Comparative Genomics

Lab & Preliminary Result

Team 1

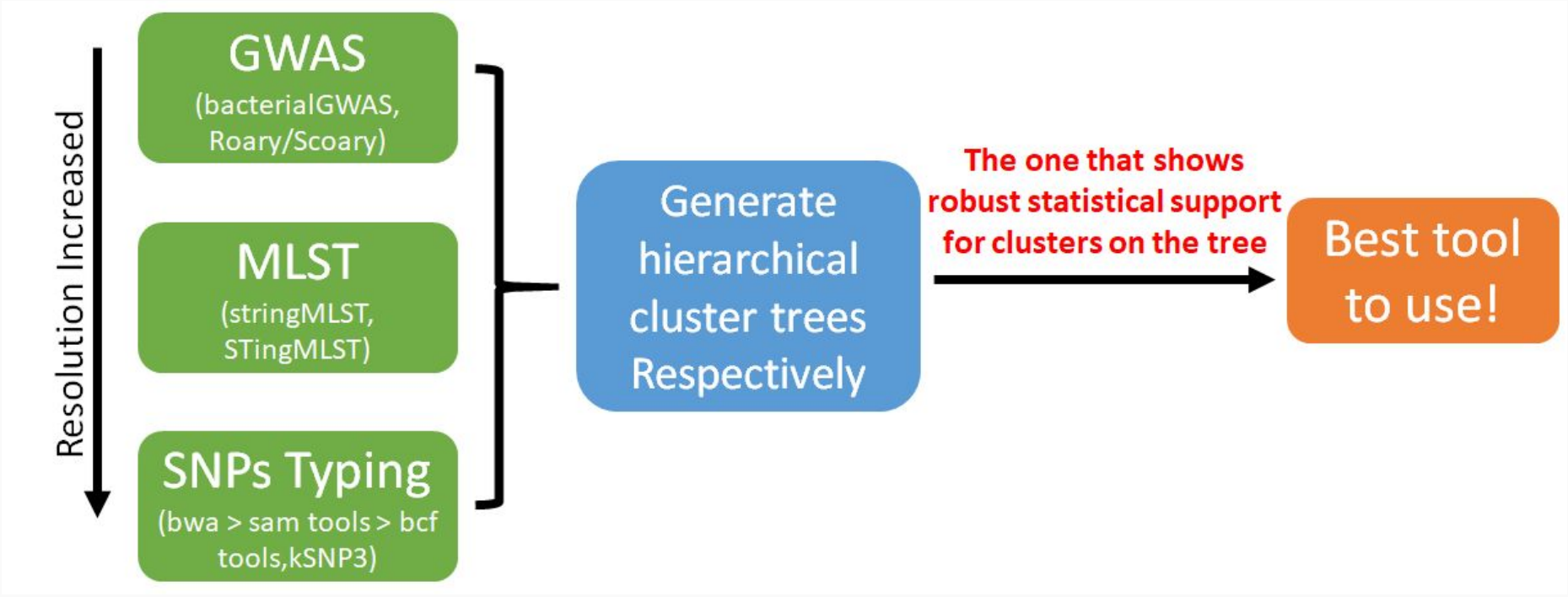
Team Members:
Frank Ambrosio, Hunter Seabolt, Vasanta Chivukula, Junyu Li, Yiqiuyi Liu, Seonggeon Cho, Yihao Ou, Qinyu Yue, Siarhei Hladyshau

Outline

Content

- Introduction
- GWAS
- MLST
- SNP calling
- Comparison & Decision

Introduction



Introduction

Samples Description

Klebsiella spp.	Susceptible	Heteroresistant	Resistant	Total
Number of sample	212	21	25	258

Test Samples

Klebsiella spp.	Susceptible	Heteroresistant	Resistant	Total
Number of sample	10	5	5	20

