WASHINGTON STATE UNIVERSITY

A Study of All Recognized Bacterial Phyla Using Network Science and Whole Proteome Clustering

Ehdieh Khaledian

Assefaw H. Gebremedhin, Kelly A. Brayton, and

Shira L. Broschat

Oct 28, 2018

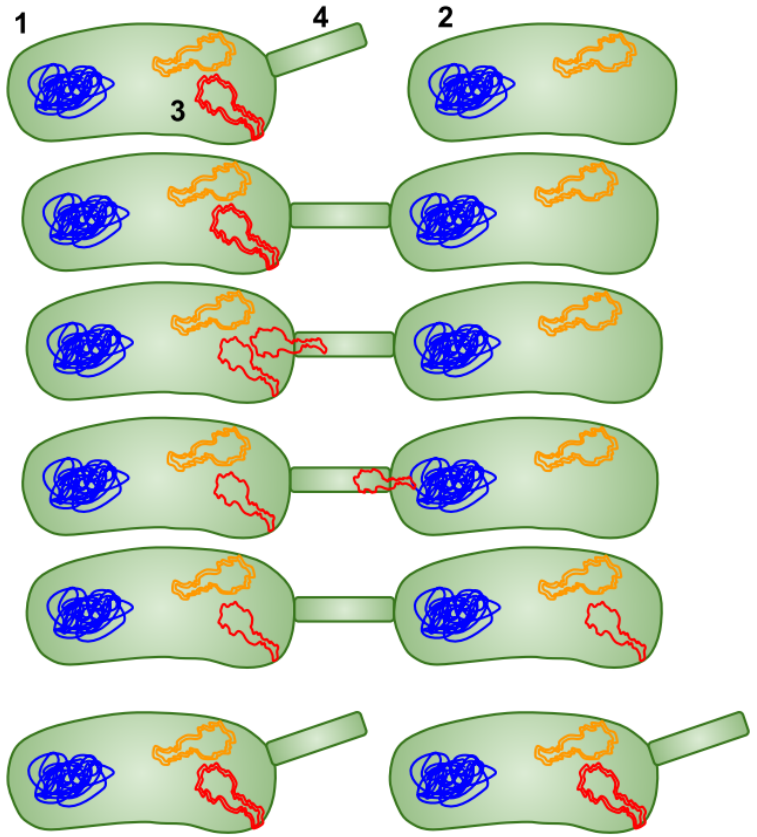
StringBio 2018

Introduction

- Organisms diversify through the process of evolution
- Biologists are interested in reconstructing the evolutionary path
- We are interested in the relationships of bacteria
- Goal: extract the information from network of Organisms and network of phyla

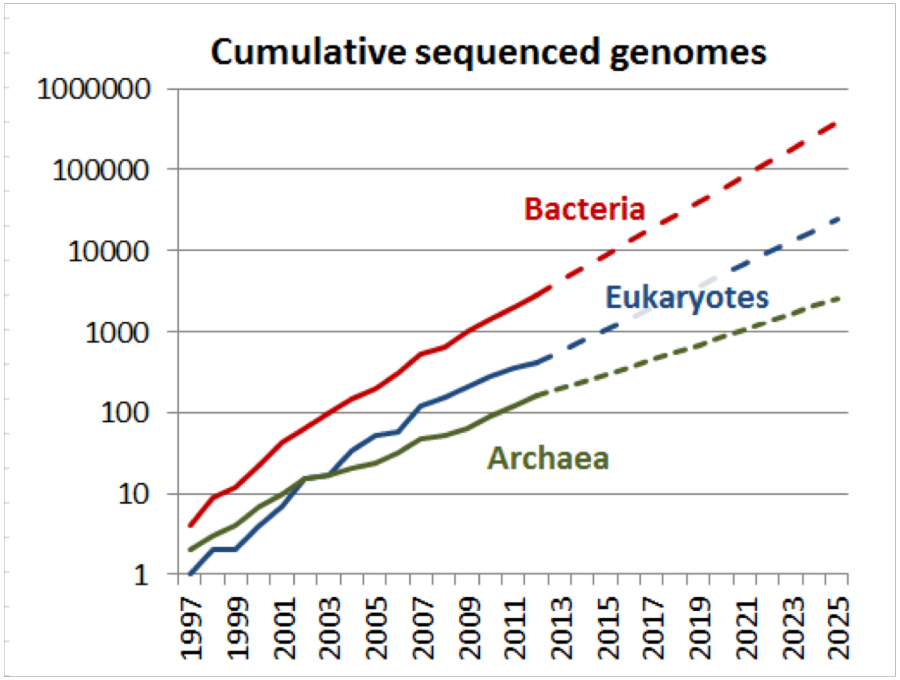
Approach

- Look at whole genome sequences
  - Old approach: A few genes
  - New approach: Entire genomes
    - Horizontal gene transfer
- Look at the network of organisms instead of tree



