Is Sharing Caring?

Elucidating the Effects of the Presence of CRISPR-Cas Systems on Rates of Horizontal Gene Transfer Using Network Analysis

Siddharth Reed
MolBiol 4C12 Thesis

Golding Lab,
Biology Department,
McMaster University

March 31, 2019

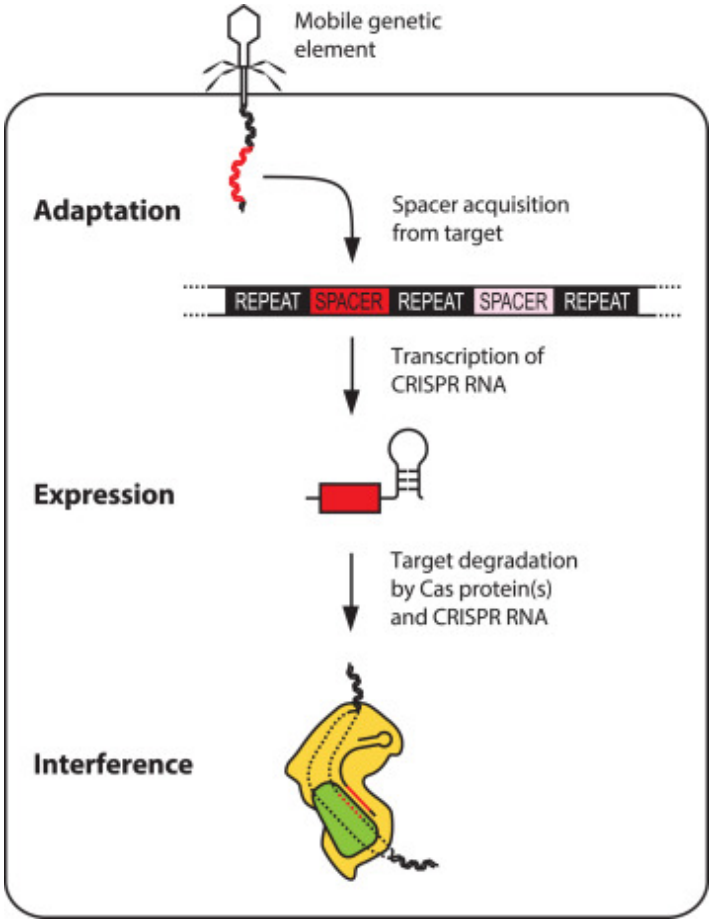
Table of Contents

CRISPR-Cas systems

What Are They?

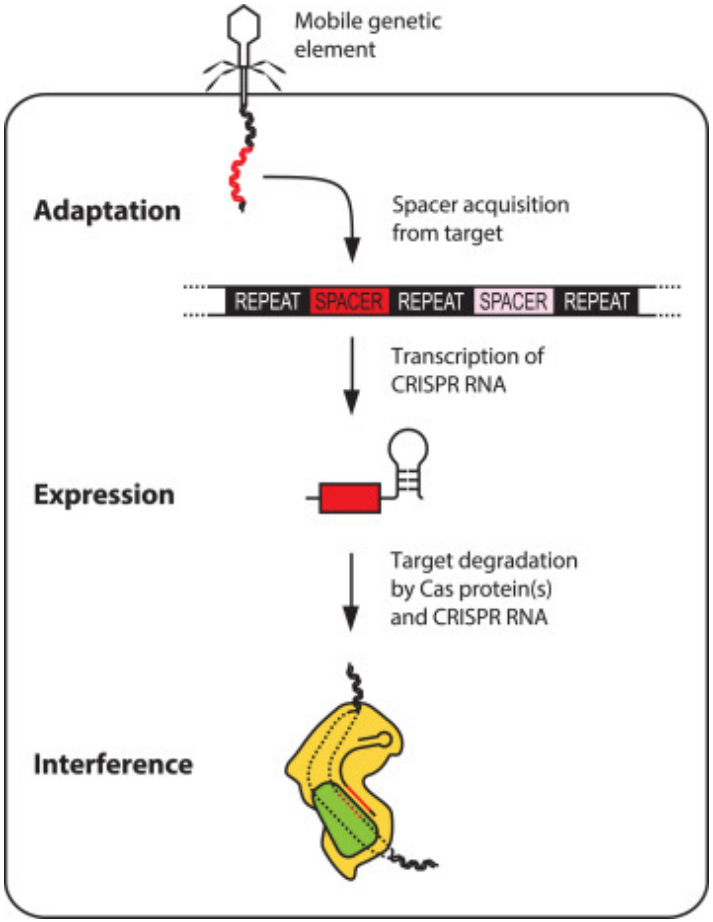
What Are They?

- Adaptive Bacterial Immune System



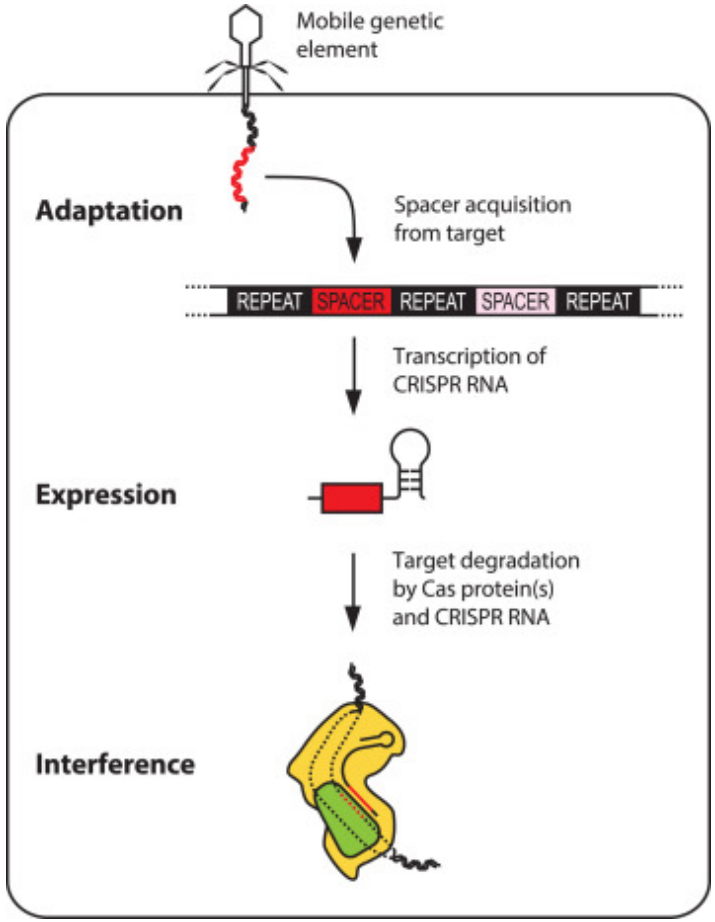
What Are They?

- Adaptive Bacterial Immune System
- Protects against foreign DNA



What Are They?

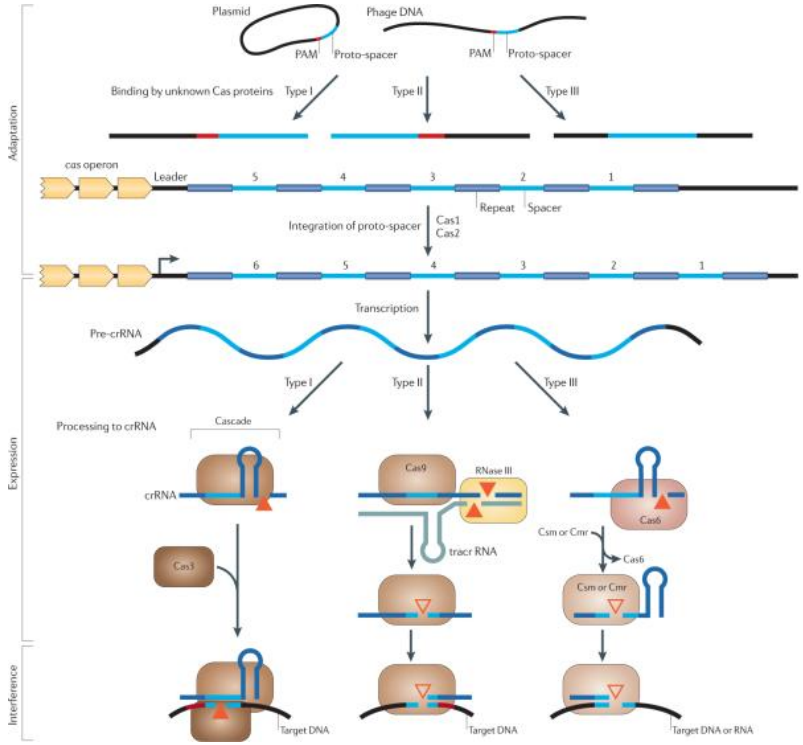
- Adaptive Bacterial Immune System
- Protects against foreign DNA
- Requires Cas proteins and CRISPR loci



Diversity & Ubiquity

Diversity & Ubiquity

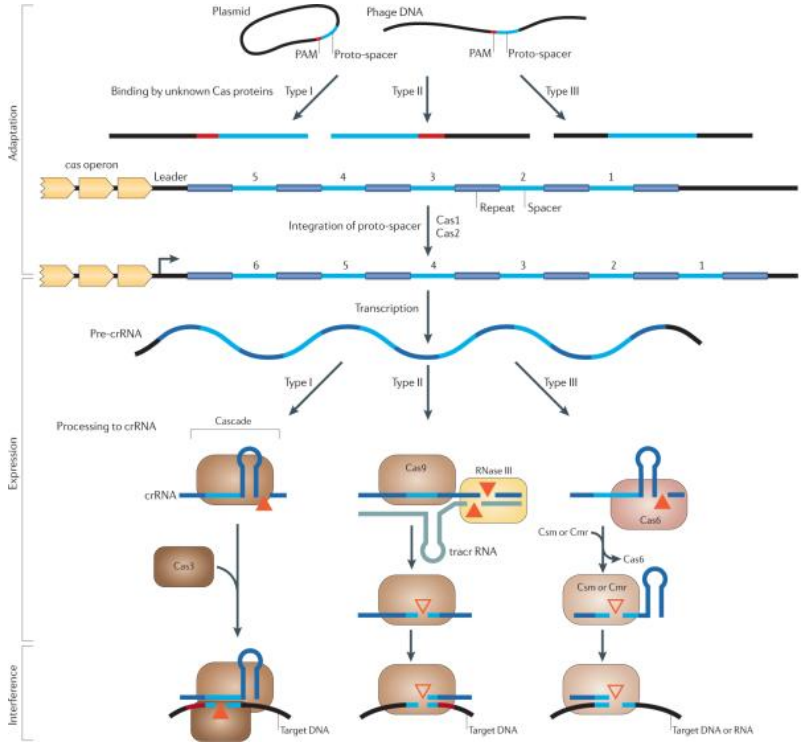
- 45% of bacteria have CRISPR loci ($n = 6782$)[2]



5

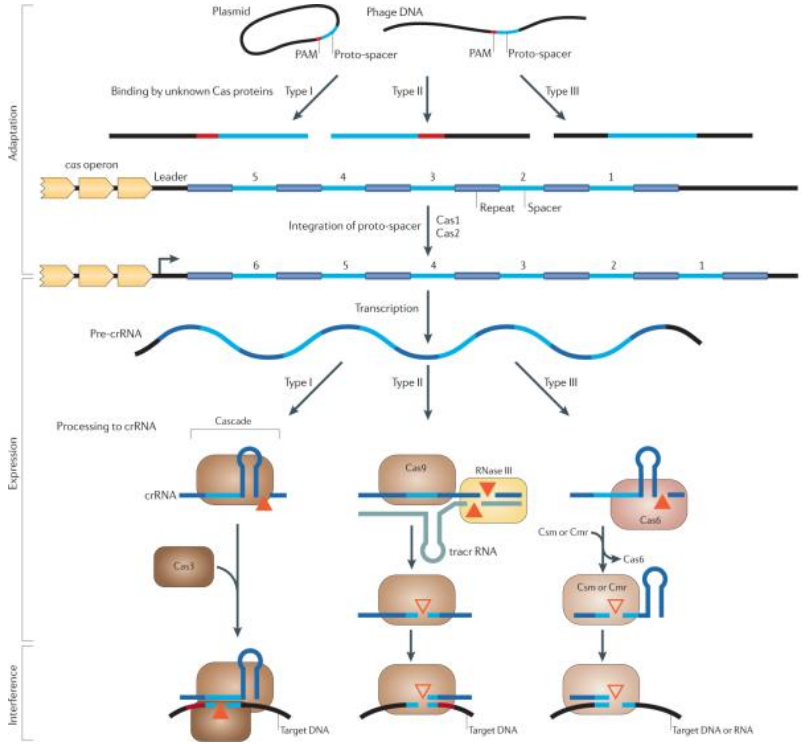
Diversity & Ubiquity

- 45% of bacteria have CRISPR loci ($n = 6782$)[2]
- 3 Main Types, multiple subtypes[3]



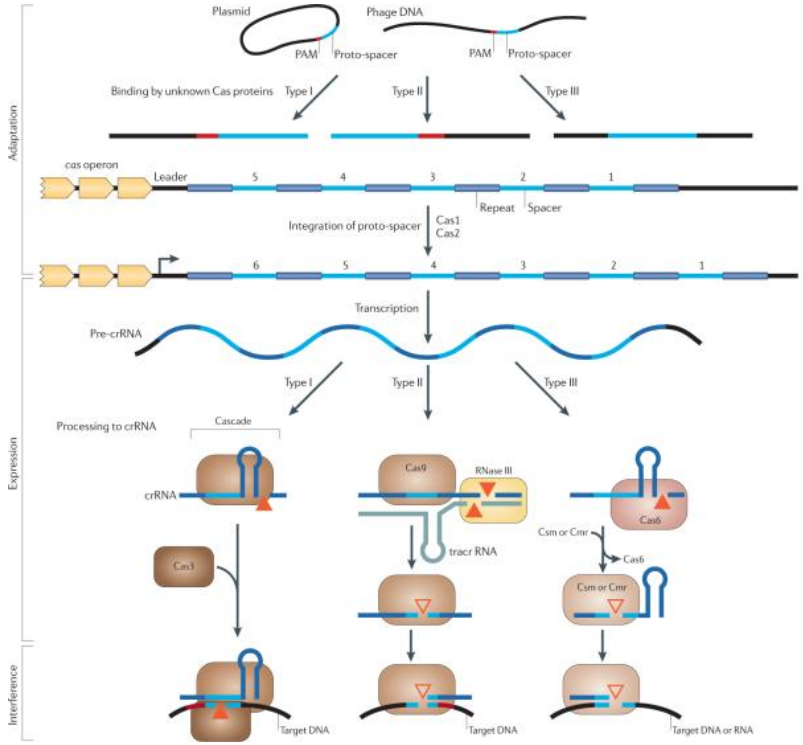
Diversity & Ubiquity

- 45% of bacteria have CRISPR loci ($n = 6782$)[2]
- 3 Main Types, multiple subtypes[3]
- CRISPR arrays represent unique life history of an organism



