

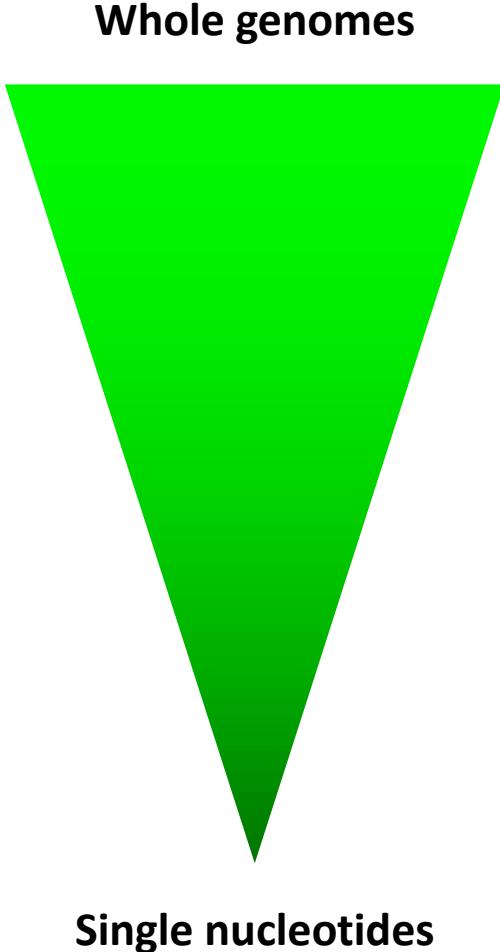
# **Module 3:**

# **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

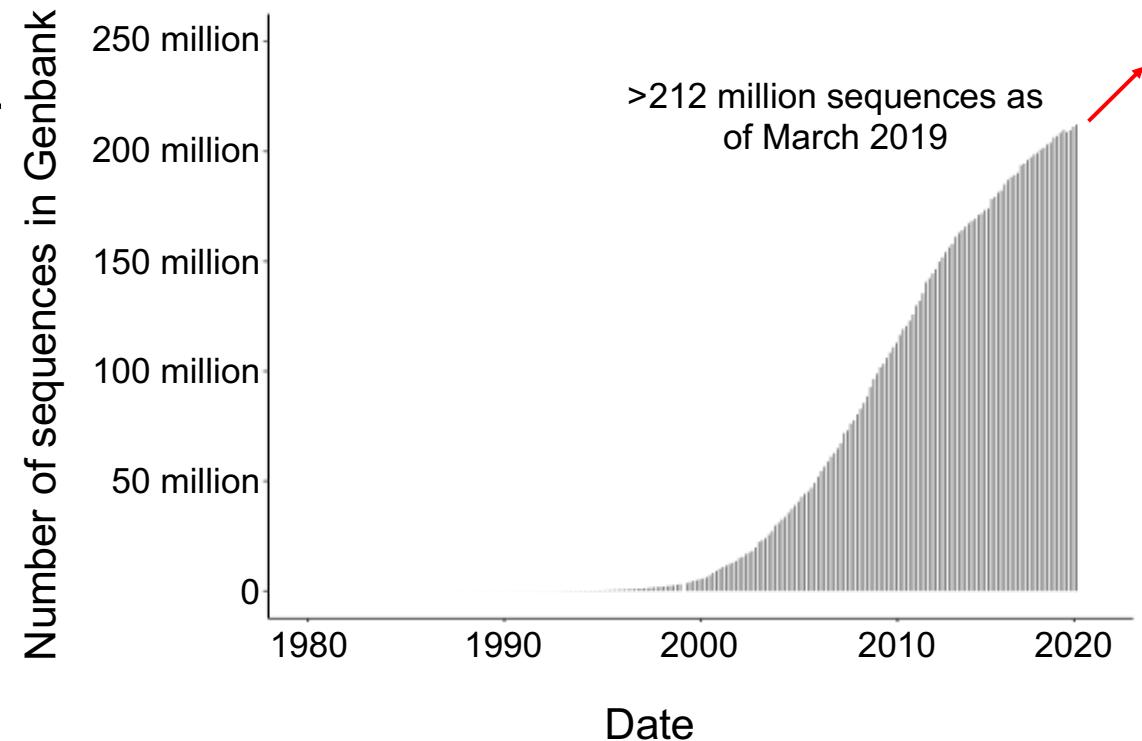


Useful over a wide variety of scales:

- 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  MNPNQKIIITIGSVCMTIGMANLILQIGNIISIWISHSIQLGNQNQIETCNQSIVITYENNT 60
          MNPNQKIIITIGSVCMTIGMANLILQIGNIISIWISHSIQLGNQNQIETCNQSIVITYENNT
Sbjct: 1   MNPNQKIIITIGSVCMTIGMANLILQIGNIISIWISHSIQLGNQNQIETCNQSURITYENNT 60