What enables us …

- Availability of whole genome sequences
  - > 7000 genomes
- Next Generation Sequencing
- Clustering using pClust software



Dataset

- **210** complete genome sequences
- **733,227** protein sequences
- **29** major recognized phyla
  - 28 available at time of study
- Overrepresentation of Proteobacteria
  - We separated this phylum into classes
- Operational taxonomic units (OTUs, **31** groups)
  - Bacteroidetes/Chlorobi grouped by NCBI

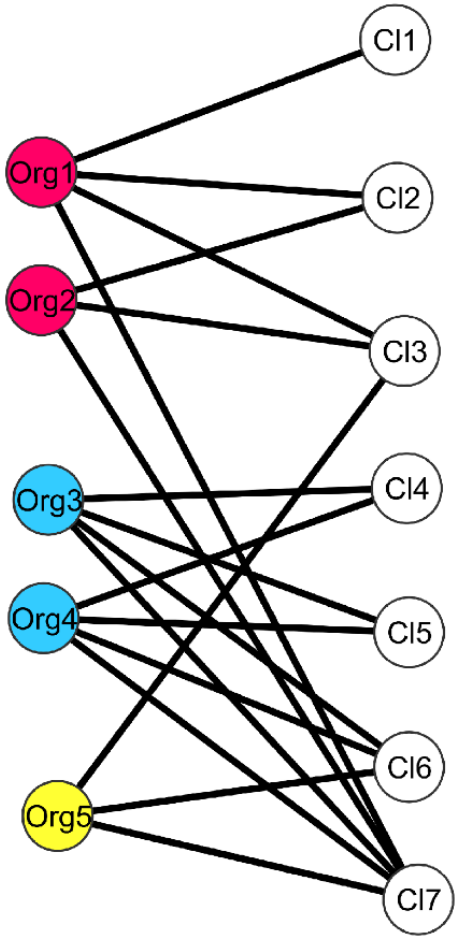
Clustering

- pClust for clustering:
  - Semi-global alignment
  - Louvain community detection
- 61,057 **non-singleton** clusters
- Exclude Singletons

Creating Networks

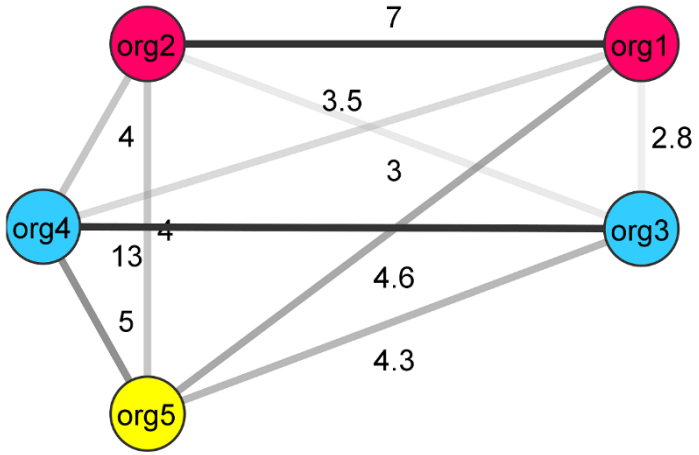
- Bipartite graph of the organism-cluster relationship:
- Compute Manhattan distance and normalize

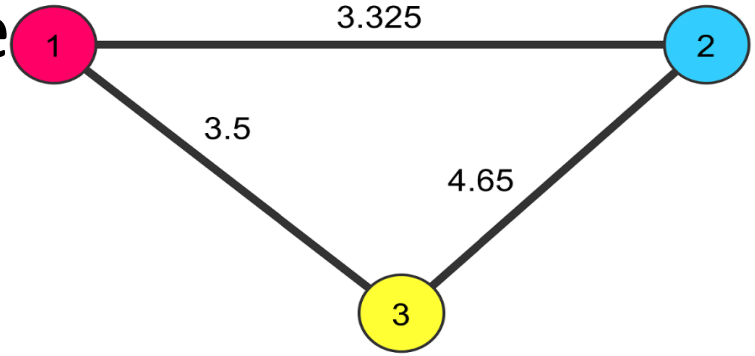
$$d(p,q) = ||p - q|| = \sum_{i=1}^{n} |p_i - q_i|$$



