Universiteit Utrecht

Bacterial GWAS

Linking phenotype to genotype in bacteria

Aldert Zomer

Microbial Genomics 2024

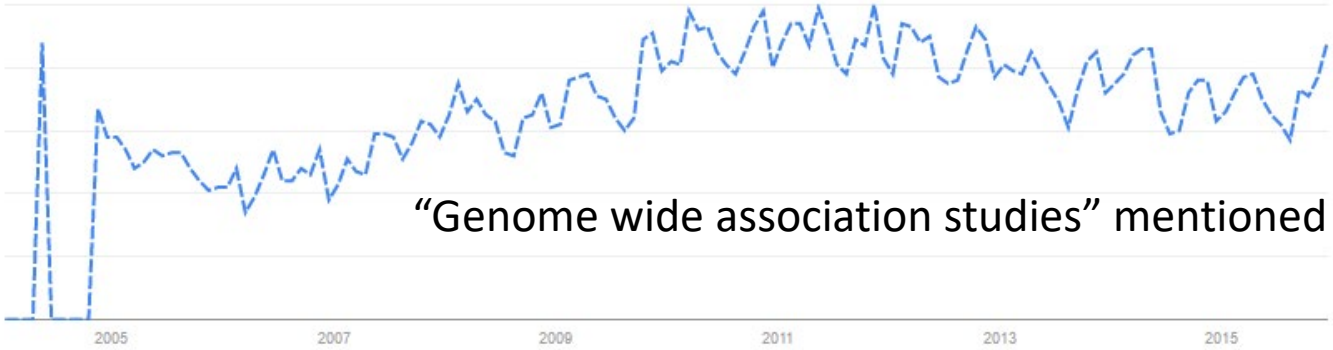
GWAS

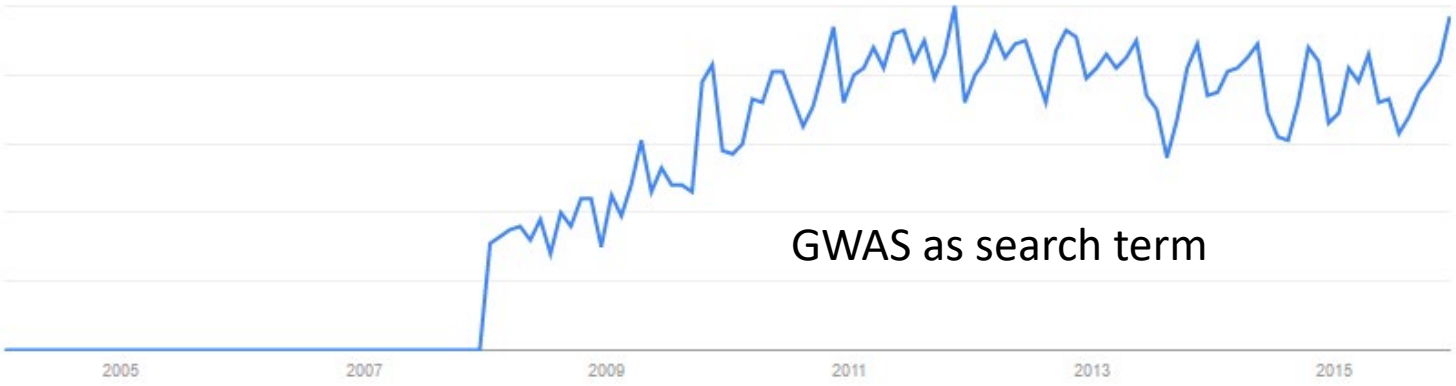
- **G**enome **W**ide **A**ssociation **S**tudies:
- Linking a phenotype to a genotype

- Phenotype: Trait, disease
- Genotype: (combinations of) Single Nucleotide Polymorphisms (SNPs), gene variants, complete genes

History

- HapMap Project (2002, 2005, 2007,2009)
- The 1000 Genomes Project (2008)





Google Trends

Basic idea

Genotype individuals/species/isolates for a large number of SNPs spread in a generally unspecified way throughout the genome. Look for association.

SNPs →

patients ↓

2	1	0	1	2	1	1	0	0	0	2	0	Control
0	1	1	0	1	2	0	1	0	0	2	1	Control
0	0	0	2	0	0	0	0	0	2	1	0	Control
0	1	1	2	1	0	1	1	1	1	2	2	Control
2	0	2	1	0	1	1	0	0	0	2	2	Control
1	1	2	1	2	2	0	1	0	0	1	1	Control
1	1	0	2	1	1	0	0	1	0	0	1	Control
0	0	1	0	2	1	0	1	2	0	1	1	Case
0	2	2	0	0	1	1	1	2	1	0	0	Case
0	0	0	2	0	2	2	0	2	2	1	2	Case
0	1	1	0	0	0	1	1	2	2	1	0	Case
2	0	2	1	1	2	2	0	2	0	2	2	Case
1	2	0	1	2	0	0	0	2	1	1	2	Case
1	1	0	0	2	2	2	0	2	0	2	0	Case

What do you see in the table? (hint: diploid)

Basic idea

Genotype individuals/species/isolates for a large number of SNPs spread in a generally unspecified way throughout the genome. Look for association.

