Bioinformatics Algorithms
COS-BIOL-530/630
Lecture06

Days & Times	Room	Meeting Dates
Tu 2:00PM - 3:50PM	Thomas Gosnell Hall (GOS)-2178	01/13/2025 - 04/28/2025
Th 2:00PM - 3:50PM	Thomas Gosnell Hall (GOS)-2178	01/13/2025 - 04/28/2025

Instructor:
Fernando Rodriguez
email: frvsbi@rit.edu
Office: Orange Hall 1311

Genome comparisons
- Lecture06-

Announcements

Week 6
Lecture06
Lab06
- Discussion 6
- Activity 6

Quiz 5

Exam 1: Lecture01-06/Lab01-06
- Thursday, February 27th 2pm.

Genome comparisons
- Lecture06-

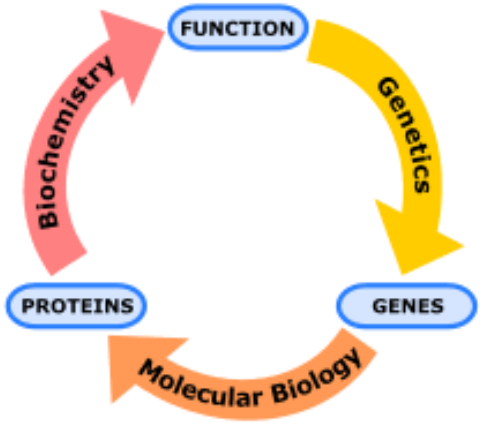
Topics:

- Genome evolution
- Synteny /Collinearity
- Tools for comparative genomics
- Orthogroups (orthologs, paralogs)
- Whole Genome alignment

Comparing genomes to understand evolution

- Phylogenetics: how to deduce the relationships in evolution using gene or protein families.

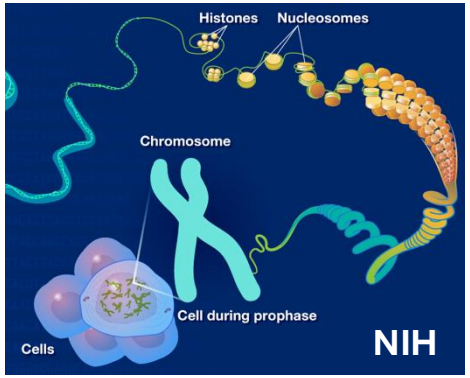
- Gene (genetics) - Levels
  - Similarities in different organisms
  - Expression levels
  - Function
  - inheritance



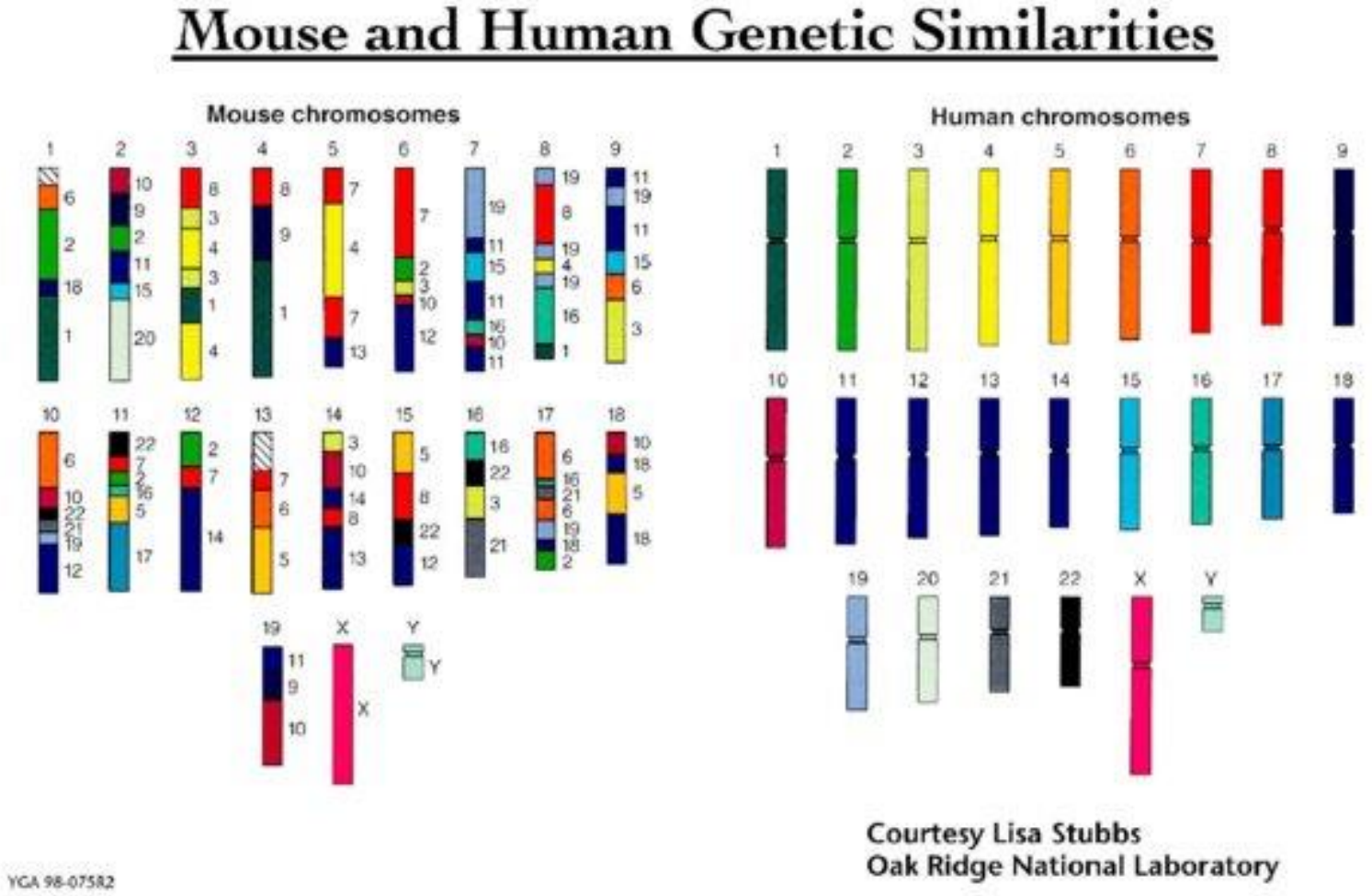
- Can we compare genomes to understand better the similarities and differences between genomes over evolutionary time?

- Genome (genomics) - Levels
  - Structure
  - Organization
  - Expression

- What makes a genome unique?
  - It's not only DNA…



The mouse genome is 14% smaller than the human genome, though 90% of mouse genes have intact human homologs.

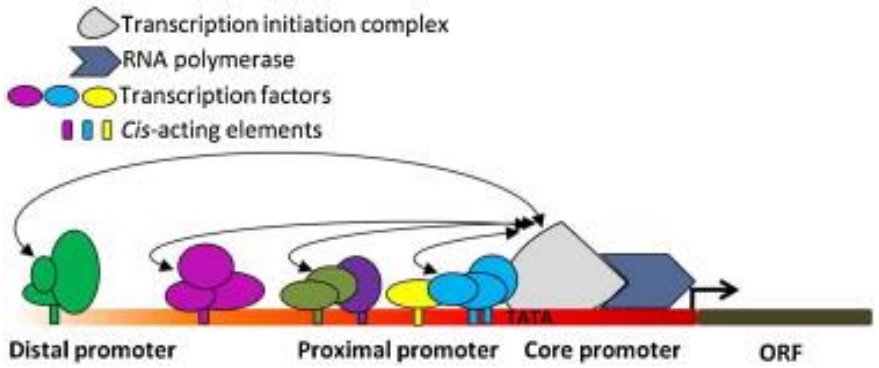


There are about 100 homologous segments. The colors and numbers next to the mouse chromosomes indicate the human counterpart.

Comparative Genomics

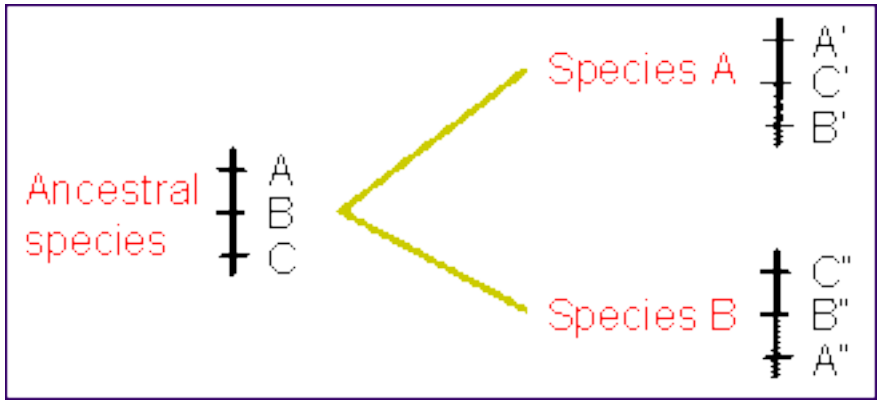
Most of the efforts in the field of computational comparative genomics focus on:

- Detect conservation within genes and intergenic regions.
- Conservation of gene order.
  - **Synteny**: the degree to which genes remain on corresponding chromosomes.
  - **Collinearity**: corresponding order of the genes
  - Both terms are related: colinear regions are syntenic / not all syntenic regions can be collinear.
- Predicting the presence and pattern of *cis*-acting regulatory elements.

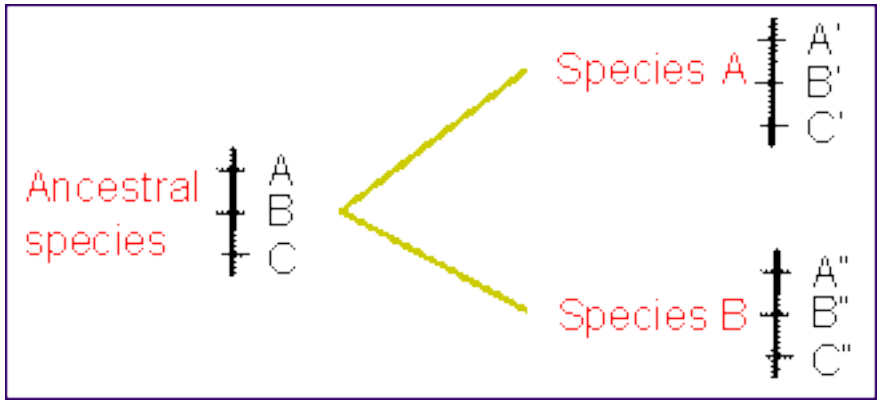


Definition: synteny and collinear

Syntenic*: a set of *loci* in two different species that are located on the same chromosome in each (not necessarily in the same order).



Collinear: a set of *loci* in two different species that are located on the same chromosome in each and are conserved in the same order.



Source: https://www.integratedbreeding.net/

Both terms are related:

- colinear regions are syntenic
- not all syntenic regions can be collinear.

* Synteny now has the same meaning as collinear, despite the different origins of the terms

Comparative Genomics

- Still, the main approach to detecting similarities on a genomic scale involves pairwise or multiple sequence alignments.

