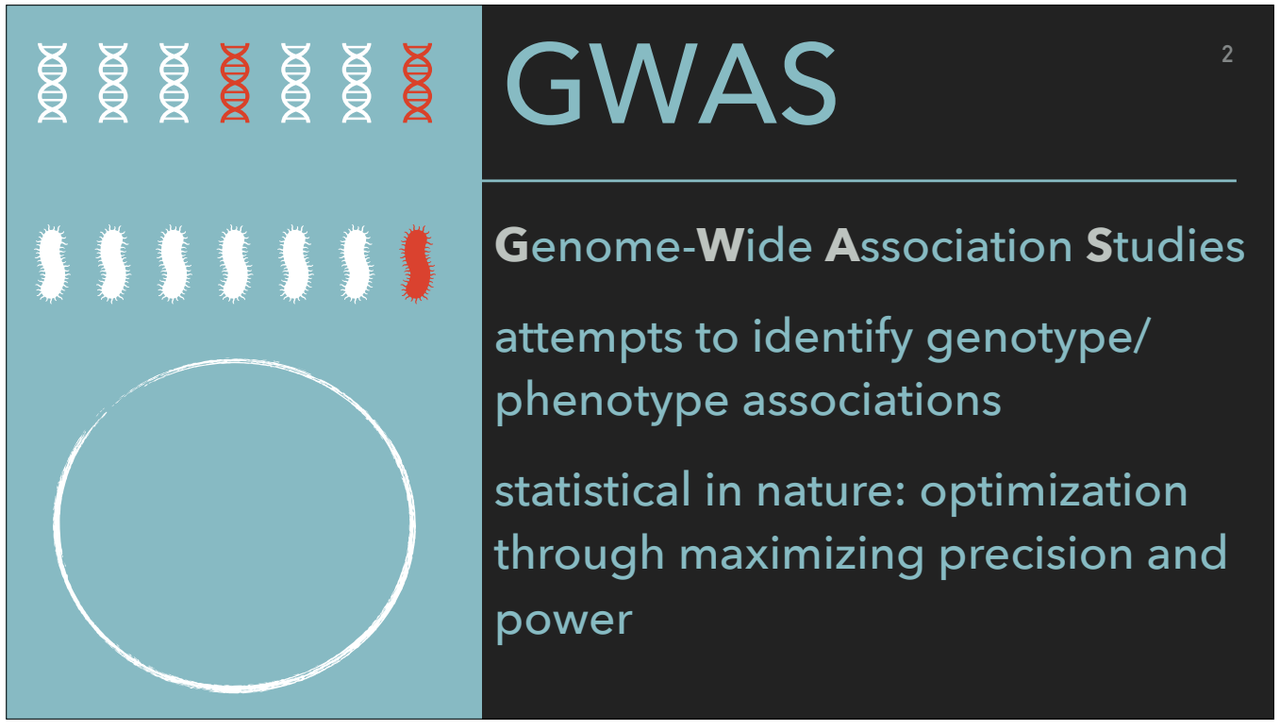
# statistics and bacterial GWAS

Jeremiah Yarmie

Biostats Final Presentation




-attempts to identify causative relationships between genetic variants and phenotypic outcomes within a population
-inherently statistical approach, concerned with maximizing precision and power in its analysis.

-top-down approach to conducting genetic research, compared to bottom-up approaches rooted in molecular biology like creating knock outs, mutant strains, etc.

# bacterial approaches



```
GTCATAACTTACCTGAGACTACTTGGAAATGTGGCTAGATC
GTCATAACTTACCTGAGACTACTTGGAAATGTGGCTAGATC
GTCATAACTTACCTGAGACTACTTGGAAATGTGGCTAGATC
GTCATAACTTACCTGAGACTACGTGGAAATGTGGCTAGATC
GTCATATCTTACCTGAGACTACTTGGAAATGTGGCTAGATC
GTCATATCTTACCTGAGACTACTTGGAAATGTGGCTAGATC
```

allele counting

-is an allele more present in cases vs controls?

homoplasy

-is there an appearance of genotype/ phenotype on multiple branches of a phylogenetic tree?

allele-counting

-looks for an increased presence of a certain allele at a locus in cases relative to controls.

homoplasy

-presence of similar genetic loci on different branches of a phylogenetic tree

-accounts for the effects of population structure and linkage disequilibrium inherently

-requires a much smaller sample size to reach statistical significance

# issues

| real world | GWAS results: association | GWAS results: unrelated |
|---|---|---|
| truly related | true positive | false negative (type II error) |
| no causation | false positive (type I error) | true negative |

two problems arise due to bacterial clonal reproduction:

