Microbial Genome Sequence Analysis

Multiple Genome Alignment with Mauve

06 June 2024

Phuc Loi Luu, PhD

Luu.p.loi@googlemail.com

Content

- Brief introduction of Microbial Genome Analysis
- Multiple Genome Alignment with Mauve
- Homework

General features of genomes

Microbial

- Small WSIWYG genomes (Mbp)
- Gene density high (>90%)
  - intergenic regions short
  - very little repetitive or non-coding DNA
  - Introns very rare
- Protein-coding genes (CDS) short (~1kbp)
- Operons with promoters just upstream
- Fewer non-coding RNAs

Human

- Very large genomes (Gbp)
- Gene density low
  - Only 25% is genes
  - Introns mean only1% codes
- Genes can span ≥30 kbp
- Genes have ~3 transcripts
  - Splicing and splice variants
- Promoter regions distant from gene

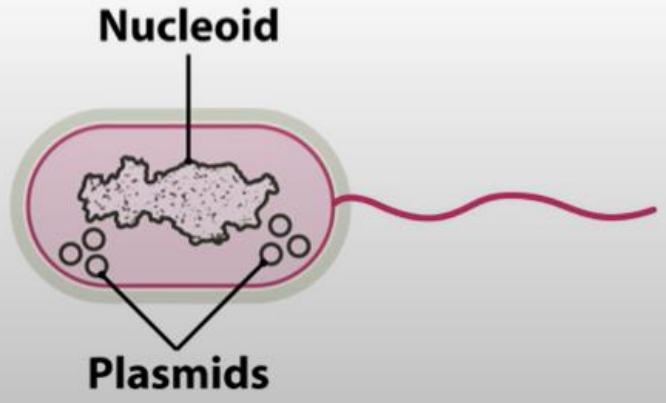
Bacterial genome organisation

Chromosomes

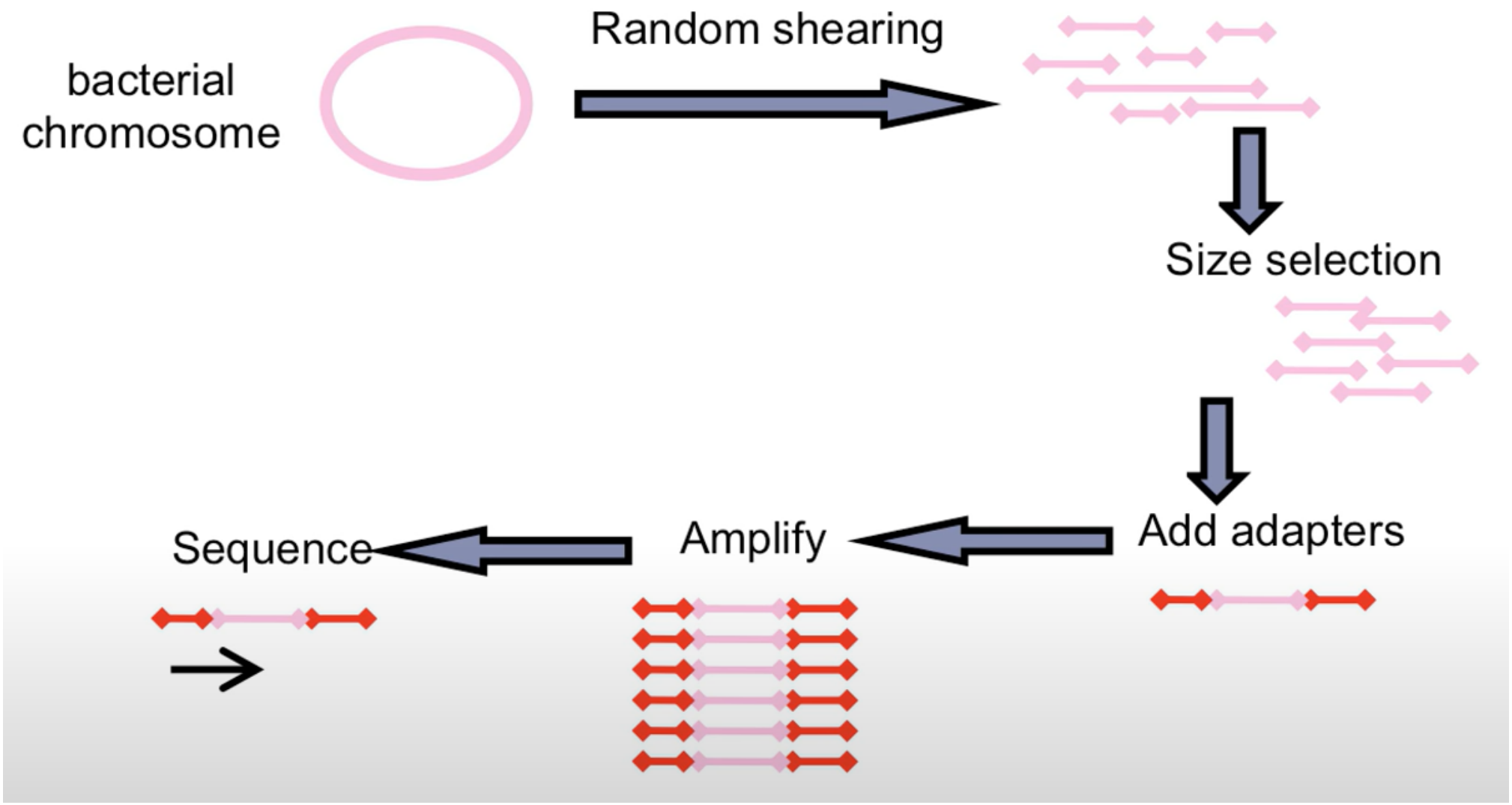
- Most commonly single circular chromosome (always DNA)
  - BUT many species have linear chromosome(s) (e.g. *Borrelia, Streptomyces, Rhodoccus*)
  - BUT a few species with two chromosomes (e.g. *Vibrio cholerae*)
- Can be mix of circular and linear (e.g. *Agrobacterium tumefaciens, B. burgdoferi*)

