Universiteit Utrecht

Bacterial GWAS

Linking phenotype to genotype in bacteria

Aldert Zomer

Microbial Genomics 2024

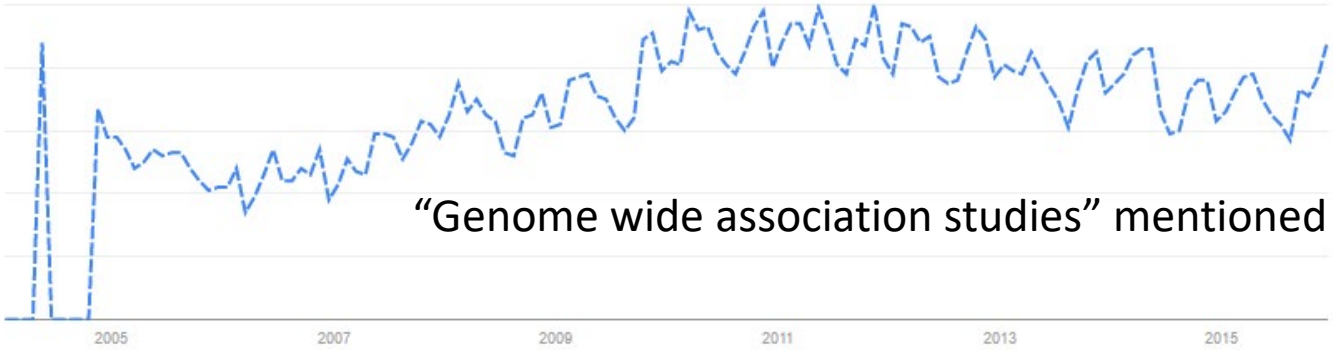
GWAS

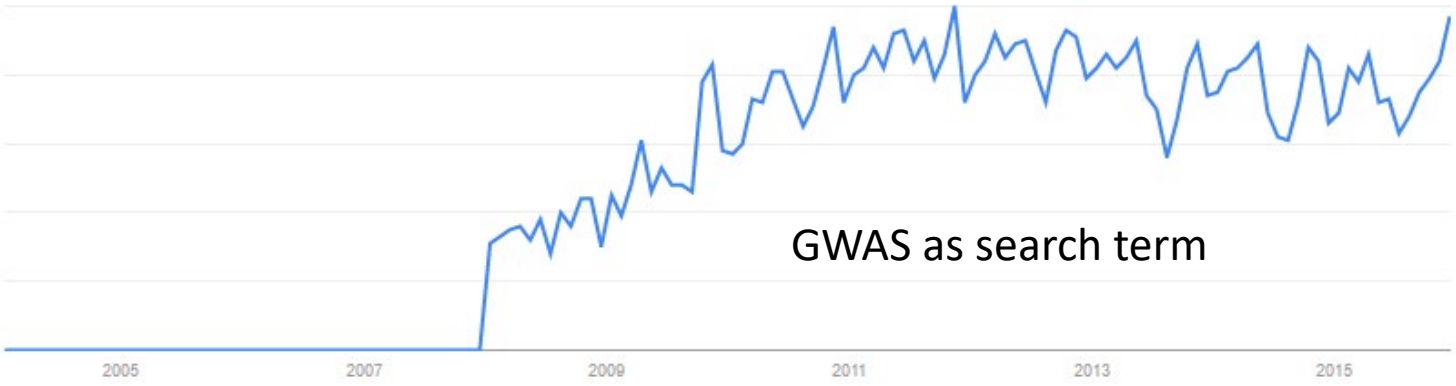
- **G**enome **W**ide **A**ssociation **S**tudies:
- Linking a phenotype to a genotype

- 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)

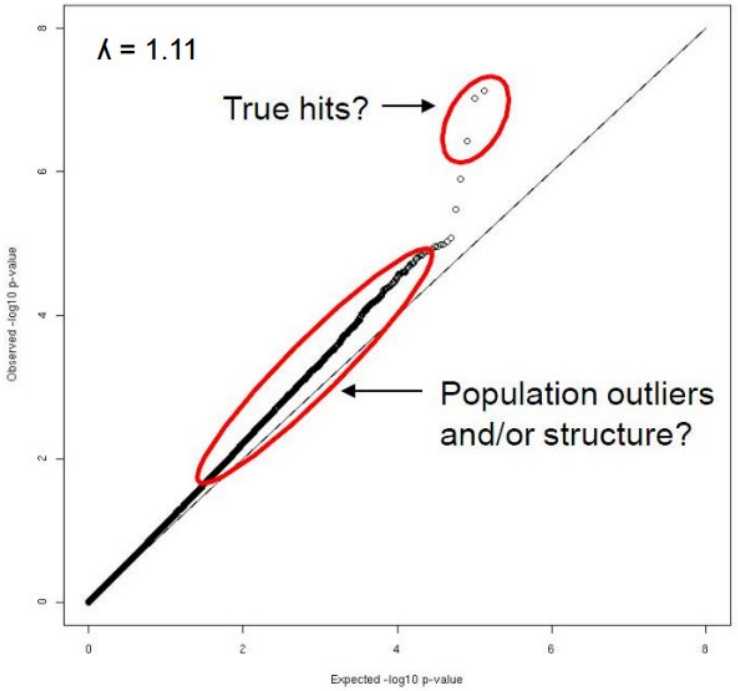
	SNP present	SNP absent
with phenotype	20	3
without phenotype	4	16

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

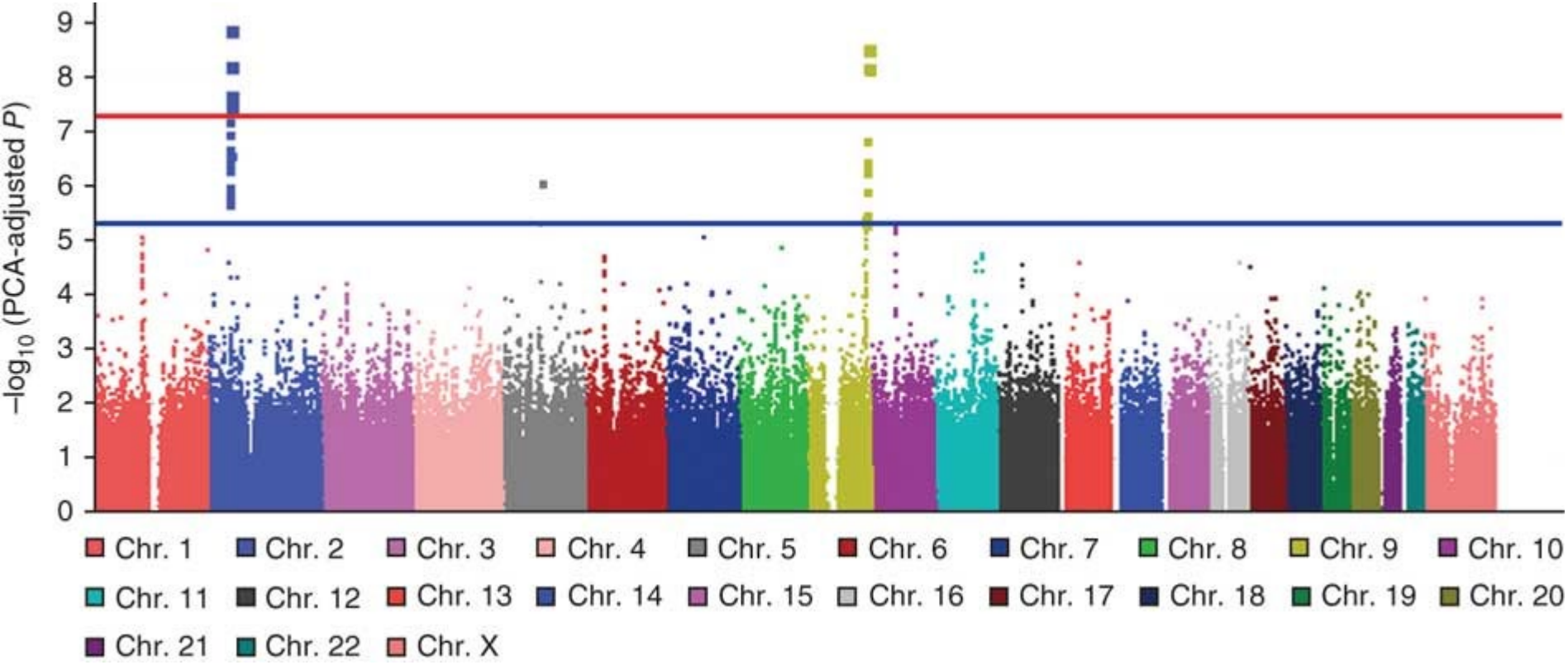
2x2 (or 3x2 in diploid genomes) contingency tests
e.g.
Fisher exact (small samplesizes, values <10)
Chi squared (large samplesizes, values >10)

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

Strong deviations are likely candidates



Basic idea (3)