Introduction

GOAL

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

- **Predict** colistin susceptibility of other *Klebsiella* spp. strains

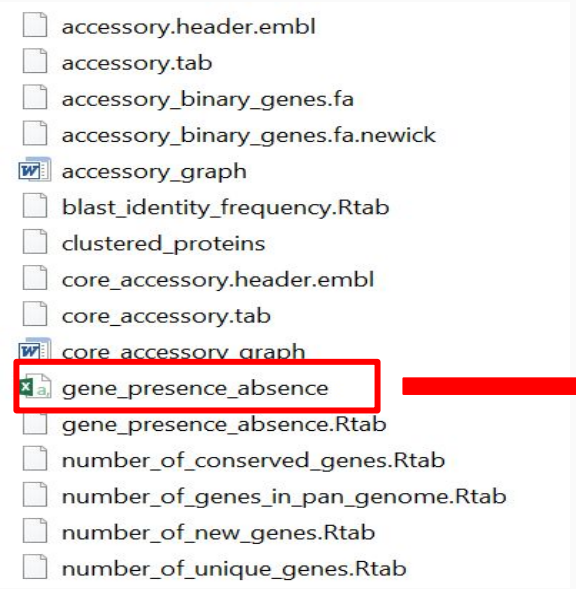
GWAS

Roary/Scory

Input:

GFF3 files, must contain the nucleotide sequence at the end of the file (output from Prokka)

Output:



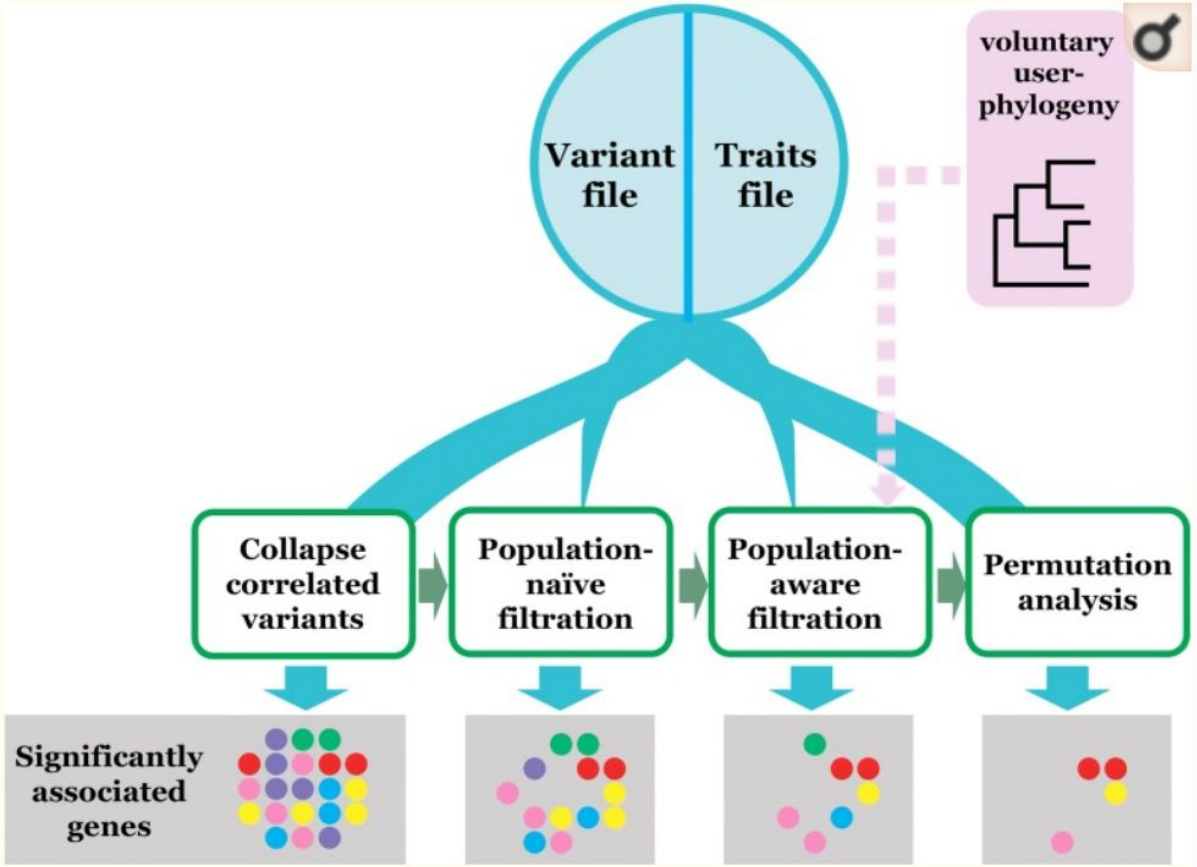
Input for Scoary

GWAS

Roary/Scory

Input:

- gene_presence_absence.csv file from Roary
- list of traits to test associations

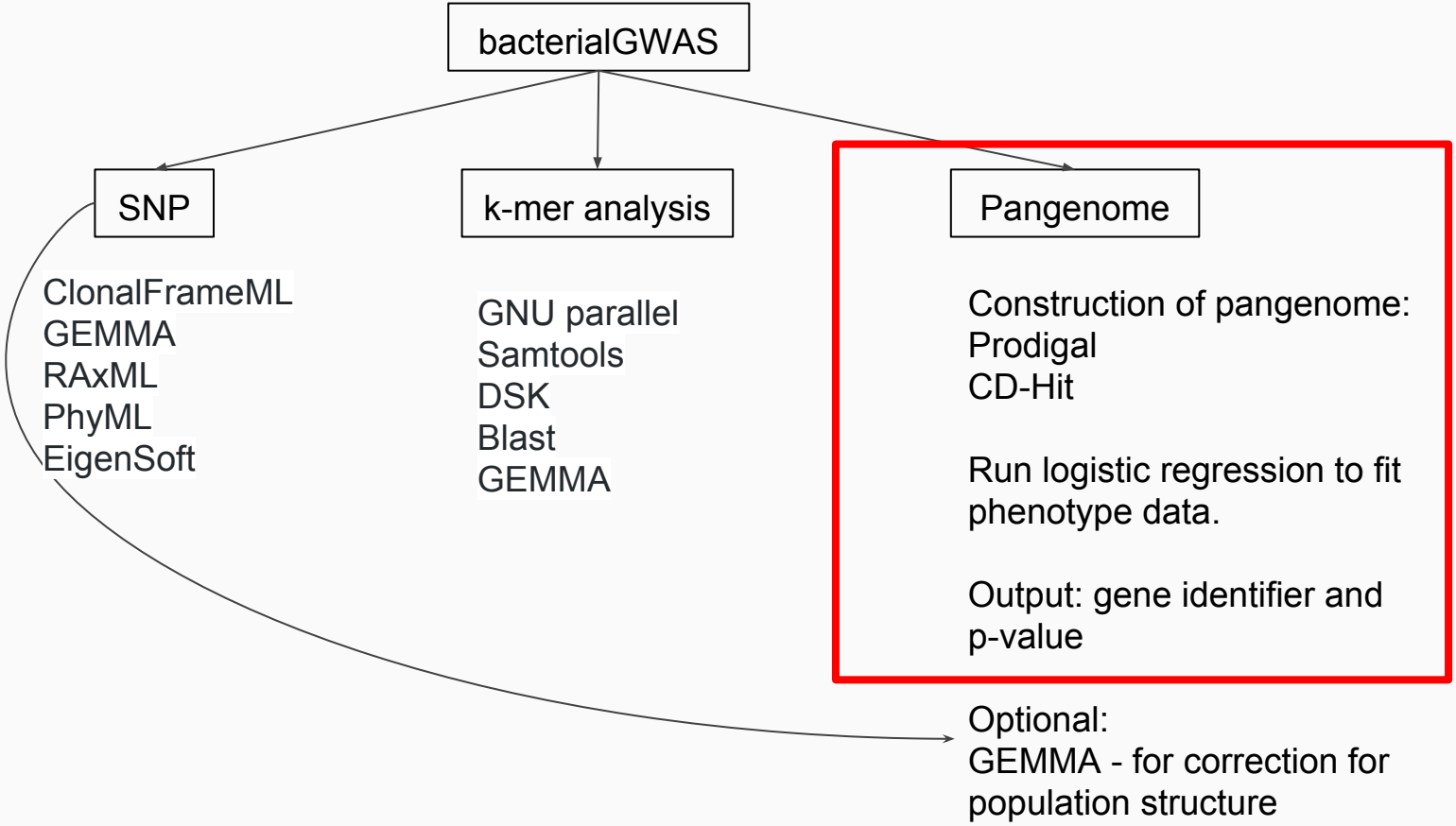


GWAS

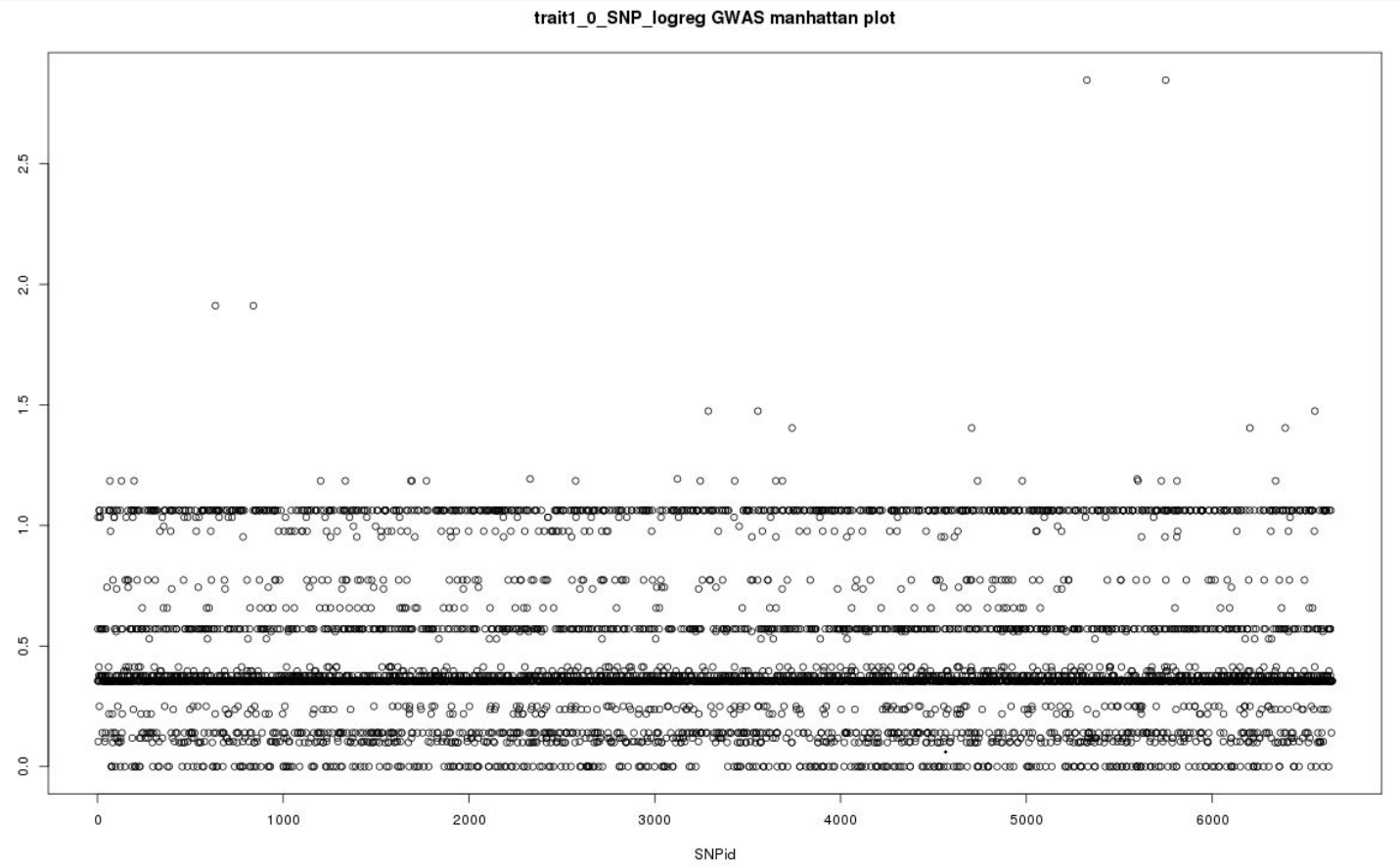
Roary/Scory

Column name	Explanation
Gene	The gene name
Non-unique gene name	The non-unique gene name
Annotation	Annotation
Number_pos_present_in	The number of trait-positive isolates this gene was found in
Number_neg_present_in	The number of trait-negative isolates this gene was found in
Number_pos_not_present_in	The number of trait-positive isolates this gene was not found in
Number_neg_not_present_in	The number of trait-negative isolates this gene was not found in
Sensitivity	The sensitivity if using the presence of this gene as a diagnostic test to determine trait-positivity
Specificity	The specificity if using the non-presence of this gene as a diagnostic test to determine trait-negativity
Odds_ratio	[Odds ratio] (https://en.wikipedia.org/wiki/Odds_ratio)
p_value	The naïve p-value for the null hypothesis that the presence/absence of this gene is unrelated to the trait status