Plasmids

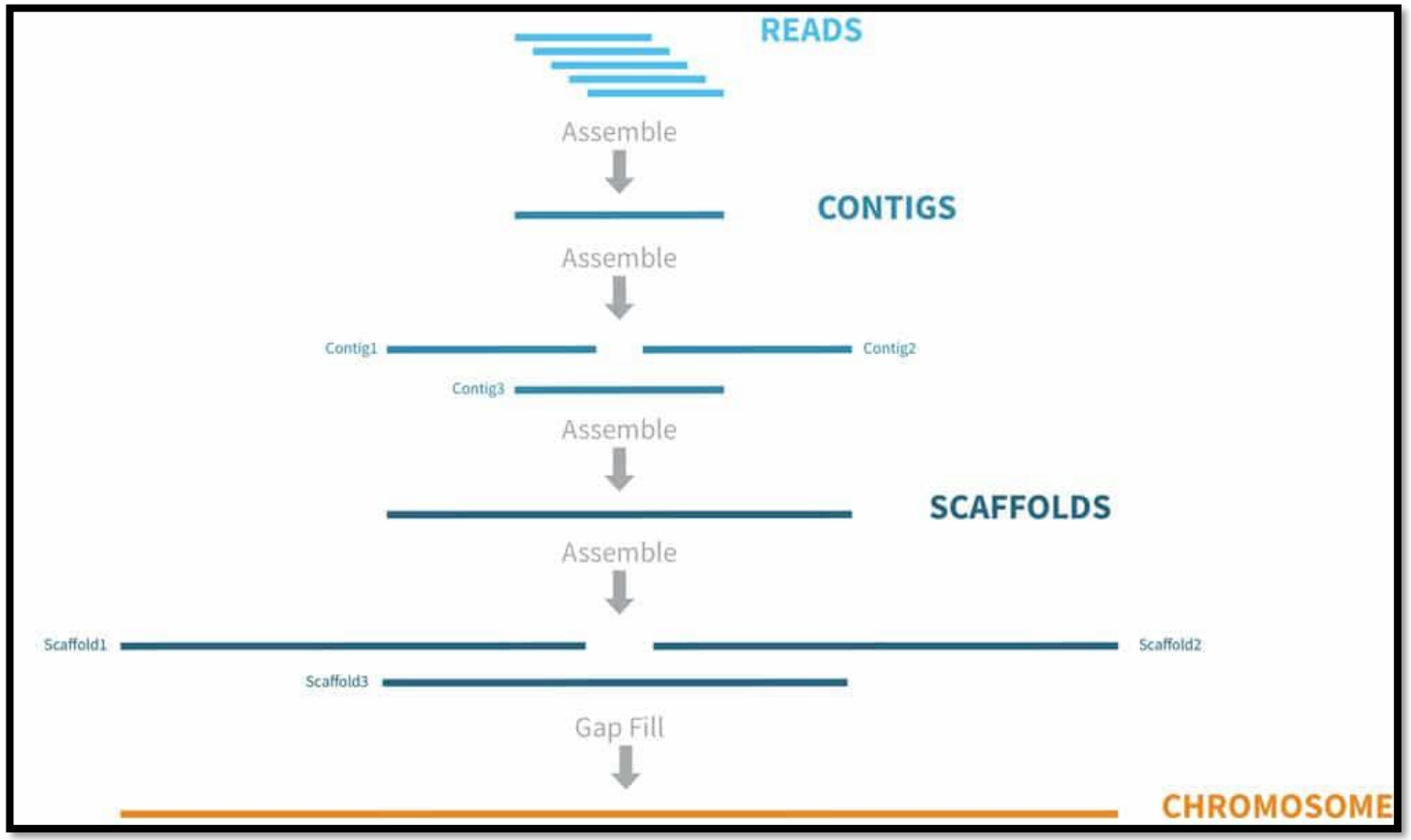
- Independent autonomous replicon, can be circular or linear
- may integrate into chromosome
- copy number varies 1 to 10s
- often carry non-essential genes that confer an adaptive advantage in certain conditions



