Parasite Genomics: Day 2

Steve Doyle & Adam Reid

Wellcome Sanger Institute / LSHTM

LSHTM Pathogen Genomics

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: Transcriptome analysis - RNAseq**

Module 3:
Comparative Genomics

Comparative Genomics

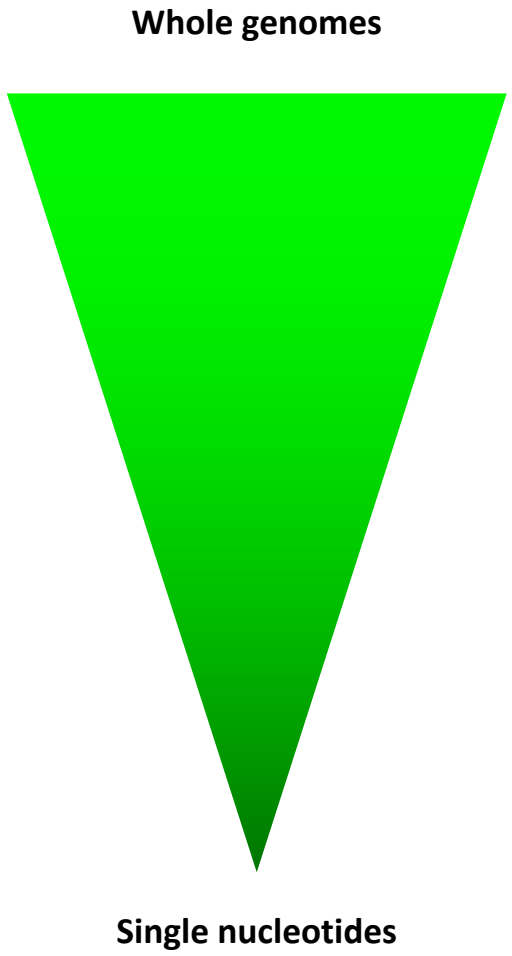
- Previously worked on individual genomes from a species - can teach us a lot about species biology

- 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

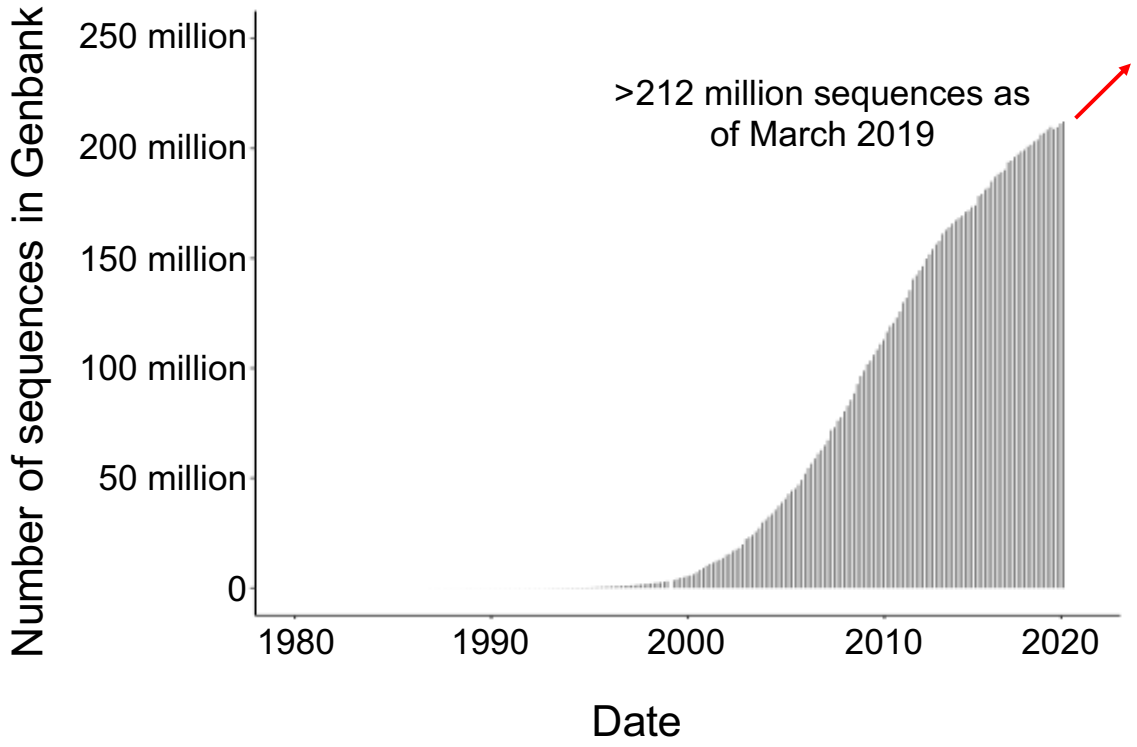


Useful over a wide variety of scales:

