Cancer Regulatory Network Modeling and Strain Level Metagenomics Analysis

Haluk Dogan
https://haluk.github.io/
hdogan@vivaldi.net

Department of Computer Science
University of Nebraska-Lincoln

September 30, 2019

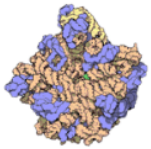
Introduction

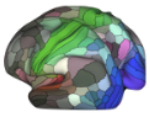
Causal modeling of cancer associated genes, miRNAs, and lncRNAs

Strain level metagenomics analysis

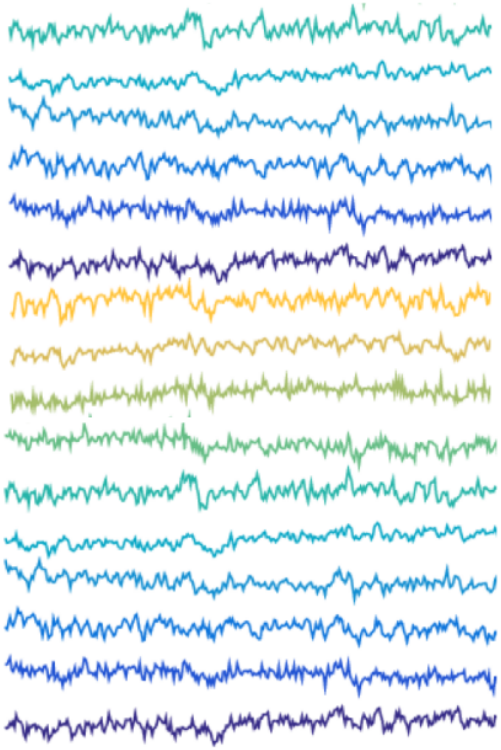
Outline

Network Analysis

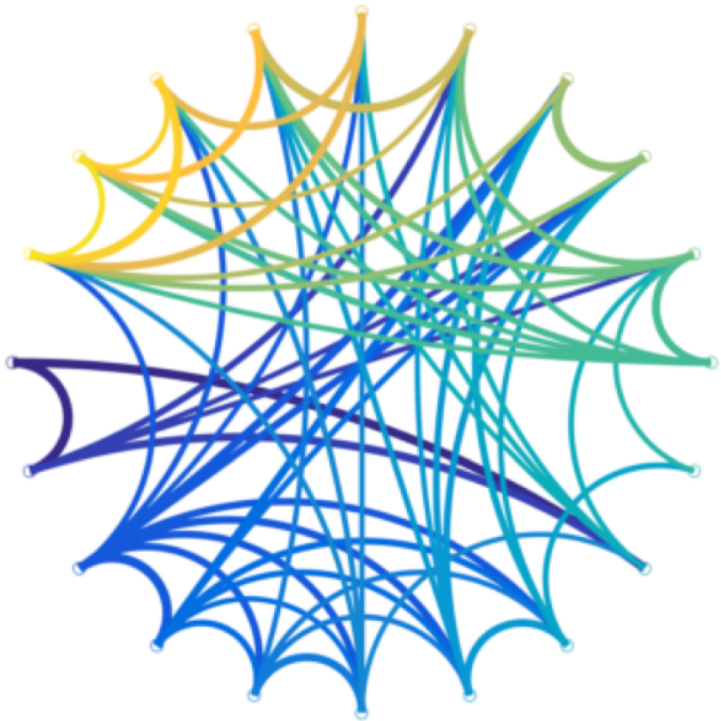




p

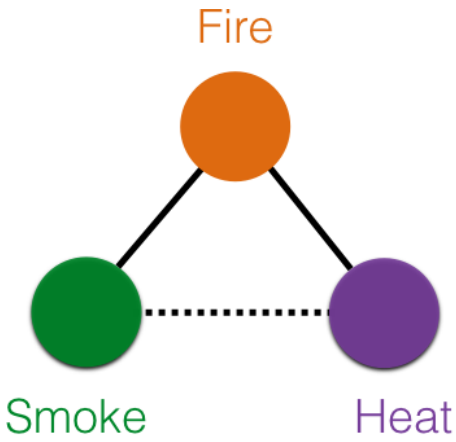


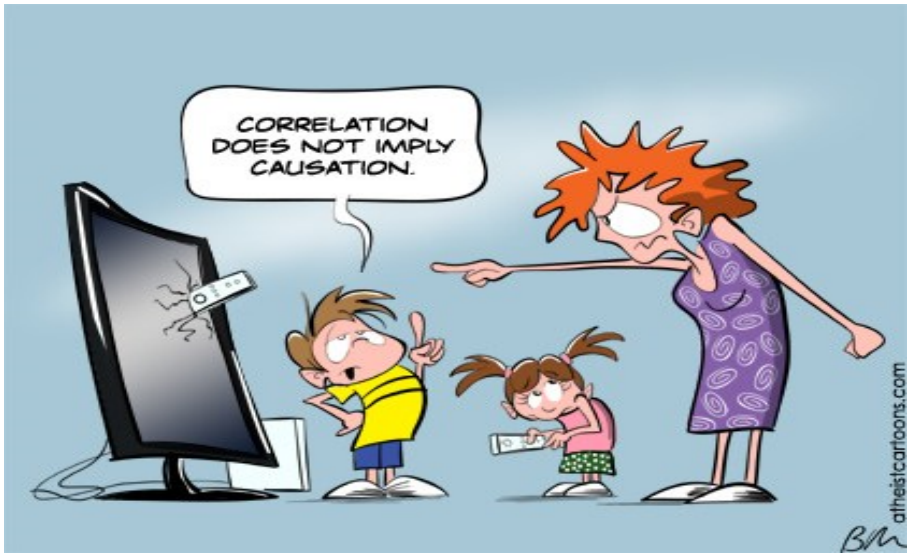
n



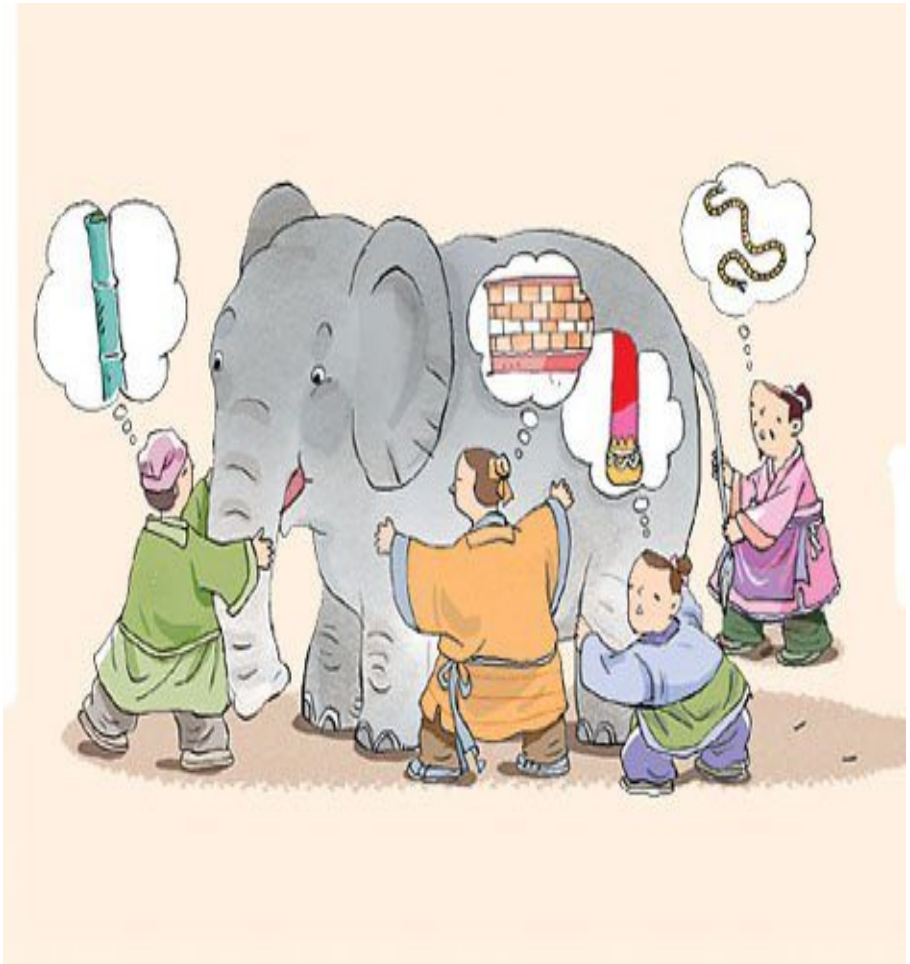
Network Analysis (cont'd)