SNPs →

patients ↓

2	1	0	1	2	1	1	0	0	0	2	0	Control
0	1	1	0	1	2	0	1	0	0	2	1	Control
0	0	0	2	0	0	0	0	0	2	1	0	Control
0	1	1	2	1	0	1	1	1	1	2	2	Control
2	0	2	1	0	1	1	0	0	0	2	2	Control
1	1	2	1	2	2	0	1	0	0	1	1	Control
1	1	0	2	1	1	0	0	1	0	0	1	Control
0	0	1	0	2	1	0	1	2	0	1	1	Case
0	2	2	0	0	1	1	1	2	1	0	0	Case
0	0	0	2	0	2	2	0	2	2	1	2	Case
0	1	1	0	0	0	1	1	2	2	1	0	Case
2	0	2	1	1	2	2	0	2	0	2	2	Case
1	2	0	1	2	0	0	0	2	1	1	2	Case
1	1	0	0	2	2	2	0	2	0	2	0	Case

homozygous for mutation: associated with case

Basic idea (2)

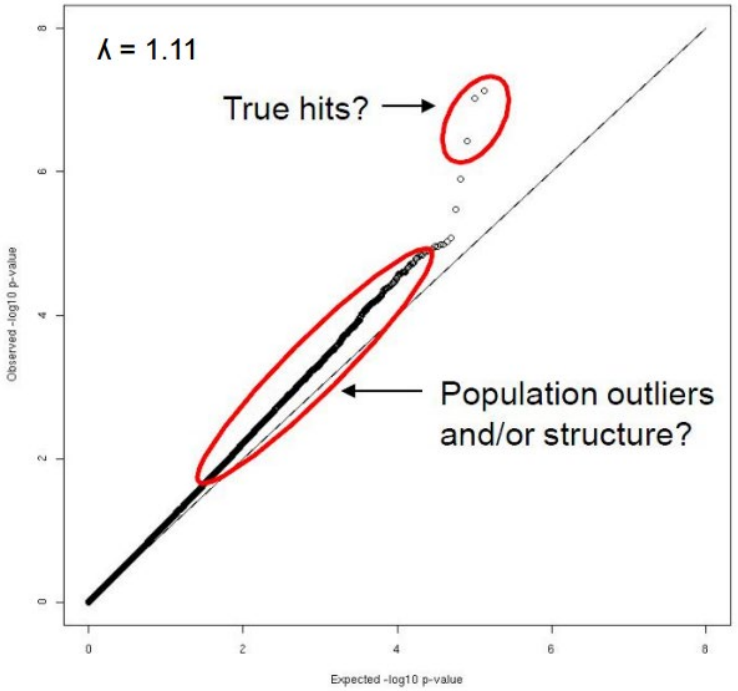
	SNP present	SNP absent
with phenotype	20	3
without phenotype	4	16

