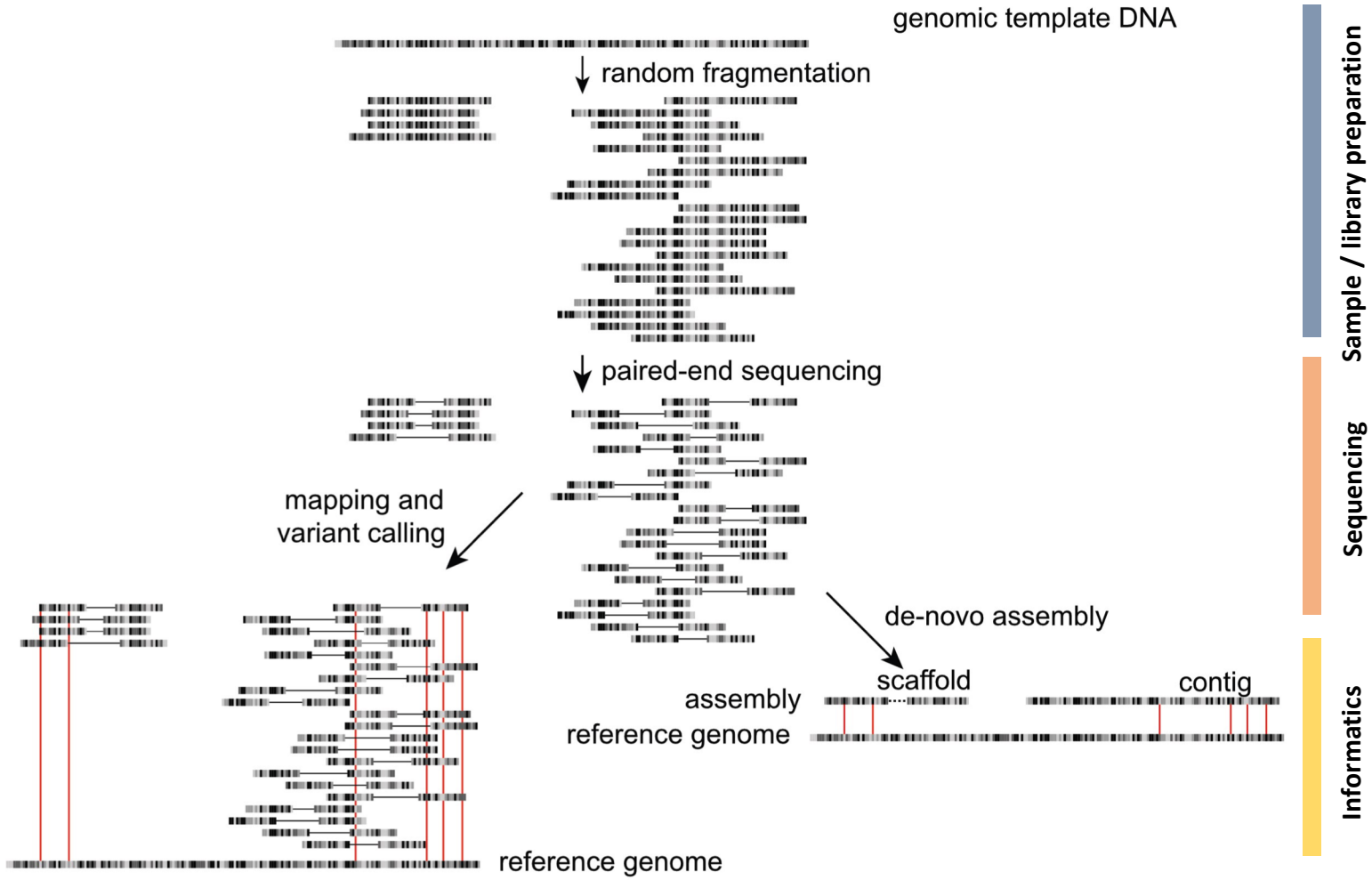
Genome sequencing is conceptually straight forward




```
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…!



