Comparative Genomics

Background and Strategy

Team II: Fosfomycin Resistance

Tomáš Brůna, Harshini Chandrashekar, Ragy Haddad, William Harvey, Jane Lew, Sini Nagpal, Eunbi Park, Prachiti P. Prabhu, Vishnu Raghuram, Qi Zhang

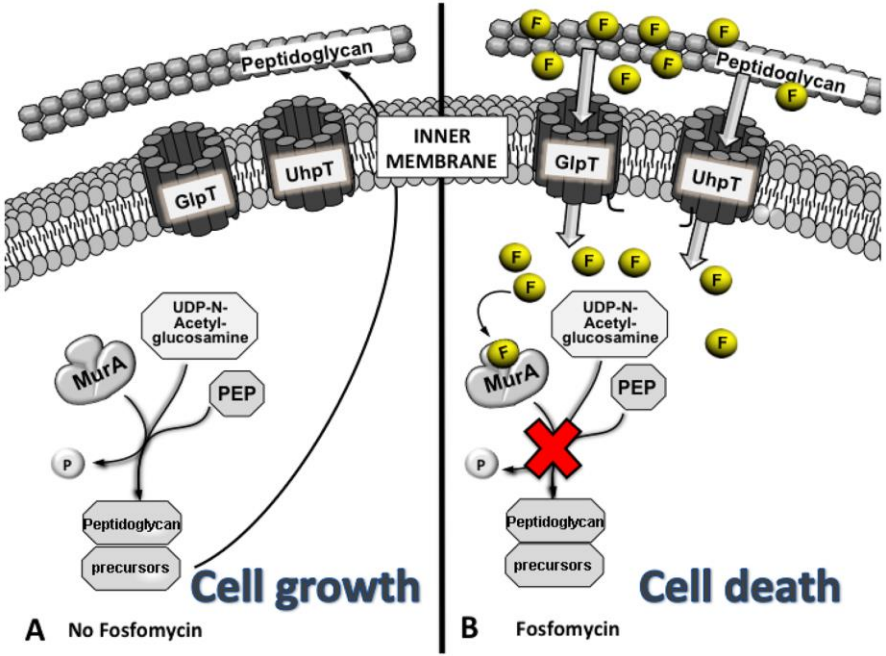
Introduction

- Comparative genomics is the study of comparing genome sequences to better understand the structure and function of genes.

- This field has explored areas ranging from organism development and behavior to metabolism and susceptibility to disease.

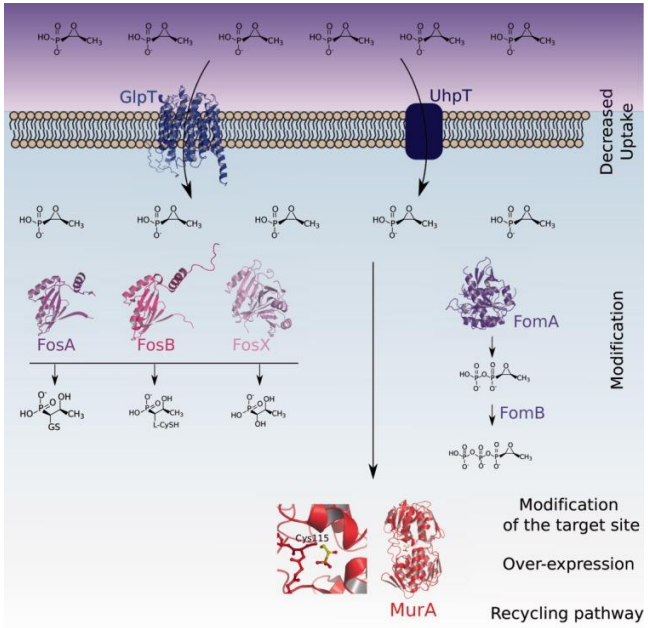
Fosfomycin - Mechanism of Action

- Bactericidal antibiotic
- Interferes with cell wall synthesis in both Gram-positive and Gram-negative bacteria
- Inhibits the initial step involving *phosphoenolpyruvate synthetase*
- Initially used to treat bladder infections



Fosfomycin - Resistance

- Reduced uptake
- Target site modification
- Expression of antibiotic-degrading enzymes
- Rescue of the UDP-MurNAc biogenesis pathway (ex. mutation within MurA enzyme)



Our Data

Resistant	27 (10%)
Heteroresistant	176 (68%)
Susceptible	9 (3%)
N/A	46 (18%)
Total	**258**

Objective

To identify genetic determinants that could be a potential cause for Fosfomycin heteroresistance in the isolates provided.

