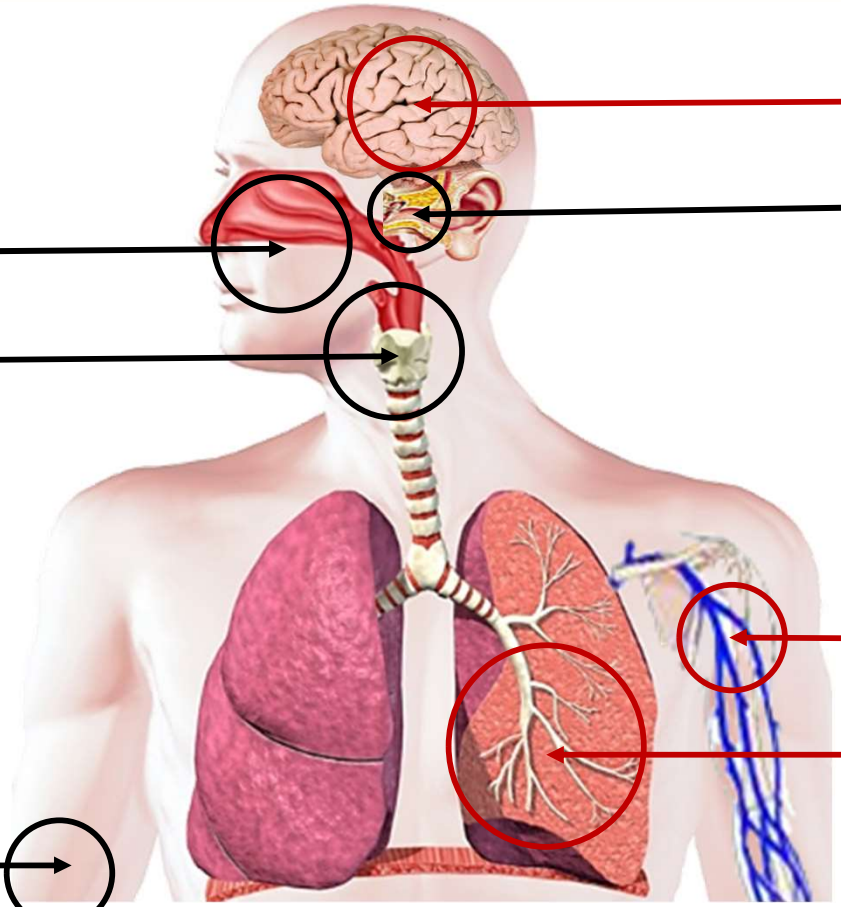
Introduction dataset – Day 1

Invasive disease caused by *Streptococcus pneumoniae* and patient mortality

Aldert Zomer – Bas Dutilh

Streptococcus pneumoniae



Meningitis

Otitis media

Sinusitis
mastoiditis

Bronchitis

Capsule

Vaccin
PCV7,10,13
2006

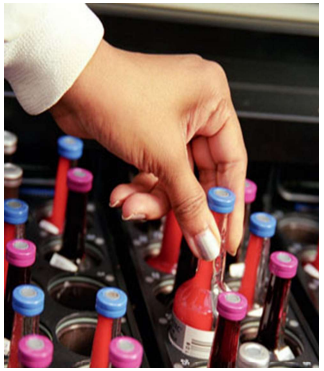
Bacteremia
Sepsis

Pneumonia

Septic arthritis

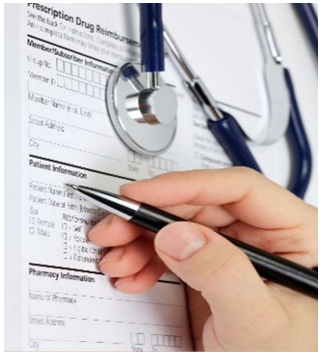
Cohort

S.pneumoniae



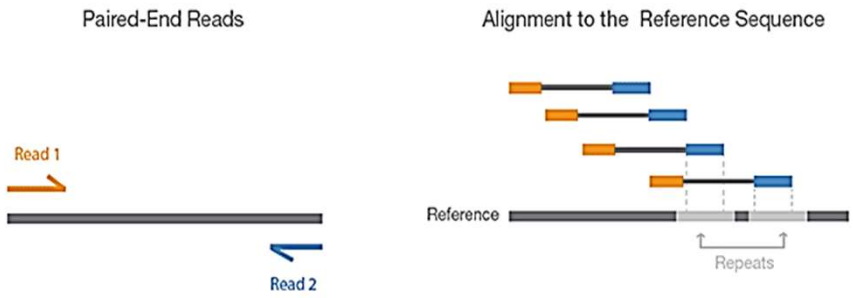
Blood culture +

Two Dutch hospitals `2001-2011`



Clinical data

Serotype

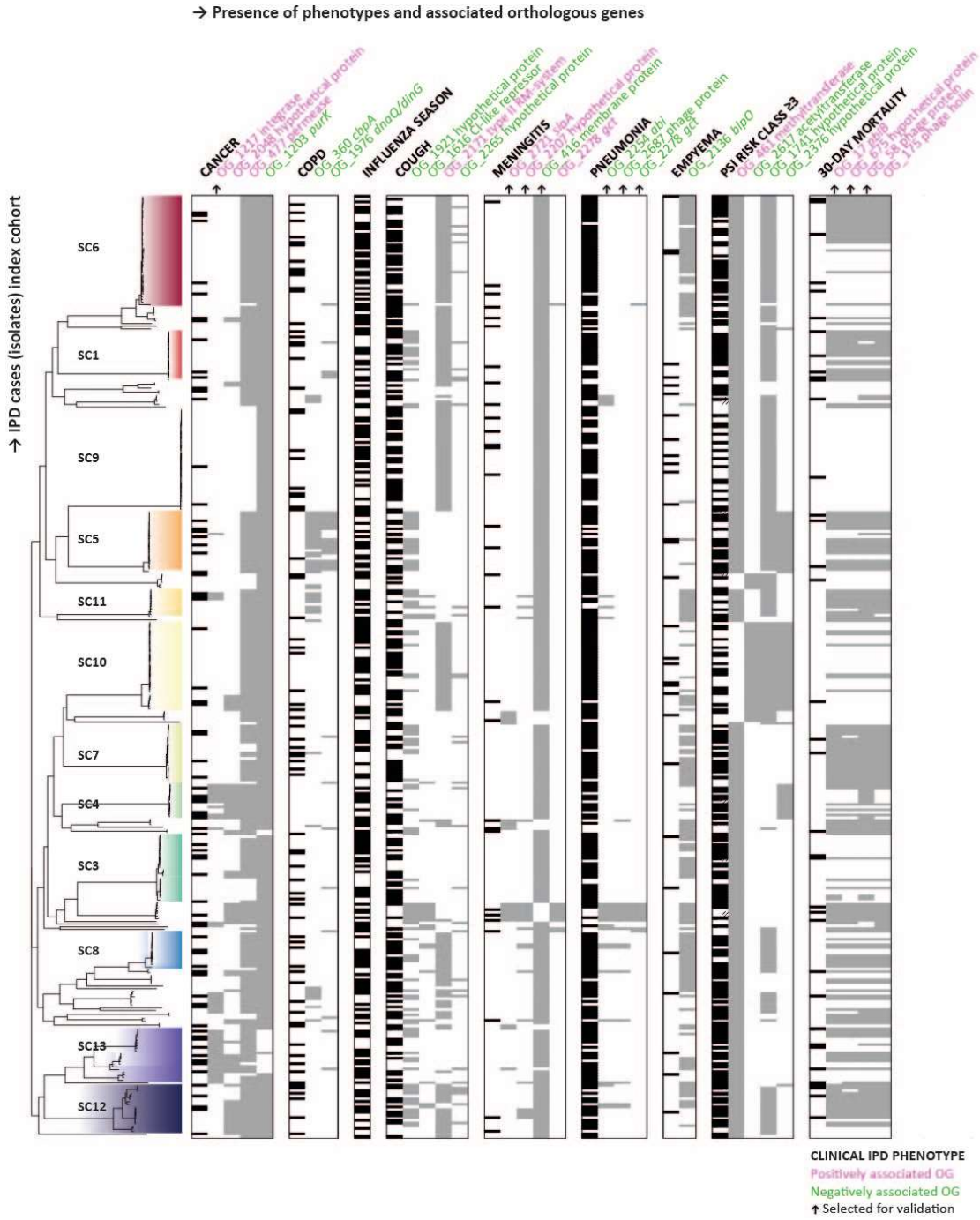


Illumina PE sequencing: Sanger Institute

Severity of disease

Are there specific bacterial clones that are associated with patient mortality?

What pneumococcal virulence genes are over-represented in the isolates from patients that died within 30 days of hospital admission and could be associated with patient mortality