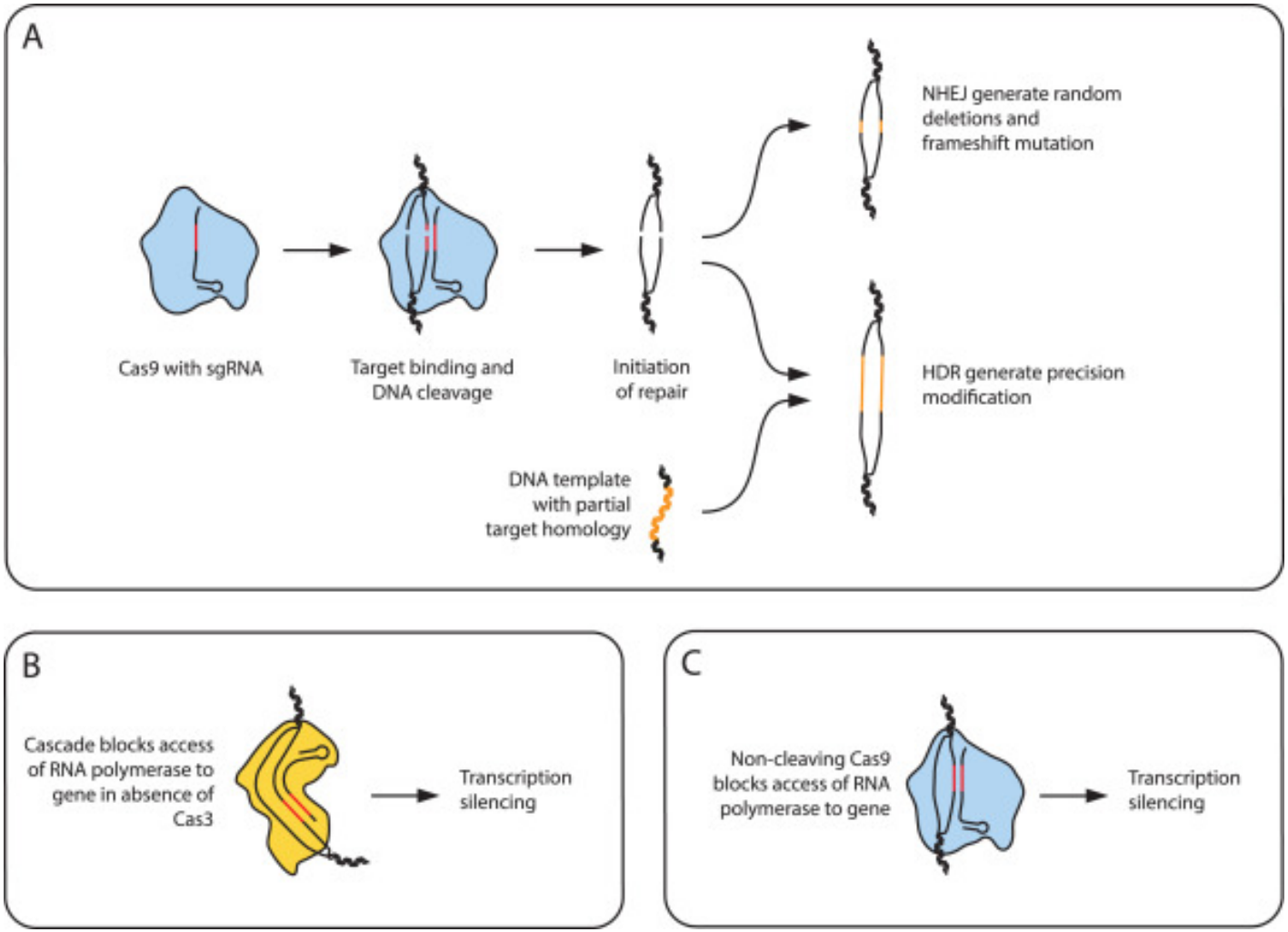
Diversity & Ubiquity

- 45% of bacteria have CRISPR loci ($n = 6782$)[2]
- 3 Main Types, multiple subtypes[3]
- CRISPR arrays represent unique life history of an organism
- $11\% - 28\%$ are false or orphaned CRISPR loci[4]



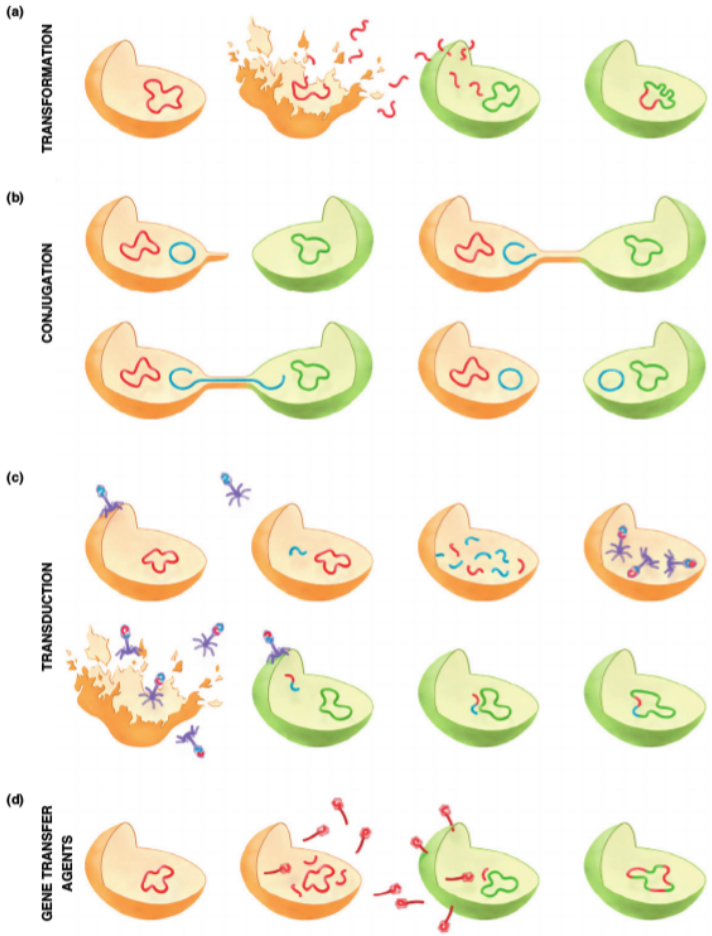
Biotech Application

Biotech Application



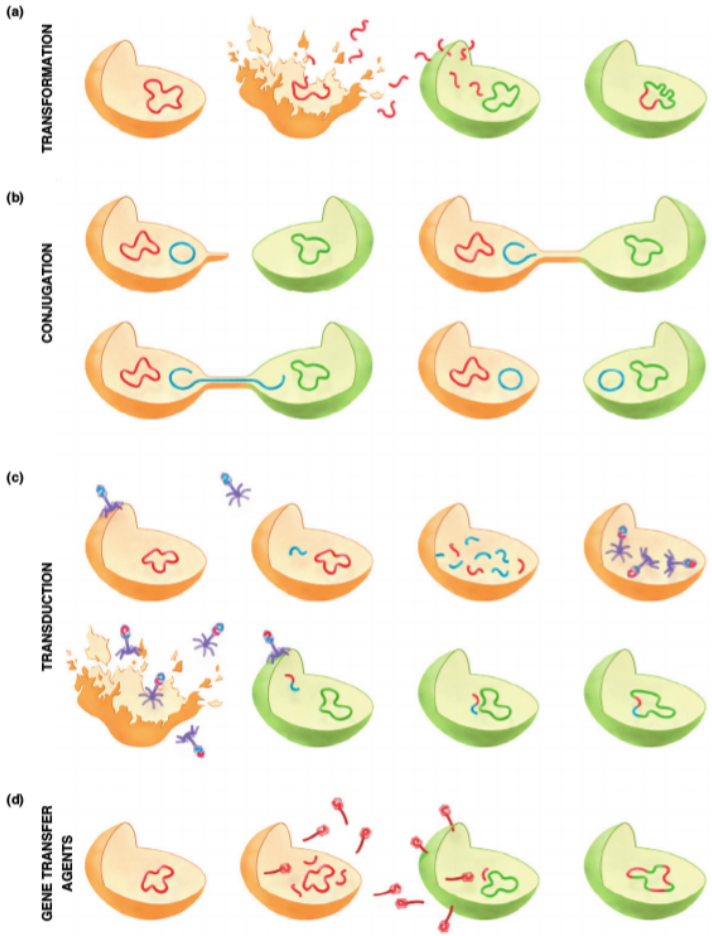
Horizontal Gene Transfer

Mechanisms



6

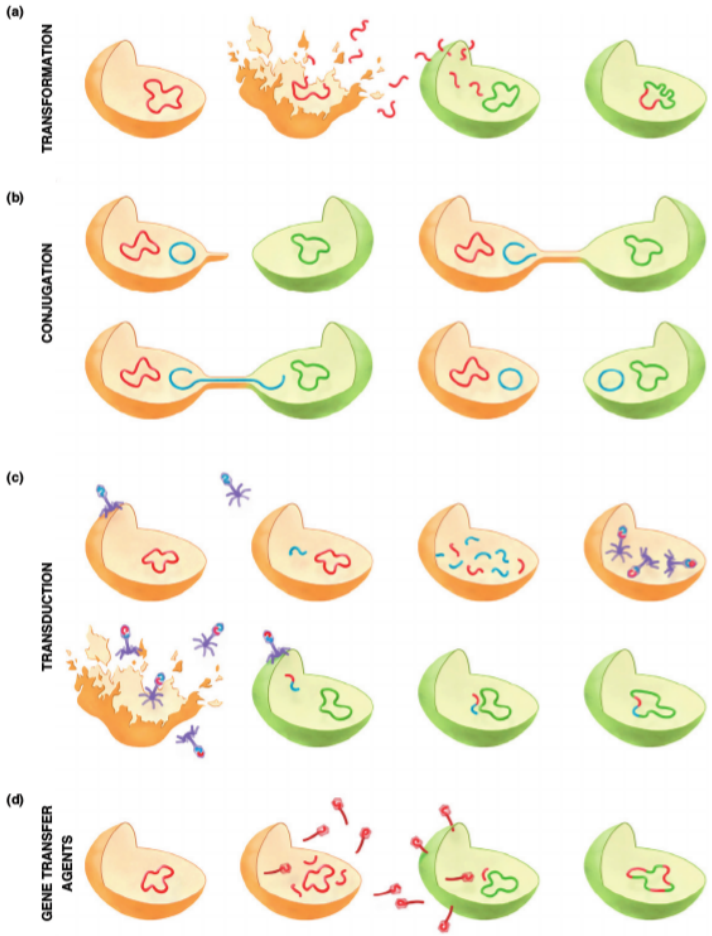
Mechanisms



- Conjugation: Transfer of DNA through cell-cell connections[6]

6

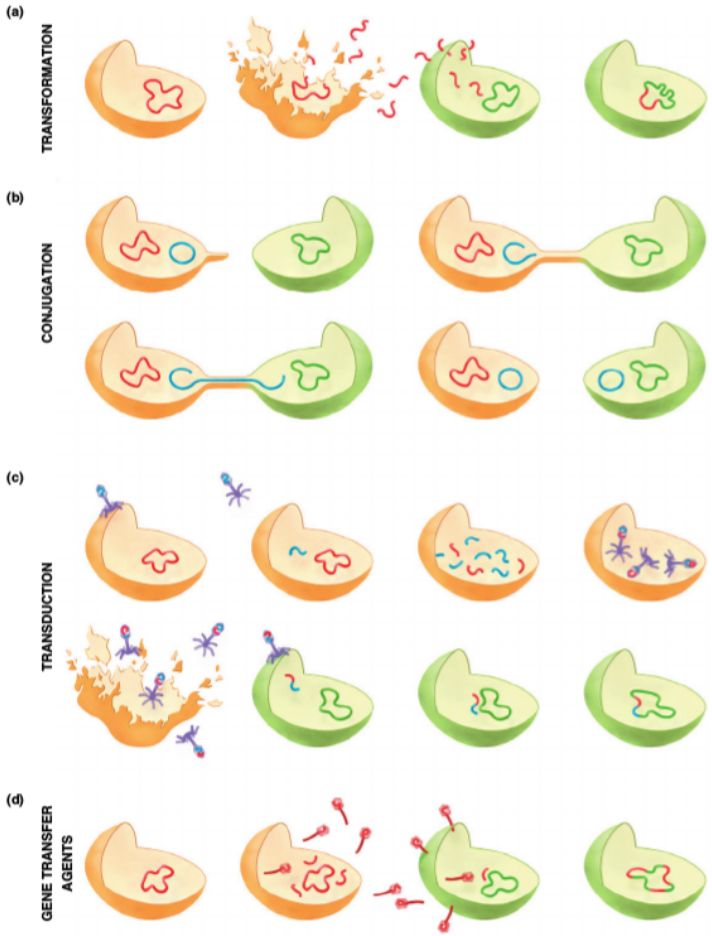
Mechanisms



- Conjugation: Transfer of DNA through cell-cell connections[6]
- Transformation: Incorportaion of free-floating DNA into the genome[6]

6

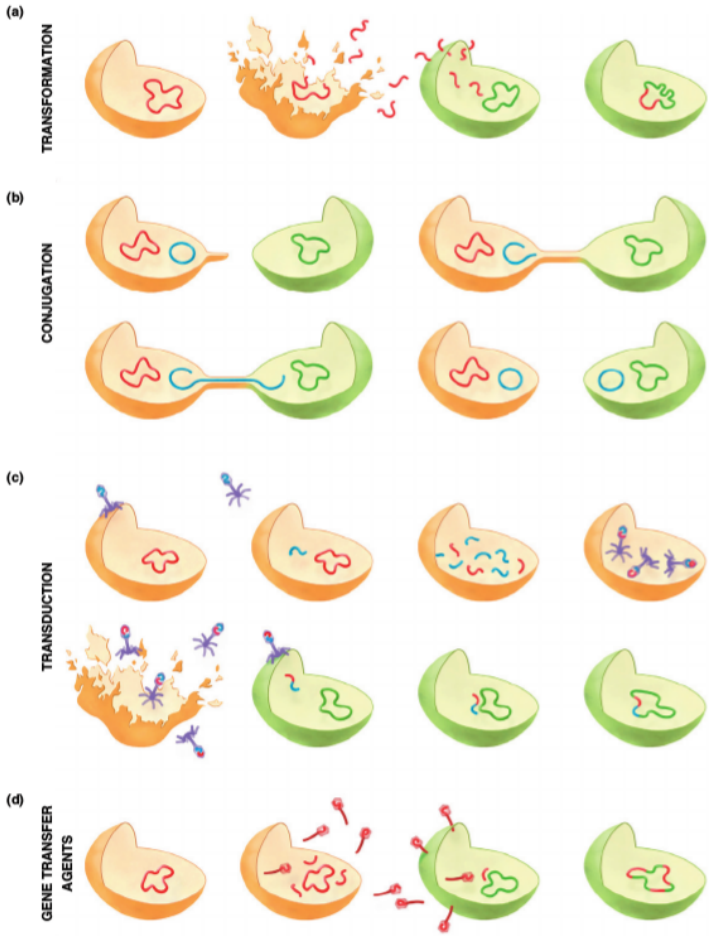
Mechanisms



6

- Conjugation: Transfer of DNA through cell-cell connections[6]
- Transformation: Incorportaion of free-floating DNA into the genome[6]
- Transduction: Transfer of DNA through phage[6]

Mechanisms

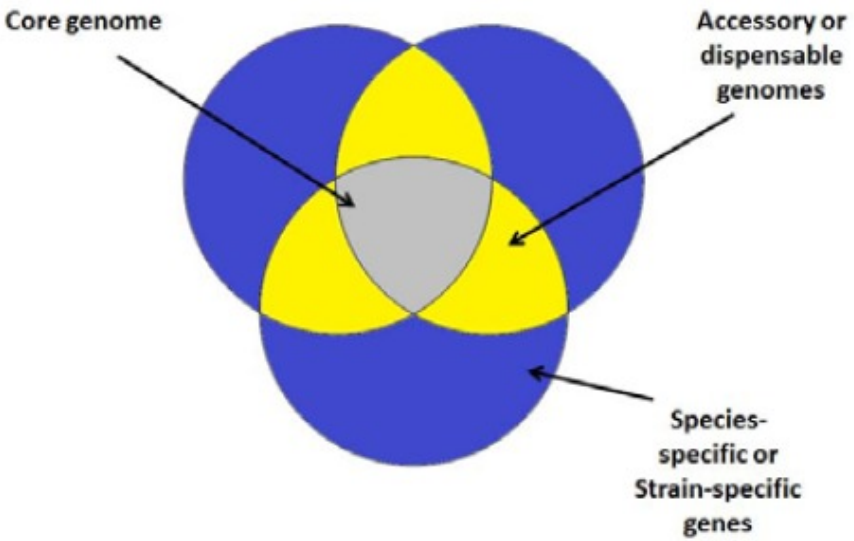


6

- Conjugation: Transfer of DNA through cell-cell connections[6]
- Transformation: Incorportaion of free-floating DNA into the genome[6]
- Transduction: Transfer of DNA through phage[6]
- **CRISPR-Cas directly affects Transduction and Transformation**[6]