Correlation network:





Occam's razor:



Gaussian Graphical Models

Gene regulatory network:

- $\boldsymbol{X} = [X_1, X_2, \ldots, X_p]$
- $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
- $P(\boldsymbol{X}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{X} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{X} - \boldsymbol{\mu})\right)$
- $\boldsymbol{\mu} \in \mathbb{R}^p, \boldsymbol{\Sigma} \in \mathbb{S}^p_{++}$
- Precision matrix: $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$
- Challenging problem when $n \ll p$
  Assumptions:
  - $\boldsymbol{\Theta}$ is sparse (ℓ_1)
  - Models for each group should not be too different (ℓ_2)
- Smoke $\perp\!\!\!\perp$ Heat $\mid$ Fire

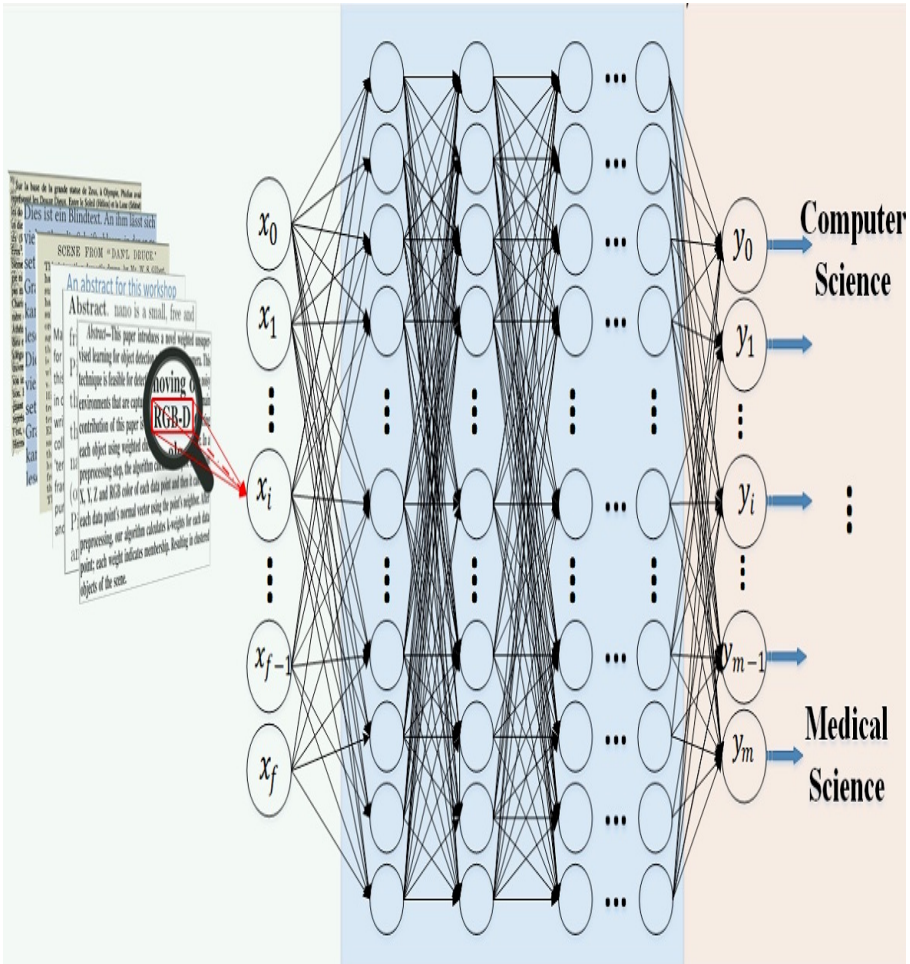
Data Augmentation

Conducting experiments:

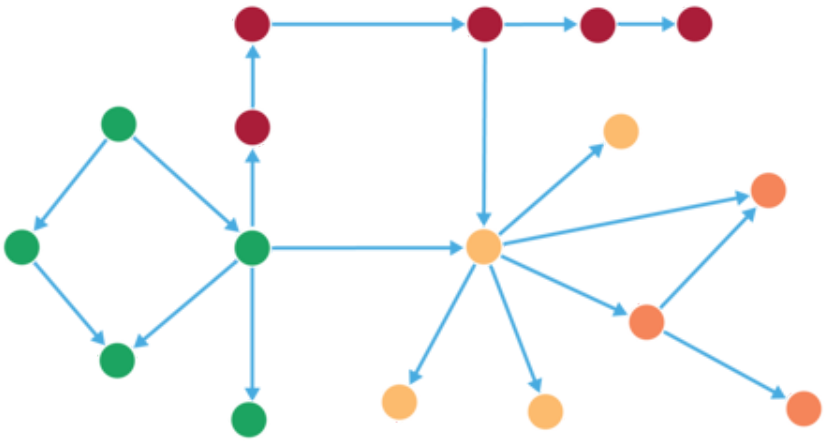
- Unethical
- Expensive
- Difficult to repeat

We used deep learning to generate more in-vitro samples:

- number of samples in groups is imbalanced
- unbiased estimator for a learner



MiRNA Co-Binding Network



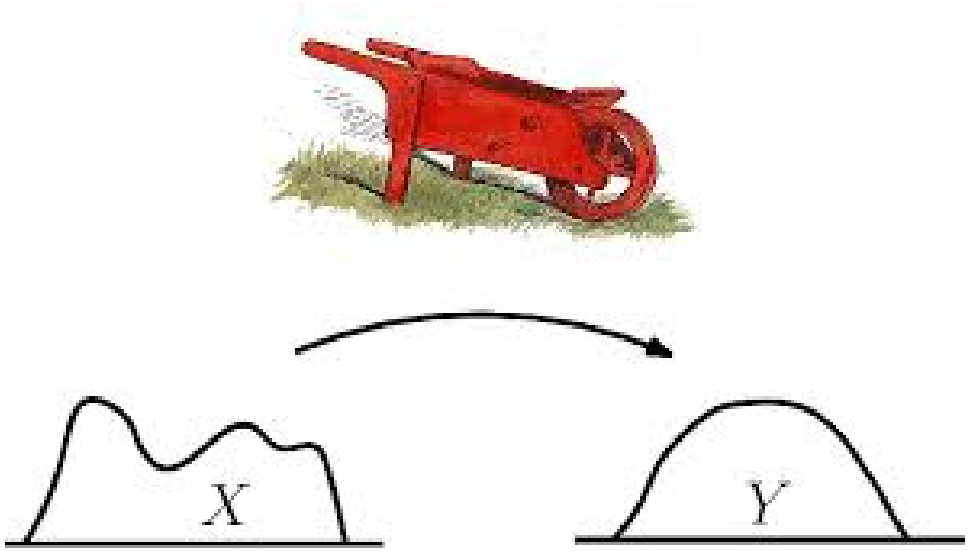
- We constructed a Bayesian network to represent miRNA co-binding relationships
- Evidence matrix is from starBase

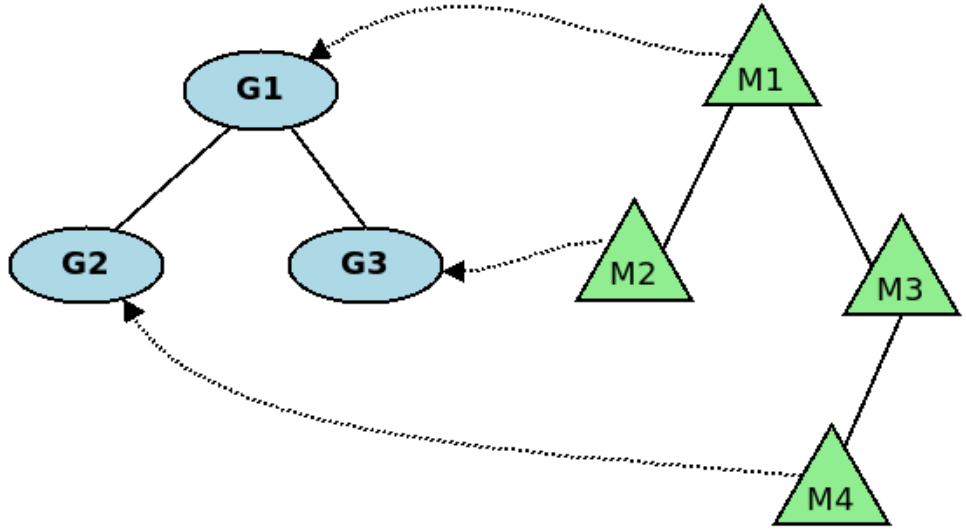
	Gene_1	Gene_2	...	Gene_n
miRNA_1	1/0			
miRNA_2				
⋮				
miRNA_p				

