

Whole genome trees



Overview

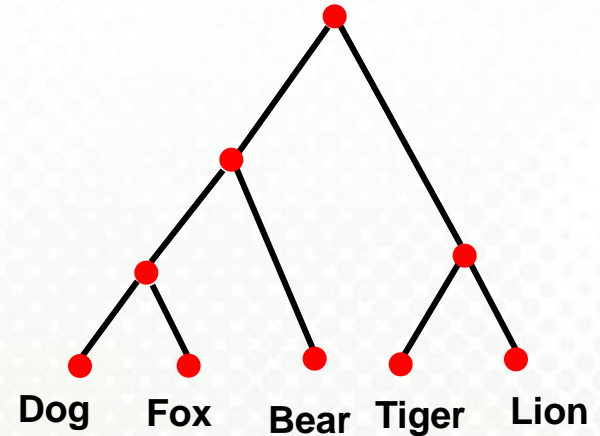
- Why trees?
- Phylogenetic trees
- Whole genome trees
 - MLST
 - Phylogenomics and Super-trees
 - Sequence statistics trees
 - Gene content trees
- Pan-genome tree

Why trees?

- Reconstructing evolutionary history
 - *"In biology nothing makes sense except in the light of evolution"*
 - The Tree of Life
 - Phylogenetic trees
- Graphical display of similarities
 - "Similarity" may have nothing to do with evolution
 - Clustering dendrograms

Phylogenetic trees

- The objects we want to study are called Operational Taxonomic Units (OTU)
- The tree describes the evolutionary relation between the OTUs
- Each OTU is a leaf in the tree
- The length of the edges reflect evolutionary distance



Ways of construction

- Maximum parsimony
 - Given some observations, we look for the simplest possible tree to explain the data
 - Simplest possible means as few evolutionary changes as possible
- Maximum likelihood
 - Requires a probabilistic model for evolutionary changes (PAM)
 - We get a likelihood function value for every tree, given the data
 - Search through “all” possible trees to find the maximum likelihood tree
- Distance based methods
 - Compute distance table
 - Tree-constructing algorithms: UPGMA, Neighbor joining, Fitch-Margoliash...

The classical approach

- Classical phylogenetic trees for prokaryotes is based on the 16S RNA sequence
 - Found in all prokaryotes
 - Slow divergence, provides a look into ancient history
 - Resistant to horizontal transfer