Methods

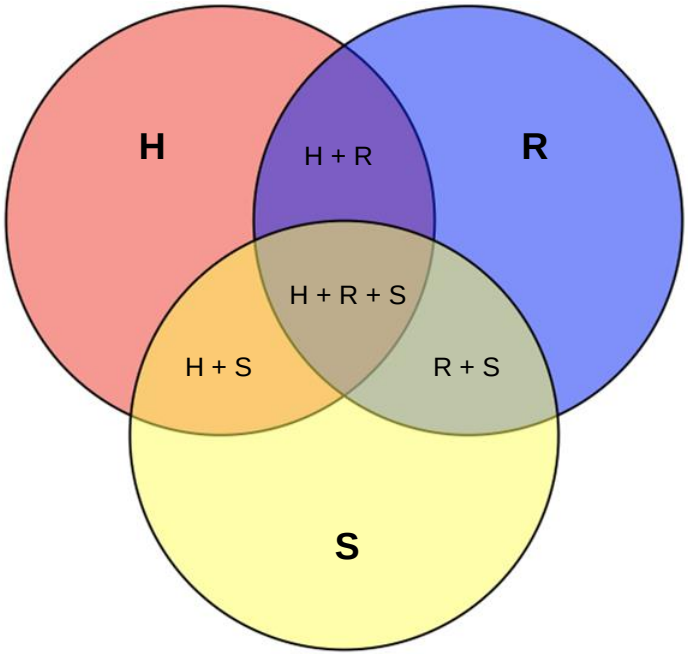
Whole Genome Approaches

Phylogeny Approaches

Whole-Genome Based Methods

Description, tools, strategy

What are we looking for?



Similarities

Identifying which of our samples are similar to each other would …

1. Help us choose representatives
2. Hint at phenotypic similarities
3. Tell us what not to care about

Differences

How are our samples different?

1. Is it an insertion/deletion/inversion /rearrangement?
2. Is it on a protein coding region/ncRNA /promoter?

Methods

- Clustering based on genome distance
- Pan Genome and core genome analysis
- SNV analysis

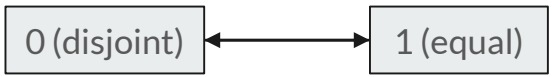
Distance computation - MinHash

To estimate the similarity between two sets based on Jaccard similarity coefficient

Jaccard similarity coefficient:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Indicator of the similarity between two sets



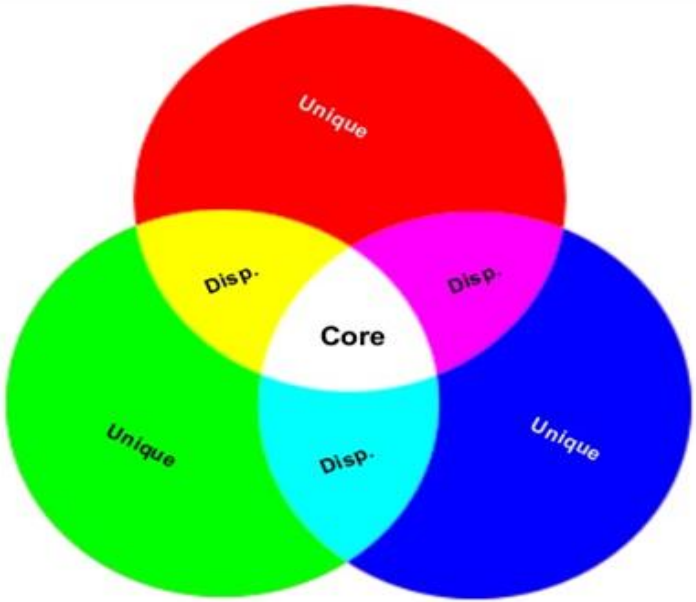
Distance computation - ANI

- Pairwise alignment of all input sequences
- Identifies matching regions
- Calculates average percentage identity of matching regions
- Accurate (>95% identity indicates species level similarity)
- Does not scale well (Quadratic number of pairwise comparisons)

Clustering

- Cluster based on distance matrix
- Hierarchical clustering
  - Informative dendrogram visualizations
  - Runtime: O(N^2 log(N)) where N is the number of samples => speed is not an issue
- We expect the main clustering to distinguish between strains
  - Focus on smaller differences = cluster the main clusters again
- Choose number of clusters which best matches antibiotic resistance labels -- maximize mutual information

Pan and core genome analysis



Three groups - Resistant, Heteroresistant and Susceptible.

Core genome: Intersect of all identified genes across 3 groups

Pan genome: Union of all identified genes across 3 groups

Variable genome: Pan genome - Core genome

Pan and core genome analysis - tools

- **ROARY**
  - Fast
  - Scales Well
  - Compatible with Prokka output
  - Takes order of genes into account
  - Generates core genome, pan genome and accessory genome
  - Several tools for visualization built in

Pan and core genome analysis

- **BPGA** - Bacterial Pan Genome Analysis Tool
  - Input processing is similar to ROARY (Clusters before pairwise blast) - also fast and scales well
  - Has a large suite of pan genome analysis tools
  - Also does functional analysis (KEGG)

Tool Summary

- Distance computation:
  - ANI, MinHash
- Clustering: Hierarchical
  - R Hclust package, MoCham
- Pan and core genome analysis:
  - ROARY
  - BPGA