$p = 1.19*10^{-05}$

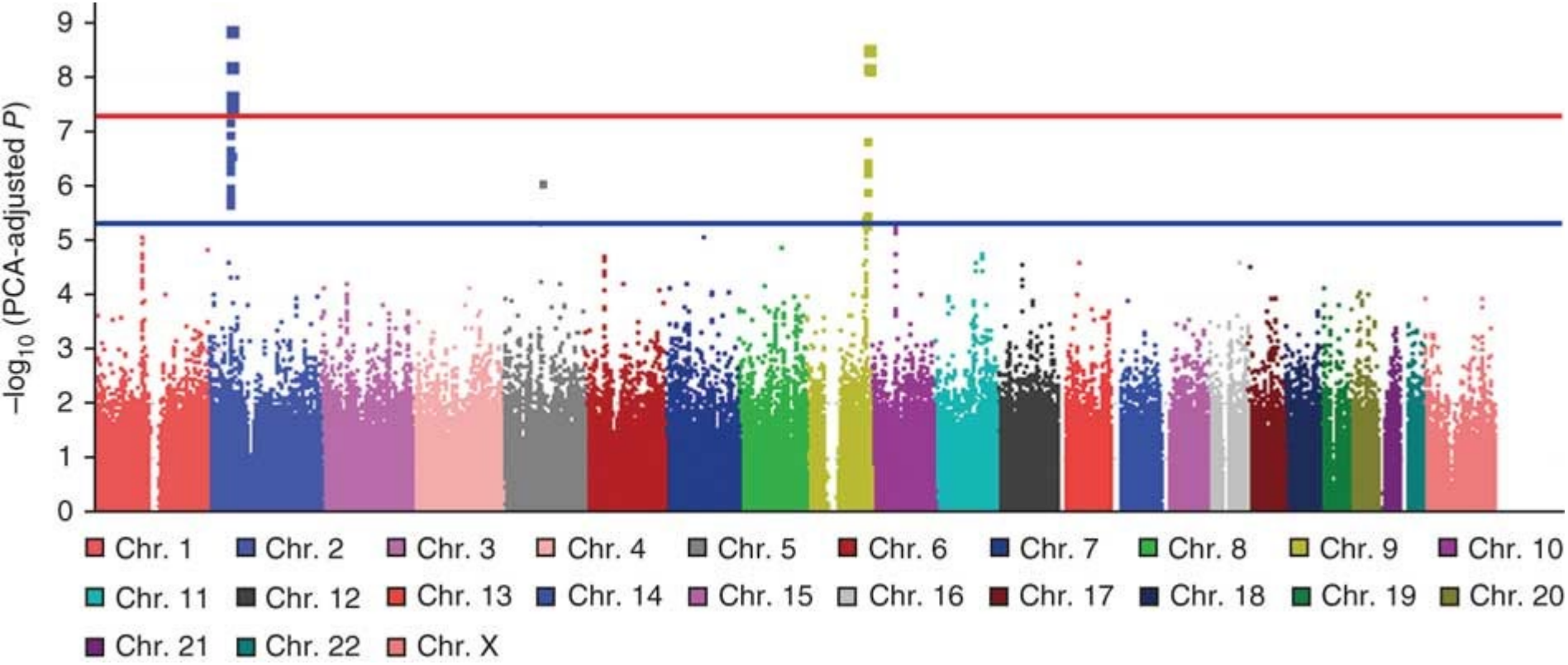
2x2 (or 3x2 in diploid genomes) contingency tests
e.g.
Fisher exact (small samplesizes, values <10)
Chi squared (large samplesizes, values >10)

QQ-plot:
Plot the expected p-values against the observed p-values

Strong deviations are likely candidates



Basic idea (3)



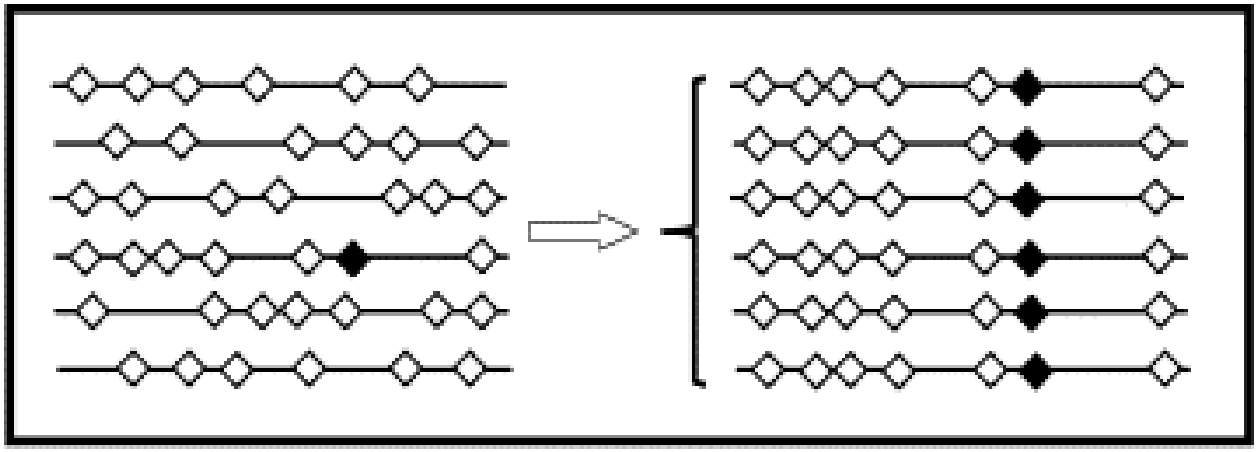
Negative log10 P-values plotted against location on genome: Manhattan plot