1) linkage disequilibrium
2) population stratification

can result in spurious associations (false positives)

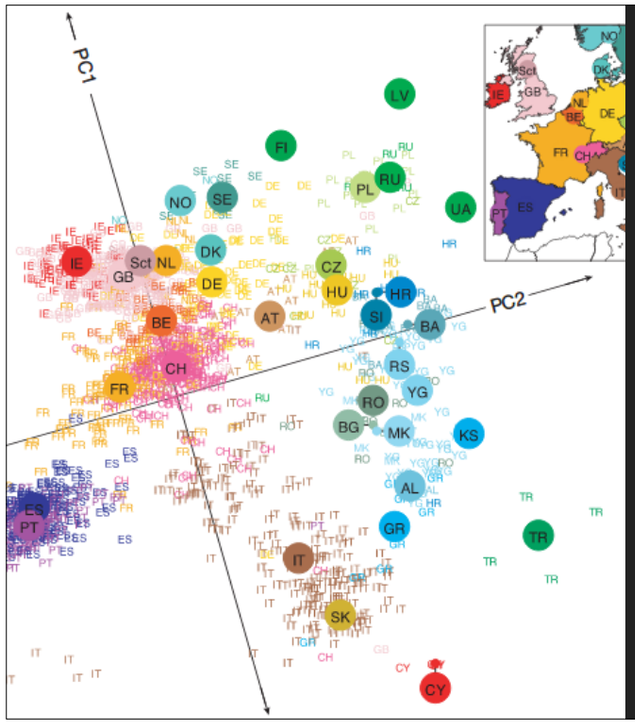
confounding effect of genetic relatedness between different bacterial isolates and strains

# clonal frame

unlike in humans,
almost the entire bacterial chromosome is under linkage disequilibrium
-a clonal frame interrupted by areas of recombination


almost the entire bacterial chromosome is under linkage disequilibrium.

-a patchwork of recombined regions on a tract of linked regions called a clonal frame
-all regions of the clonal frame are in linkage disequilibrium.






# populations

like in humans, population stratification can result in spurious associations

-if a specific population is enriched with an allele it may be deemed associated

population stratification
-certain subgroups of closely-related individuals can give rise to spurious associations with phenotypes of interest.
-members of a population and subpopulation structure contain a non-random distribution of alleles.
-problem in highly clonal and rarely recombining bacteria

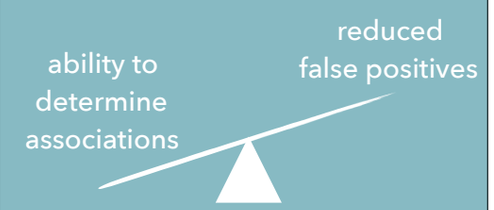
# solutions?

principal component analysis
can correct for population stratification and clonal frames

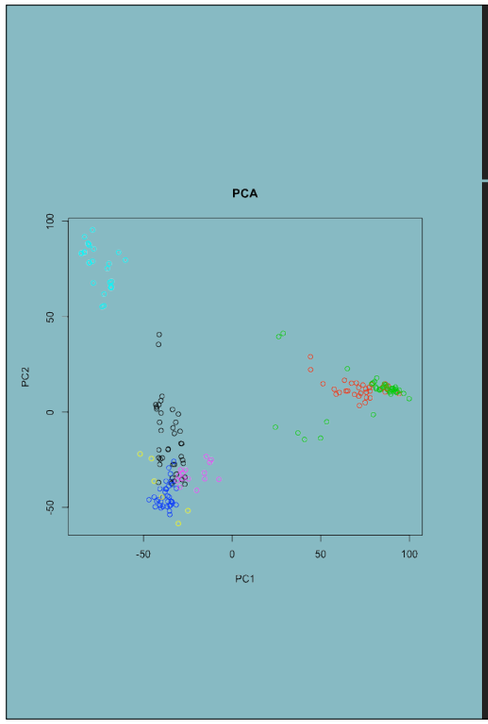
the largest principal components tend to correspond to these sources of error

but PCA trades statistical power for reduced type I error






PCA
-correct for population structure and stratification when conducing GWAS analysis.
-largest principal components tend to correspond to major population structures and strain lineages or clonal frames.