Whole genome sequencing – Shotgun method



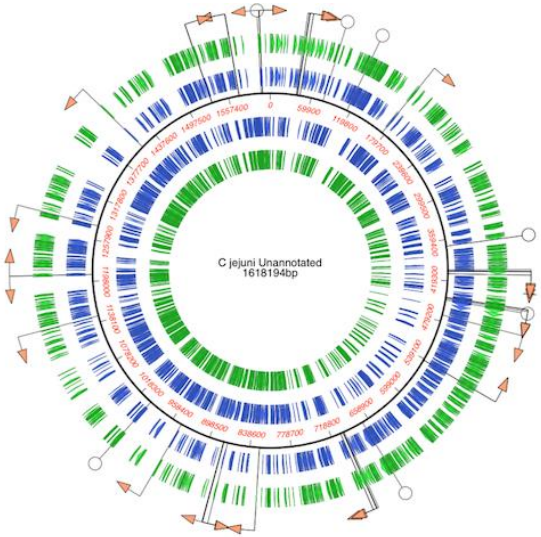
De novo assembly




```
48541 agcccttcaa agaaatgttc tcagcaggca tggagcccag gacttgctcc ctttgggtag
48601 agagccgggt tgaaggtgac tgaagtgaaa tgggacagta gaggcggggg gggtggtgag
48661 ttcctggagg tgggggtgt gggaacctgc tgtgtactga gatgcacccc tgccagttct
48721 gcctgaagat ttgaggcggg gggcaggggg gcggagtgaa gtcattttac tggtaagtaa
48781 ttttaaacct tttaatatta aagcaaacgt ggatatgtaa tgaatgaaat tcattctgga
48841 atgaaaaatt cacgtgatgt tgaaaaataa cacggggctt cagagaggac tttctggctg
48901 gcagcagact ccagattccc agggcccctg caccctcctc tgcccacagg gcaccttaat
48961 ggagaaggtg tgggaggaga gccaggccgg agtcagagca cactggtgac tccacatttg
49021 cagcgtgccc tgcctctctc ctgaggcttg gcaacgtgca atatgctaag caaactcccc
49081 ctgtccccgt ccagtttctg aggacaagag ccaccacctg tagcaaataa agacccagca
49141 acccttgac tcatctttgt gagtctctgg aatcagaggg tagccacatc gctgagaggt
49201 ggagtgaagc actcgggtga aaaggtacaa ggaagtcagg gacaggagtg tggggacatc
49261 acctagacaa tgacagagaa gaggggcaca gccgagtgag gggagagggg ccggcagtcc
49321 tacatcccct ggcctgaagc acgctccagg gcagaaggaa aaacactgtc tttggggtcc
49381 aagagacctg agttcaaatt ctggctccac cactgaccac ctgtgtaacc ttgaactgct
49441 gctgcctgaa cctcaggttc cccttctaaa aatagaggag aaaaggatgc atttctcctt
49501 gcccctgtga gaacgaaatg gtgcaagcac caaggagcct cagcaaaggt cgggcctgcc
49561 cccgcctggc caaacctttc ctcttcagga ggccacggca accgtagttt gacagaagag
49621 cagcaccttg atttaatgct tcccagcatg tgtccttgag caagtcacct aacctctctg
49681 ggctgcttcc tcattgggaa aatatggctg ccagtaaaac ctgccctgtc cacctcctgg
49741 ggcacttggc aaacagcaaa agagtccaaa tgtgcaggct gggccaggcg cagtggctca
49801 tgcctgtaat cccagcaatt taggaagcca aggtgggcgg atcacctgag gtcaggagtt
49861 tgagaccagc ctggccaaca tggtgaaacc ttgtctctac aaaaatacaa aaattagccg
49921 ggcatgatgg cgggtgcctg taatcccagt tactcgggag gctgaggcaa gagaatcgct
49981 tgaacccgga aggggaaggt tgcagtgagc caagattgtg ccactgcact ccagcctggg
50041 caacagagcg agactctgtc tcaaaaaaaa aaaaaaaaaa aaacaatgca gagctggctg
50101 tgtaaaaaac ctgttccact gcagggccca gtgtccacca ggctggggtg caggcctatg
50161 gggtgggggc ccagcatcag cctctcagca gccctgggag gcggggcgca tcccgtgccc
50221 ctcgtggtct ggatgtgttc tagcccaagt cctaggttac acctgccgtc gcctggcctc
50281 tcaggagagg cccagggtga ggaggagcat ggtaaaggtg aagctgattg ggaagtcagc
50341 tgttgggaaa gcaactcctt gcacattgga ggaaccgaga aagactgacc ccgaggacag
50401 cagccagcat ggcccttcct gggagcccat gttgggggat tcctgctgca gccaaggctc
50461 agcccttgtg gtcgcaggtg ctggtcctgg cctcttcccc tcccatgcag gagcacagga
50521 gagatggctt ctgaggacct gttgcagctg tggccctggg aatagatttg ccagggagct
50581 ttaaagcagc tgagtgtgtc atccagctaa gcctggggaa ggagcttggc tcaggtcctg
50641 acaggtgtga cagggatggg gactgggaag taagagatga aaccctggct ggaggctgtg
50701 agcttccaca gccagcgctg gacagggagg gtccagatat acccactagt gccctcacca
```

Genome annotation

- Annotation is the addition of information about the predicted sequence features to the flat file of DNA code
- Identification of potential coding sequences - CDS
- Homology searches to predict function
- Other features can be annotated as well
  - rRNAs
  - Potential promoters
  - tRNAs
  - Small non-coding RNAs
  - Repeat sequences
  - Insertion sequences (ISs), transposons, gene fragments
- Location of the origin of replication
- Determination of the number of bases, genes, and G+C%.



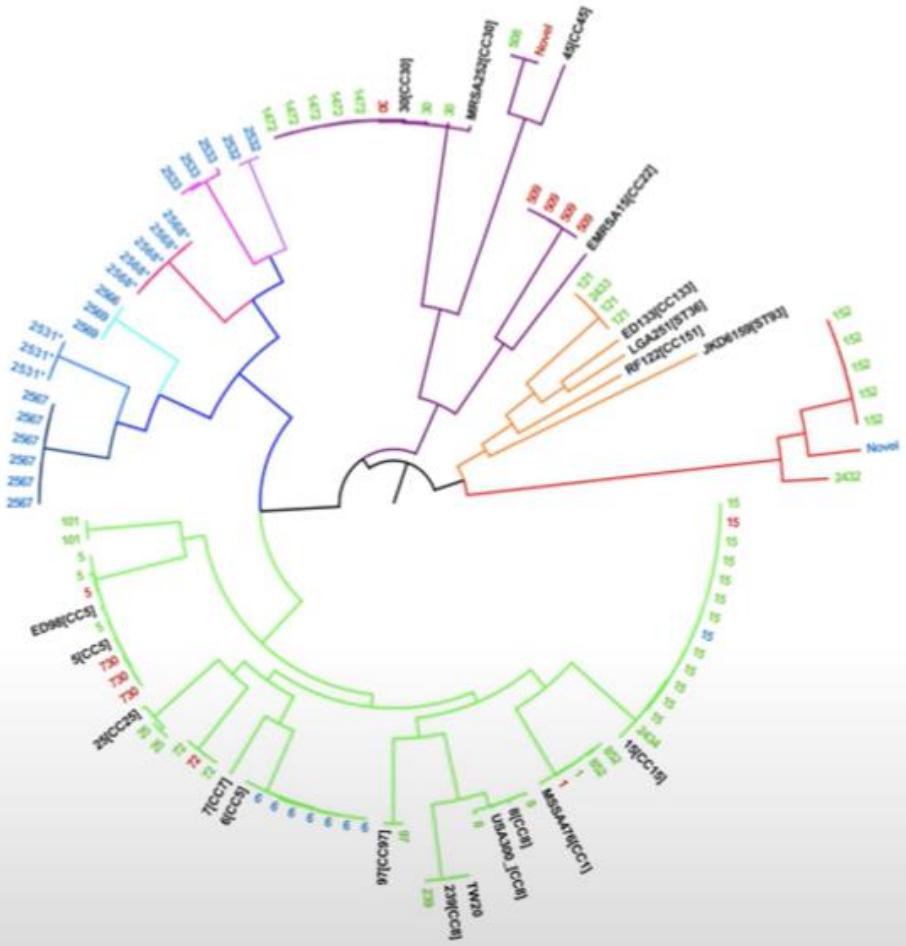
Tools for gene annotation in prokaryotes