- Intervention to a model is a lot easier

Connecting Models

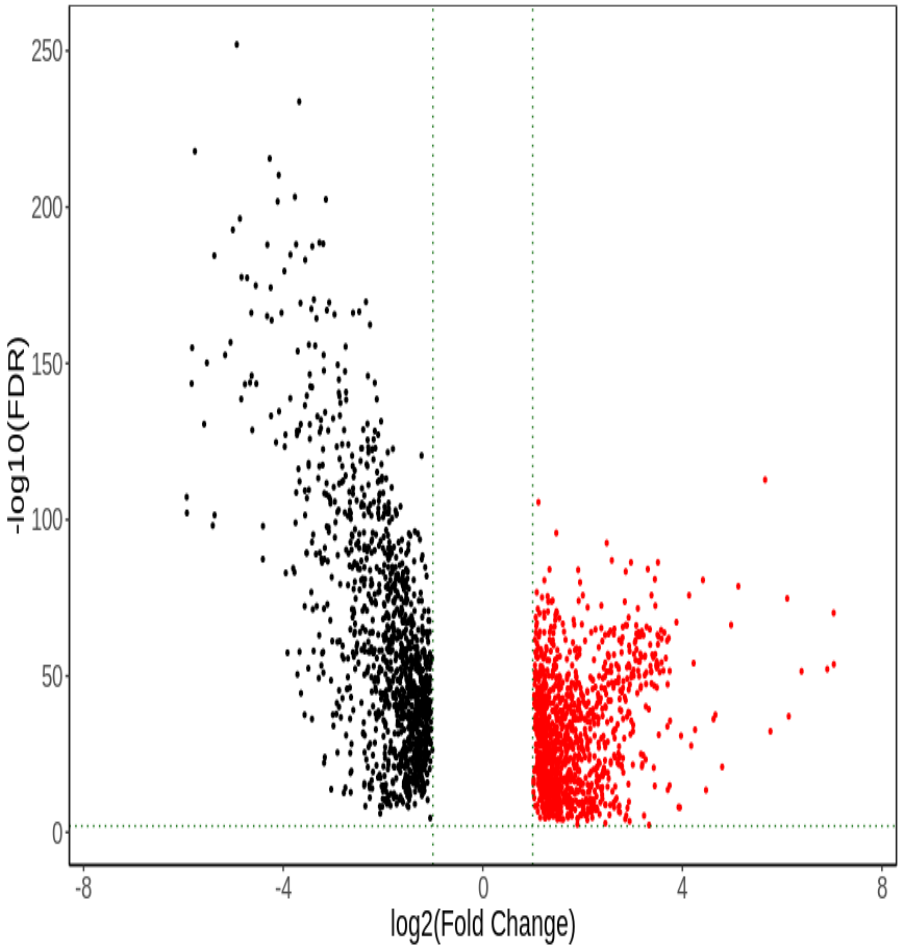
- We test each miRNA and dependencies to see if they make significant group difference
- We used probabilistic distance metric to measure the data similarity with and without testing miRNAs and their dependencies
  - Entropic Gromov Wasserstein



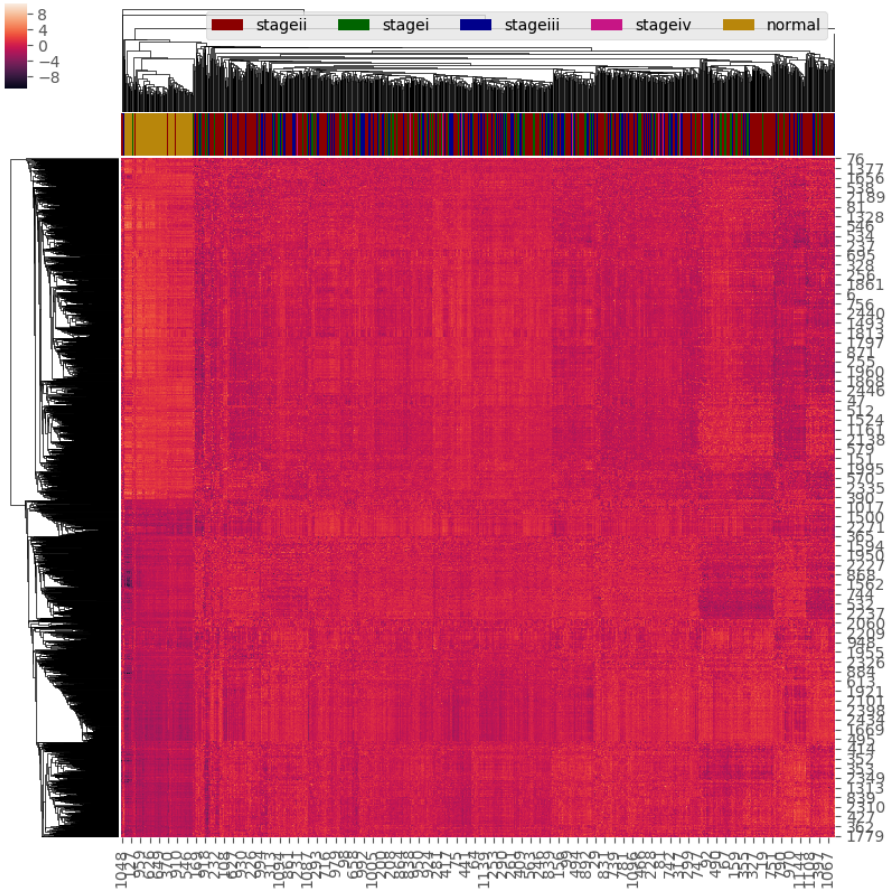


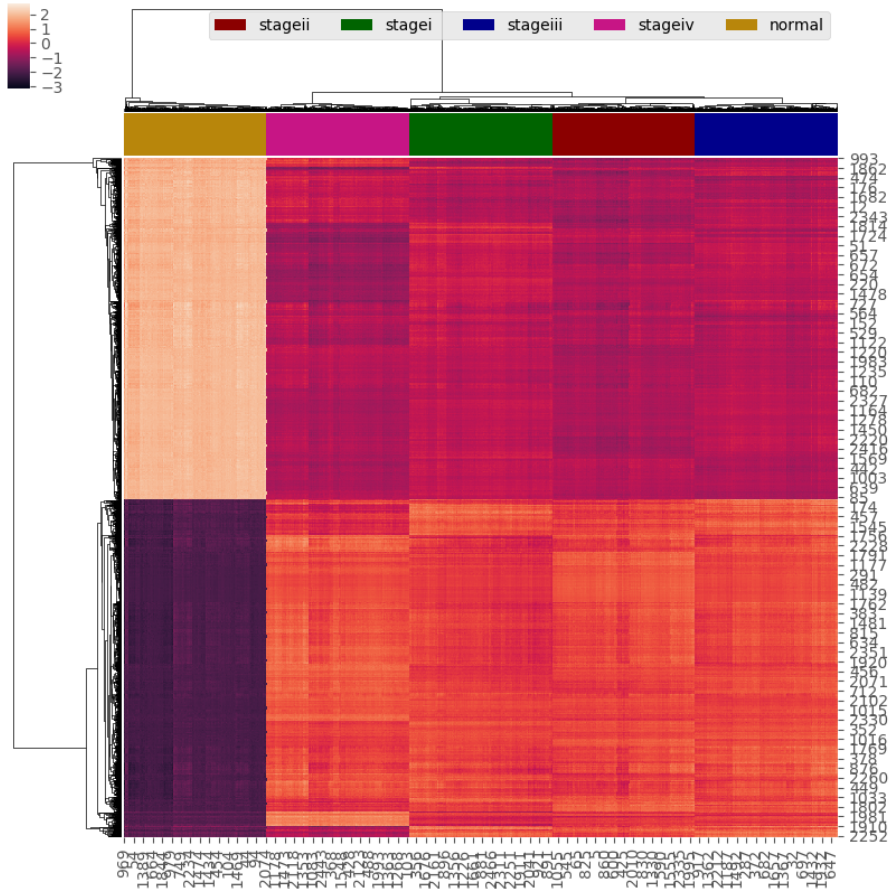
Data

- Data: TCGA-BRCA
- Samples:
  - Normal: 104
  - Stage 1: 179
  - Stage 2: 608
  - Stage 3: 242
  - Stage 4: 20
- DEGs (fold-change > 2):
  - Up: 1218
  - Down: 1236
- 617 miRNAs
  - 1,678 co-binding relationships
  - 20,441/166,669 bindings make significant group difference