# typically



standard GWAS approaches pair PCA with downstream linear mixed modelling

PCs are used as fixed effect factors in downstream linear regression analysis

Principal components identified that consider population structure are then chosen as fixed effect factors in downstream linear regression

# treeWAS

homoplasy-based tool for bacterial GWAS

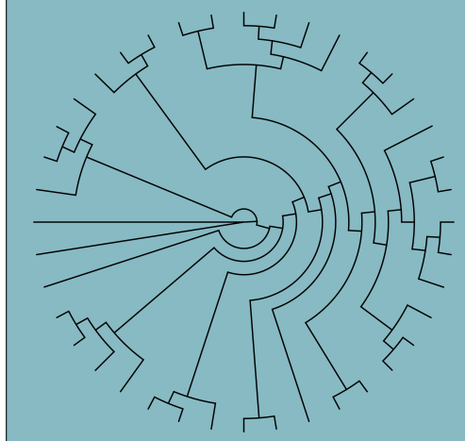
phylogenetic approach to correcting for stratification

appropriate for both clonal and recombining bacterial species



-balances a low false positive rate, high sensitivity, and high positive predictive value
-appropriate to use both with highly clonal bacteria and extremely recombinant ones (through the implementation of ClonalFrameML)

# workflow



creates null dataset representing confounding factors

conducts hypothesis tests between null and test scores to determine significance

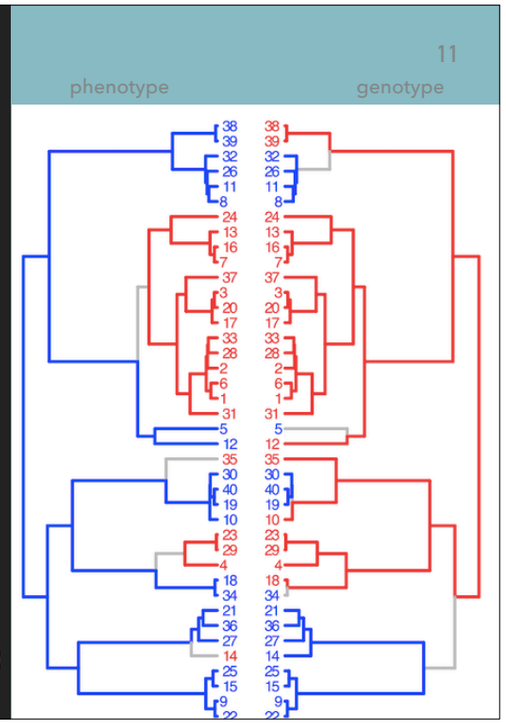
-uses a null genetic dataset to conduct hypothesis tests on the validity of associations seen in the test data
-null dataset represents test dataset except for associations (unless they arise due to the confounding factors)
-maintains: clonal genealogy, terminal phenotype, genetic composition and homoplasy

# terminal score

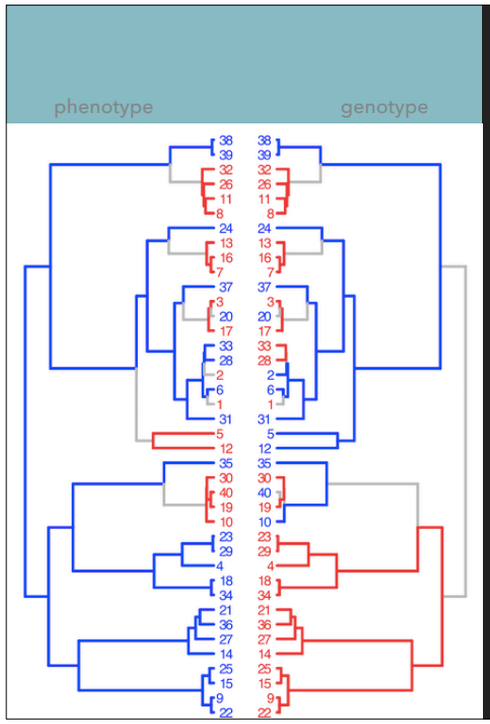
$$\textbf{Score 1} = \left| \sum_{i=1}^{n} \frac{1}{n} \left( p_i^{des} g_i^{des} + (1 - p_i^{des})(1 - g_i^{des}) - (1 - p_i^{des}) g_i^{des} - p_i^{des} (1 - g_i^{des}) \right) \right| \quad (1)$$

considers tree tips only

agnostic to ancestral states (internal nodes)




- Measures sample-wide association across the tree leaves.
- Counts all 4 terminal states p+g+, p+g-, p-g+, and p-g-
- Then determines if an allele is over-represented among a particular phenotypic state
- Only determines association at the tree termini, agnostic to ancestral states




# simultaneous score



$$\textbf{Score 2} = \left| \sum_{i=1}^{n_b} (p_i^{anc} - p_i^{des})(g_i^{anc} - g_i^{des}) \right| \quad (2)$$

considers ancestral states (internal nodes)

measures parallel change in genotype and phenotype occurring along a tree branch

-Measures the degree of parallel change in genotype and phenotype across tree branches.