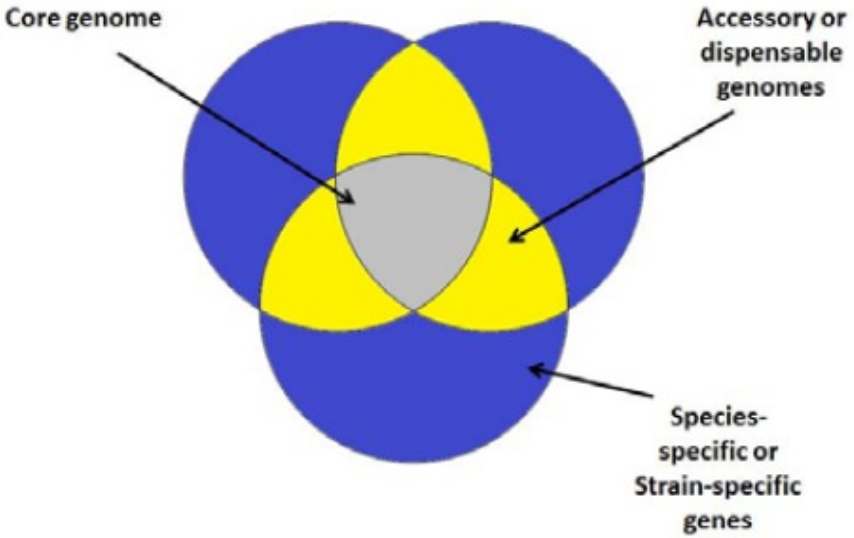
Pan-Genomes

Pan-Genomes

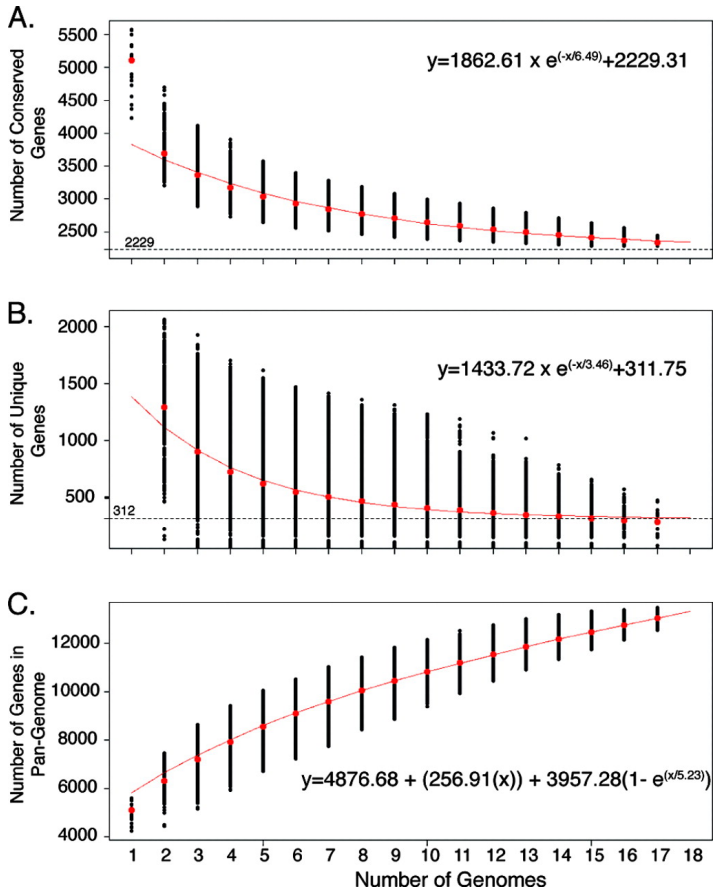


7

Pan-Genomes



7



8

Rate Influencing Factors

Rate Influencing Factors

- Amount of exogenous DNA/cell density/phage density

Rate Influencing Factors

- Amount of exogenous DNA/cell density/phage density
- Selective pressures

Rate Influencing Factors

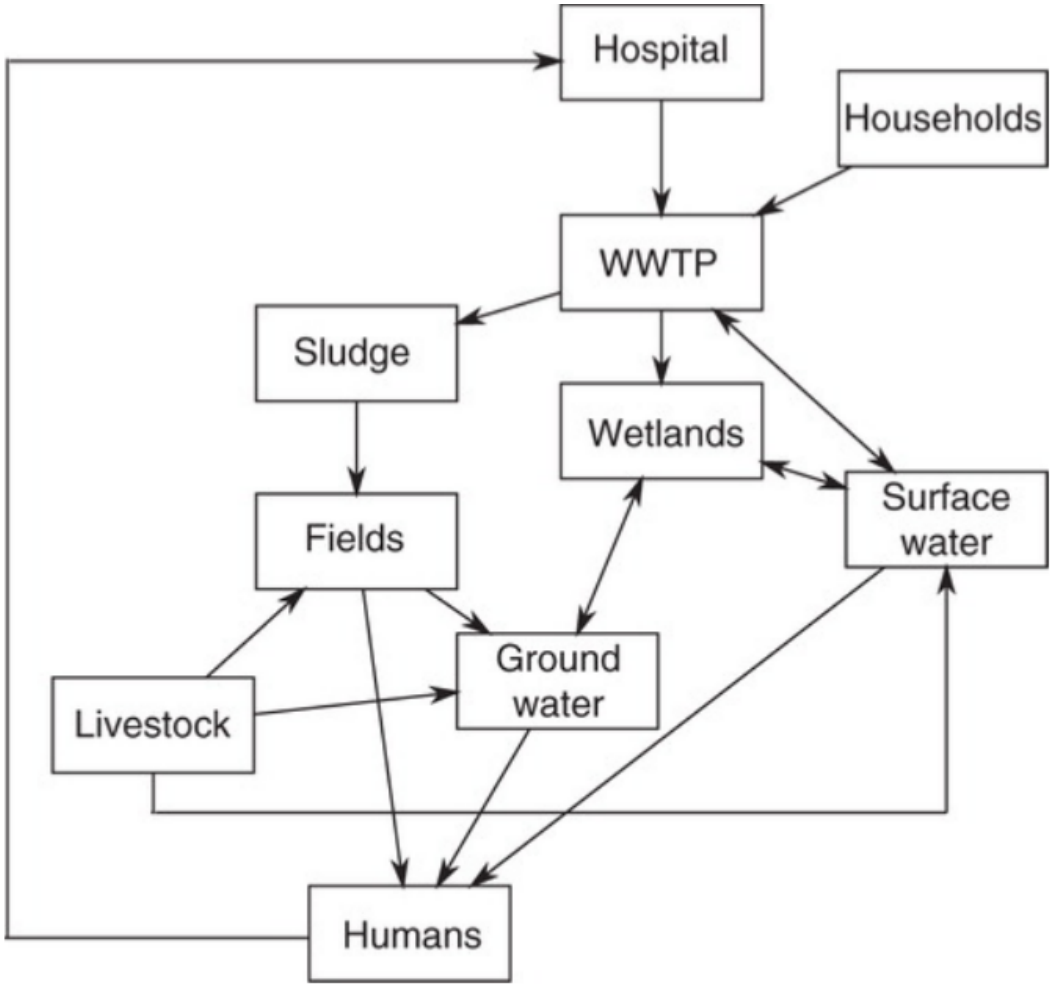
- Amount of exogenous DNA/cell density/phage density
- Selective pressures
- Metabolic costs

Rate Influencing Factors

- Amount of exogenous DNA/cell density/phage density
- Selective pressures
- Metabolic costs
- Sequence compatibility

Applications

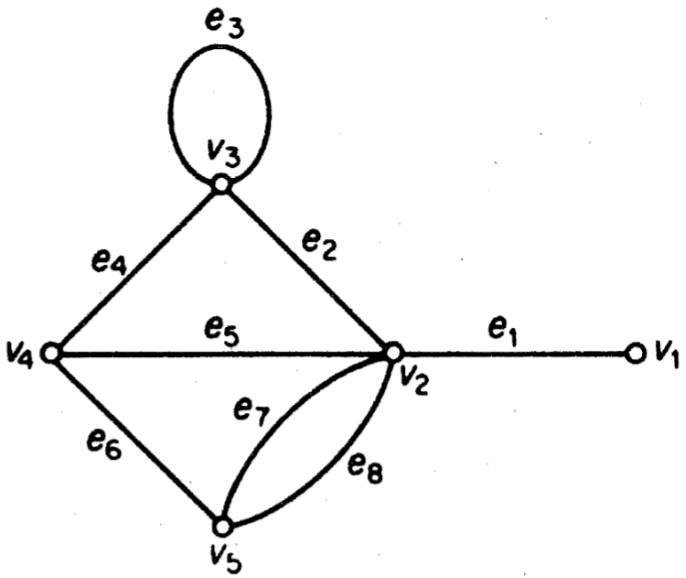
Applications



Phylogenomic Networks

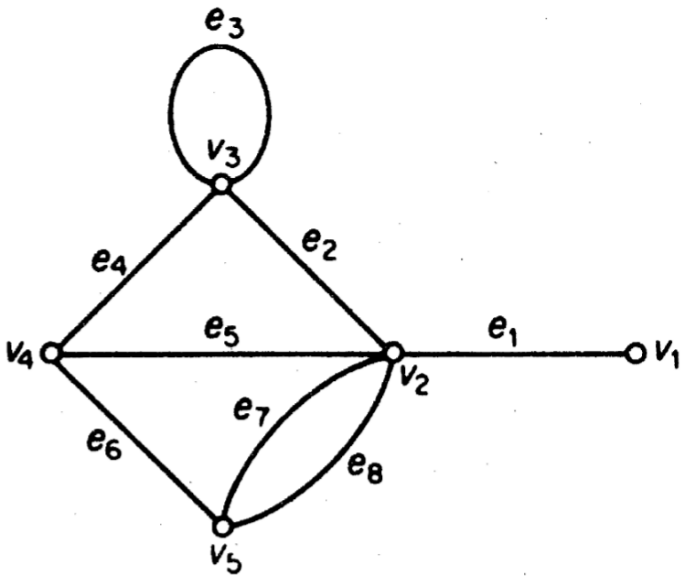
What is A Network?

What is A Network?



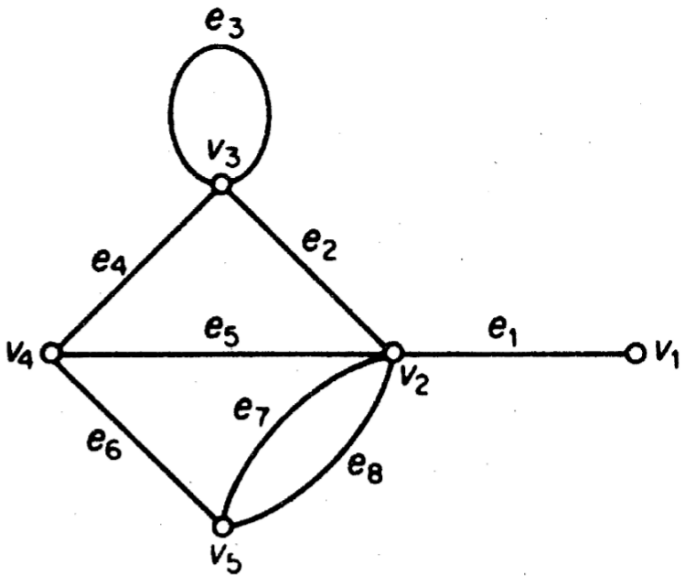
- Useful mathematical abstraction of real world system

What is A Network?



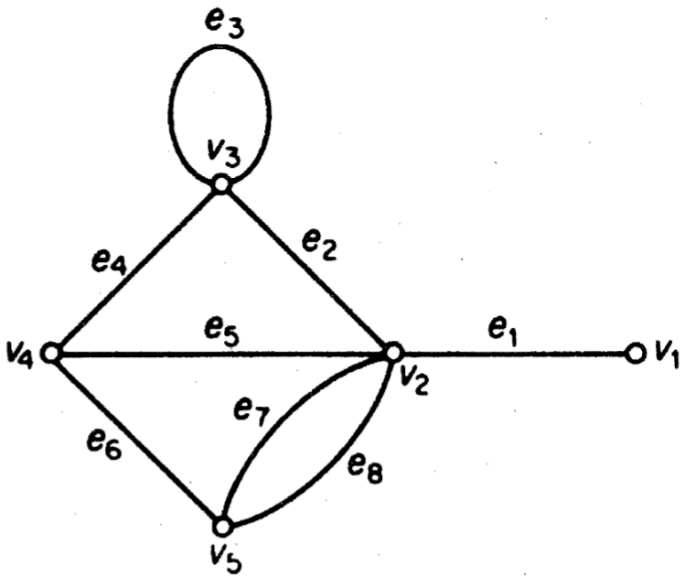
- Useful mathematical abstraction of real world system
- Nodes can have attributes

What is A Network?



- Useful mathematical abstraction of real world system
- Nodes can have attributes
- Directed or Undirected Edges

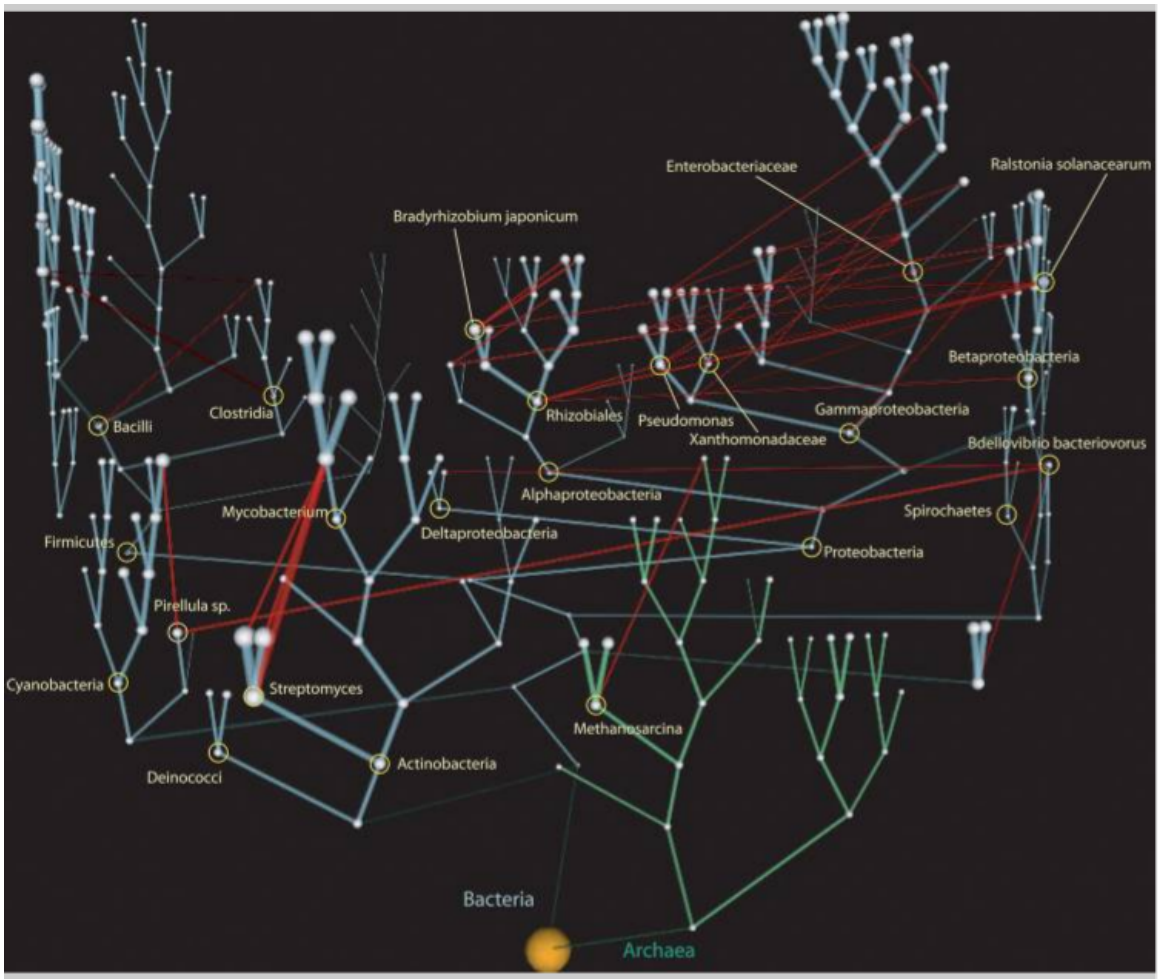
What is A Network?



