Comparison of Microbial Comparative Genomics using Bacteriophages and Mycoplasma bacteria

Presented by: Elizabeth Helton

Overview

-What is a genome, gene, and bacteriophage?

-Glimpse at Bioconductor

-What is Comparative Genomics?

-Bacteriophage Dataset

- Package ‘Find my Friends’

-examples

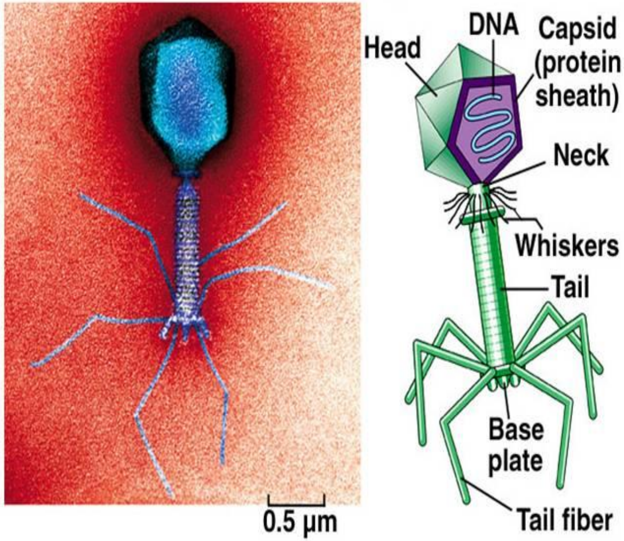
-summary

Background Info

-Genome: Organism’s complete set of DNA, which includes all of its genes and noncoding sequences

-Gene: sequence of DNA or RNA that codes for a molecule with a function (ex.proteins)

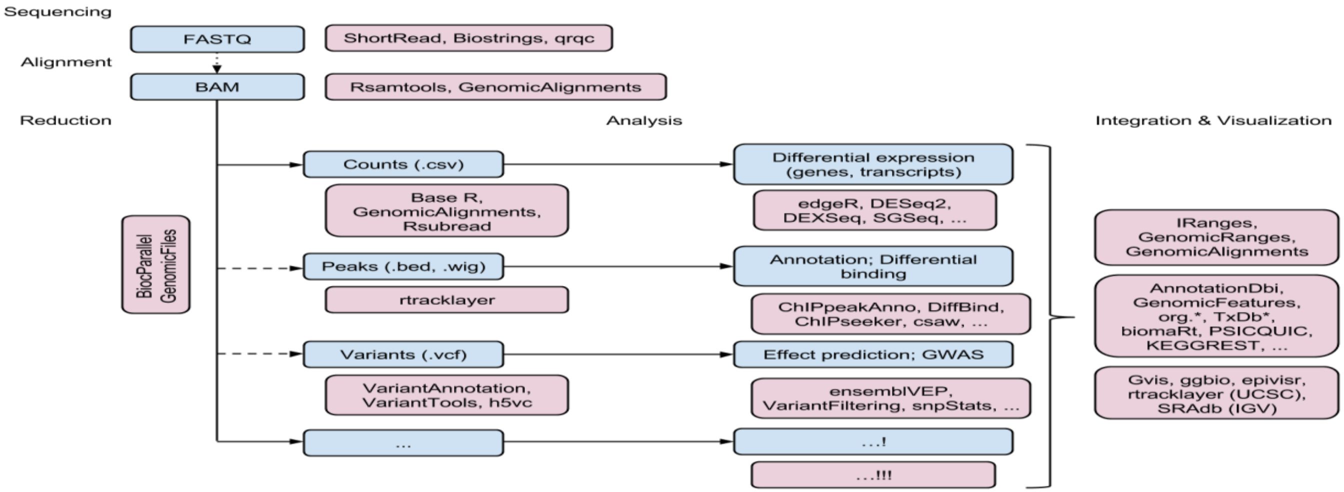
-Bacteriophage:a type of small virus that uses bacteria as a host cell, and destroys the bacteria cell



Bioconductor

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

- Used for analysis, comprehension, and visual aid of genomic data. It is an ope source and open developmental software program. It’s primarily used in R programming. Bioconductor uses packages to solve various issues.