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

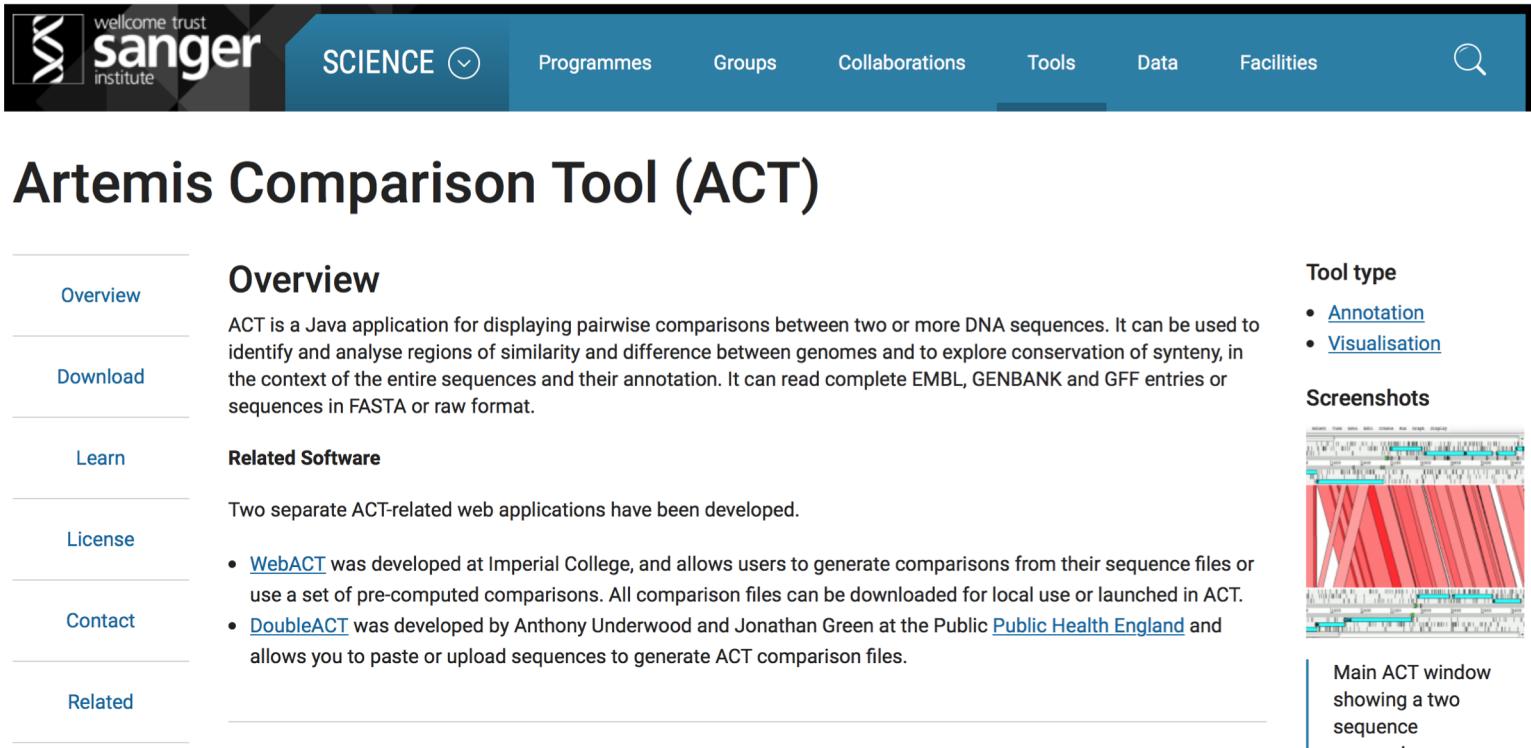
nucmer

[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?



The screenshot shows the Wellcome Trust Sanger Institute Science website. The top navigation bar includes links for Programmes, Groups, Collaborations, Tools, Data, Facilities, and a search icon. The main content area features a large heading 'Artemis Comparison Tool (ACT)' and a sidebar with links for Overview, Download, Learn, License, Contact, and Related. The main content area has two columns: 'Overview' (describing ACT as a Java application for pairwise comparisons between DNA sequences) and 'Related Software' (mentioning WebACT and DoubleACT). To the right, there's a 'Tool type' section with links to Annotation and Visualisation, and a 'Screenshots' section showing a screenshot of the ACT software interface.

## Artemis Comparison Tool (ACT)

[Overview](#)

[Download](#)

[Learn](#)

[License](#)

[Contact](#)

[Related](#)

### Overview

ACT is a Java application for displaying pairwise comparisons between two or more DNA sequences. It can be used to identify and analyse regions of similarity and difference between genomes and to explore conservation of synteny, in the context of the entire sequences and their annotation. It can read complete EMBL, GENBANK and GFF entries or sequences in FASTA or raw format.

### Related Software

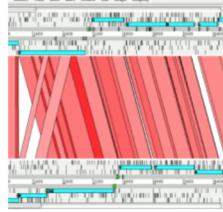
Two separate ACT-related web applications have been developed.

- [WebACT](#) was developed at Imperial College, and allows users to generate comparisons from their sequence files or use a set of pre-computed comparisons. All comparison files can be downloaded for local use or launched in ACT.
- [DoubleACT](#) was developed by Anthony Underwood and Jonathan Green at the Public [Public Health England](#) and allows you to paste or upload sequences to generate ACT comparison files.

### Tool type

- [Annotation](#)
- [Visualisation](#)