Universiteit Utrecht

Bacterial GWAS

Linking phenotype to genotype in bacteria

Aldert Zomer

Microbial Genomics

2020 – Day 2

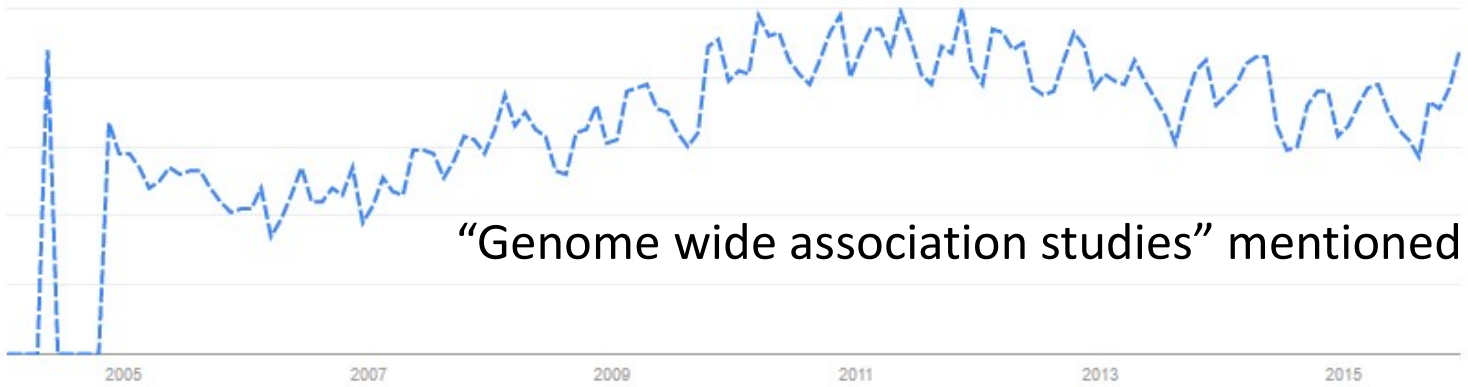
GWAS

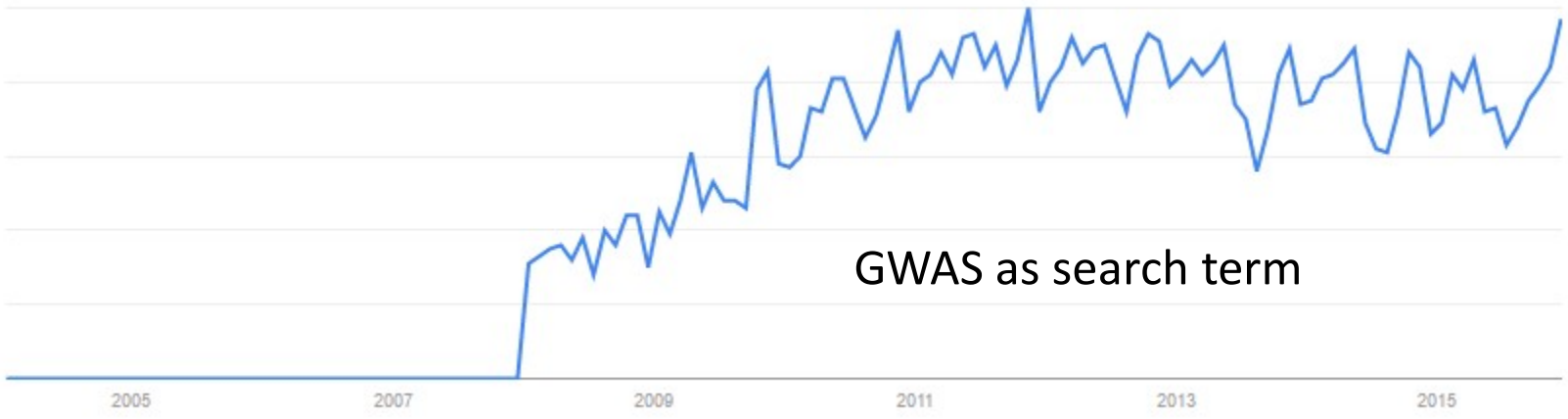
- **G**enome **W**ide **A**ssociation **S**tudies:
- Linking a phenotype to a genotype

- 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)