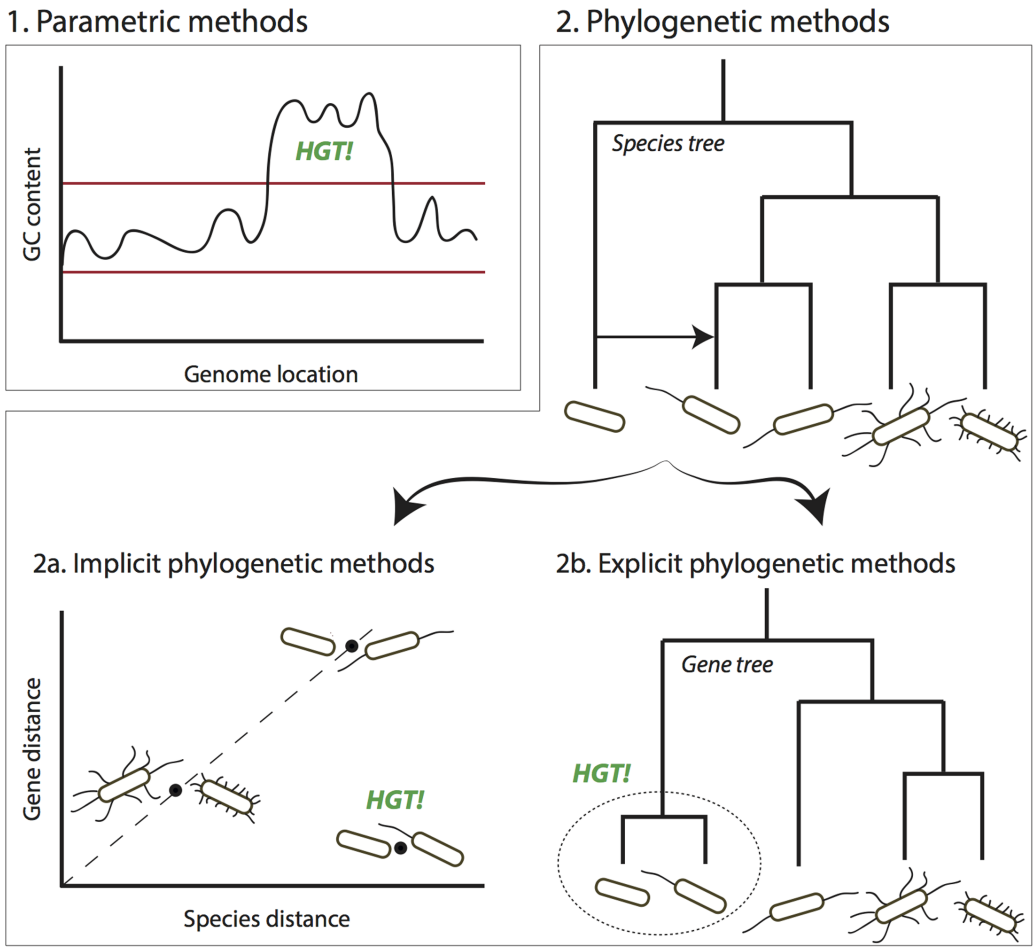
- Useful mathematical abstraction of real world system
- Nodes can have attributes
- Directed or Undirected Edges
- Weighted or Unweighted Edges

Prokaryotic "Net of Life"

Prokaryotic "Net of Life"



Construction



Do CRRISPR Systems Affect Horizontal Gene Transfer?

Yes

CRISPR Cost Complexity

CRISPR Cost Complexity

- Cost tradeoff factors:

CRISPR Cost Complexity

- Cost tradeoff factors:
  - Metabolic maintenace[1]

CRISPR Cost Complexity

- Cost tradeoff factors:
  - Metabolic maintenace[1]
  - Environmental pressures[13]

CRISPR Cost Complexity

- Cost tradeoff factors:
  - Metabolic maintenace[1]
  - Environmental pressures[13]
  - Off-target effects (autoimmune)[14]

CRISPR Cost Complexity

- Cost tradeoff factors:
  - Metabolic maintenace[1]
  - Environmental pressures[13]
  - Off-target effects (autoimmune)[14]
  - Anti-CRISPR systems[3]

CRISPR Cost Complexity

- Cost tradeoff factors:
  - Metabolic maintenace[1]
  - Environmental pressures[13]
  - Off-target effects (autoimmune)[14]
  - Anti-CRISPR systems[3]
  - Phage virulence/density[3]

CRISPR Cost Complexity

- Cost tradeoff factors:
  - Metabolic maintenace[1]
  - Environmental pressures[13]
  - Off-target effects (autoimmune)[14]
  - Anti-CRISPR systems[3]
  - Phage virulence/density[3]
  - Prophage abundance[15]

Curbing CRISPR Cost

Curbing CRISPR Cost

- CRISPRs themselves can be transfered $\Longrightarrow$ population level immunity[16]

Curbing CRISPR Cost

- CRISPRs themselves can be transfered $\Longrightarrow$ population level immunity[16]
- Selective CRISPR inactivation[1]

Curbing CRISPR Cost

- CRISPRs themselves can be transfered $\Longrightarrow$ population level immunity[16]
- Selective CRISPR inactivation[1]
- CRISPR can enhance transduction-mediated HGT[15]

Previous Findings

Previous Findings

- Gophna et al. (2015) found no realtion between the presence of CRISPR systems and HGT over short evolutionary timescales[17]

Previous Findings

- Gophna et al. (2015) found no realtion between the presence of CRISPR systems and HGT over short evolutionary timescales[17]
  - Assume all singletons arrose from HGT

Previous Findings

- Gophna et al. (2015) found no realtion between the presence of CRISPR systems and HGT over short evolutionary timescales[17]
  - Assume all singletons arrose from HGT
  - Used GC% to identify HGT

Previous Findings

- Gophna et al. (2015) found no realtion between the presence of CRISPR systems and HGT over short evolutionary timescales[17]
  - Assume all singletons arrose from HGT
  - Used GC% to identify HGT
- Contradicted by a former underdraduate thesis student

Previous Findings

- Gophna et al. (2015) found no realtion between the presence of CRISPR systems and HGT over short evolutionary timescales[17]
  - Assume all singletons arrose from HGT
  - Used GC% to identify HGT
- Contradicted by a former underdraduate thesis student
  - Can see inhibitory effects of CRIPSR on HGT over short evolutionary time scales

Previous Findings

- Gophna et al. (2015) found no realtion between the presence of CRISPR systems and HGT over short evolutionary timescales[17]
  - Assume all singletons arrose from HGT
  - Used GC% to identify HGT
- Contradicted by a former underdraduate thesis student
  - Can see inhibitory effects of CRIPSR on HGT over short evolutionary time scales
  - Higher gene indel rates for CRISPR containg genera than non-CRISPR containing outgroups

My Project

Hypothesis

Null Hypothesis

Bacterial strains or genera with known CRISPR systems will show no significant differences in network statistics compared to those strains or genera without known CRISPR systems.

Hypothesis

Null Hypothesis

Bacterial strains or genera with known CRISPR systems will show no significant differences in network statistics compared to those strains or genera without known CRISPR systems

Alternative Hypothesis

Bacterial strains or genera with known CRISPR systems will show a significant difference in at least 1 network statistic compared to those strains or genera without known CRISPR systems.

Objectives

Objectives

Within Network Comparisons

For genera with CRISPR conatining strains, compare the node statistics of CRIPSR-containing strain to non-CRISPR-containing strains.

Objectives

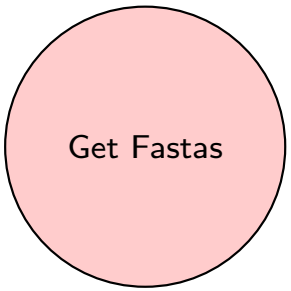
Within Network Comparisons

For genera with CRISPR conatining strains, compare the node statistics of CRIPSR-containing strain to non-CRISPR-containing strains.

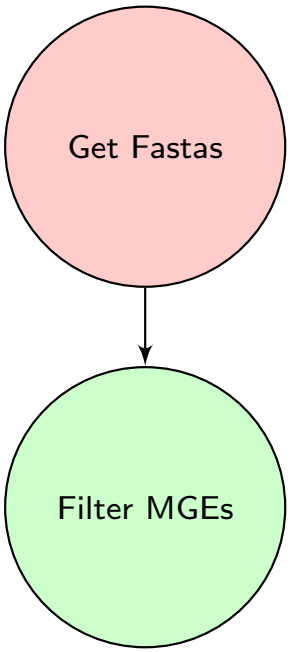
Gene Indel Rates vs. Network Statistics

Compare gene InDel rates to node/network statistics for CRISPR-containing and non-CRISPR-containing strains/genera.