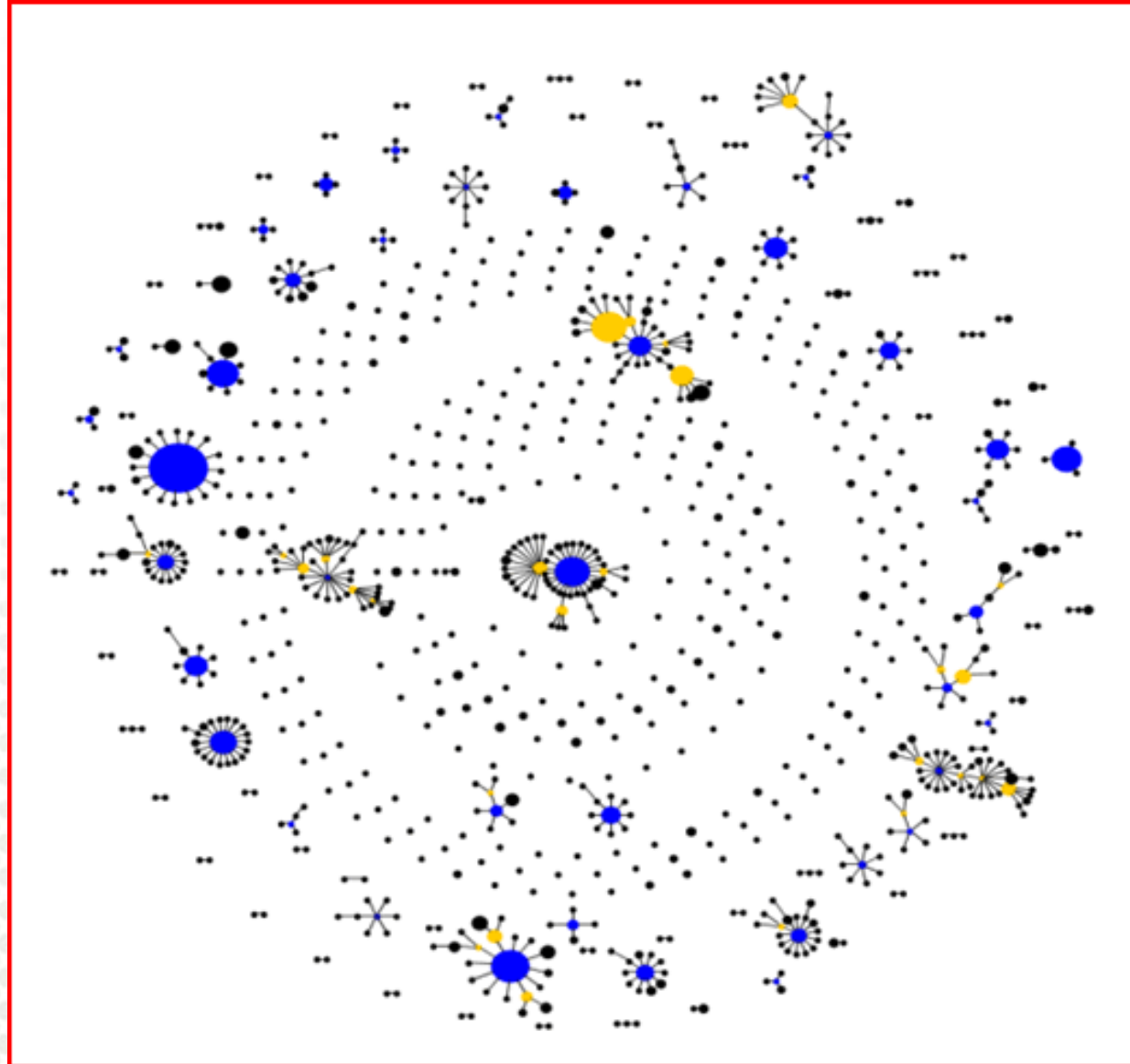
- BUT
 - Only useful for separating highly diverged species
 - How likely is it that the evolution of an organism (genome) is reflected in one single gene?

Example

Multi Locus Sequence Typing (MLST)