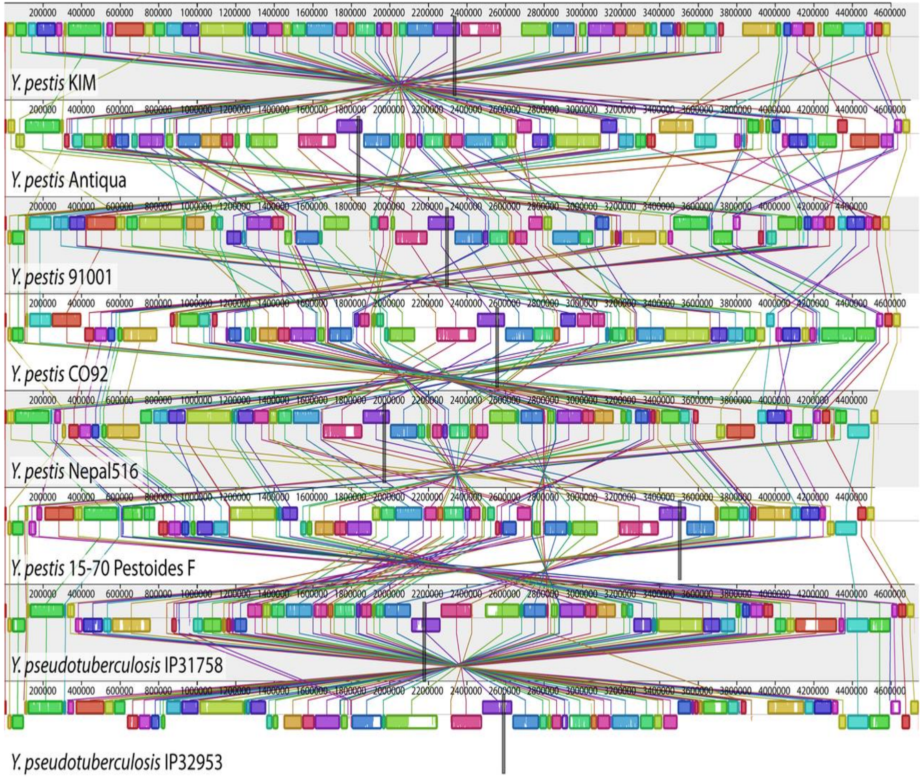

Sequencing
FASTQ
ShortRead, Biostrings, qrqc
Alignment
BAM
Rsamtools, GenomicAlignments
Reduction
Analysis
Integration & Visualization
Counts (.csv)
Base R, GenomicAlignments, Rsubread
Differential expression (genes, transcripts)
edgeR, DESeq2, DEXSeq, SGSeq, ...
BiocParallel
GenomicFiles
Peaks (.bed, .wig)
rtracklayer
Annotation; Differential binding
ChIPpeakAnno, DiffBind, ChIPseeker, csaw, ...
Variants (.vcf)
VariantAnnotation, VariantTools, h5vc
Effect prediction; GWAS
ensemblVEP, VariantFiltering, snpStats, ...
...
...!
...!!!
IRanges, GenomicRanges, GenomicAlignments
AnnotationDbi, GenomicFeatures, org.*, TxDb*, biomaRt, PSICQUIC, KEGGREST, ...
Gvis, ggbio, epivisr, rtracklayer (UCSC), SRAdb (IGV)

Comparative genomics

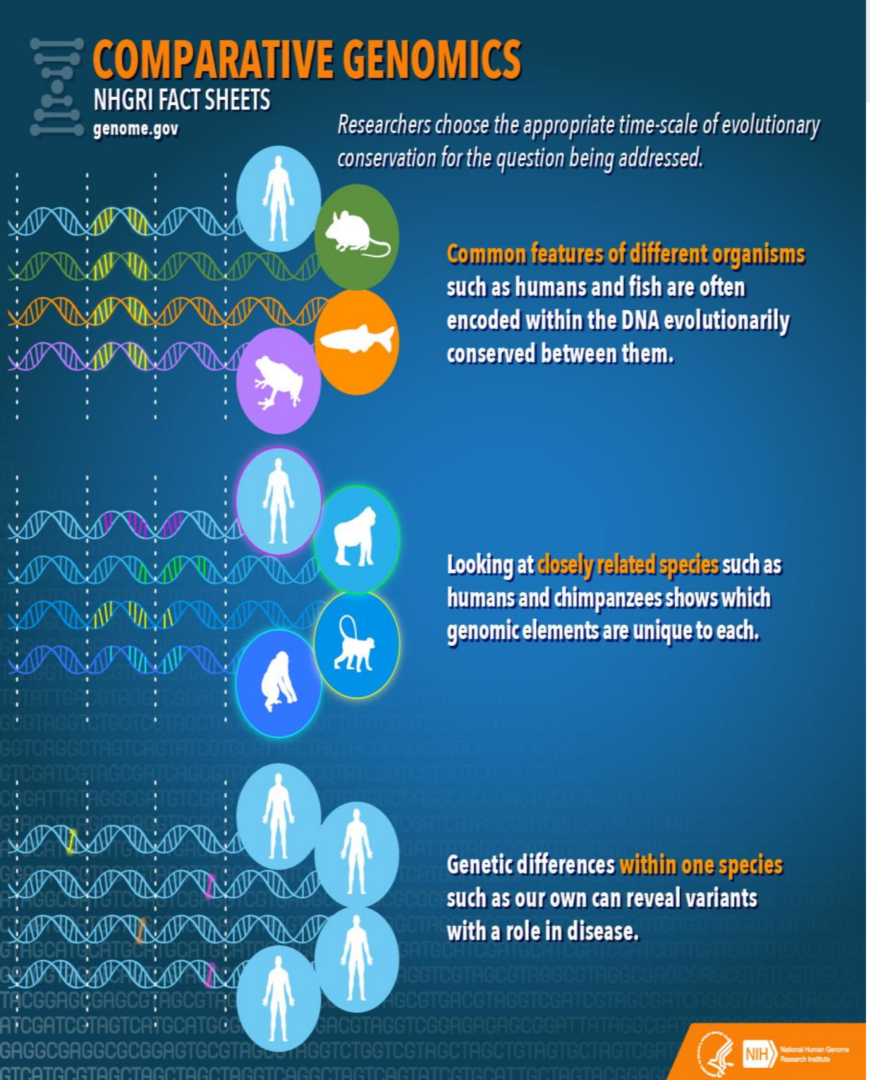
-Used to compare complete genome sequences of various species

-Able to identify regions of similarity and differences between species

-Used to better understand the structure and function of human genes and come up with new ways to fight diseases



Y. pestis KIM
Y. pestis Antiqua
Y. pestis 91001
Y. pestis CO92
Y. pestis Nepal516
Y. pestis 15-70 Pestoides F
Y. pseudotuberculosis IP31758
Y. pseudotuberculosis IP32953



COMPARATIVE GENOMICS

NHGRI FACT SHEETS

genome.gov

Researchers choose the appropriate time-scale of evolutionary conservation for the question being addressed.

Common features of different organisms such as humans and fish are often encoded within the DNA evolutionarily conserved between them.

Looking at **closely related species** such as humans and chimpanzees shows which genomic elements are unique to each.

Genetic differences **within one species** such as our own can reveal variants with a role in disease.