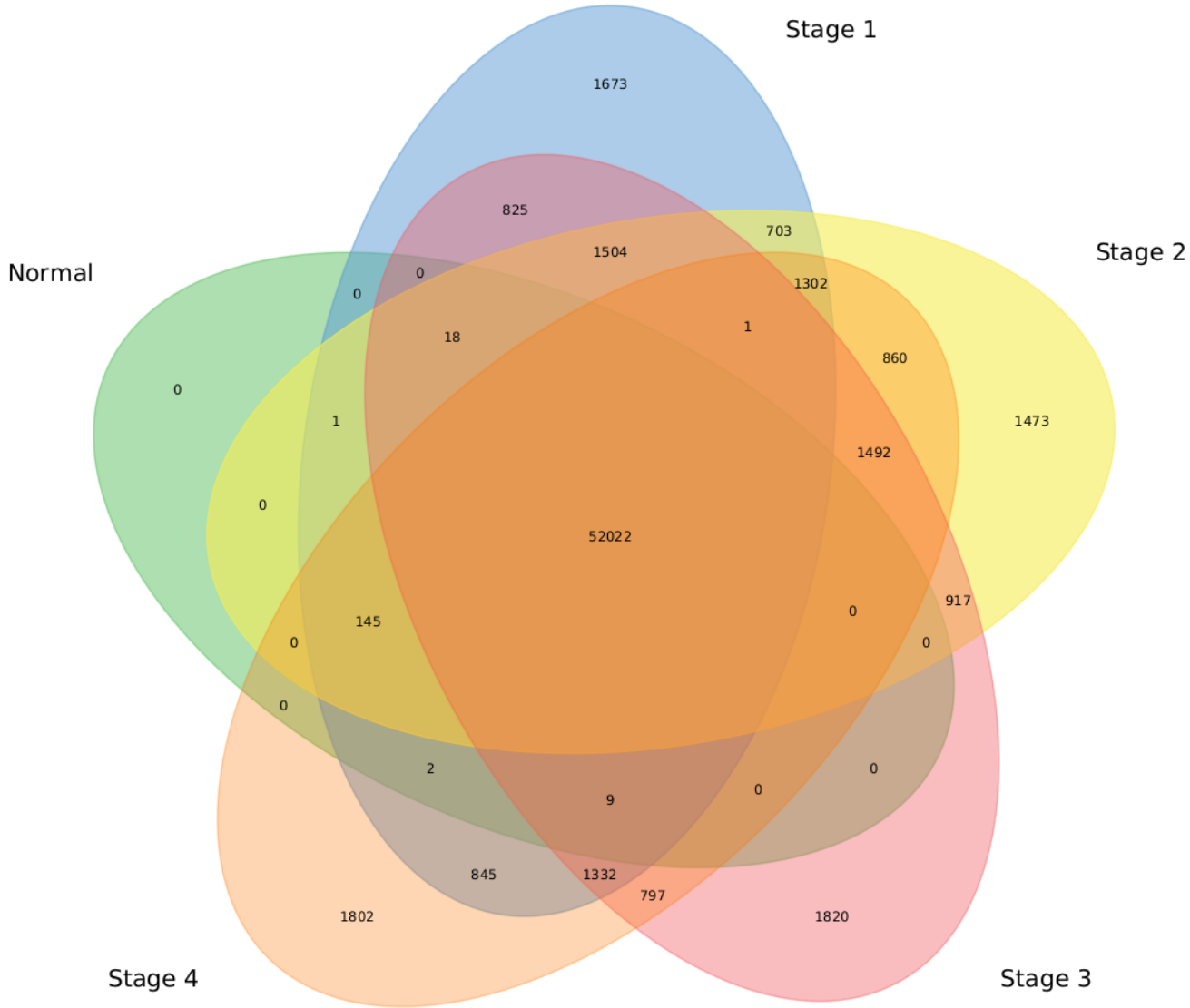


Results


8
4
0
−4
−8
stageii
stagei
stageiii
stageiv
normal


2
1
0
−1
−2
−3
stageii
stagei
stageiii
stageiv
normal

Results (cont'd)



Results (cont'd)