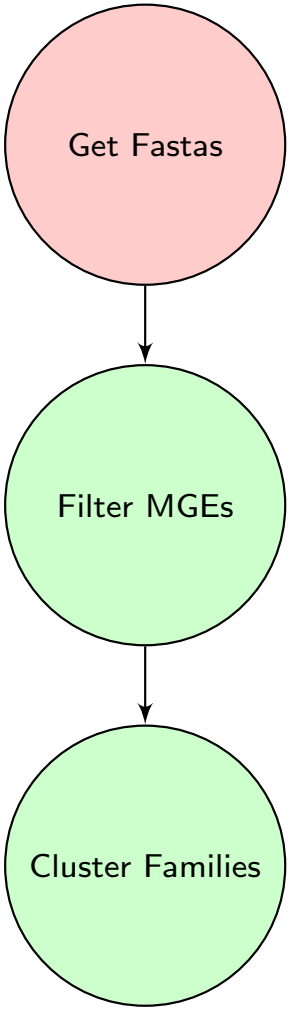
Workflow



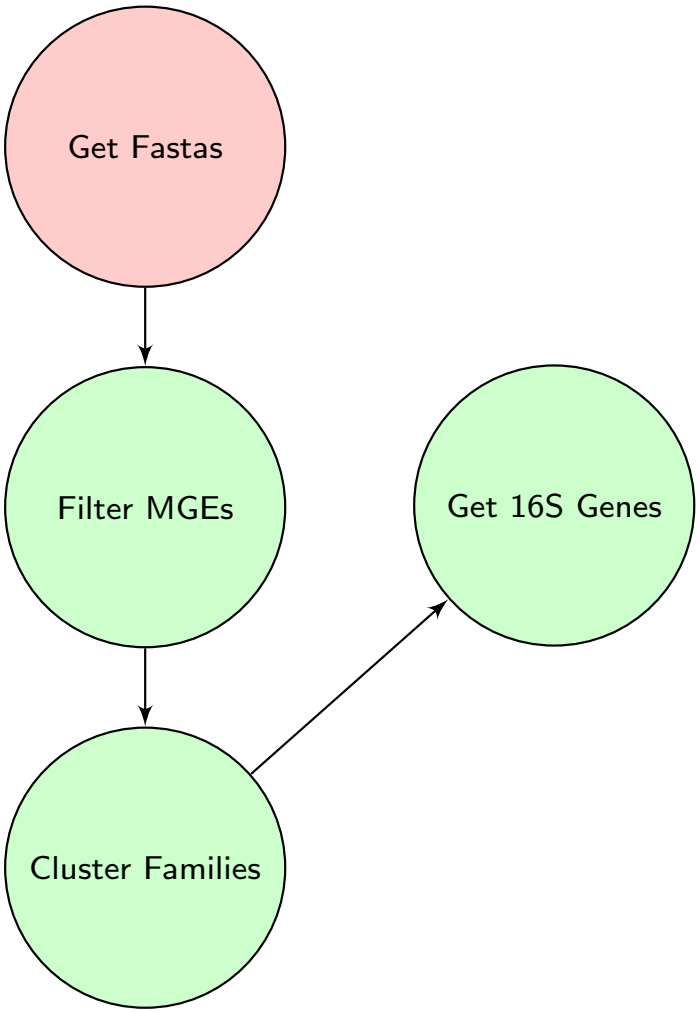
Workflow



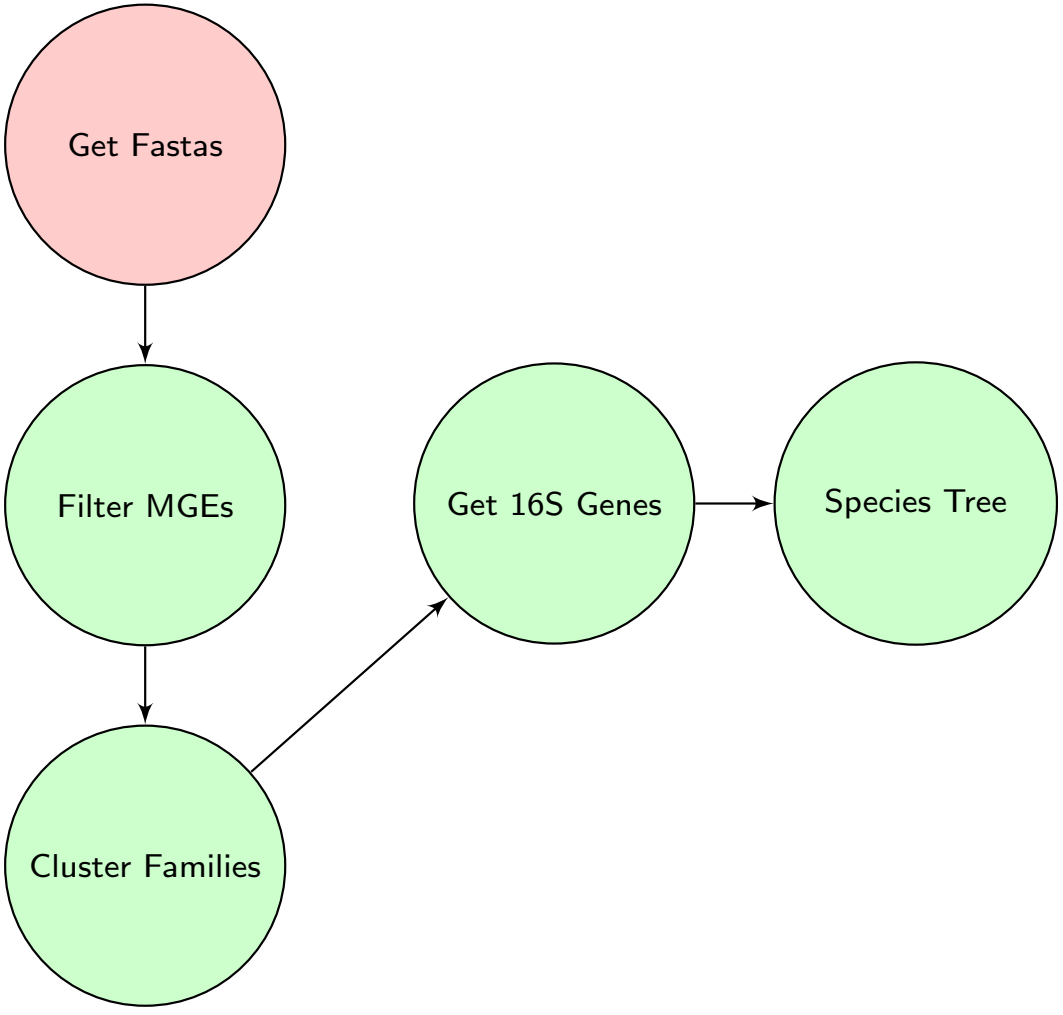
Workflow



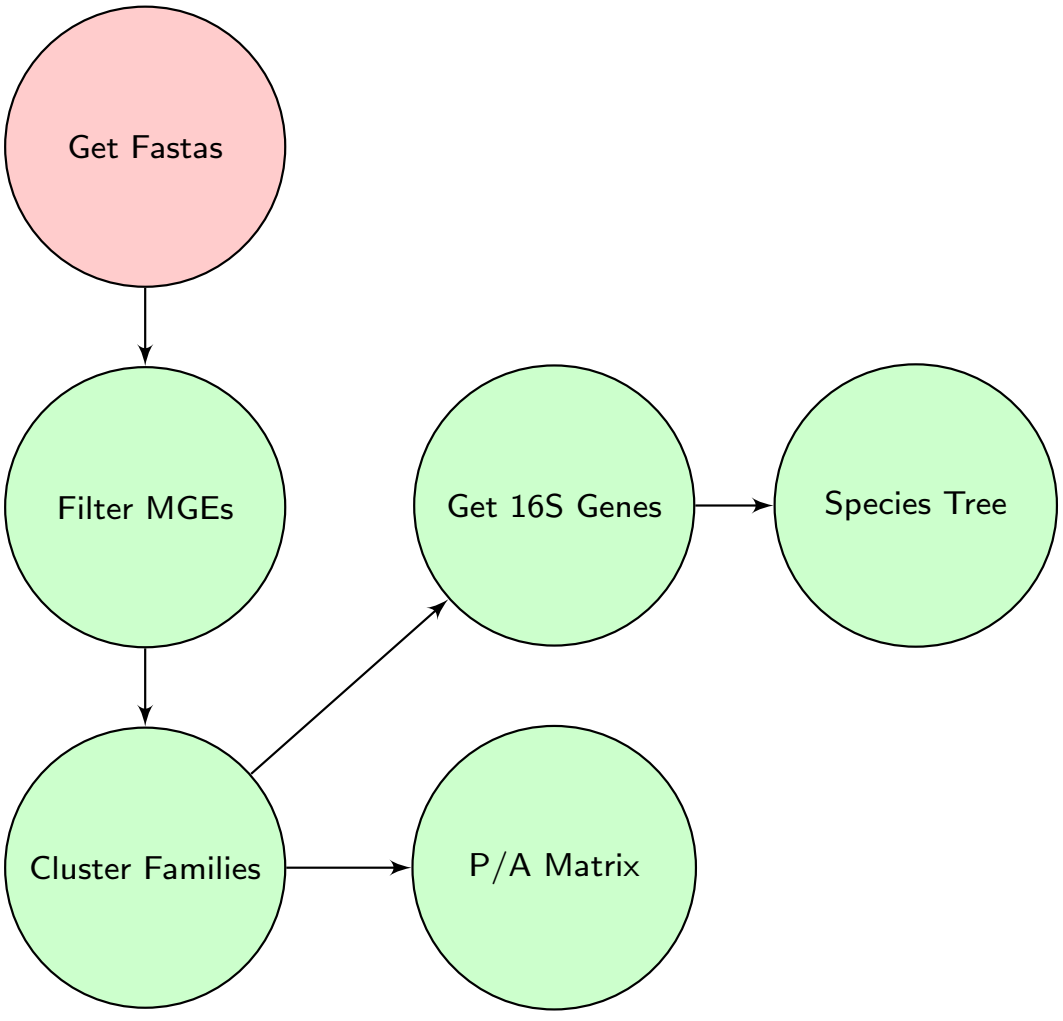
Workflow



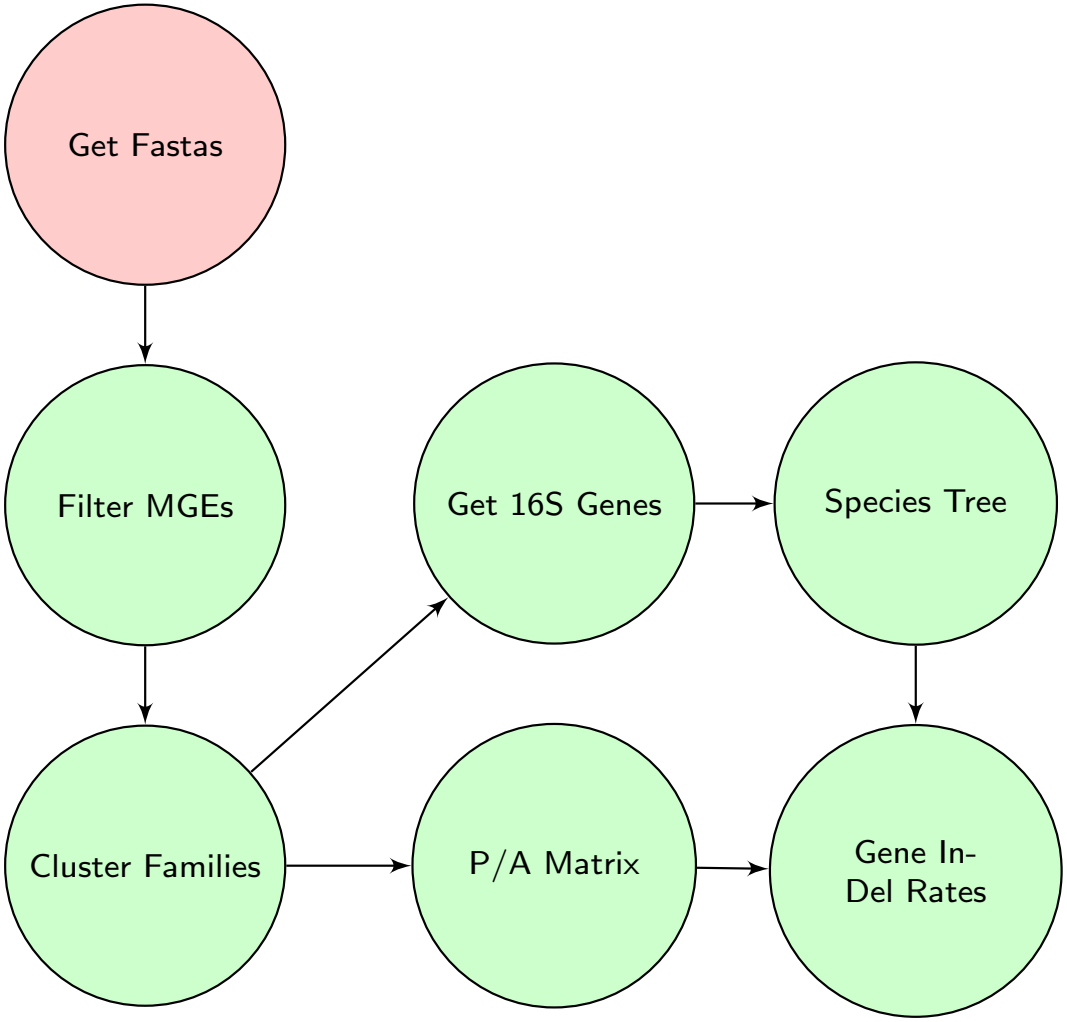
Workflow


Get Fastas
Filter MGEs
Cluster Families
Get 16S Genes
Species Tree

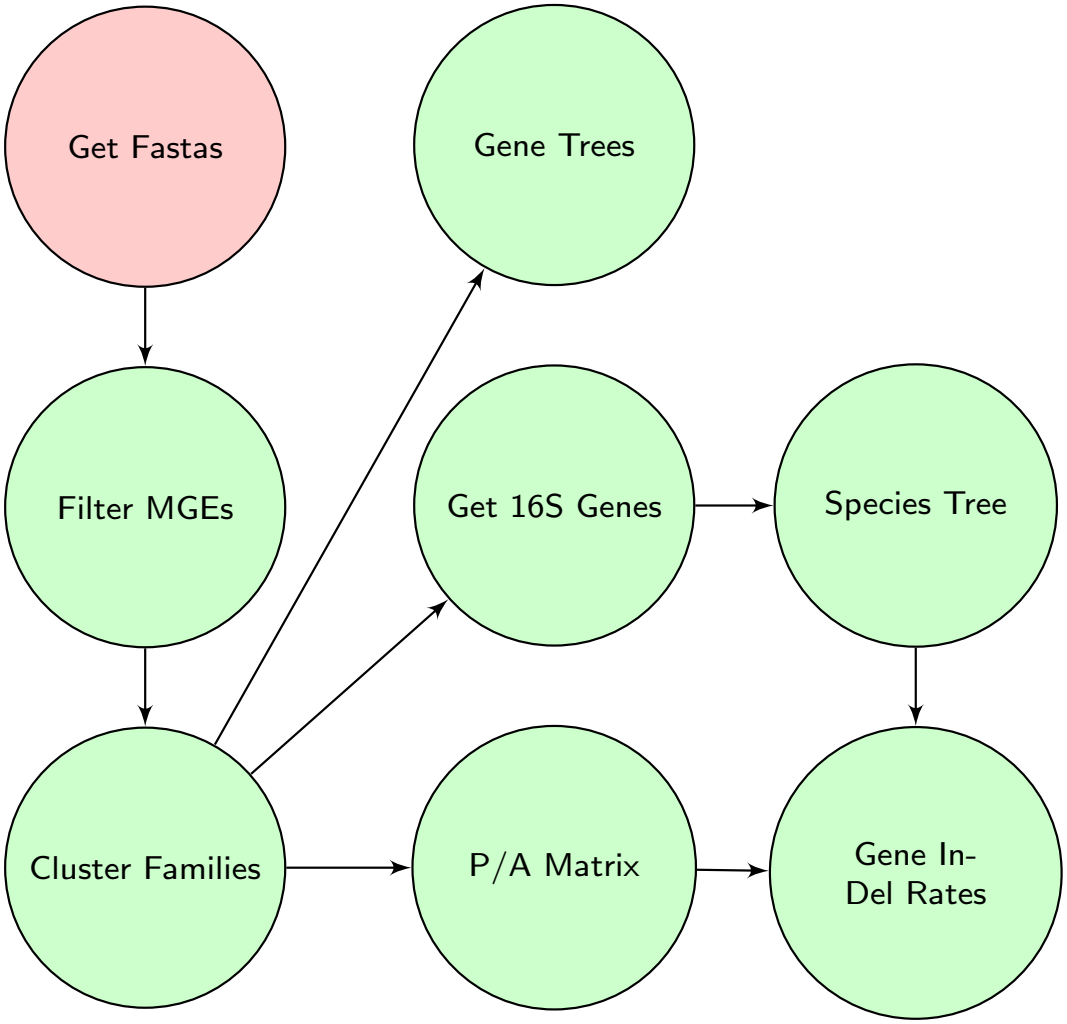
Workflow


Get Fastas
Filter MGEs
Get 16S Genes
Species Tree
Cluster Families
P/A Matrix

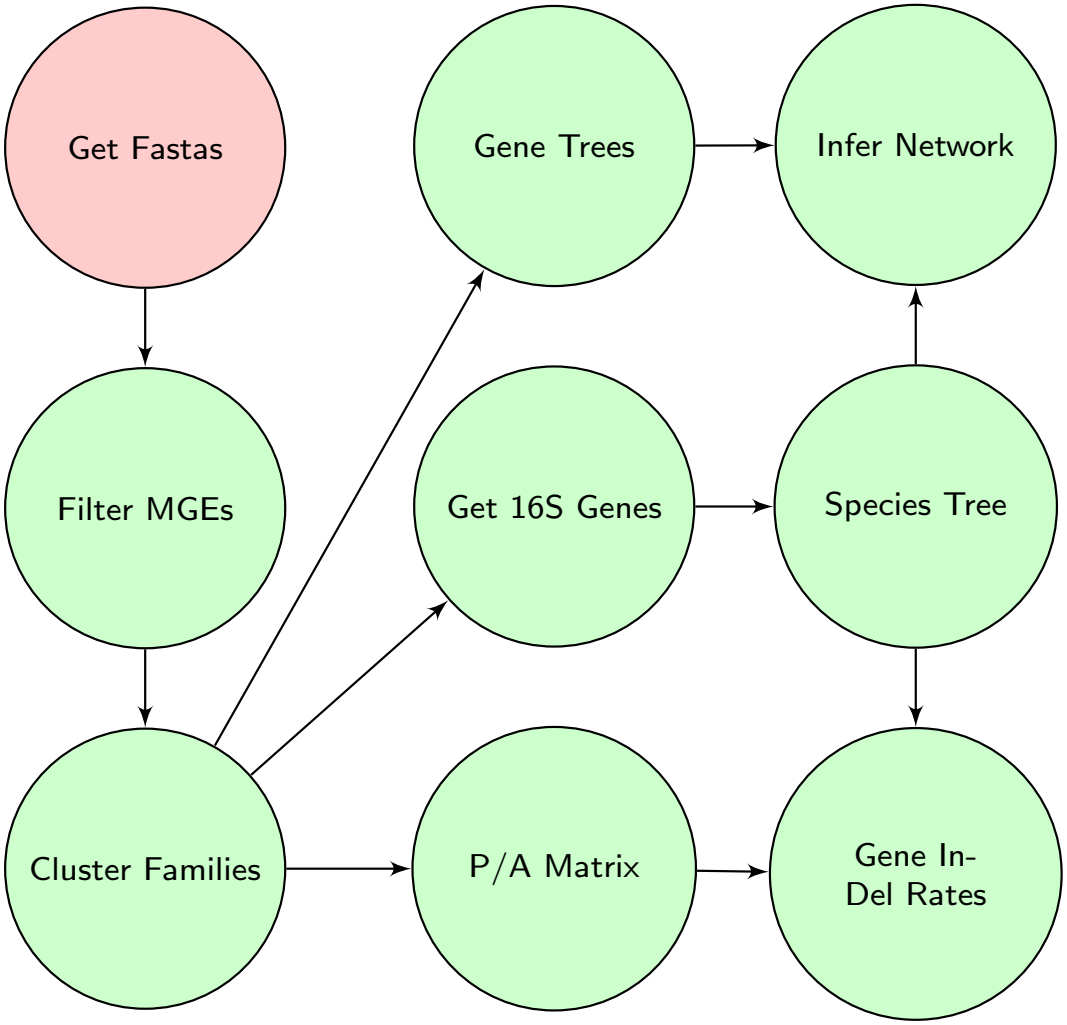
Workflow


Get Fastas
Filter MGEs
Get 16S Genes
Species Tree
Cluster Families
P/A Matrix
Gene In-
Del Rates

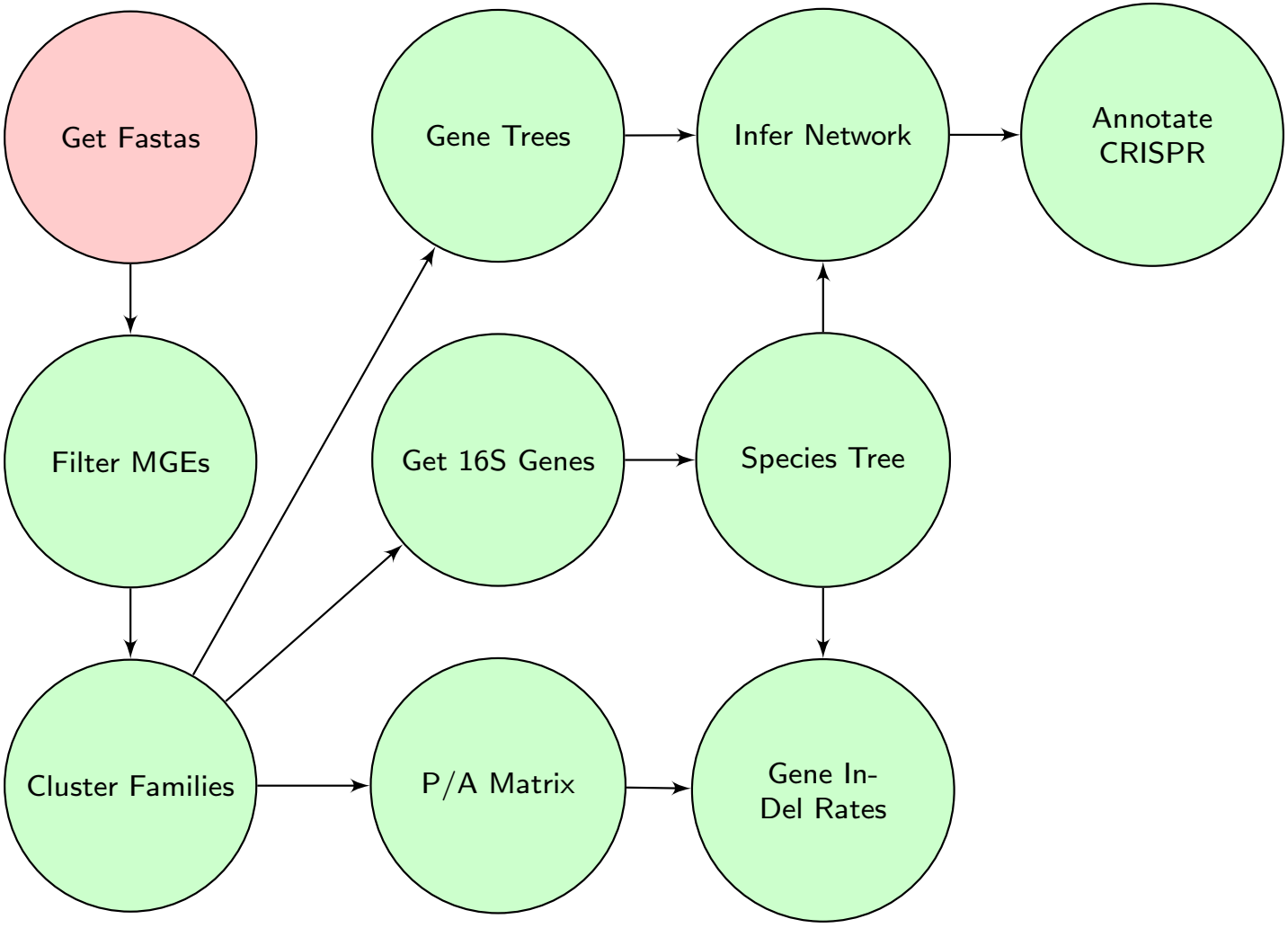
Workflow


Get Fastas
Gene Trees
Filter MGEs
Get 16S Genes
Species Tree
Cluster Families
P/A Matrix
Gene In-
Del Rates