### Screenshots



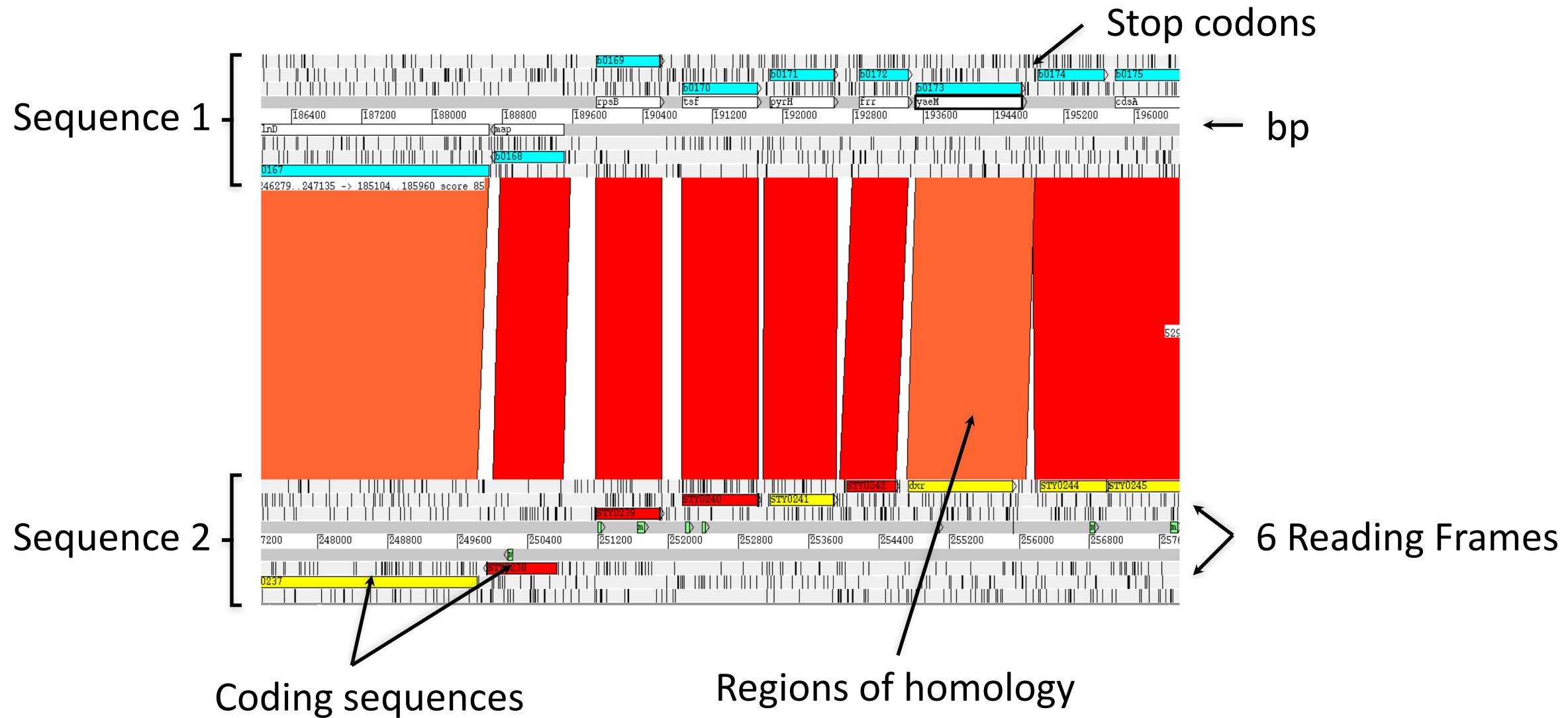
Main ACT window showing a two sequence

- 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

fasta results for STY0122 from /nfs/disk222/yeastpub3/Salmonella\_typhi/whole\_genome//old\_whole\_genome/fasta/St.tat

10	20	30	40	50	60
STY012	MQALLEHFITQSTLYSLIAVLLVAFLES	LAvgLILPGTVLMAGLGA	LIGSGELNF	WHTW	
BAA013	.X.:.....	MAVV	LVAFLES	LAvgLILPGTVLMAGLGA	LIGSGELSF
		10	20	30	40
70	80	90	100	110	120
STY012	LVGIIGCLMGDWISFWLGWRFKKPLHRWSFMKKNS	LDDKTEHALHQHS	MFTILVGR	FVG	
BAA013	LAGIIGCLMGDWISFWLGWRFKKPLHRWSFLKKNK	ALLDKTEHALHQHS	MFTILVGR	FVG	
	50	60	70	80	90 100
130	140	150	160	170	180
STY012	PTRPLVPMVAGMLDLPVAKPIGPNLIGC	LLWPPFYFLPG	GILAGAAIDIP	SDMQSGDF	KWL
BAA013	PTRPLVPMVAGMLDLPVAKPITPN	IIGC	LLWPPFYFLPG	GILAGAAIDIP	AGMOSGEFKWL
	110	120	130	140	150 160
190	200	210	220	230	240
STY012	LLATALLLWVGGWL	CWRLWERSGKAA	VDRLTAYLPRSR	LTLGIGVV	VALVVLVRHP
BAA013	LLATAVFLWVGGWL	CWRLWERSGKAT	-DRLSHYLSRG	RLLWLTP	ISAIGVVALVVLIRHP
	170	180	190	200	210 220
250					
STY012	LMPVYIDILRKVVGY				
BAA013	LMPVYIDILRKVVGV				
	230				

MQALL...

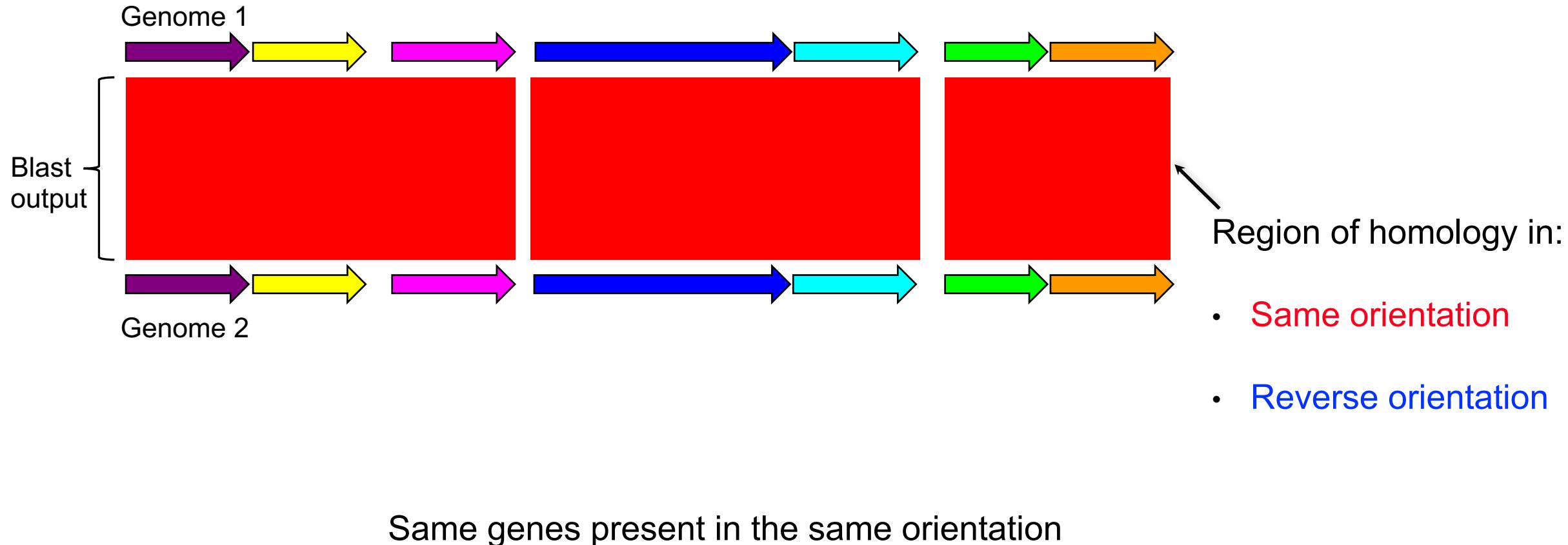
...RHP

Region sharing similarity  
(BLASTP)