Phylogeny Based Methods

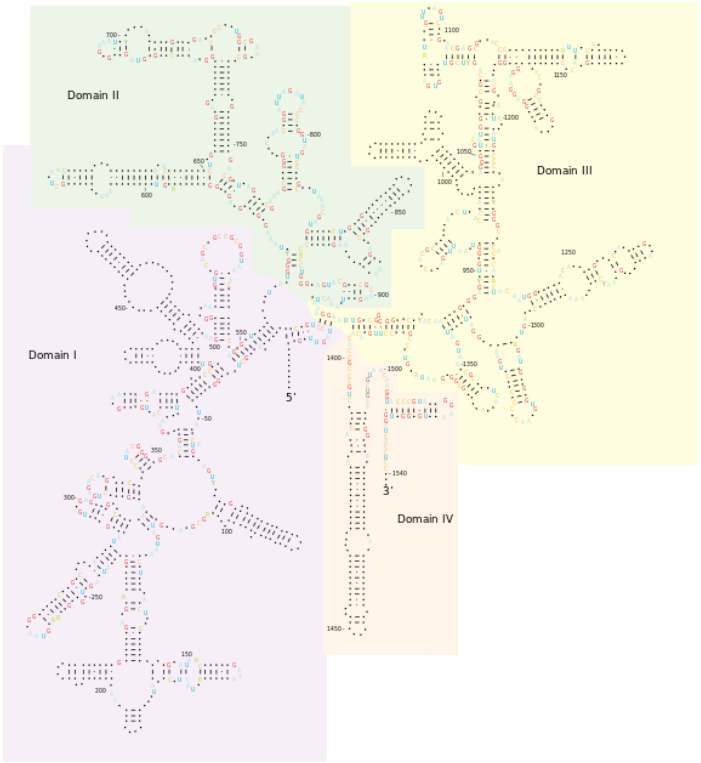
Description, Tools and Strategy

Phylogeny Based Methods

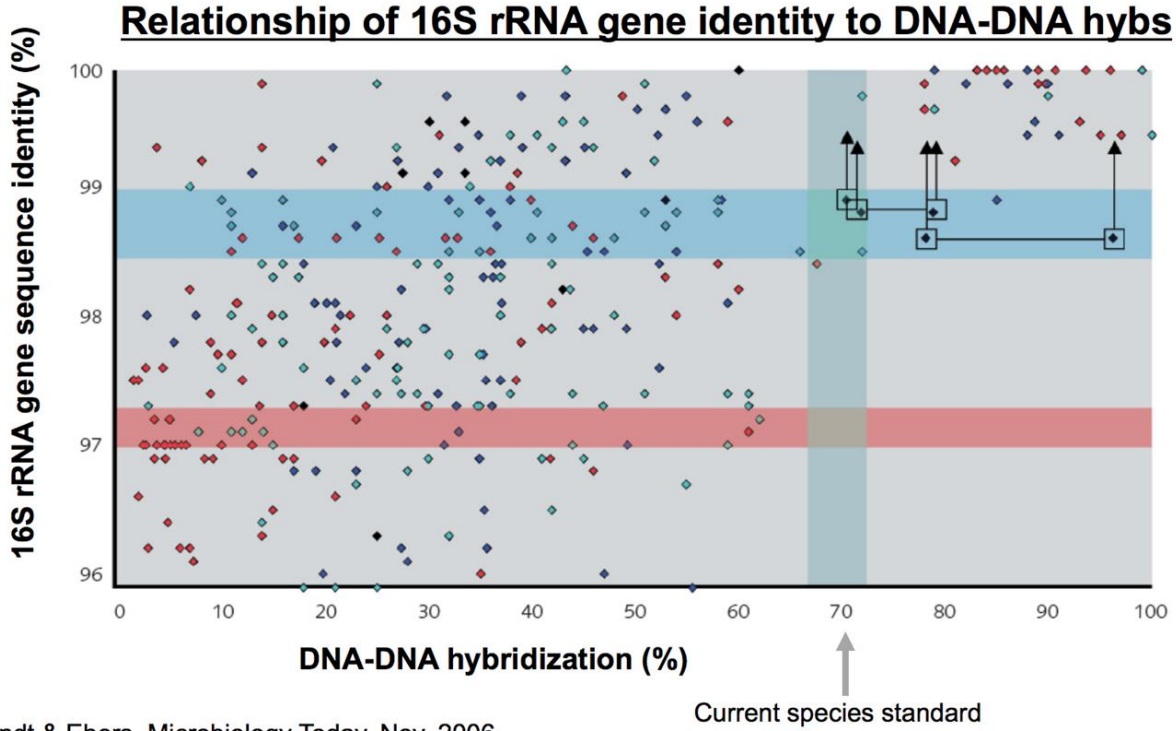
- 16S Subunit Analysis
- SNP Trees
- MLST-based Methods

16S rRNA

- Highly conserved region of the 30S subunit of prokaryotic ribosomes
- Methodology
  - Blast/Align 16S coding regions
- Benefits
  - Easy to sequence
  - Highly conserved
- Drawbacks
  - Inconclusive
  - Not functionally illuminating



16S Drawbacks



Stackebrandt & Ebers, Microbiology Today, Nov. 2006

Single Nucleotide Polymorphism Trees

- Methodology
  - Compare SNPs to identify regions which can be implicated in typing
- Benefits
  - Broad spectrum
  - Thorough
  - Illuminate indirect antibiotic resistance factors
- Drawbacks
  - Resource intensive
  - SNPs might not indicate change in functionality
  - Insensitive to large changes in the genome

Tool: kSNP3.0

- Alignment-free method
- Breaks sequences in to non-redundant k-mers
- Establishes trees and annotates sequences if desired

kSNP3.0: SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome

Shea N. Gardner[1], Tom Slezak[1] and Barry G. Hall[2,*]

[1]Computations/Global Security, Lawrence Livermore National Laboratory, Livermore, CA 94550, USA and [2]Bellingham Research Institute, Bellingham, WA 98229, USA

Table 1 Times required for kSNP analysis of 20 *E.coli* genomes

Program	Conditions	Time (h)
kSNP v2	Default (no annotation)	1.04
kSNP3.0	Default (no annotation)	0.89
kSNP v2	Annotation	11.04
kSNP3.0	Standard annotation	2.92
kSNP3.0	Full annotation	11.14

Multi Locus Sequence Typing (MLST)

- Methodology
  - Establish MLST Scheme
  - Sort in to allelic groups
- Benefits
  - Basic MLST schemes pre-established
  - Flexible
  - Does not require alignment
- Drawbacks
  - Current schemes might not be sensitive enough

MLST Scheme Explained

- Traditionally composed of a set of housekeeping genes
- Typing scheme for *Klebsiella pneumoniae* available via PubMLST
  - gapA
  - infB
  - Mdh
  - Pgi
  - phoE
  - rpoB
  - tonB
- Why use housekeeping genes?

