Comparative Genomics

Final Results

Team 1

Team Members:
Frank Ambrosio, Vasanta Chivukula, Seonggeon Cho, Siarhei Hladyshau, Junyu Li, Yiqiuyi Liu, , Yihao Ou, Hunter Seabolt, Qinyu Yue

Outline

Content

- Objectives
- Initial Analysis (MASH)
- SNP Analysis
- Roary/Scoary
- Bacterial GWAS
- MLST
- Conclusions

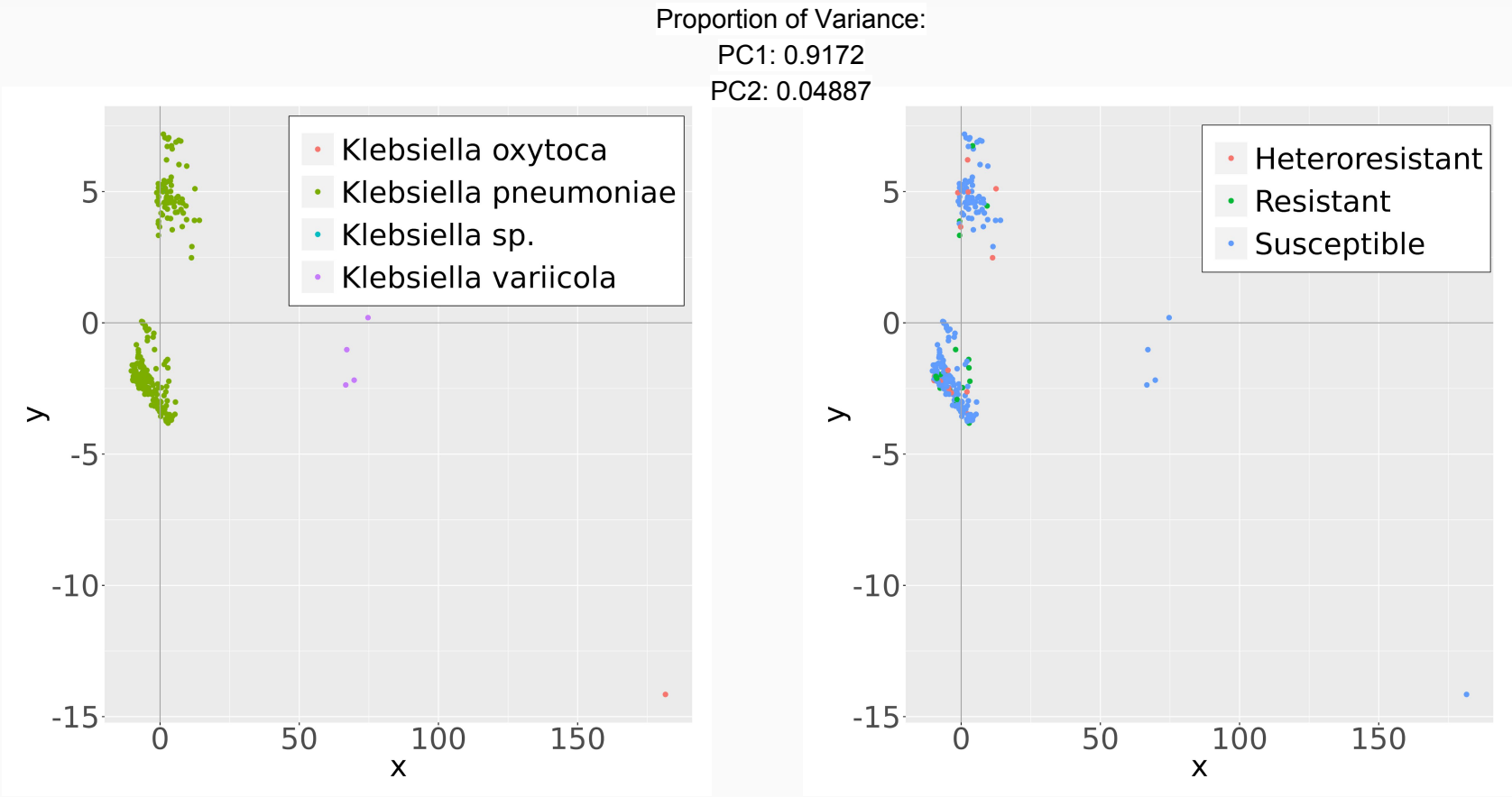
Objectives

Explore gene features in Klebsiella that confer colistin resistance. Looking for fixed genomic differences indicating a “shared” ancestry between groups.

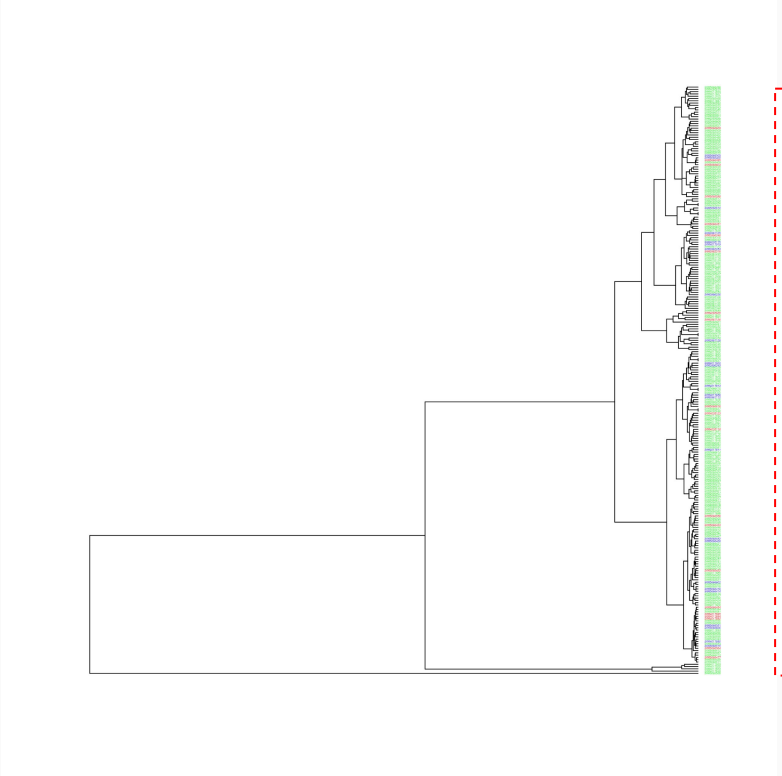
Determine if it is possible to...
Predict colistin susceptibility of other *Klebsiella* spp. strains
...using only Illumina sequencing reads

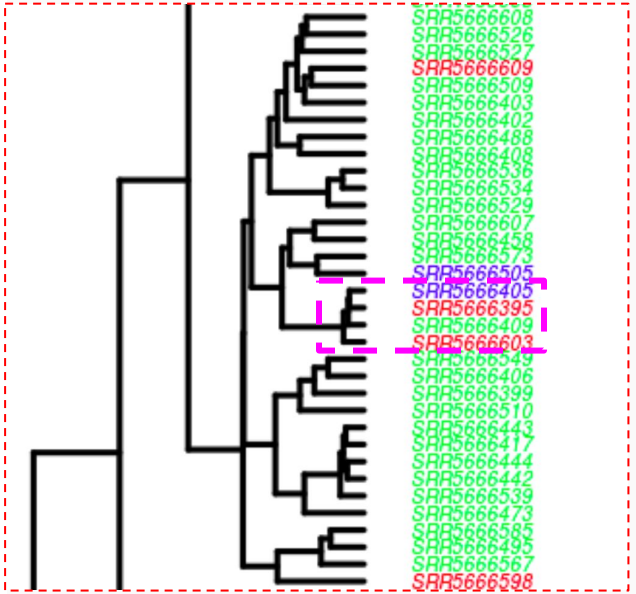
“Culture-independent Antimicrobial Resistance profiling”

MASH PCA


Proportion of Variance:
PC1: 0.9172
PC2: 0.04887
Klebsiella oxytoca
Klebsiella pneumoniae
Klebsiella sp.
Klebsiella variicola
Heteroresistant
Resistant
Susceptible
y
x
5
0
-5
-10
-15
0
50
100
150