Sequencing genomes is easy, constructing good genomes is not

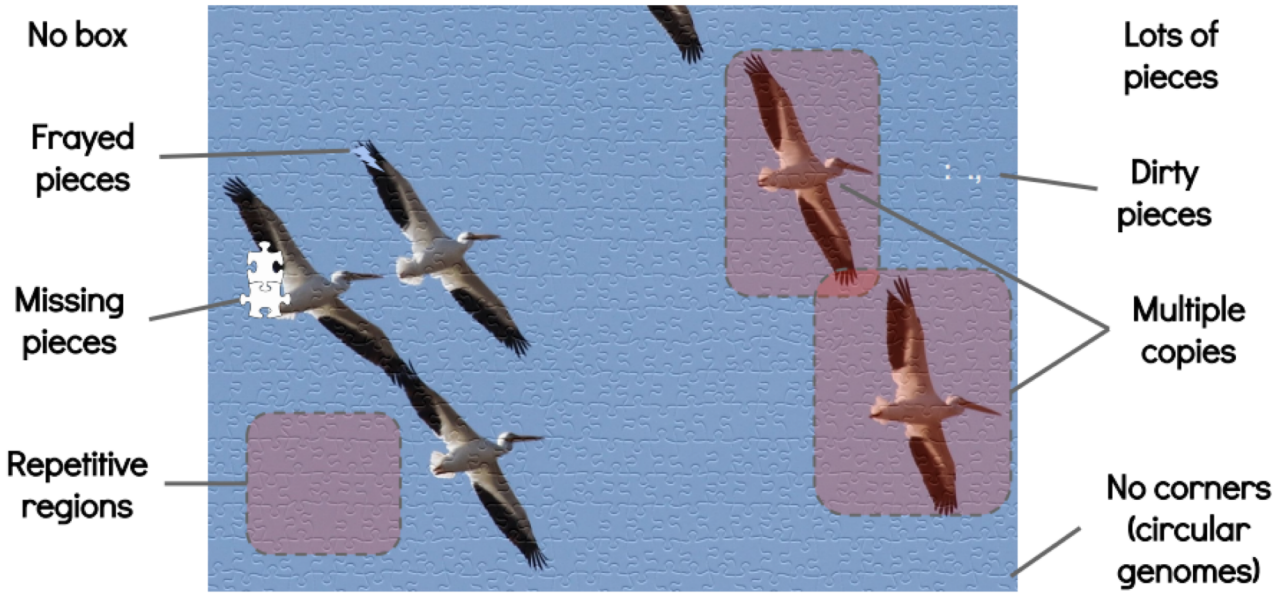
- **Genome: *biologically***
  - “the haploid set of chromosomes in a gamete or microorganism, or in each cell of a multicellular organism”
  - “the complete set of genes or genetic material present in a cell or organism”

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

wellcome sanger institute