NIH
National Human Genome Research Institute

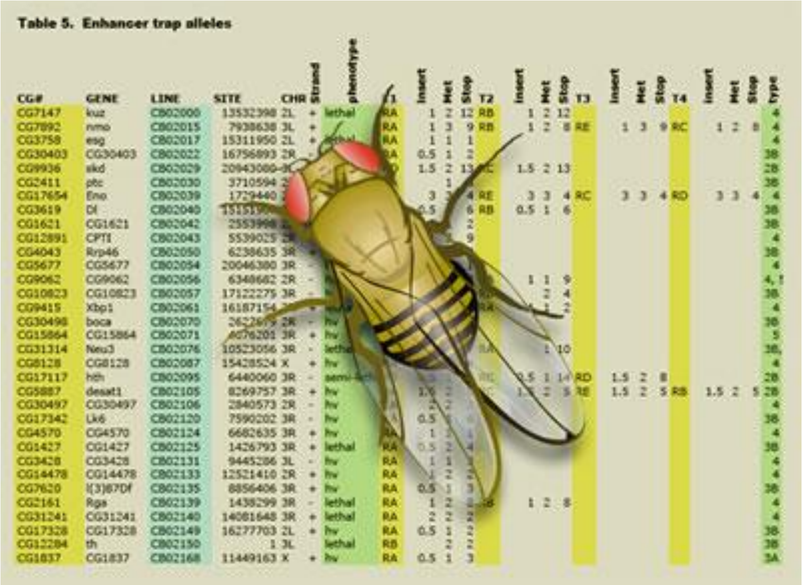


Table 5. Enhancer trap alleles

CG#	GENE	LINE	SITE	CHR	Strand	phenotype	T1	Insert	Met	Stop	T2	Insert	Met	Stop	T3	Insert	Met	Stop	T4	Insert	Met	Stop	Type
CG7147	kuz	CB02000	13532398	2L	+	lethal	RA	1	2	12	RB	1	2	12									4
CG7892	nmo	CB02015	7938638	3L	+	[illegible]	RA	1	3	9	RB	1	2	8	RE	1	3	9	RC	1	2	8	4
CG3758	esg	CB02017	15311950	2L	+	[illegible]	RA	1	1	1													4
CG30403	CG30403	CB02022	16756893	2R	-	[illegible]	[illegible]	0.5	1	2													3B
CG9936	skd	CB02029	20943080	[illegible]	[illegible]	[illegible]	[illegible]	1.5	2	13	[illegible]	1.5	2	13									2B
CG2411	ptc	CB02030	3710594	[illegible]	[illegible]	[illegible]	[illegible]		1														3B
CG17654	Eno	CB02039	1729440	[illegible]	[illegible]	[illegible]	[illegible]	3	[illegible]	4	RE	3	3	4	RC	3	3	4	RD	3	3	4	4
CG3619	Dl	CB02040	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.5	[illegible]	6	RB	0.5	1	6									3B
CG1621	CG1621	CB02042	2553998	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	2													3B
CG12891	CPTI	CB02043	5539025	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	9													4
CG4043	Rrp46	CB02050	6238635	3R	+	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]													3B
CG5677	CG5677	CB02054	20046380	3R	-	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]													4
CG9062	CG9062	CB02056	6348682	2R	-	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1	1	9									4, [illegible]
CG10823	CG10823	CB02057	17122275	3R	-	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]		2	4									3B
CG9415	Xbp1	CB02061	16187154	[illegible]	+	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	RA	[illegible]	1	2									4
CG30498	boca	CB02070	2622679	2R	-	hw	[illegible]	[illegible]	[illegible]	[illegible]													3B
CG15864	CG15864	CB02071	[illegible]	3R	+	hw	[illegible]	[illegible]	[illegible]	[illegible]													5
CG31314	Neu3	CB02076	10523056	3R	-	lethal	[illegible]	[illegible]	[illegible]	[illegible]	RA		1	10									3B,
CG8128	CG8128	CB02087	15428524	X	+	hw	[illegible]	[illegible]	[illegible]	[illegible]													4
CG17117	hth	CB02095	6440060	3R	-	semi-[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	RC	0.5	1	14	RD	1.5	2	8					2B
CG5887	desat1	CB02105	8269757	3R	+	hw	[illegible]	1.5	[illegible]	[illegible]	[illegible]	1.5	2	5	RE	1.5	2	5	RB	1.5	2	5	2B
CG30497	CG30497	CB02106	2840573	2R	-	hw	[illegible]	[illegible]	2	[illegible]													4
CG17342	Lk6	CB02120	7590202	3R	-	hw	[illegible]	0.5	[illegible]	[illegible]													3B
CG4570	CG4570	CB02124	6682635	3R	+	hw	RA	1	[illegible]	[illegible]													4
CG1427	CG1427	CB02125	1426793	3R	+	lethal	RA	0.5	2	4													3B
CG3428	CG3428	CB02131	9445286	3L	-	hw	RA	1	1	[illegible]													4
CG14478	CG14478	CB02133	12521410	2R	+	hw	RA	1	2	[illegible]													4
CG7620	l(3)87Df	CB02135	8856406	3R	+	hw	RA	0.5	1	3													3B
CG2161	Rga	CB02139	1438299	3R	-	lethal	RA	1	2	[illegible]	[illegible]	1	2	8									4
CG31241	CG31241	CB02140	14081648	3R	+	lethal	RA	2	2	2													4
CG17328	CG17328	CB02149	16277703	2L	+	hw	RA	0.5	1	2													3B
CG12284	th	CB02150	1	3L		lethal	RB		2	2													3B
CG1837	CG1837	CB02168	11449163	X	+	hw	RA	0.5	1	3													3A

Bacteriophage Dataset

- 10 Bacteriophages coming from the Mycobacterium host genus(2 of them were discovered at Webster University)
- Came from Actinobacteriophage Database
- This database shares data, pictures, protocols and analysis tools that were used in the discovery, sequencing and characterization of the phages.
- Bacteriophages Used: Bobby, Cjw1, Dori, Giles, Kalah2, Lilbit, Petra64142, ShereKhan, Spongebob, Webster2

Kalah2

Bobby

Find my Friends/ comparison

-Framework for microbial comparative genomics. Defines a class system for when working with a pangenome datasets. It allows for a transparency to the underlying sequence data while being able to handle massive collections of genomes.

-Defines a set of novel algorithms that make it possible to create a high quality and speedy pangenome sequence.

```
GATTCGATTAG  ->  ATT: 2
                 CGA: 1
                 GAT: 2
                 TAG: 1
                 TCG: 1
                 TTA: 1
                 TTC: 1
```

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|_2\|\mathbf{B}\|_2} = \frac{\sum\limits_{i=1}^{n} A_i B_i}{\sqrt{\sum\limits_{i=1}^{n} A_i^2}\sqrt{\sum\limits_{i=1}^{n} B_i^2}}$$

Find My Friends Using Bacteriophages Genomes

-cdhitGrouping used to calculate pangenomes. cdhitGrouping repeatedly combines gene groups based on lower similarity thresholds. During each step the longest member in each of the gene groups becomes the model for the next step. It is best to use the lowest threshold possible to ensure that genes that are in the same group can be clustered together

```
> mypang
An object of class pgFull

The pangenome consists of 10
genes from 10 organisms
5 gene groups defined


     Core|
Accessory|================
========================
Singleton|==========

Genes are translated
```