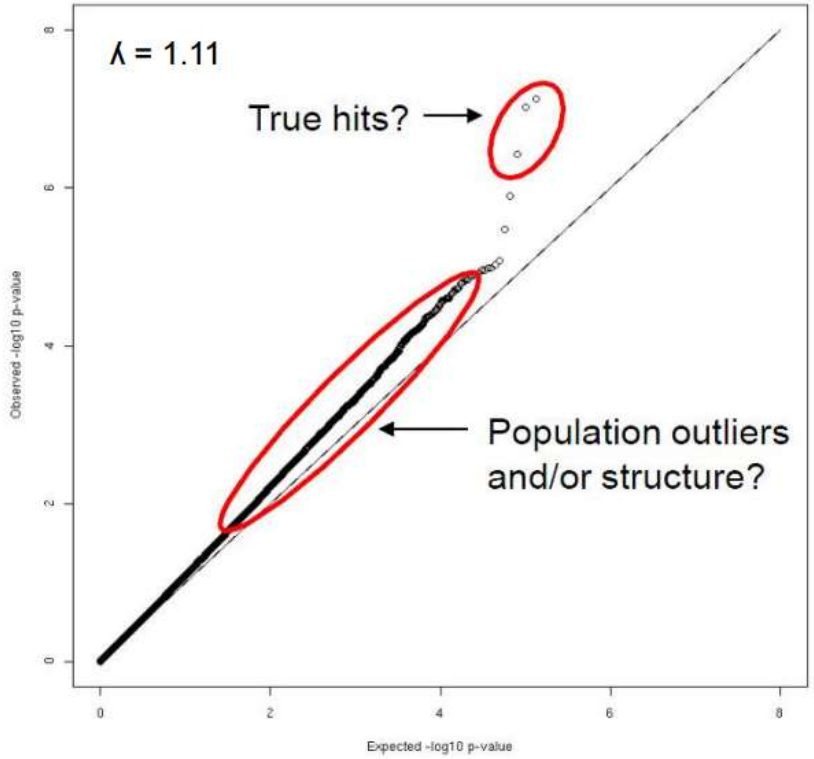
	SNP present	SNP absent
with phenotype	20	3
without phenotype	4	16

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

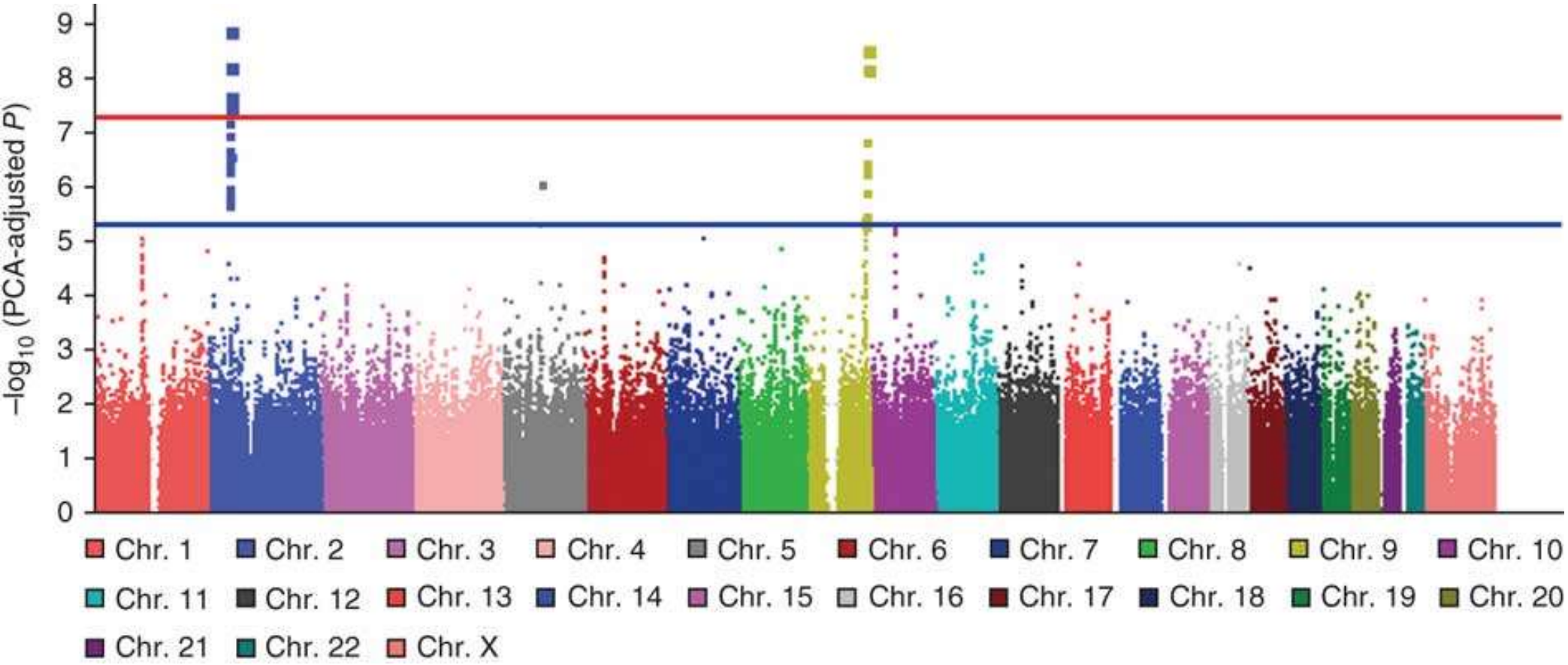
2x2 (or 3x2 in diploid genomes) contingency tests
e.g.
Fisher exact (small samplesizes, values <10)
Chi squared (large samplesizes, values >10)