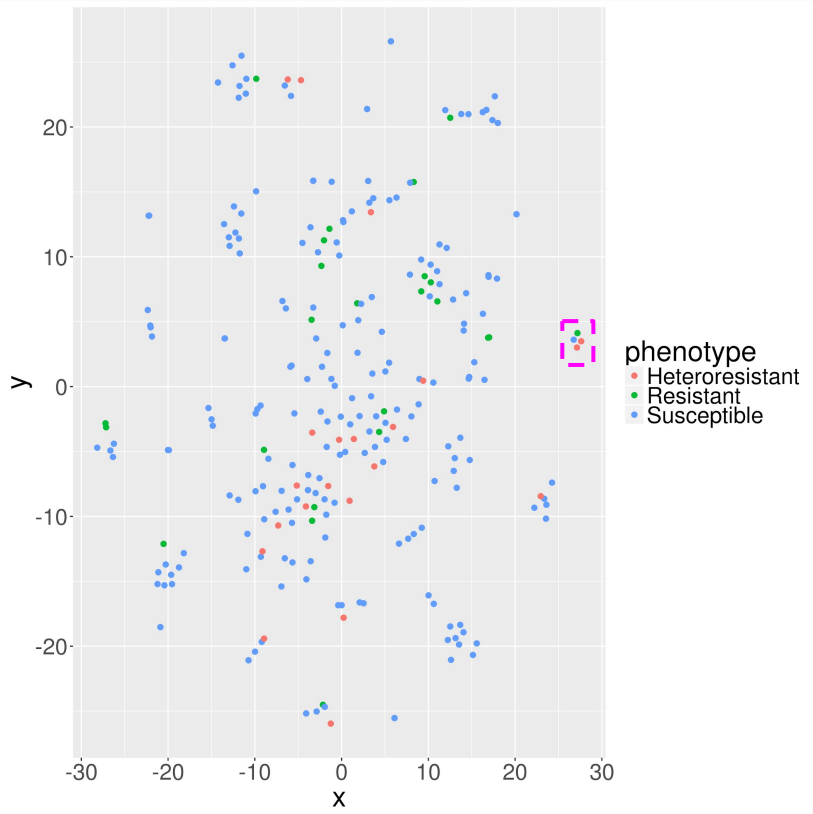
MASH hierarchical clustering

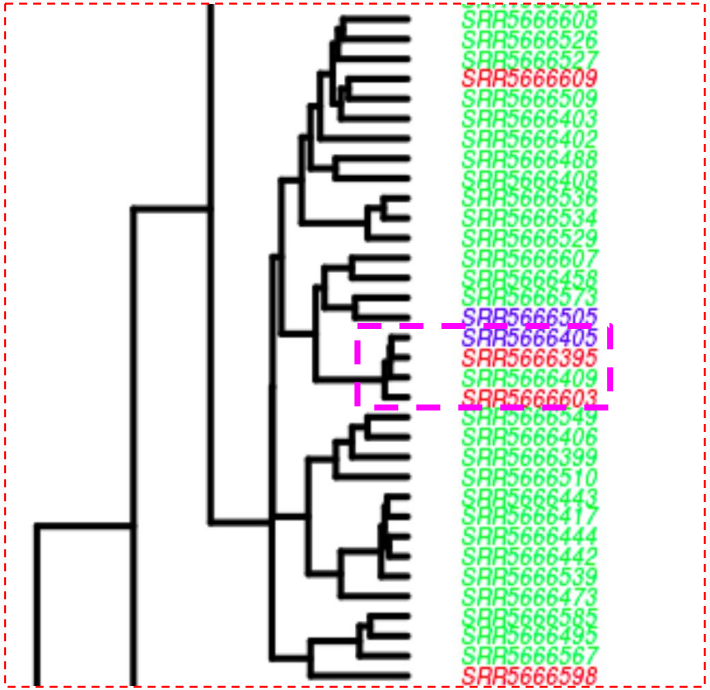




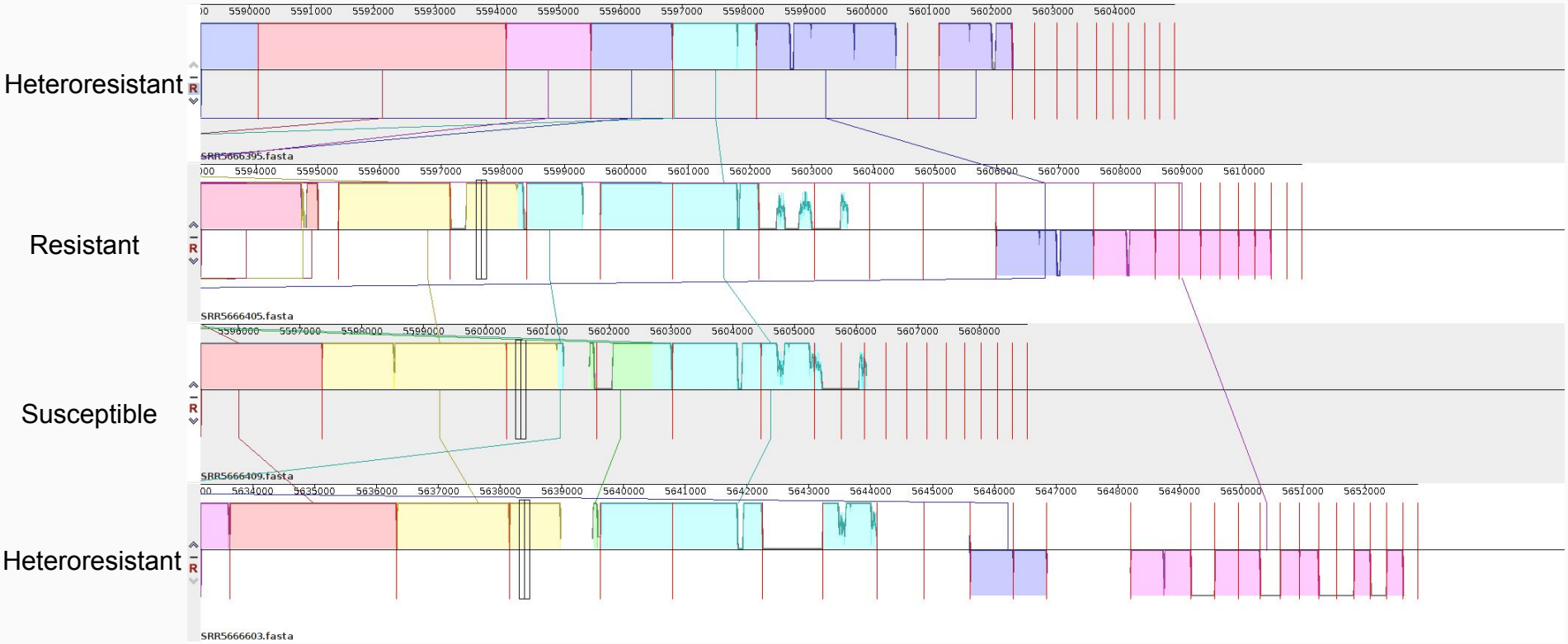
MASH -tSNE

Clustering

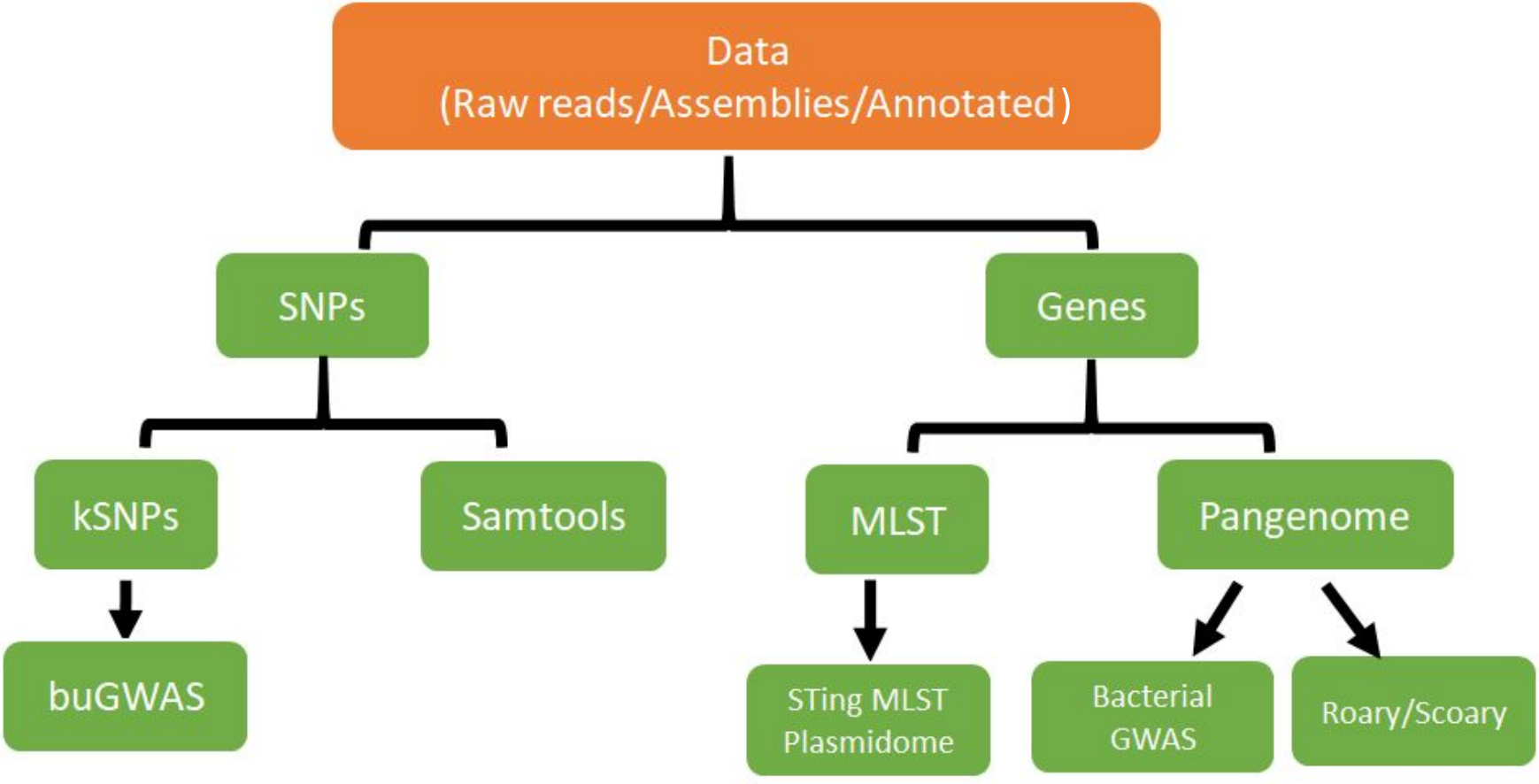




Multiple alignment of 4 genomes

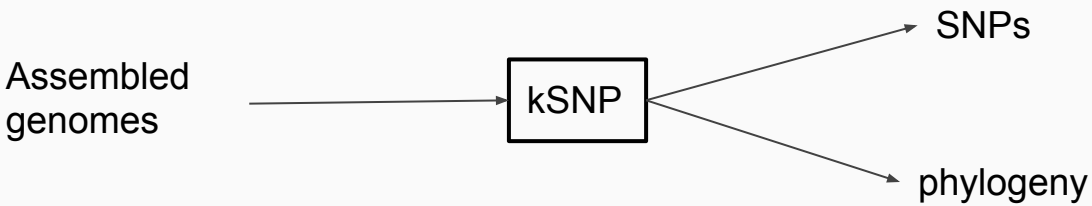


Pipeline

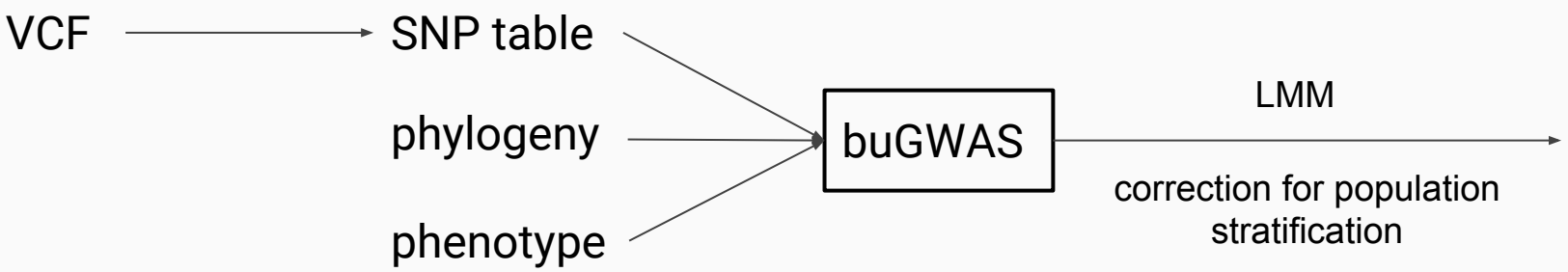

Data
(Raw reads/Assemblies/Annotated)
SNPs
Genes
kSNPs
Samtools
MLST
Pangenome
buGWAS
STing MLST
Plasmidome
Bacterial