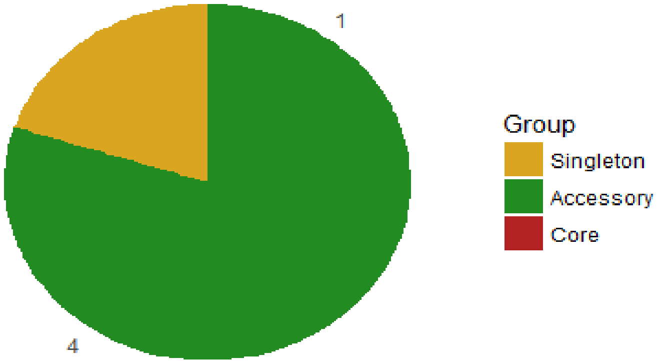
ExpressionSet

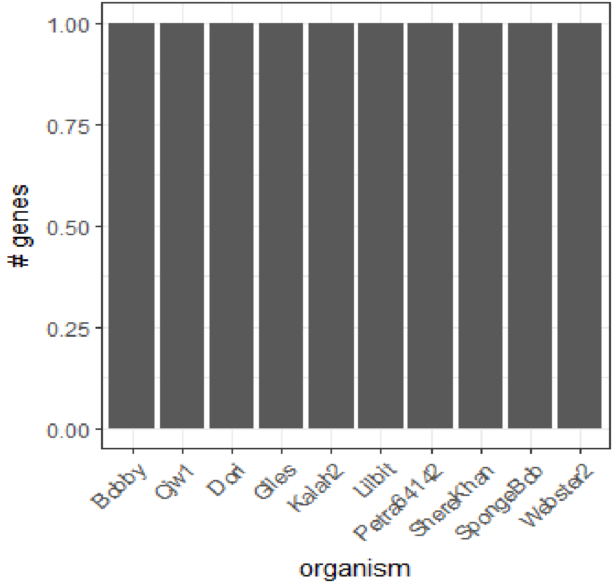
-Views the pangenome matrix as a ExpressionSet object

```
> as(mypang,'ExpressionSet')
ExpressionSet
(storageMode:lockedEnvironment)assayData: 5
features, 10 samples element names: exprs
protocolData: none
Pheno DatasampleNames: Bobby Cjw1 ... Webster2
(10 total)
varLabels: nGenes
varMetadata: labelDescription
featureData featureNames: OG1 OG2 ... OG5
(5total)
 fvarLabels: description group ... nGenes (7 total)
 fvarMetadata: labelDescription

experimentData: use 'experimentData(object)'
```

Plot Stat

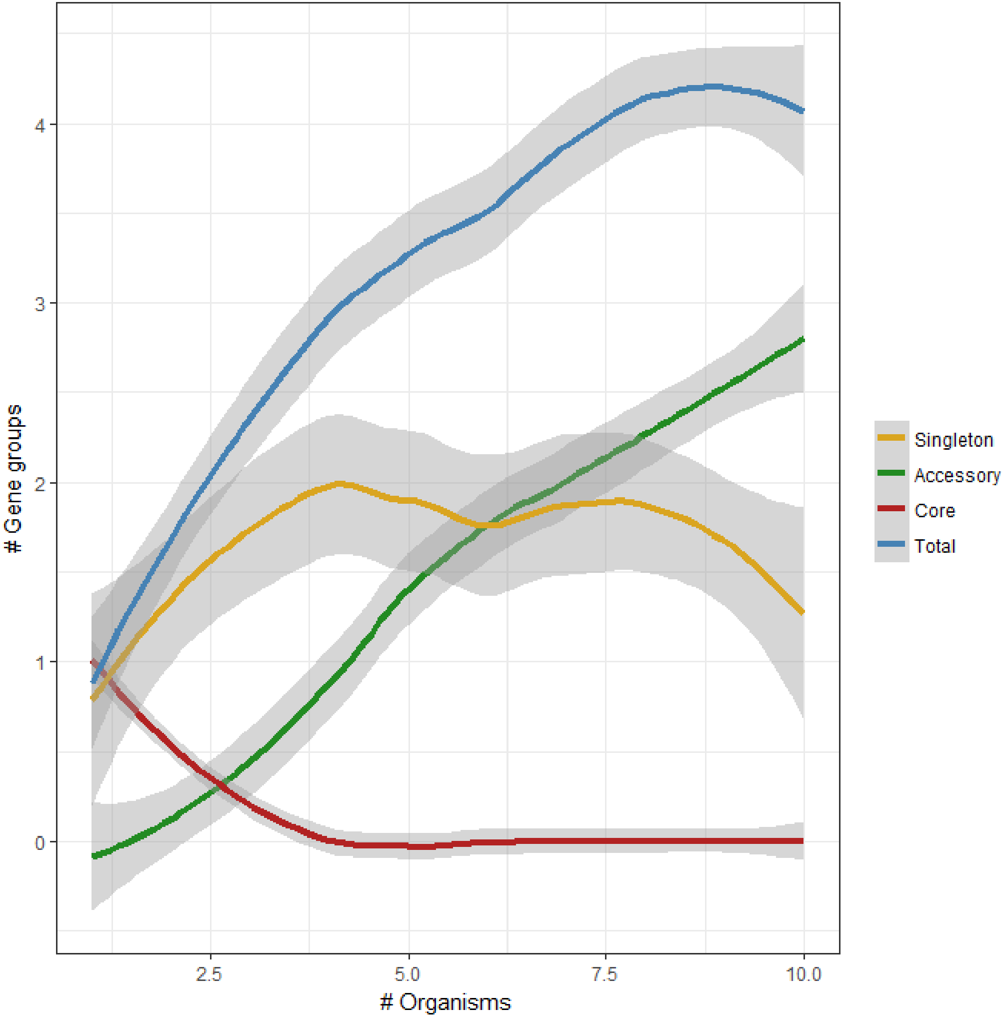

1
4
Group
🟨 Singleton
🟩 Accessory
🟥 Core


1.00
0.75
0.50
0.25
0.00
# genes
Bobby
Qiwi
Dori
Giles
Kalah2
Lilbit
Petra84142
ShereKhan
SpongeBob
Webster2
organism