- Aligning genomic-scale data sets is no different than the initial problems in Lecture02 & Lecture03:

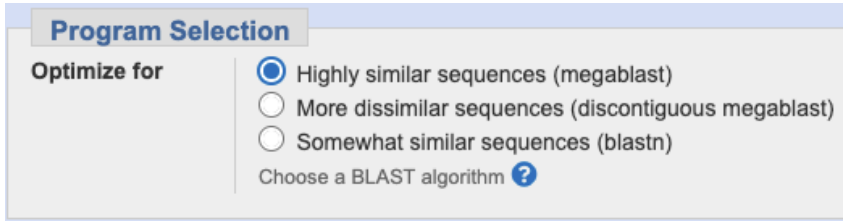
  - Global/local alignment
    - ➢ Needleman-Wunsch / Smith-Waterman algorithms (Dynamic programming)

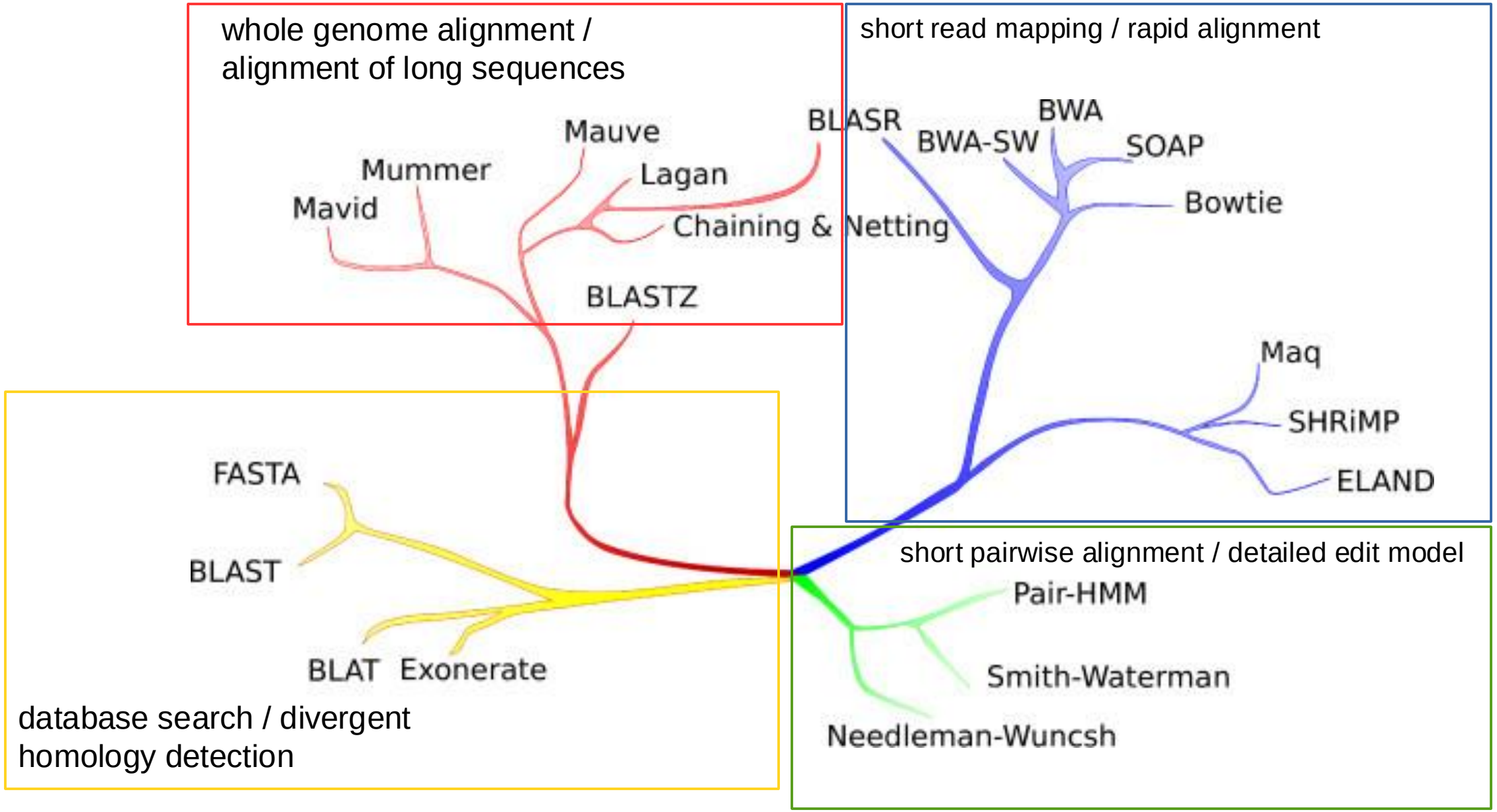
  - Heuristic search
    - ➢ FASTA/BLAST algorithms

- Given the increasing availability of complete genome sequences, numerous algorithms for genome alignment of (two or more) genomes have now become popular.

- We have seen MegaBLAST (NCBI), which is optimized to align longer nucleotide sequences with a high level of sequence similarities (>95%).

NCBI





Chaisson and Tesler, 2012

Comparative Genomics

But in reality…..

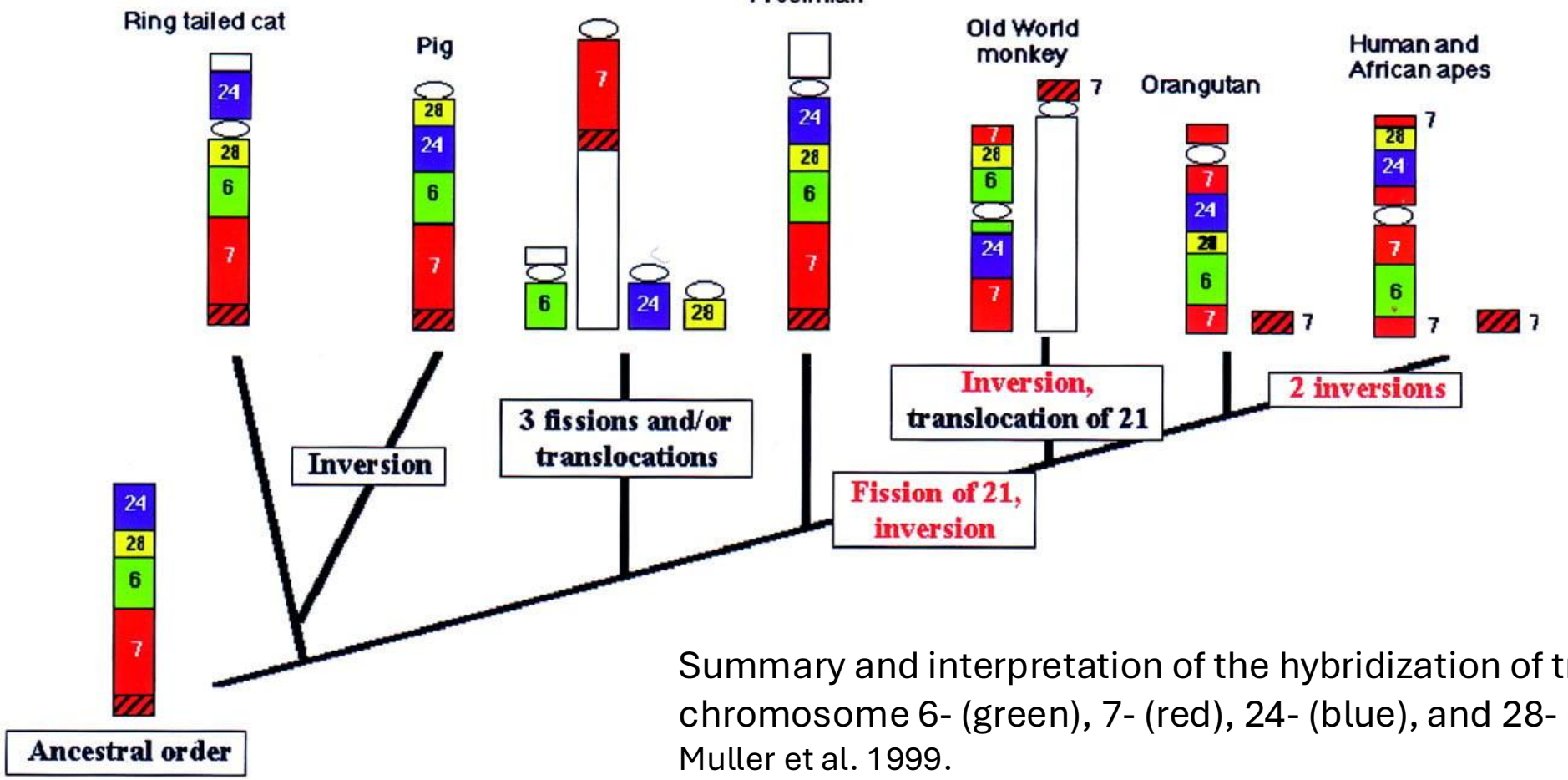
- A strict whole genome alignment between species will fail due to chromosomal rearrangements over evolutionary history.

- Computationally difficult depending on genome length, genome completeness, etc.

Comparative Genomics

Tree shrew

Prosimian



Summary and interpretation of the hybridization of tree shrew chromosome 6- (green), 7- (red), 24- (blue), and 28- (yellow). Muller et al. 1999.

Comparative Genomics

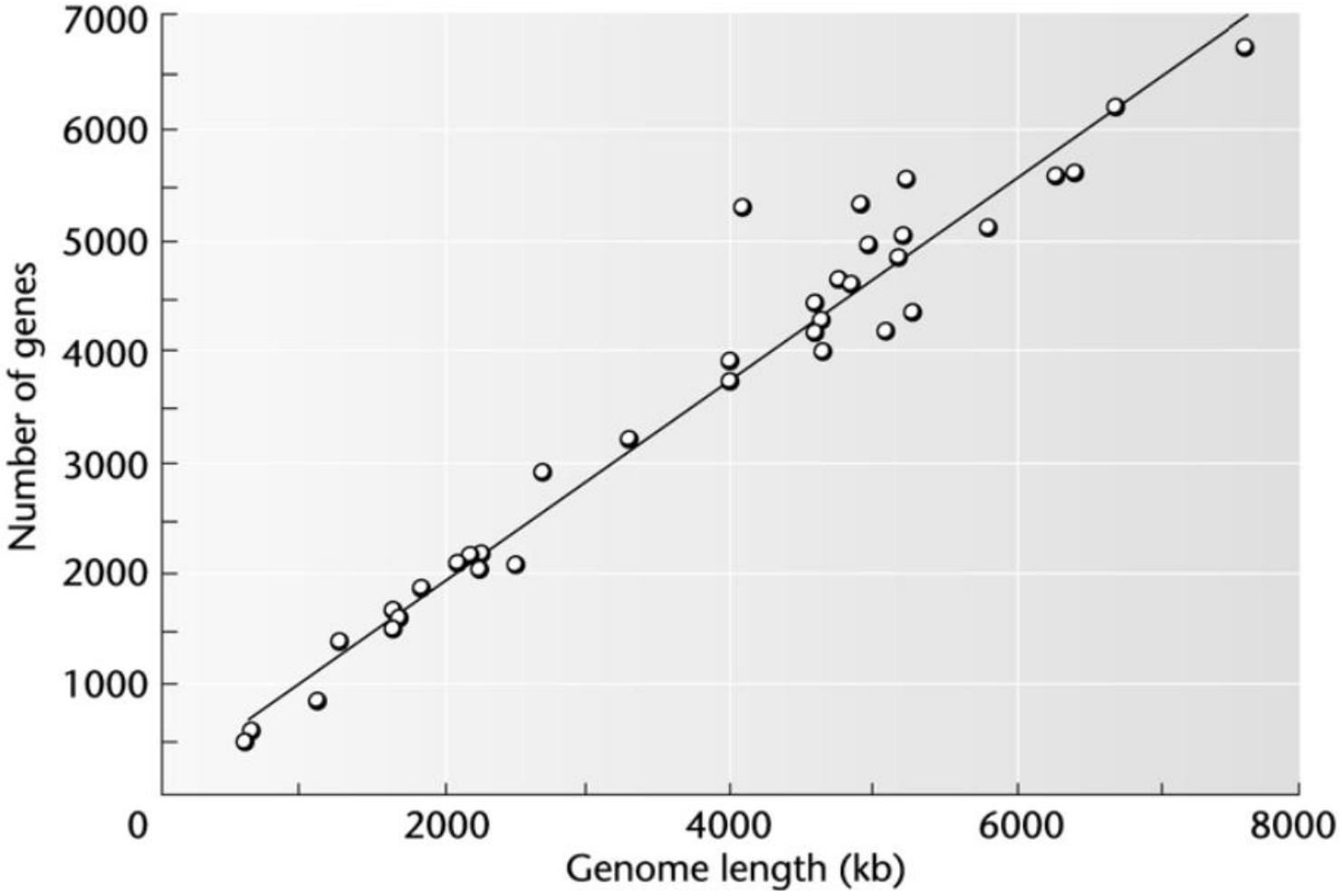
But in reality…..