GWAS
Roary/Scoary

kSNP

MakeFasta input_list.txt sh.fasta
Kchooser sh.fasta
kSNP3 -in input_list.txt -outdir final_selected_genomes -annotate annotated_list.txt -k 21 -vcf -ML | tee log.txt



buGWAS



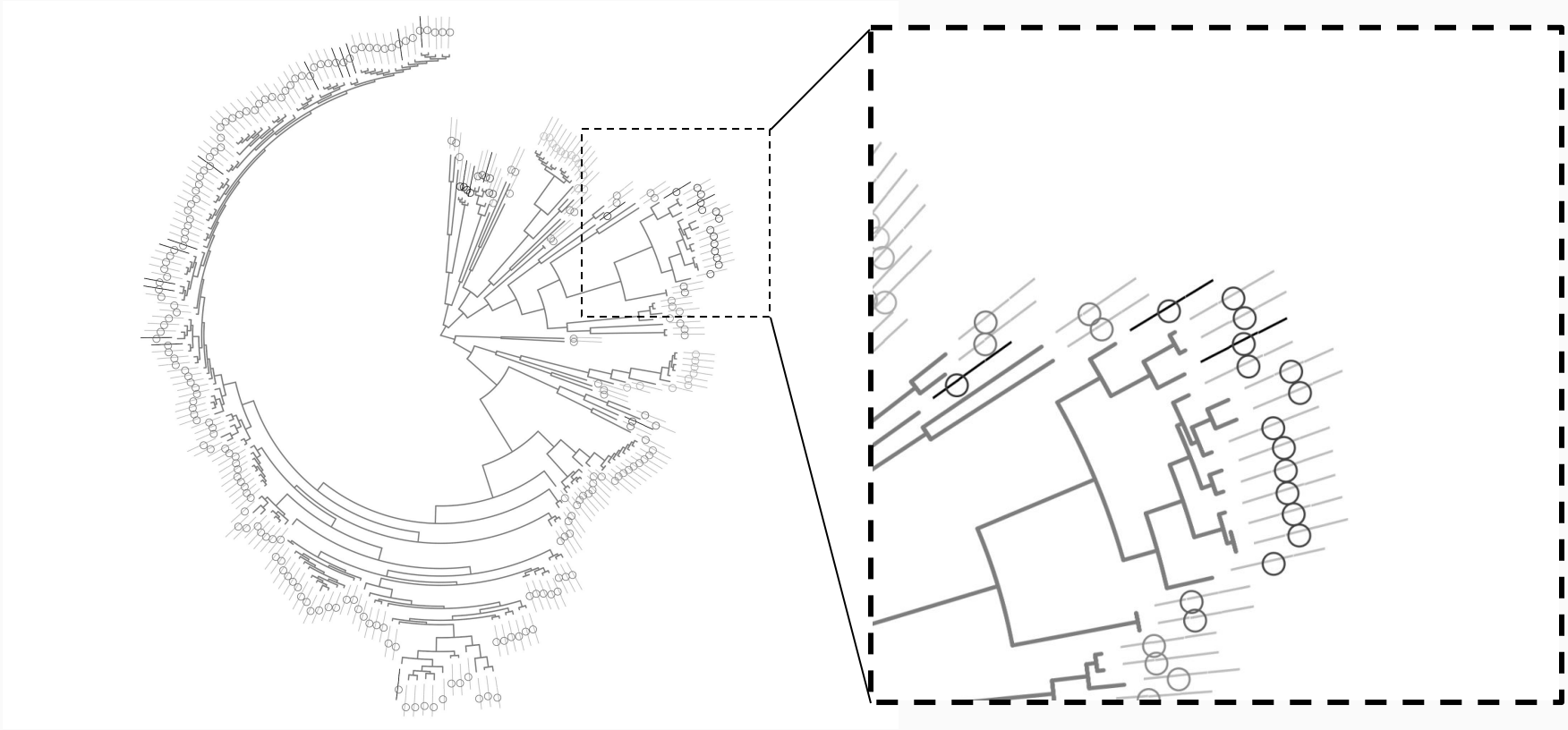
phenotype = covariates + foreground locus + background loci + environment

decomposed into lineage-level effects

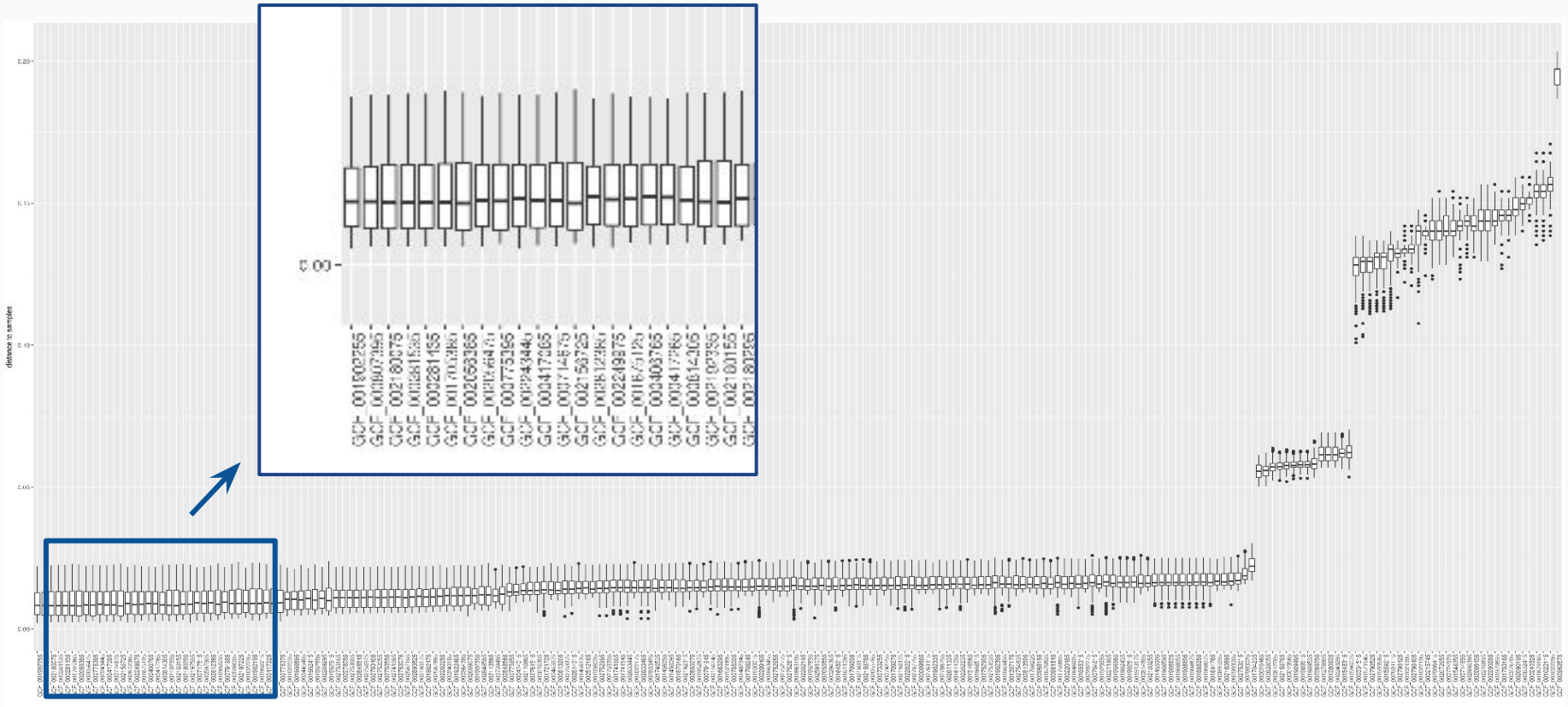
Principal components correspond to lineages in the clonal genealogy.