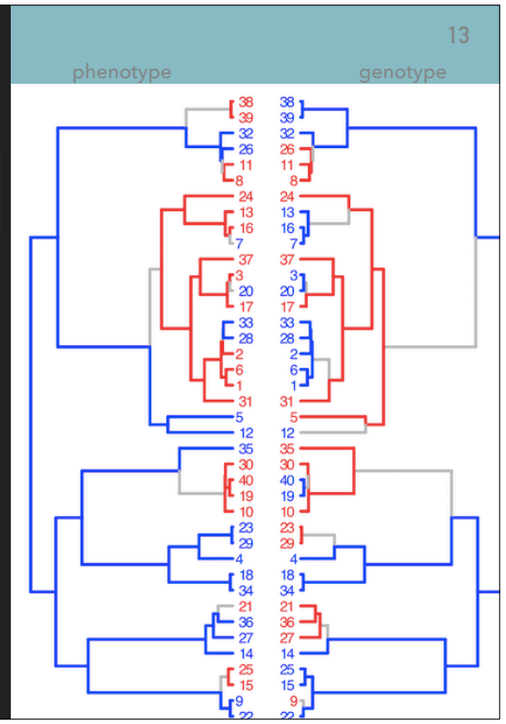
-Counts the number of branches containing a simultaneous substitution in both genotype and phenotype. -Simultaneous substitutions indicate a strong relationship between genotype and phenotype.

-Able to detect associations arising through similar or complementary pathways.

# subsequent score

$$\textbf{Score 3} = \left| \sum_{i=1}^{n_b} \frac{4}{3} p_i^{anc} g_i^{anc} + \frac{2}{3} p_i^{anc} g_i^{des} + \frac{2}{3} p_i^{des} g_i^{anc} + \frac{4}{3} p_i^{des} g_i^{des} - p_i^{anc} - p_i^{des} - g_i^{anc} - g_i^{des} + 1 \right| \quad (3)$$

measures proportion of entire tree where genotype and phenotype coincide



# workflow

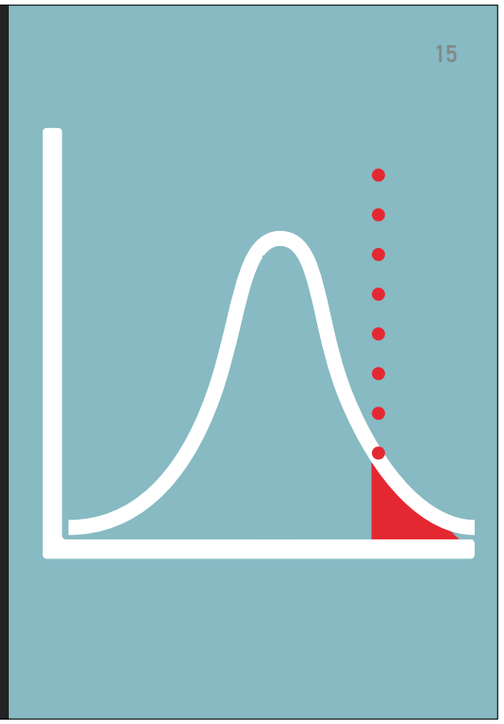
three scores are compared to null distribution determined from null loci and true phenotypes

-null distribution of the three association score statistics is calculated by measuring associations between the null loci and the true phenotypes.
-these will represent the null hypothesis of our tests.