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

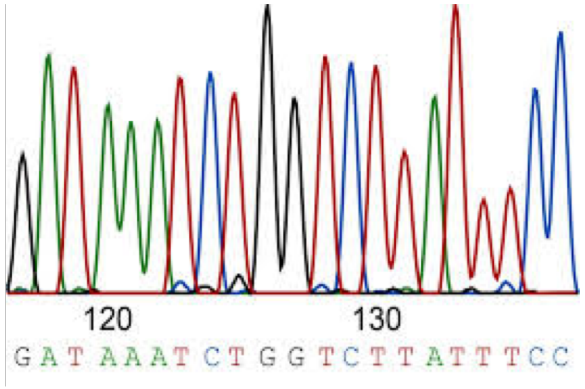
Time, money, expertise

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

wellcome sanger institute

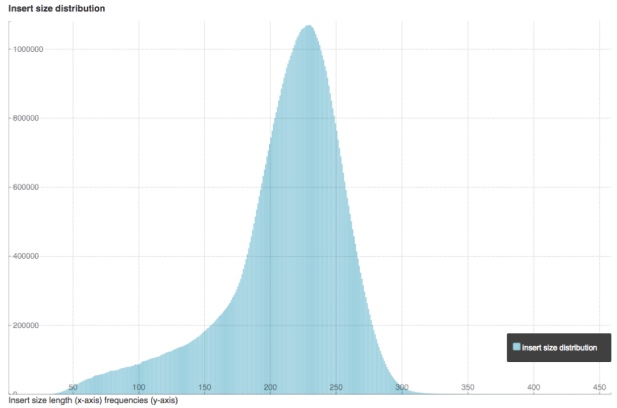
New technologies are making genome assembly easier

Sanger Sequecing: ABi



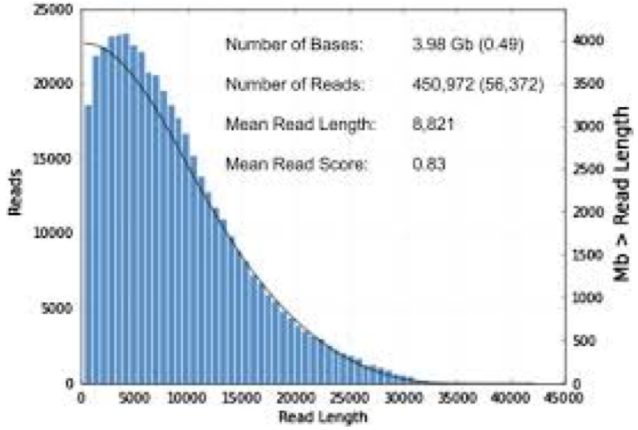
Read length:500-1000 bp

High Throughput Sequencing: Illumina