Evolution Plot

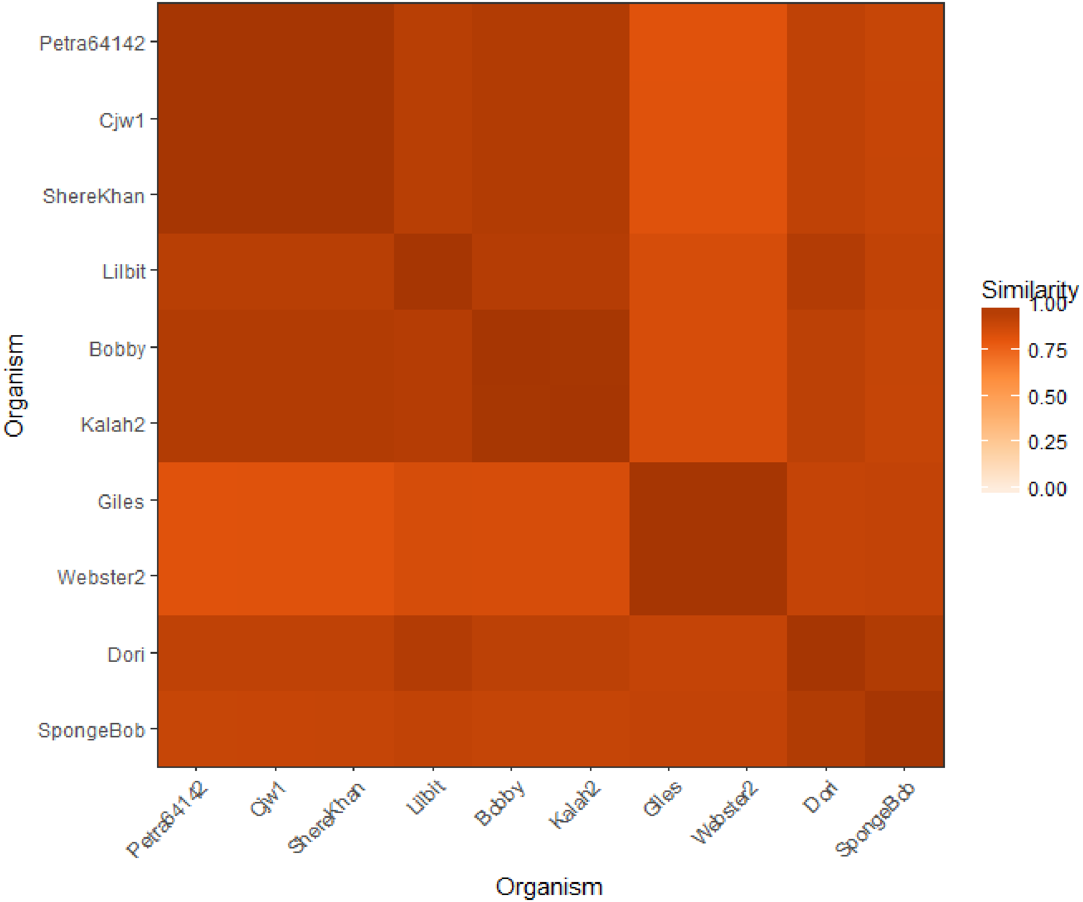
-Views number of singleton,accessory and core genes as the amount of organisms increase