- A strict whole genome alignment between species will fail due to chromosomal rearrangements over evolutionary history.
    - Synteny is the change in gene locations over evolutionary time.
    - When gene locations change, how is it going to affect the alignment?
    - Problems:
        - Large insertions/deletions
        - Gene gain/loss
        - Chromosomal rearrangements
        - Repetitive elements
    - Simple proxies in prokaryotic examples.

Length of the genome

- One way to compare genomes would be based on the assumption that those with similar lengths are more closely related.
    - In prokaryotes, length is relatively conserved.

Length of the genome



A strong correlation (*R* 0.98) between the total genome length and the estimated number of genes on bacterial genomes.

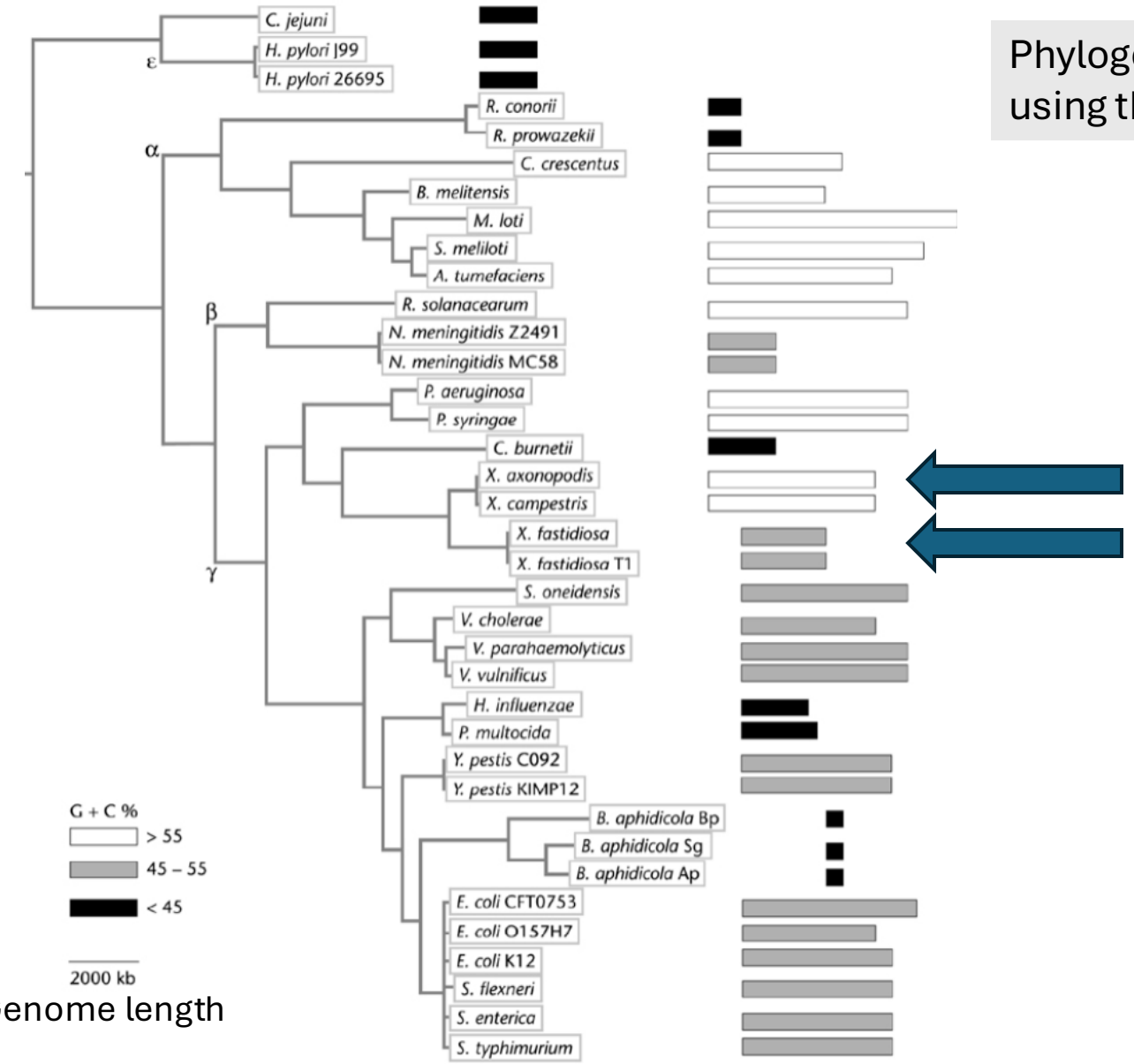
Each point is one species of proteobacteria.

Most of the space is composed by genes.

Problem: Genome length is not always similar in closely related groups.

Source: *Bioinformatics and Molecular Evolution*, Higgs and Attwood

tRNA tree vs. genome lengths



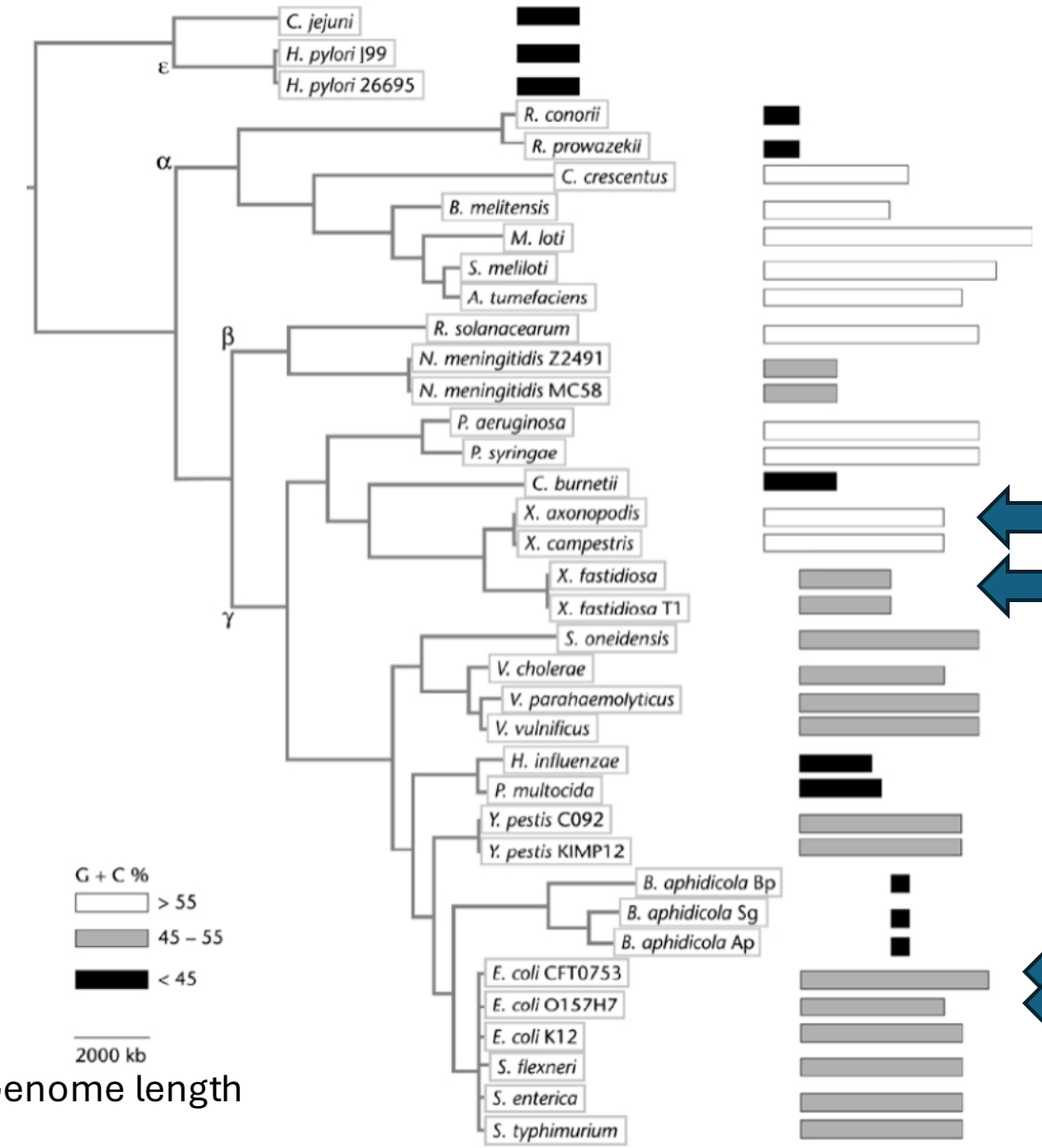
Genome length

Phylogenetic tree of proteobacteria using tRNA genes.