- National Center for Biotechnology Information (NCBI) Prokaryotic Genome Annotation Pipeline (**PGAP**)
- **Prokka** (prokaryotic annotation)
- **RAST** (Rapid Annotations using Subsystem Technology)
- **DRAM** (Distilled and Refined Annotation of Metabolism)
- **Augustus** (for eukaryotes and prokaryotes)
- **Glimmer3** (only for prokaryotes)
- **Prokaryotic Genome Annotation Guide Annotation Genome Workbench and table2asn** (the replacement of tbl2asn)

These tools start from genomes and predict genes and proteins, tRNAs, rRNAs, and perform functional annotation of the predicted proteins

SNP calling

- Comparisons between closely related strains allows identification of SNPs that are informative for
  - Identifying biologically significant changes, e.g. during evolution in lab or patient
  - Reconstructing phylogenies using neutral changes

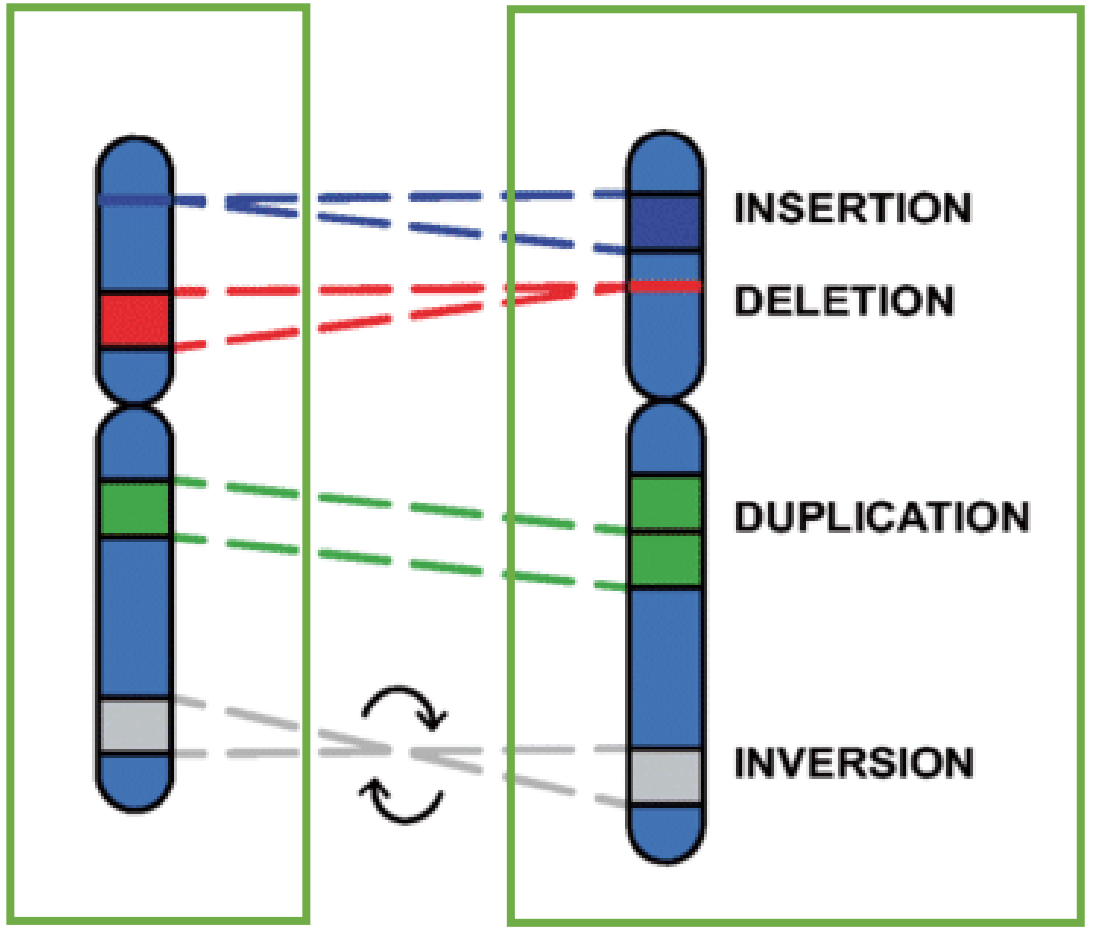


ATGTCGAGTGACCAGTGAGTAG
ATGTCGAGTCACCAGTGAGTAG

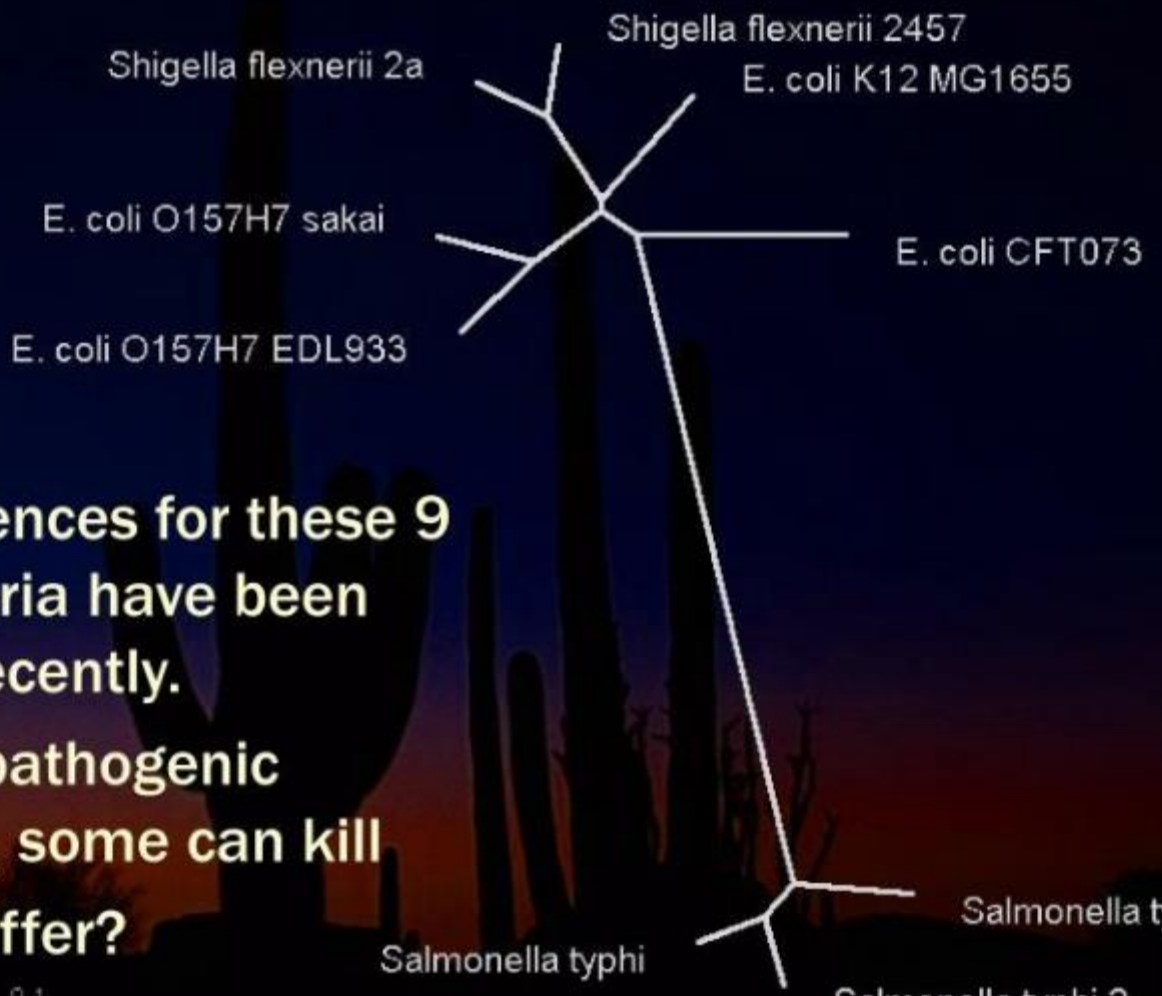
Large-scale mutation

- Inversion
- Horizontal gene transfer