Nat Microbiol. 2016 Apr 4,Earle SG et al.

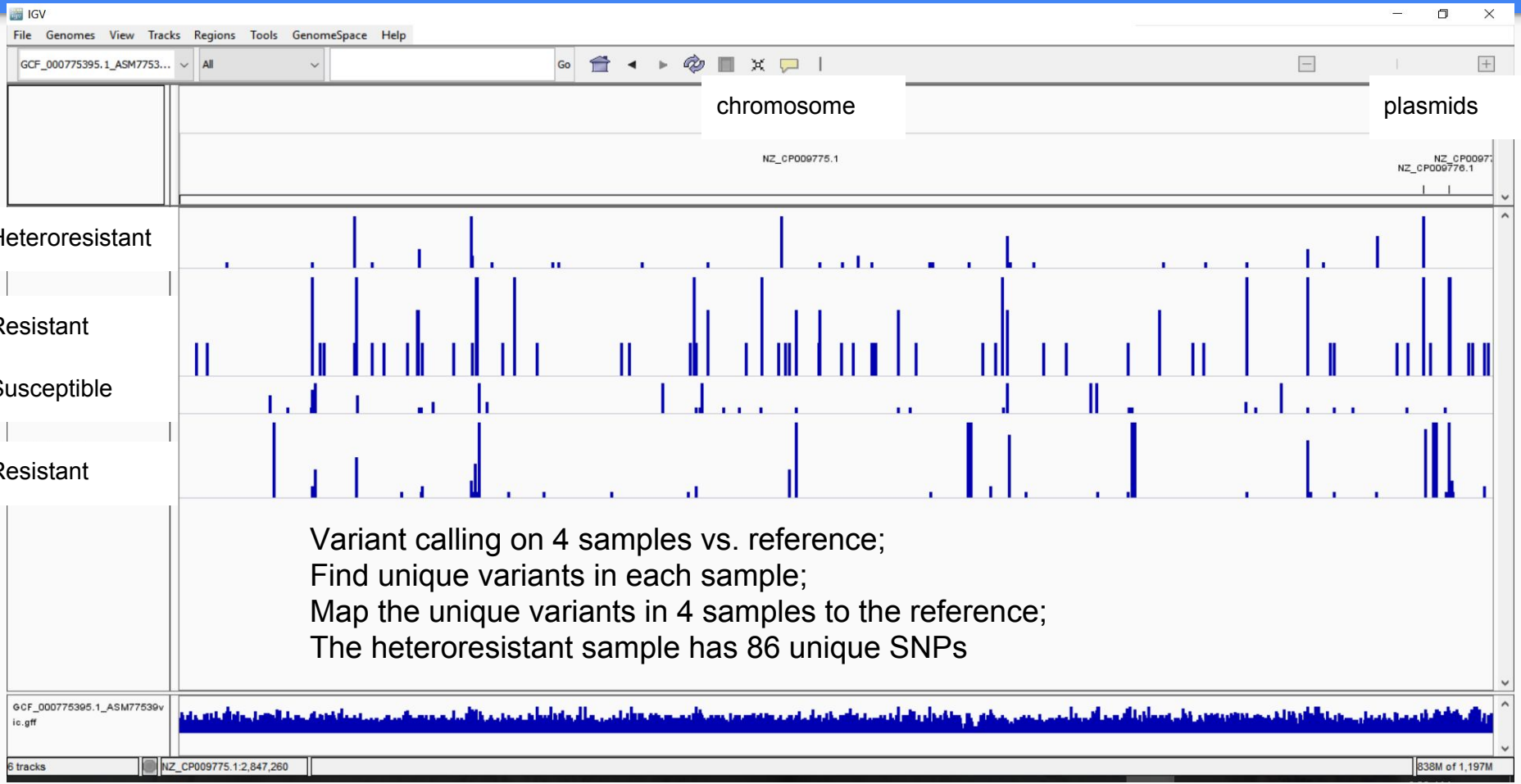
buGWAS LMM prediction



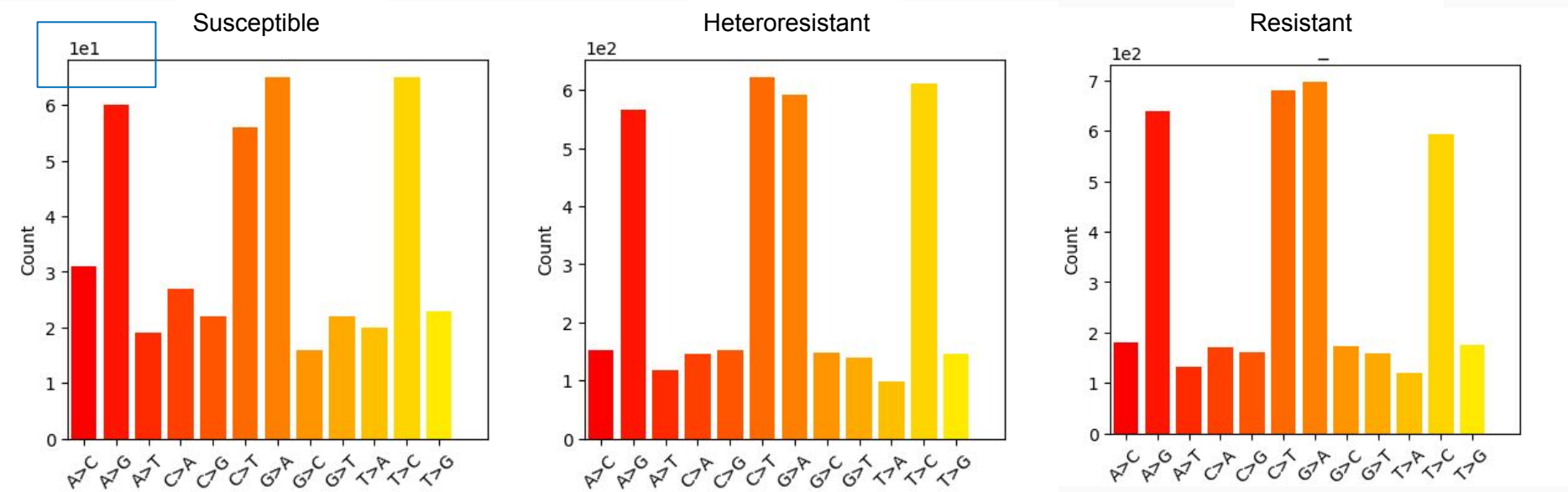
Selection of reference genome



Samtools and bcftools: unique SNPs in 4 samples vs. reference



Count of Substitutions in 10 Samples per Group



Pangenome GWAS (Features Comparison)