# multiple testing

correcting for multiple loci and three tests

then a significance threshold is drawn



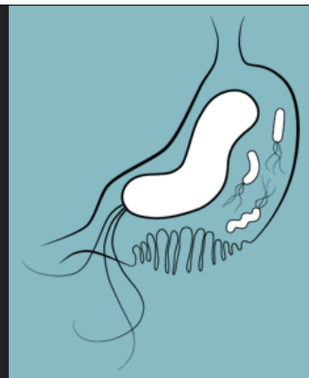

Bonferroni correction and p = 0.01 threshold

# my datasets?

*Helicobacter pylori*
extremely recombinant

*Mycobacterium tuberculosis*
extremely clonal

genomic datasets were taken from previous GWAS studies

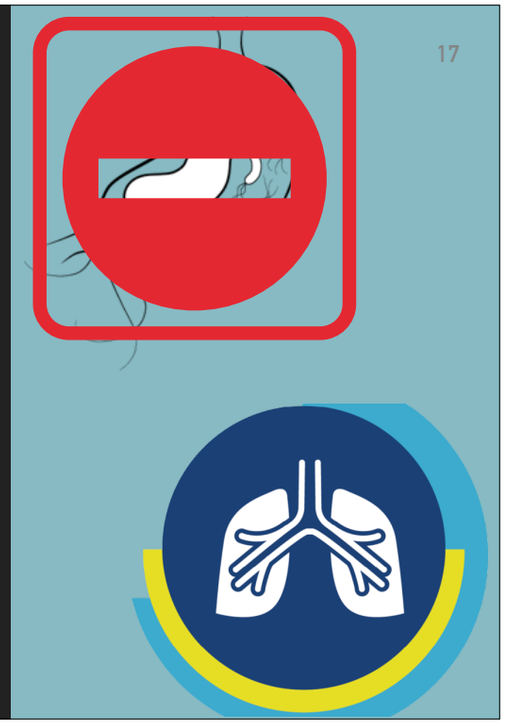


# my datasets?

*Helicobacter pylori*
extremely recombinant

*Mycobacterium tuberculosis*
extremely clonal

genomic datasets were taken from previous GWAS studies





had issues accessing the data (said it was uploading on NCBI's SRA, but it was actually deposited in contig draft genome form)

could not run the same pipeline that you can with SRA data (fastq illumina reads)

tried multiple alignments but it was too computationally intensive and didn't end up working

# dataset size

entire *H. pylori* dataset from Berthenet et al. (2018) used
-170 genomes

subset *M. tuberculosis* dataset from Farhat et al. (2019) used
-183 genomes

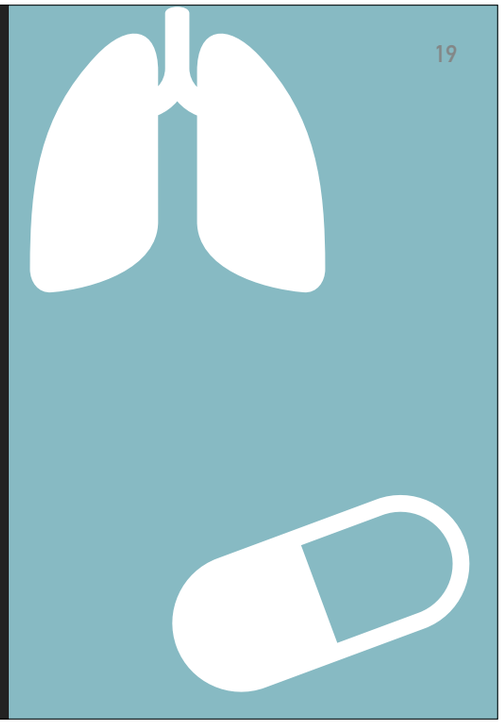
Reduction in the Mtb dataset was done predominantly due to time and computation power, as well as being similar in size to the Hp dataset.

# phenotypic data



metadata for *M. tuberculosis* is antimicrobial resistance or susceptibility (binary) for various drugs:

-isoniazid, rifampin, pyrazinamide, ethambutol, streptomycin, kanamycin, etc.



# data input

| | 6.g | 6.a | 7.c | 7.a | 16.t | 16.g | 22.c |
|---|---|---|---|---|---|---|---|
| SRR057510 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| SRR057595 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| SRR057610 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| SRR057619 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| SRR057734 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| SRR057768 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| SRR057770 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| SRR057771 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| SRR058116 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| SRR058369 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| SRR058370 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| SRR058371 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| SRR058372 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| SRR058373 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| SRR058377 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| SRR058399 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| SRR058417 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| ERR2467267 | 1 | 0 | 0 | 1 | 0 | 1 | 0 |
| ERR2467268 | 1 | 0 | 1 | 0 | 0 | 1 | 1 |
| ERR2467269 | 1 | 0 | 0 | 1 | 0 | 1 | 0 |

```r
#next we import our genetic data
mtb_dna <- read.dna(file = "mtb_data/mtb_snv_align.fasta", format = "fasta") #import our
multi-aligned fasta of all isolate genome SNVs
mtb_snps <- DNAbin2genind(mtb_dna)@tab #converts all SNVs to a DNA bin matrix
mtb_tree <- read.tree(file = "mtb_data/mtb_tree.nhx") #import out phylogenetic tree created with
FastTree
```