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

Strong deviations are likely candidates

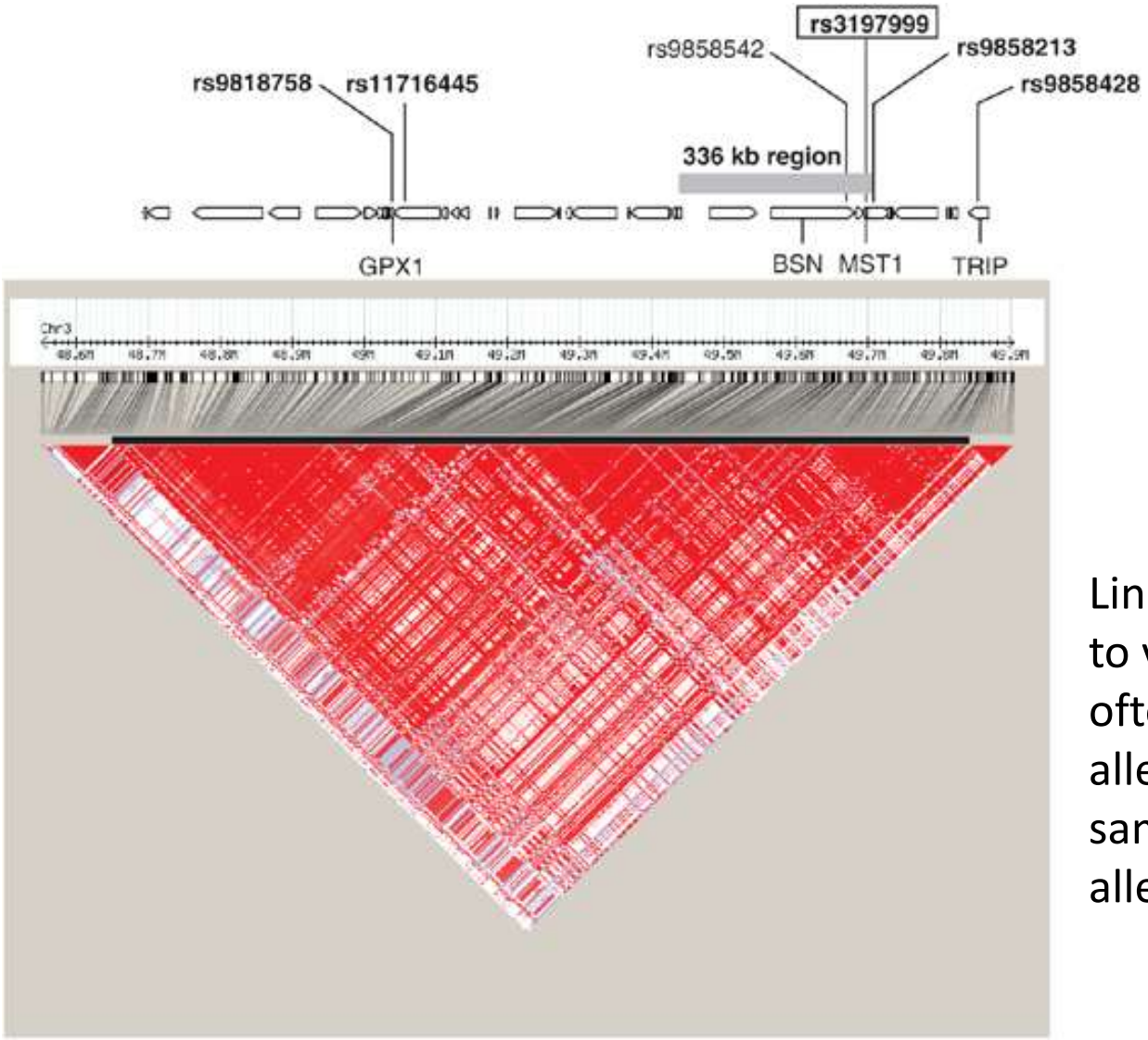


Basic idea (3)



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

Basic idea (4)