Related groups have different lengths

Source: *Bioinformatics and Molecular Evolution*, Higgs and Attwood

tRNA tree vs. genome lengths



Genome length

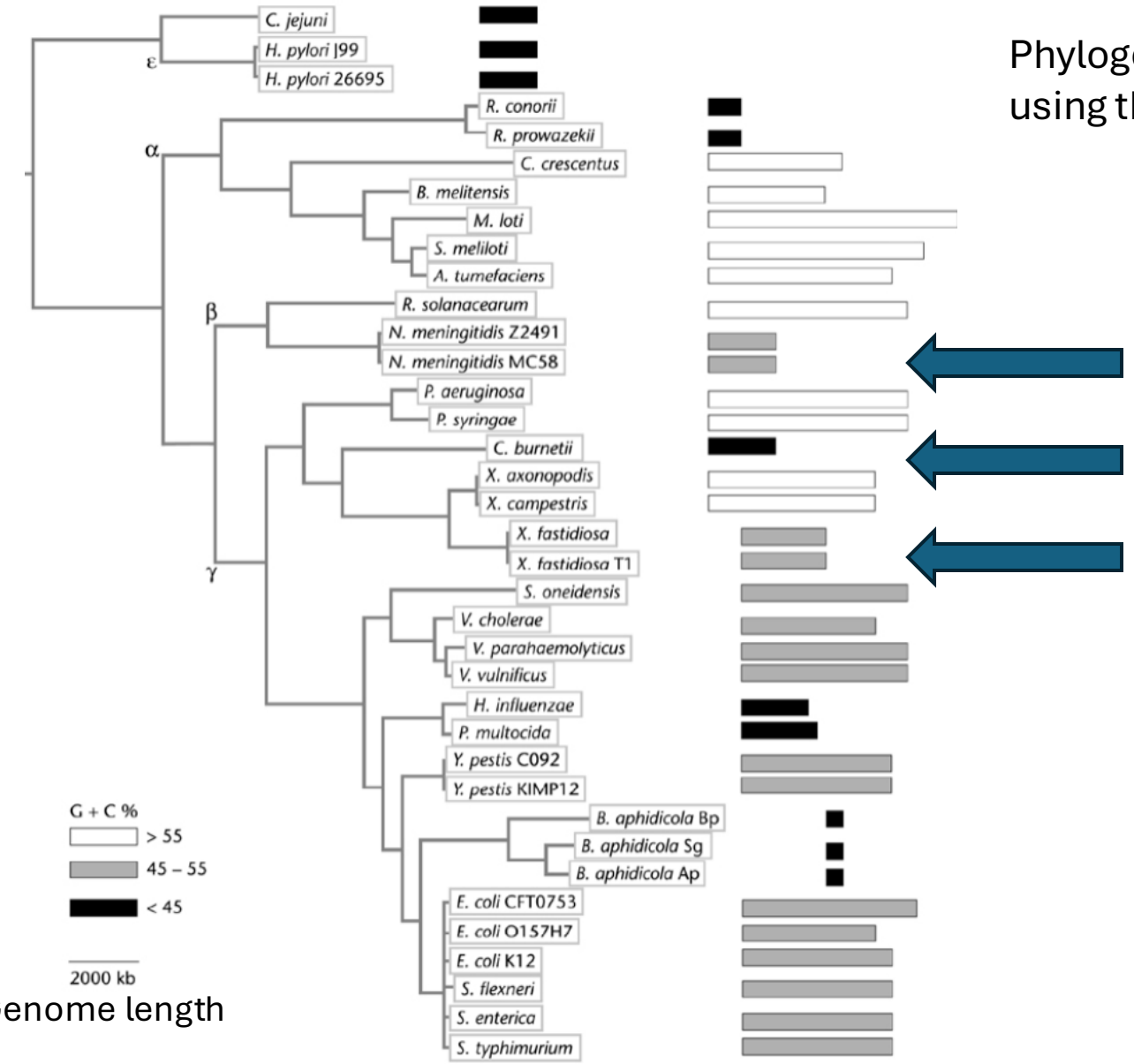
Phylogenetic tree of proteobacteria using tRNA genes.

Related groups have different lengths

Two strains of *E. coli*

Source: *Bioinformatics and Molecular Evolution*, Higgs and Attwood

tRNA tree vs. genome lengths



Genome length

Phylogenetic tree of proteobacteria using tRNA genes.

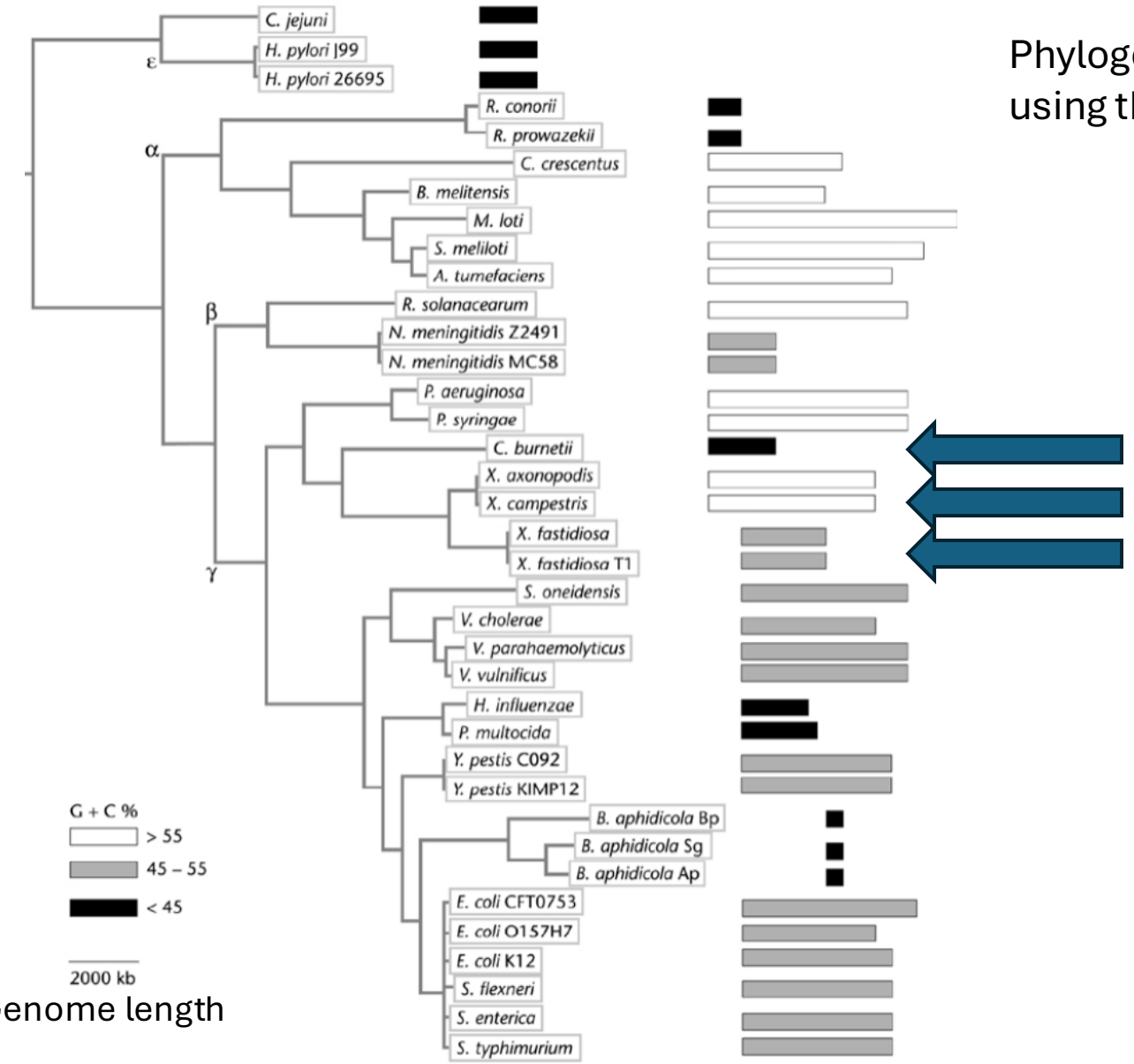
Different groups with similar lengths

Source: *Bioinformatics and Molecular Evolution*, Higgs and Attwood

GC content vs. genome lengths

- The GC content of organisms is a highly variable trait.
- GC content is usually conserved (within a range) in taxonomic groups.
  - E. coli: ~50%
  - Human: ~41%
  - Rotifers: ~30%
- Can we assume that those with similar GC content are more closely related?
- Problem: not really

GC content vs. genome lengths



Phylogenetic tree of proteobacteria using tRNA genes.

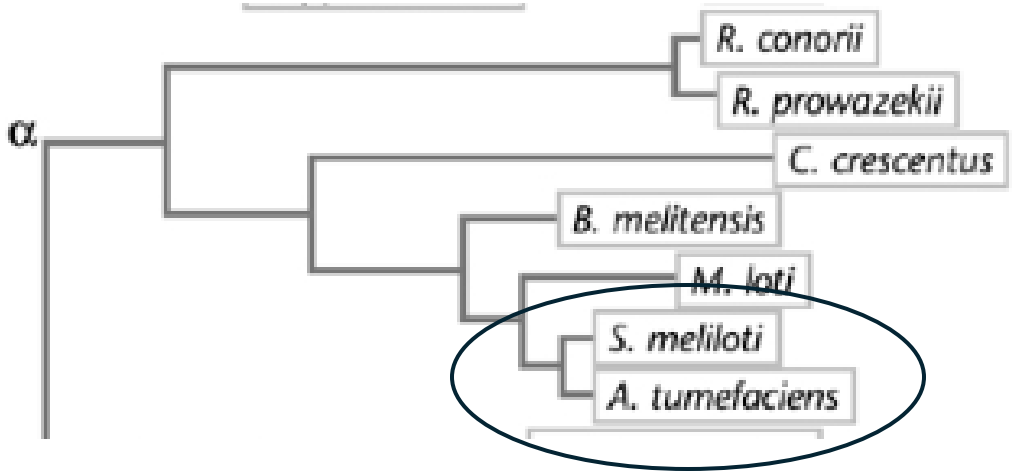
Related groups, but different CG content

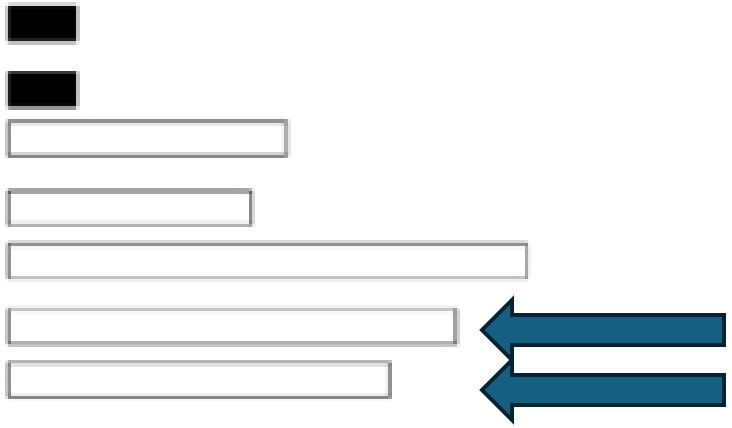
The main observations from this figure are that both **genome length and base composition** can change quite rapidly and that closely related species are not always similar in either length or base composition.