Population structure correction

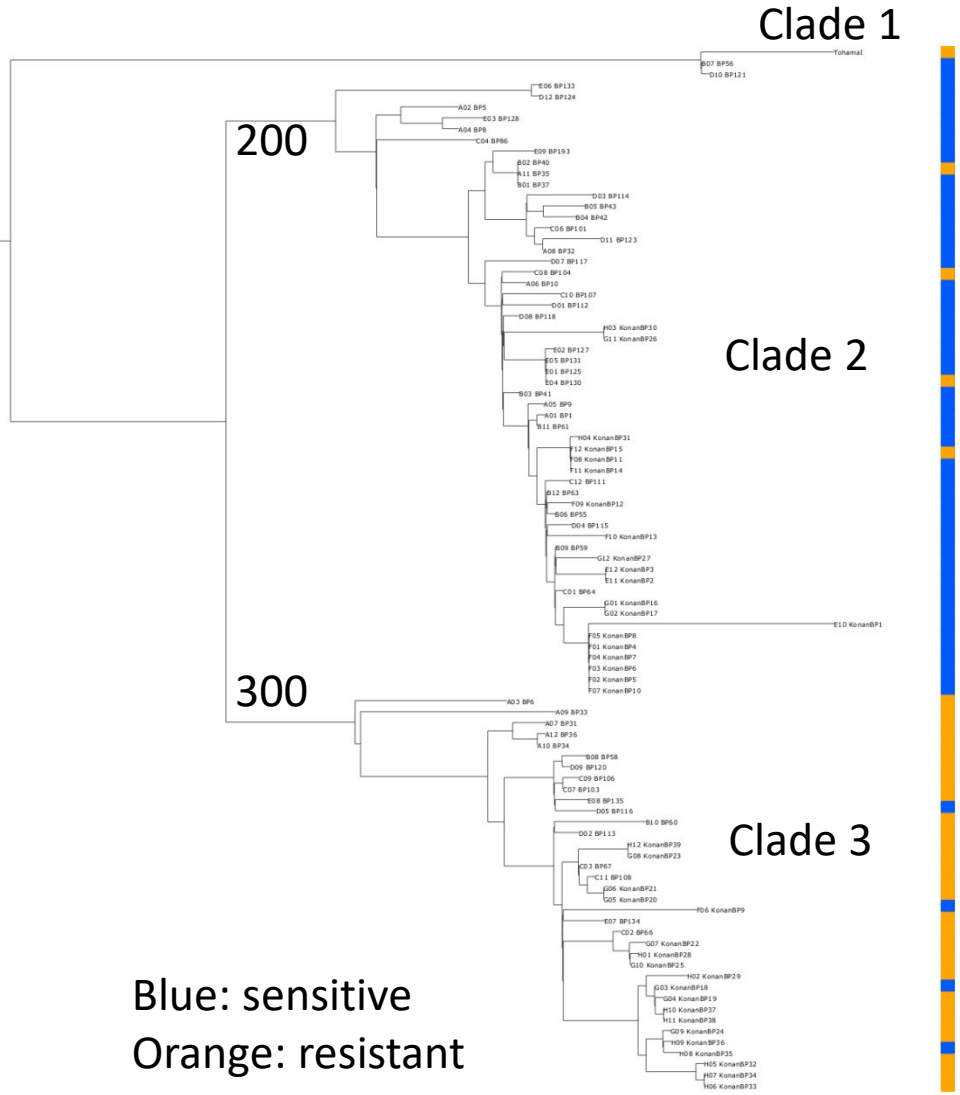
Population structure: Potentially a problem in human genetics.
A real problem in bacterial genetics

Population structure correction

- Population structure (in humans) occurs through mechanisms such as genetic drift, ancestral divergence and non-random mating
- Confounds GWAS: higher than expected allele frequencies within certain members of the study set
- Big problem in bacterial GWAS: haploid and only cell division. Genetic variants will be passed on to descendants and be in "linkage disequilibrium" with other mutations that occur in that lineage



Population structure correction



Blue: sensitive
Orange: resistant

Example:

Find the SNP associated with antimicrobial resistance

But.. Resistance against an antibiotic is primarily associated with a certain branch in the phylogenetic tree.

Standard contingency test will associate phylogenetic markers with resistance, 100s of SNPs (clade 3 defining SNPs) (Fisher Exact test in Scoary)