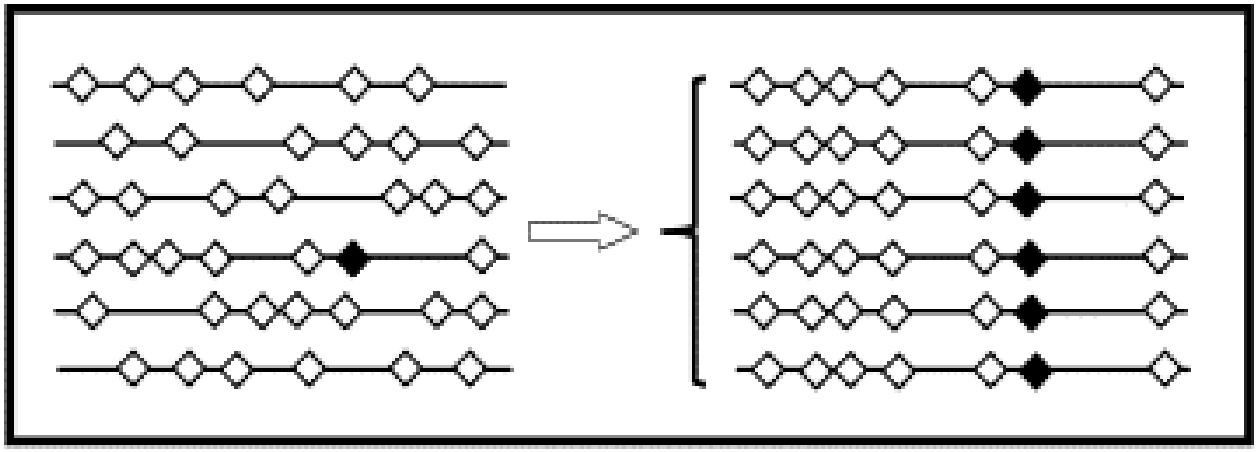
Negative log10 P-values plotted against location on genome: Manhattan plot

Population structure correction

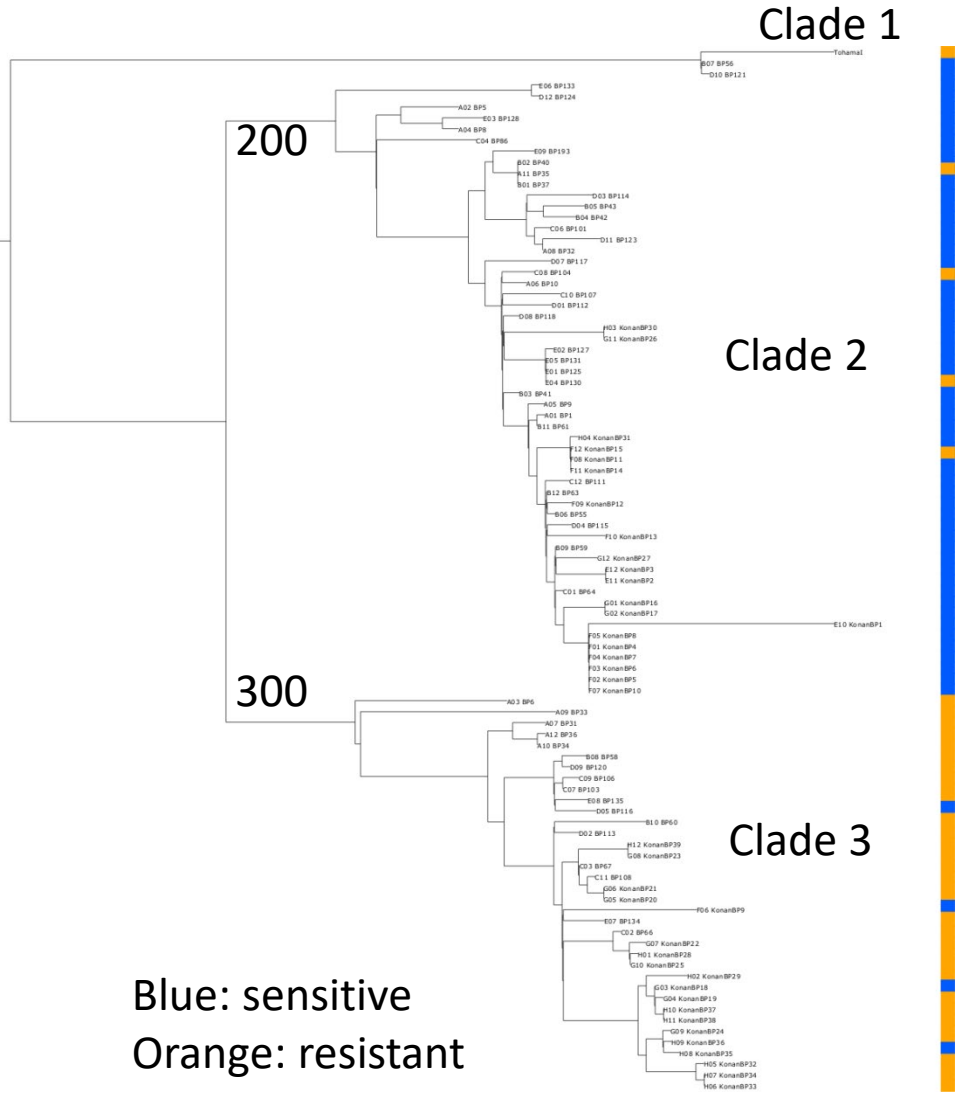
Population structure: Potentially a problem in human genetics.
A real problem in bacterial genetics

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)

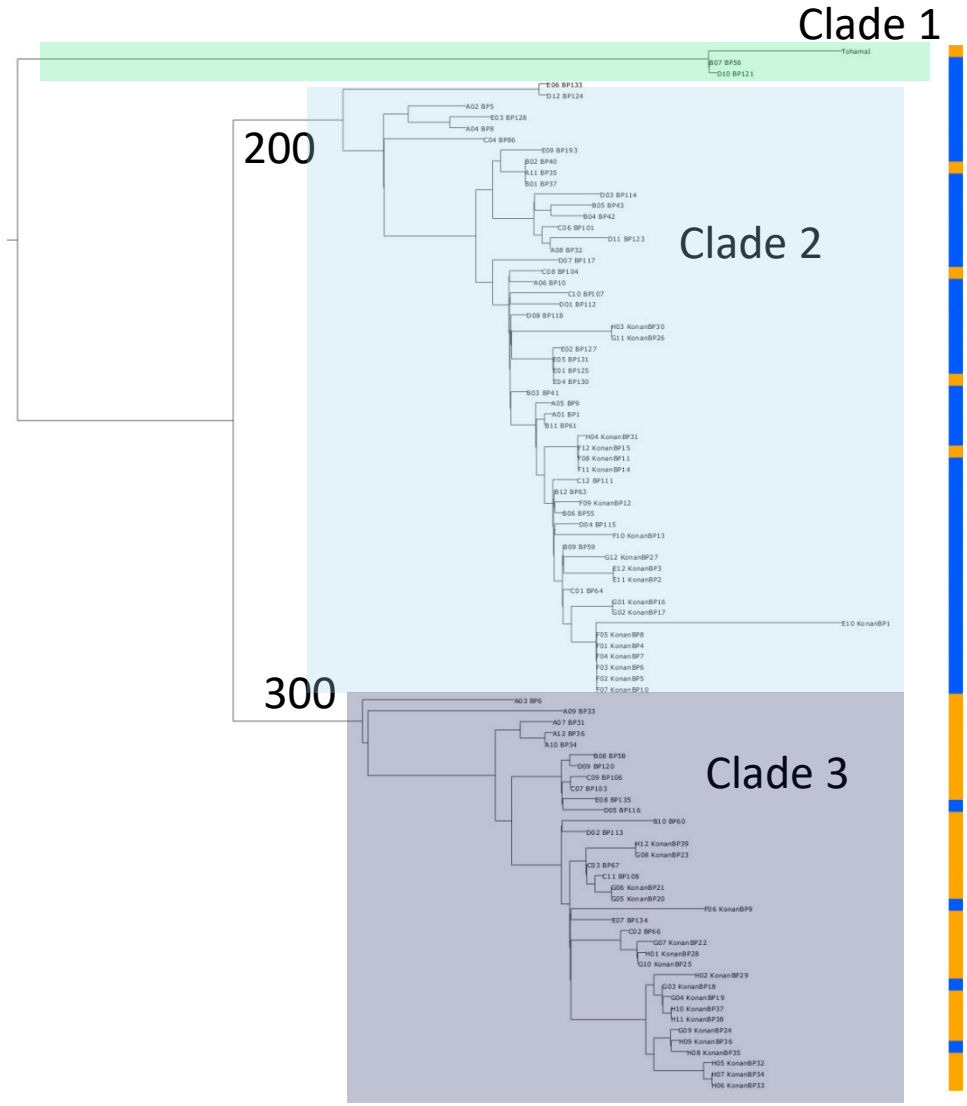
Population structure correction

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)