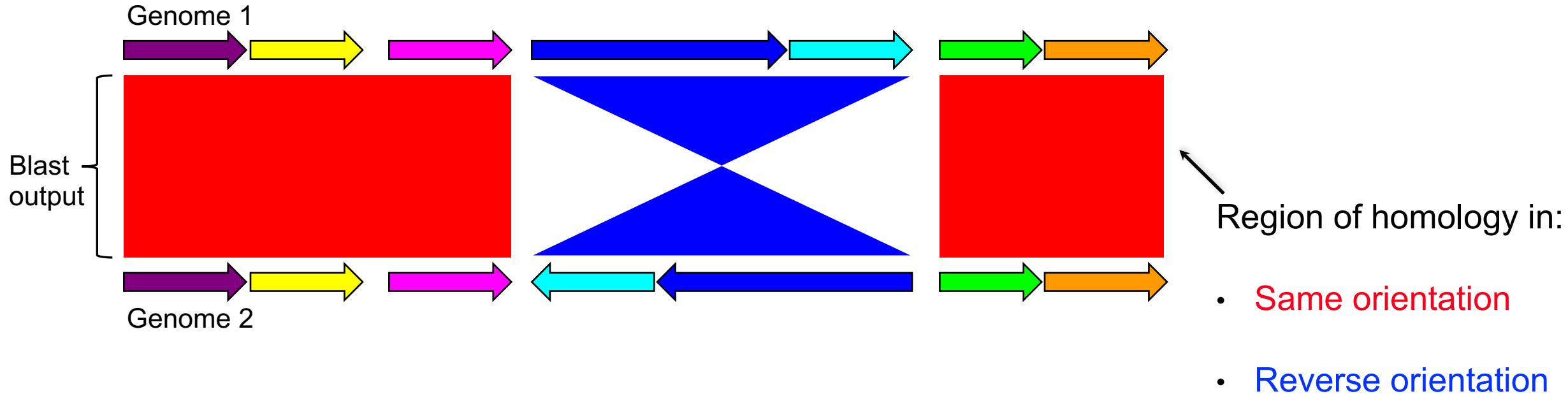
MAVV...