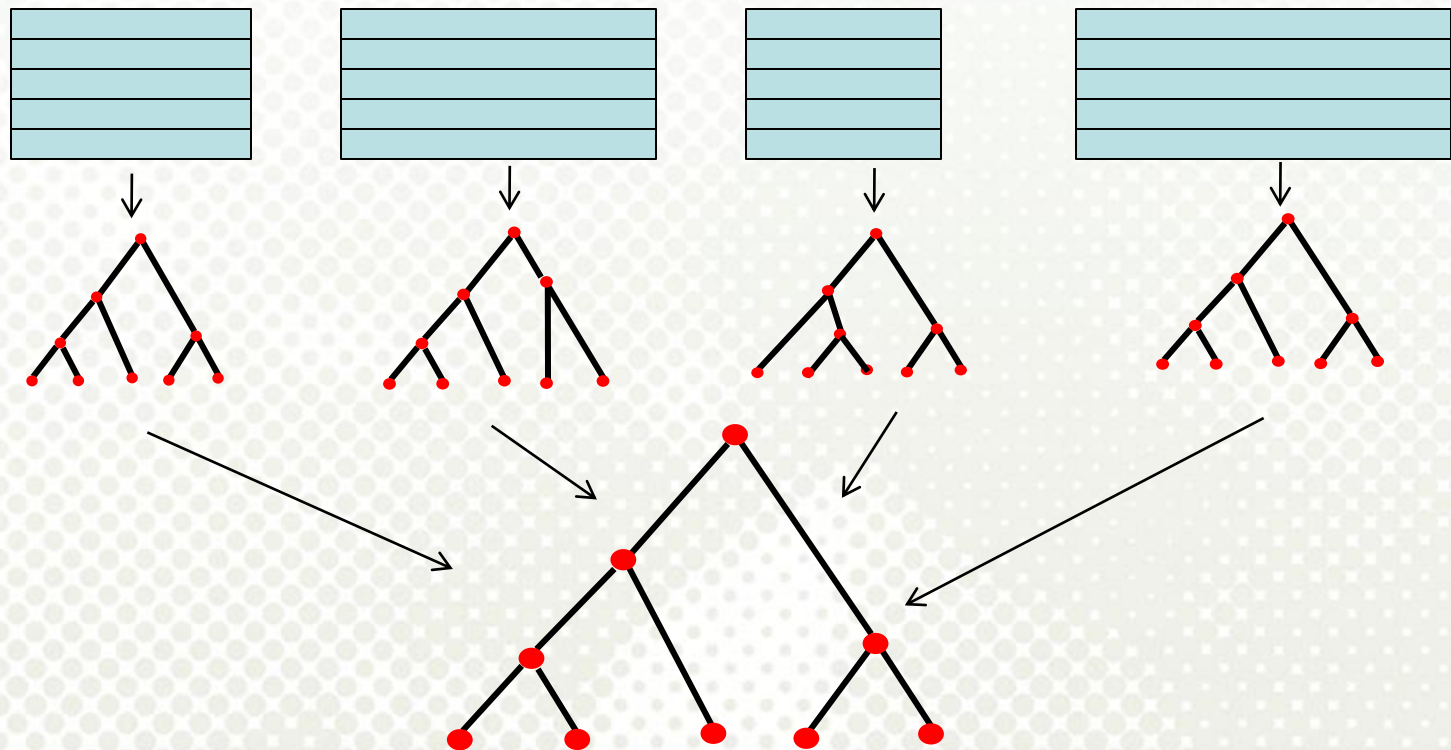
- Based on a set of (usually 7) “housekeeping” genes
 - All strains within a species have these genes, but with some variation
- Each sequence type of each gene is numbered 1,2,3,...
 - The numbers do not reflect how many/big mutations we observe
- Each genome has a sequence of 7 digits
 - Genome 1: (2, 6, 4, 3, 3, 1, 9)
 - Genome 2: (3, 1, 4, 2, 3, 8, 7)
 - Genome 3: (1, 5, 5, 1, 2, 1, 3)
- Distance between genomes = Hamming distance
 $H(1,2)$ = Number of different sequence types

Graphical display – star-shaped trees



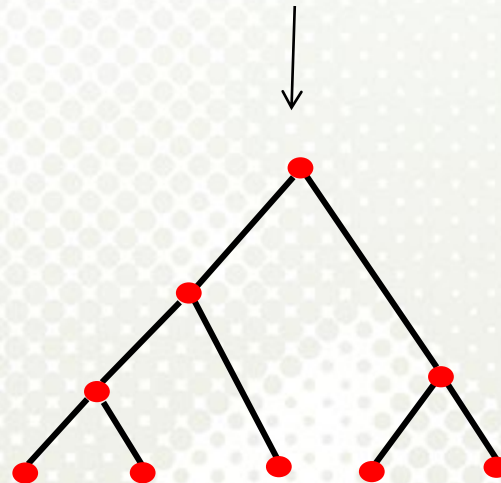
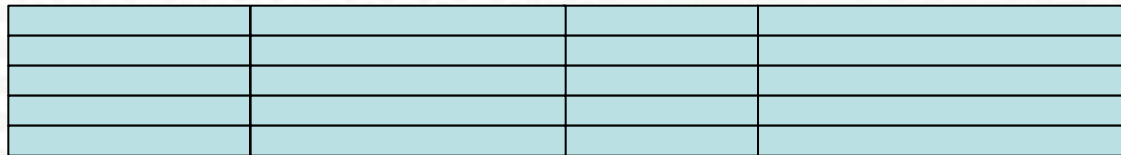
Phylogenomics and Super-trees

- Select a number of genes conserved in the population
- Construct a phylogenetic tree for each gene
- Merge the trees into a super-tree for the entire genome



Alternative approach

- Build one “super-alignment” for the conserved part of the genome
- Construct tree more or less according to classical approaches