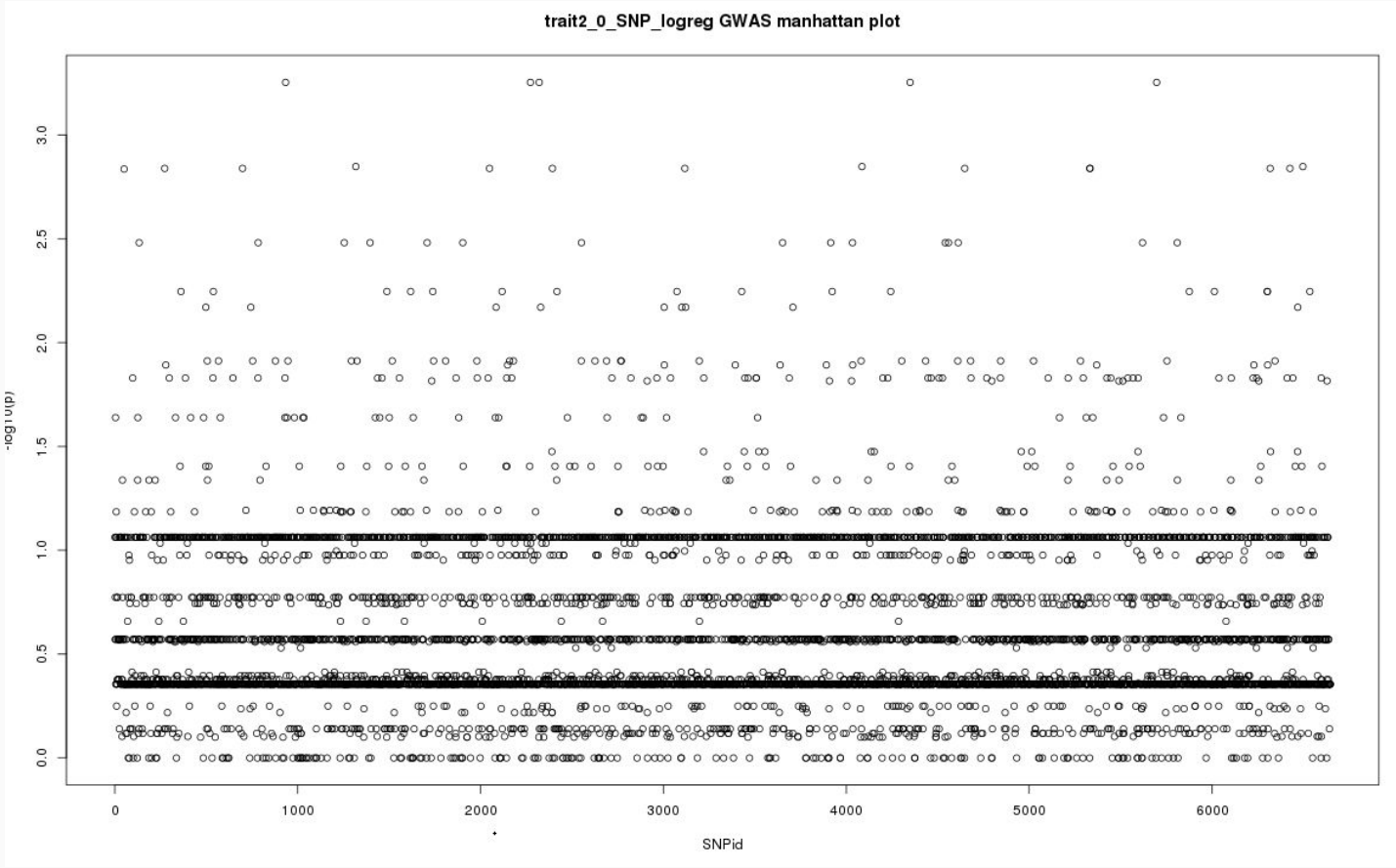
GWAS


bacterialGWAS
SNP
k-mer analysis
Pangenome
ClonalFrameML
GEMMA
RAxML
PhyML
EigenSoft
GNU parallel
Samtools
DSK
Blast
GEMMA
Construction of pangenome:
Prodigal
CD-Hit
Run logistic regression to fit phenotype data.
Output: gene identifier and p-value
Optional:
GEMMA - for correction for population structure