- 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

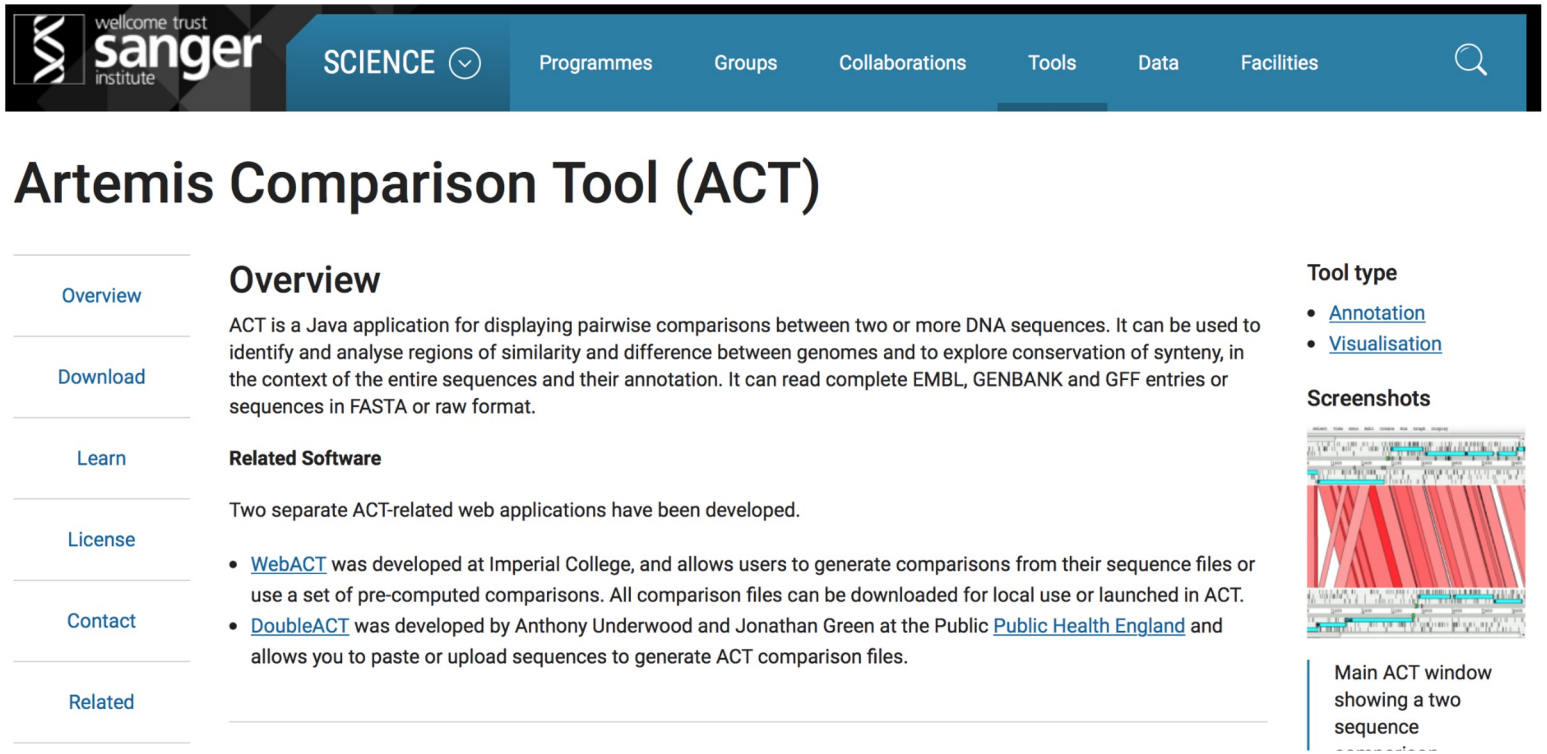
[S1]	[E1]	[S2]	[E2]	[LEN 1]	[LEN 2]	[% IDY]	[TAGS]	
1	364851	595755	960570	364851	364816	99.99	Wolbachia_onchocerca_volvulus|SC|contig000001	OVOC.WOLBACHIA
1811	1974	6260008	6260171	164	164	99.39	Wolbachia_onchocerca_volvulus|SC|contig000001	OVOC.OM2
1955	2338	1503891	1503509	384	383	97.14	Wolbachia_onchocerca_volvulus|SC|contig000001	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?

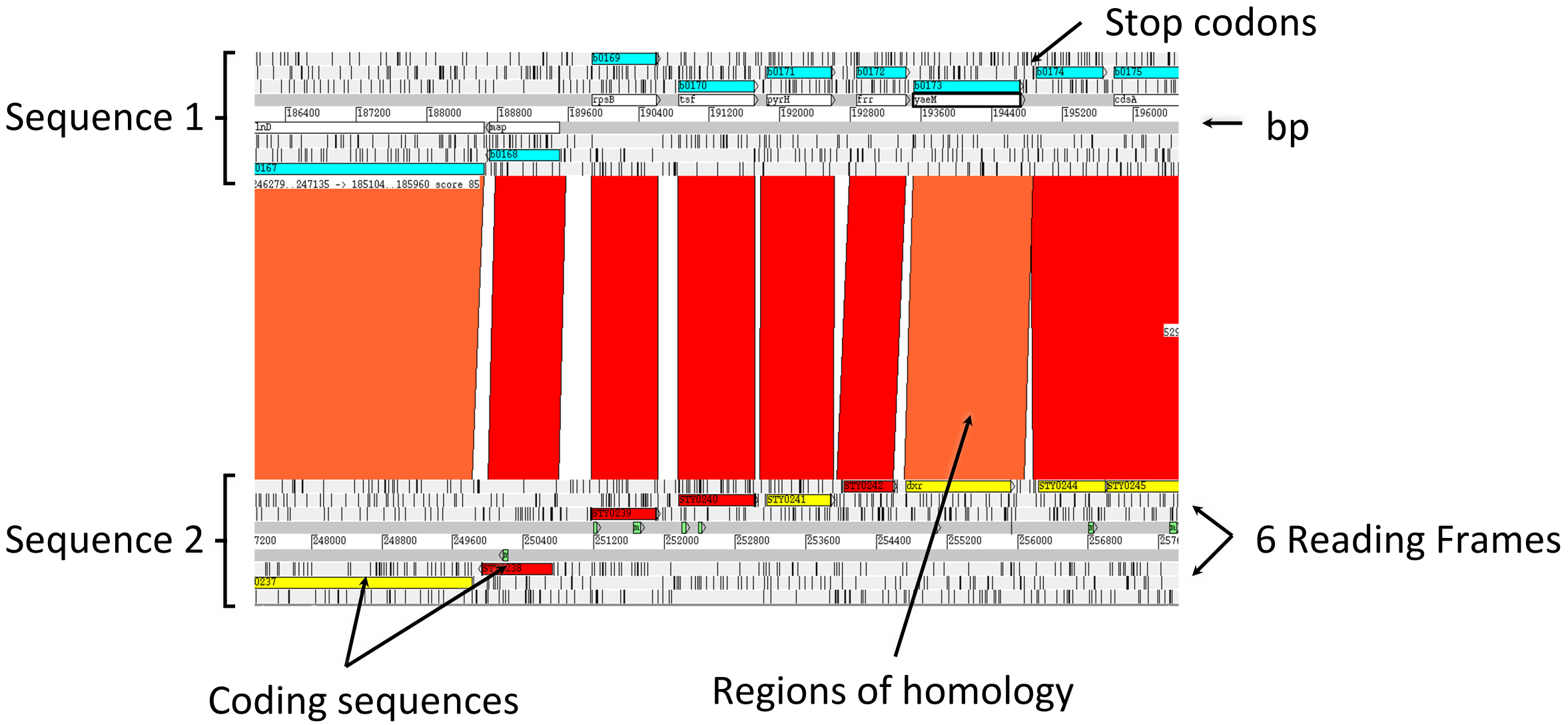


- Freely available (unix, mac, win)
  - WEB: http://www.sanger.ac.uk/science/tools/artemis-comparison-tool-act
  - MANUAL: ftp://ftp.sanger.ac.uk/pub/resources/software/act/act.pdf
  - PAPER: Carver et al. (2005) Bioinformatics. 21;16;3422-3