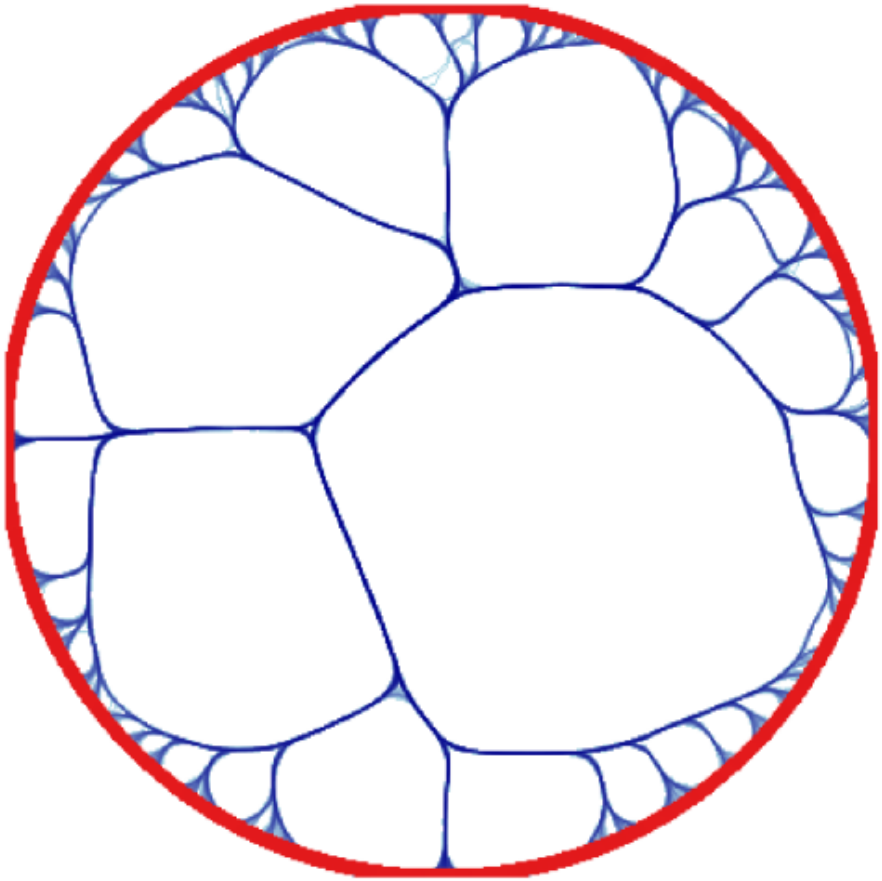


Figure: Stage 1

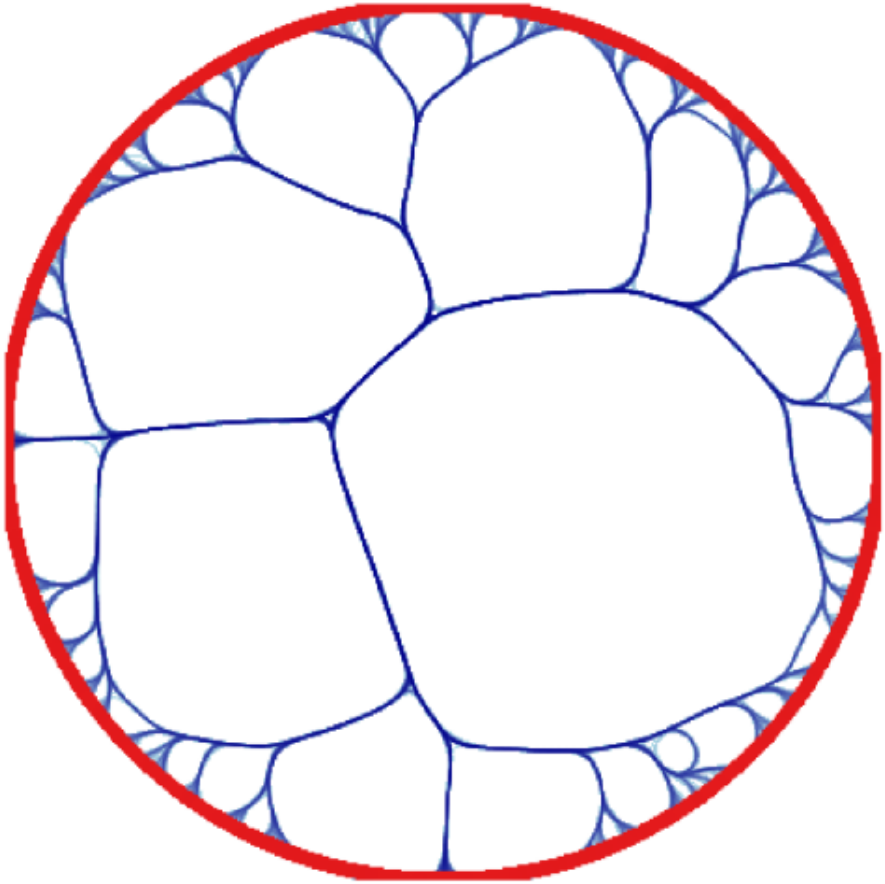


Figure: Stage 2

Results (cont'd)

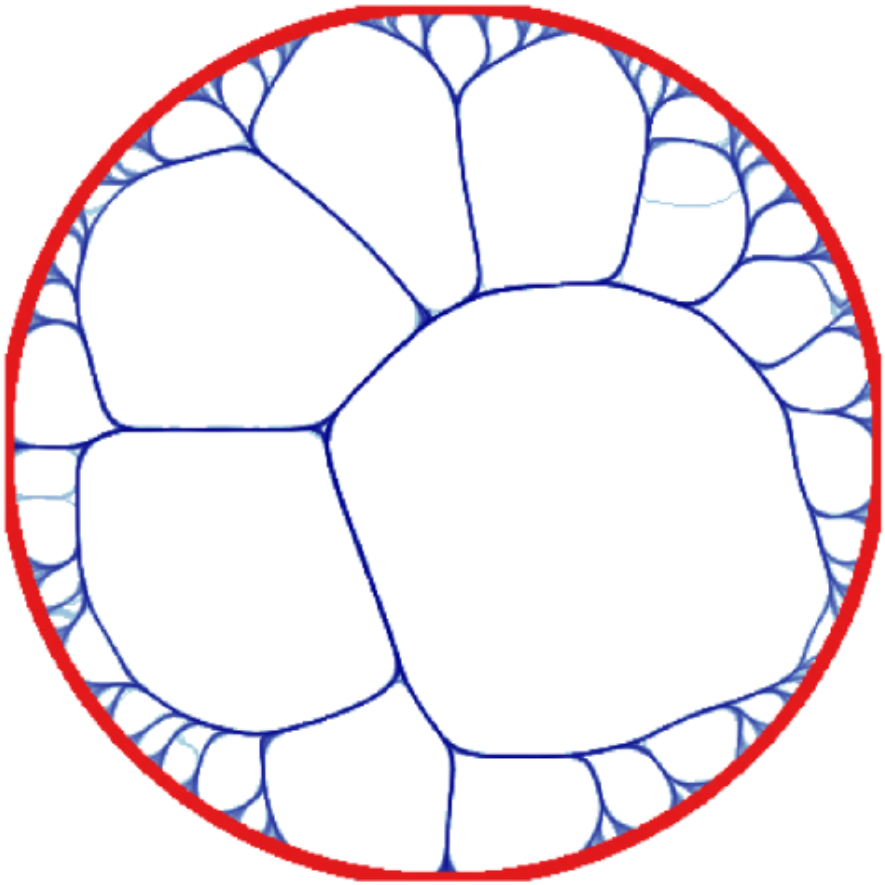


Figure: Stage 3

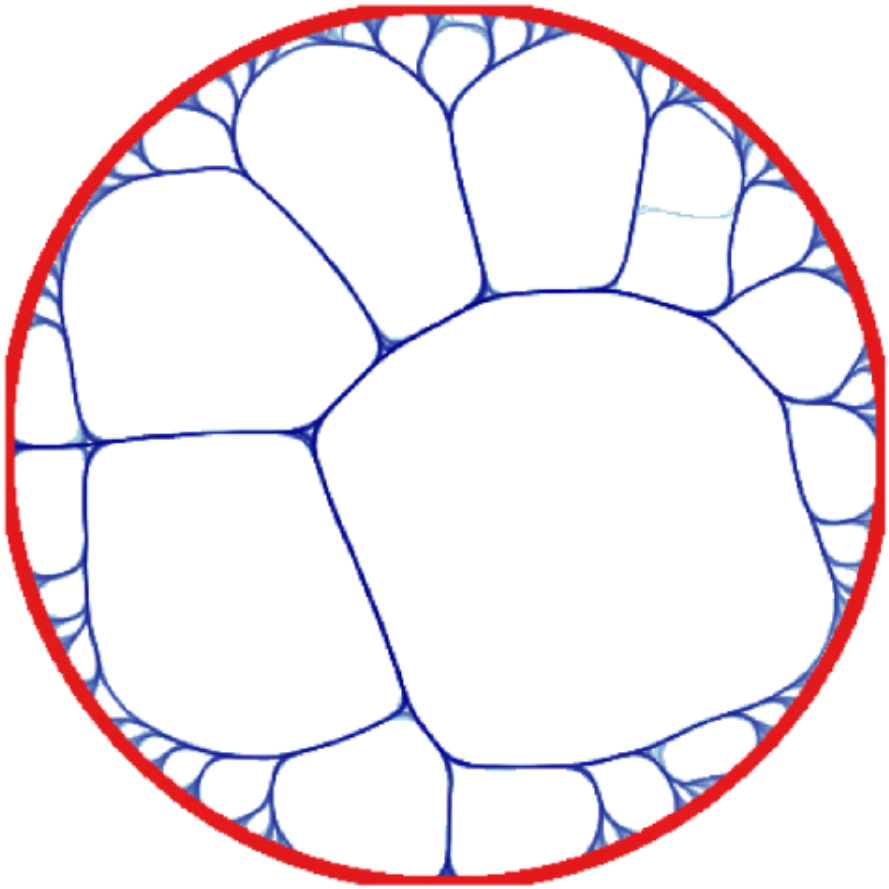


Figure: Stage 4

Results (cont'd)

Activated signalling pathways:

- Amphetamine addiction
- Bacterial invasion of epithelial cells
- Chemokine signaling pathway
- Complement and coagulation cascades
- Cytokine-cytokine receptor interaction
- Dilated cardiomyopathy
- ECM-receptor interaction
- Fanconi anemia pathway
- Focal adhesion
- HTLV-I infection
- Insulin signaling pathway
- Oocyte meiosis
- p53 signaling pathway
- Salmonella infection
- Serotonergic synapse

Results (cont'd)

Inhibited signalling pathways:

- Adipocytokine signaling pathway
- Alcoholism
- Amoebiasis
- Cell cycle
- Fc gamma R-mediated phagocytosis
- Focal adhesion
- HTLV-I infection
- Influenza A
- Malaria
- Measles
- Pancreatic cancer
- Pathways in cancer
- PPAR signaling pathway
- Progesterone-mediated oocyte maturation
- Systemic lupus erythematosus
- Tight junction

Results (cont'd)

Disease Enrichment on lncRNAs:

- DOID:2449 acromegaly and **GATA3-AS1**
- DOID:299 adenocarcinoma and **PVT1**
- DOID:3355 fibrosarcoma, DOID:8791 breast carcinoma in situ, DOID:8719 in situ carcinoma and **LINC00987**