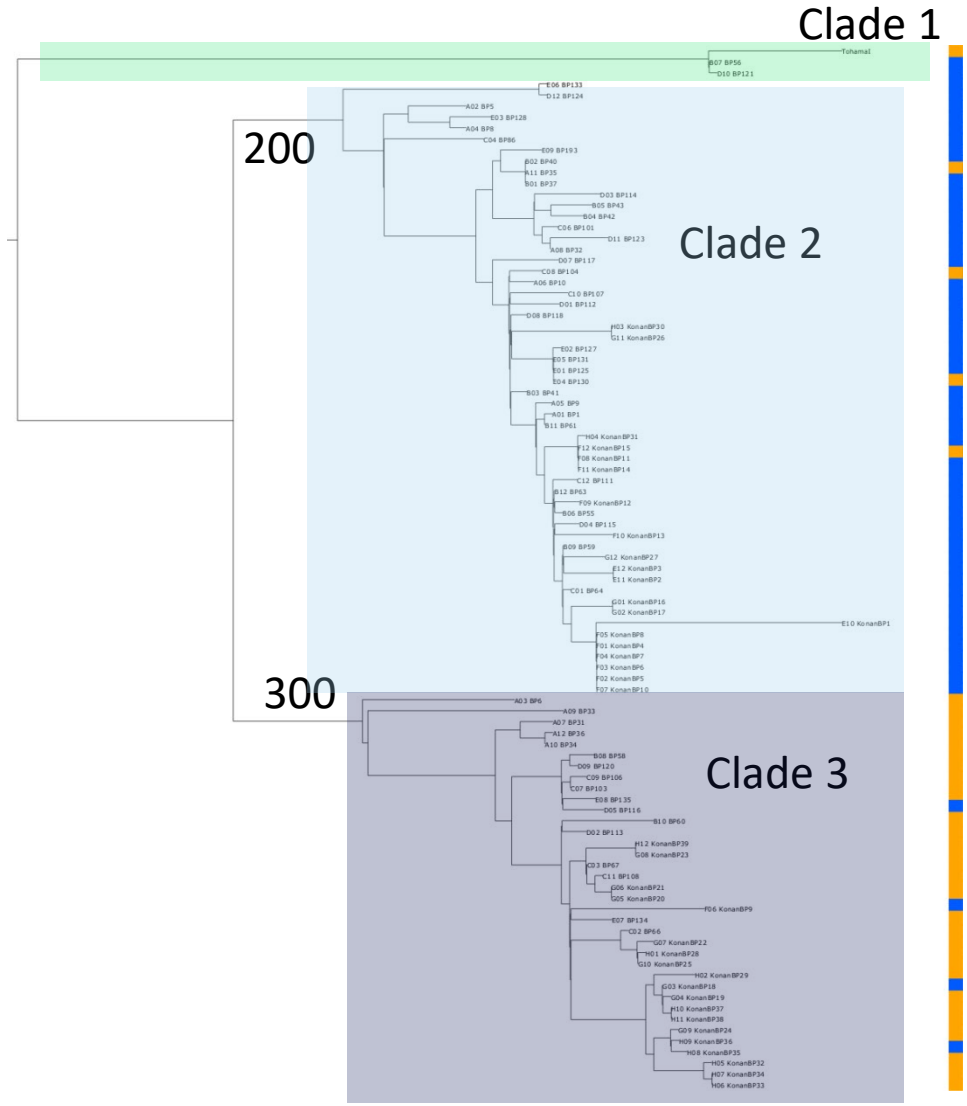
Population structure correction

Determine population groups:

- Pre-existing knowledge from e.g. MLST
- multi-dimensional scaling in PLINK
- principal component analysis in EIGENSTRAT
- Bayesian analysis of genetic population structure: BAPS
- Infer clones based on branch lengths in phylogenetic tree
- Many others..

Use the groups as covariates in association testing (e.g. with the Cochran-Mantel-Haenszel test)