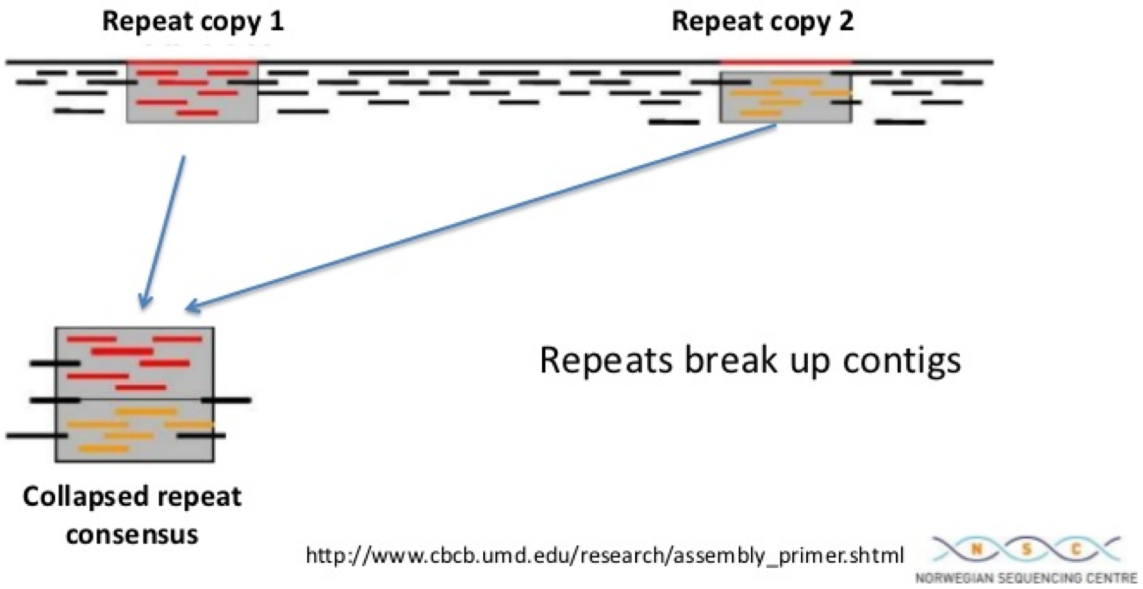
Read lengths: 100-300 bp
Insert lengths: ave 300-500 bp

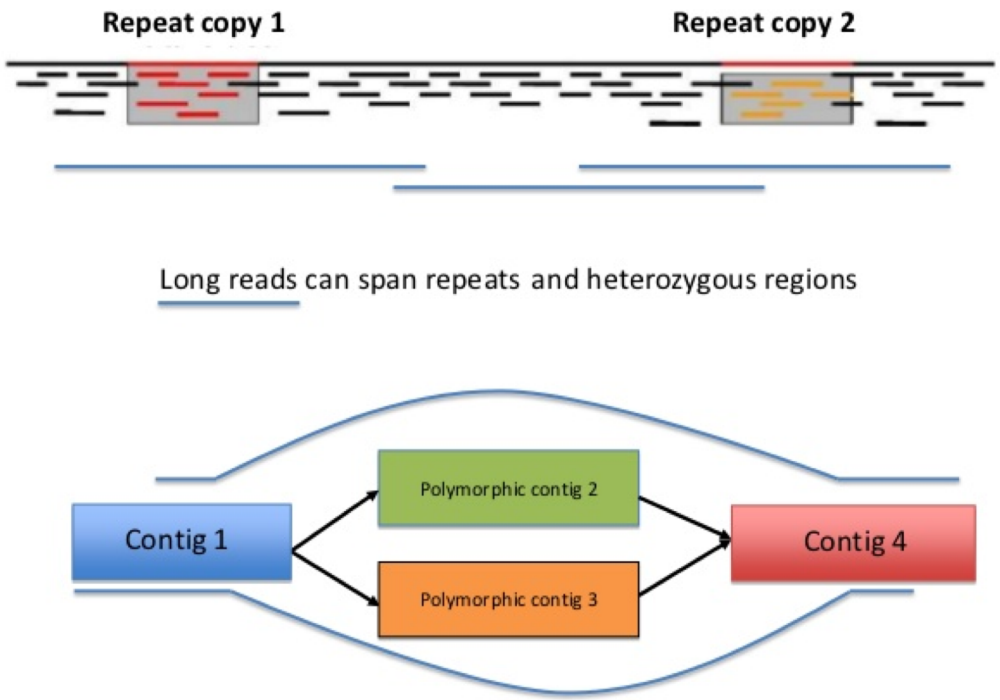
Long read sequencing: Pacbio & Nanopore



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

Repeats / polymorphic loci can break genomes



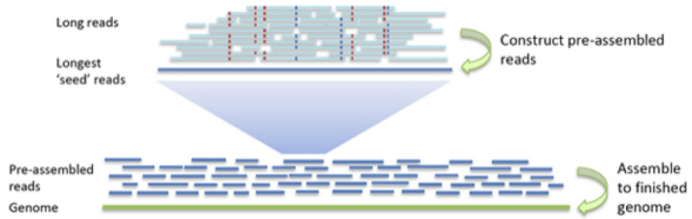


https://www.slideshare.net/flxlex/combining-pacbio-with-short-read-technology-for-improved-de-novo-g