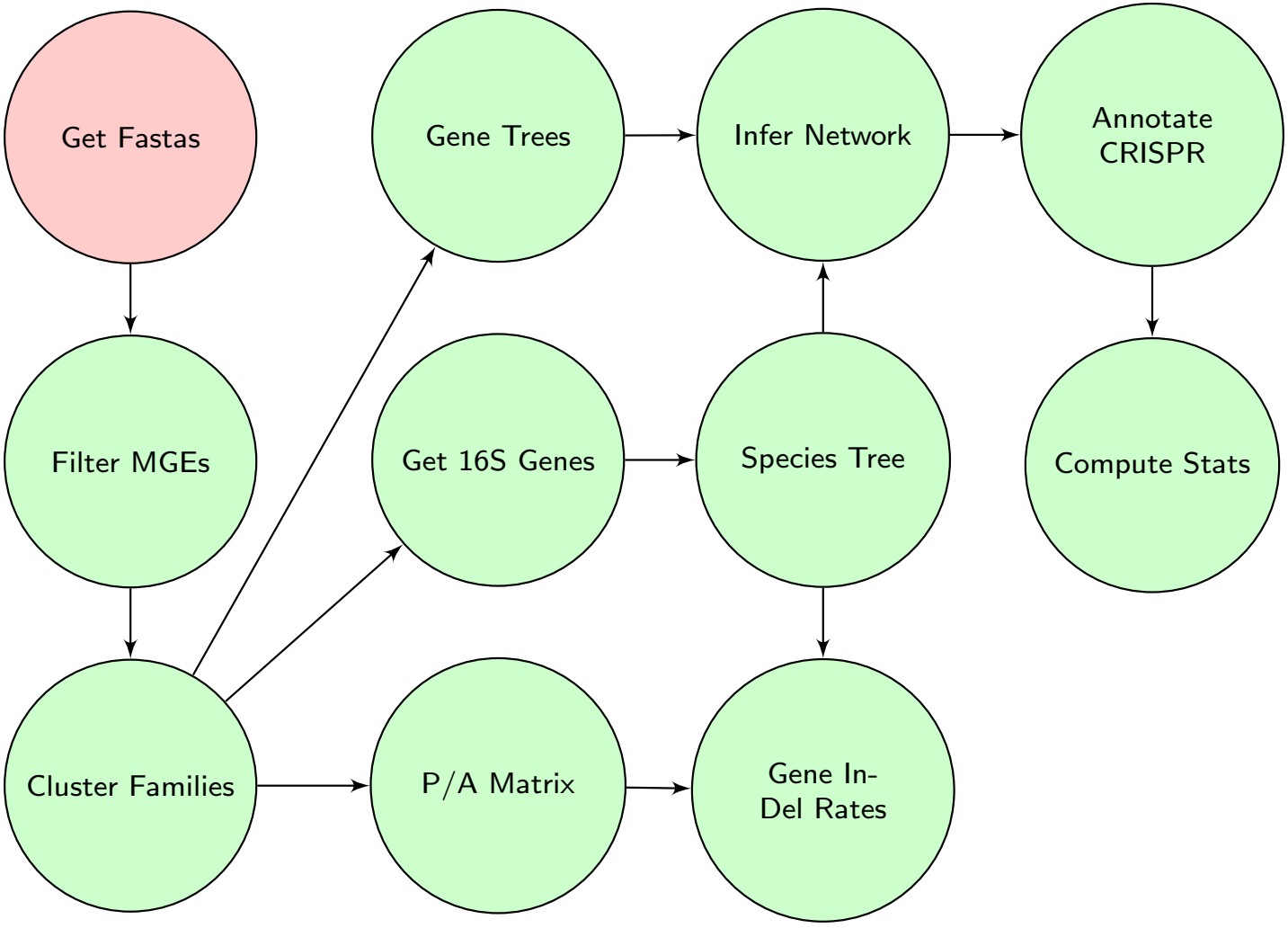
Workflow


Get Fastas
Gene Trees
Infer Network
Filter MGEs
Get 16S Genes
Species Tree
Cluster Families
P/A Matrix
Gene In-
Del Rates

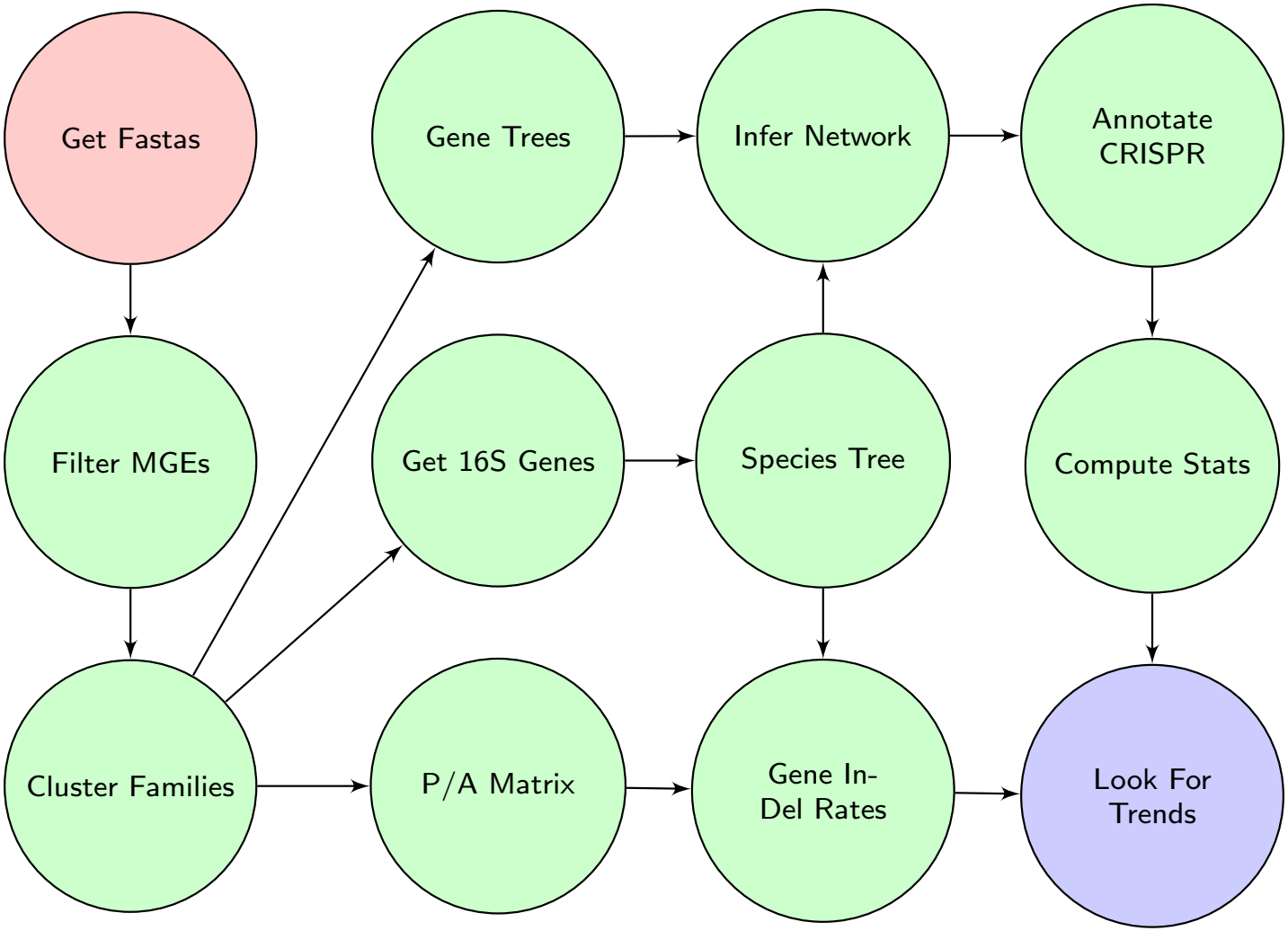
Workflow


Get Fastas
Gene Trees
Infer Network
Annotate CRISPR
Filter MGEs
Get 16S Genes
Species Tree
Cluster Families
P/A Matrix
Gene In-Del Rates

Workflow


Get Fastas
Gene Trees
Infer Network
Annotate CRISPR
Filter MGEs
Get 16S Genes
Species Tree
Compute Stats
Cluster Families
P/A Matrix
Gene In-Del Rates

Workflow


Get Fastas
Gene Trees
Infer Network
Annotate CRISPR
Filter MGEs
Get 16S Genes
Species Tree
Compute Stats
Cluster Families
P/A Matrix
Gene In-Del Rates
Look For Trends

Network Statistics

Network Statistics

- **Average Node Degree**: $\frac{1}{|N_u|} \sum_{uv}^{N_u} w_{uv}$ where N_u is the set of nodes incidenent to u

Network Statistics

- **Average Node Degree**: $\frac{1}{|N_u|}\sum_{uv}^{N_u} w_{uv}$ where N_u is the set of nodes incidenent to u
- **Average Edge Weight**: $\frac{1}{N_c}\sum_i w_i$, The average edge weight for all nodes with CRISPR or without CRISPR

Network Statistics

- **Average Node Degree**: $\frac{1}{|N_u|}\sum_{uv}^{N_u} w_{uv}$ where N_u is the set of nodes incidenent to u
- **Average Edge Weight**: $\frac{1}{N_c}\sum_i w_i$, The average edge weight for all nodes with CRISPR or without CRISPR
- **Node Clustering Coefficient**: $\frac{1}{k_u(k_u-1)}\sum_{vw}^{T(u)}\left(\hat{w}_{uw}\hat{w}_{vw}\hat{w}_{uv}\right)^{\frac{1}{3}}$ where $T(u)$ is the set of traingles containing u[18]

Network Statistics

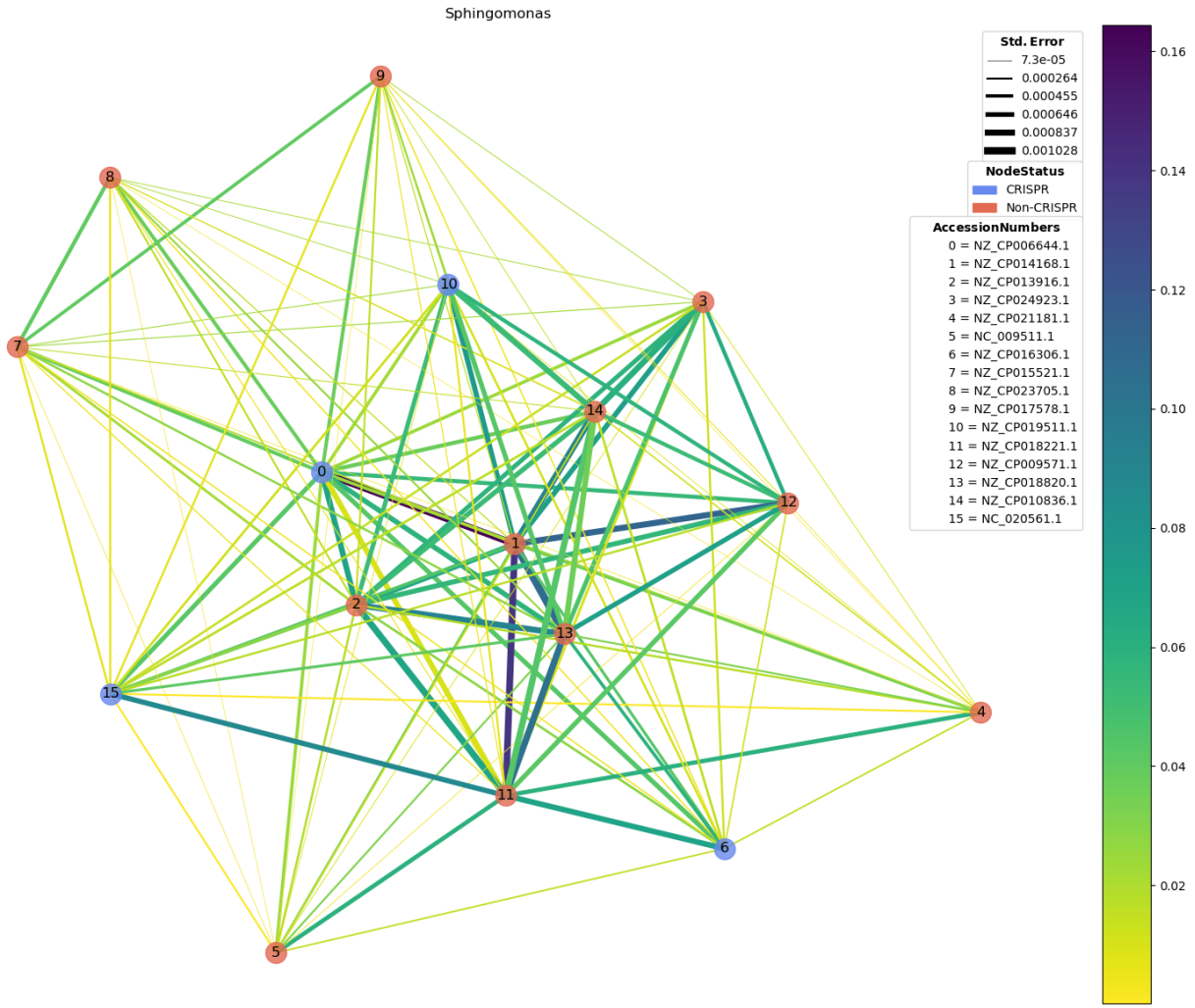
- **Average Node Degree**: $\frac{1}{|N_u|}\sum_{uv}^{N_u} w_{uv}$ where N_u is the set of nodes incidenent to u
- **Average Edge Weight**: $\frac{1}{N_c}\sum_i w_i$, The average edge weight for all nodes with CRISPR or without CRISPR
- **Node Clustering Coefficient**: $\frac{1}{k_u(k_u-1)}\sum_{vw}^{T(u)}(\hat{w}_{uw}\hat{w}_{vw}\hat{w}_{uv})^{\frac{1}{3}}$ where $T(u)$ is the set of traingles containing u[18]
- **Node Assortativity**: $\frac{Tr(M)-||M^2||}{1-||M^2||}$ Where M is the mixing matrix of a given attribute and $||M||$ is the sum of all elements of M.[19]

