Comparative Genomics

Background & Strategy

Team 1

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

Outline

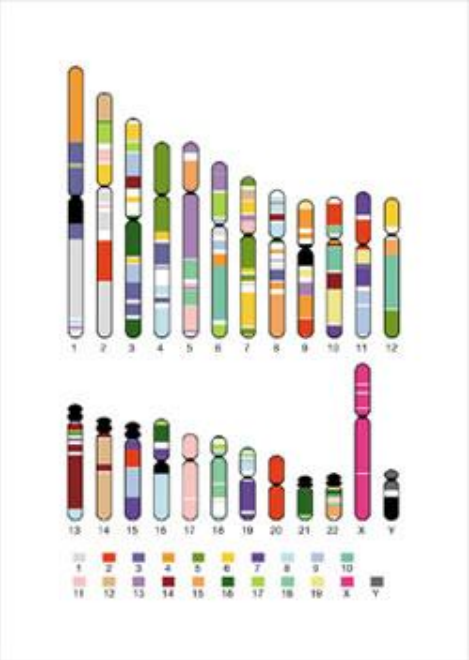
Content

- Background
- Goals
- Methods & Tools
- Prediction
- Strategy

Comparative Genomics: Background

- Comparative Genomics helps understand the roles of:
  - Common features of different organisms
  - Variants among closely related species
  - Variants within one species

- Different phylogenetic distances → different questions
  - distant genomes → general question

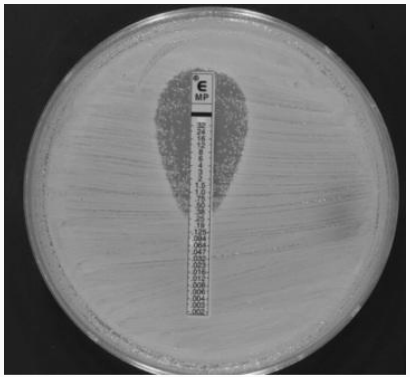


Conserved segments in the human and mouse genome (Lander, E. S. et al. 2001)

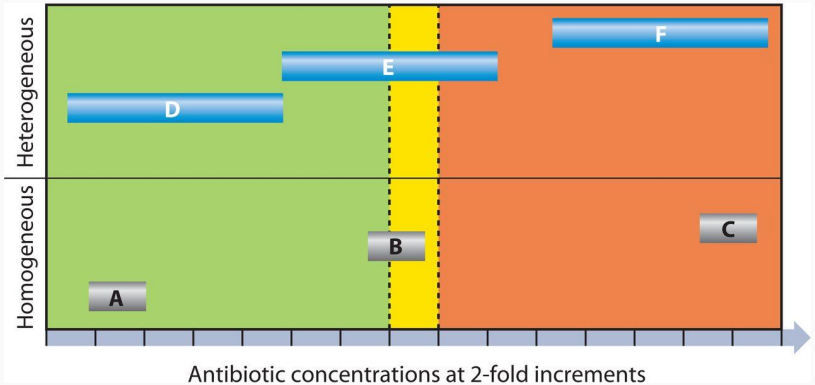
Comparative Genomics: Background

Heteroresistance

- The emergence of a subpopulation with lesser susceptibility in an otherwise susceptible population
- Determined via Etest for Antimicrobial susceptibility (AST) / Minimum Inhibitory Concentration (MIC)



An Etest strip of a heteroresistant strain (Weiss Lab)



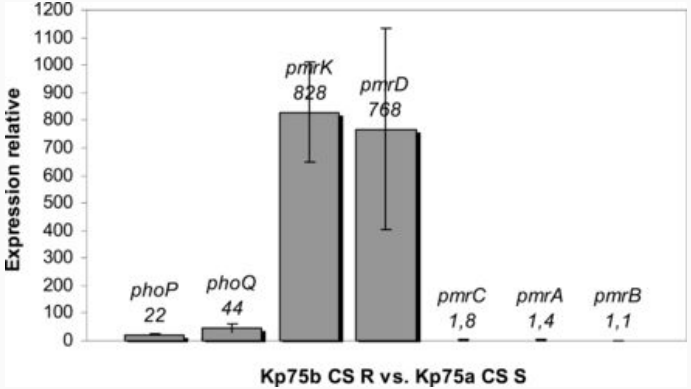
Heterogeneous versus homogeneous responses to antibiotics (El-Halfawy et al., 2015)