-Can be biased toward order of organisms

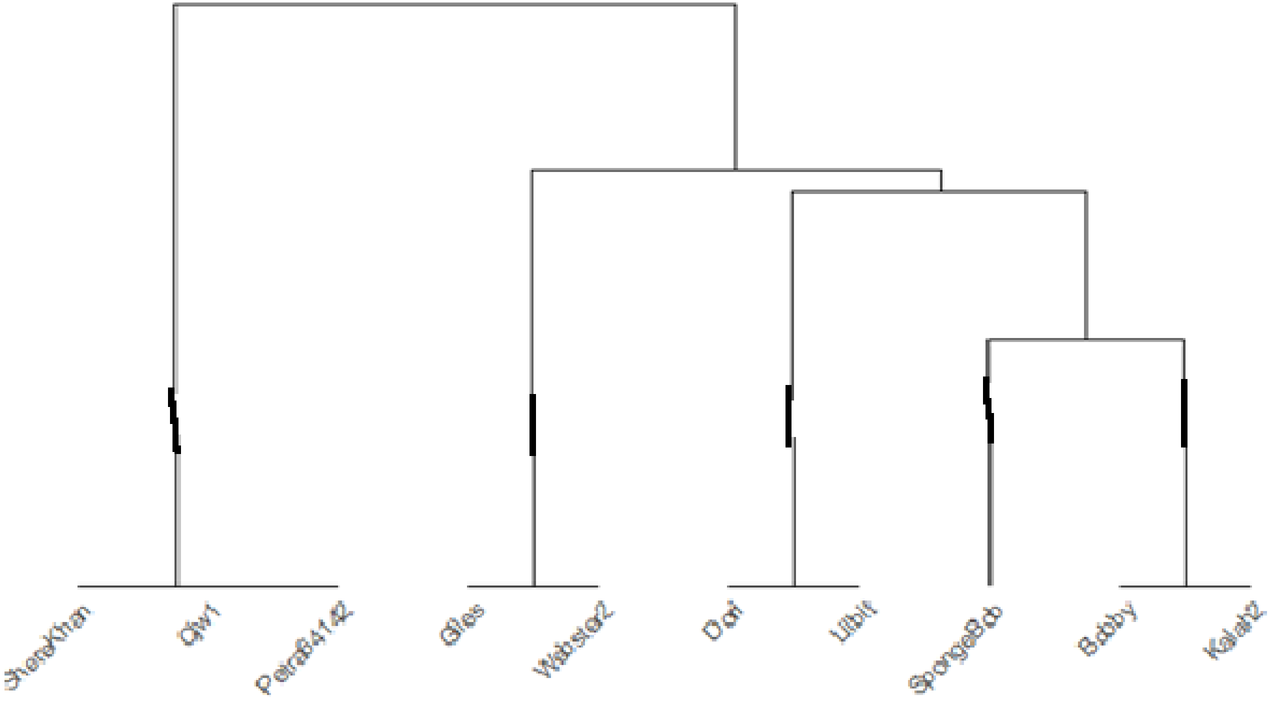


Kmer heatplot

-Comparison of Kmer values to each organism



Dendrogram

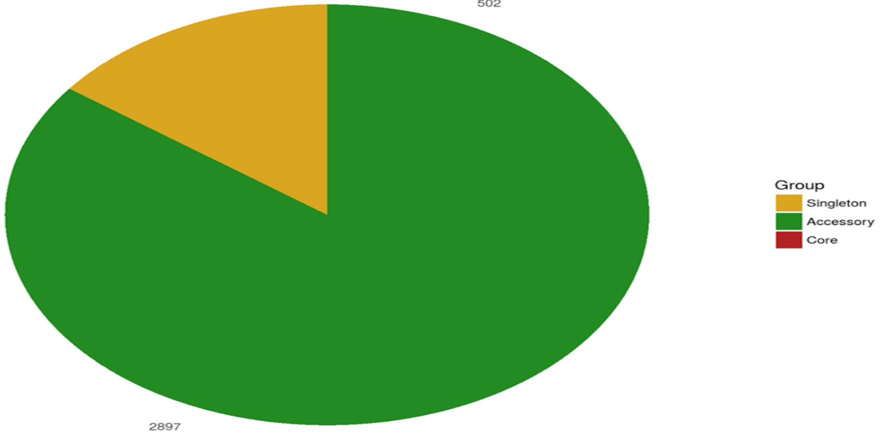

ShereKhan
Qiwi
Perro84142
Giles
Webster2
Dori
Lilbit
SpongeBob
Bobby
Kalah2

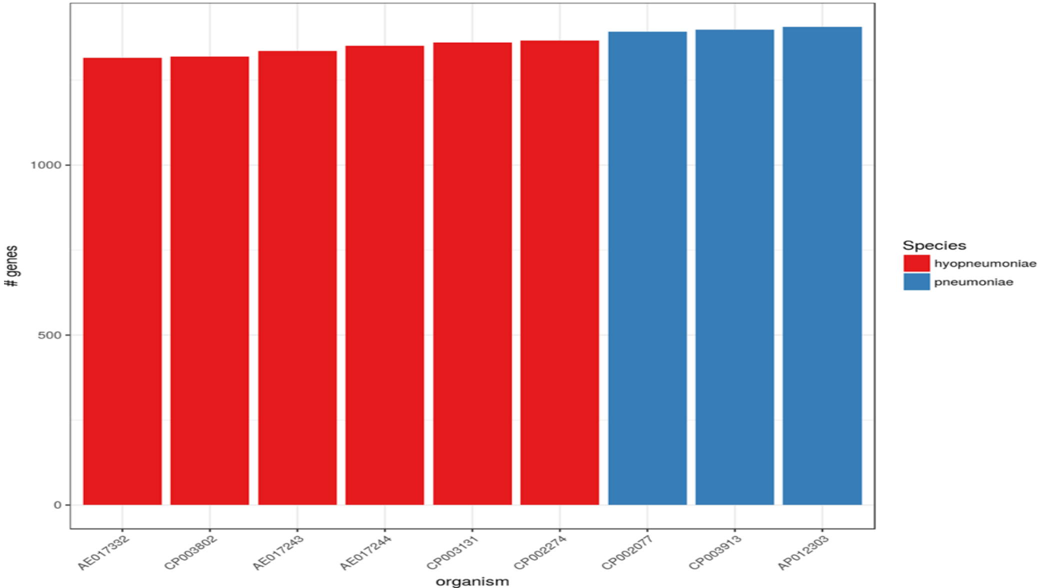
FindMyFriends Using Mycoplasma

```
mycoPan
## An object of class pgFullLoc
##
## The pangenome consists of 12247 genes from 9 organisms
## 3141 gene groups defined
##     Core|
##Accessory|=============================================:
## Singleton|======
## Genes are translated
```

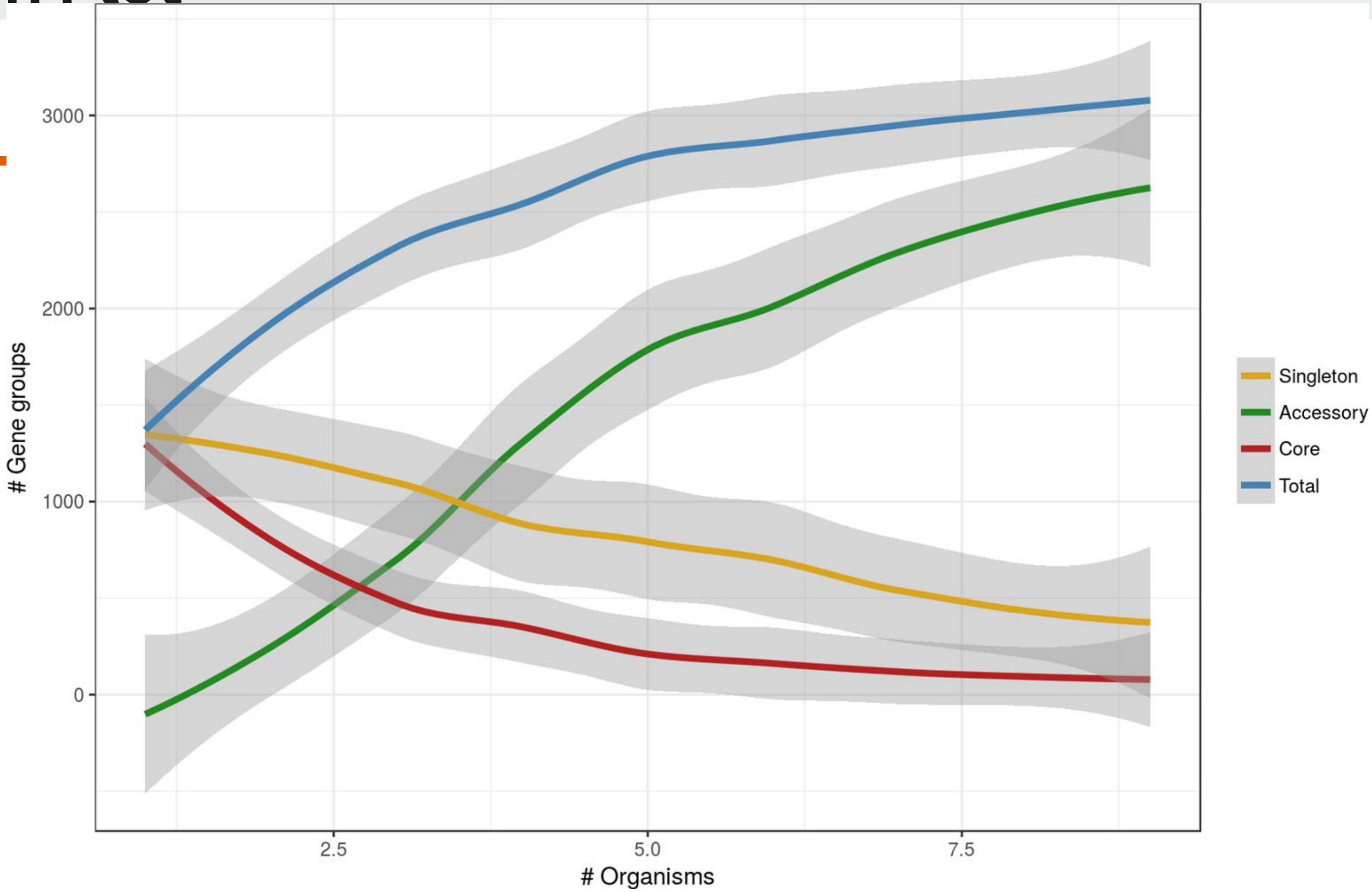
Pangenome as ExpressionSet

```
## ExpressionSet (storageMode: lockedEnvironment)
## assayData: 3399 features, 9 samples
##   element names: exprs
## protocolData: none
## phenoData
##   sampleNames: AE017243 AE017244 ... CP003913 (9 total)
##   varLabels: nGenes Id ... GenBankDivision (14 total)
##   varMetadata: labelDescription
## featureData
##   featureNames: OG1 OG2 ... OG3399 (3399 total)
##   fvarLabels: description group ... nGenes (7 total)
##   fvarMetadata: labelDescription
## experimentData: use 'experimentData(object)'
```