Network Statistics

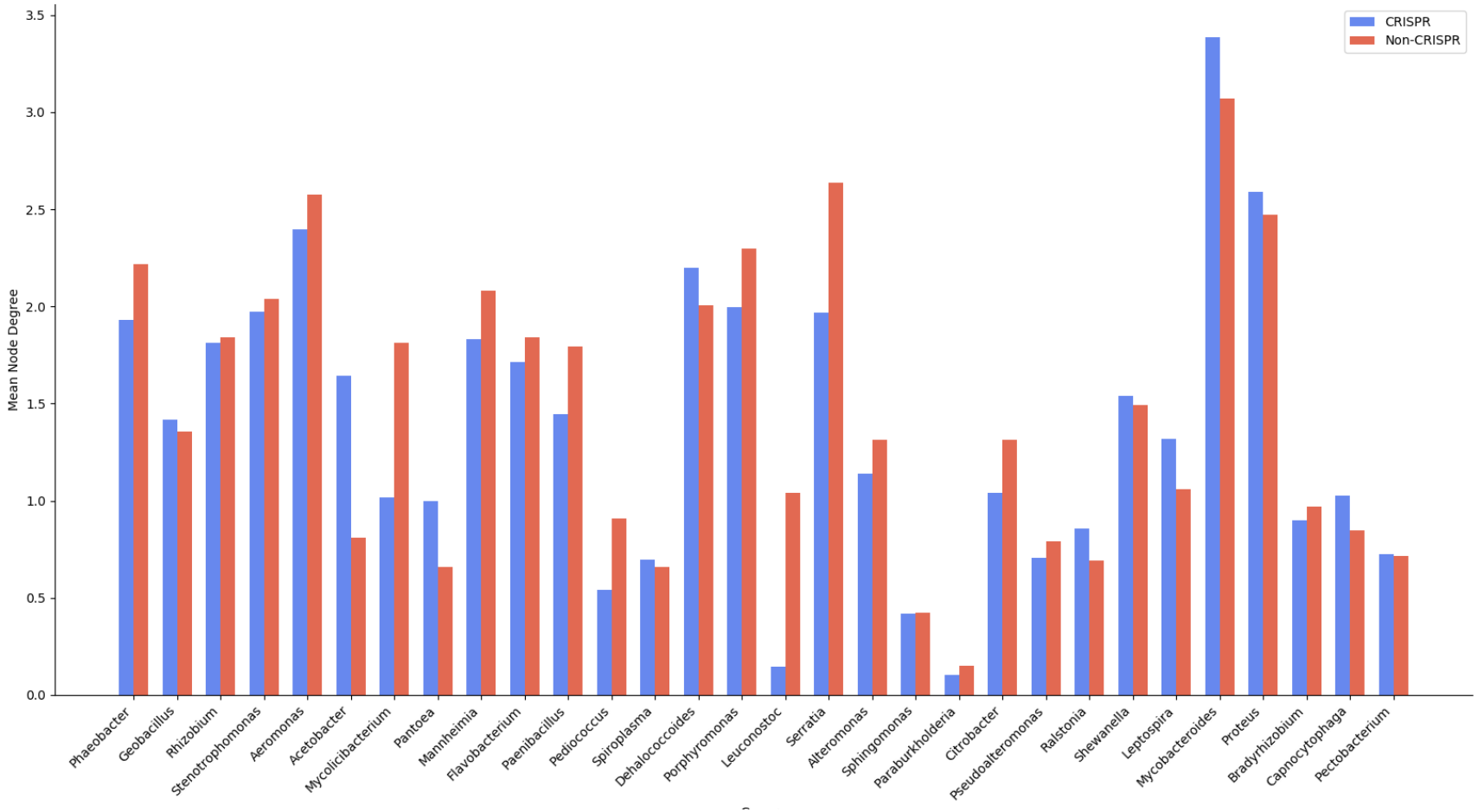
- **Average Node Degree**: $\frac{1}{|N_u|}\sum_{uv}^{N_u} w_{uv}$ where N_u is the set of nodes incidenent to u
- **Average Edge Weight**: $\frac{1}{N_c}\sum_i w_i$, The average edge weight for all nodes with CRISPR or without CRISPR
- **Node Clustering Coefficient**: $\frac{1}{k_u(k_u-1)}\sum_{vw}^{T(u)}(\hat{w}_{uw}\hat{w}_{vw}\hat{w}_{uv})^{\frac{1}{3}}$ where $T(u)$ is the set of traingles containing u[18]
- **Node Assortativity**: $\frac{Tr(M)-||M^2||}{1-||M^2||}$ Where M is the mixing matrix of a given attribute and $||M||$ is the sum of all elements of M.[19]
- **Network Modularity**: $Q = \frac{1}{2m}\sum_{uv}^{W}[W_{uv} - \frac{k_u k_v}{2m}]\delta(u,v)$ where m is the total weight of alledges, k_u is the degree of u and $\delta(u,v)$ is 1 if u and v both have or do not have CRISPR systems and 0 otherwise. $Q \in [-1,1]$[20]