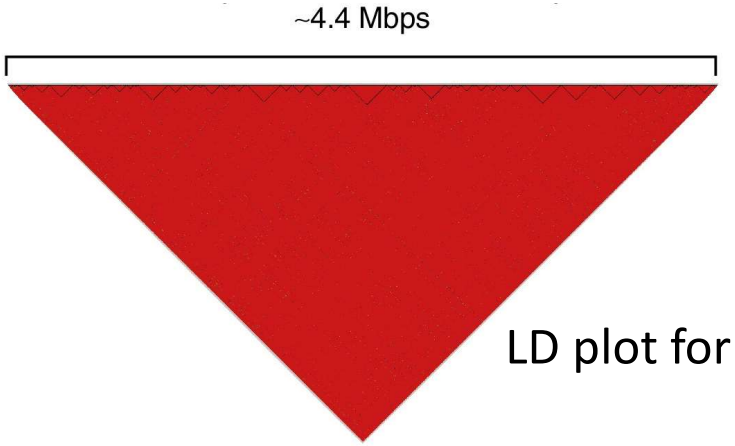
Linkage disequilibrium plots are used to visualize alleles that co-occur more often than what would be expected if alleles were independently, randomly sampled, based on their individual allele frequencies.

Population structure correction

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

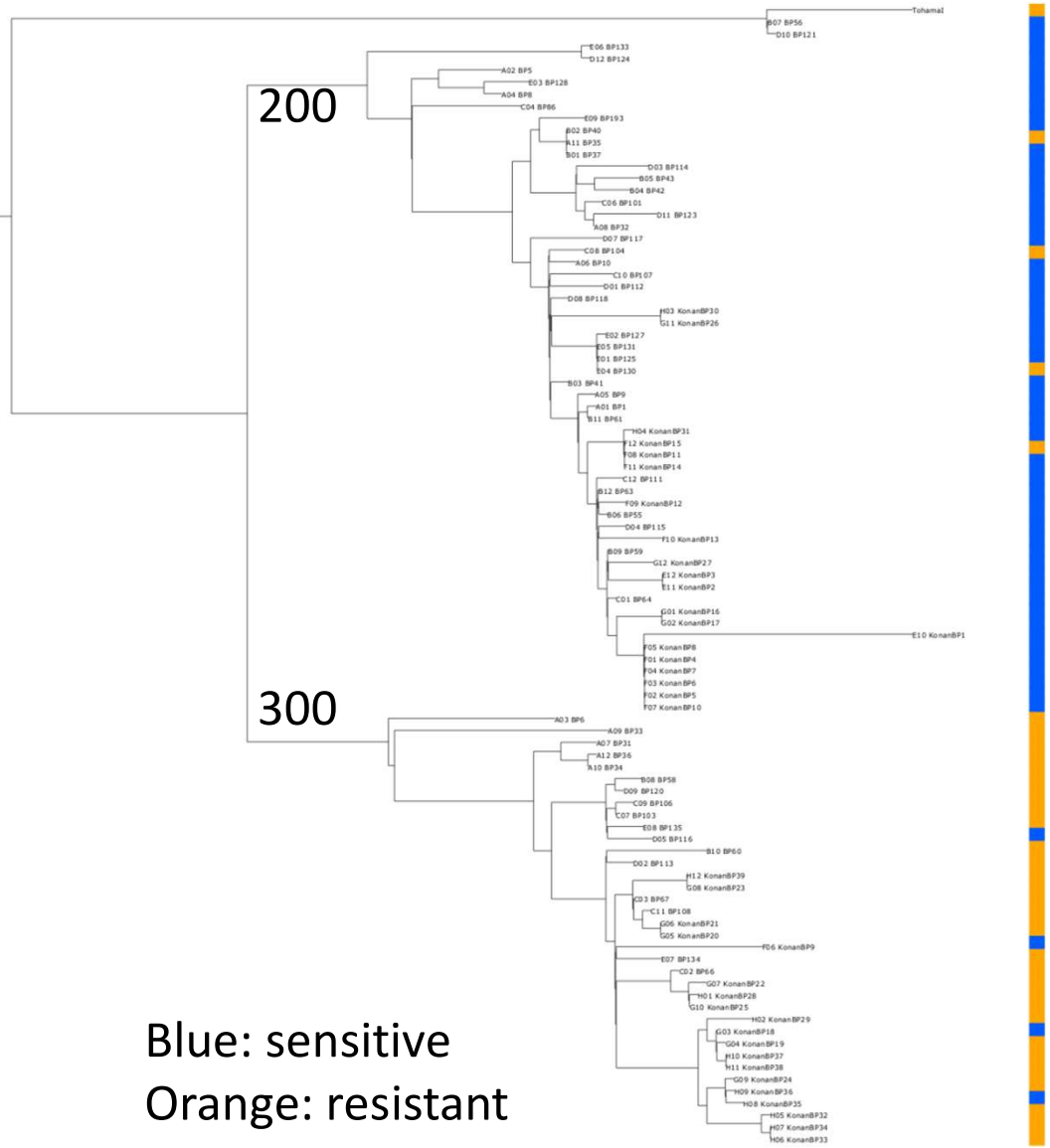
Population structure correction

- 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
- Problematic in bacterial GWAS: haploid and asexual. In the absence of recombination, all fixed genetic variants will be passed on to descendants and be in linkage disequilibrium with other mutations that occur in that lineage.



LD plot for *M. tuberculosis*.