multi-aligned SNV fasta

newick tree

# data input



| isolates | INH | RIF | EMB | STR | CIP | CYS | CAP | ETA | KAN | OFLX | PAS | PZA | AMK |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SRR671787 | R | R | R | R | NA | NA | S | R | S | R | NA | NA | NA |
| SRR671788 | S | S | S | S | NA | NA | S | S | S | S | NA | NA | NA |
| SRR671789 | R | R | S | R | NA | NA | S | S | S | R | NA | NA | NA |
| SRR671790 | R | R | R | S | NA | NA | S | S | S | R | NA | NA | NA |
| SRR671791 | R | R | S | R | NA | NA | S | R | S | R | NA | NA | NA |
| SRR671792 | S | S | S | S | NA | NA | S | S | S | S | NA | NA | NA |
| SRR671793 | S | S | S | S | NA | NA | S | S | S | S | NA | NA | NA |
| SRR671794 | R | R | R | R | NA | NA | R | R | R | R | NA | NA | NA |
| SRR671795 | R | R | S | R | NA | NA | S | S | S | S | NA | NA | NA |
| SRR671796 | R | R | R | R | NA | NA | S | R | S | R | NA | NA | NA |
| SRR671797 | R | R | R | R | NA | NA | S | R | S | R | NA | NA | NA |
| SRR671798 | R | R | S | R | NA | NA | S | S | S | S | NA | NA | NA |
| SRR671799 | R | R | R | R | NA | NA | R | R | R | R | NA | NA | NA |
| SRR671800 | S | S | S | S | NA | NA | S | S | S | S | NA | NA | NA |
| SRR671801 | R | R | S | R | NA | NA | S | S | S | R | NA | NA | NA |
| SRR671802 | S | S | S | S | NA | NA | S | S | S | S | NA | NA | NA |
| SRR671803 | S | S | S | S | NA | NA | S | S | S | S | NA | NA | NA |
| SRR671804 | R | R | R | R | NA | NA | S | S | S | R | NA | NA | NA |
| SRR671805 | S | S | S | S | NA | NA | S | S | S | S | NA | NA | NA |

```
mtb_phen_rif <- as_vector(mtb_isos_meta$RIF)
names(mtb_phen_rif) <- mtb_isos_meta$isolates
```

| | V1 |
|---|---|
| SRR671800 | S |
| SRR671801 | S |
| SRR671802 | S |
| SRR671803 | S |
| SRR671804 | S |
| SRR671805 | S |
| SRR671806 | S |
| SRR671807 | R |
| SRR671808 | S |
| SRR671809 | S |
| SRR671810 | S |
| SRR671811 | R |
| SRR671812 | S |
| SRR671813 | S |
| SRR671814 | S |
| SRR671815 | R |
| SRR671816 | S |
| SRR671817 | S |
| SRR671818 | S |
| SRR671819 | S |

vector of phenotypic data

# calling treeWAS

```
str <- treeWAS(snps = mtb_snps,
               phen = mtb_phen_str,
               tree = mtb_tree,
               seed = 1)
```

just one line of code

# data output



```
> emb$treeWAS.combined
$treeWAS.combined
[1] "17249.c" "3696.c"  "3696.t"  "3833.a"  "3833.g"

$treeWAS
$treeWAS$terminal
[1] "3696.c"  "3696.t"  "3833.a"  "3833.g"  "17249.c"

$treeWAS$simultaneous
[1] "3696.c" "3696.t"

$treeWAS$subsequent
[1] "3696.c" "3696.t"
```

```
9147  191   30   26    6    5
  21                        6
   0                        0

attr(
[1] "
> emb$
  treeWAS.combined
  terminal
  simultaneous
  subsequent
  dat
```

# *M. tuberculosis* results



| gene | name | function | in LMM GWAS? |
|---|---|---|---|
| rpoB | RNA polymerase B | transcription | Yes |
| rpsL | small ribosomal protein | translation | No |
| fabG1 | 3-oxoacyl-ACP reductase | mycolic acid production | Yes |
| lldD2 | L-lactate dehydrogrenase | pyruvate biosynthesis | No |
| embA | arabinosyl transferase | mycolic acid production | Yes |

# ease of use

i used a very small subset of the total genomes analyzed using PCA-LMM GWAS

this is not a true validation study

think of it as a “pilot” or proof of concept

# discussion

each analysis only took about 4 minutes to run, so scaling up should be easy

data pre-processing will increase in time and computational demand with more genomes, however

# input

no support for non-binary categorical phenotypic data

only binary categorical and continuous phenotypes supported

# output

Bonferroni corrected p-value $<10^{-5}$

```
$sig.snps
        SNP.locus p.value      score G1P1 G0P0 G1P0 G0P1
3696.c       1895       0  0.4285714   94   36   31   21
3696.t       1896       0 -0.4285714   21   31   36   94
3833.a       1967       0  0.3296703   95   26   41   20
3833.g       1968       0 -0.3296703   20   41   26   95
17249.c      8882       0  0.3296703  113    8   59    2

$min.p.value
p-values listed as 0 are less than:
                        1.060783e-05
```

it would be nice for the p-values to be printed with precision, especially since it may come up in journal guidelines or with reviewers

# references


Berthenet, E., Yahara, K., Thorell, K., Pascoe, B., Meric, G., Mikhail, J. M., … Sheppard, S. K. (2018). A GWAS on Helicobacter pylori strains points to genetic variants associated with gastric cancer risk. *BMC biology* 16(1), 84. doi:10.1186/s12915-018-0550-3

Chen, P. E., and Shapiro, B. J. (2015) The advent of genome-wide association studies for bacteria. *Curr Opin Microbiol*. 25:17-24. doi: 10.1016/j.mib.2015.03.002.