GWAS


trait1_0_SNP_logreg GWAS manhattan plot
0.0
0.5
1.0
1.5
2.0
2.5
0
1000
2000
3000
4000
5000
6000
SNPid

GWAS


trait2_0_SNP_logreg GWAS manhattan plot
-log10(p)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1000
2000
3000
4000
5000
6000
SNPid

GWAS

	Roary/Scoary	BacterialGWAS
Input	GFF3 file from Prokka (annotated gene)	Can be assembly or Customized pangenome
Output	Variant and p-value	Variant and p-value
Pros/Cons	Developed specifically for microorganism	Developed specifically for bacteria
	Takes input strait from Prokka, no need for extra work	customized pangenome is an option and should be able to generate by annotation group
	Easy to installed and run	Required a lot of dependence
	population stratification is taken into account	correction for population stratification using SNP data

MLST

stringMLST:

Assembly-free k mer-based MLST tool

Input: pre-assembled FASTQ files, database

Output: sequence type (ST) of the sample and allele profile corresponding to the ST

MLST

Running stringMLST:

```
stringMLST.py --buildDB -c klebsiella-pneumoniae_config.txt -k 35 -P PN

stringMLST.py --predict -1 forward.fastq.gz -2 reverse.fastq.gz -k 35 -p --prefix PN
```

```
[loci]
gapA  mlst_dbs/klebsiella-pneumoniae/klebsiella-pneumoniae_gapA.tfa
mdh   mlst_dbs/klebsiella-pneumoniae/klebsiella-pneumoniae_mdh.tfa
tonB  mlst_dbs/klebsiella-pneumoniae/klebsiella-pneumoniae_tonB.tfa
infB  mlst_dbs/klebsiella-pneumoniae/klebsiella-pneumoniae_infB.tfa
pgi   mlst_dbs/klebsiella-pneumoniae/klebsiella-pneumoniae_pgi.tfa
rpoB  mlst_dbs/klebsiella-pneumoniae/klebsiella-pneumoniae_rpoB.tfa
phoE  mlst_dbs/klebsiella-pneumoniae/klebsiella-pneumoniae_phoE.tfa
[profile]
profile    mlst_dbs/klebsiella-pneumoniae/klebsiella-pneumoniae_profile.txt
```

multi-FASTA files with allele sequence

ST and the allele profile corresponding to the ST

MLST

stringMLST Output:

	Sample	gapA	infB	mdh	pgi	phoE	rpoB	tonB	ST
1	SRR3982229	3	3	1	1	1	1	79	258
2	SRR3982316	3	3	1	1	1	1	79	258
3	SRR4035118	3	3	1	1	1	1	79	258
4	SRR5666552	3	3	1	1	1	1	79	258
5	SRR5666401	3	4	6	1	7	4	4	273
6	SRR5666402	2	1	1	1	4	4	4	17
7	SRR5666403	2	1	2	1	4	4	87	294
8	SRR5666404	3	3	1	1	1	1	79	258
9	SRR5666405	2	5	2	2	7	1	10	48
10	SRR5666406	2	1	2	1	4	4	4	16
11	SRR5666407	3	3	1	1	1	1	18	340
12	SRR5666408	2	1	5	1	17	4	42	111
13	SRR5666409	2	5	2	2	7	1	10	48
14	SRR5666410	3	3	1	1	1	1	4	11

MLST

Current problem: database

- 7 housekeeping genes is not nearly enough

Next step:

- new database with stringMLST & Gene detection function with STing

List of genes of interest:

- AMR genes, colistin-associated genes, efflux pump associated genes, the plasmid mediated resistance mcr family of genes

STing

Functionality:

- **Database Construction**
- Sequence Typing
- **Gene Detection**

List of genes of interest:

- AMR genes, colistin-associated genes, efflux pump associated genes, the plasmid mediated resistance mcr family of genes

STing

Database Construction

- Much easier for STing Gene Detection
- Colistin resistance associated genes pulled from CARD DB for testing
- Expect to see high occurrence of MCR gene family in resistance isolates