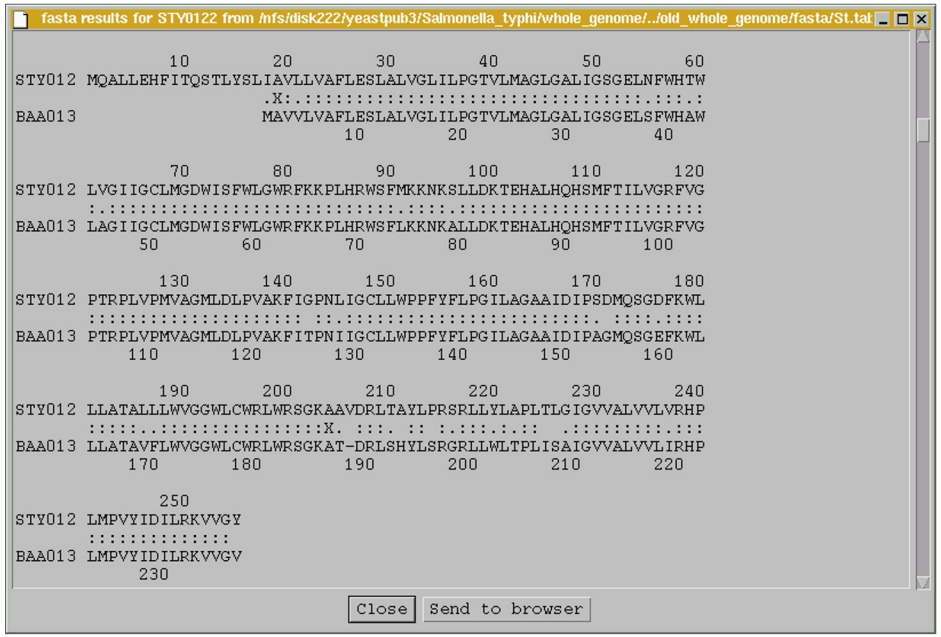
What do you need to run ACT?

- 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

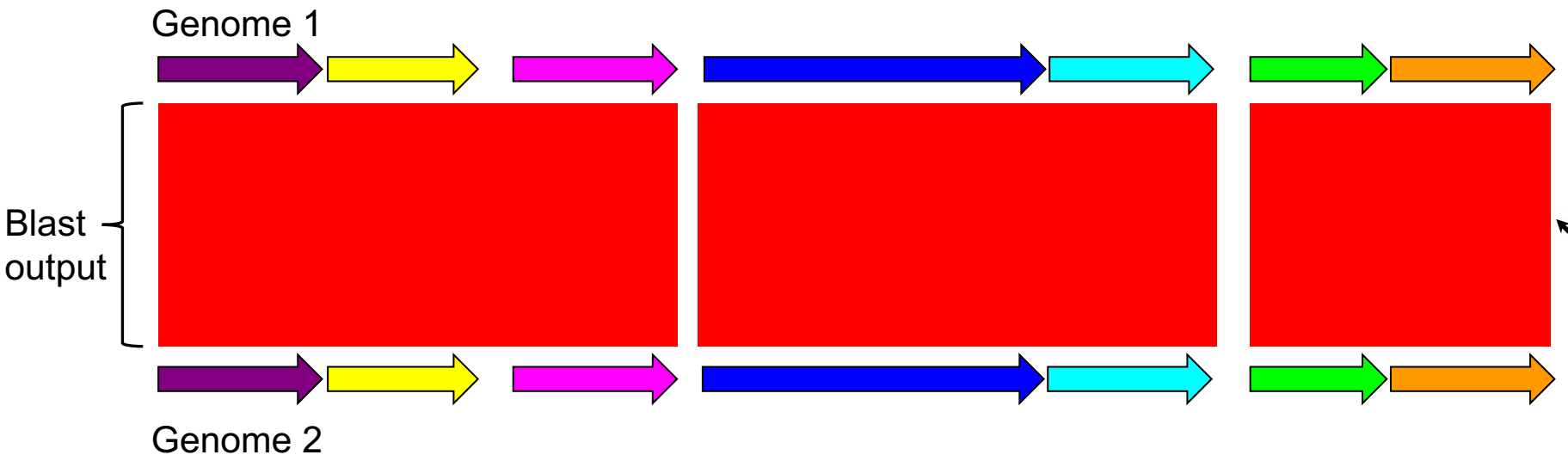
Basic setup of an ACT session



Visualising sequence similarity



Visualising genome rearrangements

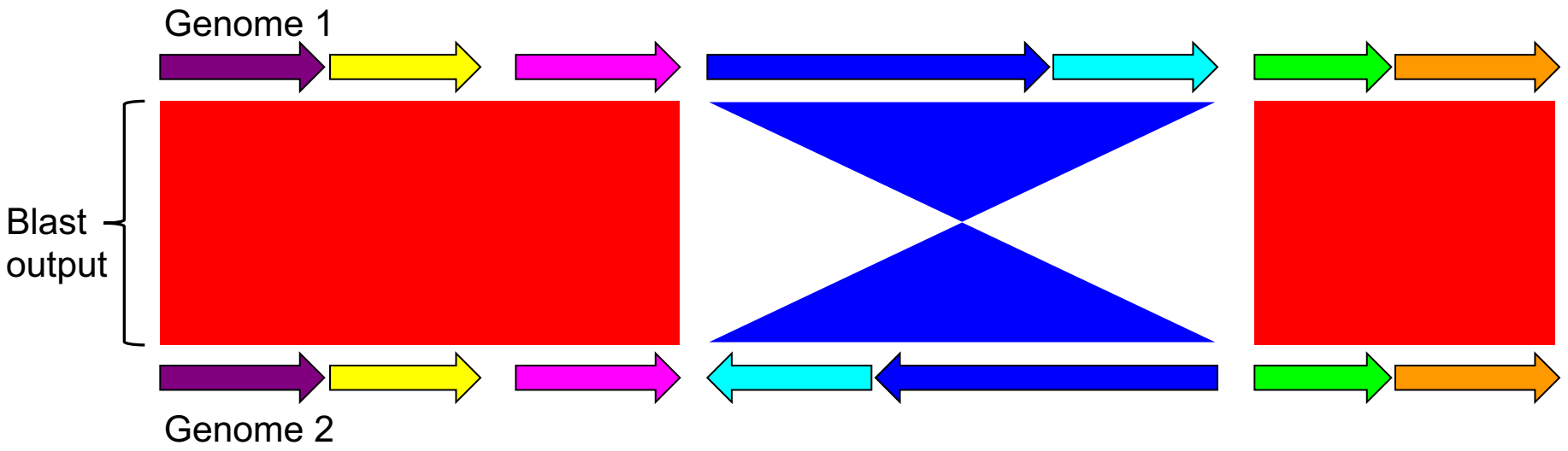


Region of homology in:

- Same orientation
- Reverse orientation

Same genes present in the same orientation

Visualising genome rearrangements

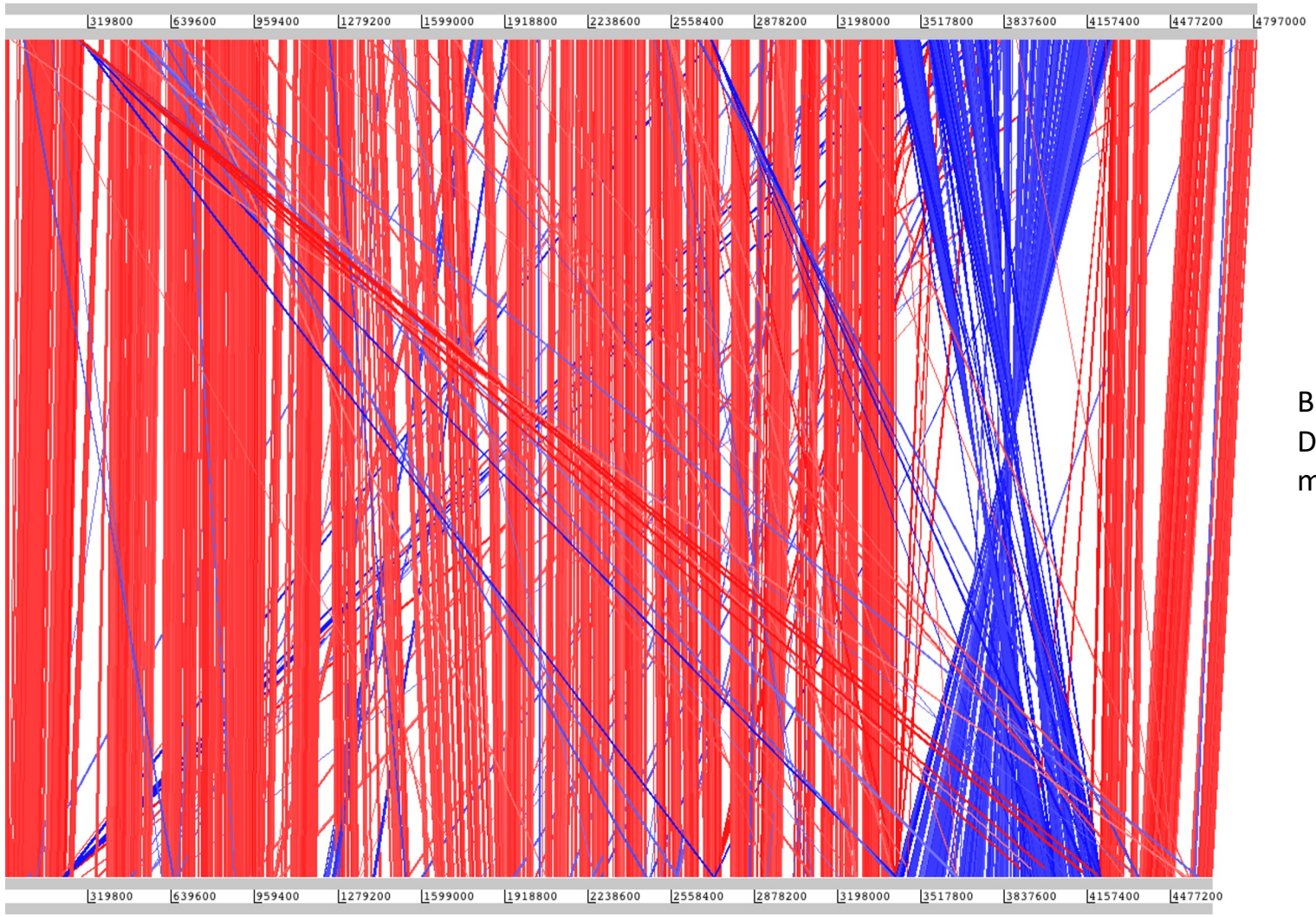


Region of homology in:

- Same orientation
- Reverse orientation

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

Visualising genome rearrangements

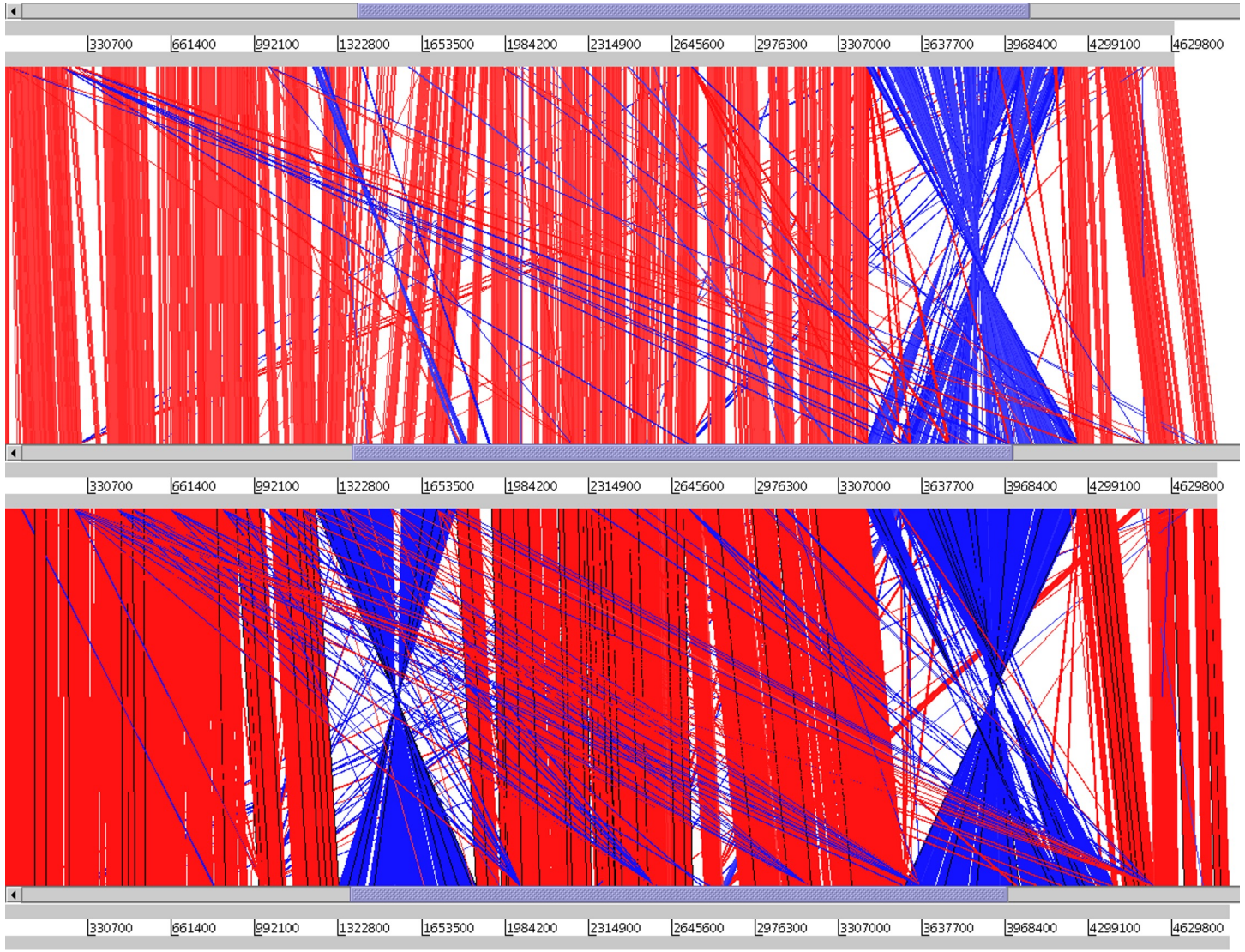


S. typhi

BLASTN
DNA-DNA
matches

E. coli

Visualising genome rearrangements

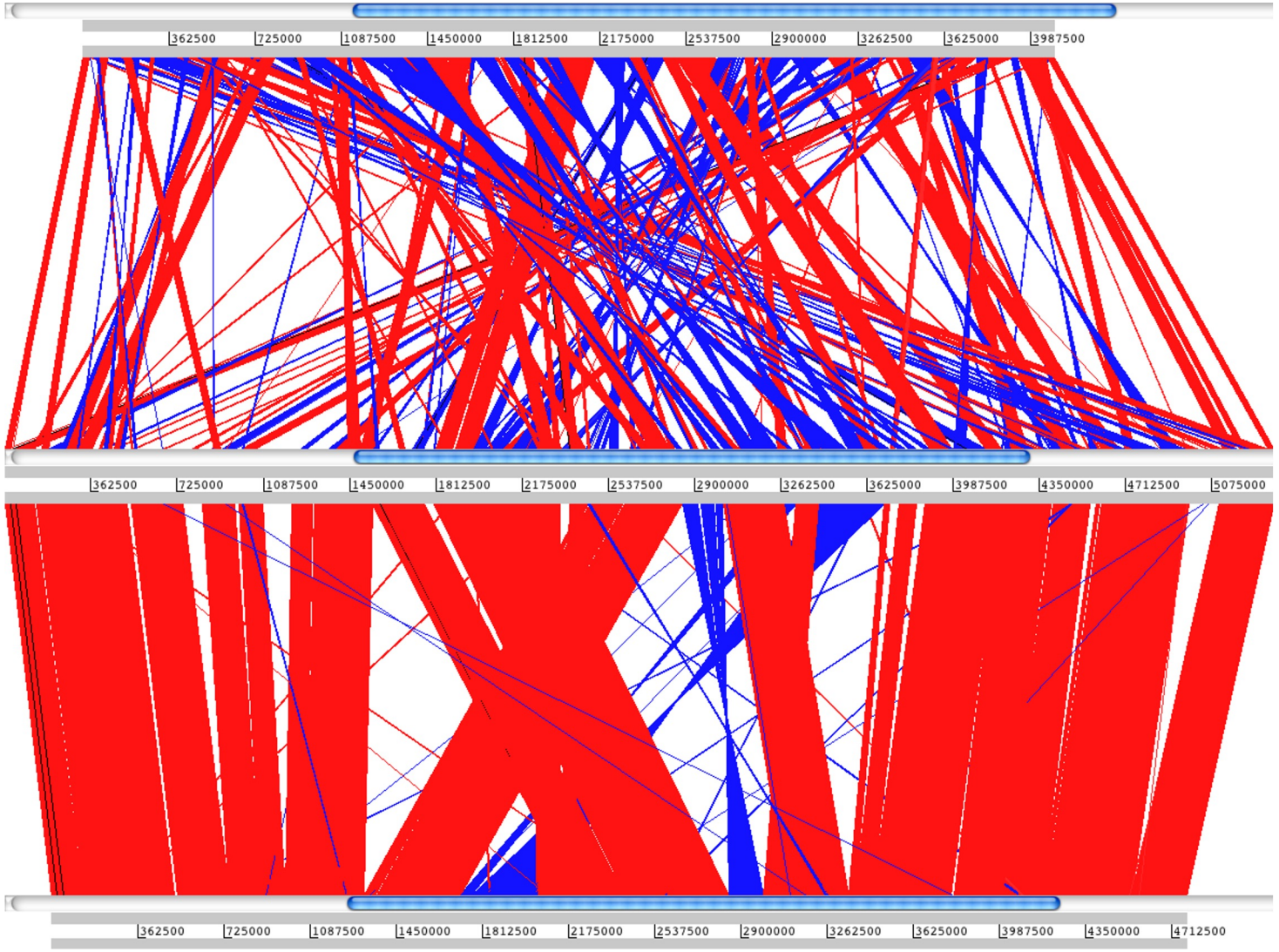