...RHP

# Visualising genome rearrangements

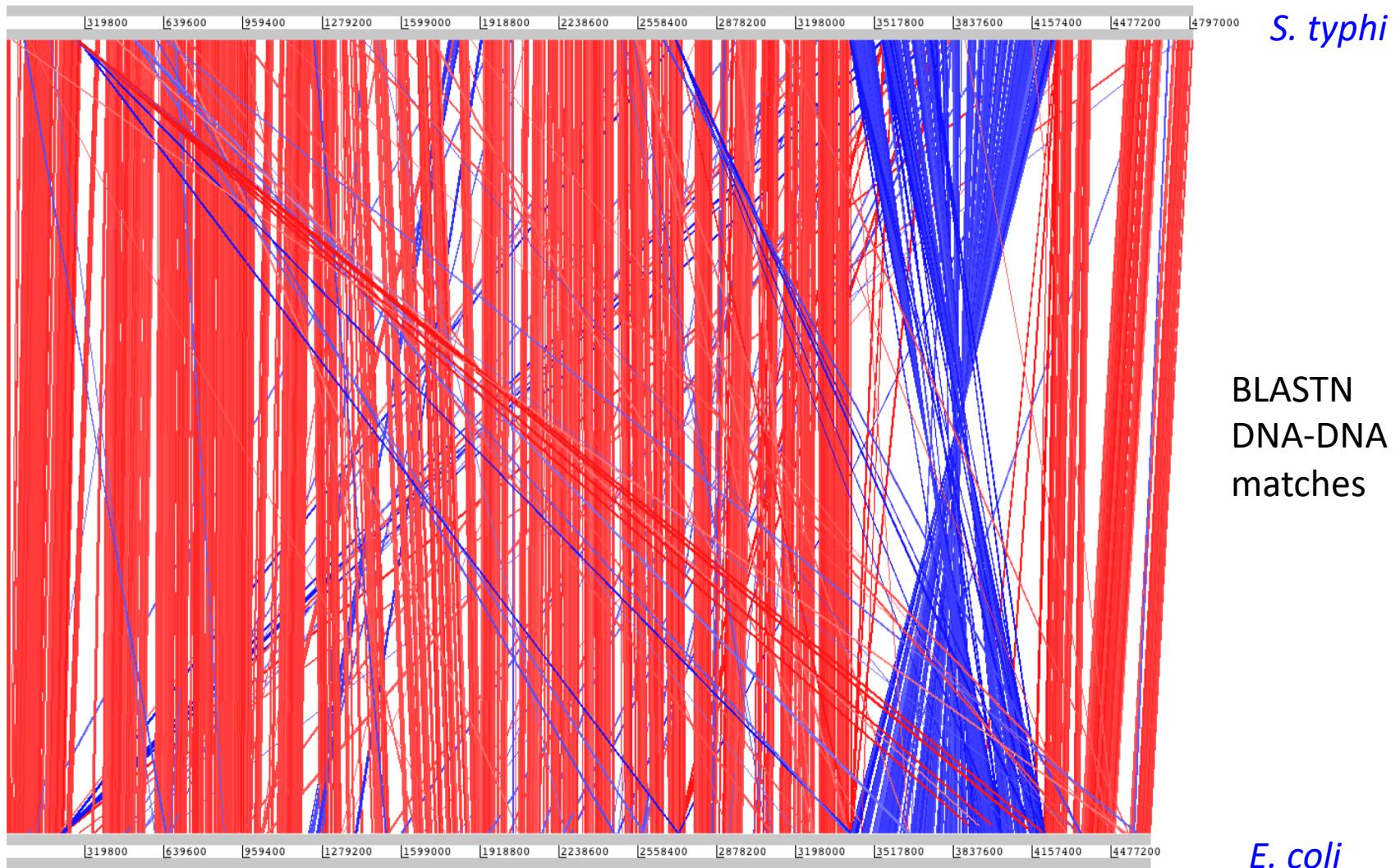


# Visualising genome rearrangements

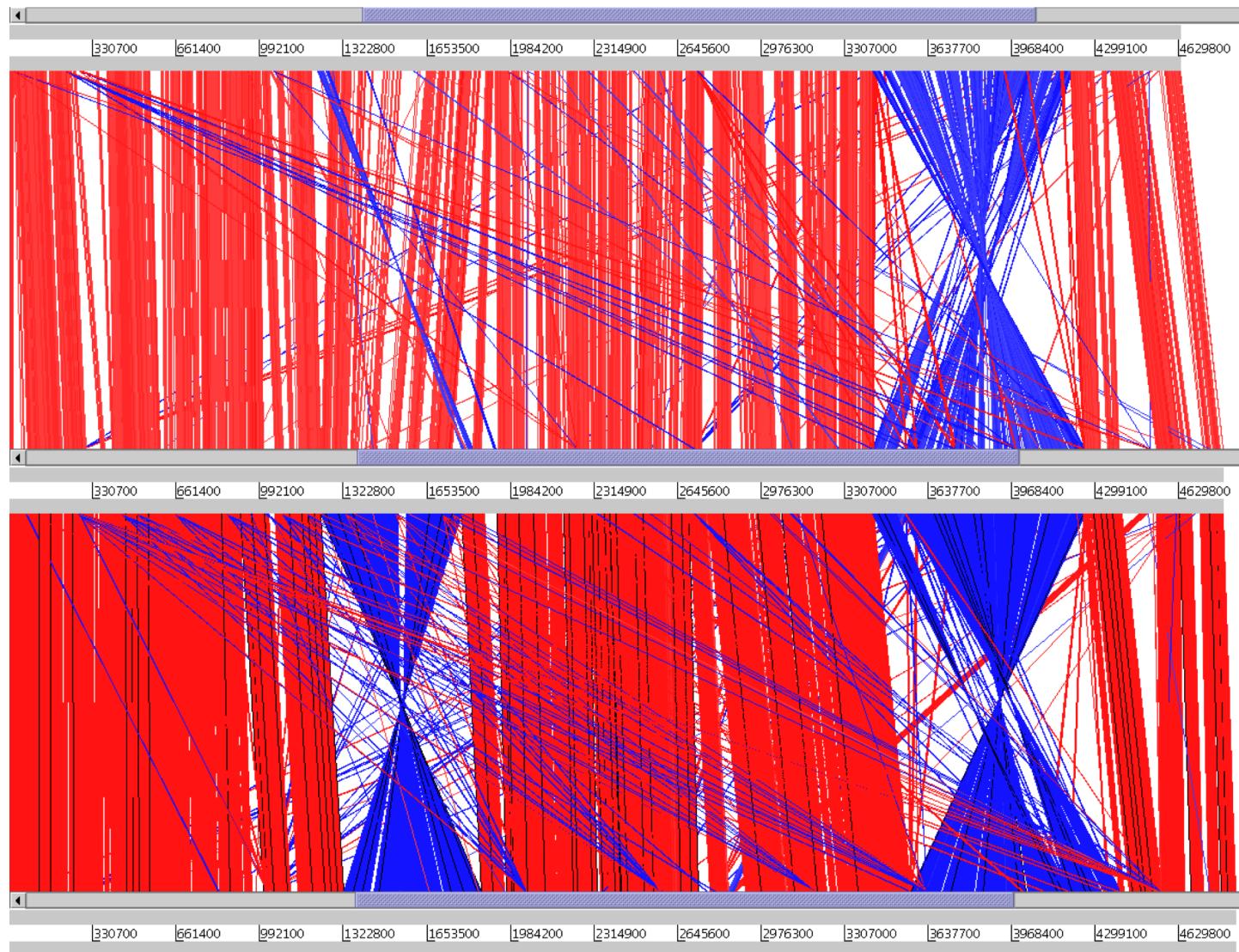


The same genes are present but the blue genes have undergone a rearrangement

# Visualising genome rearrangements



# Visualising genome