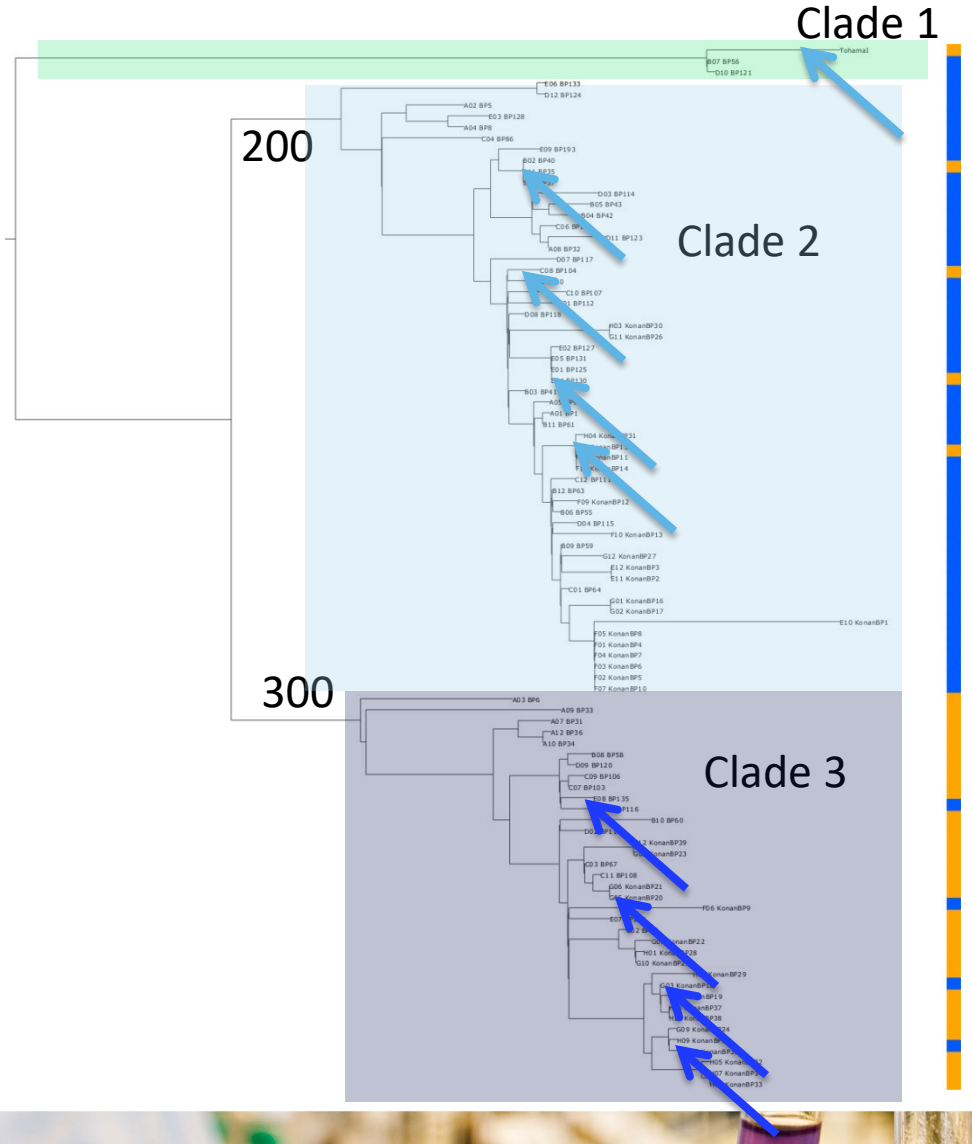
Population structure correction



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

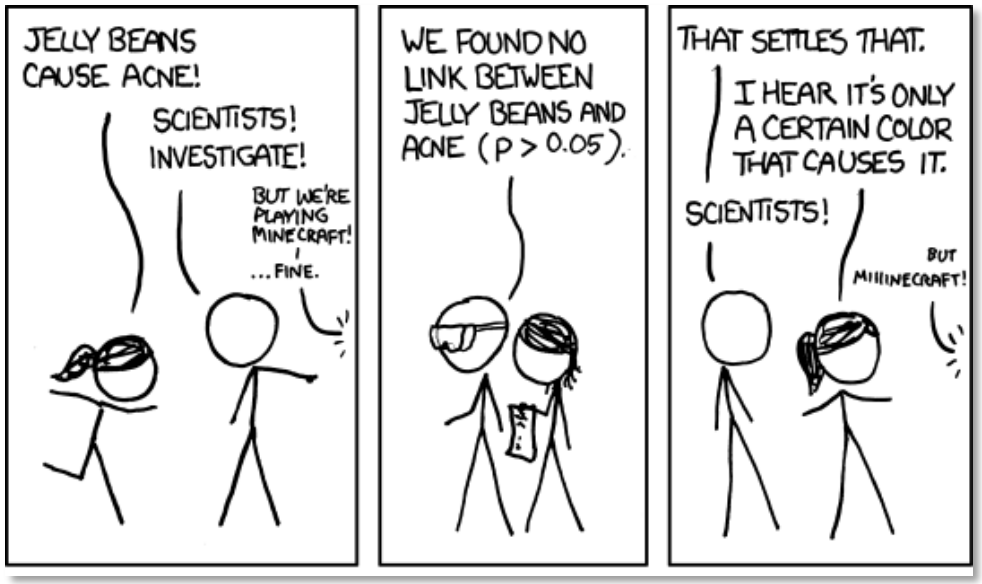
Alternatively:

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

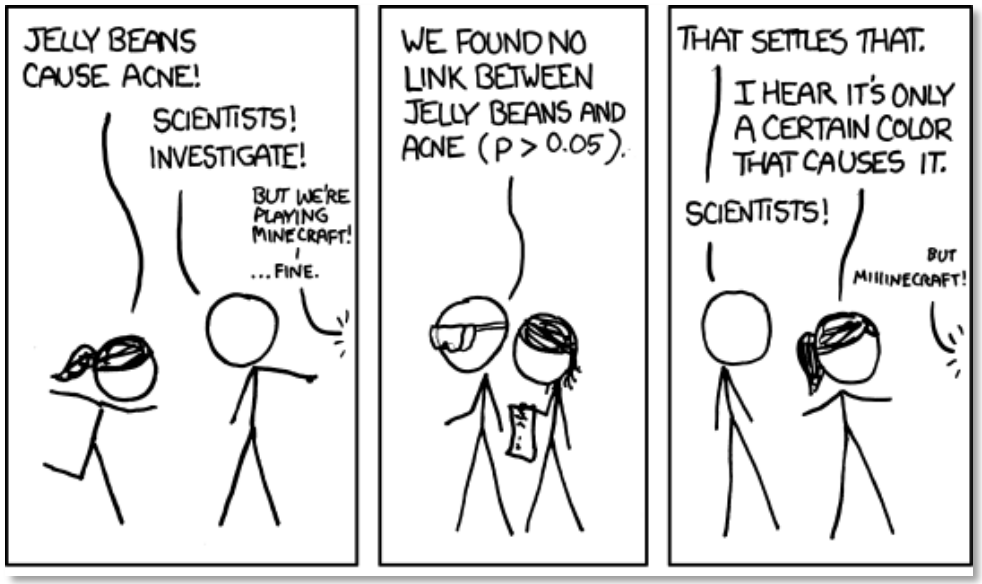
1000 SNPs have a p-value < 0.05. Are they all true positives?

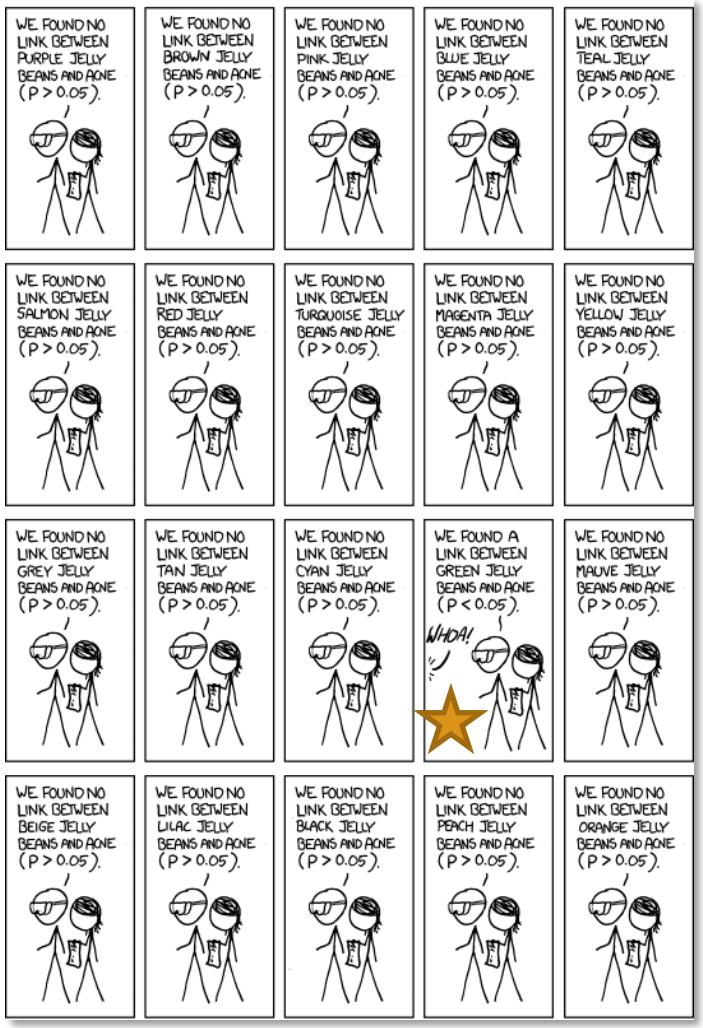
Multiple testing problem



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

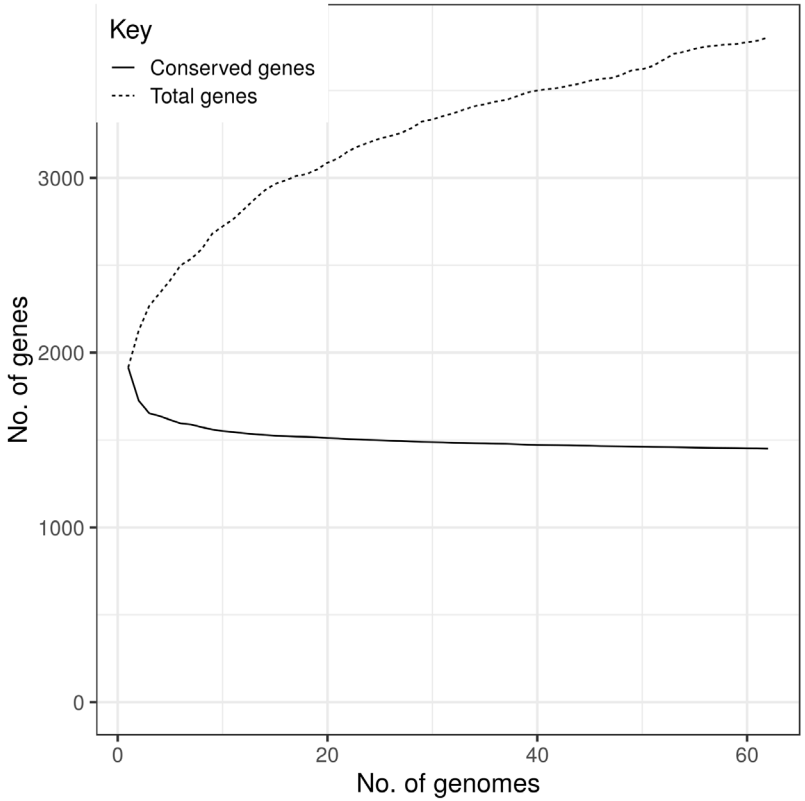
Phage-Derived Protein Induces Increased Platelet Activation and Is Associated with Mortality in Patients with Invasive Pneumococcal Disease.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5241397/