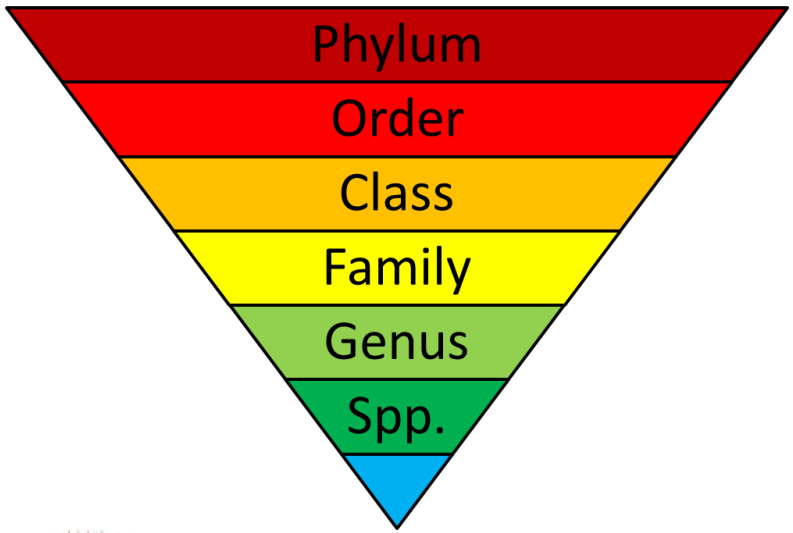
Results (cont'd)

Low degree miRNAs that make group difference:

- hsa-miR-129-5p
- hsa-miR-140-3p
- hsa-miR-146b-5p
- hsa-miR-188-5p
- hsa-miR-193a-5p
- hsa-miR-28
- hsa-miR-346
- hsa-miR-3605-3p
- hsa-miR-361
- hsa-miR-455-5p
- hsa-miR-671-3p
- hsa-miR-320b
- hsa-miR-193a-3p
- hsa-miR-326
- hsa-miR-330
- hsa-miR-501-3p

Strain Level Metagenomics Analysis

- Genetic content often varies even within a species
- PanPhlAn
  - Identifies which genes are present in the strains from your sample
- StrainPhlAn
  - Extracts species-specific marker genes from reads
  - Aligns the markers against reference genomes
  - Phylogenetic relatedness between strains from different samples



(a) E. coli O157:H7

(b) E. coli Nissle 1917

Data

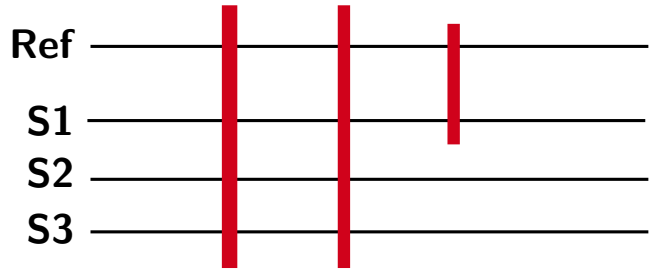
Sample reads are from Bovine milk exosomes

- Negative group:
  - 3 samples
  - ERS, exosome/RNA-sufficient: minimal essential media supplemented with the equivalent of exosomes from 0.5 L bovine milk distributed in the total intestinal water space in an adult, normalized for mice
- Positive group:
  - 3 samples
  - ERD, exosome/RNA-depleted: no exosomes added

Results

Identified species from reads:

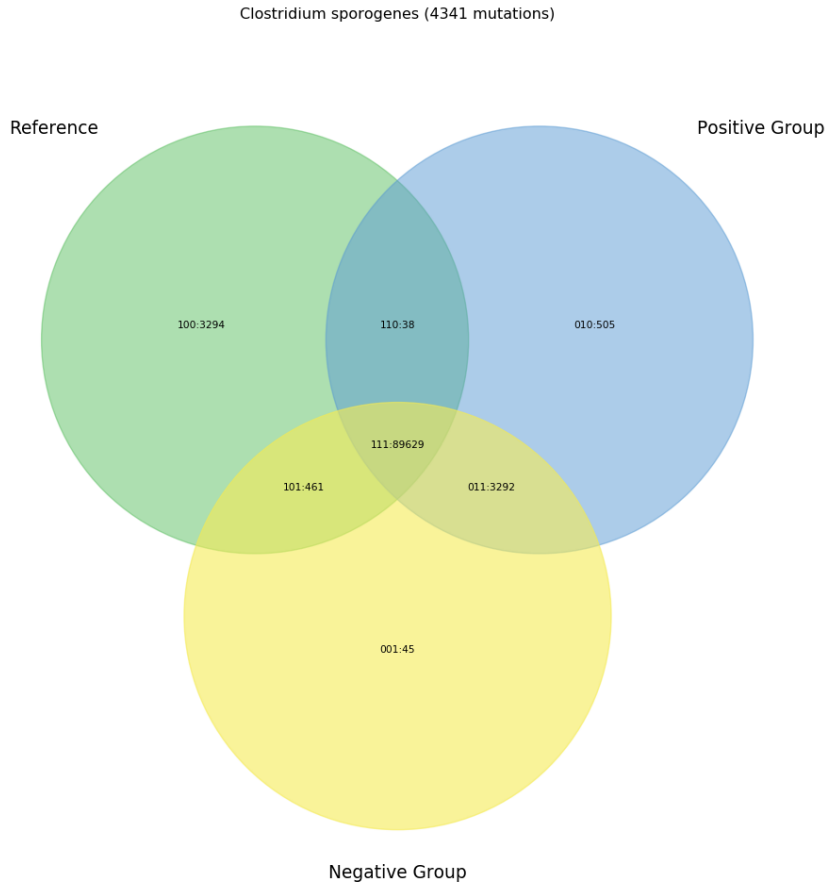
1. s__Aneurinibacillus_aneurinilyticus
2. s__Anaerotruncus_sp_G3_2012
3. s__Bacillus_cereus_thuringiensis
4. **s__Clostridium_sporogenes**
5. s__Desulfotomaculum_ruminis
6. **s__Enterobacteria_phage_lambda**
7. s__Enterobacteria_phage_phiX174_sensu_lato
8. **s__Enterococcus_faecalis**
9. s__Human_endogenous_retrovirus_K
10. **s__Lactobacillus_johnsonii**
11. s__Oscillibacter_sp_1_3
12. s__Saccharomyces_cerevisiae_killer_virus_M1