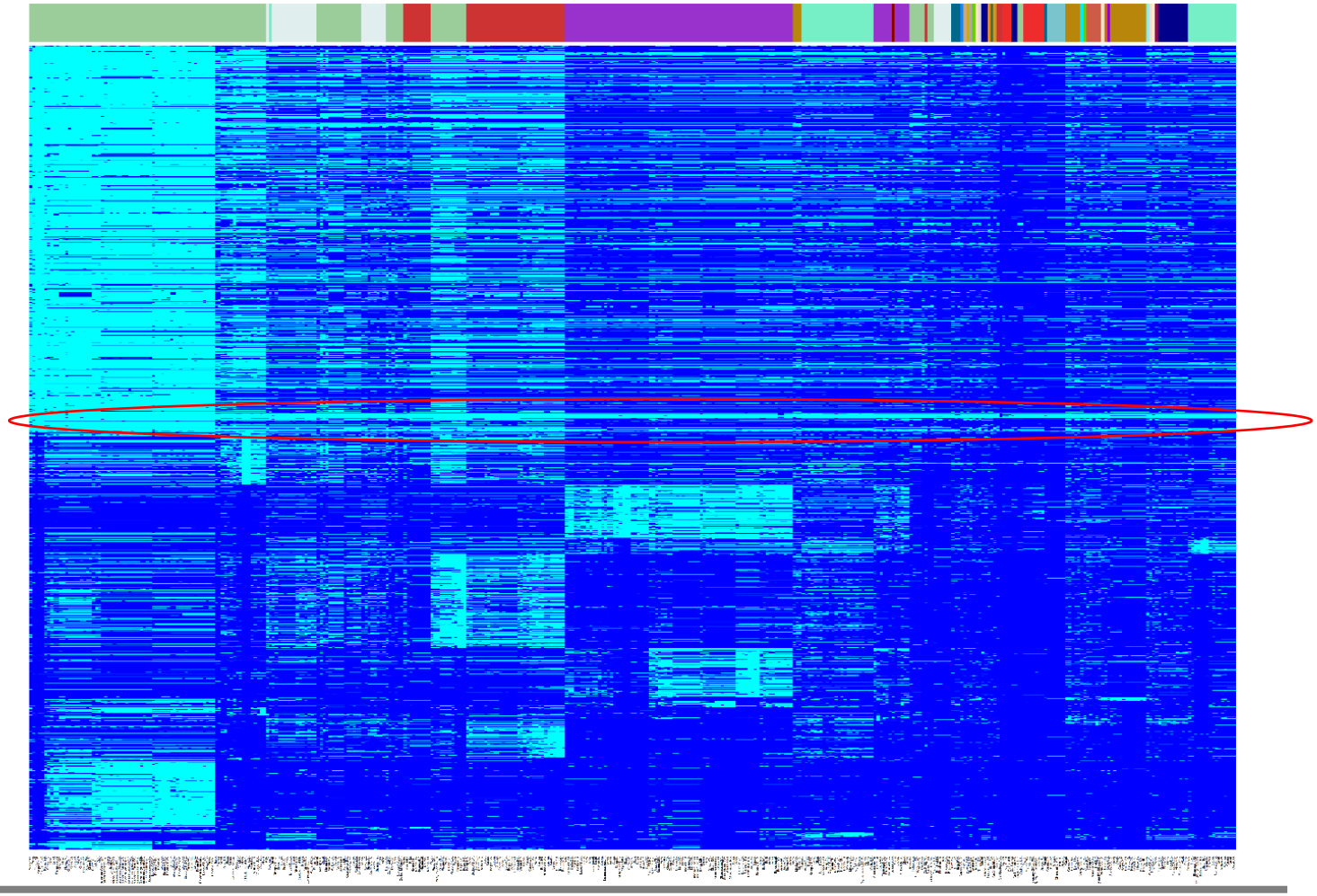
Creating Networks

- Create Network of Phyla from Network of Organisms using average distance
- Apply BFS to extract the Tree of Phyla