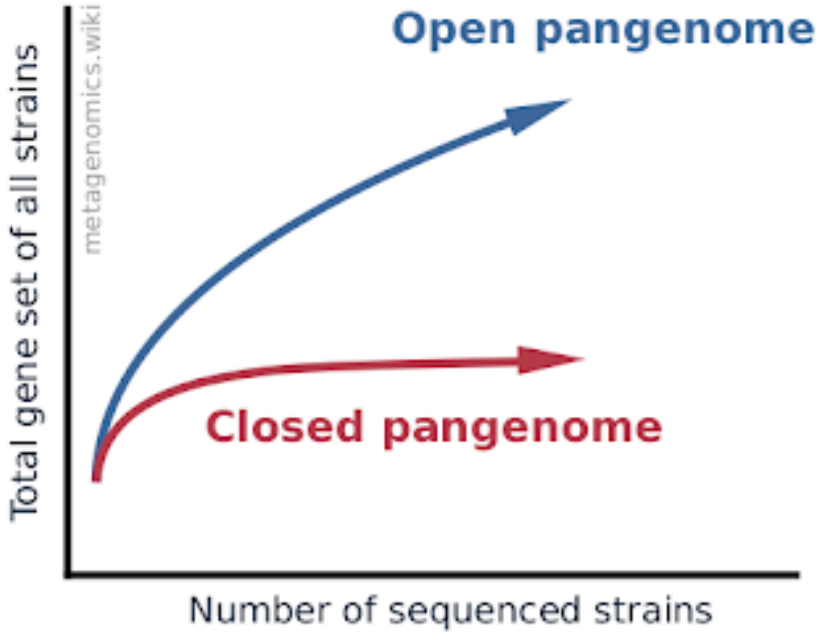
Comprehensive identification of single nucleotide polymorphisms associated with beta-lactam resistance within pneumococcal mosaic genes.
http://www.ncbi.nlm.nih.gov/pubmed/25101644

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

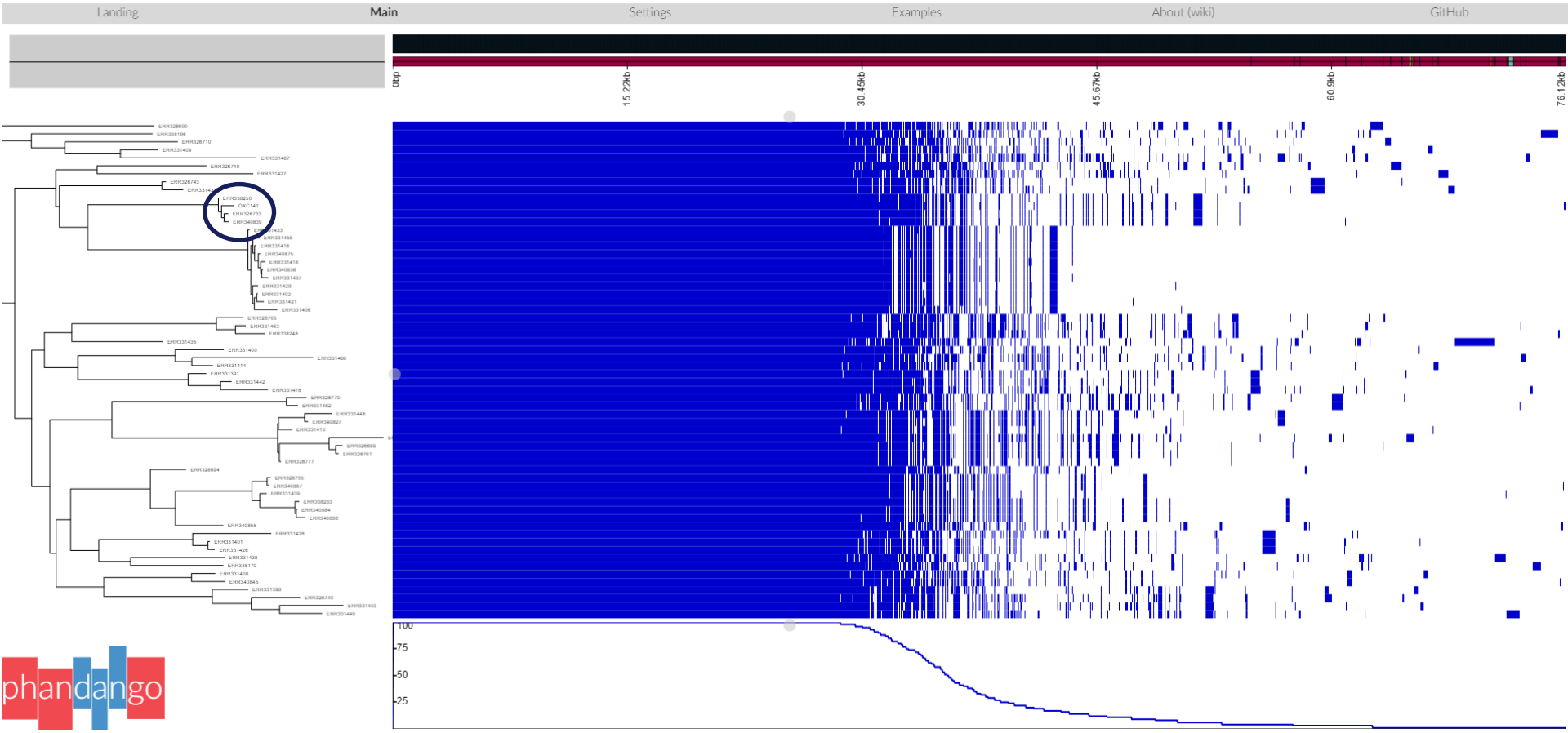
Day 4 results

Open or closed pan genome?