Trees based on sequence statistics

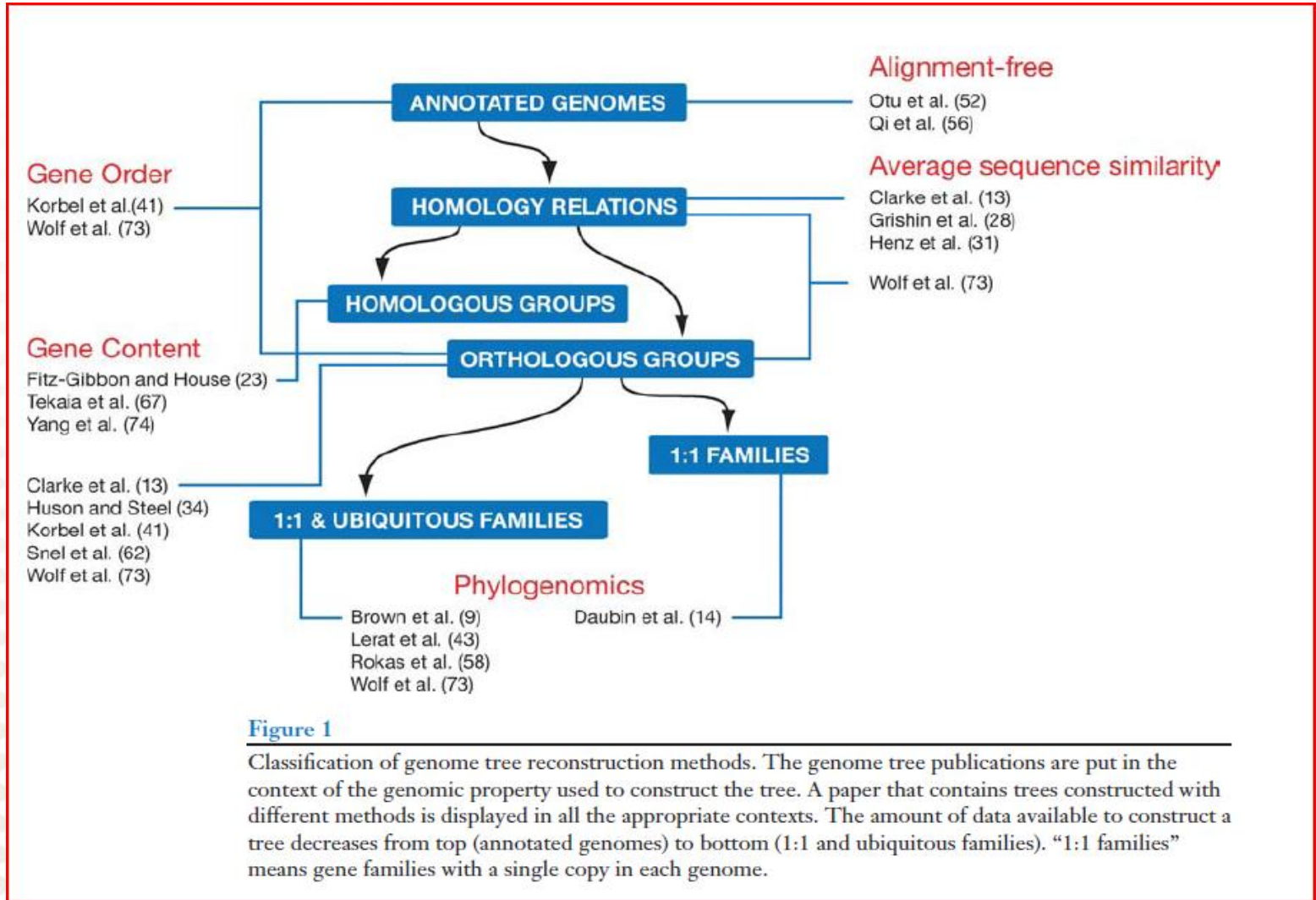
- Compare genomes by counting
 - GC-content
 - dinucleotide frequencies
 - Codon usage
 - Codon bias
 - Etc...
- Alignment free approach
- Use the entire genome sequences, not only the conserved parts
- Difficult to interpret, especially in an evolutionary context