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)

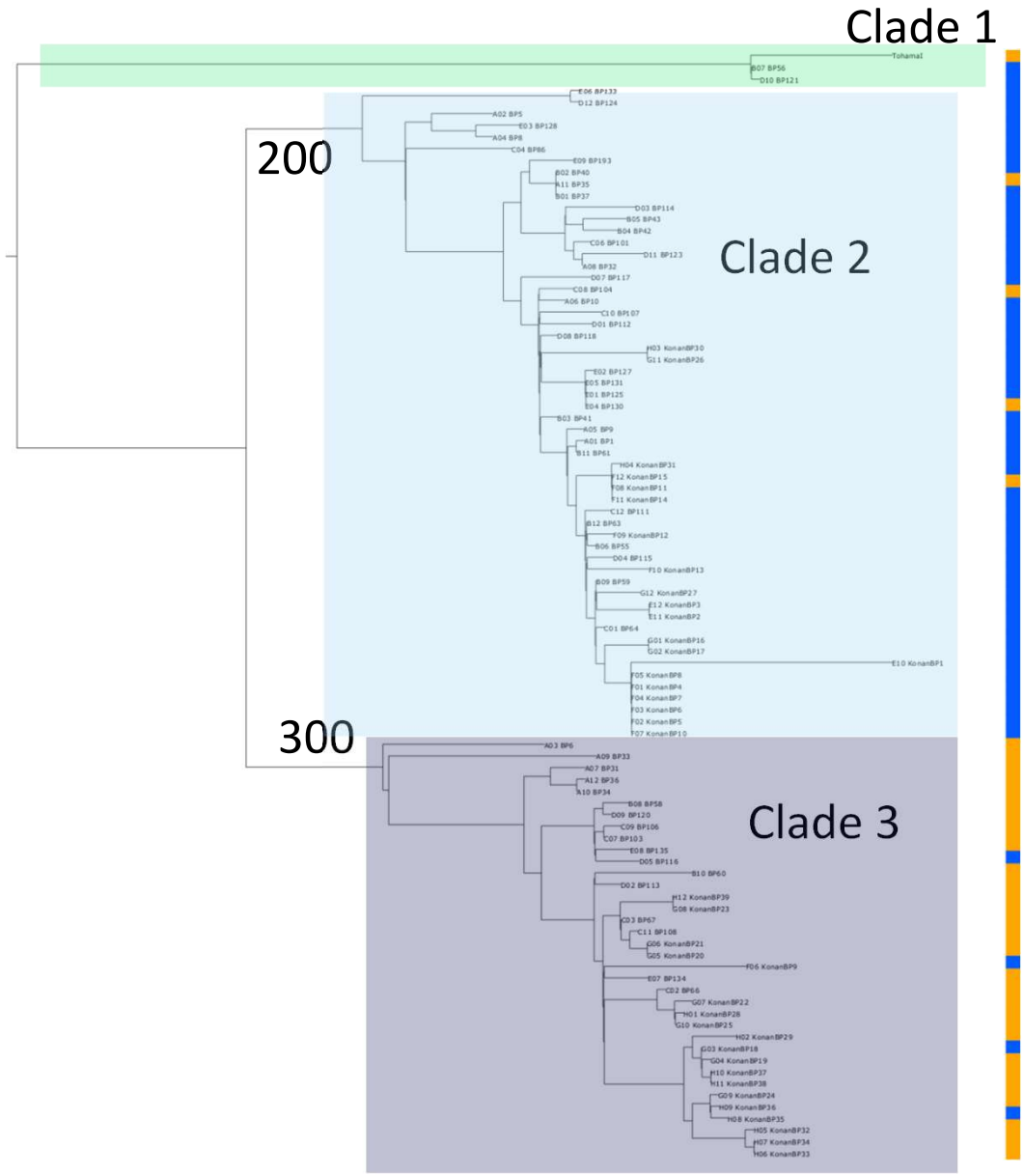
Population structure correction

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)