Source: *Bioinformatics and Molecular Evolution*, Higgs and Attwood

Collinearity

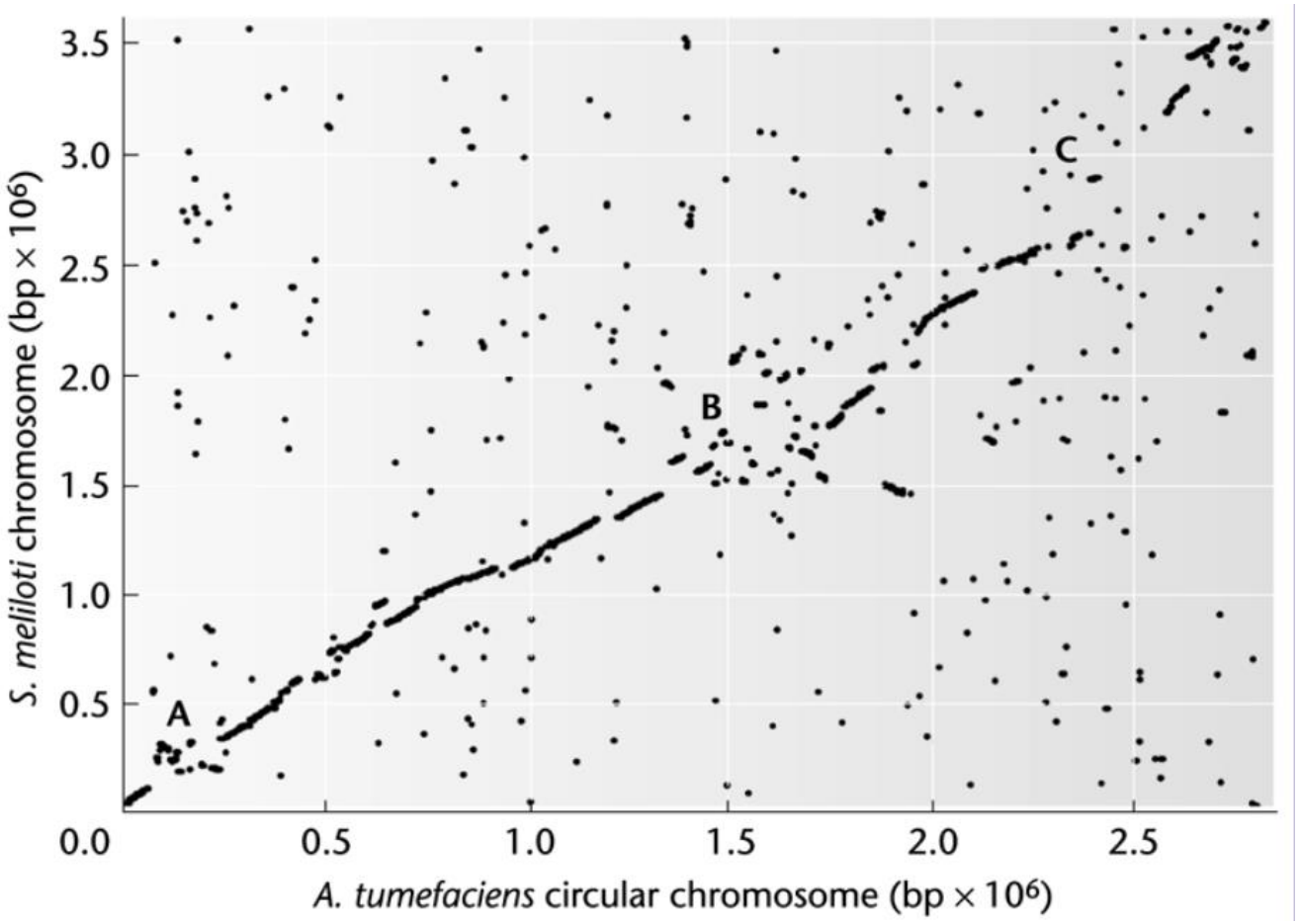
Instead of genome length, we can check collinearity as a measure for closely related species of prokaryotes.





Let’s check collinearity in these two species

Collinearity



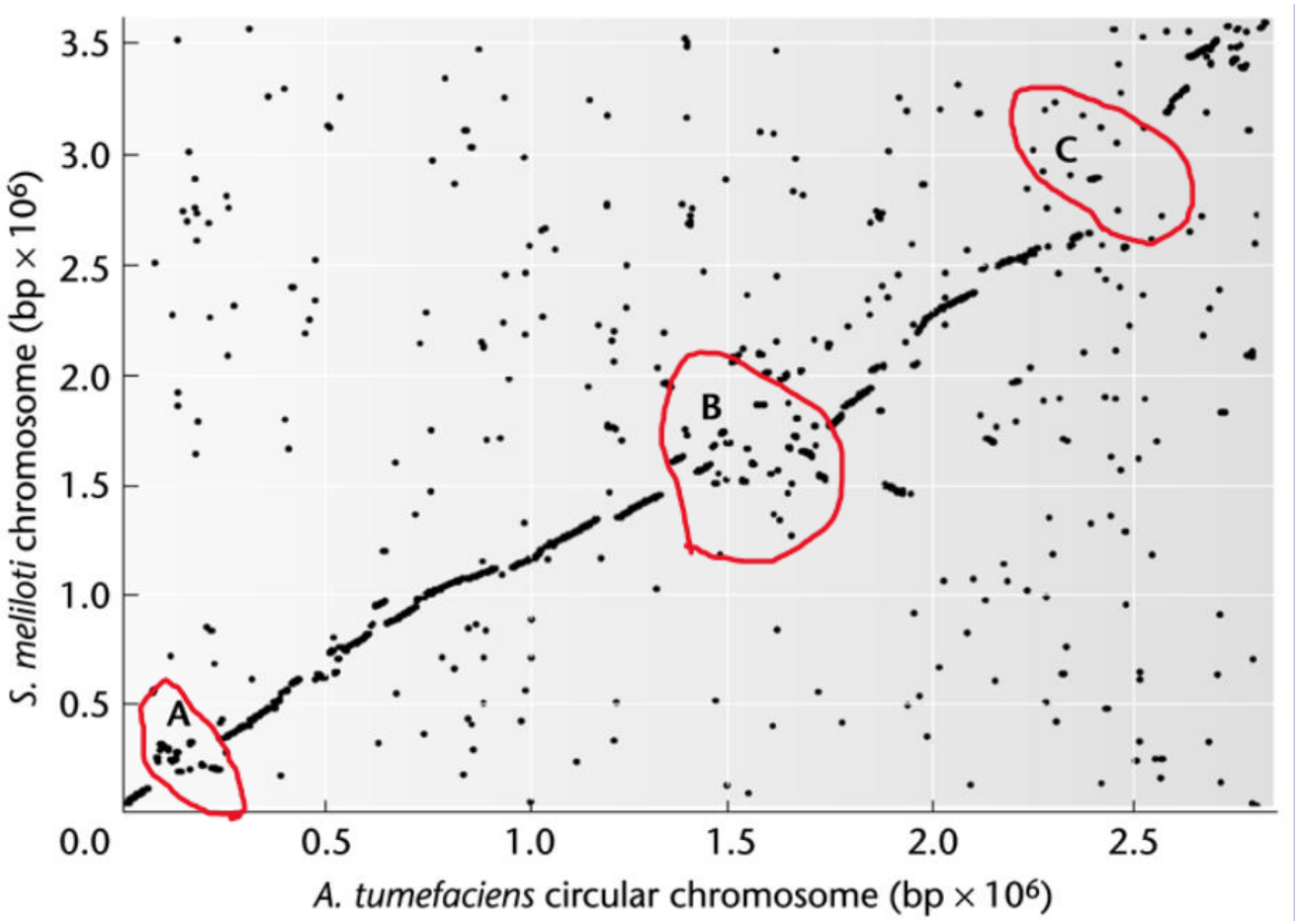
A dot-blot illustrating regions of Collinearity between the two bacterial chromosomes.

Each **dot** represents the **best blast hit** (*blastp*) between protein sequences.

There are three (A, B, C) regions lacking collinearity.

Source: *Bioinformatics and Molecular Evolution*, Higgs and Attwood

Collinearity



A dot-blot illustrating regions of collinearity between the two bacterial chromosomes.

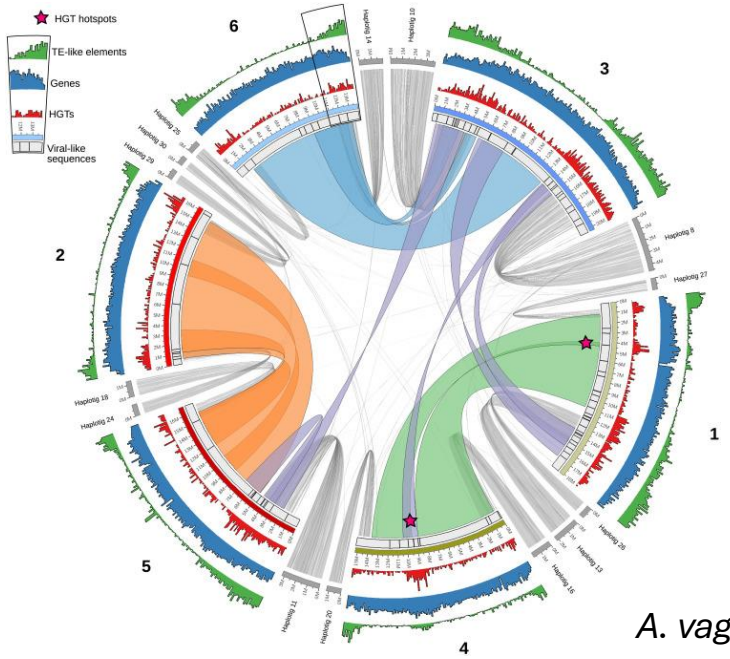
Each **dot** represents the **best blast hit** (*blastp*) between protein sequences.

There are three (A, B, C) regions lacking collinearity.

Source: *Bioinformatics and Molecular Evolution*, Higgs and Attwood

Collinearity

- Collinearity provides some information for closely related species.
- However, a strict whole genome alignment will not work.
  - In prokaryotes, rearrangements are common (previous slide).
  - In eukaryotes, there are many gene translocations to different chromosomes (during and after speciation).



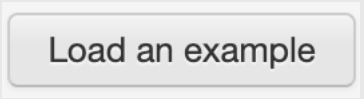
Solution: We should investigate a set of genes only present in different related species (like gene overlap).

A. vaga rotifer genome (Simion *et al*. 2021)

Small in-class activity!!

synteny block analysis

Synteny Portal: a web-based application portal for synteny block analysis (Lee et al. 2016)
https://doi.org/10.1093/nar/gkw310

- Go to http://bioinfo.konkuk.ac.kr/synteny_portal/
- Explore the different options.
- In SynSearcher:
  - Click in 

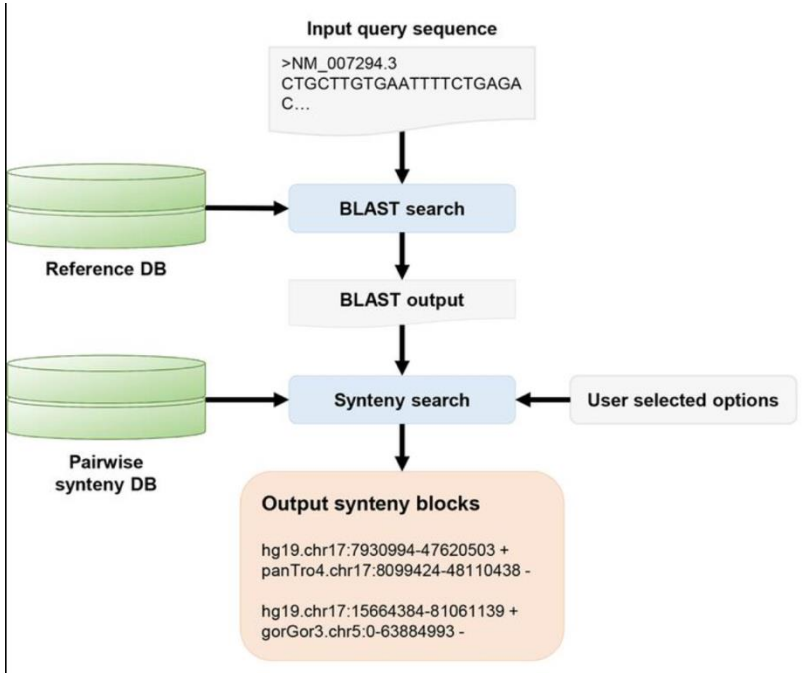
  - Submit
  - In JobStatus, click **View**
  - Feel free to change the Target Sequences, Resolution