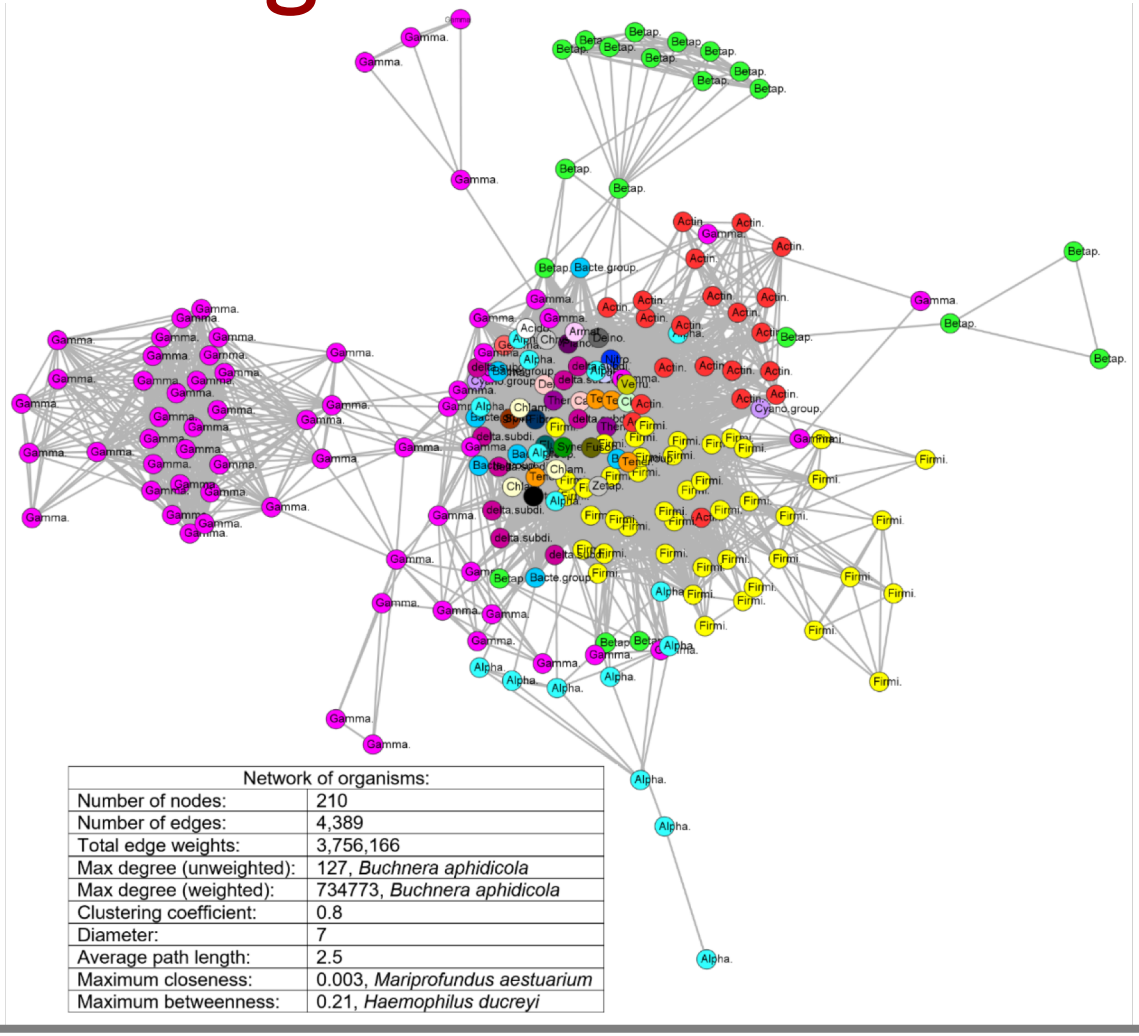
Heatmap



Minimum gene set

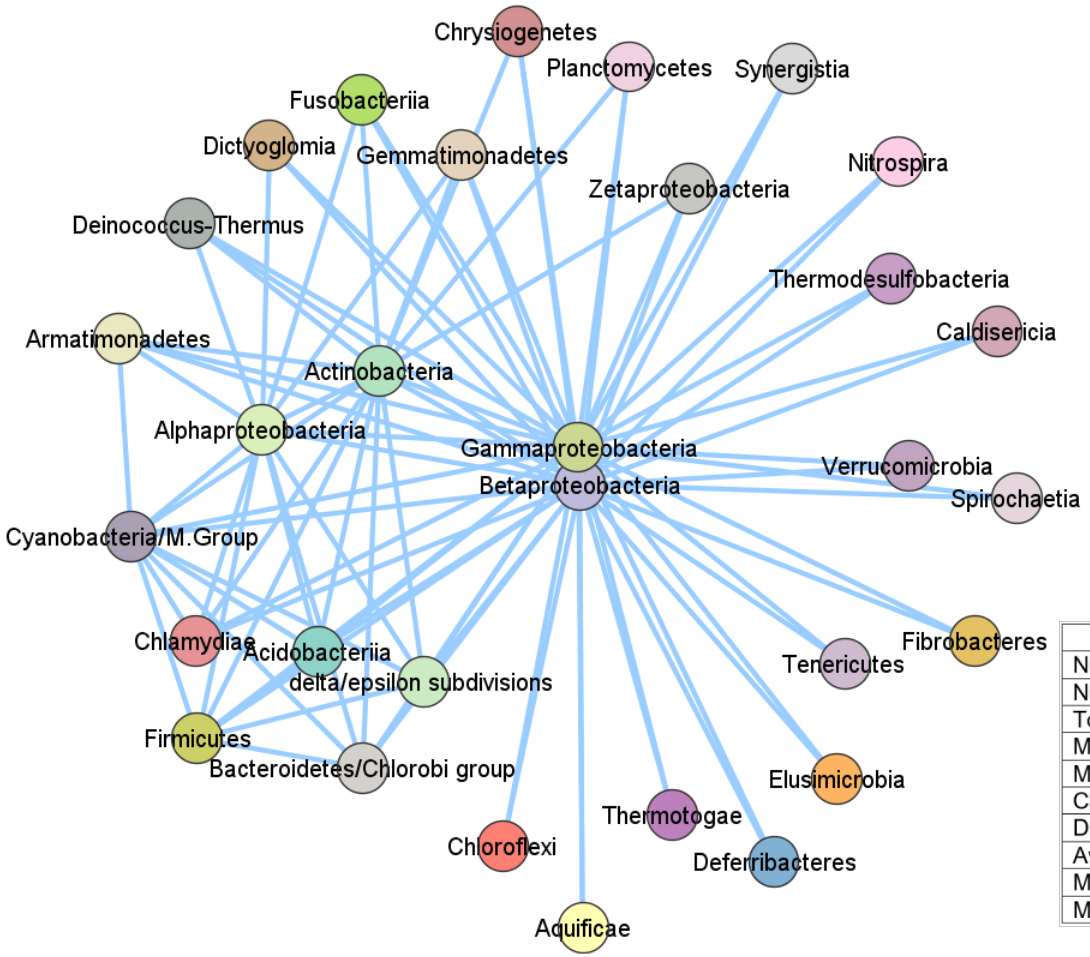
Number of Sequences in Cluster	Proteins
400	30S ribosomal protein S13
413	30S ribosomal protein S3
402	30S ribosomal protein S5
403	30S ribosomal protein S7
403	30S ribosomal protein S8
400	30S ribosomal protein S9
409	30S ribosomal subunit protein S4
400	50S ribosomal protein L14
399	50S ribosomal protein L16
399	50S ribosomal protein L17
442	50S ribosomal protein L2 putative dihydrodipicolinate synthase
402	50S ribosomal protein L20
400	50S ribosomal protein L5
402	50S ribosomal protein L6
510	DNA gyrase, subunit A DNA topoisomerase IV
694	DNA gyrase, subunit B DNA gyrase subunit B, type II topoisomerase topoisomerase IV subunit B
531	GTP-binding protein chain elongation factor EF-G translation elongation factor 2 (EF-2/EF-G)
778	peptide chain release factor RF-1, peptide chain release factor RF-2
596	protein chain elongation factor EF-Tu
410	seryl-tRNA synthetase