	Roary/Scoary	BacterialGWAS
Input	Annotation.gff & Trait file	Assembly.fasta & Trait file
Step before generate pangenome	N/A	Using Prodigal to predict gene
Gene Clustering	CD-Hit	CD-Hit
Association with phenotype	Logistic Regression	Logistic Regression
Result Analysis	Provided	Blast to check for function of significant genes

bacterialGWAS pangenome

Assembled genomes
Trait file

→

Gene prediction
Prodigal

→

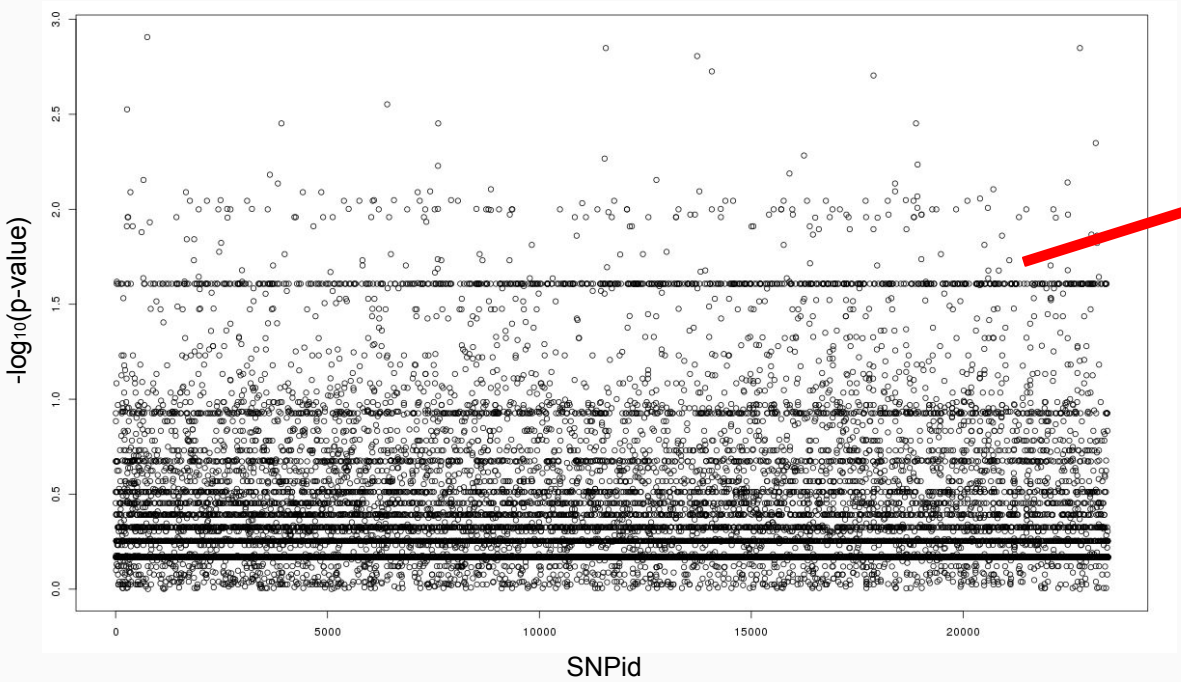
Genes Clustering
CD-Hit

→

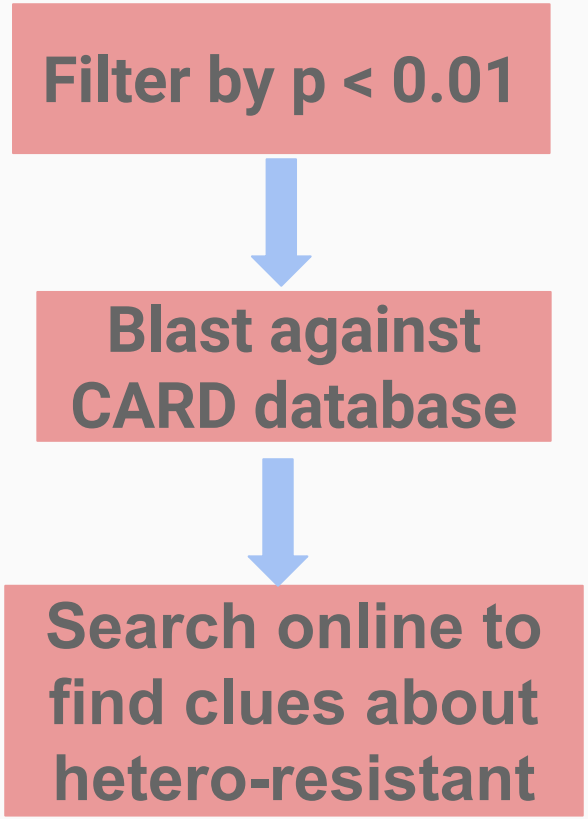
Association with phenotype
(logistic regression)

Case:
Hetero-resistant

Control:
Susceptible
Resistant



bacterialGWAS pangenome, obtain significant genes



bacterialGWAS pangenome, significant genes

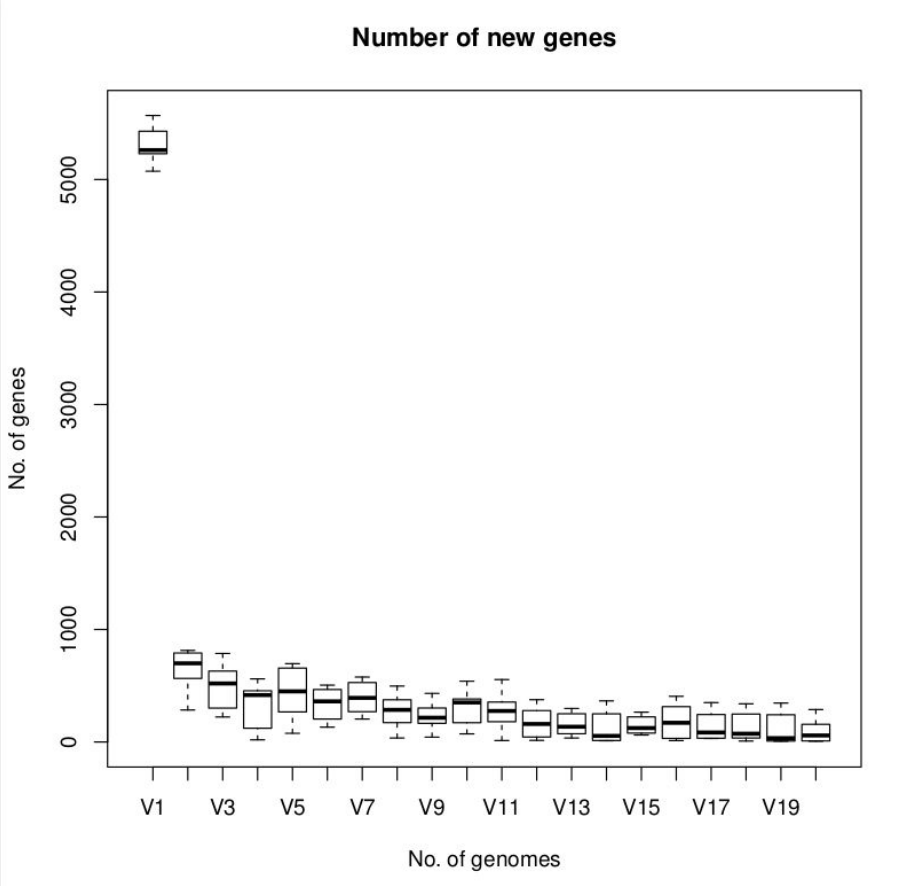
tet(59) tet(A) tet(B) tet(31) tet(E) tet(G) tet(J) tet(H) tet(Y)
tet(Z) tet(41) tet(39) tet(33) tet(30)
ANT(2'')-Ia mexK
TriC adeJ amrB mexI MuxB acrD mdtC ceoB
smeE MexB mdtN optrA TaeA salA carA srmB
oleB vgaB tlrC mel
OXA-9 OXA-18
APH(3')-IIa APH(3')-IIb APH(3')-IIc adeL

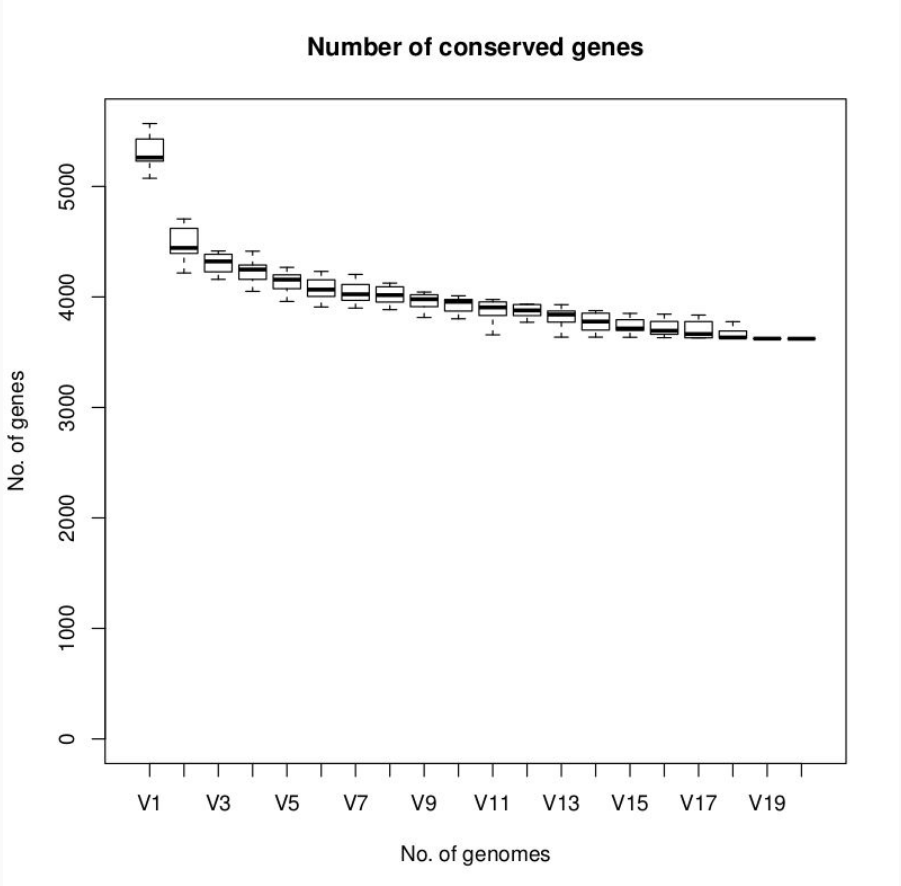
bacterialGWAS pangenome, significant genes