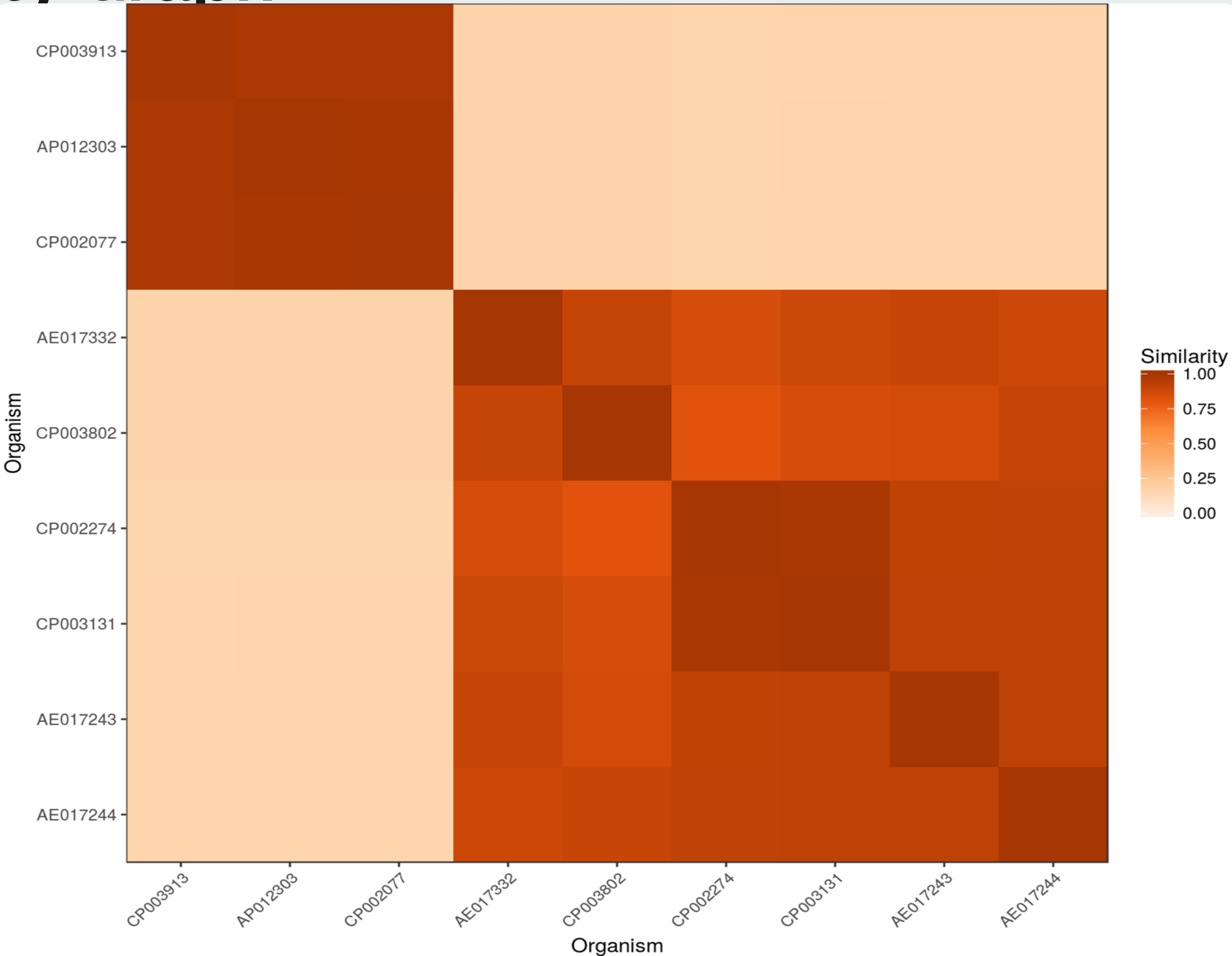

502
2897
Group
Singleton
Accessory
Core


# genes
1000
500
0
AE017332
CP003802
AE017243
AE017244
CP003131
CP002274
CP002077
CP003913
AP012303
organism
Species
hyopneumoniae
pneumoniae