Comparative Genomics: Background

Colistin (polymyxin E) heteroresistance in *Klebsiella pneumoniae*

In a 2015 study by Jayol, Aurélie, et al.,

- Colistin-susceptible subpopulation has a partial deletion (25 bp) and a substitution in the *phoP* gene compared to colistin-resistant subpopulation.
- Leading to a single amino acid substitution in protein PhoP in the colistin-resistant subpopulation
- PhoP is part of the PhoPQ two-component system responsible for **l-aminoarabinose synthesis** and **polymyxin resistance**.



Upregulation in the resistant Kp75b (MIC=12 µg/ml) vs. susceptible isolate Kp75a (MIC=0.12 µg/ml)

Jayol, Aurélie, et al. "Heteroresistance to colistin in *Klebsiella pneumoniae* associated with alterations in the PhoPQ regulatory system." *Antimicrobial agents and chemotherapy* (2015): 2780-2784.

Comparative Genomics: Background

We use the following data:

- Assemblies of 258 *Klebsiella* spp. strains
- Predicted genes and ncRNAs on each genome
- Pan-genome annotation

We also have the colistin-resistance phenotypic data of each strain

- 212 susceptible
- 25 resistant
- 21 heteroresistant

Comparative Genomics: Goals

- **Explore** gene features in *Klebsiella* that confer colistin resistance.
- **Infer** colistin susceptibility of other *Klebsiella* spp. strains

Comparative Genomics: Method

- **Genome Wide Association Study (GWAS)**
- **Whole/Core/Pan/Accessory Genome Sequence Typing**
- **Single Nucleotide Polymorphism (SNP) Typing**

Comparative Genomics: Method (GWAS)

GWAS - association of phenotypic traits with genetic variants (SNPs, CNVs, polygenic features)

genotype ⟷ phenotype

Linear regression (beta)
Logistic regression (odd ratio)
Random forests

Pay attention to:

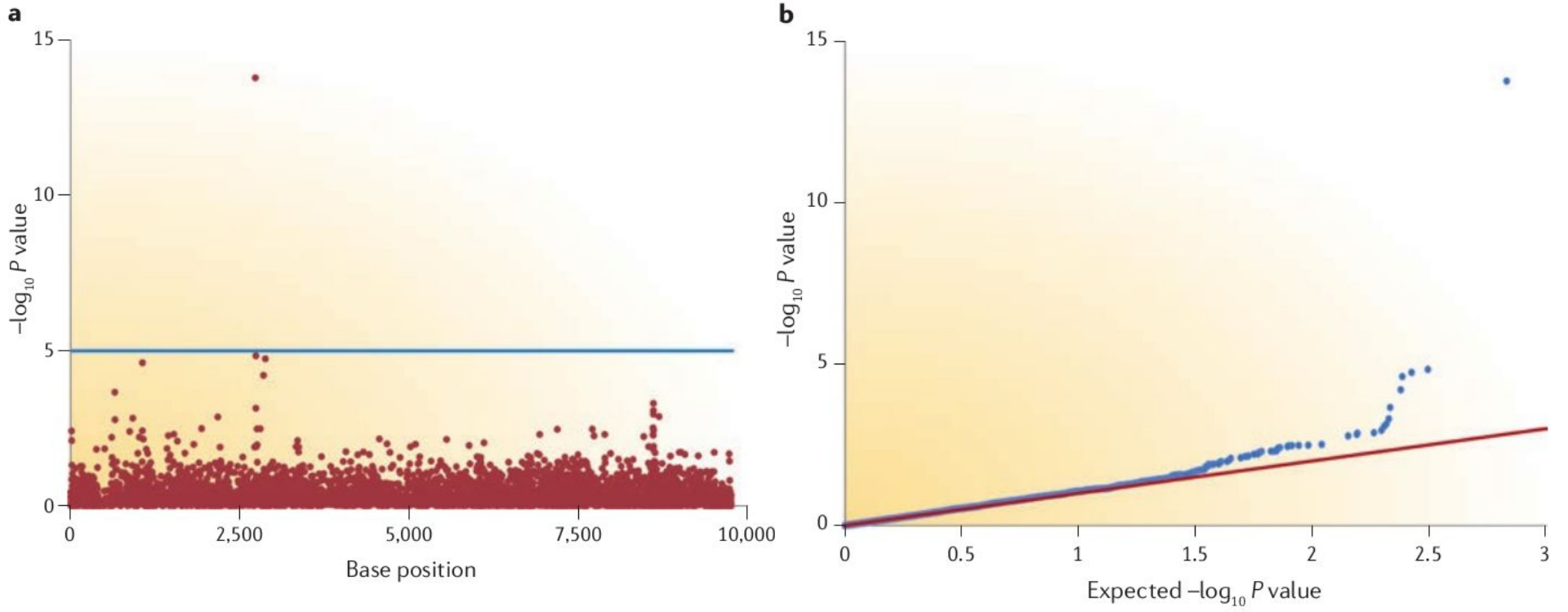
- Epistatic interactions
- Linkage disequilibrium
- Number of alleles
- Recombination in bacterial genomes is a usual thing!