Gene-content trees

- Compare genomes by gene content
- We need a list of gene families present in each genome
 - Genome 1: (gfam1, gfam2, gfam3, gfam4, gfam5,...)
 - Genome 2: (gfam1, gfam2, gfam4, gfam7, gfam8,...)
- Jaccard distance between genome 1 and 2

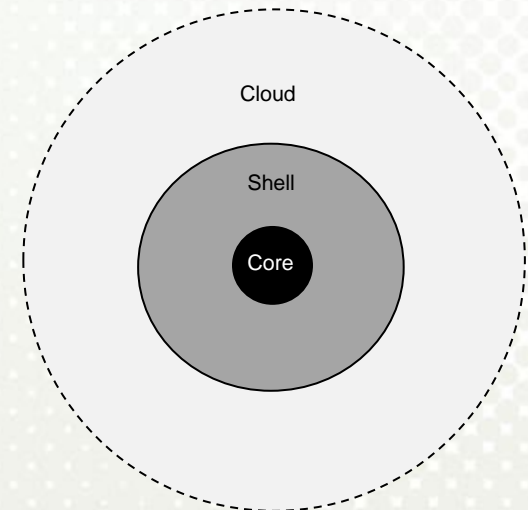
$$J(1,2) = 1 - [\text{number of gfams present in both}]/[\text{total number of unique gfams in both}]$$

Other approaches



Pan-genomes

- The pan-genome of a species (genus) is the set of unique gene families found in at least one genome within the population
- In a pan-genome context genes are classified as
 - **Core genes:** Always present in all genomes (few)
 - **Shell genes:** Usually present in most genomes (many)
 - **Cloud genes:** Rarely present (HUGE pool)



The pan-matrix

	Genome 1	Genome 2	Genome 3	Genome 4	Genome 5
Gfam 1	1	1	0	1	0
Gfam 2	0	1	1	0	0
Gfam 3	1	1	1	1	1
Gfam 4	0	1	0	1	0
Etc...	1	0	0	0	0

- Distance between genomes

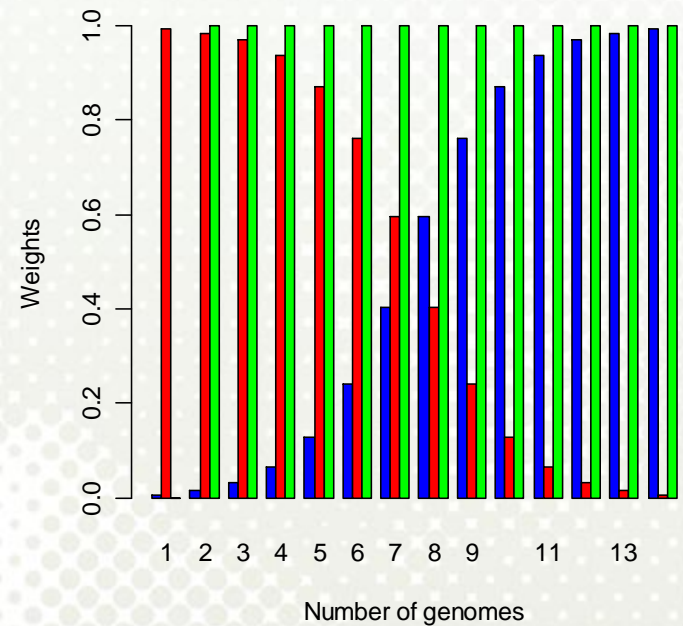
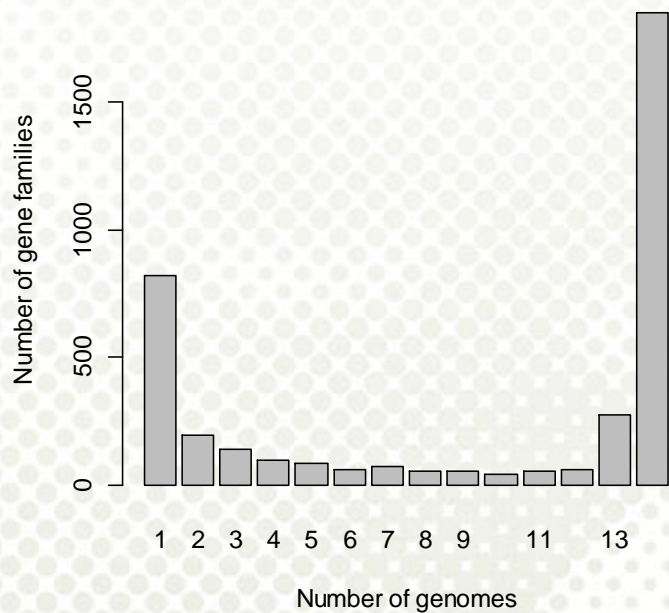
$$D(1,2) = [\text{number of gfam-wise differences}]/[\text{total number of gfams}]$$