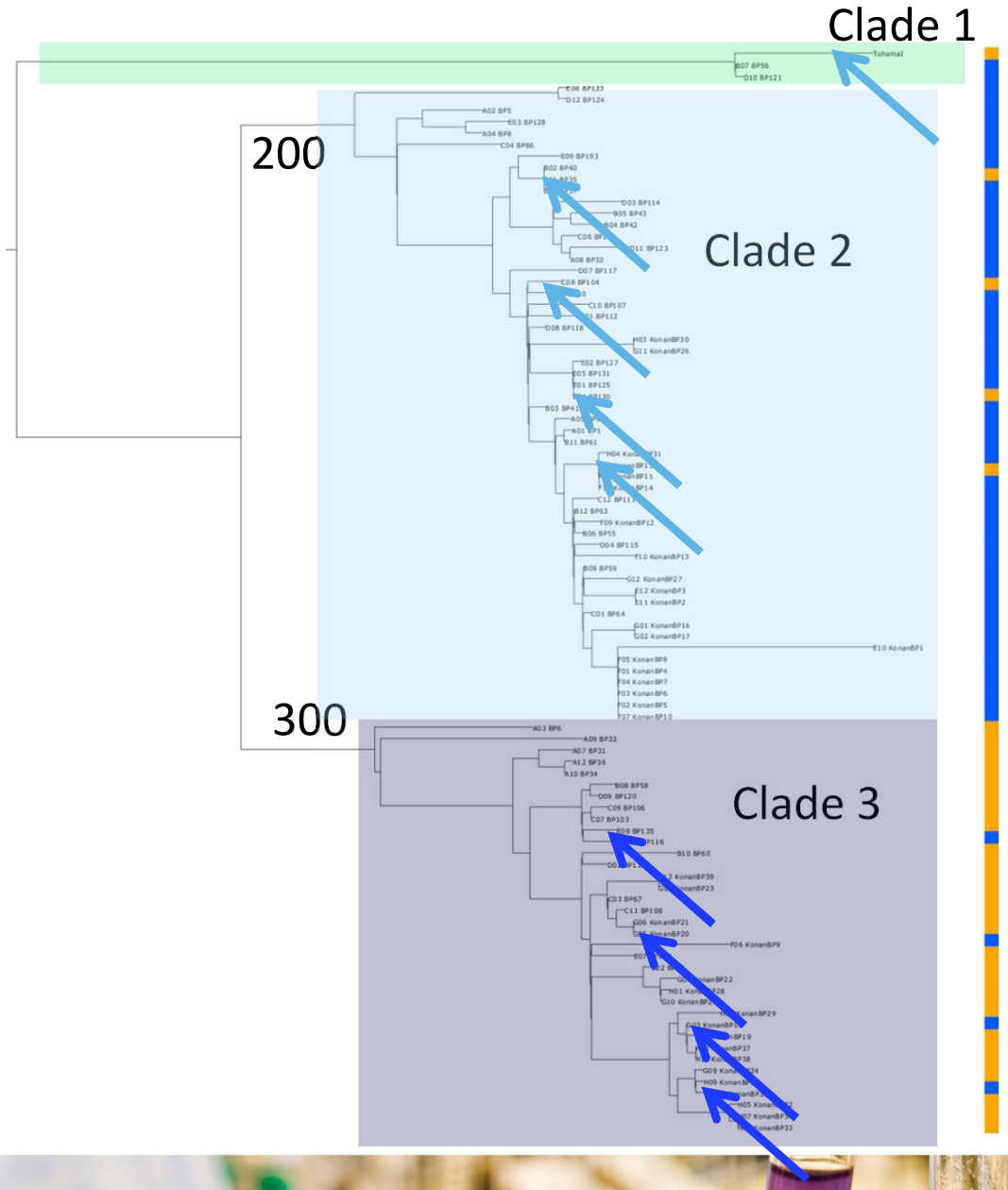
Population structure correction



Cochran-Mantel-Haenszel:

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

Population structure correction



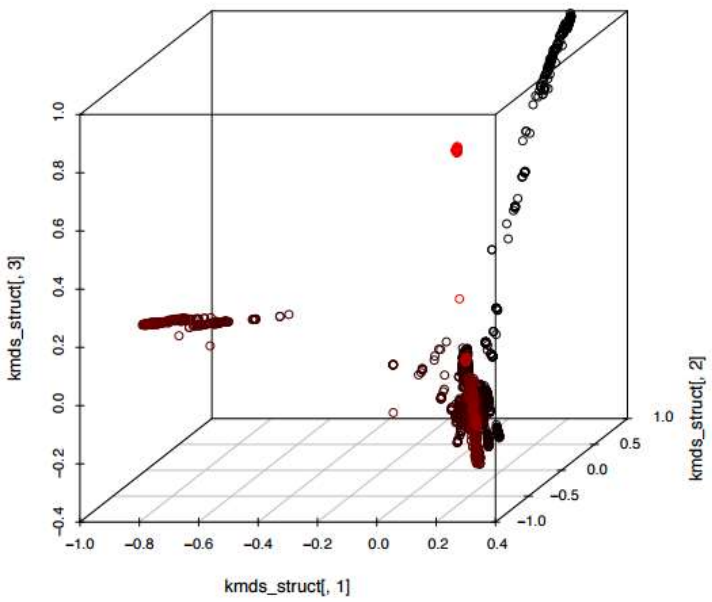
Cochran-Mantel-Haenszel:

- 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)
Implemented in Scoary

Newer methods



MENU

nature COMMUNICATIONS

Altmetric: 18 Views: 1,806 More detail »

Article | OPEN

Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes

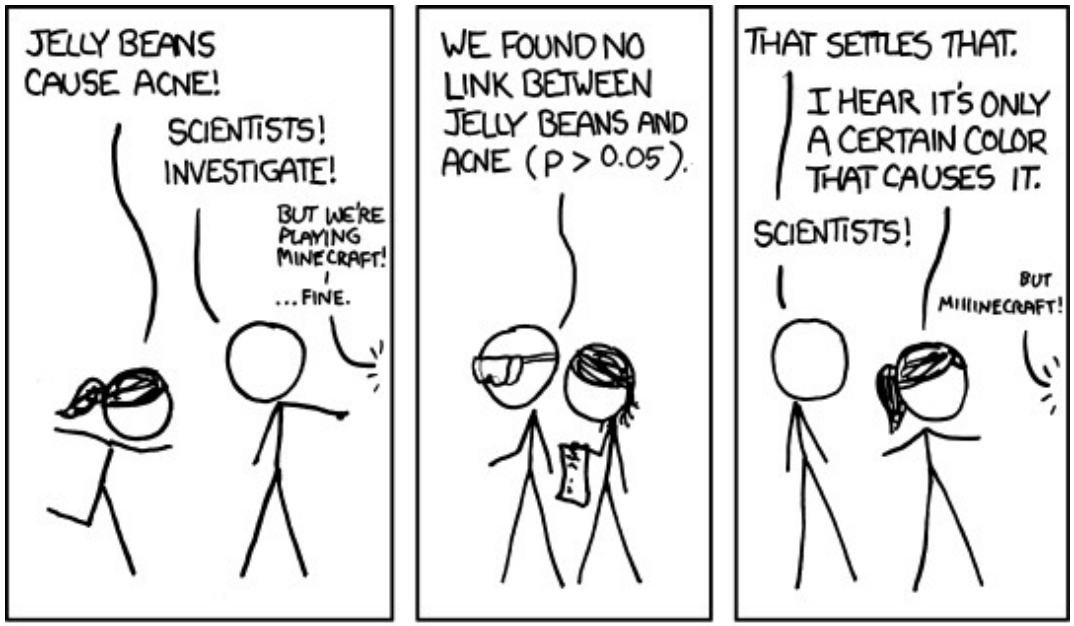
John A. Lees, Minna Vehkala, Niko Välimäki, Simon R. Harris, Claire Chewapreecha, Nicholas J. Croucher, Pekka Marttinen, Mark R. Davies, Andrew C. Steer, Steven Y. C. Tong, Antti Honkela, Julian Parkhill, Stephen D. Bentley & Jukka Corander