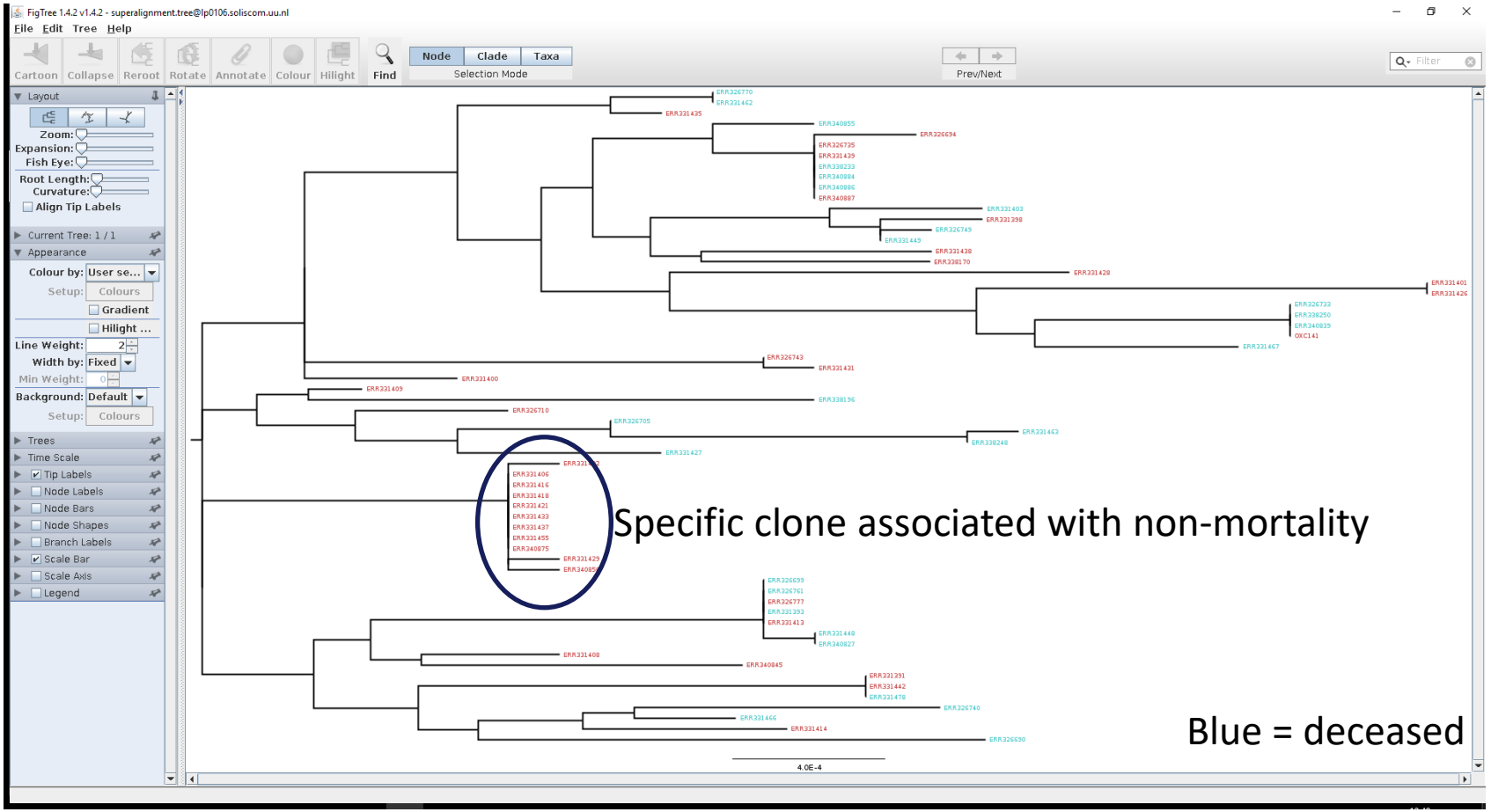




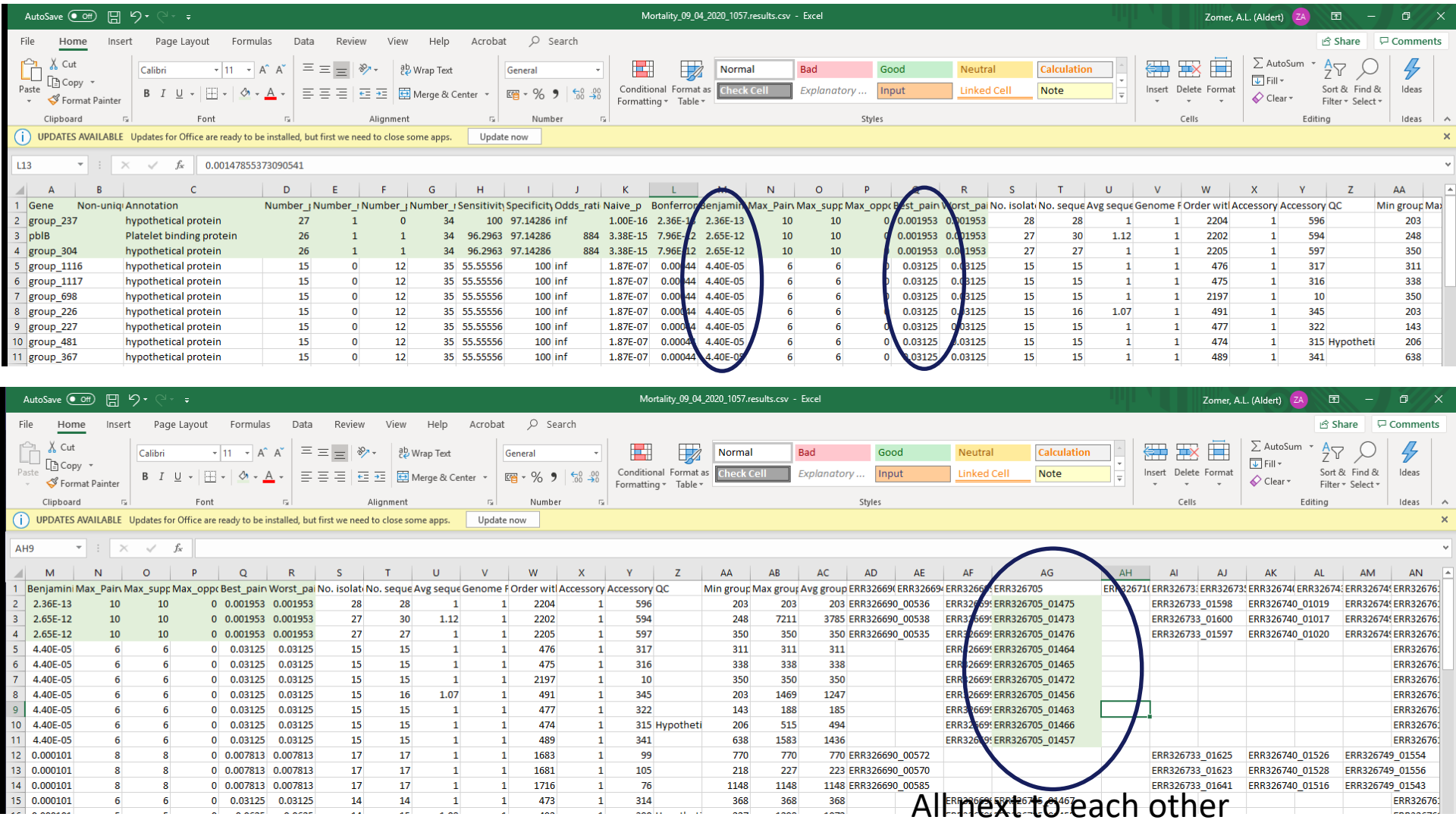
Phandango – which isolates are related to the reference