- NOTE:

- Shared absence is just as important as shared presence
- Core genes have no impact at all

Weighted distances

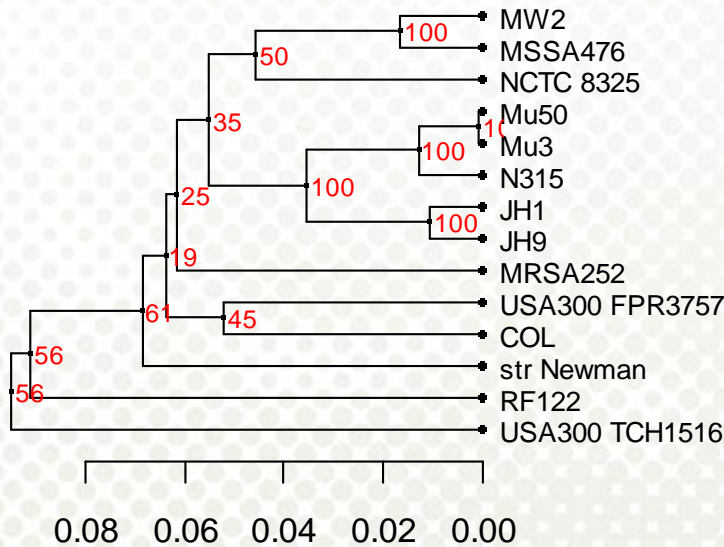
- Down-weighting less important genes
 - ORFans
 - Cloud genes
 - Shell genes



The pan-genome tree

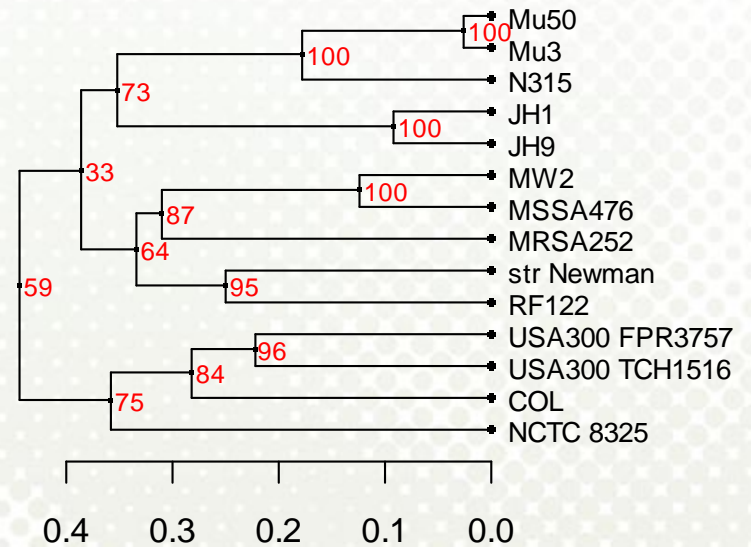
- Display of similarity in gene content inside pan-genomes
- Possibly weighted to emphasize parts of the pan-genome

S. aureus, shell



Relative manhattan distance

S. aureus, cloud



Relative manhattan distance

Streptococcus pangenome tree