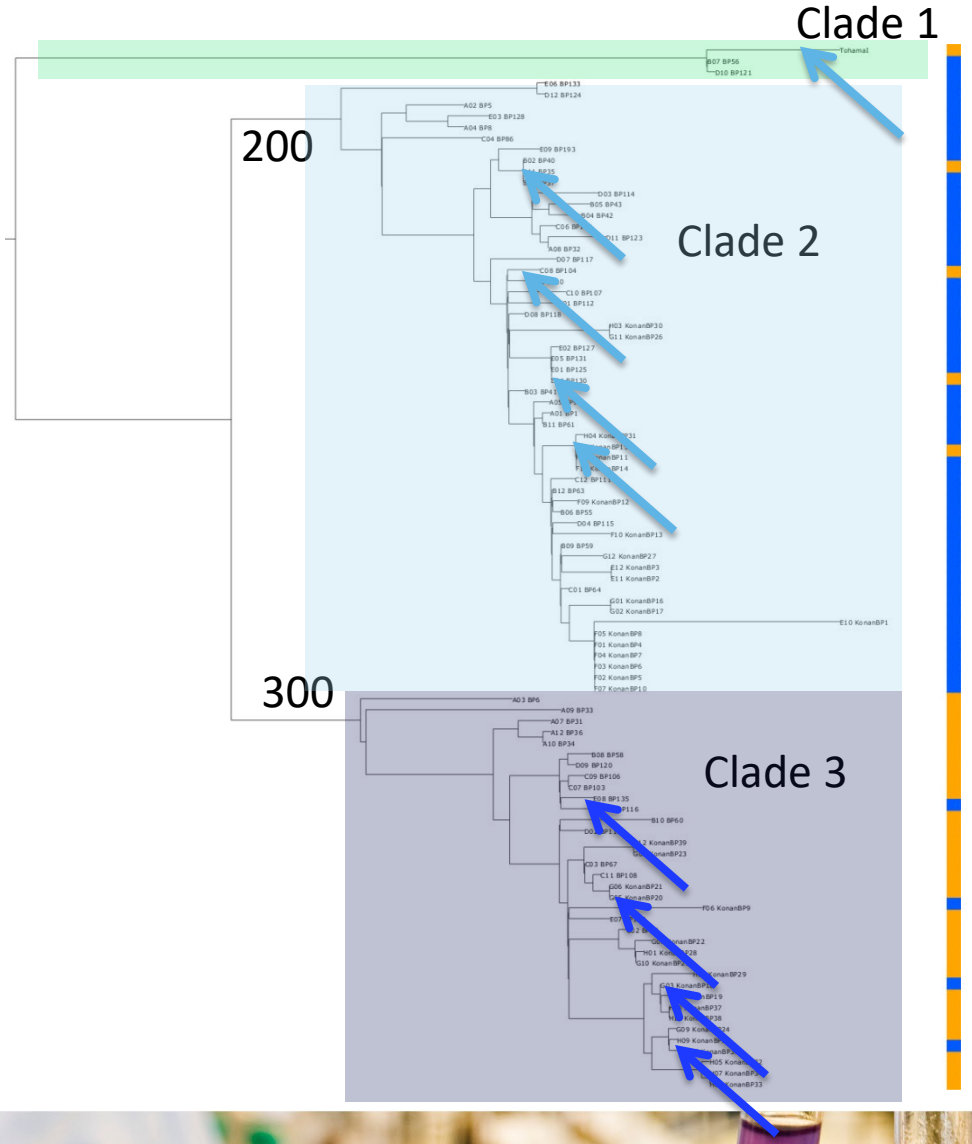
Population structure correction



Cochran-Mantel-Haenszel:

- Performs association testing per clade
- Computes a weighted p value

Population structure correction



Cochran-Mantel-Haenszel:

- Performs association testing per clade
- Computes a weighted p value

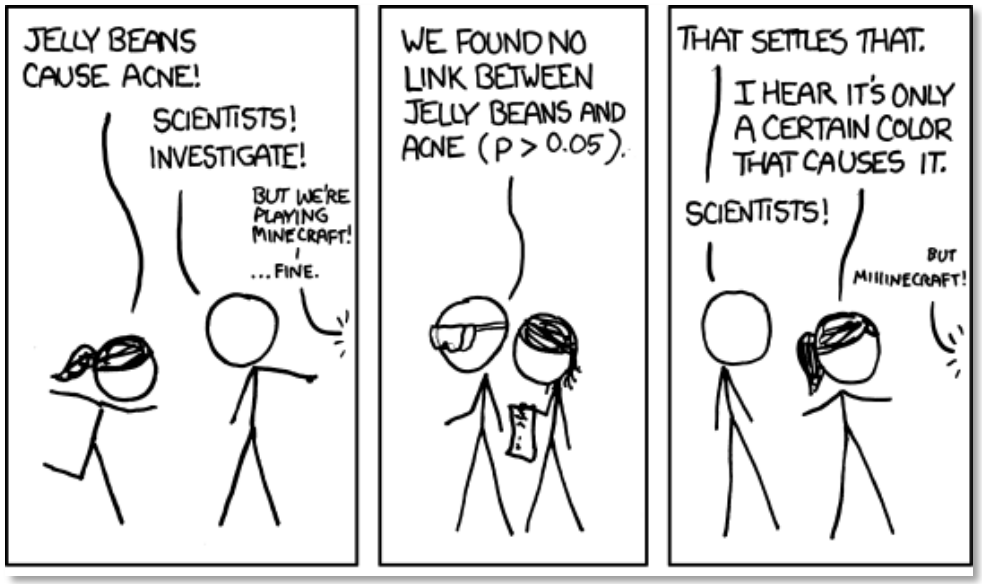
Alternatively:

Count repeated and independently emerged mutations occurring more often on branches of cases relative to controls
(PhyC: Farhat et al Nat Genet. 2013, this is implemented in Scoary)