Clones associated with mortality

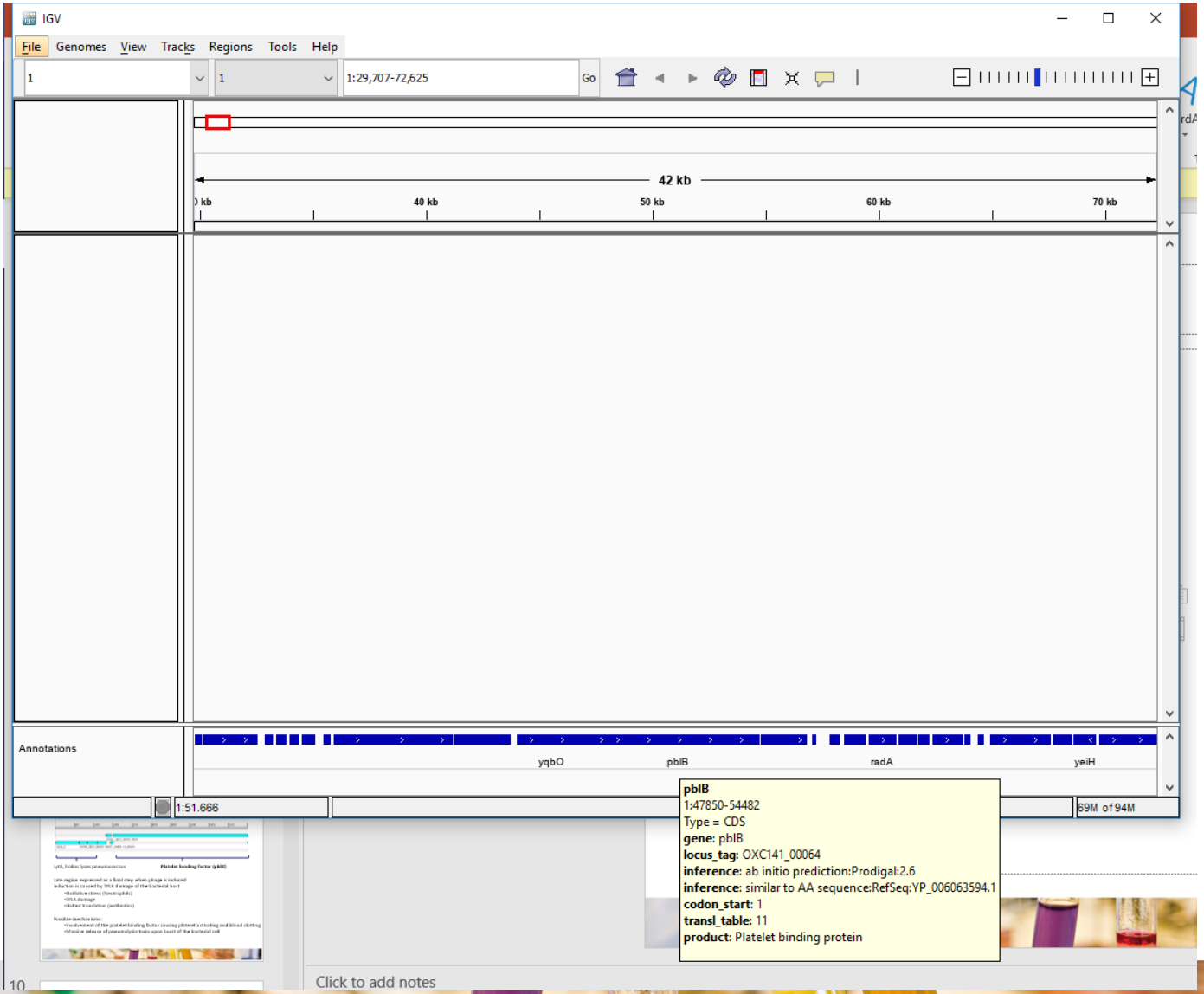


Genes associated with mortality

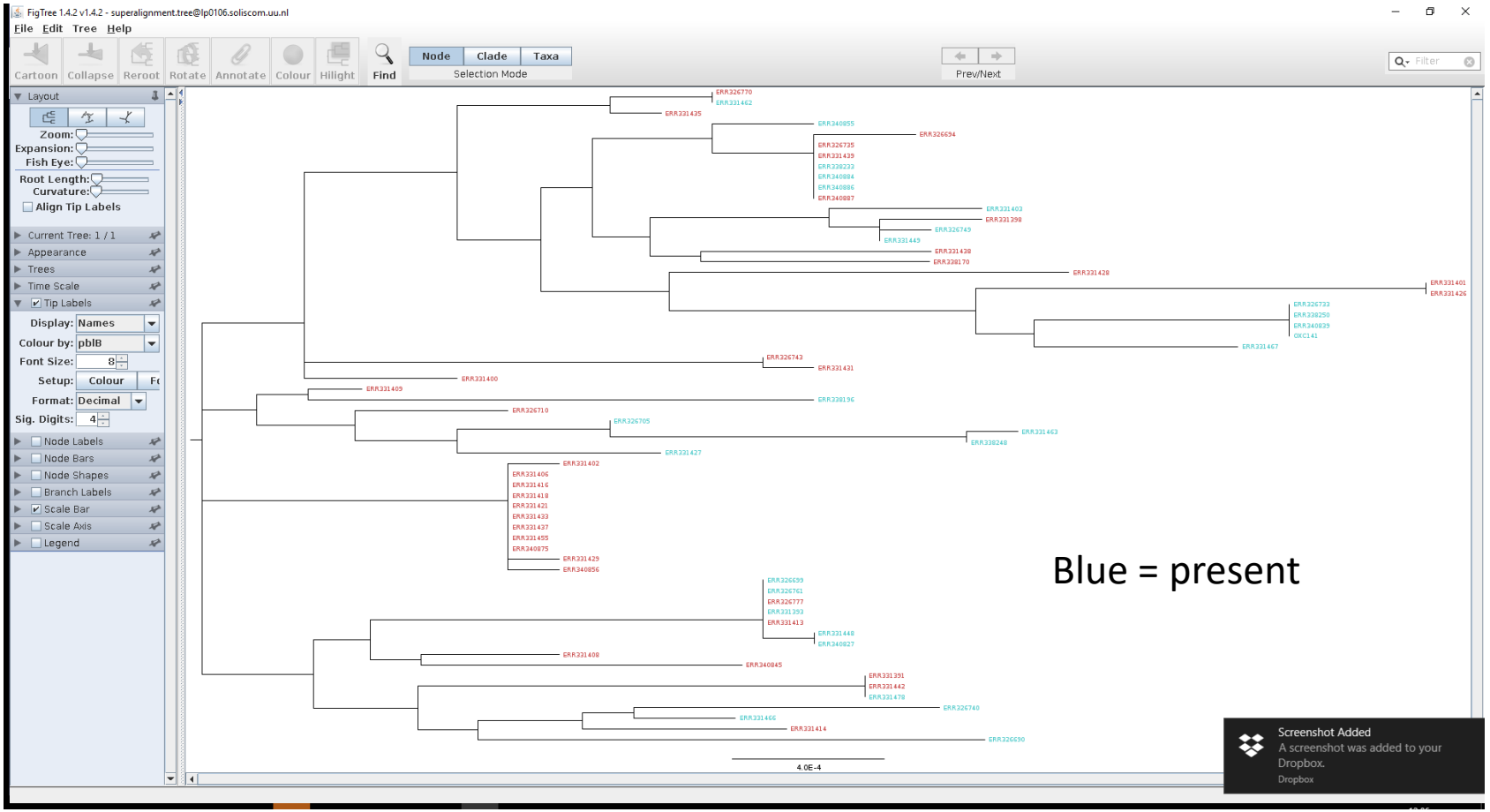


All next to each other

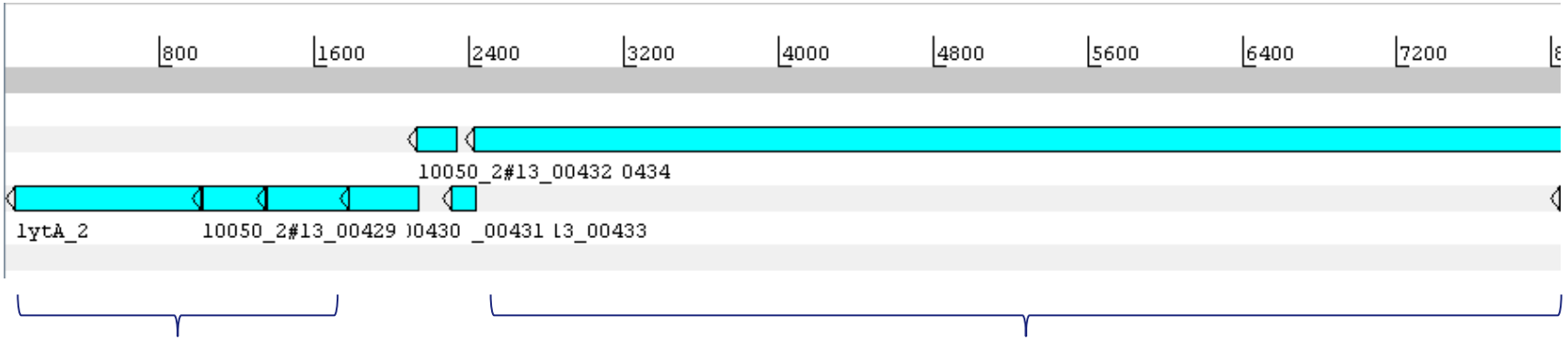
Location


IGV
File Genomes View Tracks Regions Tools Help
1:29,707-72,625
Go
42 kb
40 kb
50 kb
60 kb
70 kb
Annotations
yqbO
pblB
radA
yeiH
1:51.666
69M of 94M
pblB
1:47850-54482
Type = CDS
gene: pblB
locus_tag: OXC141_00064
inference: ab initio prediction:Prodigal:2.6
inference: similar to AA sequence:RefSeq:YP_006063594.1
codon_start: 1
transl_table: 11
product: Platelet binding protein
Click to add notes

PblB presence absence



Gene cluster associated with mortality



LytA, holins lyses pneumococcus

Platelet binding factor

Gene cluster is part of a bacteriophage on the bacterial genome. Can be expressed when bacteria is stressed

- Oxidative stress (neutrophils, immune system of the host)
- DNA damage (antibiotics)
- Halted translation (antibiotics)

Hypotheses about pathogenicity:

- Platelet binding factor causing platelet activating and blood clotting

or

- Massive release of pneumolysin toxin upon burst of the bacterial cell due to phage expression

AMERICAN SOCIETY FOR MICROBIOLOGY | mBio®

RESEARCH ARTICLE

Phage-Derived Protein Induces Increased Platelet Activation and Is Associated with Mortality in Patients with Invasive Pneumococcal Disease

Rahajeng N. Tunjungputri,[a,b] Fredrick M. Mobegi,[c] Amelieke J. Cremers,[c,d] Christa E. van der Gaast-de Jongh,[c] Gerben Ferwerda,[c] Jacques F. Meis,[d,e] Nel Roeleveld,[f,g] Stephen D. Bentley,[h] Alexander S. Pastura,[c] Sacha A. F. T. van Hijum,[i] Andre J. van der Ven,[a] Quirijn de Mast,[a] Aldert Zomer,[i,j] Marien I. de Jonge[c]

Department of Internal Medicine, Radboud Institute for Molecular Life Sciences, Radboud University Medical Center, Nijmegen, The Netherlands[a]; Center for Tropical and Infectious Diseases (CENTRID), Faculty of Medicine, Diponegoro University, and Dr. Kariadi Hospital, Semarang, Indonesia[b]; Laboratory of Pediatric Infectious Diseases, Department of Pediatrics, Radboud Institute for Molecular Life Sciences, Radboud University Medical Center, Nijmegen, The Netherlands[c]; Department of Medical Microbiology and Infectious Diseases, Canisius-Wilhelmina Hospital, Nijmegen, The Netherlands[d]; Department of Medical Microbiology, Radboud University Medical Center, Nijmegen, The Netherlands[e]; Department for Health Evidence, Radboud Institute for Health Sciences, Radboud University Medical Center, Nijmegen, The Netherlands[f]; Department of Pediatrics, Radboudumc Amalia Children's Hospital, Radboud University Medical Center, Nijmegen, The Netherlands[g]; Wellcome Trust Sanger Institute, Pathogen Genomics Group, Hinxton, Cambridgeshire, United Kingdom[h]; Center for Molecular and Biomolecular Informatics, Radboud Institute for Molecular Life Sciences, Radboud University Medical Center, Nijmegen, The Netherlands[i]; Department of Infectious Diseases and Immunology, Faculty of Veterinary Medicine, Utrecht University, Utrecht, The Netherlands[j]

ABSTRACT To improve our understanding about the severity of invasive pneumococcal disease (IPD), we investigated the association between the genotype of *Streptococcus pneumoniae* and disease outcomes for 349 bacteremic patients. A pneumococcal genome-wide association study (GWAS) demonstrated a strong correlation between 30-day mortality and the presence of the phage-derived gene *pblB*, encoding a platelet-binding protein whose effects on platelet activation were previously unknown. Platelets are increasingly recognized as key players of the innate immune system, and in sepsis, excessive platelet activation contributes to microvascular obstruction, tissue hypoperfusion, and finally multiorgan failure, leading to mortality. Our *in vitro* studies revealed that *pblB* expression was induced by fluoroquinolones but not by the beta-lactam antibiotic penicillin G. Subsequently, we determined *pblB* induction and platelet activation by incubating whole blood with the wild type or a *pblB* knockout mutant in the presence or absence of antibiotics commonly administered to our patient cohort. *pblB*-dependent enhancement of platelet activation, as measured by increased expression of the α-granule protein P-selectin, the binding of fibrinogen to the activated αIIbβ3 receptor, and the formation of platelet-monocyte complex occurred irrespective of antibiotic exposure. In conclusion, the presence of *pblB* on the pneumococcal chromosome potentially leads to increased mortality in patients with an invasive *S. pneumoniae* infection, which may be explained by enhanced platelet activation. This study highlights the clinical utility of a bacterial GWAS, followed by functional characterization, to identify bacterial factors involved in disease severity.