Collins, C., and Didelot, X. (2018) A phylogenetic method to perform genome-wide association studies in microbes that accounts for population structure and recombination. *PLOS Computational Biology* 14(2): e1005958. https://doi.org/10.1371/journal.pcbi.1005958

Earle, S., Wu, C., Charlesworth, J. et al. Identifying lineage effects when controlling for population structure improves power in bacterial association studies. *Nat Microbiol* 1, 16041 (2016) doi:10.1038/nmicrobiol.2016.41

Falush, D. Bacterial genomics: Microbial GWAS coming of age. *Nat Microbiol* 1, 16059 (2016) doi:10.1038/nmicrobiol.2016.59

Farhat, M.R., Freschi, L., Calderon, R. et al. GWAS for quantitative resistance phenotypes in Mycobacterium tuberculosis reveals resistance genes and regulatory regions. *Nat Commun* 10, 2128 (2019) doi:10.1038/s41467-019-10110-6

Lees, J., Bentley, S. Bacterial GWAS: not just gilding the lily. *Nat Rev Microbiol* 14, 406 (2016) doi:10.1038/nrmicro.2016.82

Price, A., Patterson, N., Plenge, R. et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 38, 904–909 (2006) doi:10.1038/ng1847

Saber, M. M. and Shapiro, J. (2019) Benchmarking bacterial genome-wide association study (GWAS) methods using simulated genomes and phenotypes. bioRxiv 795492; doi: https://doi.org/10.1101/795492 (pre-print)

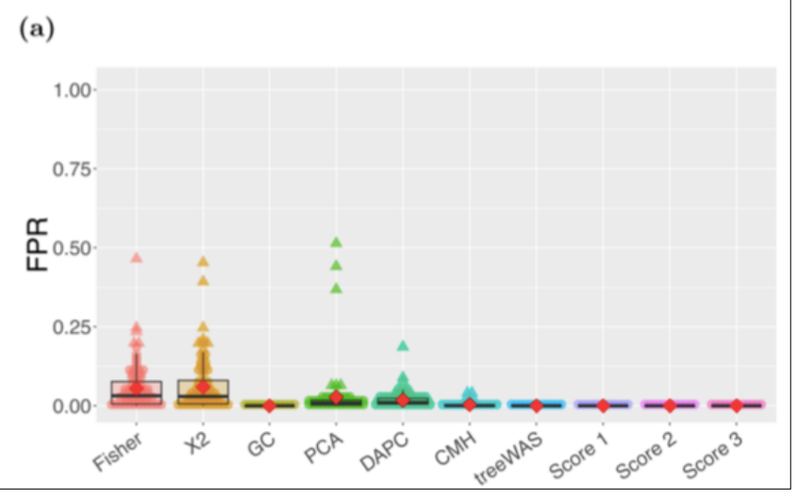
# false positives



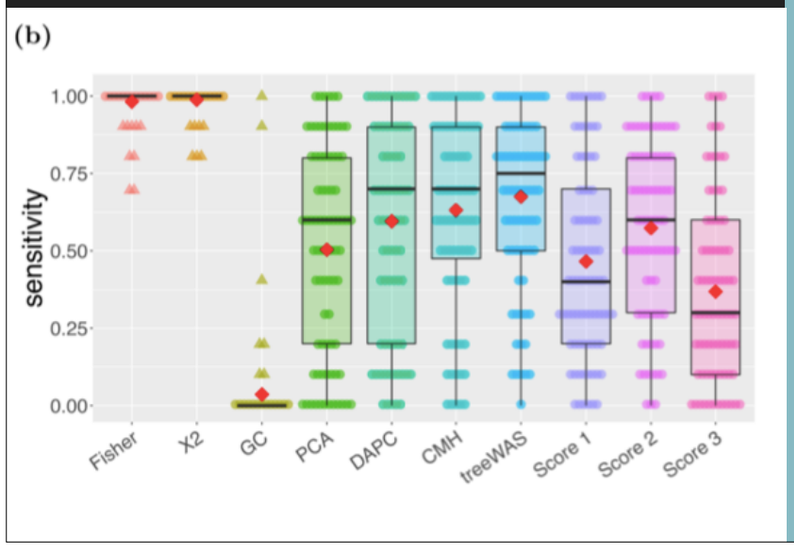
type I error

detected when it doesn't exist

$$\text{fpr} = \frac{\text{false positives}}{\text{(false positives + true negatives)}}$$