Multiple testing correction

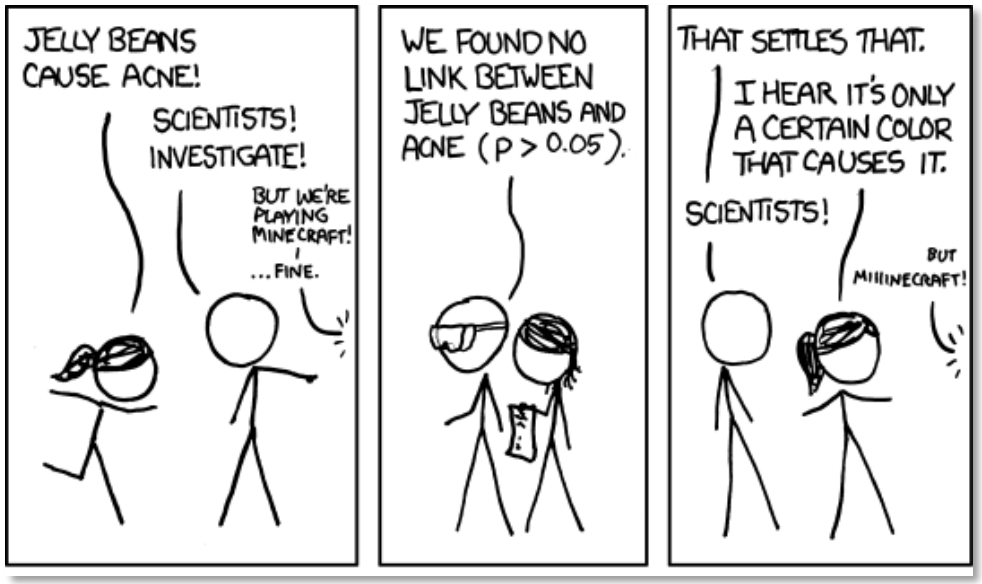
1000 SNPs have a p-value < 0.05. Are they all true positives?

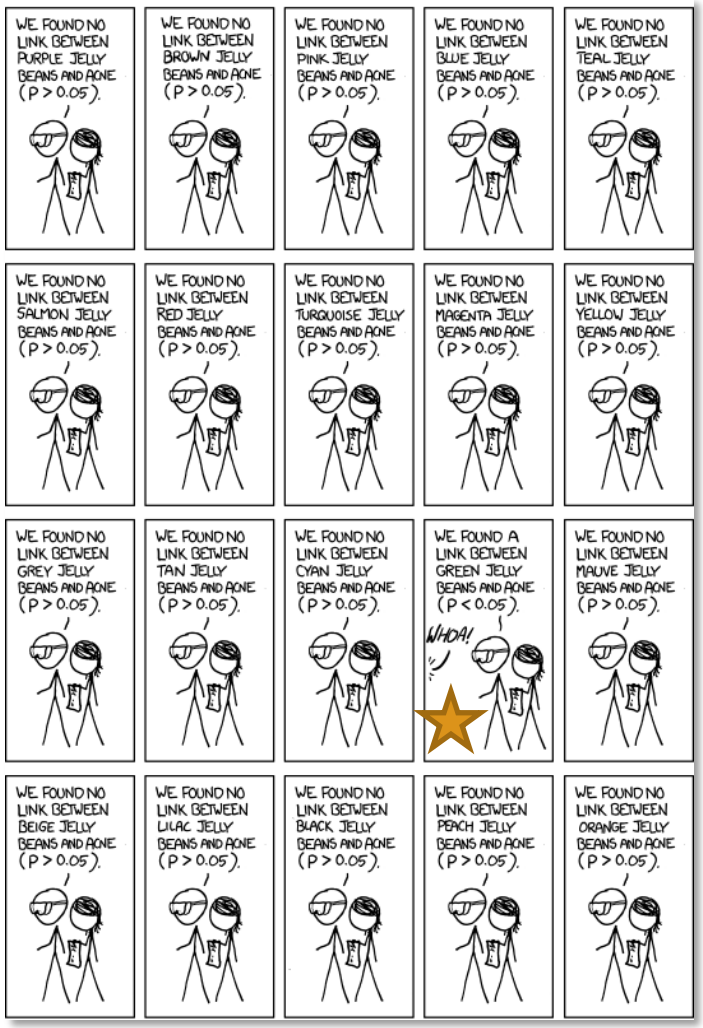
Multiple testing problem



https://xkcd.com/882/

Multiple testing problem





https://xkcd.com/882/

Multiple testing problem

https://xkcd.com/882/

Multiple testing: Adjusting

- Significance threshold must adjust for Type I error (a false positive); spurious statistical significance arising from multiple comparisons involving hundreds of thousands of SNPs

Dudbridge F, Gusnanto A (2008) Estimation of significance thresholds for genome-wide association scans. Genetic Epidemiology 32:227-34

Pe'er I, Yelensky R, Altshuler D, Daly MJ, (2008) Estimation of the multiple testing burden for genome-wide association studies of nearly all common variants. Genetic Epidemiology, May;32(4):381-5

Multiple testing: Adjusting

- Bonferroni correction
- Benjamini Hochberg (false discovery rate, FDR, in Scoary) or Storey Tibshirani (newer method)
- Permutation – computationally demanding (in Scoary)
- Bayesian approaches - computationally demanding

Multiple testing: Adjusting

- Easiest is **Bonferroni** correction. The conventional level of p (0.05) is divided by the number of tests performed (e.g. 0.05/100,000).