E. coli

S. typhi

S. typhimurium

Visualising genome rearrangements: *Bordetella spp.*

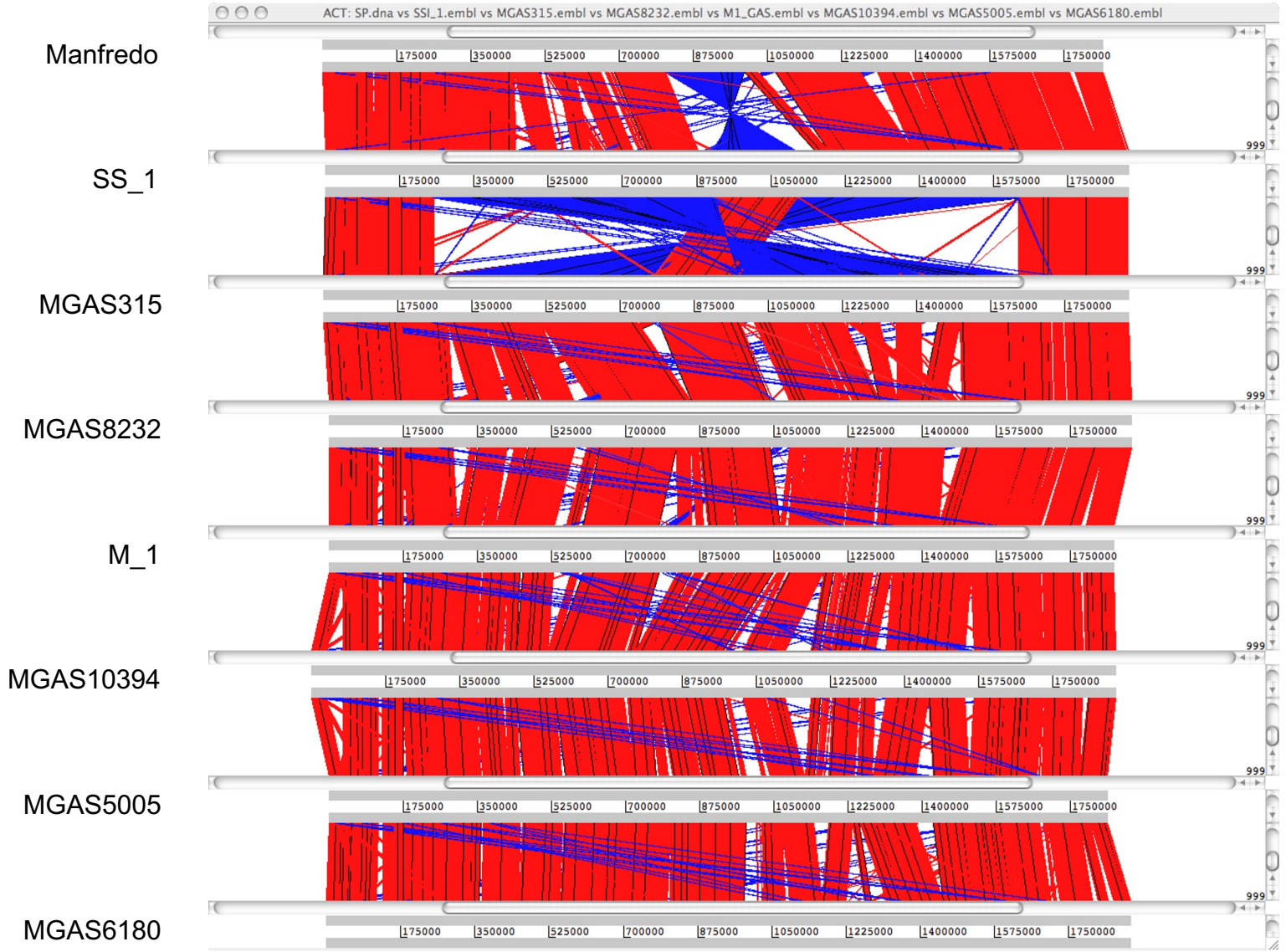


Bordetella pertussis

B. bronchiseptica

B. parapertussis

Comparison of strain diversity: *Streptococcus pyogenes*



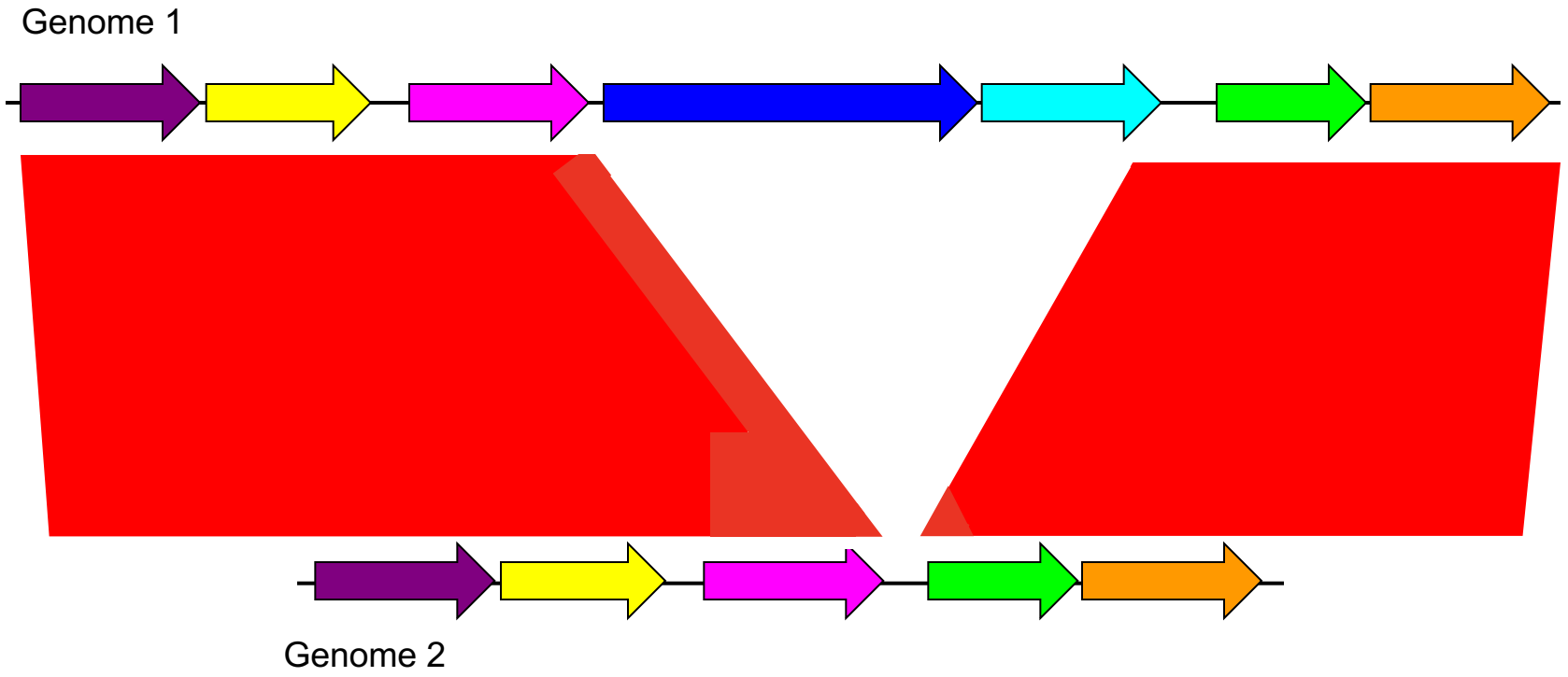