Network of Organisms



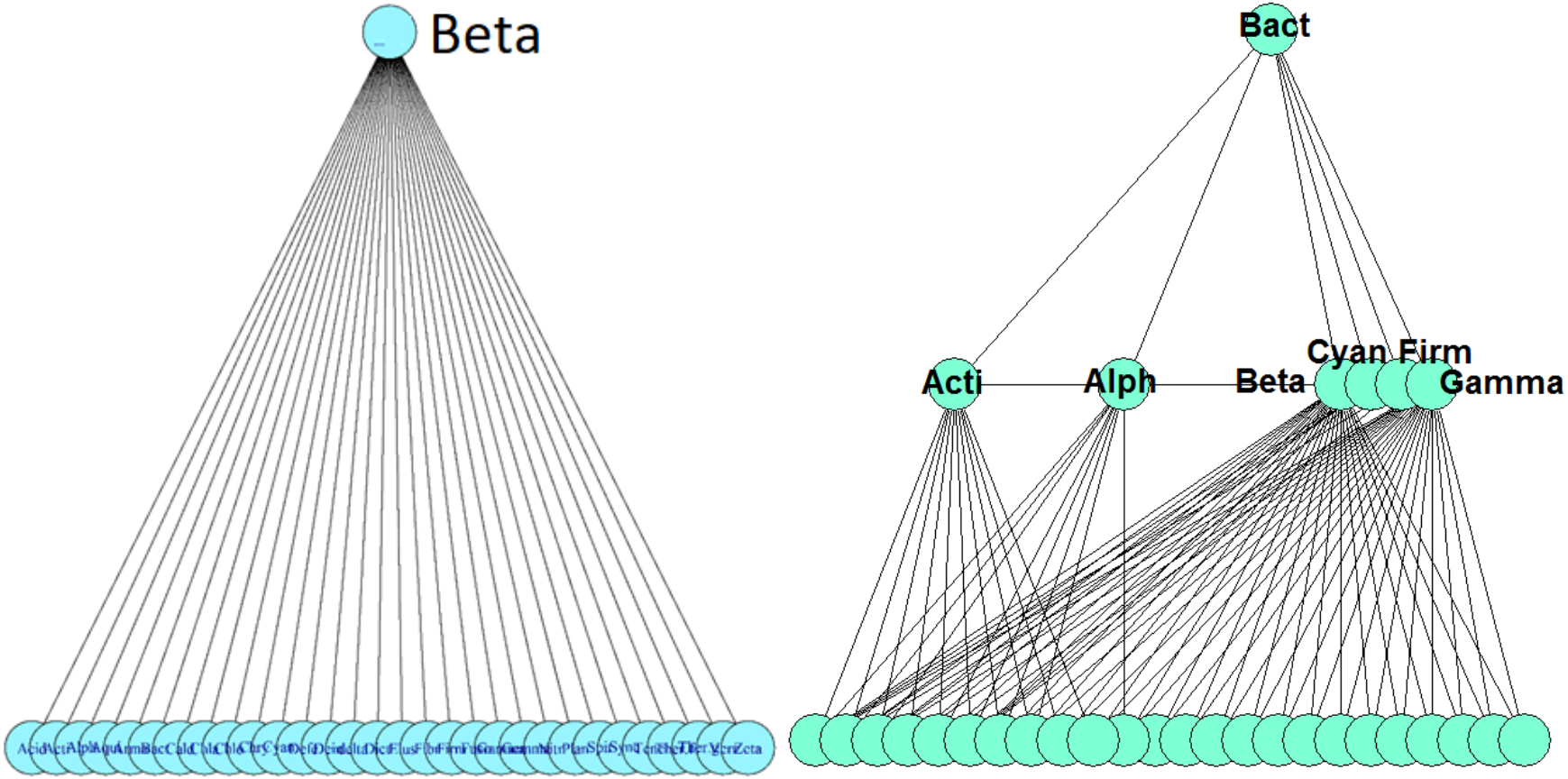
Network of organisms:	
Number of nodes:	210
Number of edges:	4,389
Total edge weights:	3,756,166
Max degree (unweighted):	127, *Buchnera aphidicola*
Max degree (weighted):	734773, *Buchnera aphidicola*
Clustering coefficient:	0.8
Diameter:	7
Average path length:	2.5
Maximum closeness:	0.003, *Mariprofundus aestuarium*
Maximum betweenness:	0.21, *Haemophilus ducreyi*