P-values are our friends!

Comparative Genomics: Method (GWAS)

Manhattan plot

Quantile-quantile plot



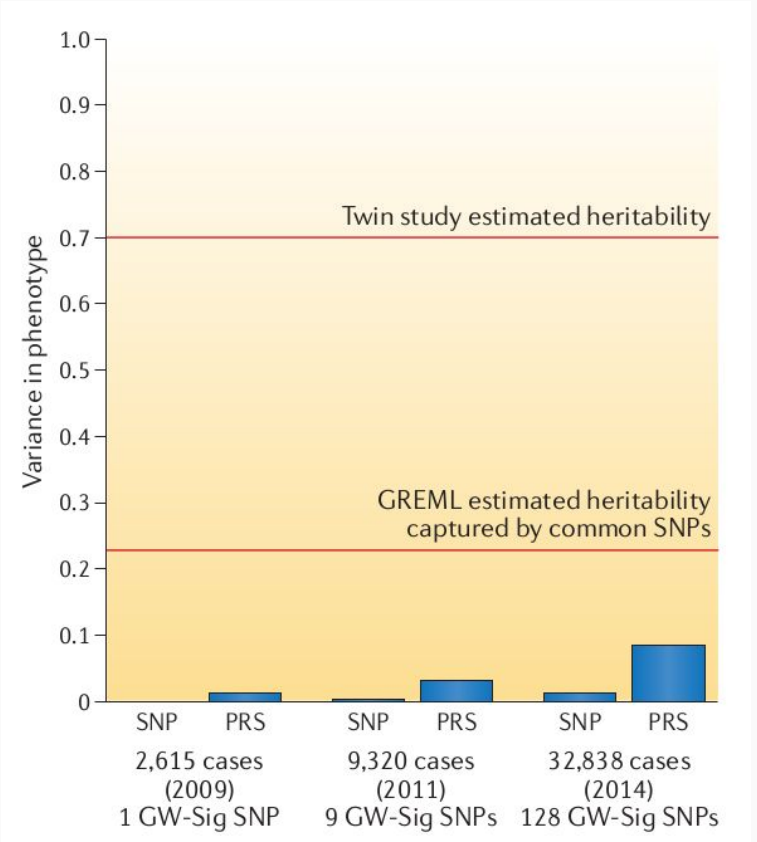
Nat Rev Genet. 2017 Jan; Power RA, Parkhill J, de Oliveira T

Comparative Genomics: Method (GWAS)

Always remember about false-positives
Possible confounding factors:

- Population stratification
  (even in homogeneous population)
- Homologous recombination
  (bias in linkage disequilibrium)
- Selection pressure on phenotype from population itself.

SNPs may predict very small fraction of variation in population.
Solution -- polygenic methods (polygenic risk scores).