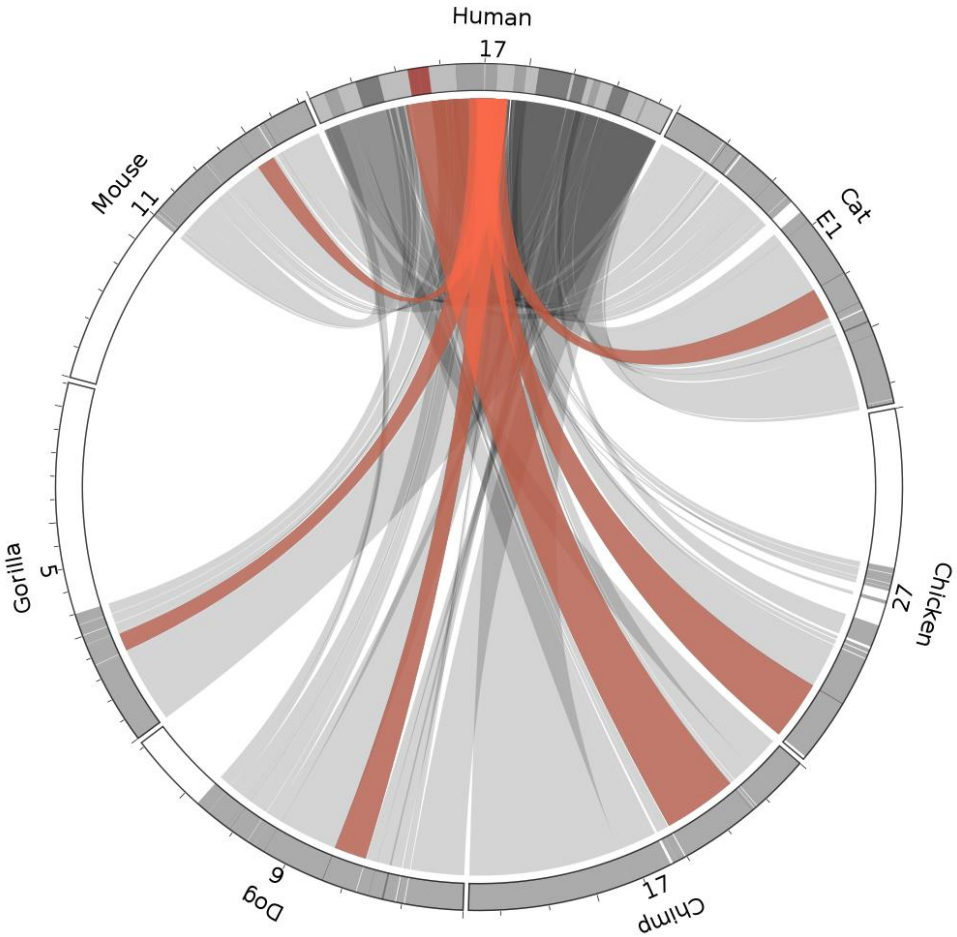
Job ID	Description	Status	Result	Circos
1	Input sequence: Click, Sequence type: DNA Resolution: 150000 bp Reference: Human (hg38) Target species: Cat (felCat9), Chicken (galGal6), Chimp (panTro6), Dog (canFam5), Gorilla (gorGor6), Mouse (mm39)	Complete	**View** Download	**View** Download

synteny block analysis

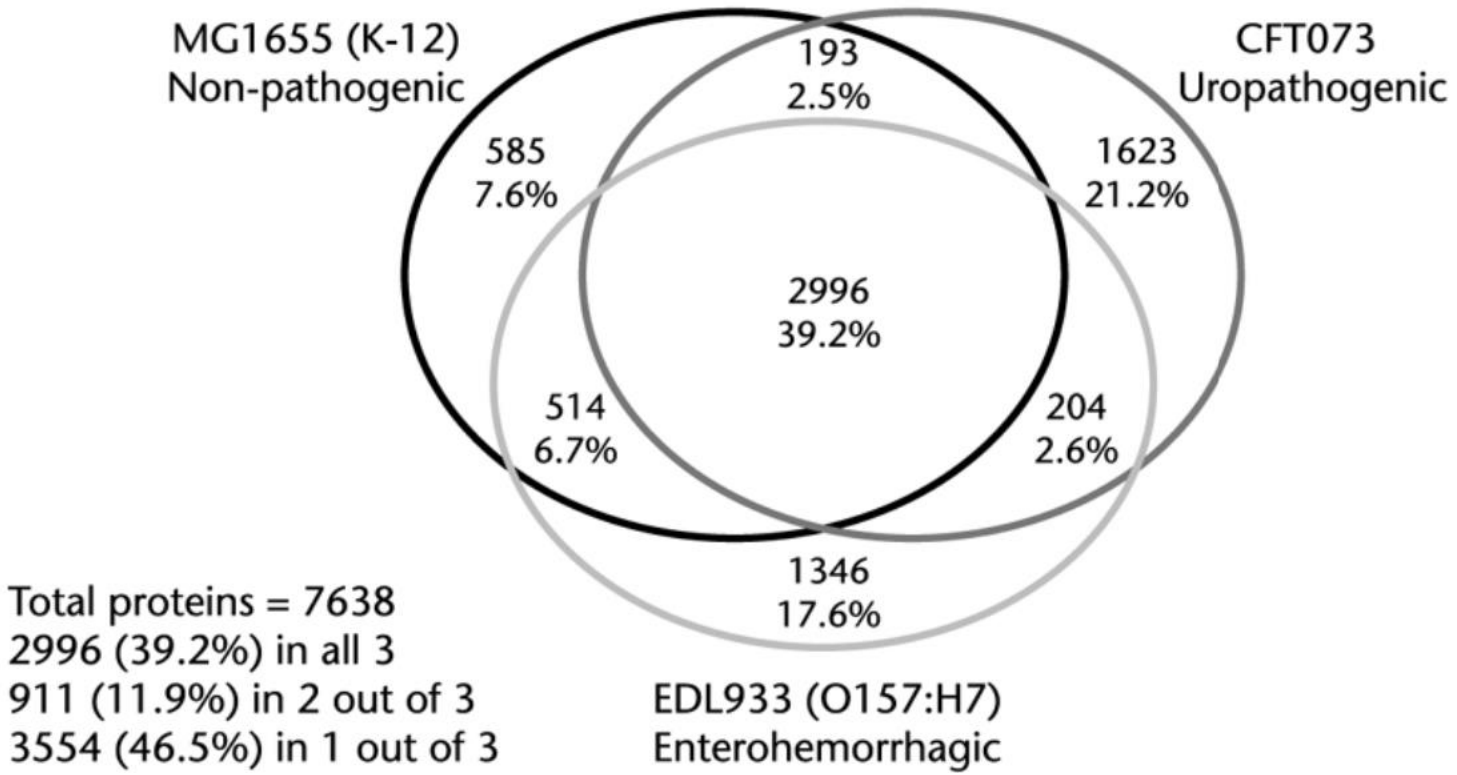
```
>>Top BLAST search result
RefChr  RefStart        RefEnd  QueryID QueryStart      QueryEnd        Rawscore        Bitscore        Evalue
chr17   43094860        43091435        gi|237757283|ref|NM_007294.3|:903-4328  1       3426    3426    6327    0.0

>>Synteny blocks
hg38.chr17:38195151-45561535 +
felCat9.chrE1:39626587-45398911 +
```





Gene content



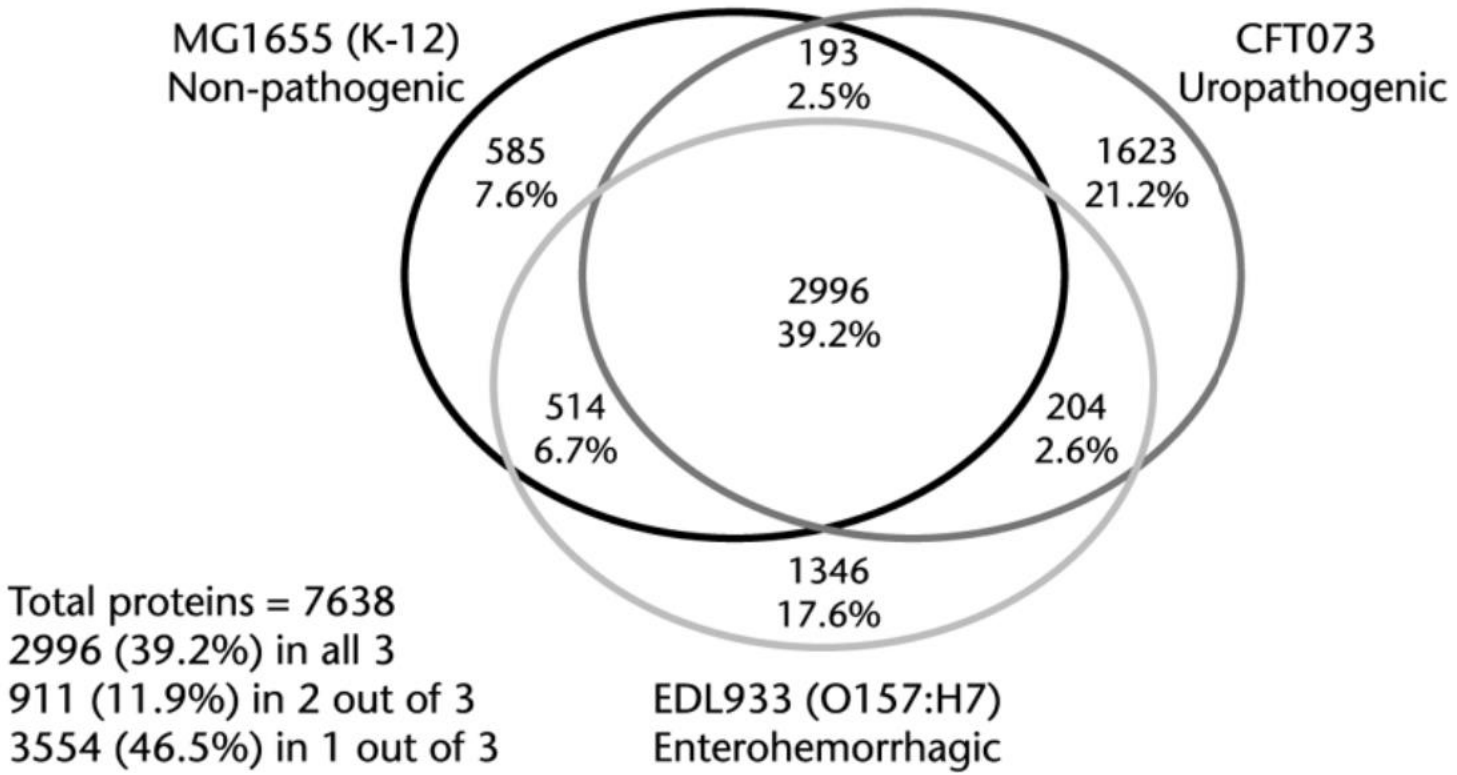
Source: *Bioinformatics and Molecular Evolution*, Higgs and Attwood

<u>Comparison of gene content in three strains of E. coli.</u>

The regions in the diagram illustrate the numbers of predicted proteins present in 1, 2, or all the three strains.

Only 39% of genes are shared between all three strains.

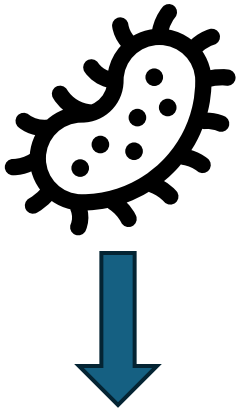
Gene content



<u>Comparison of gene content in three strains of E. coli.</u>

The regions in the diagram illustrate the numbers of predicted proteins present in 1, 2, or all the three strains.

Only 39% of genes are shared between all three strains.