STing

- Discards Irrelevant Reads Based Off MLST Scheme
- Picks Representative Sequence from Allele
- Assigns Reads to Allele Profiles

Georgia Tech Bioinformatics
Interdisciplinary Graduate Program

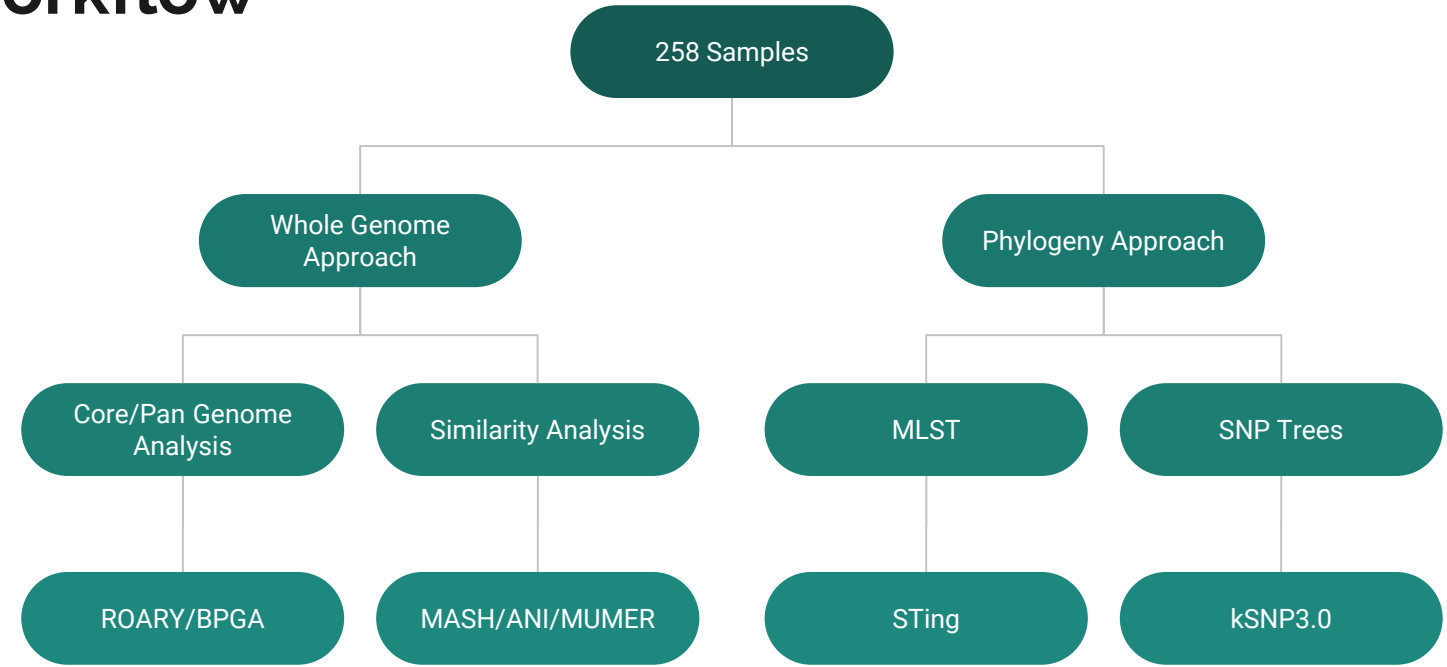
ABiL
Applied BioInformatics Laboratory

PABI