1. tet(A): PMID: 28261566 doi: 10.3389/fcimb.2017.00037
2. tet(B): PMID: 25268178 DOI: 10.1179/1973947814Y.0000000213
3. adeJ: http://aac.asm.org/content/54/12/5021.full; https://peerj.com/preprints/2655.pdf
4. amrB: doi: 10.3389/fmicb.2016.01846 PMID: 27920760
5. MexB: http://cmr.asm.org/content/28/1/191.full
6. optrA: http://aac.asm.org/content/early/2018/01/09/AAC.02007-17.full.pdf
7. carA: https://www.researchgate.net/profile/Josiah_Onaolapo/publication/281476167_Plasmid_Profile_of_Antibiotics_Heteroresistant_Escherichia_coli_Isolates_from_Diarrhoeic_Children_Attending_Ahmadu_Bello_University_Teaching_Hospital_Shika_Zaria_Nigeria/links/5669a4e208ae1a797e3762e4/Plasmid-Profile-of-Antibiotics-Heteroresistant-Escherichia-coli-Isolates-from-Diarrhoeic-Children-Attending-Ahmadu-Bello-University-Teaching-Hospital-Shika-Zaria-Nigeria.pdf

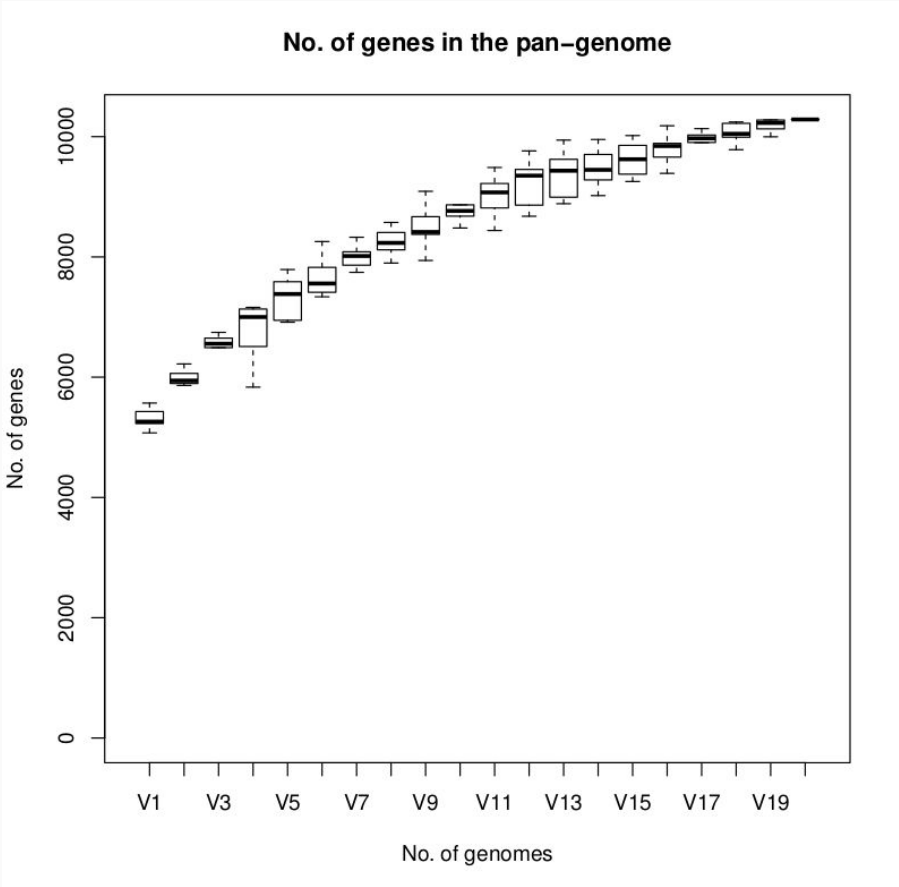
8. APH(3')-IIa: http://aac.asm.org/content/early/2017/12/05/AAC.01601-17.short?rss=1

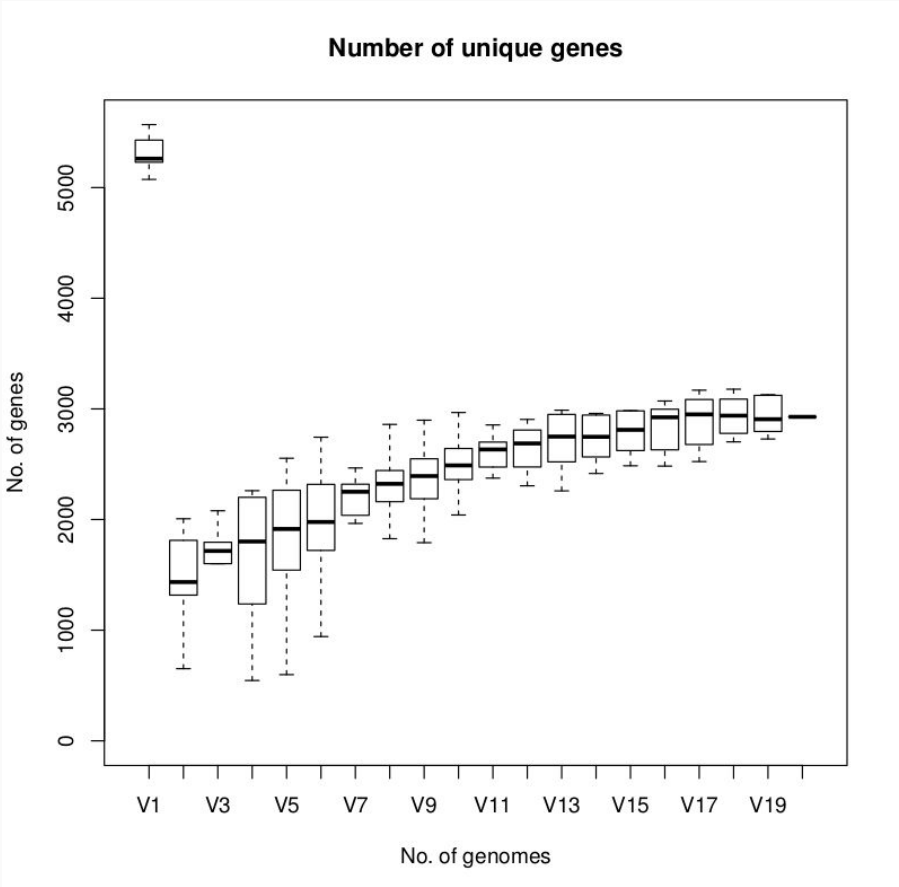
Advance Visualization from Roary





Advance Visualization from Roary





Comparison of Plasmids

Goals:

-- Assemble plasmidomes for each sample, plus pan-plasmidome

-- Look for gene duplications, plasmid CNVs using STing

-- Identify presence/absence of known AR genes located on plasmids

Plasmidome Assembly

plasmidSPAdes used for isolates
-- uses read depth outliers from the median depth to identify potential plasmids in of the genome, then *de novo* assembles those with a de Bruijn graph.

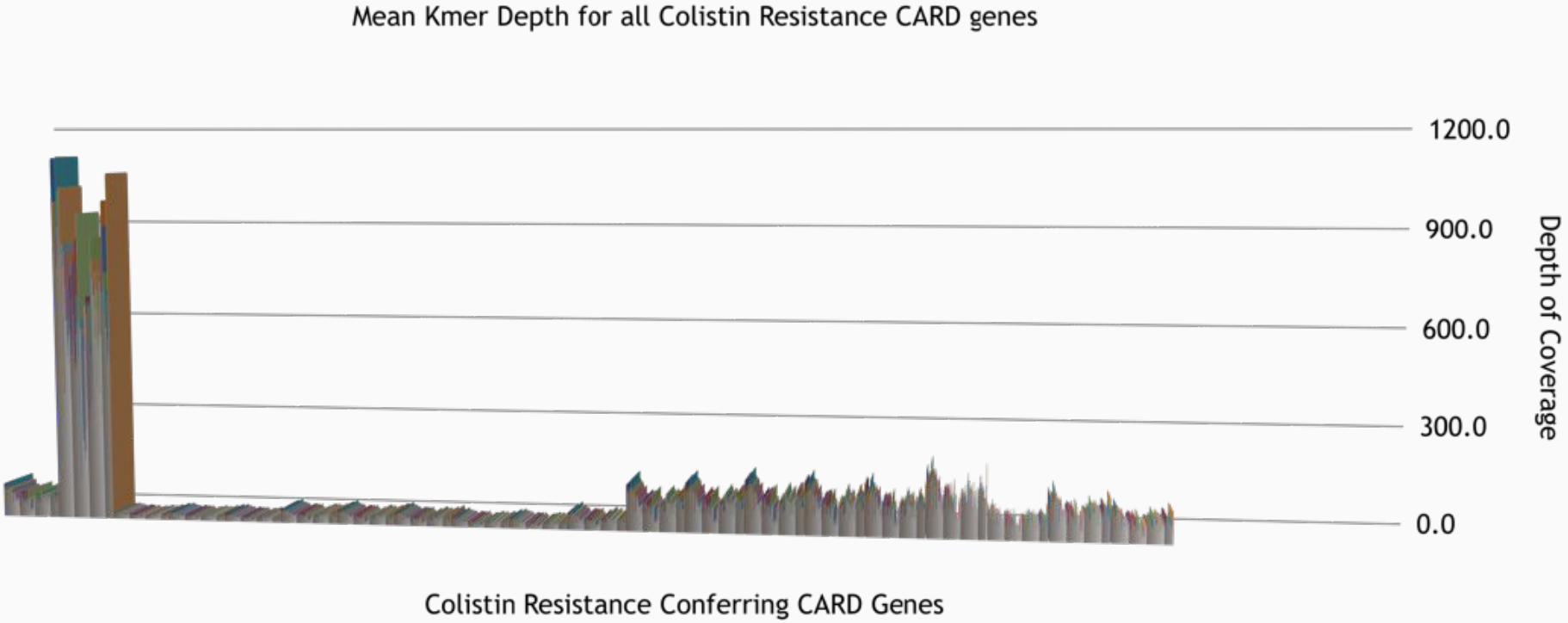
CISA used for pan-plasmidome assembly
-- merges all assemblies into one, then aligns contigs to remove duplicates.
Identified 42 plasmids, mean length: 109 Kbp