Comparative Genomics: Tools (GWAS)

Software	Analysis	Population structure adjustment
PLINK	Linear and logistic regression of allele count at SNPs	Ancestry informative principal components and other covariate inclusion
PhyC	Identifies SNPs undergoing recent convergent evolution	Based on phylogeny, so inherent
ROADTRIPS	Association analysis of SNP effect, allowing random variables to account for sample relatedness	Corrects for provided or derived relatedness between samples
FaST-LMM	Association analysis of SNP effect, allowing random variables to account for sample relatedness	Derives relatedness matrix and corrects as random effect. Principal components can be included as covariates
SEER	Linear and logistic regression using k-mers, simultaneously testing SNPs and gene presence or absence	Identifies relatedness from data using multidimensional scaling and generates covariates for regression

Comparative Genomics: Tools (GWAS)

- **BacterialGWAS**
  - A pipeline for performing genome-wide association tests for bacterial genomes.
  - Use LMM to address population stratification bias
  - Have implementation of different approaches (SNPs, k-mers, pangenomes)
  - Already applied for study of antibiotic resistance in *Klebsiella*!

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

Comparative Genomics: Tools (GWAS)

- **Roary/Scoary**
  - Roary :
    - Rapid large-scale prokaryote pan genome analysis
    - 128 samples can be done in less than 1 hour
    - Using 1G of RAM and a single processor
  - Scoary:
    - Designed to take the gene_presence_absence.csv file from Roary, along with the trait file provided by user
    - Reports a list of genes sorted by strength of association per trait

Comparative Genomics: Methods (MLST)

- **MLST** - used to sort genomes based on alleles and variants of sequence fragments found at particular loci
  - Can be performed at various resolutions
    - Traditional MLST - Uses 7 housekeeping genes
    - Ribosomal MLST - Uses 53 highly conserved loci
    - Whole Genome/Pan Genome MLST - Uses loci from all across the genome based on their correlation to the phenotypes of interest (100s to 1000s of loci)

Comparative Genomics: Methods (MLST)

- **Whole/Core/Pan/Accessory Genome Sequence Typing**
  - wgMLST
  - Core Genome MLST
  - Pan Genome MLST
  - Accessory Genome MLST

Comparative Genomics: Tools (stringMLST)

- **K-mer max-hit matching algorithm**
  - Started with center k-mer of each read
    - Only looks at reads which have passed QC

- **Allelic typing schemes are defined in a database**
  - PubMLST database does not contain a scheme for *Klebsiella* spp.
  - PubMLST database does not contain a scheme for *K. pneumoniae*
  - PubMLST database K. pneumoniae **DOES** contain a scheme for *K. aerogenes*
  - PubMLST database K. pneumoniae **DOES** contain a scheme for *K. oxytoca*

Comparative Genomics: Tools (stringMLST)

- The ability of stringML**ST** to discover **S**equence **T**ypes related to our phenotypic classes will depend on the quality of our MLST scheme database.

- The quality of our MLST scheme database will depend on the resolution of our variant calling analysis of the isolate data.

Comparative Genomics: Tools (stringMLST)

- **Filtering**
  - Discards sequence reads in middle k-mer does not match database
- **Counting**
  - K-mer/database matches (reads) are k-merized
  - Each match between a k-mer and the database is counted
- **Reporting**
  - Alleles at each locus with the maximum k-mer matches are used to define allelic profiles

Comparative Genomics: Tools (stringMLST)

- **We should aim to create an MLST scheme for:**
  - *K. pneumoniae*
- **And determine if it is possible to create an MLST scheme for:**
  - *Klebsiella* spp.
  - Colistin Resistant *Klebsiella* spp.
  - Isolates with a 90-10 ratio of reads from susceptible and resistant strains respectively

Comparative Genomics: Tools (stringMLST)

- **Assembly Free**
  - Might be good to go straight from raw reads so as to not lose one our competing allelic populations within each heteroresistant isolate
- **Alignment Free**
- **Literally Free ($0)**
- **Relies on exact k-mer pattern matching pattern**
- **Developed by an extremely reputable bioinformatics laboratory**

Comparative Genomics: Methods (SNP Typing)