- Homologous recombination
- Gene Duplication/Loss

→ Problem: Genetic elements may not have conserved order and orientation in the other genomes



Case Study: 9 Enterobacteria



Genome sequences for these 9 Enterobacteria have been published recently.

Diverse (non) pathogenic phenotypes, some can kill

Why do they differ?

Anchored genome alignment tools

Multi-LAGAN – align two or more heavily diverged genomes, assuming no differential gene content and no rearrangements (Brudno et. al. 2003)

MAVID – Like Multi-LAGAN, but also infer the branching structure of the organism's phylogeny (Bray et. al. 2004)

Shuffle-LAGAN – align two genomes that may contain repeats and rearrangements, no differential gene content (Brudno et. al. 2003)

Mauve – align two or more closely related genomes that have rearrangements, differential content in conserved order and orientation (Darling et. al. 2004)

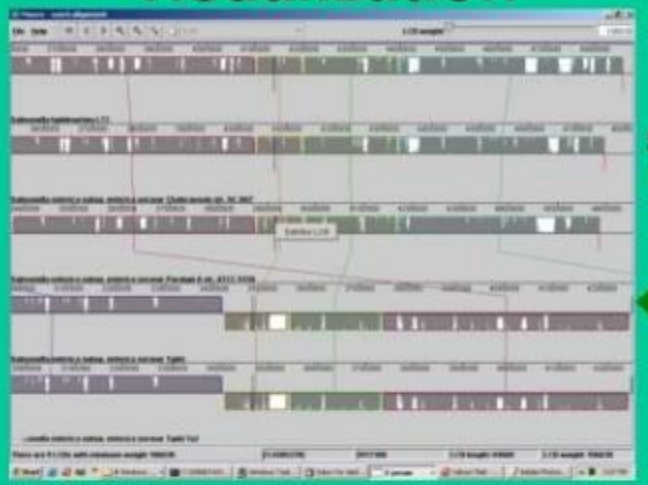
Mulan – align two or more closely related genomes, possibly with differential gene content (Ovcharenko et. al. 2005)