Next steps:

- Construct a new database with for STing Gene Detection
  - Include all colistin resistance associated genes from all available AMR databases
- Construct a database for STing Sequence Typing and String MLST

SNP

kSNP3:

Alignment-free kmer-based SNP calling tool

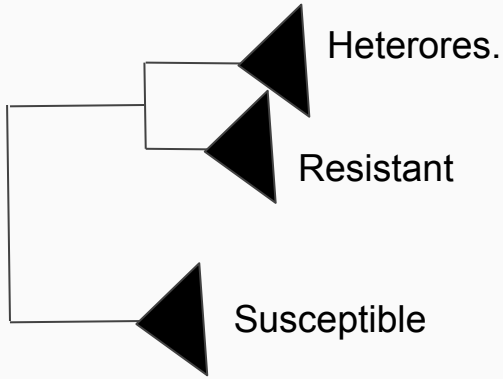
Input: (wg) Fasta file, model/bootstrap parameters

Output: Newick trees, etc.

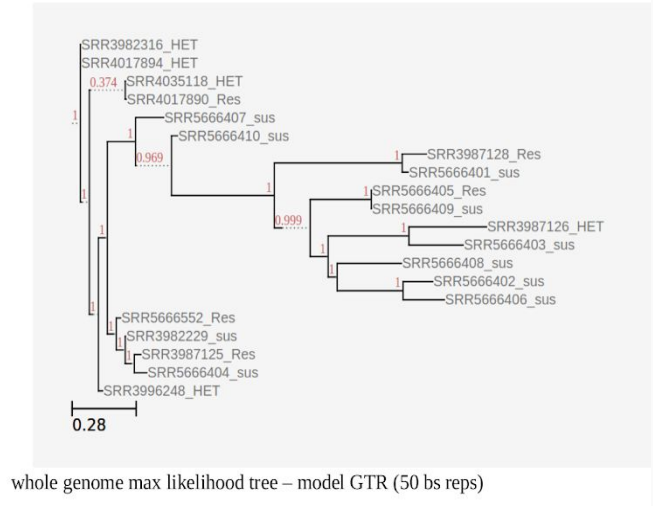
Expected(Hoped for) output:

3 monophyletic clusters (het, res, sus)

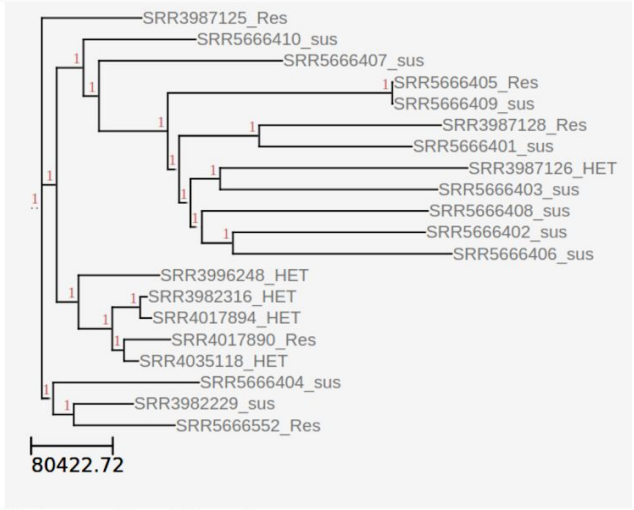
With good statistical support



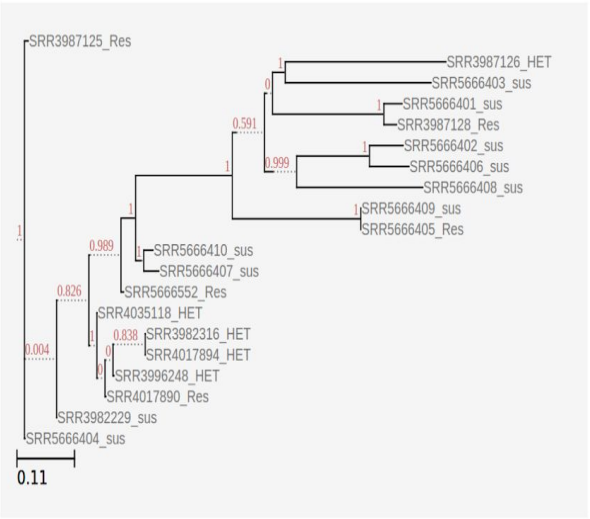
SNP



whole genome max likelihood tree – model GTR (50 bs reps)