PanAmerican Bioinformatics Institute

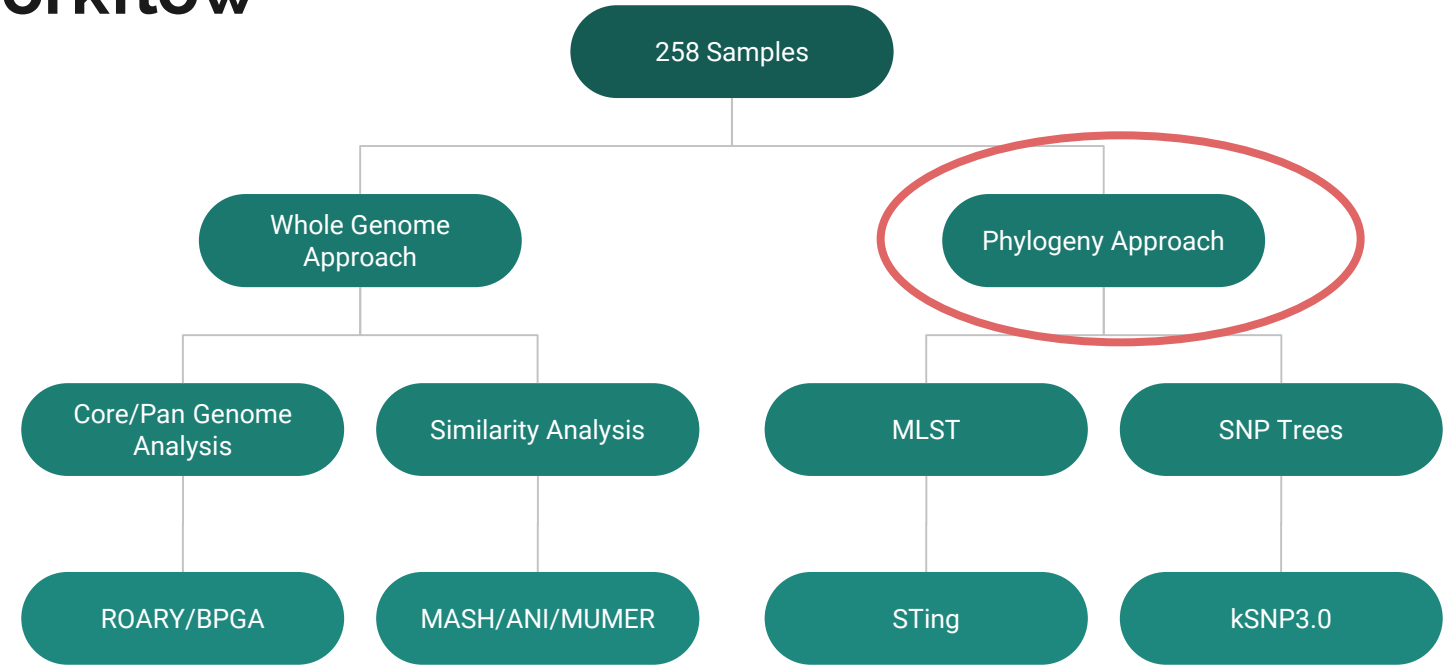
Proposed Strategy

Workflow Diagram

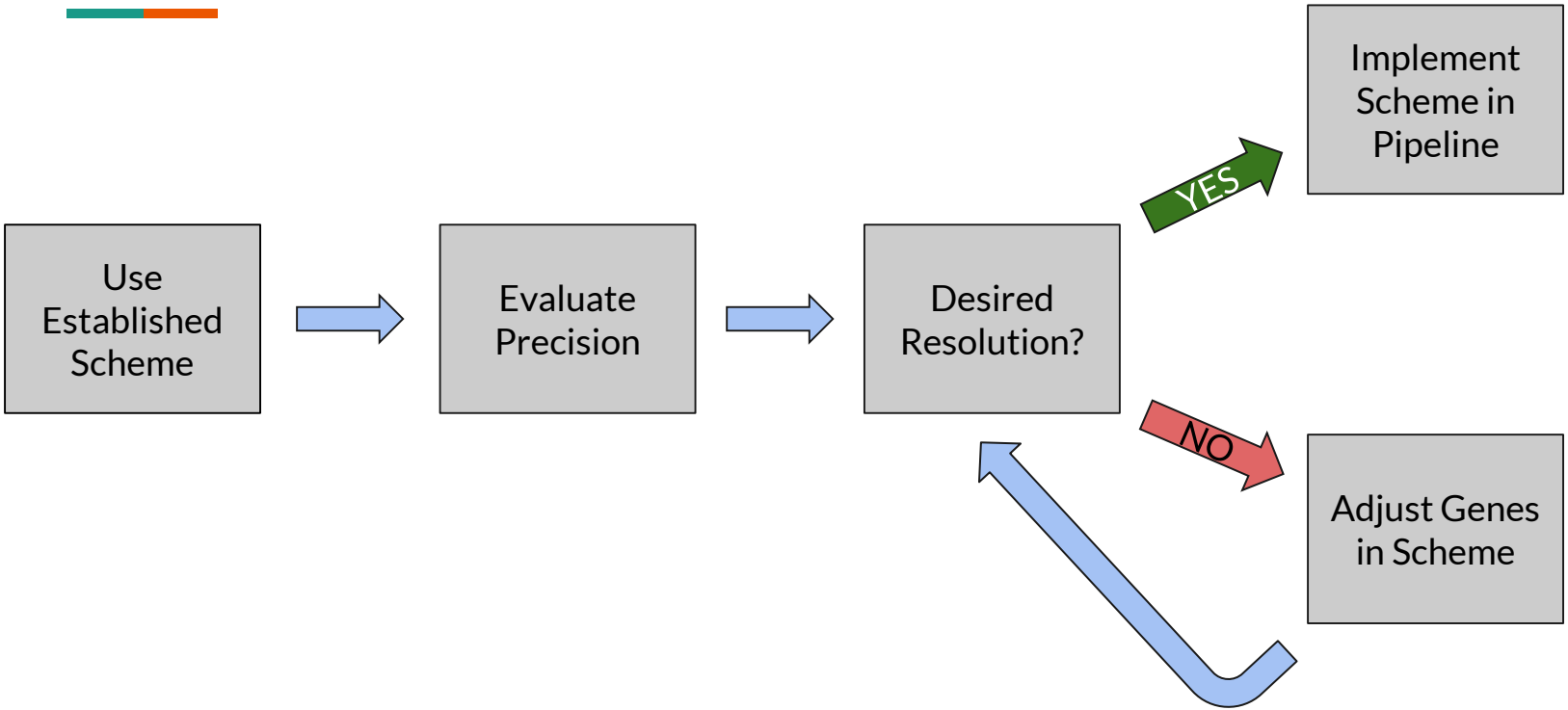
Workflow


258 Samples
Whole Genome Approach
Phylogeny Approach
Core/Pan Genome Analysis
Similarity Analysis
MLST
SNP Trees
ROARY/BPGA
MASH/ANI/MUMER
STing
kSNP3.0