wellcome sanger institute

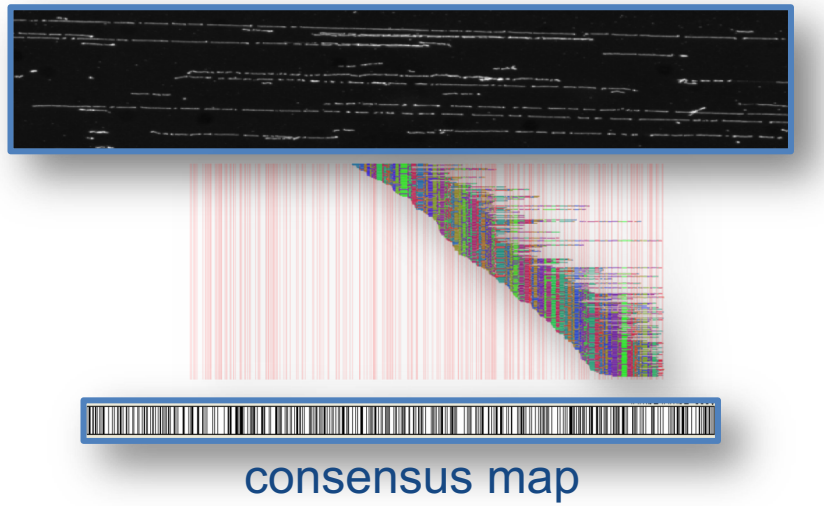
Long read / range sequencing is key to good genomes

Pacific Biosciences (PacBio)



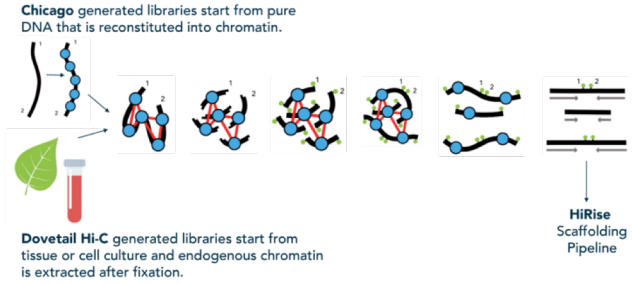
Oxford Nanopore

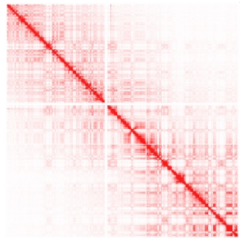
Optical Mapping (OpGen, Bionano genomics)



Linked reads (10X Genomics)

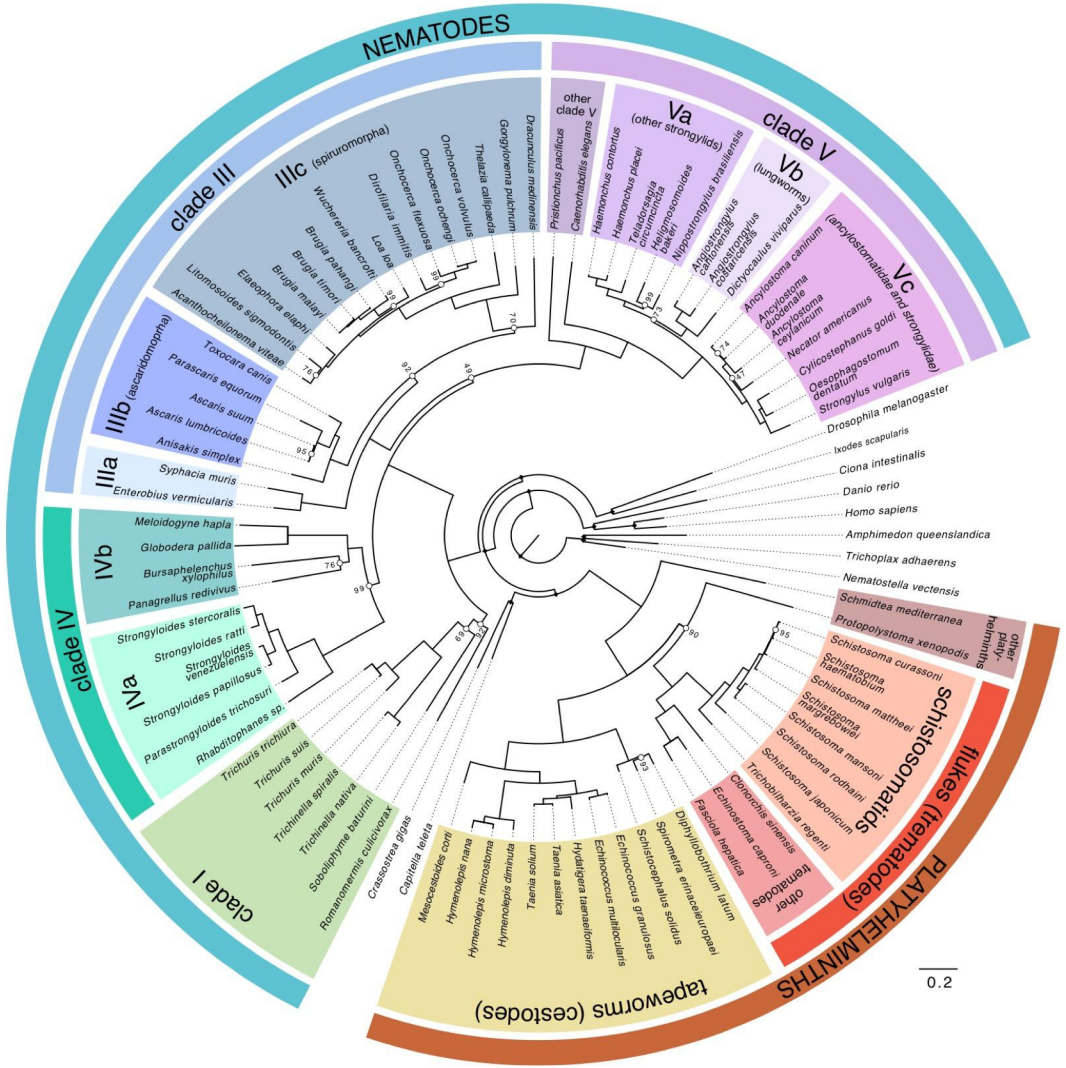