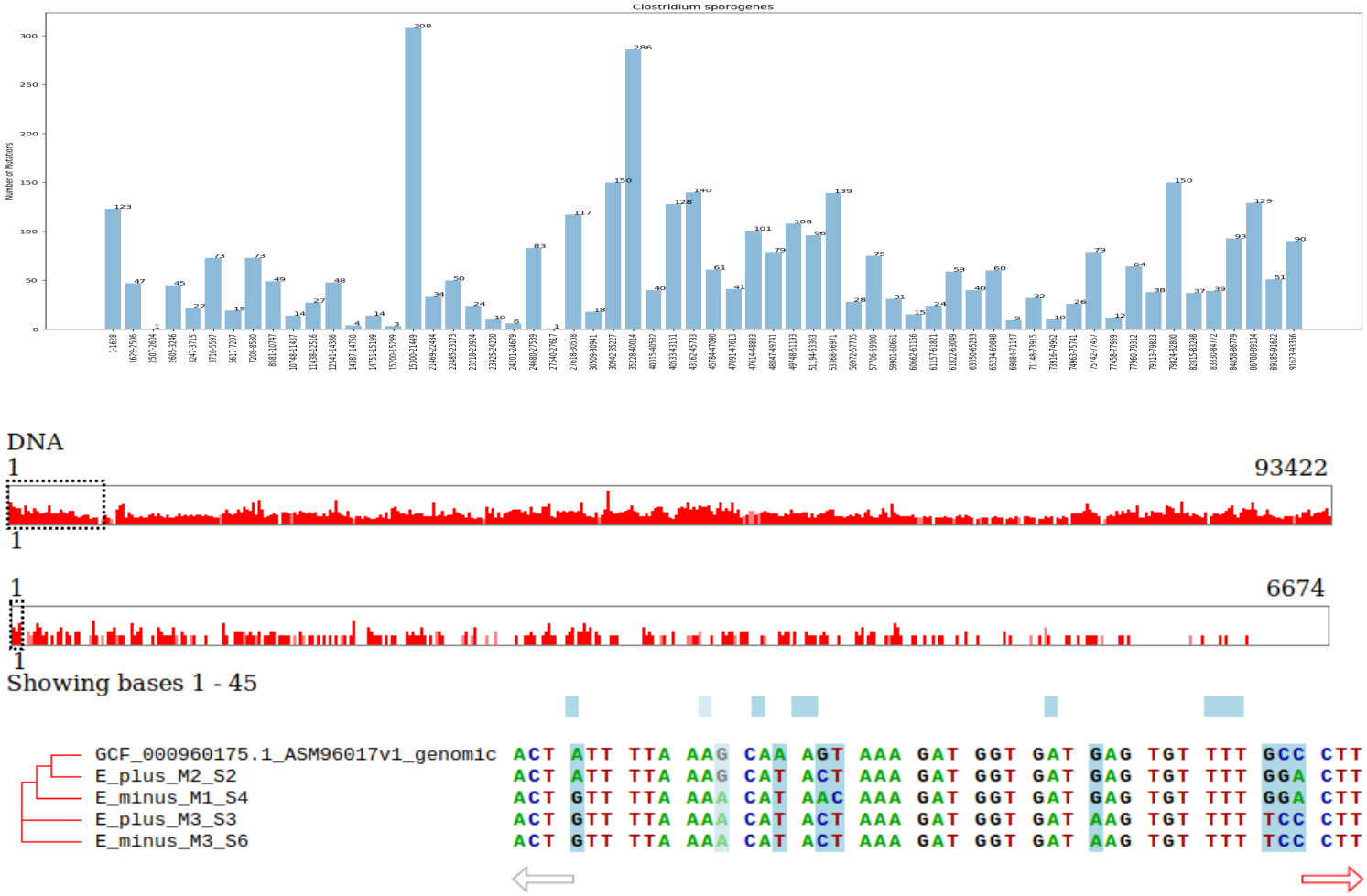
Results (cont'd)

s__Clostridium_sporogenes

- 110: 38/38 are in coding region
- 101: 459/461 are in coding region (*)
- 011: 3279/3292 are in coding region (*)
- 001: 44/45 are in coding region (*)
- 010: 502/505 are in coding region (*)



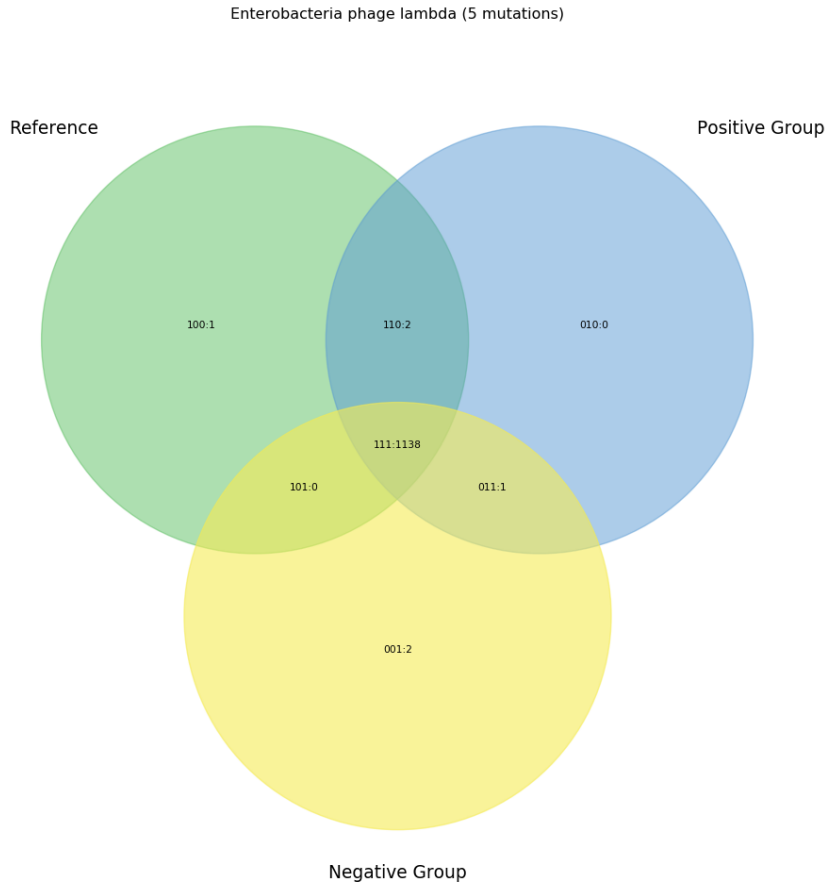
Results (cont'd)



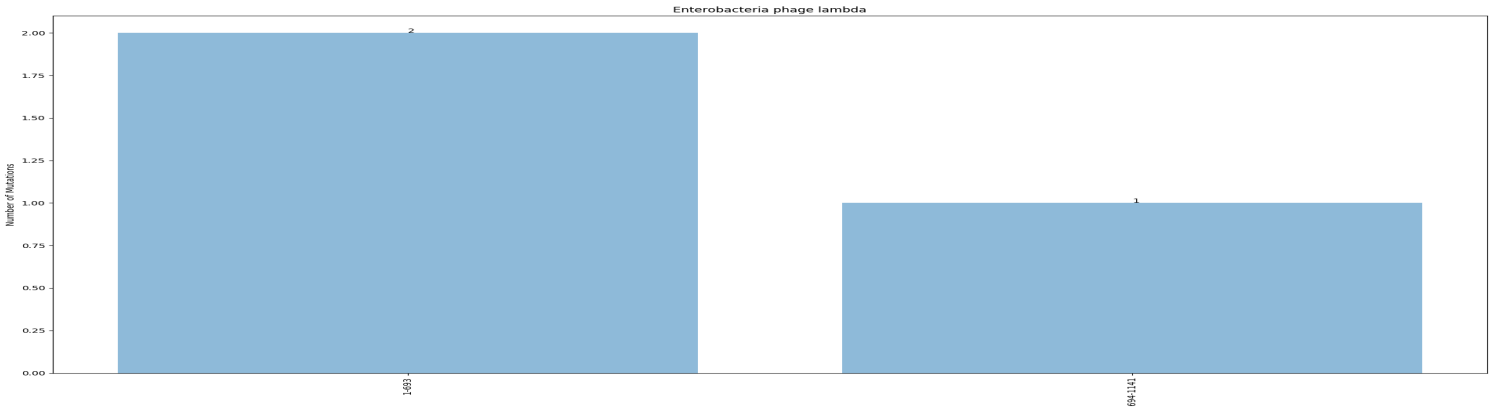
Results (cont'd)

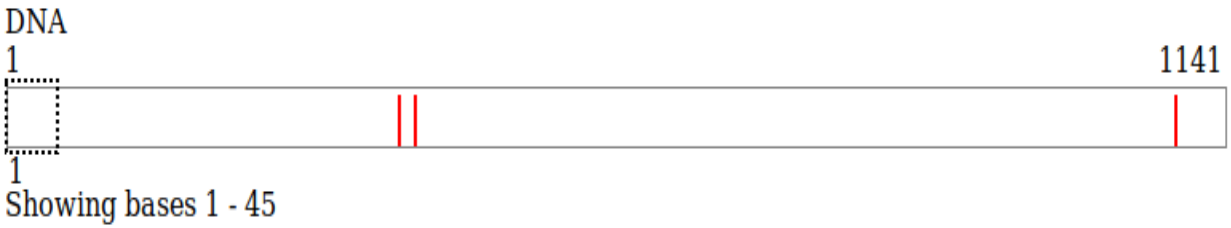
s__Enterobacteria_phage_lambda

- 110: 2/2 are in coding region
- 101: 0/0 are in coding region
- 011: 1/1 are in coding region
- 001: 2/2 are in coding region
- 010: 0/0 are in coding region



Results (cont'd)