rearrangements

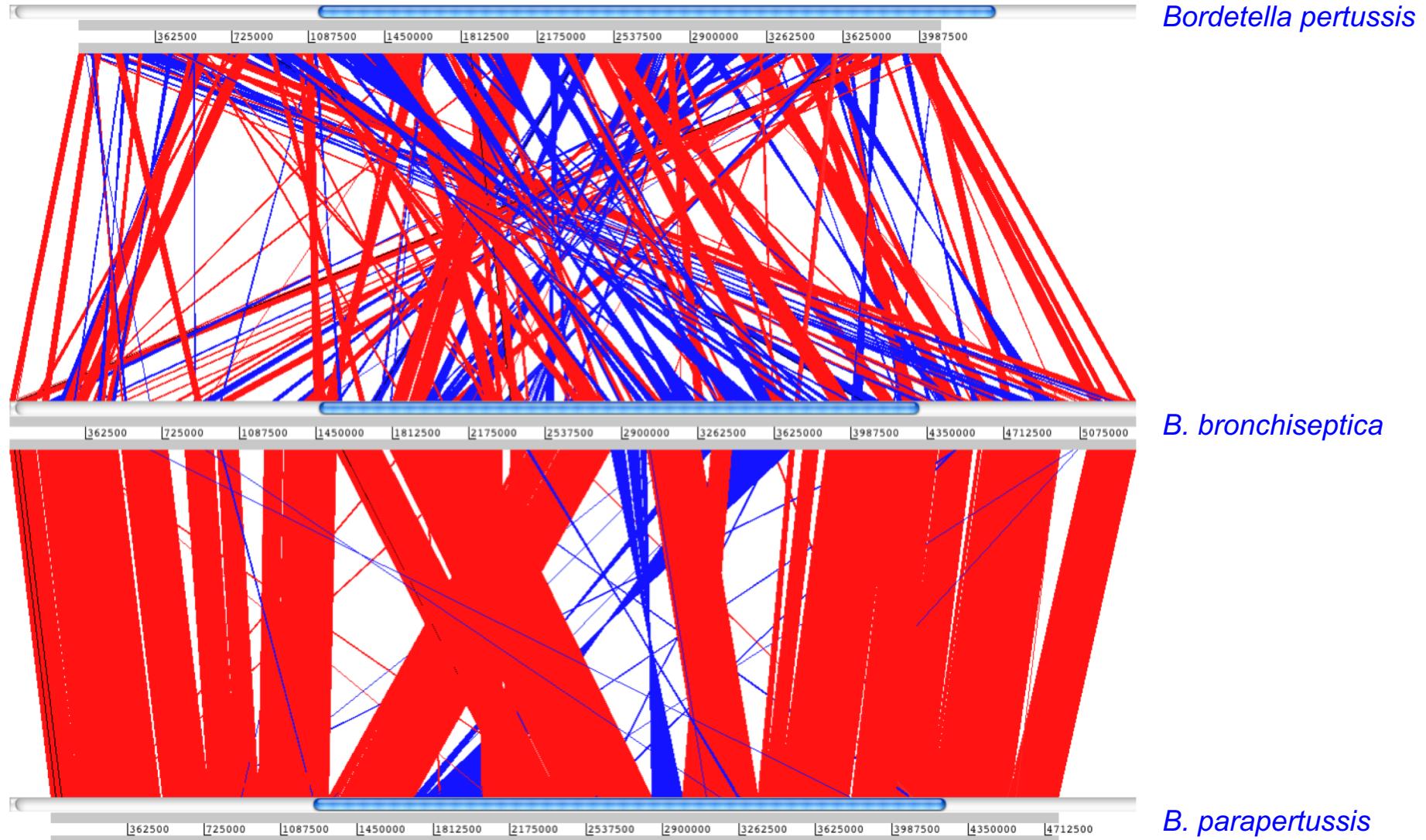


*E. coli*

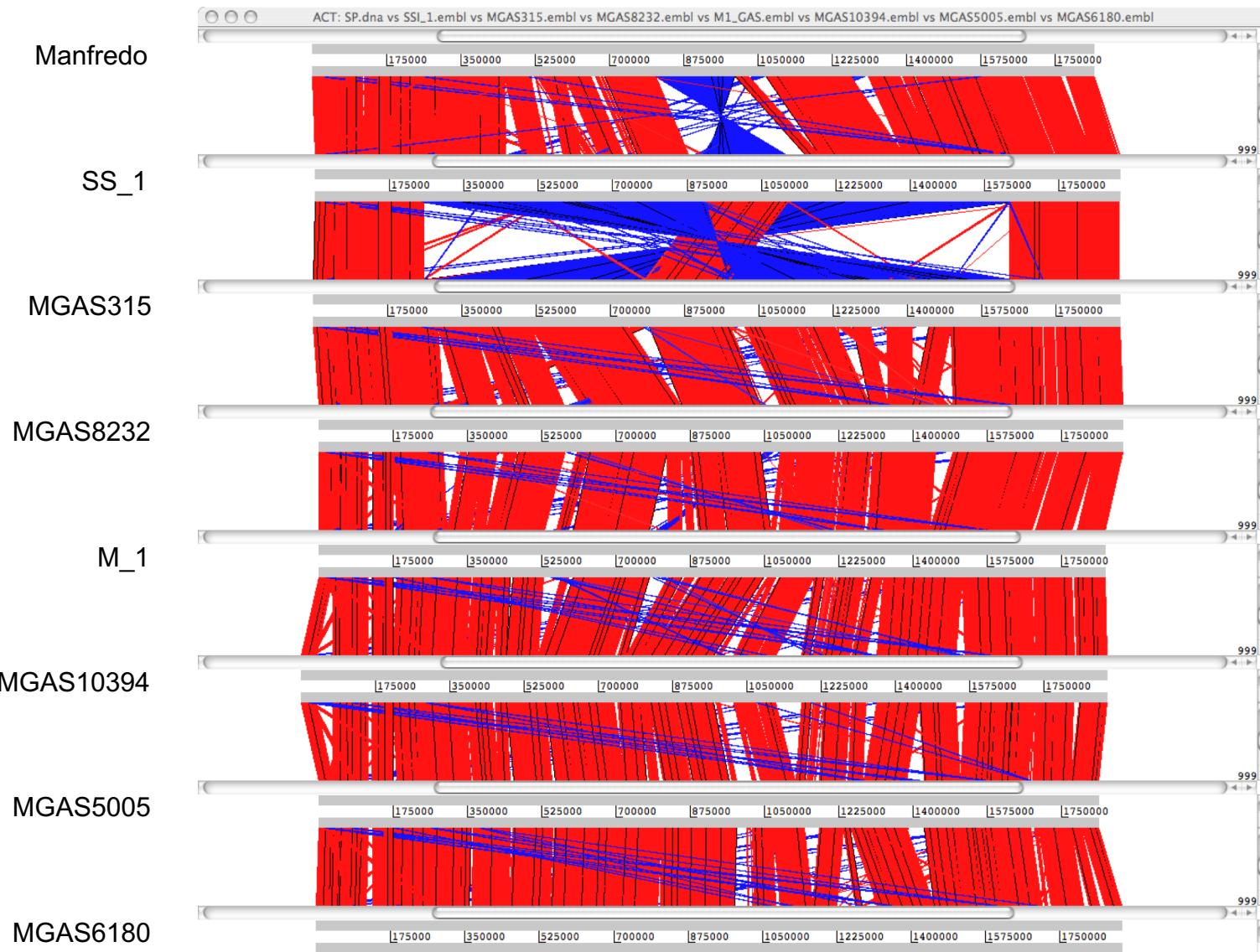
*S. typhi*

*S. typhimurium*

# Visualising genome rearrangements: *Bordetella* spp.



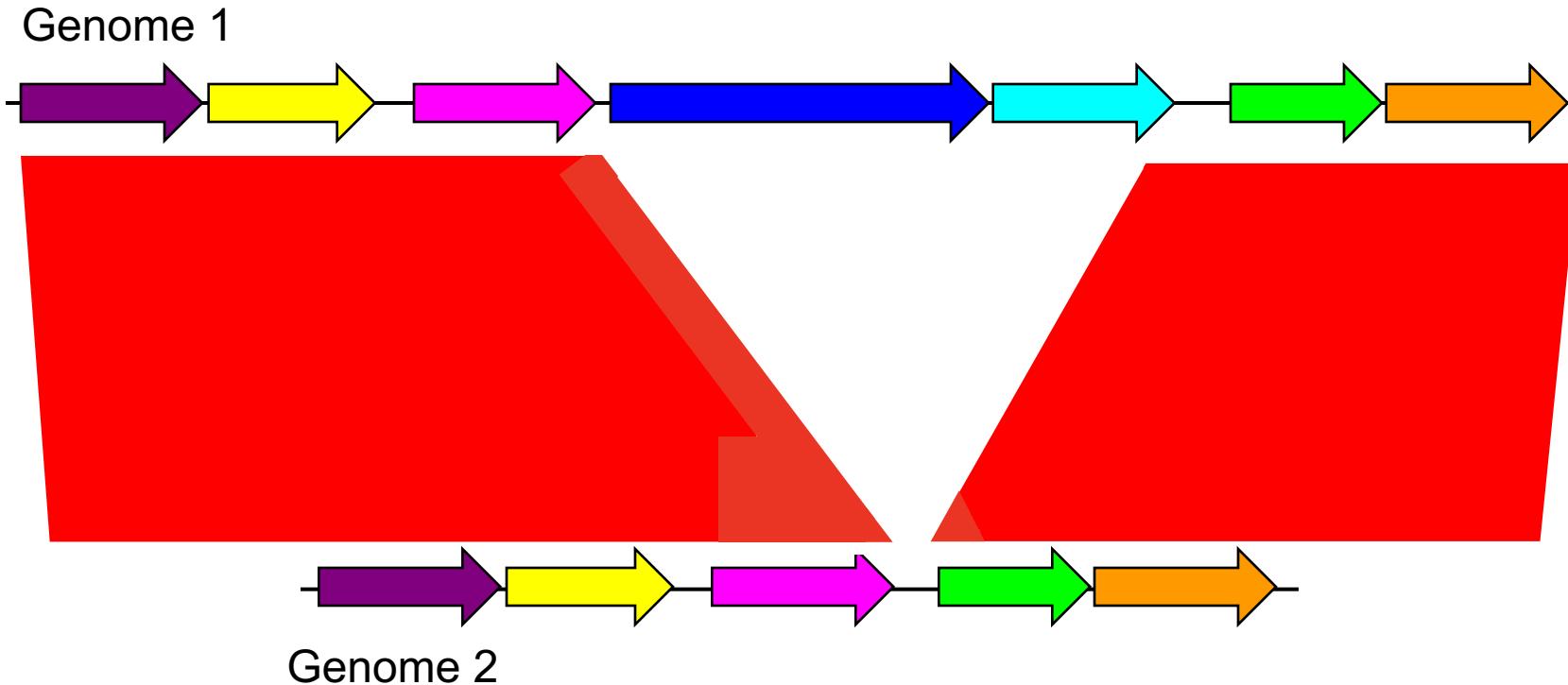
# Comparison of strain diversity: *Streptococcus pyogenes*