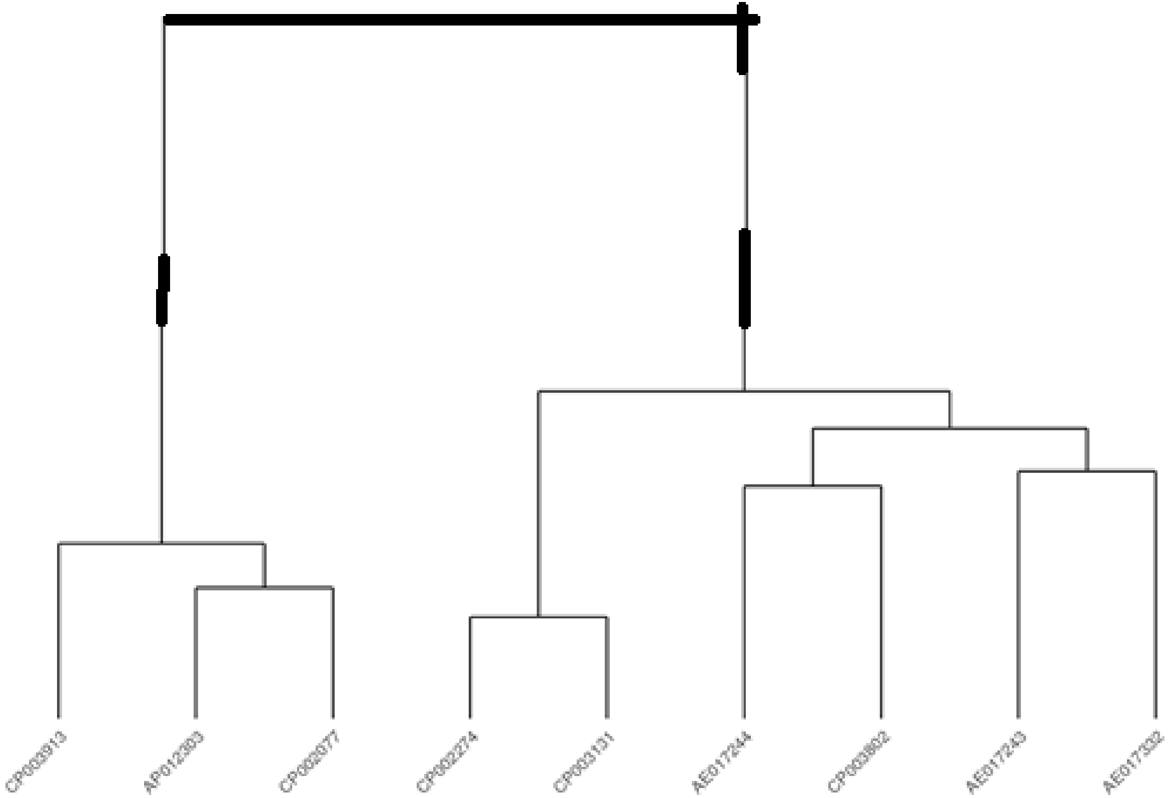
Evolution Plot



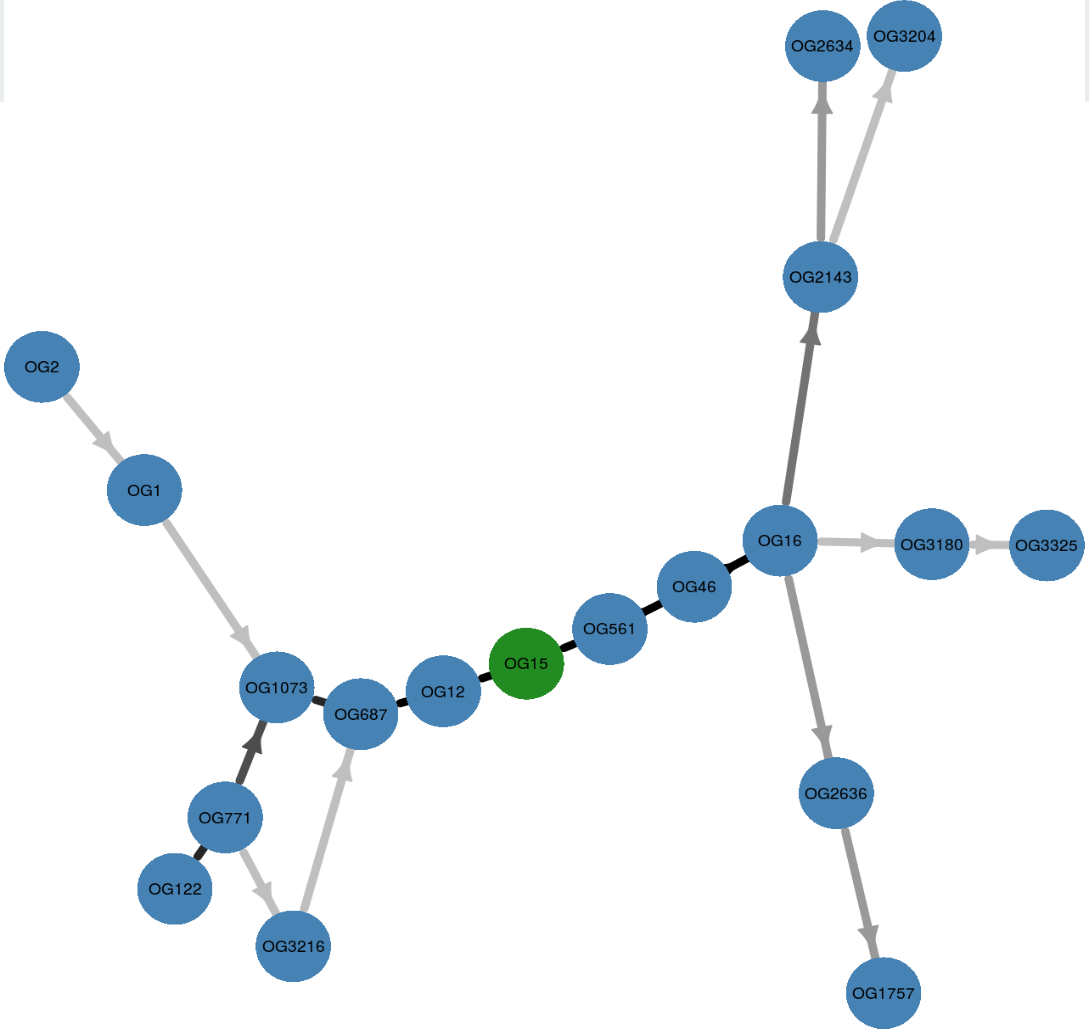
Kmer Similarity Graph



Dendogram of Pangenome



Neighborhood



References

Pictures on genomes: Google Images

“Actinobacteriophages.” *The Actinobacteriophage Database* , 28 Nov. 2017, phagesdb.org/.

Pedersen, Thomas Lin. “FindMyFriends.” *Bioconductor*, 2003, bioconductor.org/packages/release/bioc/html/FindMyFriends.html.

Pedersen, Thomas Lin. “Creating Pangenomes Using FindMyFriends.” *Bioconductor*, 30 Oct. 2017, www.bioconductor.org/packages/devel/bioc/vignettes/FindMyFriends/inst/doc/FindMyFriends_intro.html.

NIH. “Comparative Genomics Fact Sheet.” *National Human Genome Research Institute (NHGRI)*, 3 Nov. 2015, www.genome.gov/11509542/comparative-genomics-fact-sheet/.