Chromosome confirmation capture, ie Hi-C (Dovetail Genomics)





wellcome sanger institute

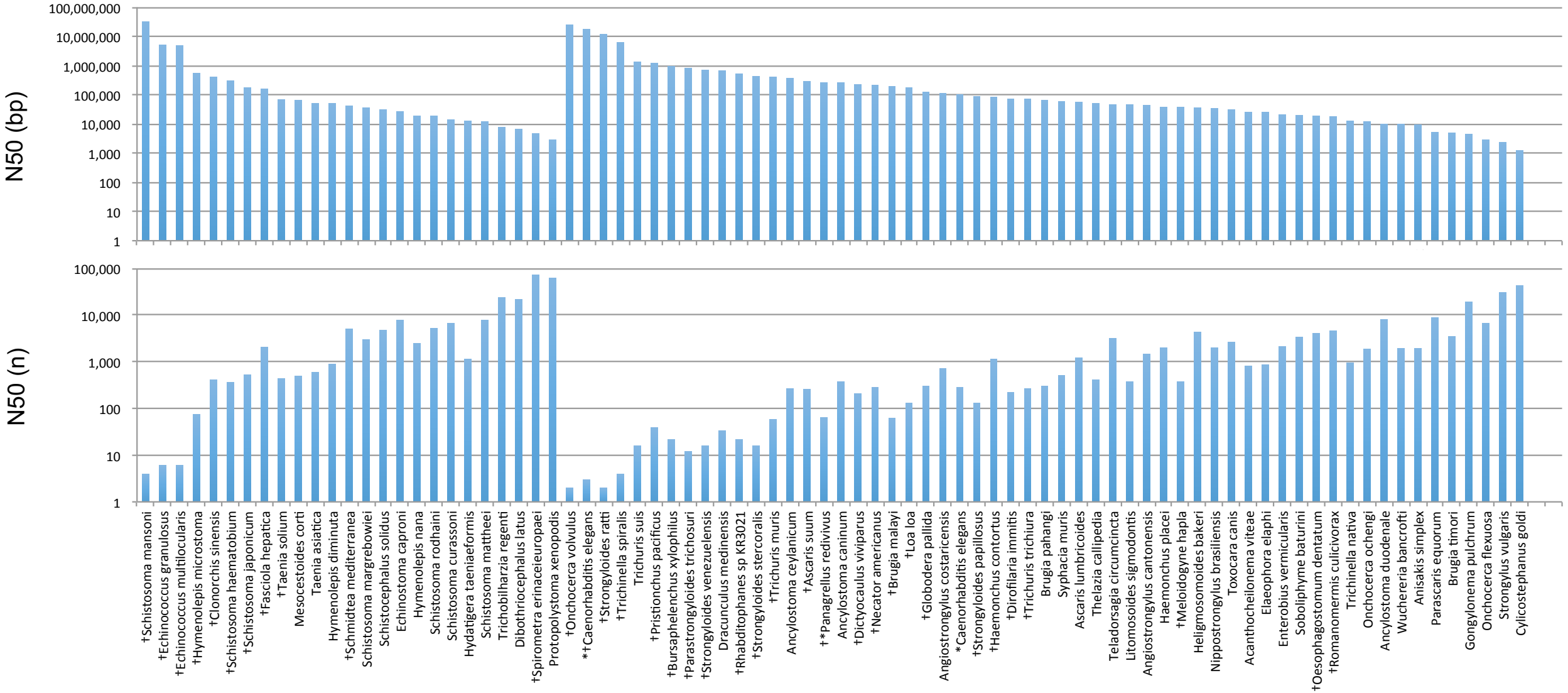
Parasite Genomics @ Sanger: 50 Helminth Genome Project



- 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 sanger institute

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

Sanger institute

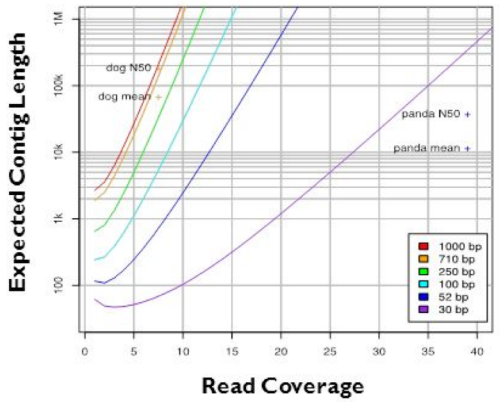
Chromosomal scale helminth assemblies