M-GCAT – align two or more closely related genomes with rearrangements and other changes (Treangen et. al. 2005)

The two component architecture of Mauve

Mauve
Multiple Genome Alignment

Java 1.4+ Interactive Visualization



GenBank or FastA sequences

alignments

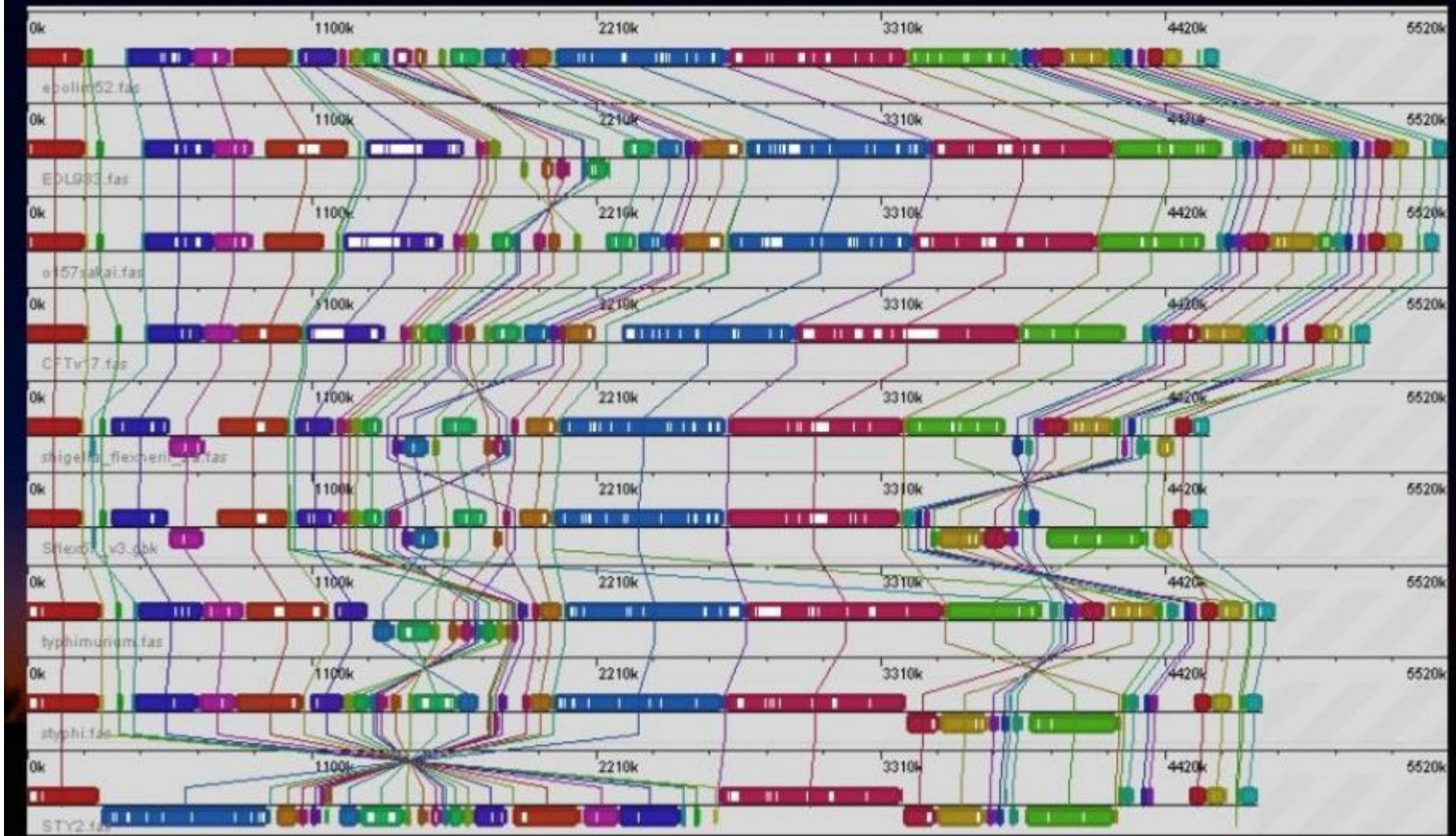
C++ command-line aligner

Windows, Linux, Mac OS X

100% Free/Open Source Software

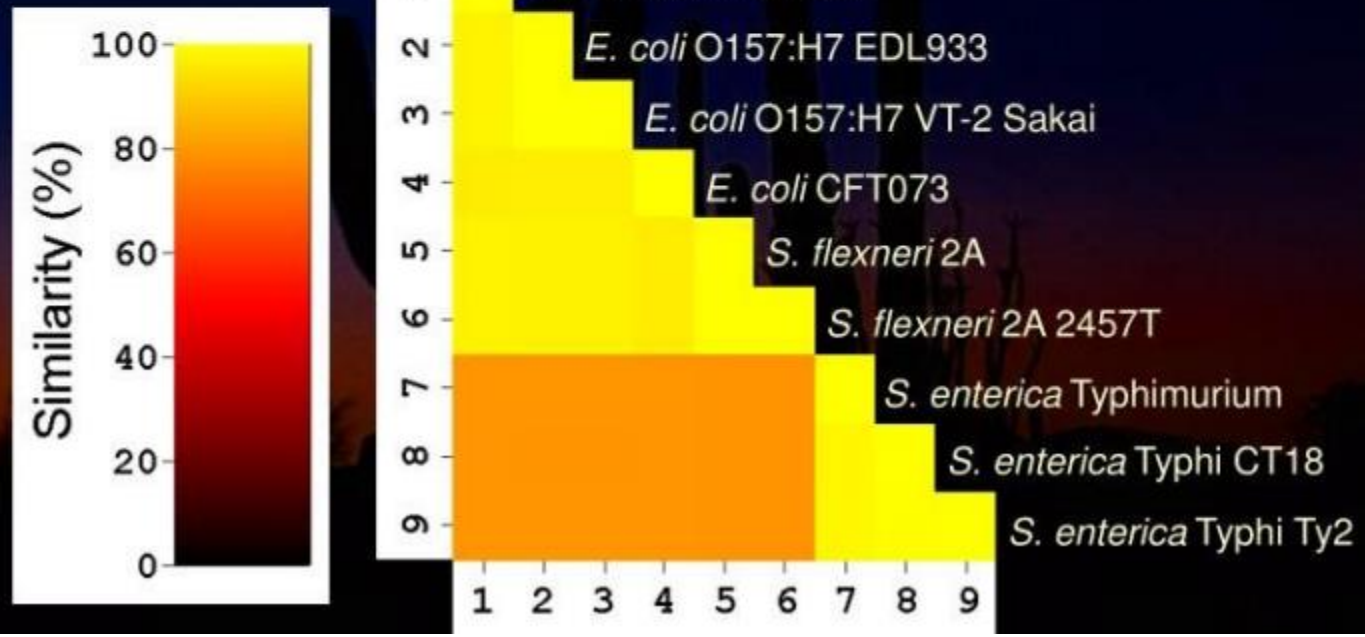
We use each language for what it does best–C++ for efficient algorithm implementation, Java for a cross platform GUI

Results: Homology structure of 9 Enterobacteria



Alignment of 9 Enterobacteria

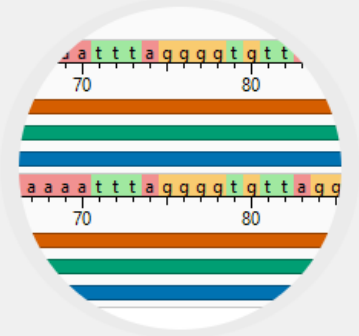
- 45 locally collinear blocks (LCBs)
- 2.86Mbp of backbone sequence – only 58% of average genome size
- Backbone is any region shared among all genomes
- Diverse phenotypes caused by horizontal gene transfer
- 3 hours compute time on a 1.6 GHz Linux PC