# Comparison of genic diversity

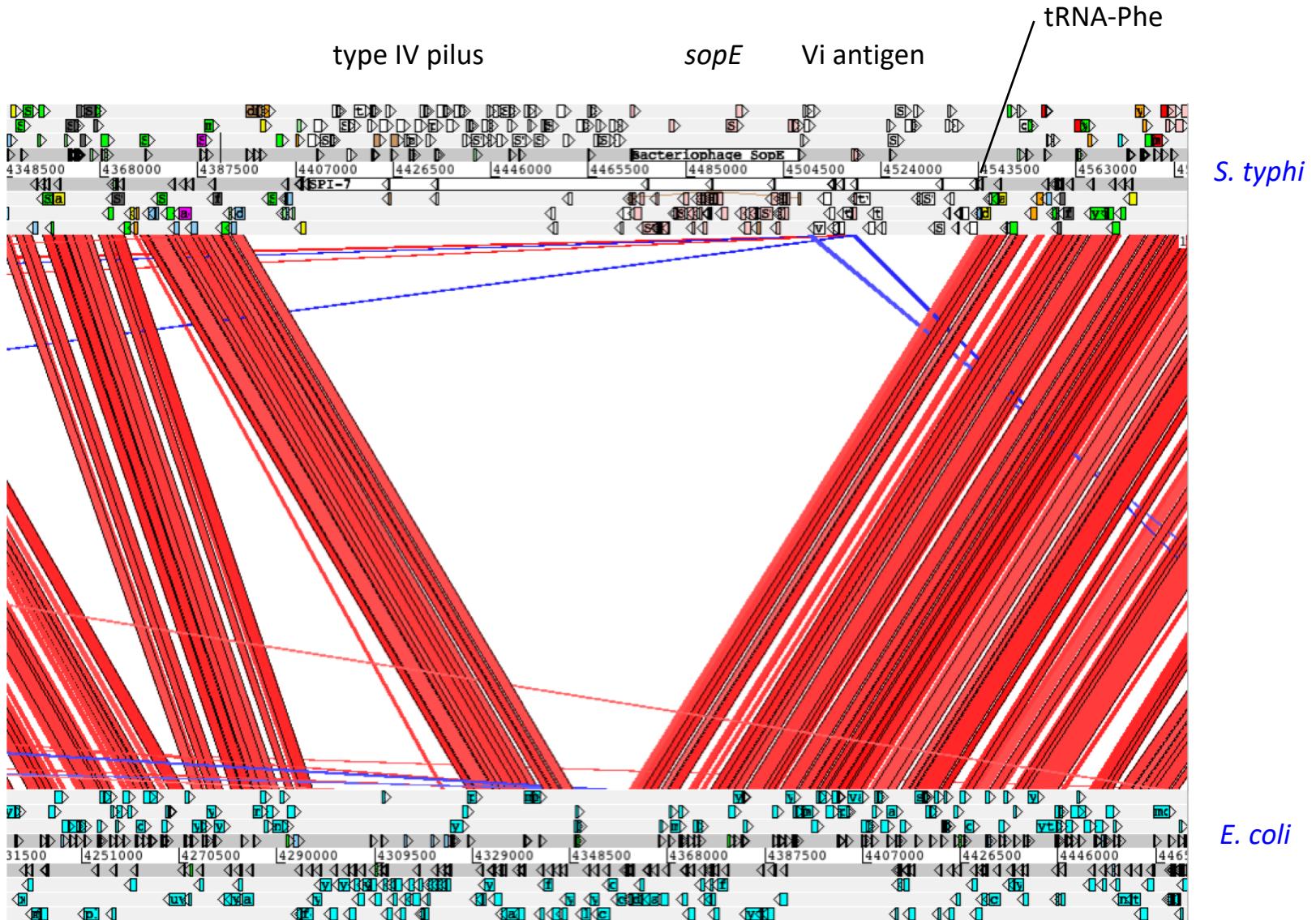
- Gene gain and/or loss
  - Duplication, horizontal gene transfer, indels
- Gene function modification
  - Loss of function (pseudogenes, gene fission)
  - Gain of function (gene fusion)
  - Mutation accumulation (SNPs)