STing Analysis

- STing was first used to determine average k-mer depth across colistin resistance genes (from CARD)

- Looking for clear distinction between phenotypic classes
  - Copy number of all CARD genes
  - Full plasmid copies
  - CARD gene copies within a plasmid

STing Analysis

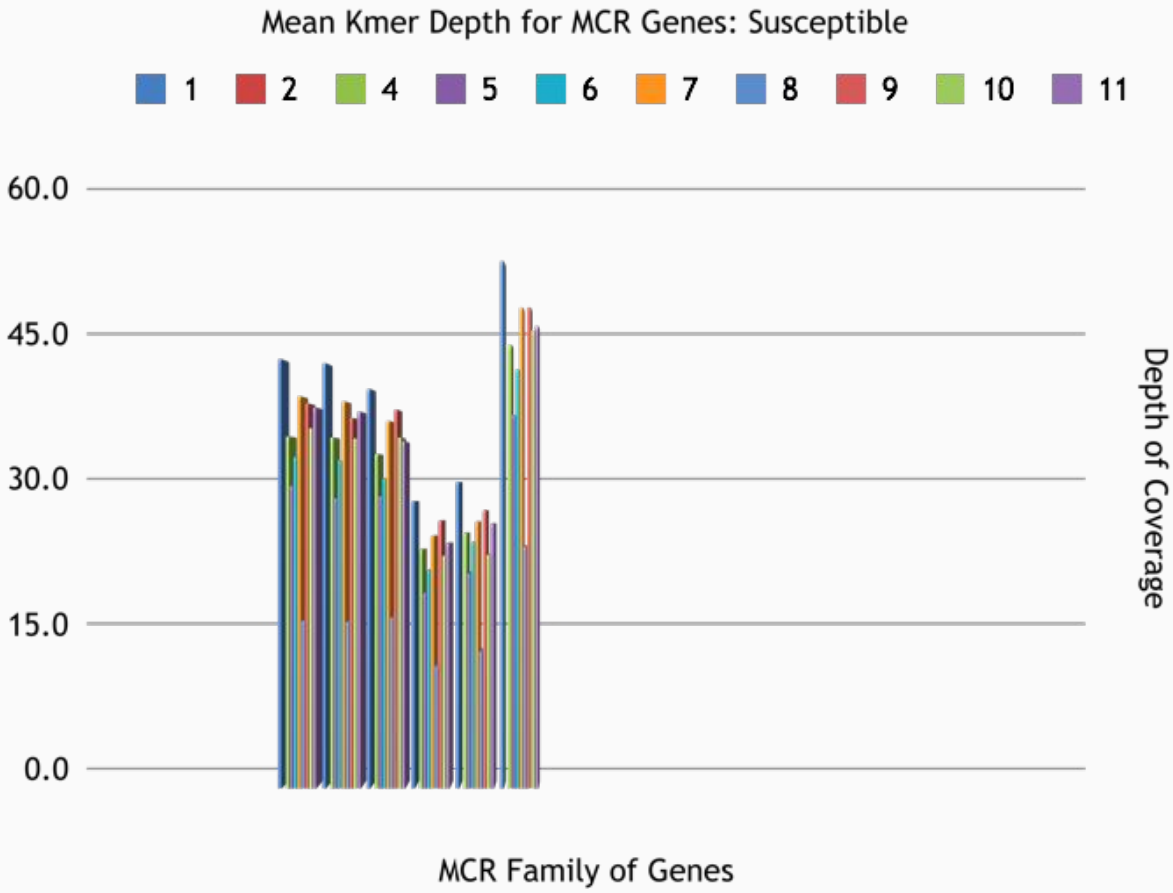


STing Analysis: Narrowing Focus

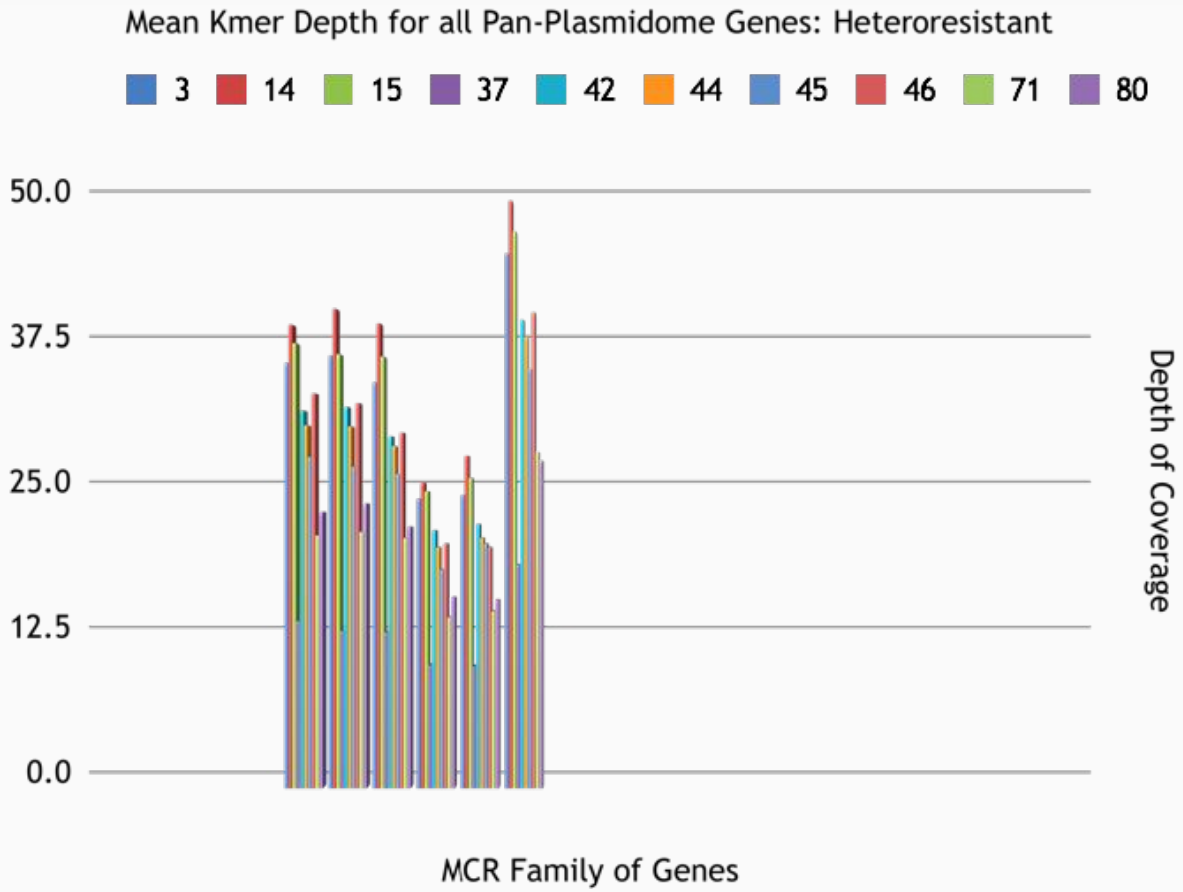
- K-mer depths across a representative group of colistin resistance conferring genes pulled out for further investigation

- Looking for differences between the raw abundance of k-mers mapping to particular genes between phenotypic classes