SEER: Also infers population structure automatically and uses position in an nMDS plot as cofactor. Nat Commun. 2016 Sep 16;7:12797

Multiple testing correction

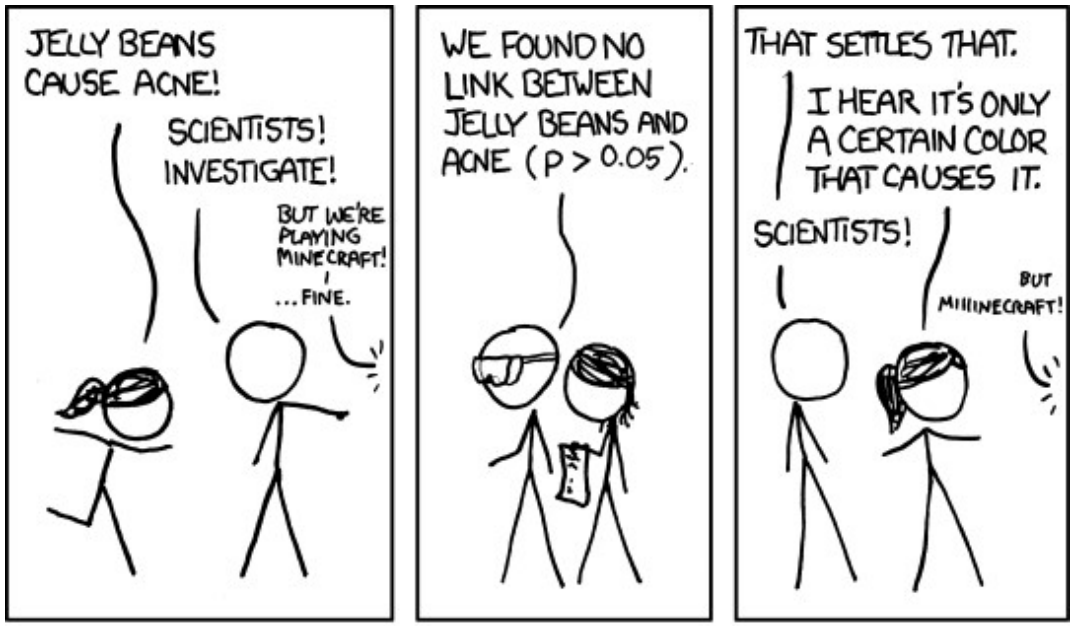
1000 SNPs have a p-value < 0.05. Are they all true positives?

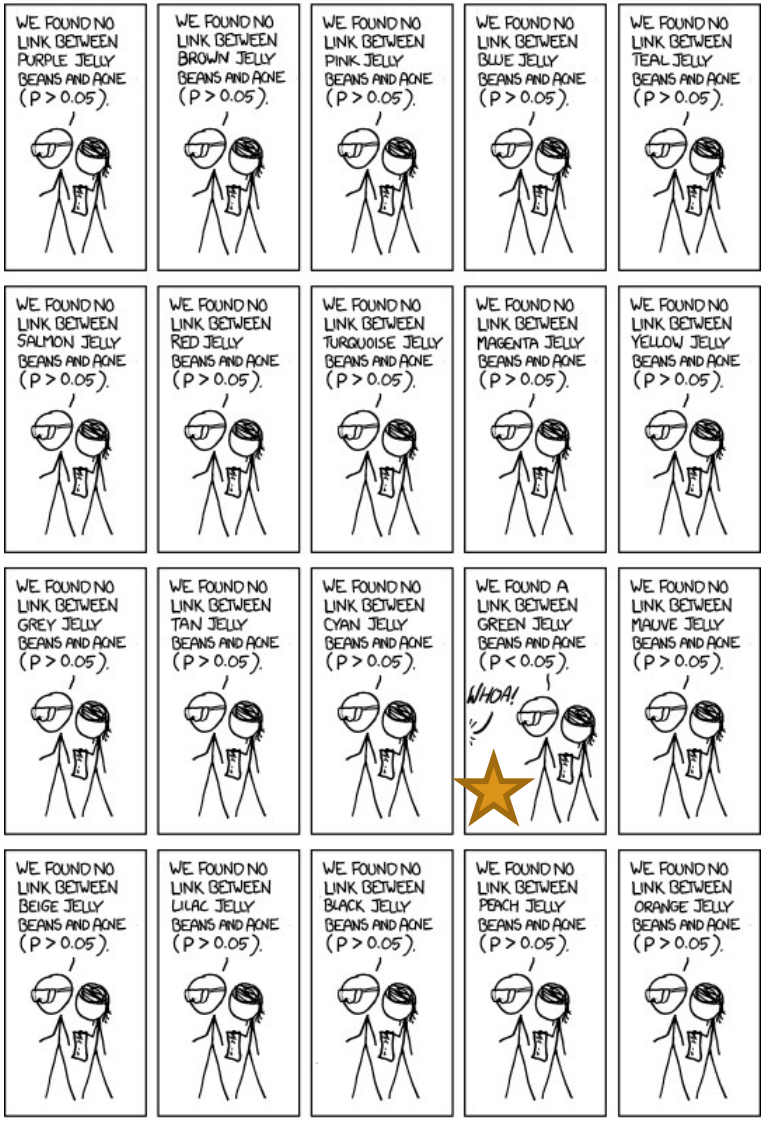
Multiple testing problem



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)
- Permutation – computationally demanding
- Bayesian approaches - computationally demanding

Severity of disease

Are there specific bacterial clones that are associated with patient mortality?

What pneumococcal virulence genes are over-represented in the isolates from patients that died within 30 days of hospital admission and could be associated with patient mortality