Results

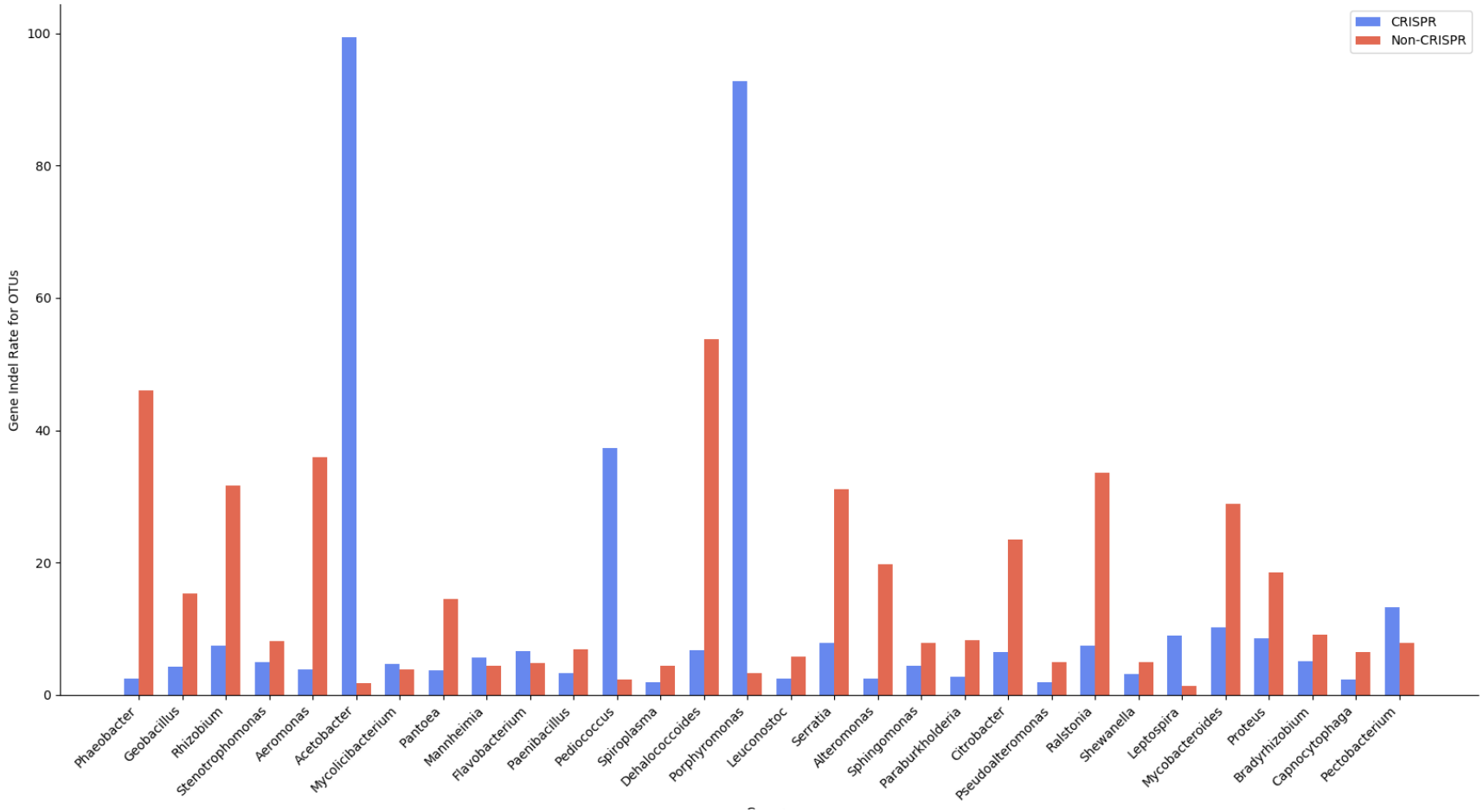
Example "Consensus" Network



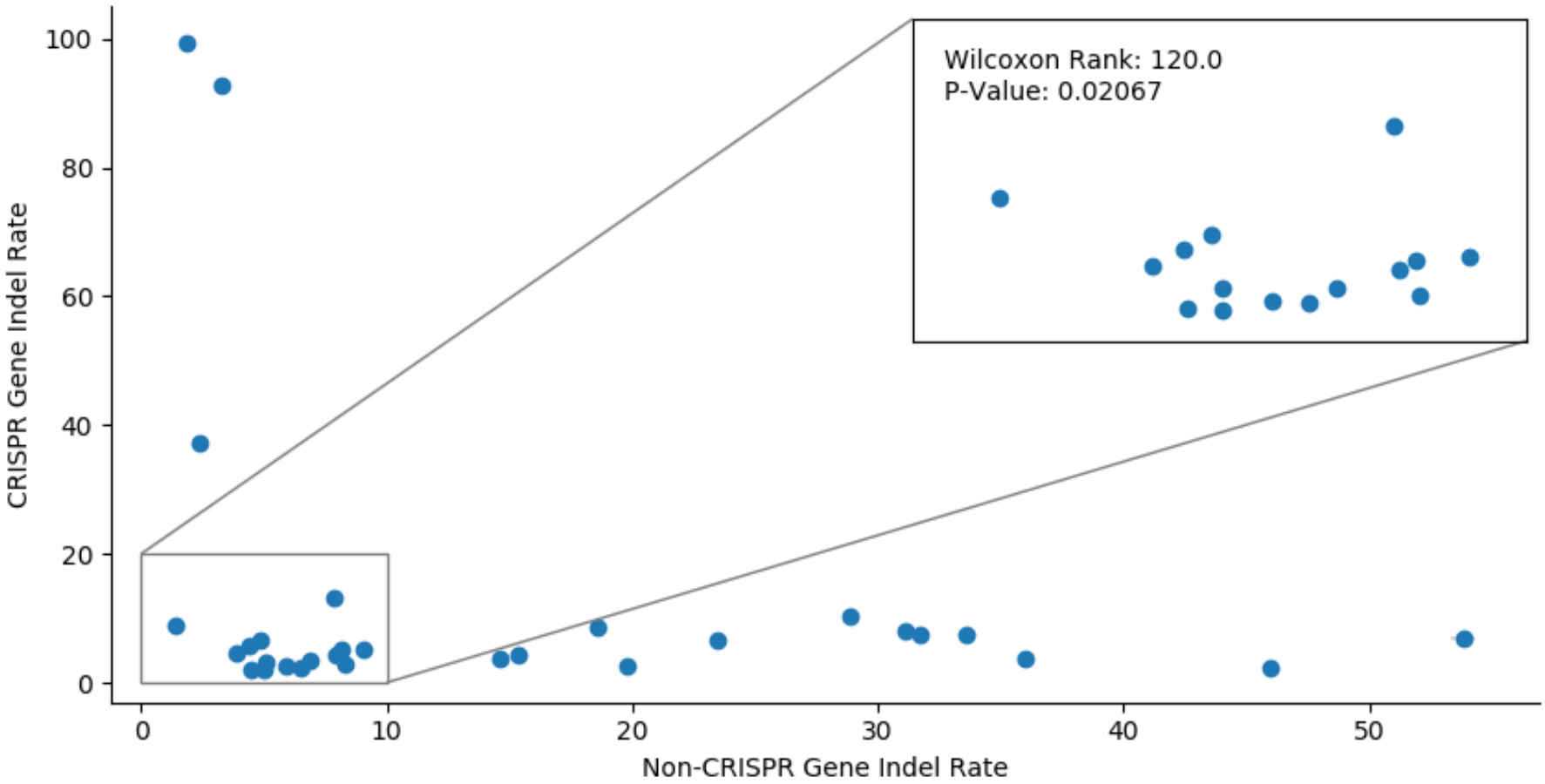
Mean Node Degree



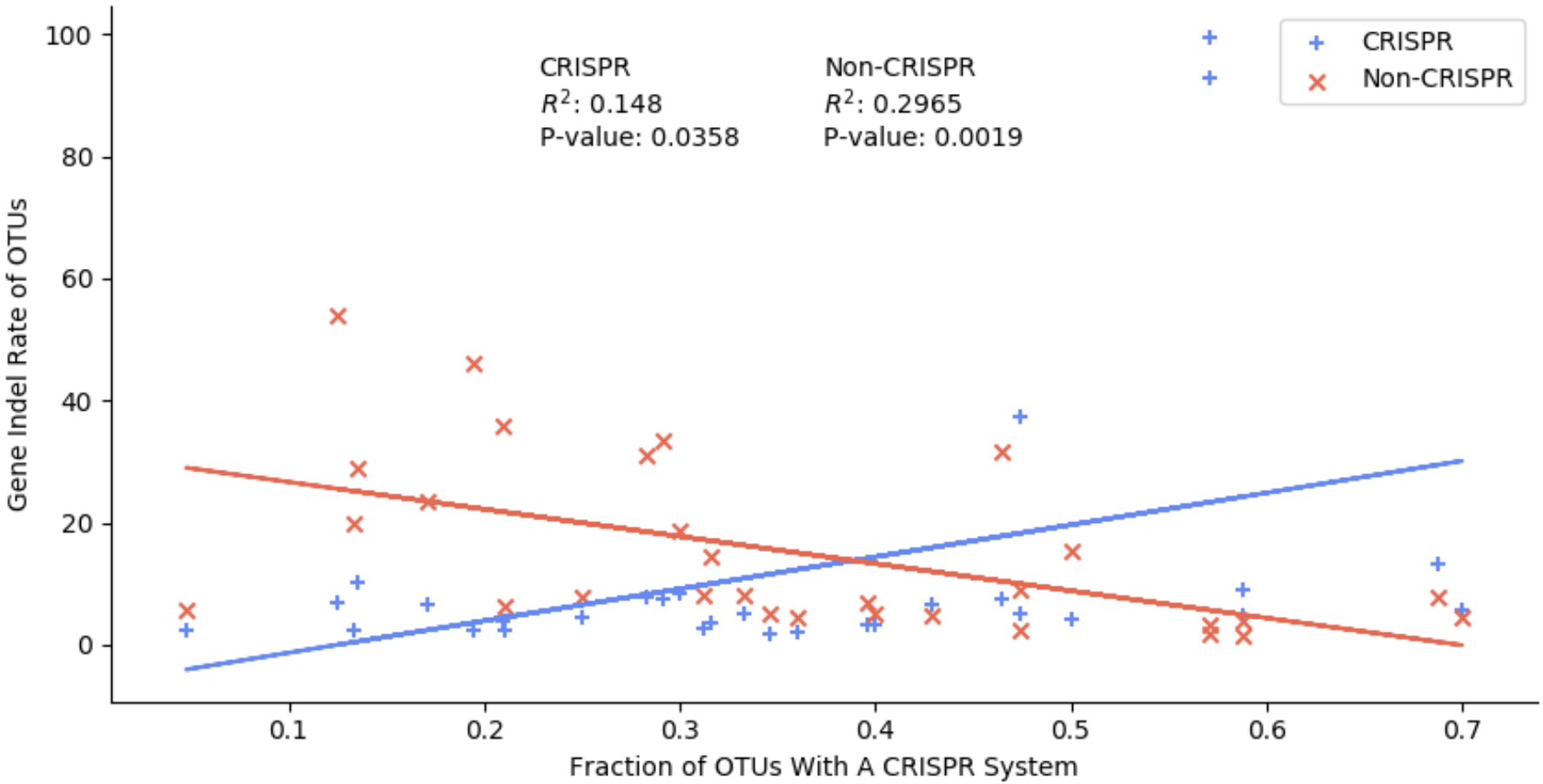
Gene Indel Rates



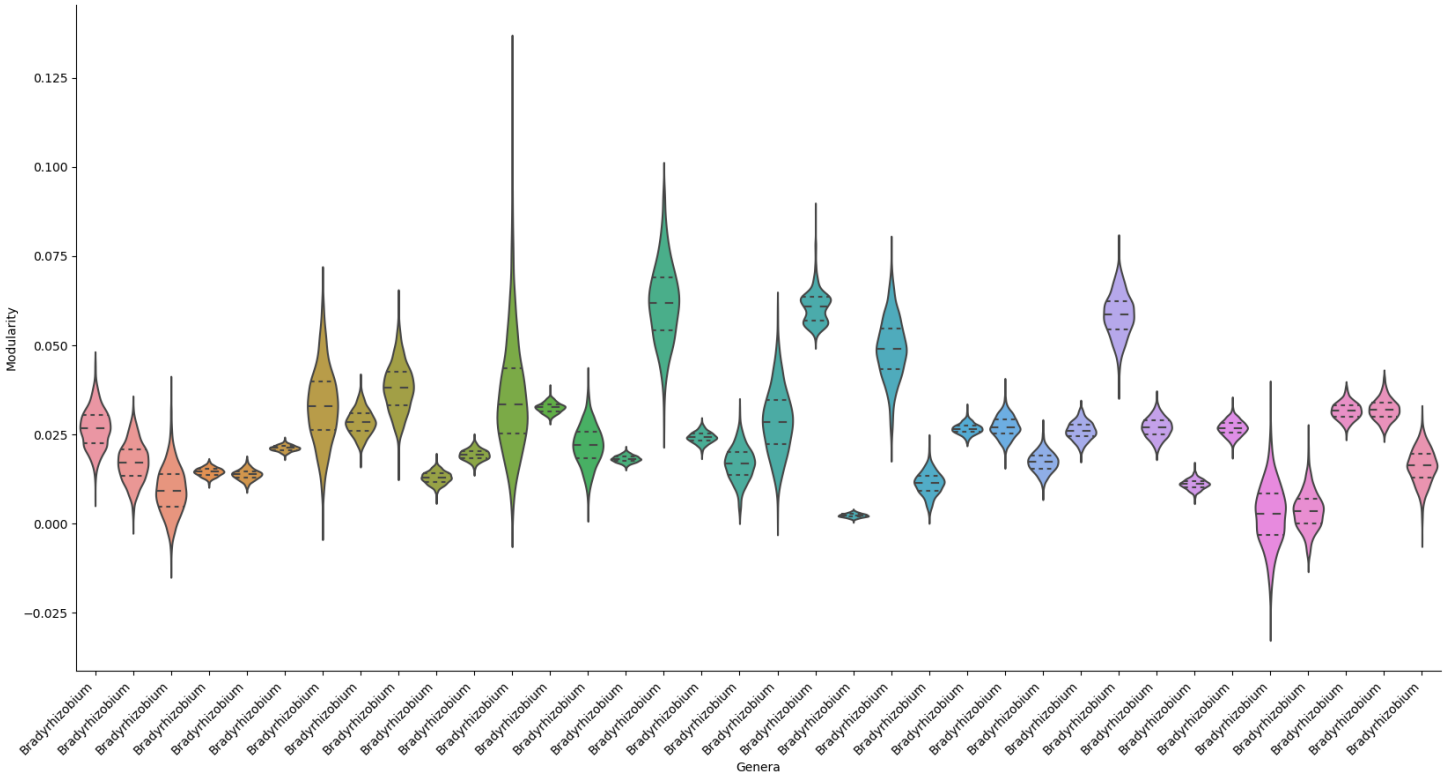
Gene Indel Rates



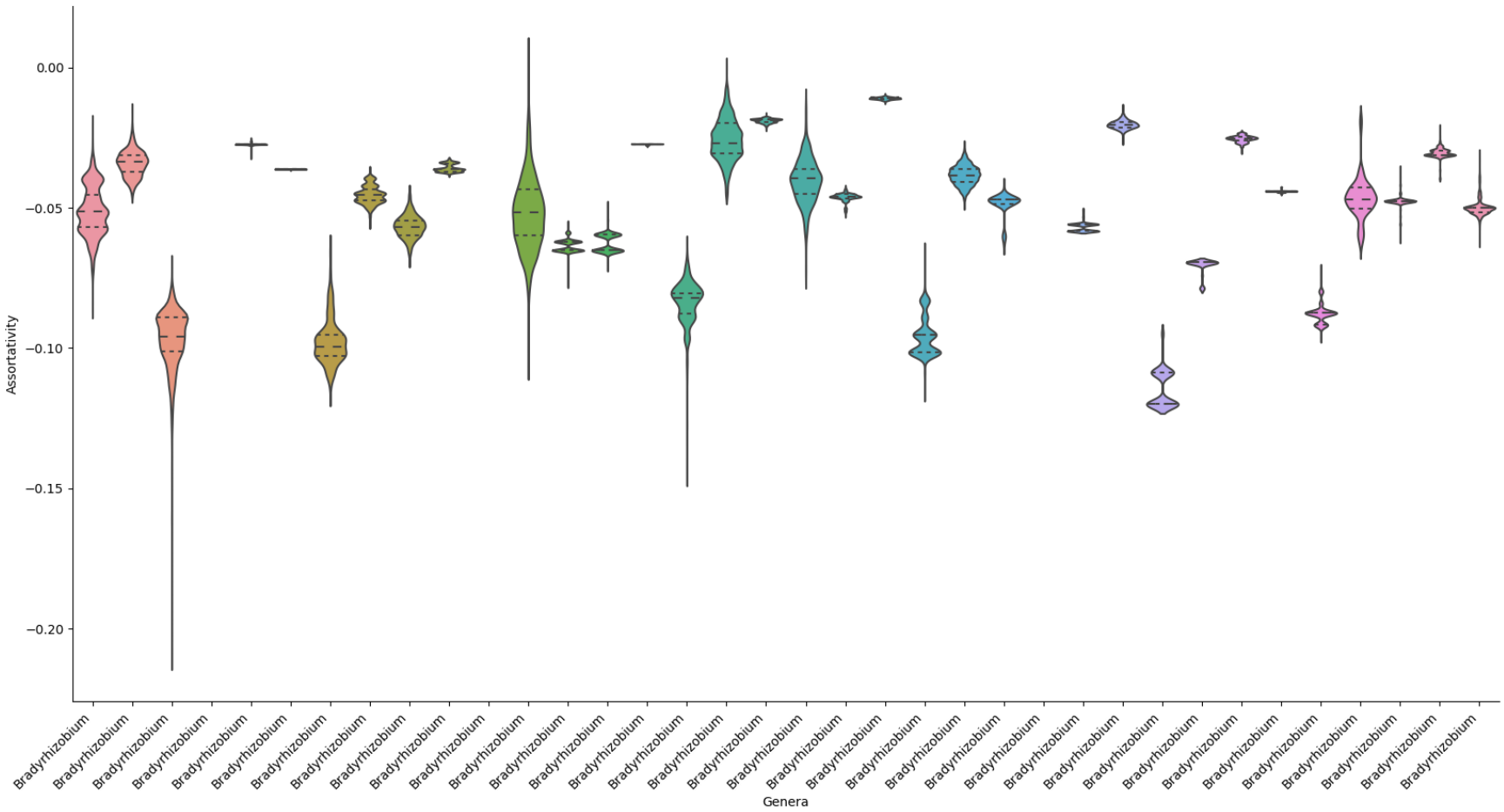
Gene Indel Rate Difference Vs. Fraction CRISPR Species



Modularity Distributions



Assortativity Distributions



Limitations & Caveats

Limitations & Caveats

- **Ignored Singletons**: Genes that did not cluster into any families were ingnored from future steps, but may have still represented horiznotally transferred genes

Limitations & Caveats

- **Ignored Singletons**: Genes that did not cluster into any families were ingnored from future steps, but may have still represented horiznotally transferred genes
- **Ignored Some Gene Families**: For time considerations, only 1500 gene trees were generated for each genus

Limitations & Caveats

- **Ignored Singletons**: Genes that did not cluster into any families were ingnored from future steps, but may have still represented horiznotally transferred genes
- **Ignored Some Gene Families**: For time considerations, only 1500 gene trees were generated for each genus
- **Signifigance Testing**: Samples are not necessairily independant in a network, further node statistics can only be tested for genera with > 20 CRISPR and non-CRISPR OTUs.

Limitations & Caveats

- **Ignored Singletons**: Genes that did not cluster into any families were ingnored from future steps, but may have still represented horiznotally transferred genes
- **Ignored Some Gene Families**: For time considerations, only 1500 gene trees were generated for each genus
- **Signifigance Testing**: Samples are not necessairily independant in a network, further node statistics can only be tested for genera with > 20 CRISPR and non-CRISPR OTUs.
- **Taxonomic Mistakes**: Inconsistencies in taxonomic labelling can result in ignored or misplaced OTUs.

Limitations & Caveats

- **Ignored Singletons**: Genes that did not cluster into any families were ingnored from future steps, but may have still represented horiznotally transferred genes
- **Ignored Some Gene Families**: For time considerations, only 1500 gene trees were generated for each genus
- **Signifigance Testing**: Samples are not necessairily independant in a network, further node statistics can only be tested for genera with > 20 CRISPR and non-CRISPR OTUs.
- **Taxonomic Mistakes**: Inconsistencies in taxonomic labelling can result in ignored or misplaced OTUs.
- **Multifurcation Error**: Some species trees contained multifurcations, which were resolved randomly to generate a bifurcating tree. Estimating this error by examining variance over different resolutions is possible.

Possible Future Directions

Possible Future Directions

- **Infering direction**: Directed networks have a host of available analytic tools undirected networks do not

Possible Future Directions

- **Infering direction**: Directed networks have a host of available analytic tools undirected networks do not
- **Gene function analysis**: Considering the transfer dynamics of different functional classes of genes

Possible Future Directions

- **Infering direction**: Directed networks have a host of available analytic tools undirected networks do not
- **Gene function analysis**: Considering the transfer dynamics of different functional classes of genes
- **Studying movement of CRISPR systems**: Studying how frequently CRISPR systems themselves are transfered from arrays, cas genes

Possible Future Directions

- **Infering direction**: Directed networks have a host of available analytic tools undirected networks do not
- **Gene function analysis**: Considering the transfer dynamics of different functional classes of genes
- **Studying movement of CRISPR systems**: Studying how frequently CRISPR systems themselves are transfered from arrays, cas genes
- **Intergenic comparisons**: Combine any set of fasta files for OTUs for analyzing transfer dynamics

Possible Future Directions

- **Infering direction**: Directed networks have a host of available analytic tools undirected networks do not
- **Gene function analysis**: Considering the transfer dynamics of different functional classes of genes
- **Studying movement of CRISPR systems**: Studying how frequently CRISPR systems themselves are transfered from arrays, cas genes
- **Intergenic comparisons**: Combine any set of fasta files for OTUs for analyzing transfer dynamics
- **Continuous CRISPR activity**: Labelling nodes by estimated CRISPR activity (array length, transciptomic data, etc.)

Possible Future Directions

- **Infering direction**: Directed networks have a host of available analytic tools undirected networks do not
- **Gene function analysis**: Considering the transfer dynamics of different functional classes of genes
- **Studying movement of CRISPR systems**: Studying how frequently CRISPR systems themselves are transfered from arrays, cas genes
- **Intergenic comparisons**: Combine any set of fasta files for OTUs for analyzing transfer dynamics
- **Continuous CRISPR activity**: Labelling nodes by estimated CRISPR activity (array length, transciptomic data, etc.)
- **Considering bacterial ecology and environments**: Consider geographically close OTUs or differences between networks due to environmental factors

Conclusion

Thanks

Thank you to

- Dr. G. Brian Golding
- Dr. Ben Evans
- The Golding lab
  - Caitlin Simopoulos
  - Daniella Lato
  - Zachery Dickson
  - Sam Long
  - Geoge Long
  - Lucy Zhang
  - Brianne Laverty
  - Nicole Zhang
- Everyone here for listening

All code used for this project is available at https://github.com/DJSiddharthVader/thesis_SidReed

References (1)


Devashish Rath et al. "The CRISPR-Cas immune system: Biology, mechanisms and applications". In: *Biochimie* 117 (2015). Special Issue: Regulatory RNAs, pp. 119–128. ISSN: 0300-9084.


GRissa, I. and Drevet, C. and Couvin, D. *CRISPRdb*. http://crispr.i2bc.paris-saclay.fr/. Online; accessed 22 October 2018. 2017.


J. Bondy-Denomy and A. R. Davidson. "To Acquire Or Resist:The Complex Biological Effects Of CRISPR-Cas systems". In: *Trends Microbio.* 22.4 (2014), pp. 218–25.


Quan Zhang and Yuzhen Ye. "Not all predicted CRISPR–Cas systems are equal: isolated cas genes and classes of CRISPR like elements". In: *BMC Bioinformatics* 18.1 (Feb. 2017), p. 92. ISSN: 1471-2105.

References (2)


K. S. Makarova et al. "Evolution and classification of the CRISPR-Cas systems". In: *Nat. Rev. Microbiol.* 9.6 (2011), pp. 467–477.


Ovidiu Popa and Tal Dagan. "Trends and barriers to lateral gene transfer in prokaryotes". In: *Current Opinion in Microbiology* 14.5 (2011). Antimicrobials/Genomics, pp. 615–623. ISSN: 1369-5274.


L. C. Guimaraes et al. "Inside the Pan-genome - Methods and Software Overview". In: *Curr. Genomics* 16.4 (2015), pp. 245–252.


David A. Rasko et al. "The Pangenome Structure of Escherichia coli: Comparative Genomic Analysis of E. coli Commensal and Pathogenic Isolates". In: *Journal of Bacteriology* 190.20 (2008), pp. 6881–6893. ISSN: 0021-9193.

References (3)


Björn Berglund. "Environmental dissemination of antibiotic resistance genes and correlation to anthropogenic contamination with antibiotics". In: *Infection Ecology & Epidemiology* 5.1 (2015), p. 28564.


J. A. Bondy and U. S. R. Murty. *Graph theory with applications*. Wiley, 2002.


V. Kunin et al. "The net of life: reconstructing the microbial phylogenetic network". In: *Genome Res.* 15.7 (2005), pp. 954–959.


Matt Ravenhall et al. "Inferring Horizontal Gene Transfer". In: *PLOS Computational Biology* 11.5 (May 2015), pp. 1–16.


Senka Dzidic and Vladimir Bedeković. "Horizontal gene transfer-emerging multidrug resistance in hospital bacteria". In: *Acta pharmacologica Sinica* 24.6 (2003), pp. 519–526.

References (4)

Adi Stern et al. "Self-targeting by CRISPR: gene regulation or autoimmunity?" In: *Trends in Genetics* 26.8 (2010), pp. 335–340. ISSN: 0168-9525.

Bridget N. J. Watson, Raymond H. J. Staals, and Peter C. Fineran. "CRISPR-Cas-Mediated Phage Resistance Enhances Horizontal Gene Transfer by Transduction". In: *mBio* 9.1 (2018). Ed. by Joseph Bondy-Denomy and Michael S. Gilmore.

James S. Godde and Amanda Bickerton. "The Repetitive DNA Elements Called CRISPRs and Their Associated Genes: Evidence of Horizontal Transfer Among Prokaryotes". In: *Journal of Molecular Evolution* 62.6 (June 2006), pp. 718–729. ISSN: 1432-1432.

References (5)


U. Gophna et al. "No evidence of inhibition of horizontal gene transfer by CRISPR-Cas on evolutionary timescales". In: *ISME J* 9.9 (2015), pp. 2021–2027.


J. P. Onnela et al. "Intensity and coherence of motifs in weighted complex networks". In: *Phys Rev E Stat Nonlin Soft Matter Phys* 71.6 Pt 2 (2005), p. 065103.


M. E. Newman. "Assortative mixing in networks". In: *Phys. Rev. Lett.* 89.20 (2002), p. 208701.


M. E. Newman. "Analysis of weighted networks". In: *Phys Rev E Stat Nonlin Soft Matter Phys* 70.5 Pt 2 (2004), p. 056131.