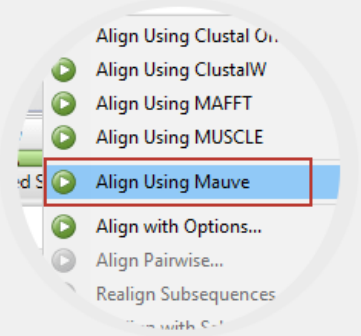


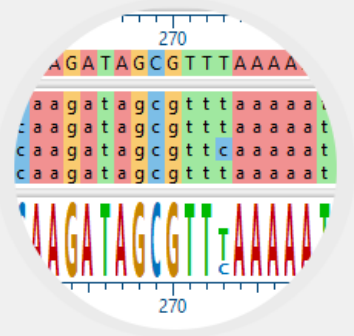
Multiple Genome Alignment with Mauve

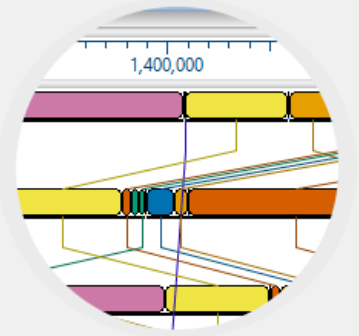
mauve

Multiple Genome Alignment









Step 1

Add genome sequences

Step 2

Perform alignment using Mauve method

Step 3

View and analyze results in multiple views

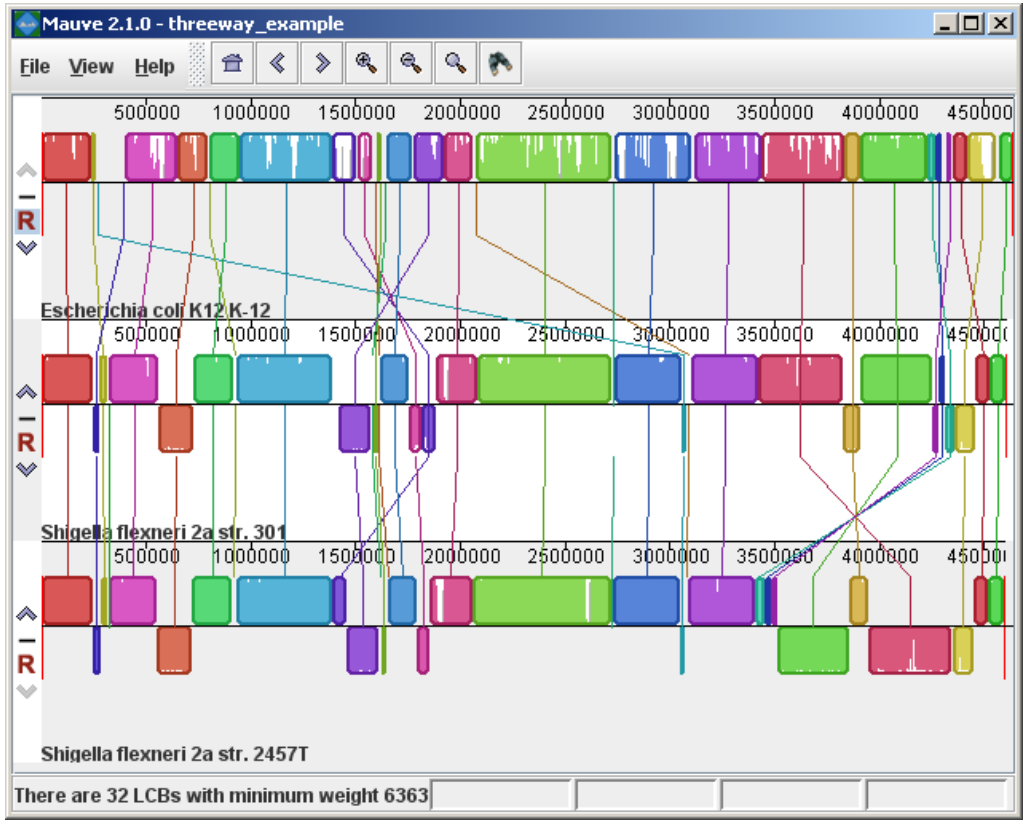
Step 4

Customize appearance and export for publication

https://darlinglab.org/mauve/mauve.html

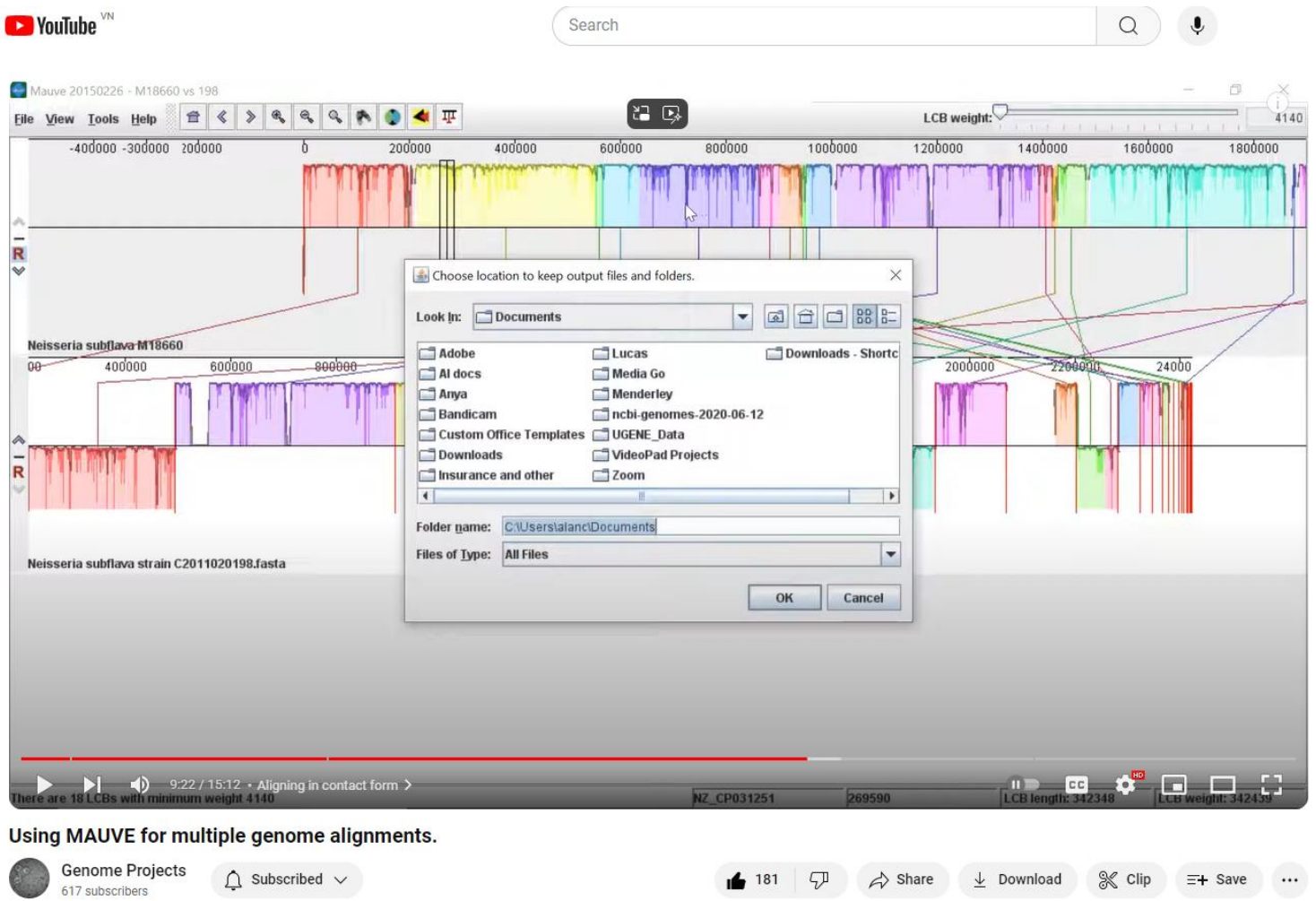
Localized Co-linear Blocks (LCB)

- Genomes diverge from an ancestral species into two or more modern species, chromosomes undergo many sorts of rearrangements, including insertions, deletions, inversions and translocations.
- Much of the same sequences remain even after long periods of evolutionary history, but in different arrangements from one species to the next.
- LCBs are regions of chromosomes that appear to be conserved across all species being examined.
- If an entire chromosome is unchanged, then the entire chromosome would be a single LCB in all species in the alignment.
- Many chromosomes will be a mosiac of several LCBs.
- In other cases, an LCB on one chromosome will be found on a different chromosome, indicating a translocation.
- The purpose of Mauve, then is to calculate the LCBs shared between two or more genomes, and to display them visually.



https://home.cc.umanitoba.ca/~psgendb/tutorials/bioLegato/getgenome/Mauve.html

Tutorial of using MAUVE for multiple genome alignments



https://www.youtube.com/watch?v=KGTSn80XzBo

Homework

1. Search for the sequences of E. coli CFT073, E.coli K12 MG1655 and E. coli O157H7
2. How long these genomes?
3. Using MAUVE for multiple genome alignments of these 3 genomes