Workflow


258 Samples
Whole Genome Approach
Phylogeny Approach
Core/Pan Genome Analysis
Similarity Analysis
MLST
SNP Trees
ROARY/BPGA
MASH/ANI/MUMER
STing
kSNP3.0

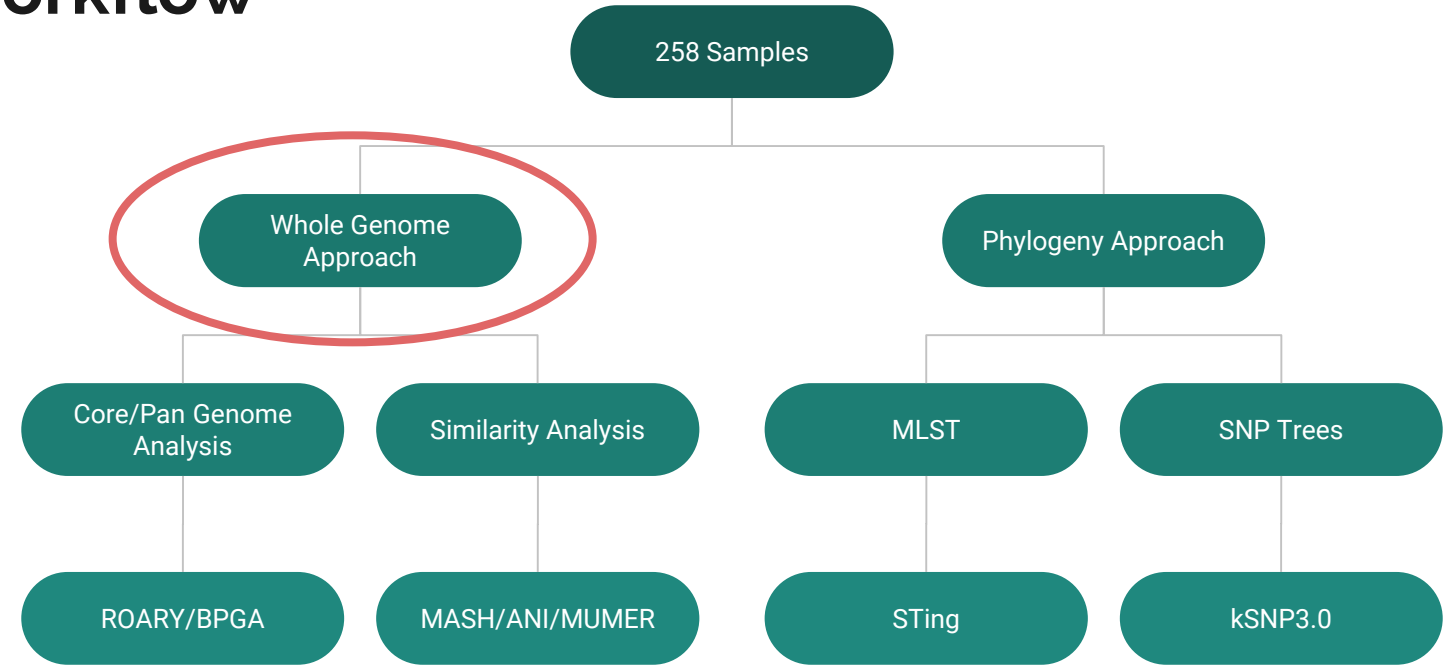
MLST Workflow



Improving MLST Precision

- Add antibiotic resistance genes
  - fosA6
  - aph6id
  - mgrB
- Compile alleles from functional annotation
  - To Be Determined