	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		

wellcome sanger institute

Recipe for a good genome assembly

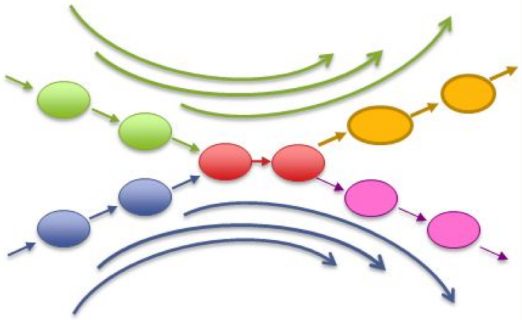
Coverage



High coverage is required

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

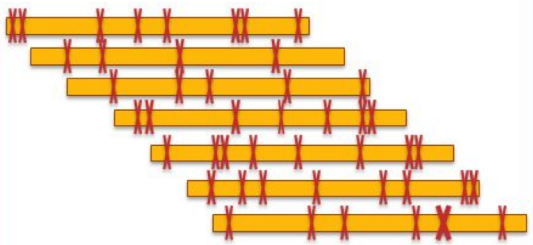
Read Length



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



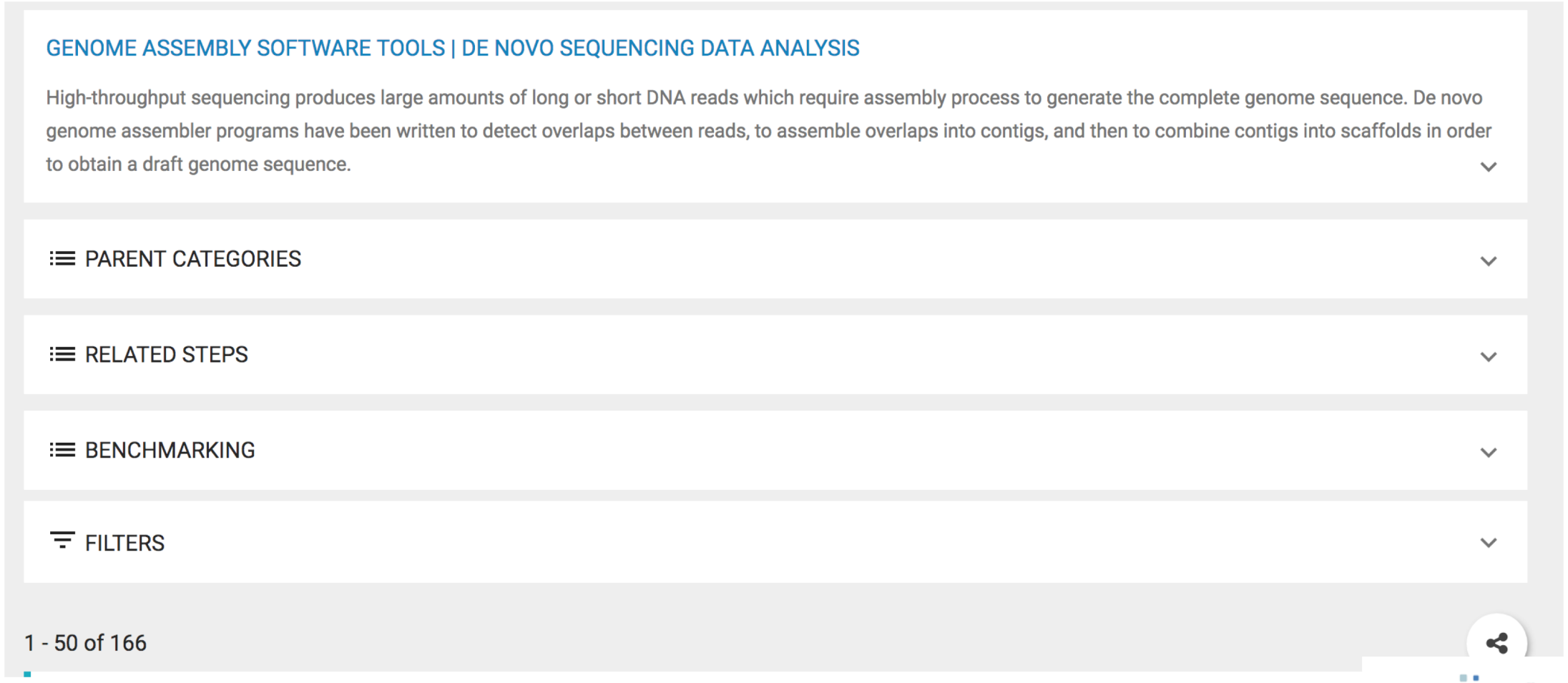
Errors obscure overlaps