Received 28 October 2016 **Accepted** 19 December 2016 **Published** 17 January 2017

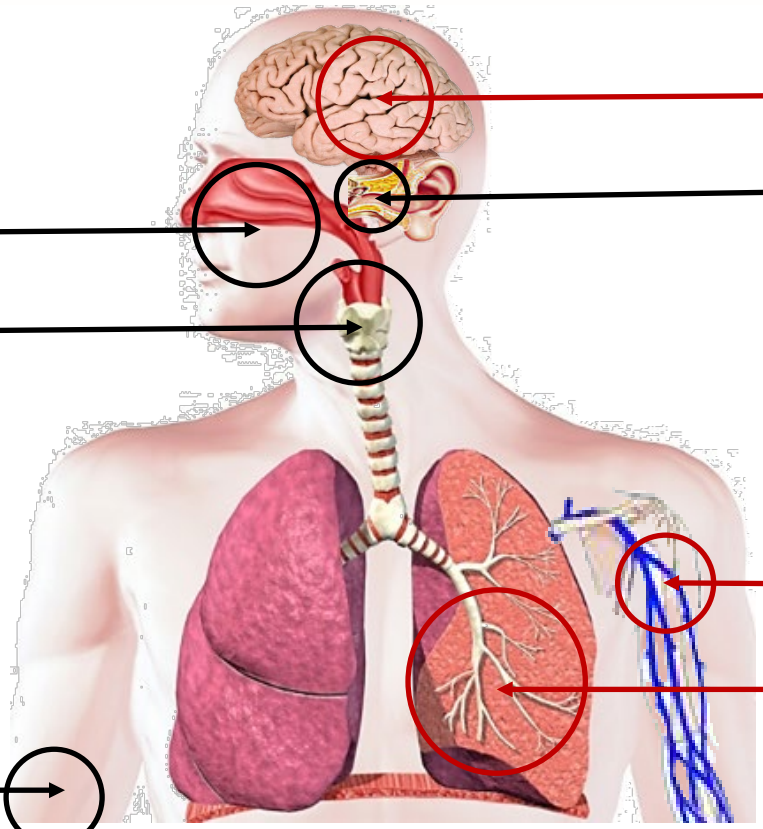
Citation Tunjungputri RN, Mobegi FM, Cremers AJ, van der Gaast-de Jongh CE, Ferwerda G, Meis JF, Roeleveld N, Bentley SD, Pastura AS, van Hijum SAFT, van der Ven AJ, de Mast Q, Zomer A, de Jonge MI. 2017. Phage-derived protein induces increased platelet activation and is associated with mortality in patients with invasive pneumococcal disease. mBio 8:e01984-16. https://doi.org/10.1128/mBio.01984-16.

Editor Lise-anne Pirofski, Albert Einstein College of Medicine

Address correspondence to Marien I. de Jonge, marien.dejonge@radboudumc.nl.

R.N.T. and F.M.M. contributed equally to this work.

Streptococcus pneumoniae



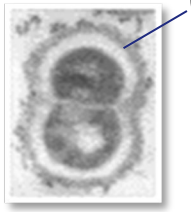
Meningitis

Otitis media

Sinusitis
mastoiditis

Bronchitis

Releases pneumolysin, a toxin, when the bacterial cell lyses



Capsule

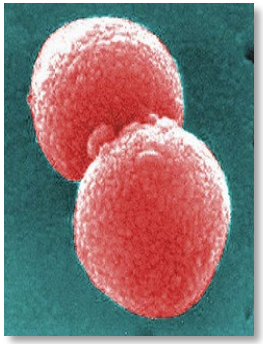
Vaccin
PCV7,10,13
2006

Bacteremia
Sepsis

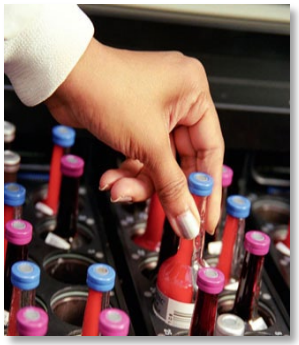
Pneumonia

Septic arthritis

Cohort

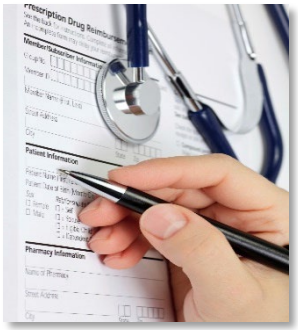


S.pneumoniae



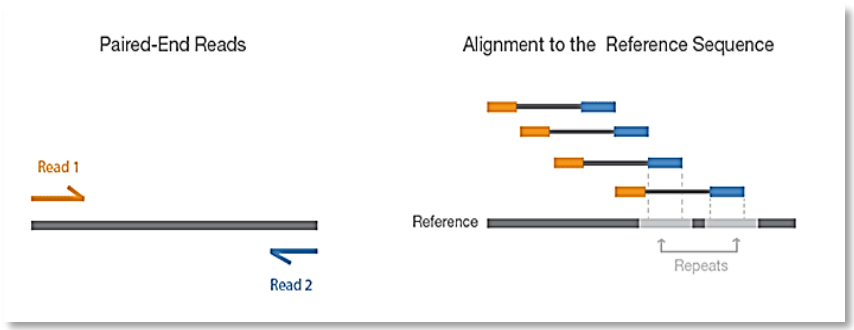
Blood culture +

Two Dutch hospitals `2001-2011`



Clinical data

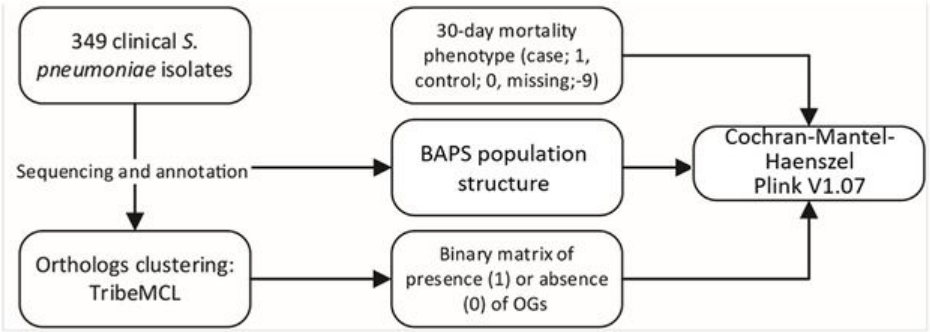
Serotype



Illumina PE sequencing: Sanger Institute

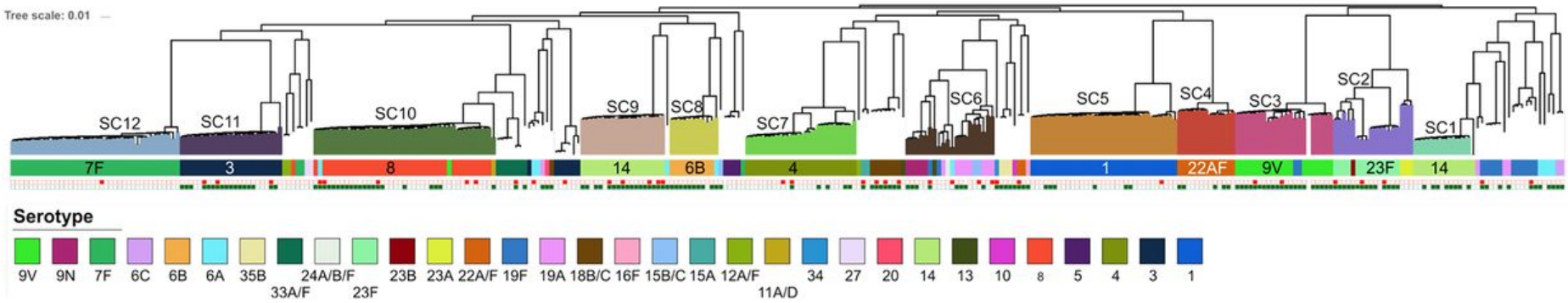
Bioinformatics analysis

A



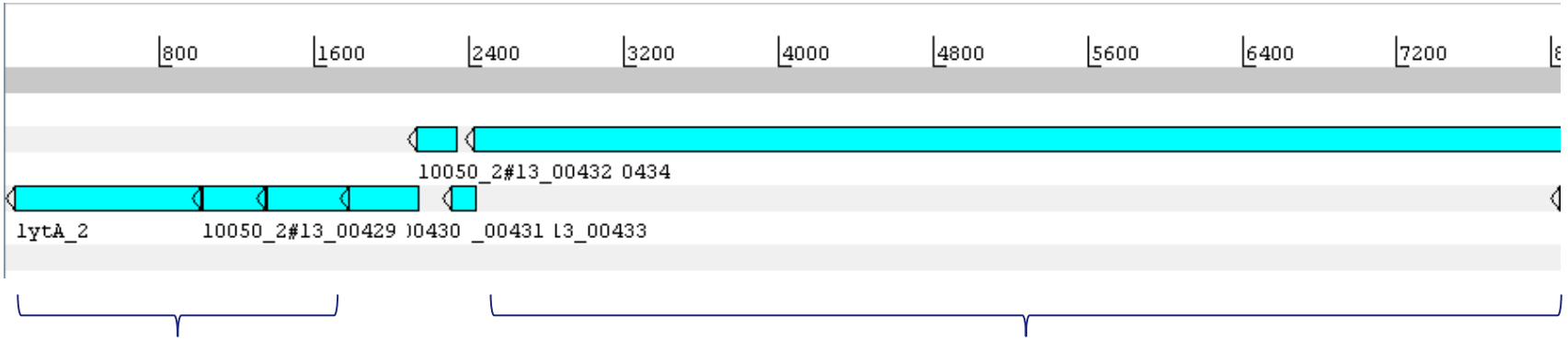
What pneumococcal virulence genes are over-represented in the isolates from patients that died within 30 days of hospital admission?