Workflow

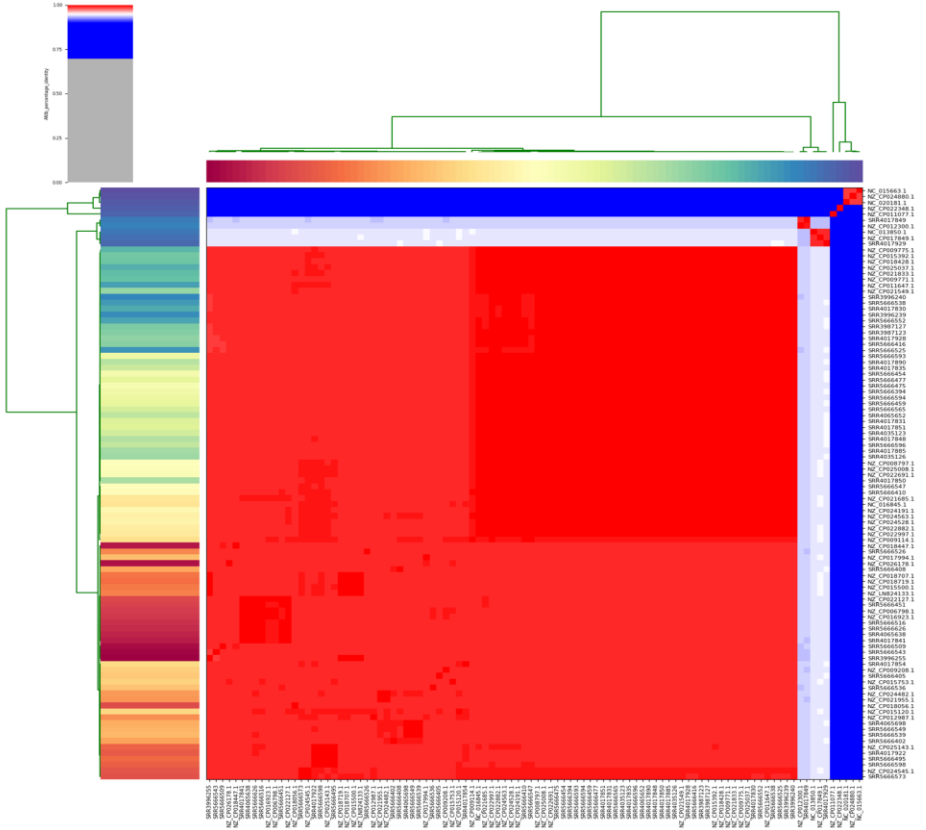

258 Samples
Whole Genome Approach
Phylogeny Approach
Core/Pan Genome Analysis
Similarity Analysis
MLST
SNP Trees
ROARY/BPGA
MASH/ANI/MUMER
STing
kSNP3.0

Similarity Analysis

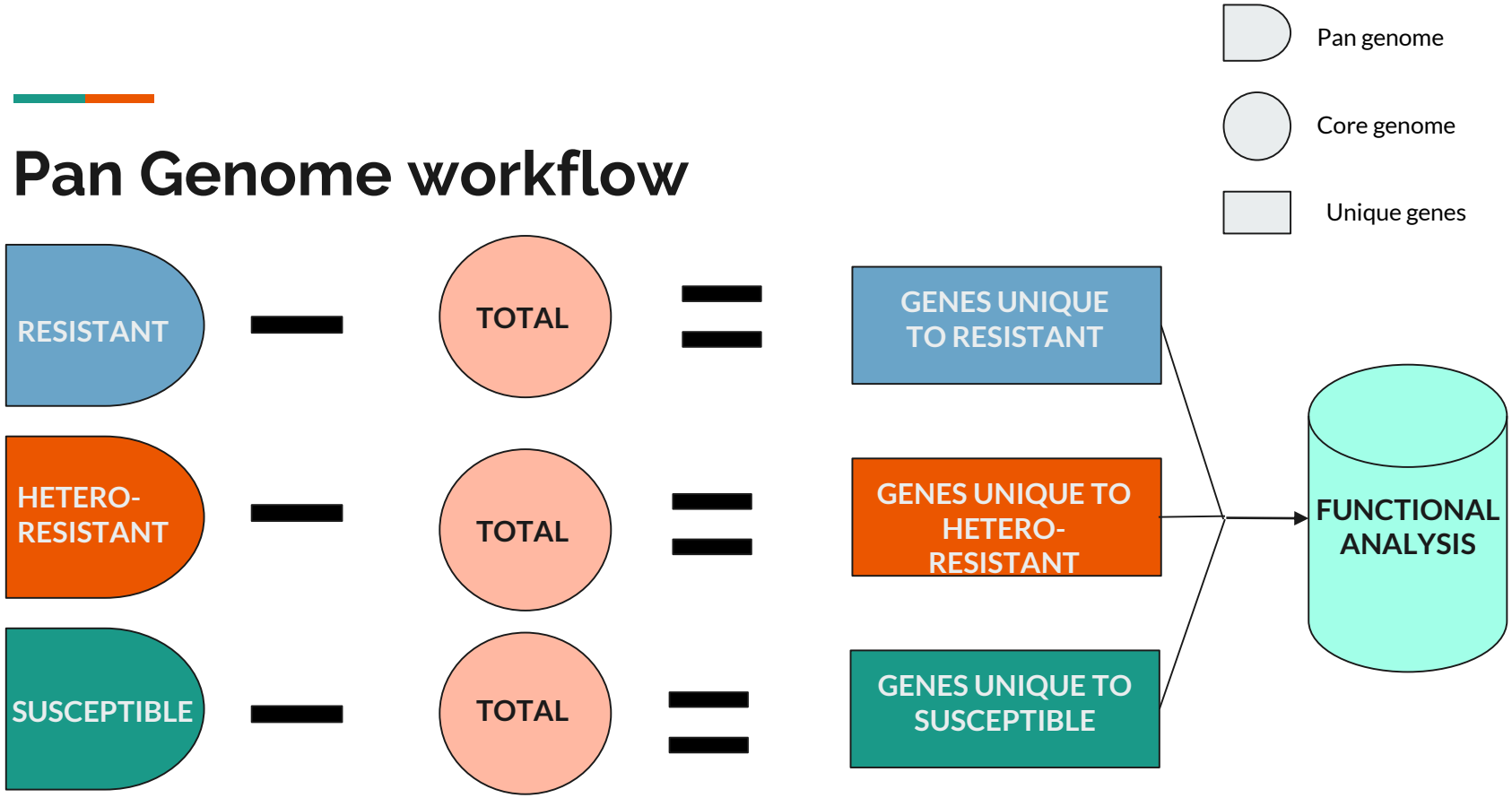
- Gives species information - but we already know this

For higher resolution

- Set a higher threshold
- *Might* cluster based on Ab phenotype (mostly wont)
- Will still be informative