- 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

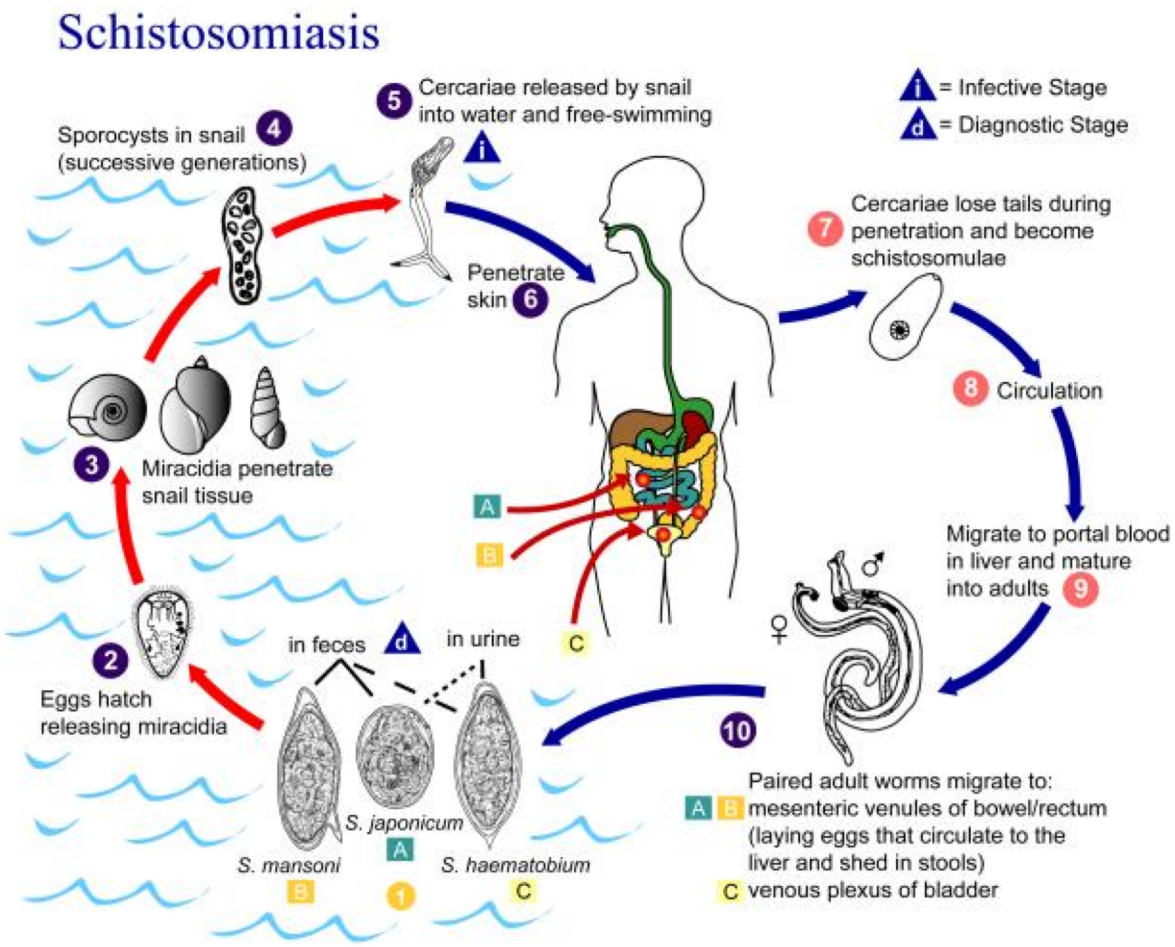
What tool(s) should I use?



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

wellcome sanger institute

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

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

wellcome sanger institute

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.