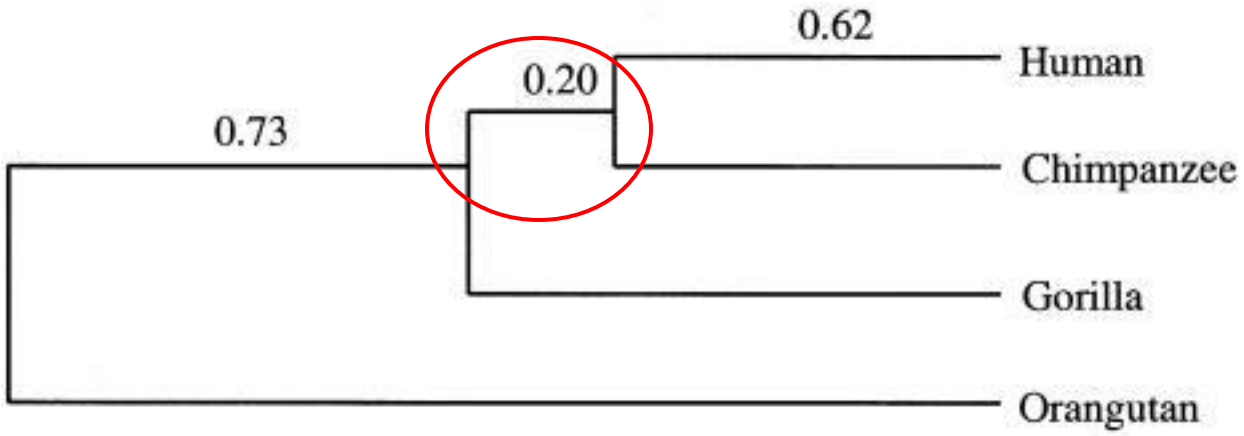


Gene duplications and horizontal gene transfers

Source: *Bioinformatics and Molecular Evolution*, Higgs and Attwood

Intergenic regions

- Chen *et al*. 2001: Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees.
- They compared a subset of 53 intergenic regions (total of24,234 bp in length).
- The majority of segments support Homo-Chimp clade
- However, the support for Gorilla-[Homo-Chimp] is not that good.

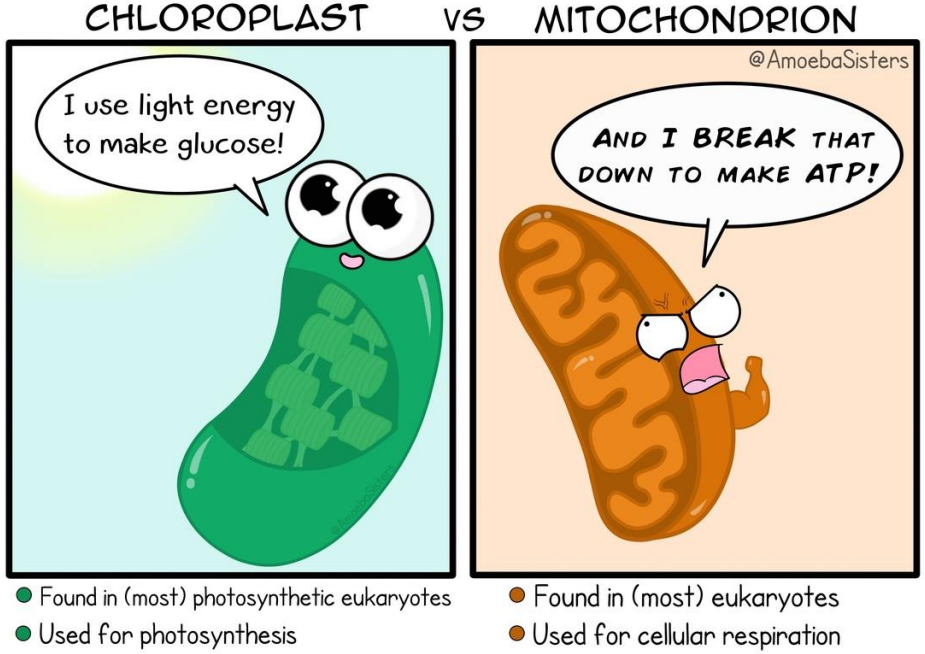


Phylogeny of hominoids. The branch lengths (Jukes-Cantor distances) are computed under the assumption of rate constancy and used for estimating divergence dates.

Source: https://doi.org/10.1086/318206

Phylogenies based on organellar genomes

- Both mitochondria and chloroplasts are organelles in eukaryotic cells with their own membranes and genomes.
- Endosymbiosis theory.
- Organellar genomes have become increasingly essential for studying genetic diversity, phylogenetics, and evolutionary histories.
    - Eg. mitochondrial DNA divergence in chamois (*Rupicapra*).
- They have structural rearrangements, too.
    - Eg. Dotplot activity with two chloroplast genomes.



Phylogenies based on shared gene content

- Sometimes, phylogenies based on different genes are not consistent with one another -> inaccuracies or lack of resolutions.

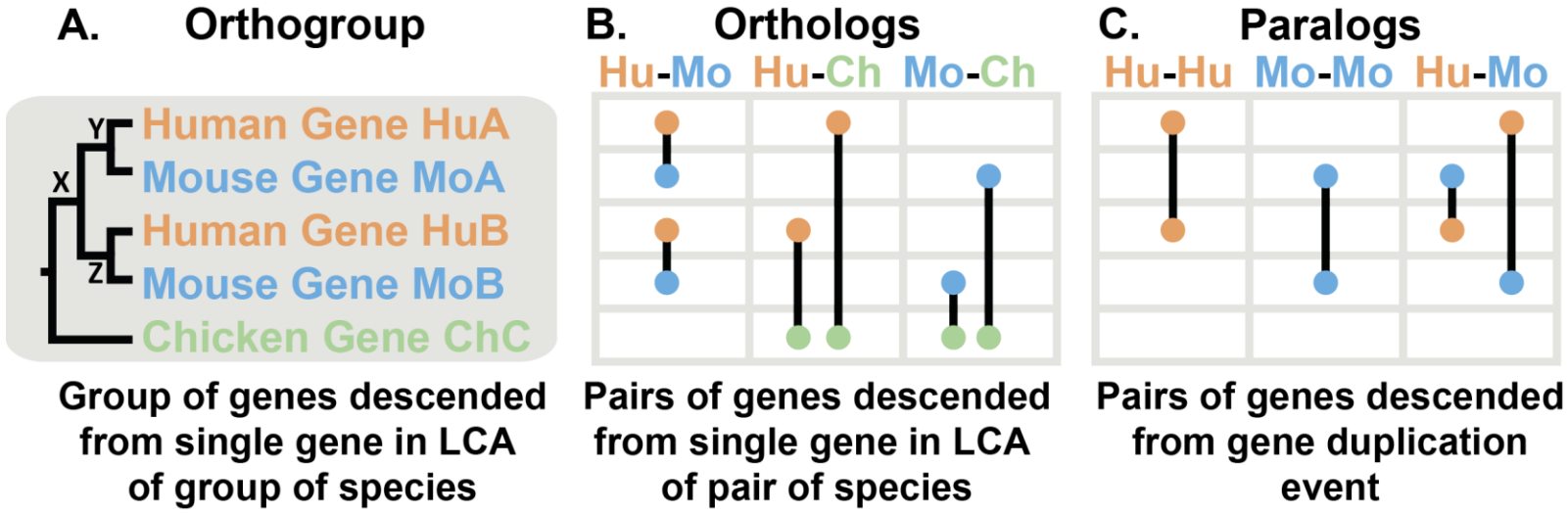
- After all, genome similarities would give a better degree of similarities than one gene comparison.

- We need to find **orthologs**: genes from two different genomes that evolved from the same sequence (ancestor).
  - **Orthologs** would have a (reciprocal) best BLAST hit and significant E-value (0.01) if we blast the two genomes.
  - **Paralogs** are generated as a result of genomic duplication events; this generates two copies of the same sequence in the same lineage.

- However, there are biases due to
  - Different genome sizes.
  - Proportion shared genes *vs*. total number of genes.
  - Gene deletions and horizontal transfers.

Phylogenies based on shared gene content

Orthogroups, Orthologs & Paralogs



LCA: Last common ancestor

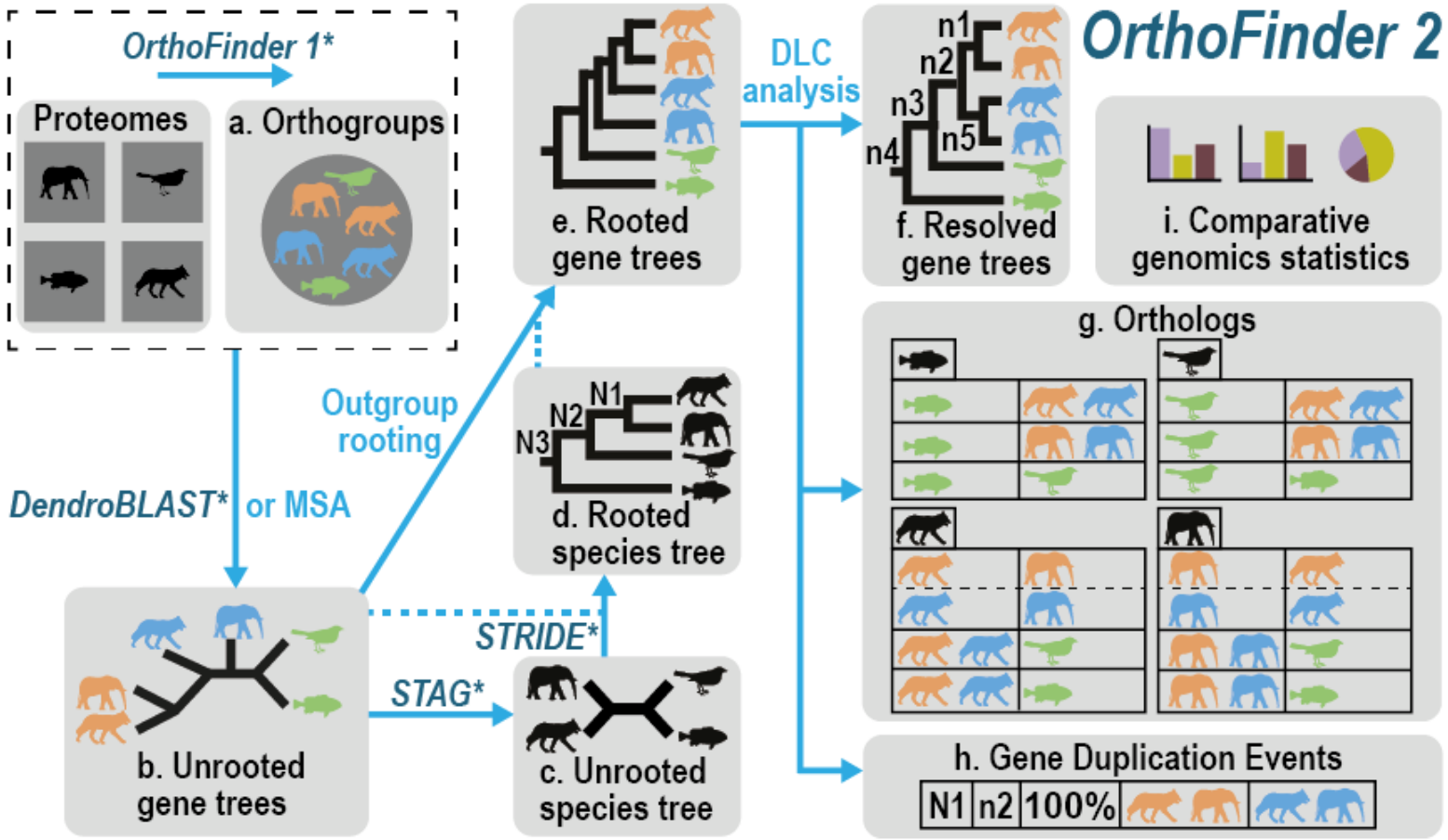
- Gene tree for three species: human, mouse and chicken.
- Orthologs are pairs of genes that descend from a single gene in LCA of two species.
- An **orthogroup** is the extension of this concept to groups of species. An orthogroup is the group of genes descended from a single gene in the LCA of a group of species.
- Genes within an orthogroup may be orthologs of one another or they may be paralogs.