Comparison of genic diversity

- Gene gain and/or loss

  Duplication, horizontal gene transfer, indels

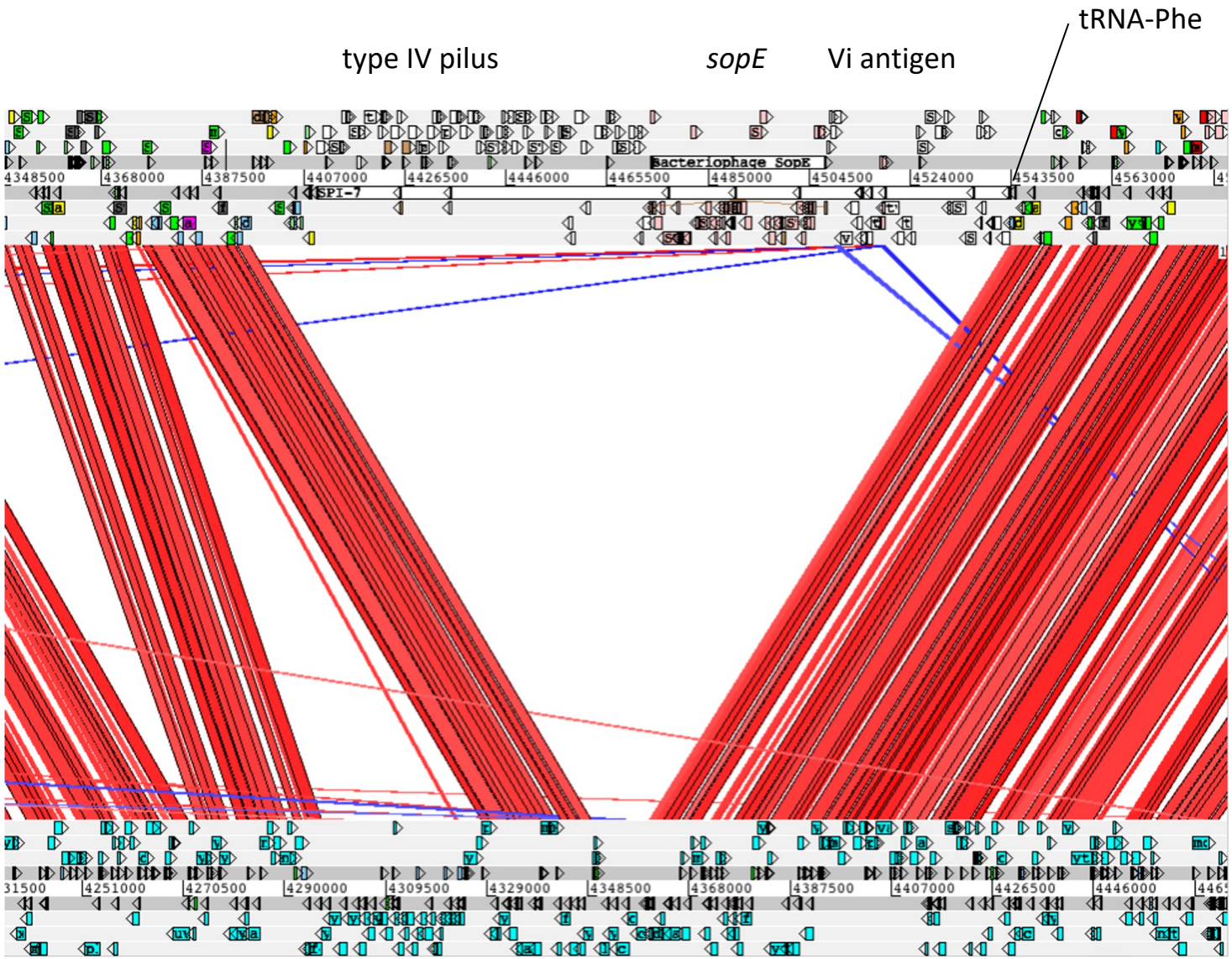
- Gene function modification
  - Loss of function (pseudogenes, gene fission)
  - Gain of function (gene fusion)
  - Mutation accumulation (SNPs)