Network of Bacterial Phyla

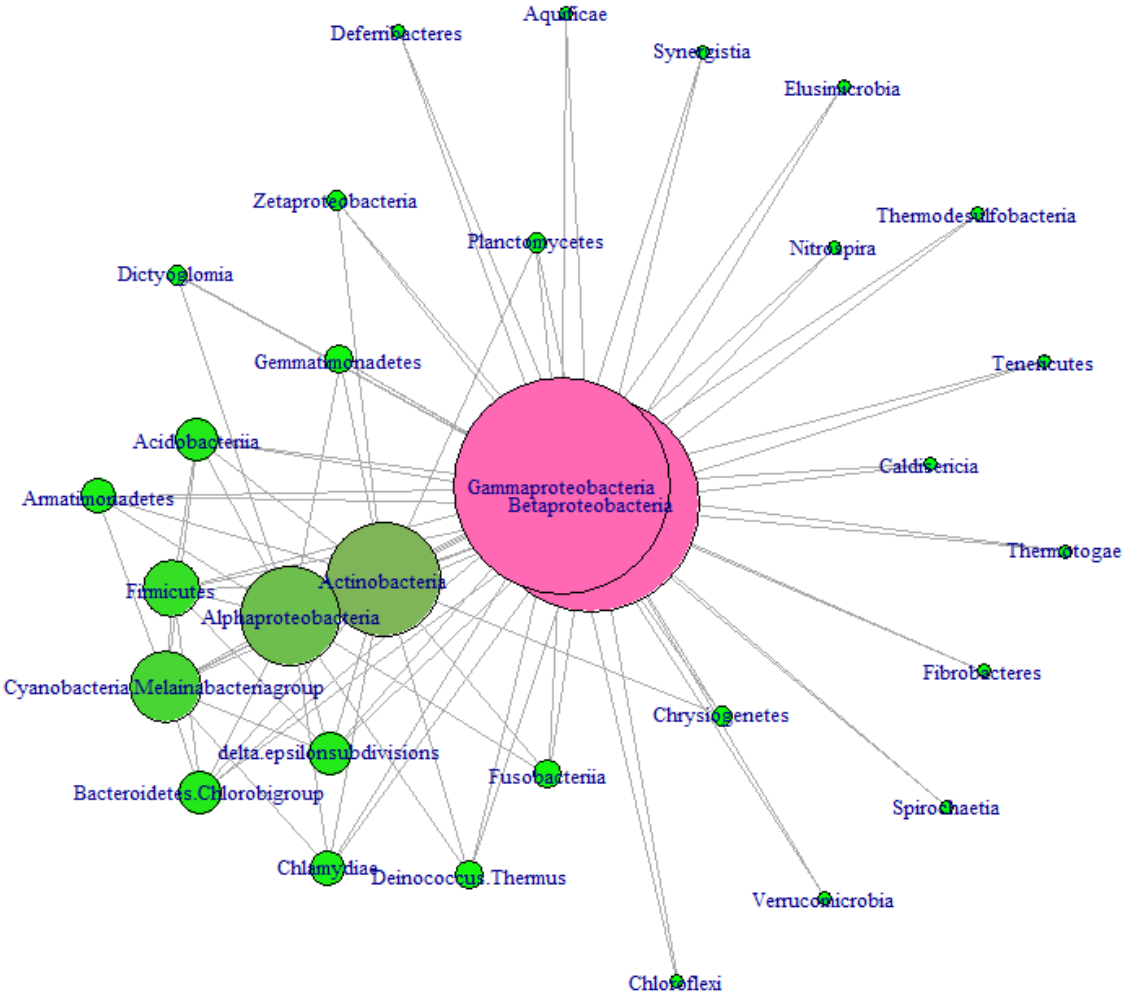


Network of bacterial phyla:	
Number of nodes:	31
Number of edges:	93
Total edge weights:	823,870
Max degree (unweighted):	30, Beta- and Gammaproteobacteria
Max degree (weighted):	106,208, Gammaproteobacteria
Clustering coefficient:	0.3
Diameter:	2
Average path length:	1.8
Maximum closeness:	1, Beta- and Gammaproteobacteria
Maximum betweenness:	0.38, Beta- and Gammaproteobacteria

Tree of Bacterial Phyla



Centrality Analysis for Network of OTUs



Conclusion

- **Centrality measures show**:
  - Most important organism is ***B. aphidicola***
    - Endosymbiont belonging to Gammaproteobacteria
    - Small genome (< 1 Mb) with a low G+C content
    - Reduced genome
    - Relatively conserved genes/proteins

Conclusions (cont.)

- **Application of breadth-first search (BFS) gives**:
  - Oldest OTUs
    - Betaproteobacteria
    - Gammaproteobacteria
  - Proteobacteria as the oldest phylum

Thank you