- Computationally simple. Low chance of false positives, but too stringent?

"Bonferroni adjustments are, at best, unnecessary and, at worst, deleterious to sound statistical inference" Perneger (1998)

Multiple testing: Adjusting

- FDR
- Or ≤ pm
- The r wh o (j/
- De ant

Rank	p value	calculation	adj. p (q)
1	0.0008	0.0008 * 11 / 1 =	0.0088
2	0.009	0.009 * 11 / 2 =	0.0495
3	0.02	0.02 * 11 / 3 =	0.073333
4	0.205	0.205 * 11 / 4 =	0.56375
5	0.396	0.396 * 11 / 5 =	0.8712
6	0.45	0.45 * 11 / 6 =	0.825
11	1	1 * 11 / 11 =	1

If SNP X has a q-value of 0.0495 it means that 4.95% of genes that show p values at least as small as SNP X are false positives

$$P(j) = \frac{}{m}$$

Bacterial GWAS - recap

- Gene level (accessory genome)
  - Predict all genes in genomes
  - Predict orthologs of the genes
  - Associate gene presence/absence with phenotype

- SNP level (primarily core genome)
  - Find all SNPs
  - Associate SNP with phenotype
  - SNP location reveals which gene is affected

- K-mer approach (core and accessory genome)
  - Find all posssible k-mers (ie 30 bp fragments)
  - Associate presence absence with phenotype
  - Map k-mer to reference genomes to identify genes

Bacterial GWAS - recap

- Population structure prediction using PCA, BAPS, others
- Use population structure ("clonality, MLST") as covariate in your statistical test
- Alternatively count repeated and independently emerged mutations occurring more often on branches of cases relative to controls: PhyC

Bacterial GWAS - recap

- Control the False Discovery Rate:
  - Bonferroni correction (very strict)
  - Benjamini Hochberg (FDR, often used)
  - Storey Tibshirani (newer FDR method)

Literature

The HAPMAP project
http://hapmap.ncbi.nlm.nih.gov/

The 1000 genomes project
http://www.1000genomes.org/

Linkage disequilibrium
http://www.nature.com/nrg/journal/v9/n6/full/nrg2361.html

Whole genome association analysis toolset
http://pngu.mgh.harvard.edu/~purcell/plink/

Eigenstrat/Eigensoft
http://genetics.med.harvard.edu/reich/Reich_Lab/Software.html

Explaining microbial phenotypes on a genomic scale: GWAS for microbes
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3743258/

High-throughput sequencing for the study of bacterial pathogen biology
http://www.sciencedirect.com/science/article/pii/S1369527414000708

The advent of genome-wide association studies for bacteria
http://www.ncbi.nlm.nih.gov/pubmed/25835153

Characterizing the genetic basis of bacterial phenotypes using genome-wide association studies: a new direction for bacteriology
http://www.genomemedicine.com/content/6/11/109

Genome-wide association mapping in bacteria?
http://www.ncbi.nlm.nih.gov/pubmed/16782339?dopt=Abstract

Estimation of significance thresholds for genome-wide association scans.
http://www.ncbi.nlm.nih.gov/pubmed/18300295

Estimation of the multiple testing burden for genome-wide association studies of nearly all common variants
http://www.ncbi.nlm.nih.gov/pubmed/18348202

Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes. https://www.ncbi.nlm.nih.gov/pubmed/27633831

What's wrong with Bonferroni adjustments
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1112991/

Statistical significance for genomewide studies
http://www.pnas.org/content/100/16/9440.full

A phylogeny-based sampling strategy and power calculator informs genome-wide associations study design for microbial pathogens
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4256898/

Genomic Analysis Identifies Targets of Convergent Positive Selection in Drug Resistant Mycobacterium tuberculosis
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3887553/

Hierarchical and spatially explicit clustering of DNA sequences with BAPS software.
http://www.ncbi.nlm.nih.gov/pubmed/23408797

A novel methodology for large-scale phylogeny partition.
http://www.ncbi.nlm.nih.gov/pubmed/21610724

Phage-Derived Protein Induces Increased Platelet Activation and Is Associated with Mortality in Patients with Invasive Pneumococcal Disease.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5241397/

Comprehensive identification of single nucleotide polymorphisms associated with beta-lactam resistance within pneumococcal mosaic genes.
http://www.ncbi.nlm.nih.gov/pubmed/25101644

Deciphering the distance to antibiotic resistance for the pneumococcus using genome sequencing data.
https://www.ncbi.nlm.nih.gov/pubmed/28205635