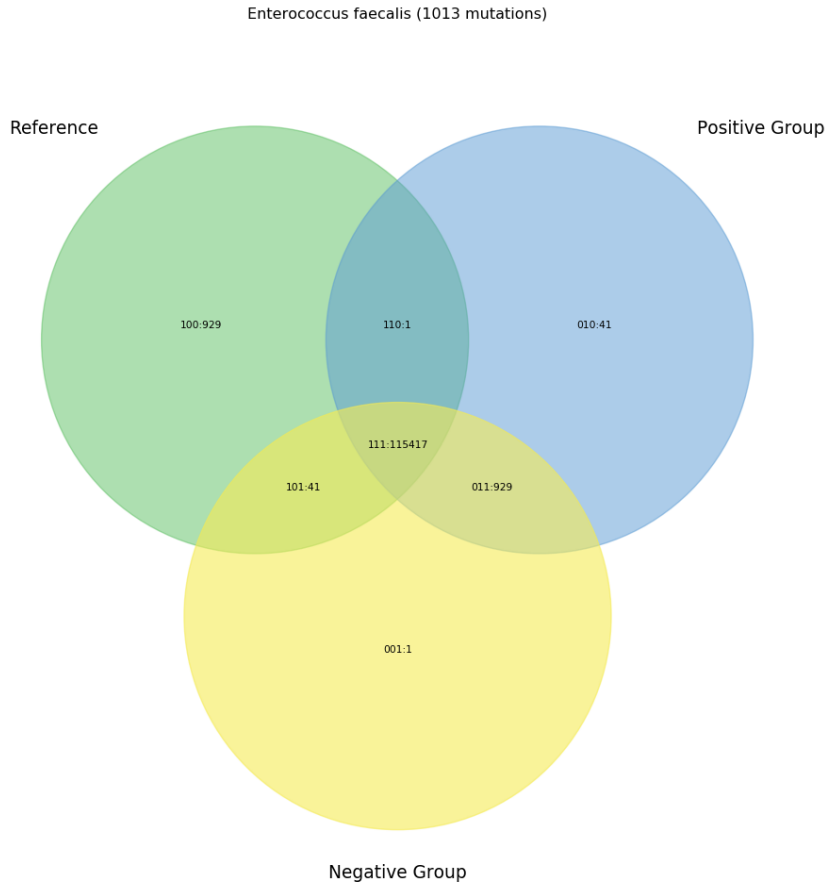




Results (cont'd)

s__Enterococcus_faecalis

- 110: 1/1 are in coding region
- 101: 41/41 are in coding region
- 011: 929/929 are in coding region
- 001: 1/1 are in coding region
- 010: 41/41 are in coding region

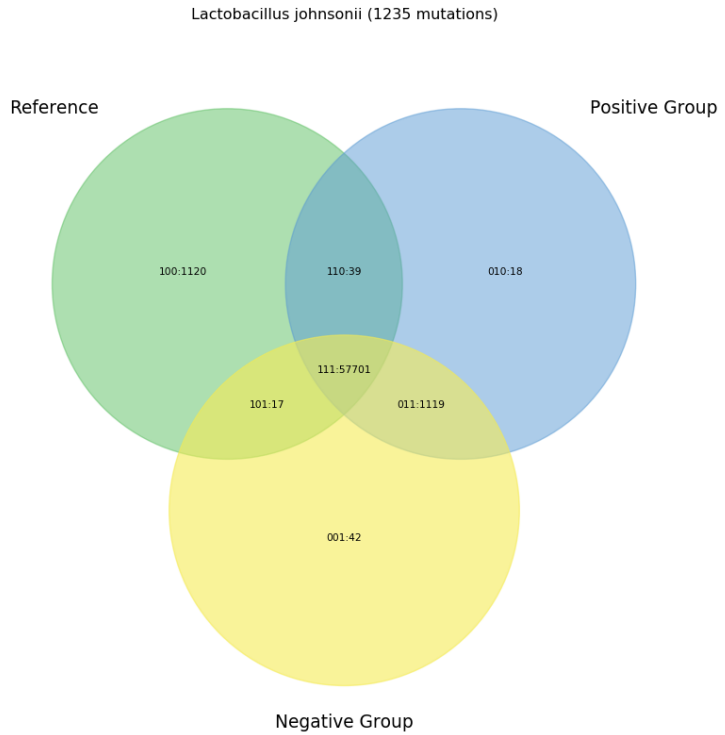


Results (cont'd)

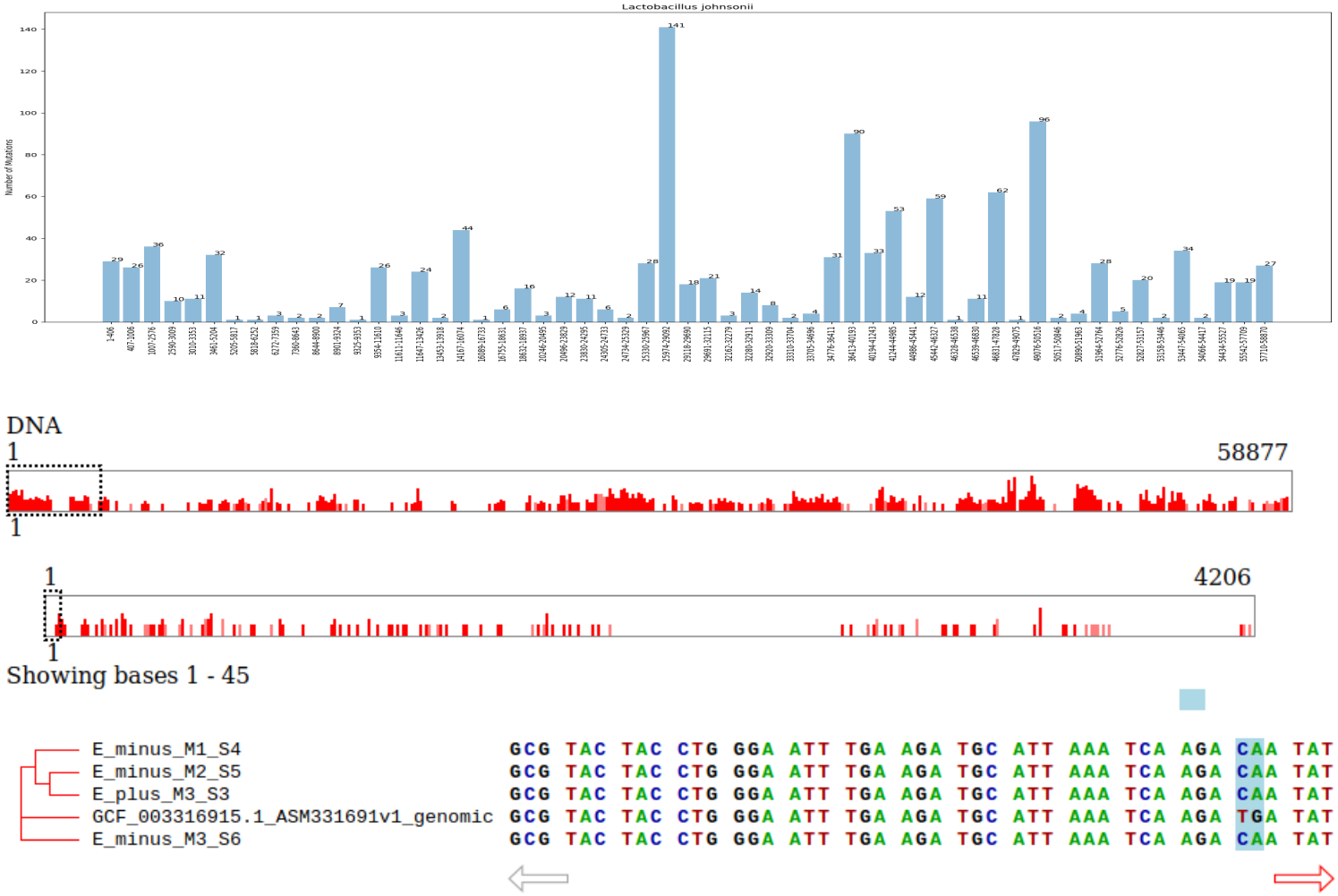
Results (cont'd)

s__Lactobacillus_johnsonii

- 110: 39/39 are in coding region
- 101: 17/17 are in coding region
- 011: 1113/1119 are in coding region (*)
- 001: 42/42 are in coding region
- 010: 18/18 are in coding region



Results (cont'd)



Questions

Questions?