Pan Genome workflow



Pan genome workflow - concerns

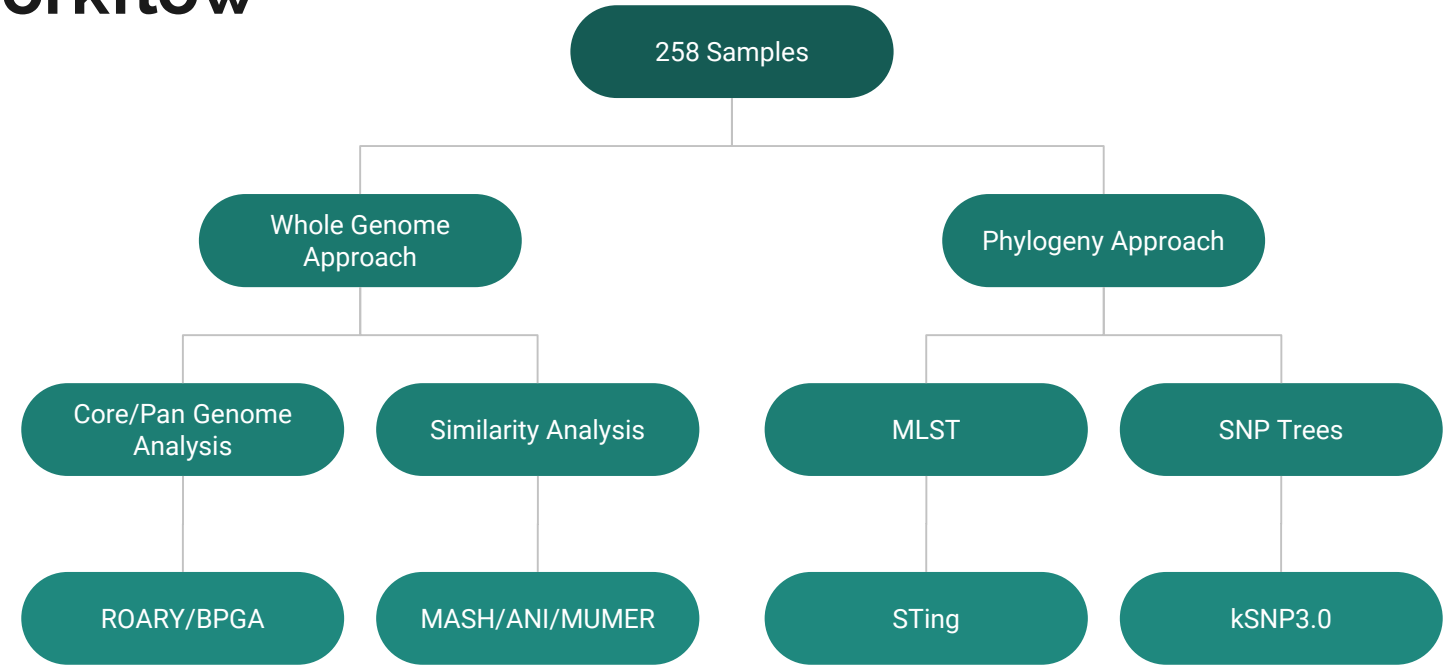
1. **We lose SNV information:** Subtracting core genome from each group means we would lose any potentially interesting single nucleotide variants that exist in the core genome of some samples. **Potential fix: Construct “consensus” genomes for each group and look for SNVs**.

1. **Multiple Alleles:** Multiple copies of the same gene might not be preserved while building pan-genomes.

   Potential fix: ?

However the MLST based methods would identify both SNPs and multiple alleles.

Workflow


258 Samples
Whole Genome Approach
Phylogeny Approach
Core/Pan Genome Analysis
Similarity Analysis
MLST
SNP Trees
ROARY/BPGA
MASH/ANI/MUMER
STing
kSNP3.0

References

1. Castañeda-García, Alfredo, Jesús Blázquez, and Alexandro Rodríguez-Rojas. "Molecular mechanisms and clinical impact of acquired and intrinsic fosfomycin resistance." *Antibiotics* 2.2 (2013): 217-236.
2. Nikolaidis I, Favini-Stabile S, Dessen A. 2014. Resistance to antibiotics targeted to the bacterial cell wall. *Protein Sci* 23: 243–259.
3. Kidd, Timothy J et al. “A *Klebsiella Pneumoniae* Antibiotic Resistance Mechanism That Subdues Host Defences and Promotes Virulence.” *EMBO Molecular Medicine* 9.4 (2017): 430–447.
4. Guo, Qinglan et al. “Glutathione-*S*-Transferase FosA6 of *Klebsiella Pneumoniae* Origin Conferring Fosfomycin Resistance in ESBL-Producing *Escherichia Coli*.” *Journal of Antimicrobial Chemotherapy* 71.9 (2016): 2460–2465.
5. Gardner, Shea N., Tom Slezak, and Barry G. Hall. "kSNP3. 0: SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome." *Bioinformatics*31.17 (2015): 2877-2878.
6. Shea N Gardner, Tom Slezak, Barry G. Hall; kSNP3.0: SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome, *Bioinformatics*, Volume 31, Issue 17, 1 September 2015, Pages 2877–2878.
7. https://doi.org/10.1093/bioinformatics/btv271
8. Kim, Mincheol, et al. "Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes." *International journal of systematic and evolutionary microbiology* 64.2 (2014): 346-351.