B



7 genes associated with 30 day mortality. All next to eachother

Gene cluster associated with mortality



LytA, holins lyses pneumococcus

Platelet binding factor

Gene cluster is part of a bacteriophage on the bacterial genome. Can be expressed when bacteria is stressed

- Oxidative stress (neutrophils, immune system of the host)
- DNA damage (antibiotics)
- Halted translation (antibiotics)

Hypotheses about pathogenicity:

- Platelet binding factor causing platelet activating and blood clotting

or

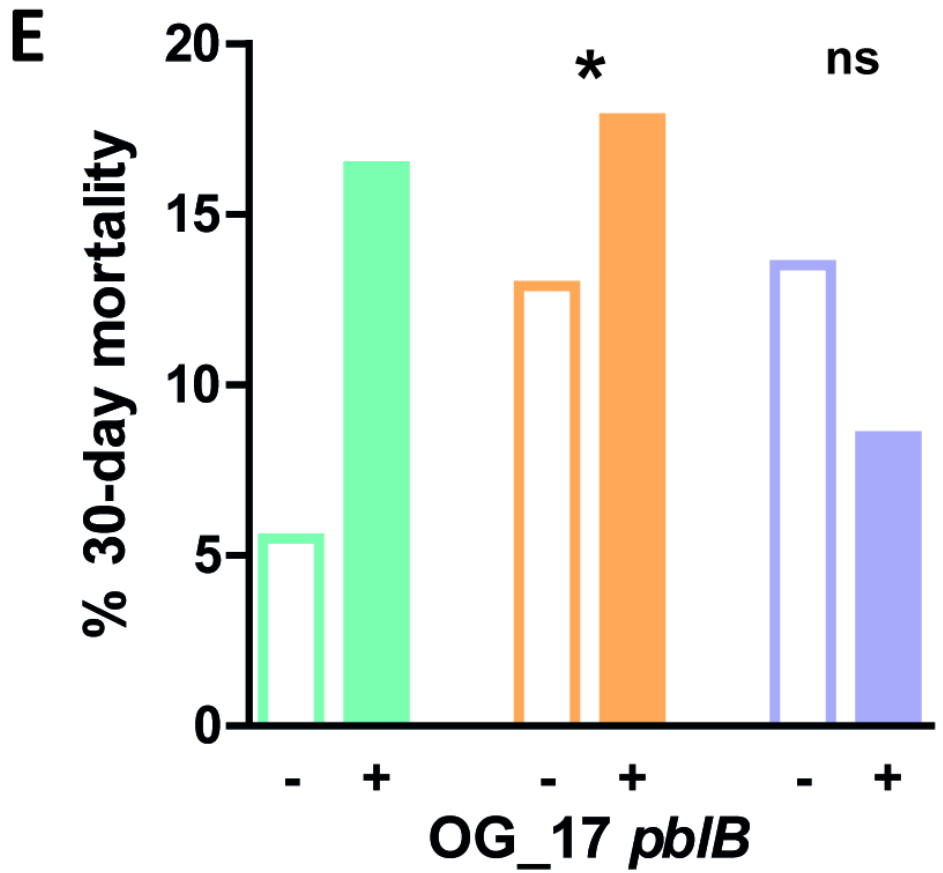
- Massive release of pneumolysin toxin upon burst of the bacterial cell due to phage expression

Severity of disease

Validated findings with qPCR in other datasets

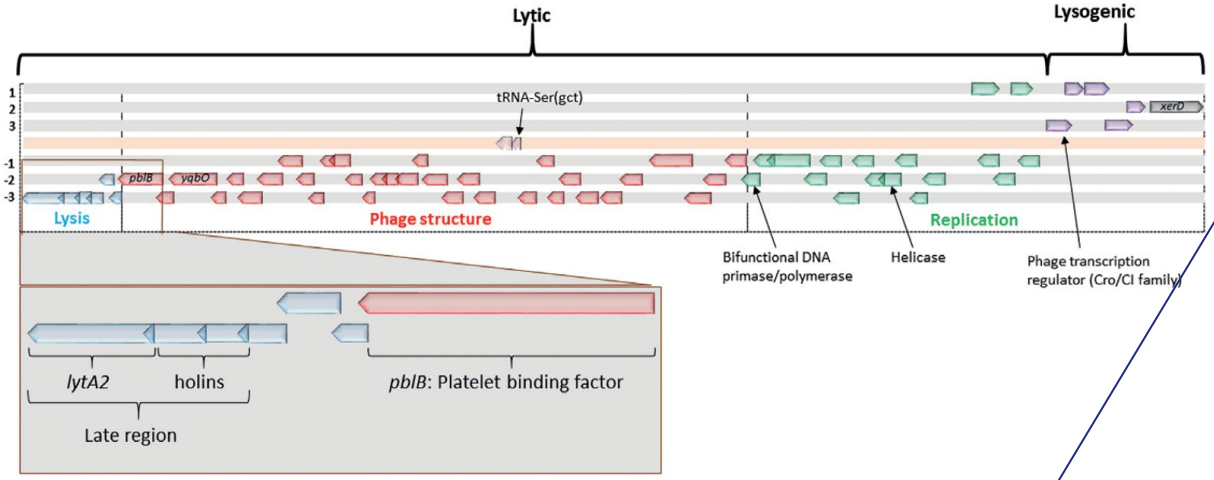
PblB (platelet binding factor). Bacterial carriage associated with increased 30 day mortality

Platelet activation -> multi organ failure



Index cohort (n=349)
Geographical validation cohort (n=482)
Temporal validation cohort (n=121)

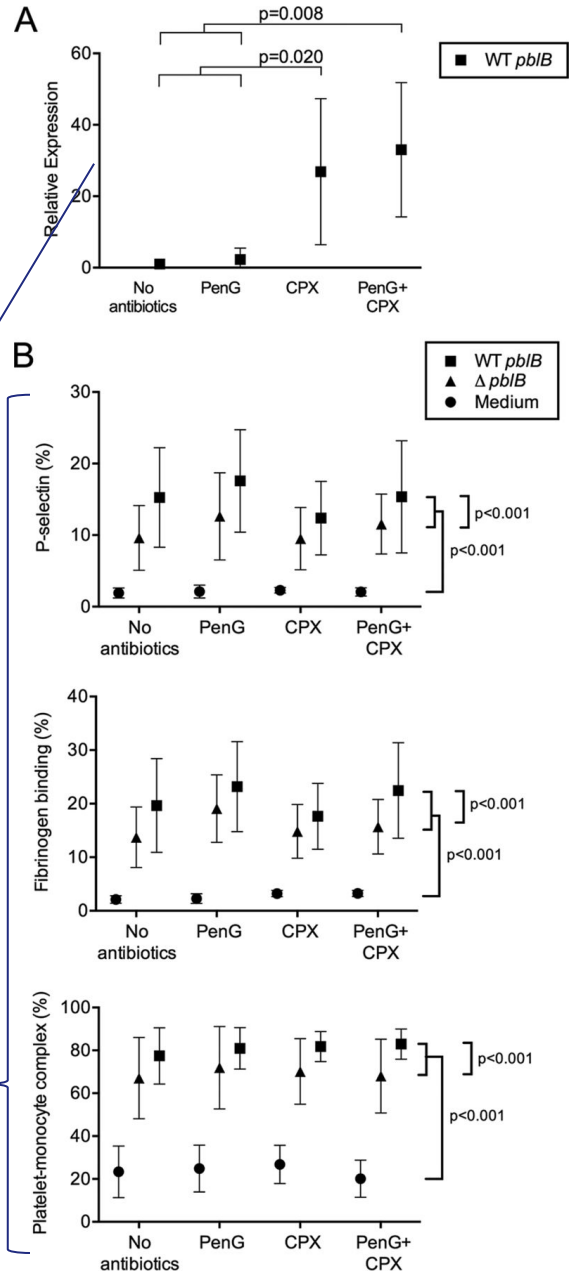
PblB – Biological mechanism



Carried on a phage – induced by DNA damage

PblB expression induced when exposed to specific antibiotics (MitC but also fluoroquinolones)

PblB activates platelets



Tunjungputri RN, Mobegi FM, Cremers AJ, van der Gaast-de Jongh CE, Ferwerda G, Meis JF, Roeleveld N, Bentley SD, Pastura AS, van Hijum SA, van der Ven AJ, de Mast Q, Zomer A, de Jonge MI. Phage-Derived Protein Induces Increased Platelet Activation and Is Associated with Mortality in Patients with Invasive Pneumococcal Disease. MBio. 2017 Jan 17;8(1). pii: e01984-16.