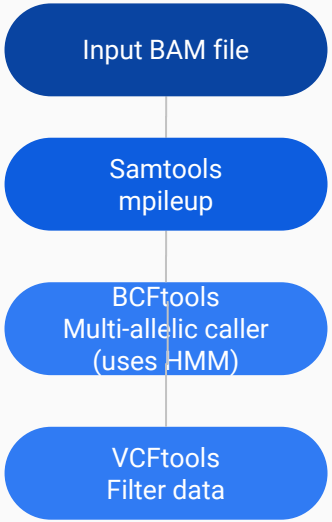
- **SNP:** Single nucleotide polymorphism (ie. a point mutation that is shared between a subset of the population due to common ancestry)

- **SNP Typing**: useful for phylogeny reconstruction, cluster detection
  - Offers extremely fine scale resolution
  - Useful here for cluster detection between highly similar genomes (but falls apart between more distantly related genomes)

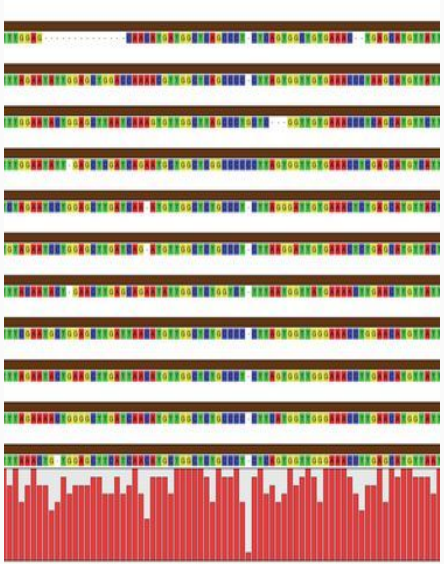
- **Widely used in public health applications**
  - PulseNet, CaliciNet, CryptoNet (CDC)

Comparative Genomics: Methods (SNP Typing)

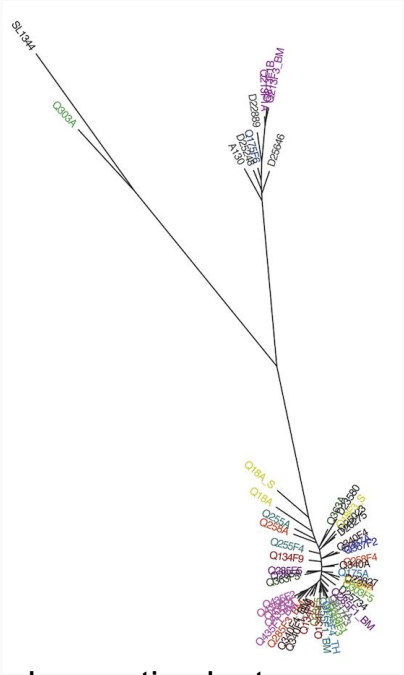
Example SNP cluster detection methodology:



Call variants from each genome



Construct an alignment matrix using SNP data from each genome



Phylogenetic cluster detection

Comparative Genomics: Tools (SNP Typing)

- **Varcap**
  - Integrate several variant calling software tools (to detect all types of variants):
    - SAMtools/bcftools
    - GATK
    - VarScan2
    - LoFreq-Star
    - FreeBayes
  - Fully automated workflow
  - Automatically performs quality filtering, mapping, variant calling and post-filtering of the predicted variants.

Comparative Genomics: Other Methods

- **Allelic Disequilibrium Analysis**
- **ANI**
- **Alignment-Free Methods**
- **Horizontal Gene Transfer (via Conjugation)**

Comparative Genomics: Prediction

- Species ID
- Gene/Allele Presence/Absence
- MLST
- Pan Genome MLST
- stringMLST
- SNP Calling/Variant Detection in CDS's which are involved in some AMR associated cellular function.

Comparative Genomics: Prediction

- **Determining Biological Mechanism Driving Resistance:**
  - For each variant that we have determined to confer resistance we must perform a literature search to determine the biological mechanism

- **Use Newly Discovered Associations to Predict Resistance:**
  - Using our (potential) discoveries of new features or variants which confer resistance we can make more accurate predictions of colistin resistance

Comparative Genomics: Prediction

- **Machine Learning**
  - PATRIC3
  - DeepARG
  - Back Propagating Artificial Neural Network

Comparative Genomics: Strategy