# Visualising gene gain or loss

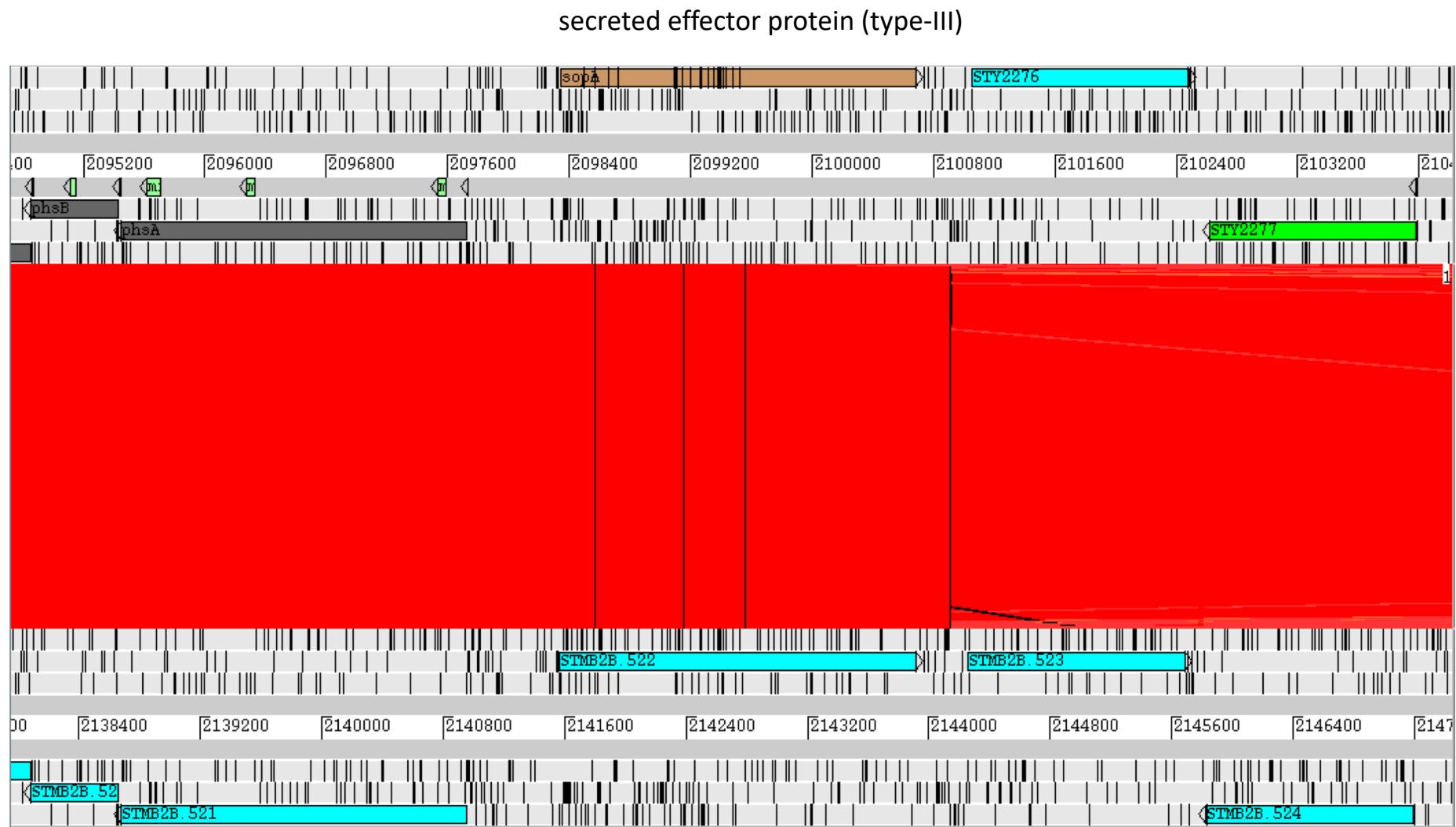


Example of an insertion or deletion

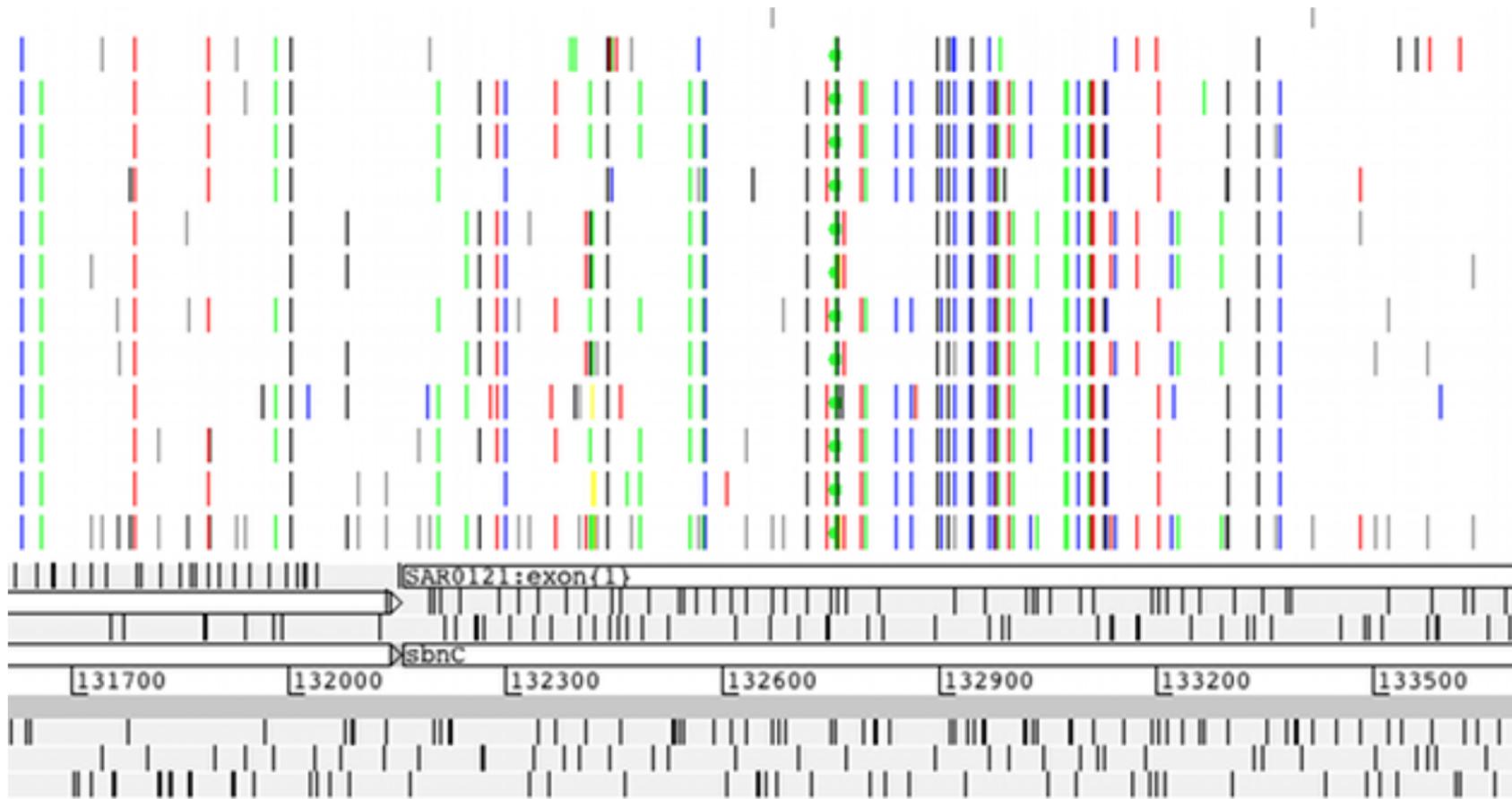
# Visualising gene gain or loss



# Loss of function: pseudogenes



# Genetic variation among populations



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

# Aims of this module

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