OrthoFinder: phylogenetic orthology inference for comparative genomics

- There are several pipelines/platforms for comparative genomics.
- Orthofinder
  - finds **orthogroups** and **orthologs**
  - infers **rooted gene trees** for all orthogroups
  - identifies all of the **gene duplication events**
  - provides **comprehensive statistics** for comparative genomic analyses
- You need decent genome annotations (sequence and proteome)

https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1832-y (in Perusal)

OrthoFinder: phylogenetic orthology inference for comparative genomics



-Algorithms are shown in italic.
-MSA, multiple sequence alignment-based tree inference.
-DLC, duplication-loss-coalescence

https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1832-y

Summary

- There are many different measures of genome similarity.
- Many will have pros/cons
- We need to understand the underlying assumptions when deciding which measure is appropriate for a particular question.

Whole Genome Alignment

- A critical process in comparative genomics, facilitating the detection of genetic variants and aiding our understanding of evolution.

- Classification of Whole Genome Alignment Algorithms:

    - Suffix Tree-Based Methods
        - MUMmer Technique

    - Anchor based methods
        - LAGAN
        - Mauve
        - BLASTZ
        - LASTZ
        - Minimap2

Whole Genome ~~Alignment~~ Comparison

- A critical process in comparative genomics, facilitating the detection of genetic variants and aiding our understanding of evolution.

- Classification of Whole Genome Comparison Algorithms:

  - Suffix Tree-Based Methods
    - **MUMmer** Technique

  - Anchor based methods
    - **LAGAN**
    - Mauve
    - BLASTZ
    - LASTZ
    - Minimap2

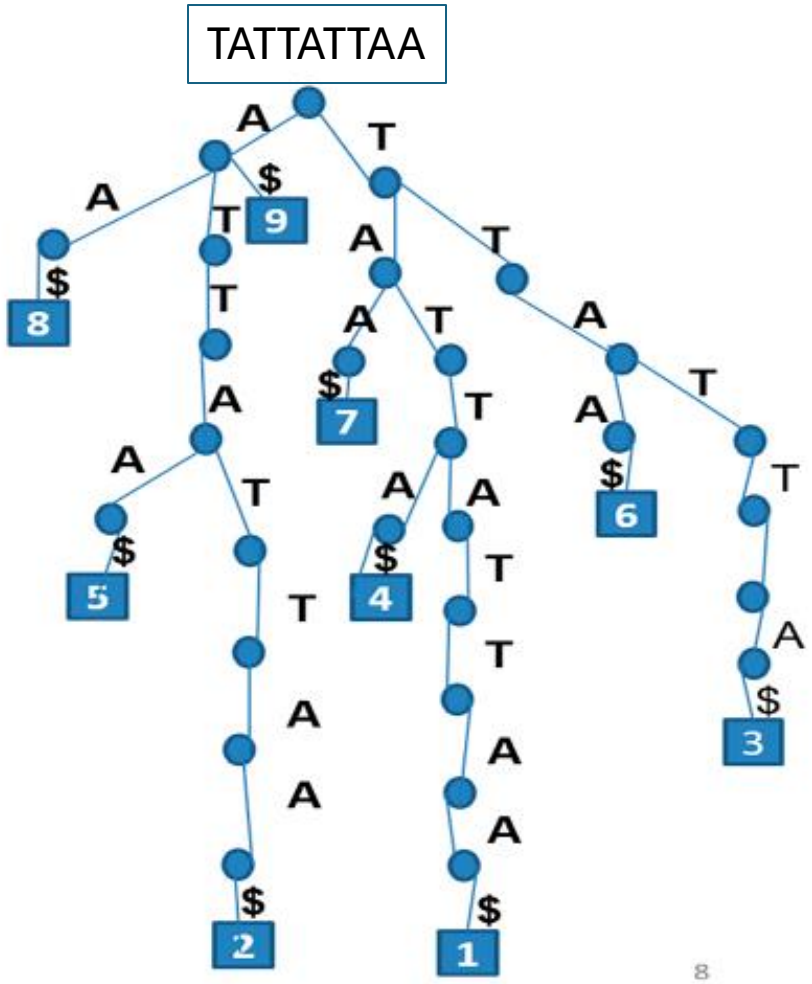
  - Hash-based methods

  - Graph-based methods

There are more!!

Whole Genome Comparison

- Suffix Tree-Based Alignment Methods



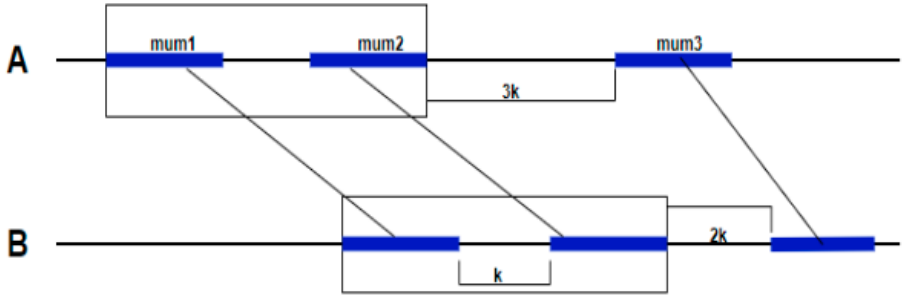
- It is a compressed data tree structure that represents all the suffixes of a given string.
- The trees save their positions in the text as well as their values.
- TATTATTAA sequence tree with 9 suffixes and $ marking the termination of each suffix with a value.
- Data structured in this way provide fast implementations for string operation.

Whole Genome Alignment

- Suffix Tree-Based Alignment Methods:
  - MUMmer uses a suffix tree-based algorithm called Maximal Unique Match (MUM) finding algorithms.
    - The idea is to perform a MUM decomposition of the two genomes.
    - Identify (cluster together) all MUMs between the two genomes.

Genome A: tcgatcGACGATCGCGGCCGTAGATCGAATAACGAGAGAGCATAAcgactta
Genome B: gcattaGACGATCGCGGCCGTAGATCGAATAACGAGAGAGCATAAtccagag

MUM of 39 nucleotides (shown in uppercase)
Any extension of the MUM will result in a mismatch.



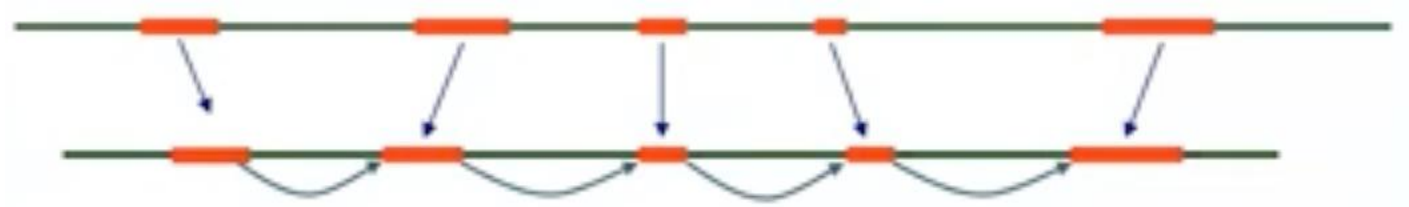
Cluster detected with MUMer v2

Whole Genome Alignment Algorithms

- Anchor based methods

    1. Anchoring: identify regions of strong similarity (local alignment)

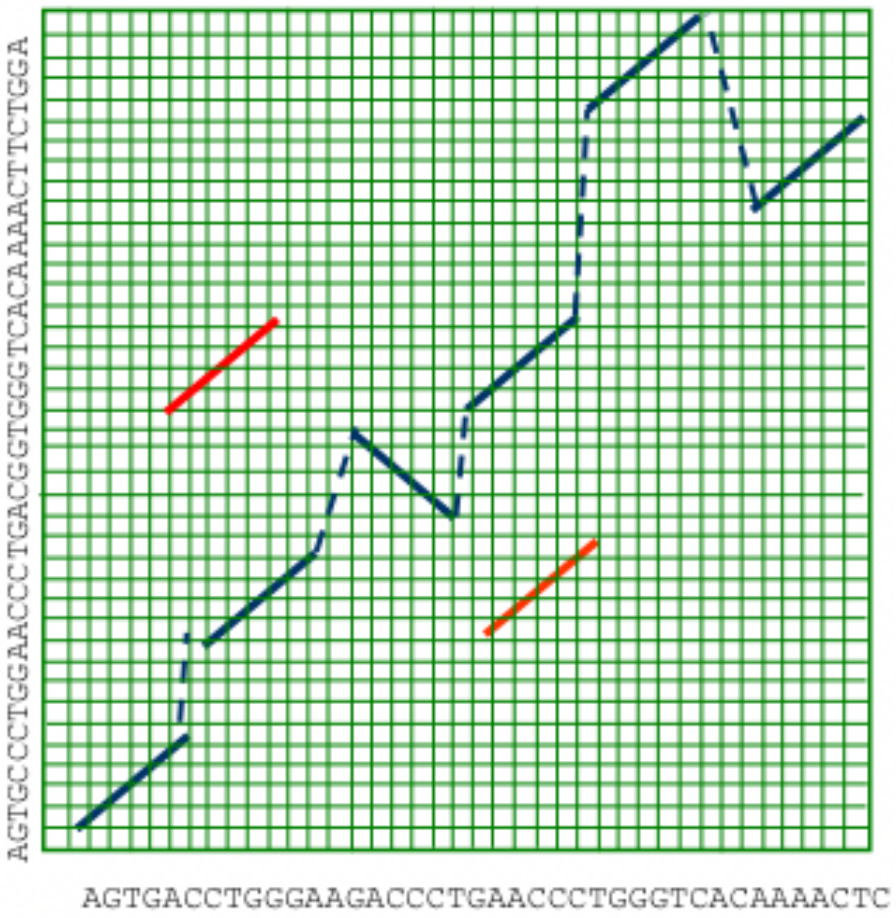
    2. Chaining: join regions of weak or no similarity



- **LAGAN**:
    - Limited Area Global Alignment of Nucleotides algorithm allows for pair-wise alignment of genomic scale sequences.

    - It is an algorithm reminiscent of the FASTA algorithm; a general approach to performing a global alignment.

Whole Genome Alignment Algorithms



- LAGAN steps:
  - Finding local alignments.
  - Chaining local alignments along the diagonal.
  - Perform restricted dynamic programming to find the optimal path.
- Multi-LAGAN: same approach but to multiple species alignment. In the algorithm, we need to prove the genomes and the corresponding phylogenetic tree. It performs pairwise guide with a tree.
- The pros of using global in whole genome alignments are that it allows the identification of duplications, inversions, and translocations.

Source: https://bio.libretexts.org/

Viewing precomputed genome alignments

Vista Alignment Tools https://genome.lbl.gov/vista/index.shtml

The Vista Point suite of tools provides an integrated set of resources for visualizing and exploring Vista DNA alignments. This toolkit allows you to switch among three visualization modes for examining the same alignment:

1.  VistaPoint allows you to navigate within an alignment and examine detailed features of each aligned region.
2.  Vista Synteny Viewer offers a multi-tiered graphical display of a pairwise alignment at three different levels of resolution
3.  VistaDot provides an interactive dot-plot viewer for a pairwise alignment.

Check the help website:
https://pipeline.lbl.gov/vista_help/help.html

VISTA Browser might not work (MACOS)