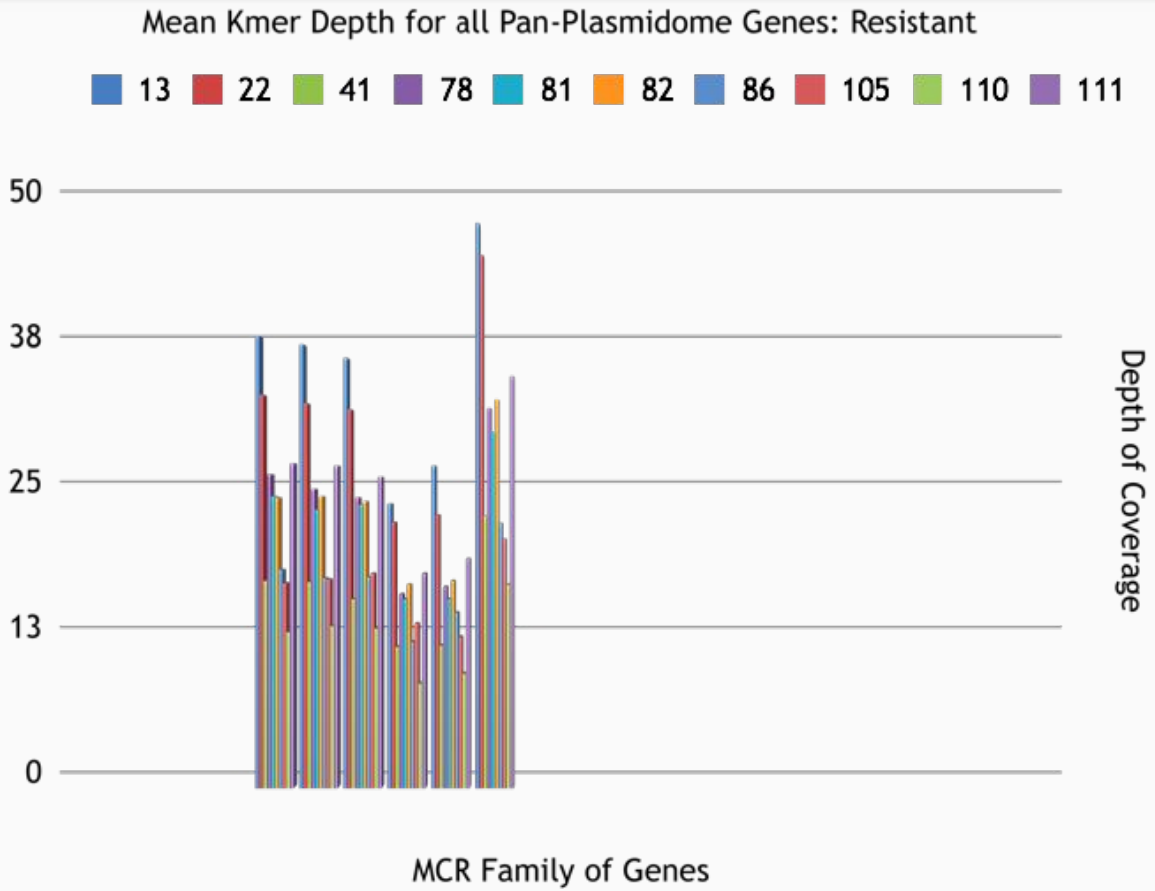
Plasmidome Analysis



Plasmidome Analysis



Plasmidome Analysis



STing Analysis: Pan-Plasmidome Analysis

- Using a pan-plasmidome a STing GDETECT database was created
- STing then calculated the per nucleotide k-mer depth
  - The per nucleotide k-mer depth can be used to determine copy number, and relative abundance

STing Analysis: Pan-Plasmidome Analysis

Coming soon: Some sort of perbase coverage histogram for genes in the pan-genome

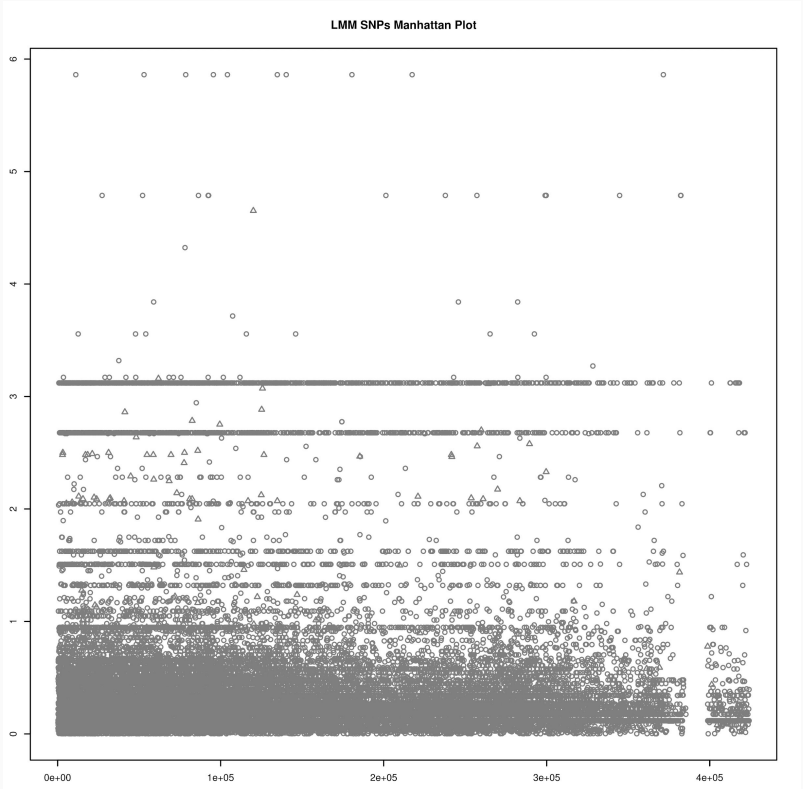
Recommendations for Continuing Analysis

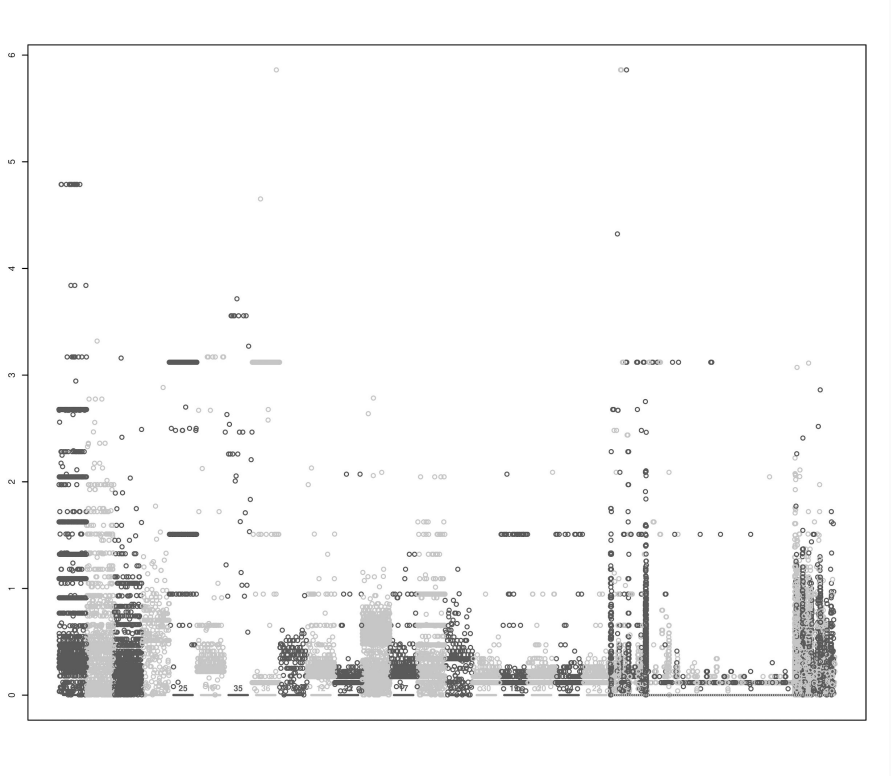
1. Create a STing GDETECT database including all pan-plasmidome genes, all colistin-CARD genes, 16S housekeeping genes and the origins of replication from our plasmids
2. Using STing determine the mean and per base k-mer depth across all genes
3. Normalize relative abundance for genomic genes based on HK genes, and plasmid genes ORI genes

References

1. Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., and Phillippy, A. M. (2016) Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biology* **17**, 132
2. Andrew J. Page, Carla A. Cummins, Martin Hunt, Vanessa K. Wong, Sandra Reuter, Matthew T.G. Holden, Maria Fookes, Daniel Falush, Jacqueline A. Keane, Julian Parkhill; Roary: rapid large-scale prokaryote pan genome analysis, *Bioinformatics*, Volume 31, Issue 22, 15 November 2015, Pages 3691–3693, https://doi.org/10.1093/bioinformatics/btv421
3. https://github.com/sgearle/bugwas/tree/master/bugwas
4. Falush, D. (2016) Bacterial genomics: Microbial GWAS coming of age. *Nature Microbiology* **1**, 16059
5. Olson, Nathan D., et al. "Best practices for evaluating single nucleotide variant calling methods for microbial genomics." *Frontiers in genetics* 6 (2015): 235.
6. Power RA, Parkhill J, de Oliveira T., Microbial genome-wide association studies: lessons from human GWAS. Nat Rev Genet. 2017 Jan;18(1):41-50. doi: 10.1038/nrg.2016.132. Epub 2016 Nov 14.

buGWAS LMM Manhattan Plot





buGWAS Bayesian Wald Test