Visualising gene gain or loss



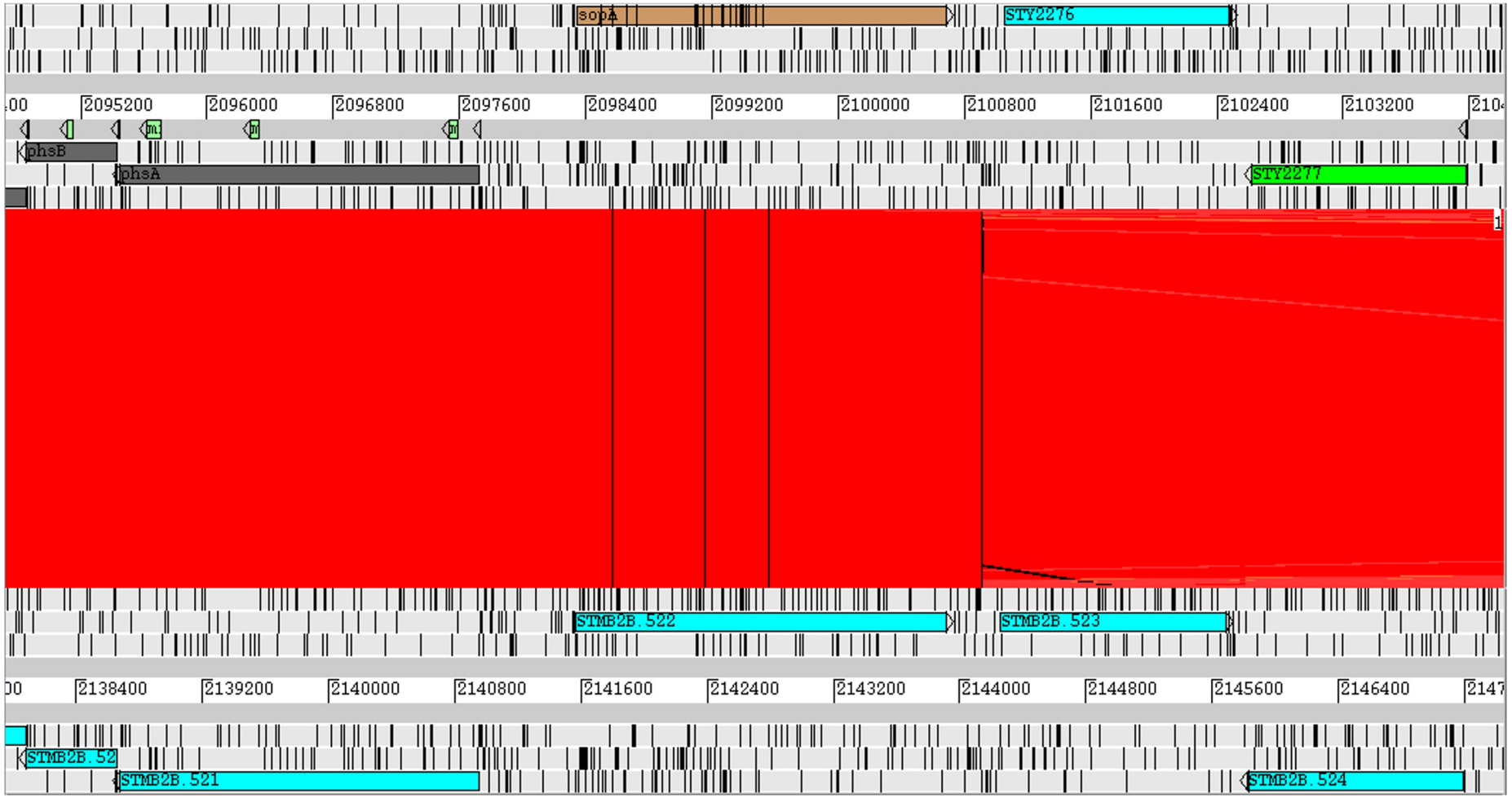
Example of an insertion or deletion

Visualising gene gain or loss

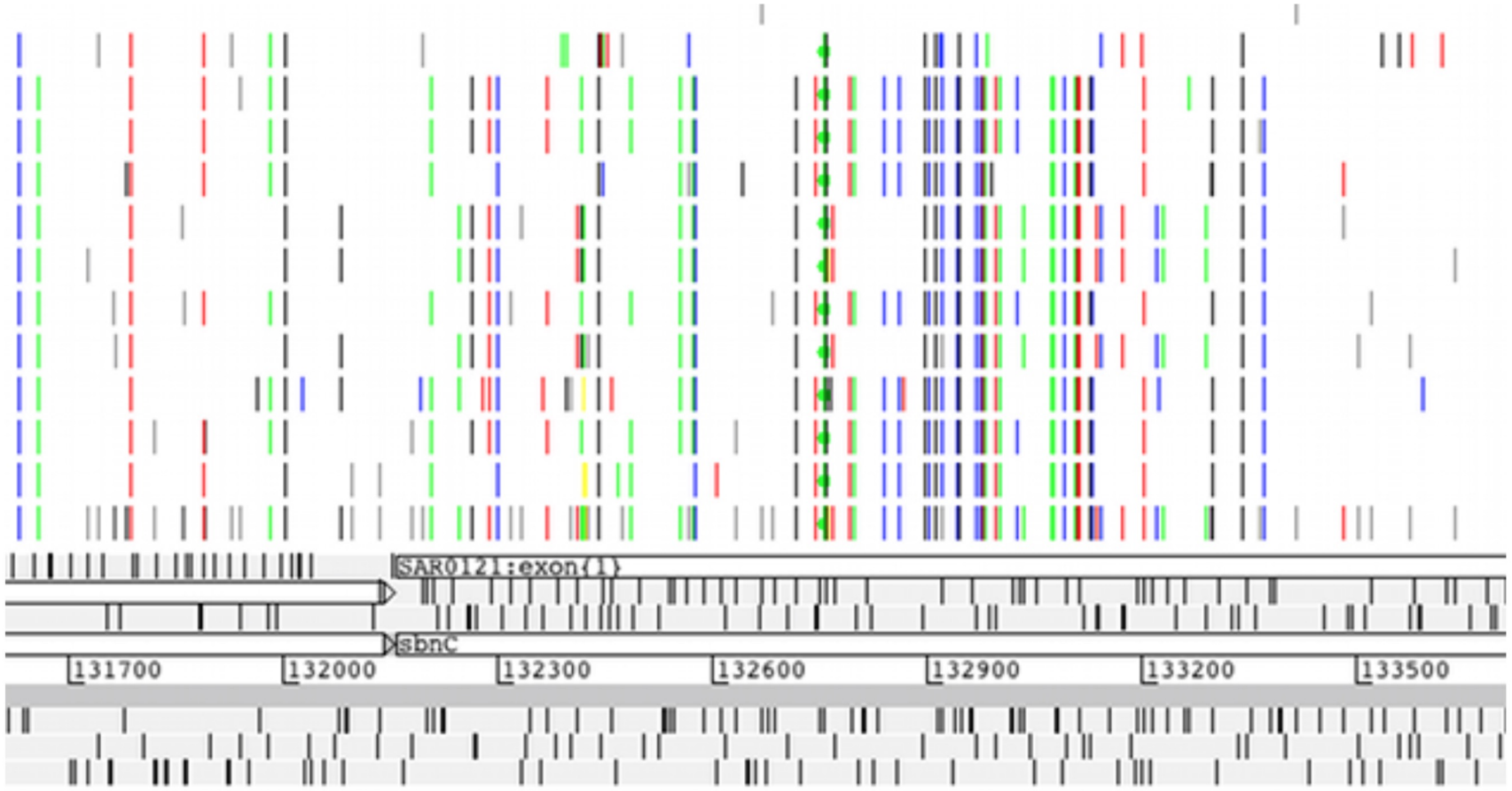


Loss of function: pseudogenes

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