Whole genome NJ tree (50 bs reps)



core genome SNP ML tree – model GTR – 50 bs reps

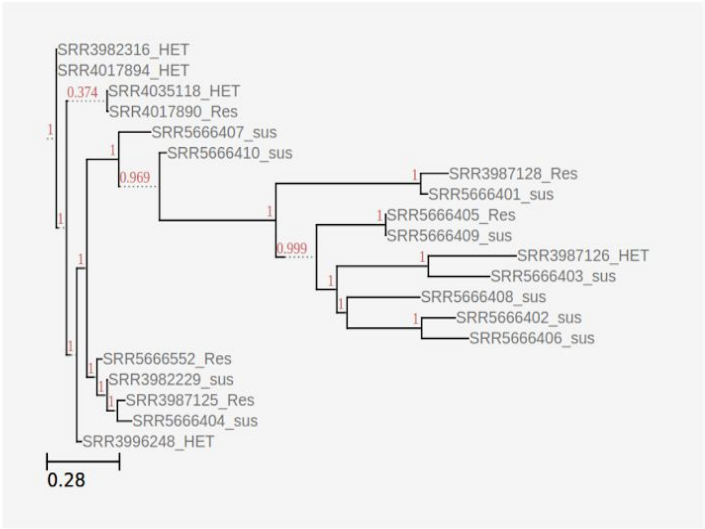
Input: de novo assembled Skesa fasta files

Notes from the trees:

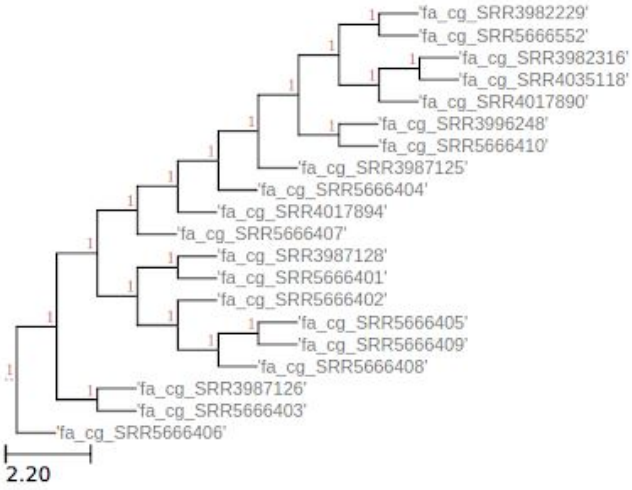
Same clusters appearing on multiple phylogenies with strong bootstrap support

No clear distinction between het/res/sus groups.

Comparison



whole genome max likelihood tree – model GTR (50 bs reps) SNP



GWAS -- UPGMA tree

Trees show very similar topologies -- phylogenetic signal is present, at differing levels of resolution (SNP trees have higher resolution).

Comparison

GOAL RECAP

- **Explore** gene features in *Klebsiella* that confer colistin resistance. Looking for fixed genomic differences indicating a "shared" ancestry between groups.
  - Little to no evidence of fixed genomic features based on SNP calling, GWAS, MLST
  - Cannot explain the AMR patterns observed with current anaylses (need to look for alternative hypotheses)
- **Predict** colistin susceptibility of other *Klebsiella* spp. strains

Citation

- "Roary: Rapid large-scale prokaryote pan genome analysis",
  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,
  Bioinformatics, (2015). doi: http://dx.doi.org/10.1093/bioinformatics/btv421

- Brynildsrud O, Bohlin J, Scheffer L, Eldholm V. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. Genome Biol. 2016;17:238 DOI: 10.1186/s13059-016-1108-8

- Fadista, J., Manning, A. K., Florez, J. C., & Groop, L. (2016). The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants. European Journal of Human Genetics, 24(8), 1202-1205. doi:10.1038/ejhg.2015.269

- Gardner, S. N., Slezak, T., & Hall, B. G. (2015). kSNP3. 0: SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome. Bioinformatics, 31(17), 2877-2878.