# sensitivity




true positive rate

proportion of hits that are real

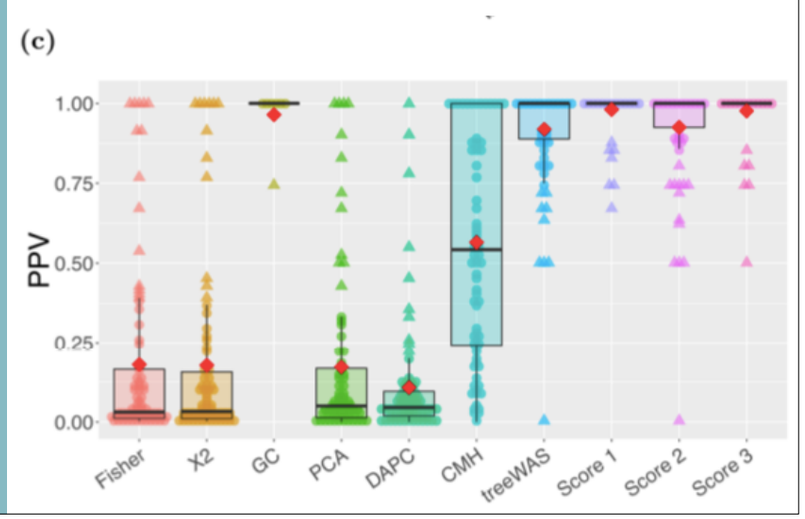
$$tpr = \frac{\text{true positives}}{(\text{true positives} + \text{false negatives})}$$

# positive predictive value



proportion of results that are true positives and negatives

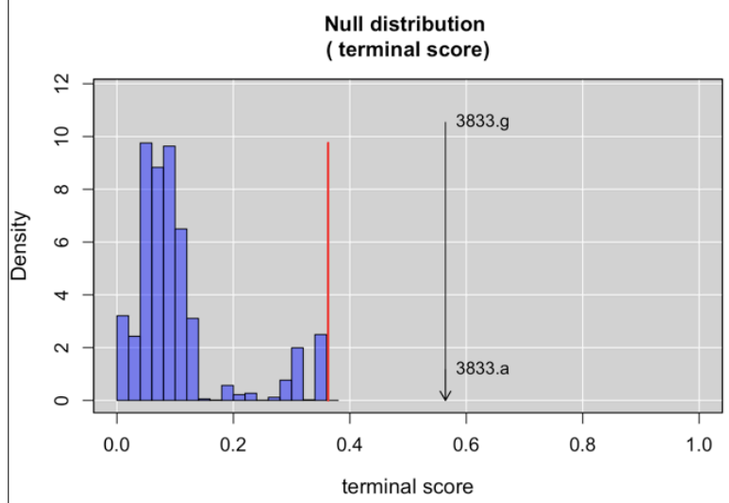
ppv = true positives / (false positives + true positives)



# output



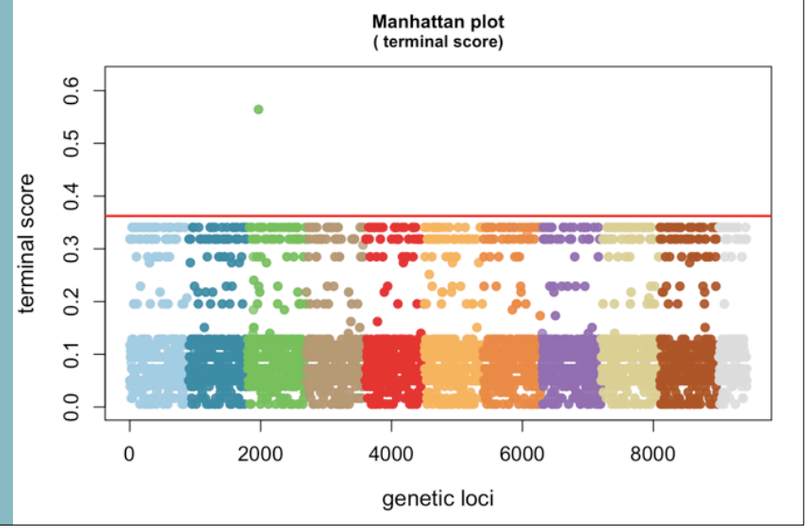
## null distribution



# output



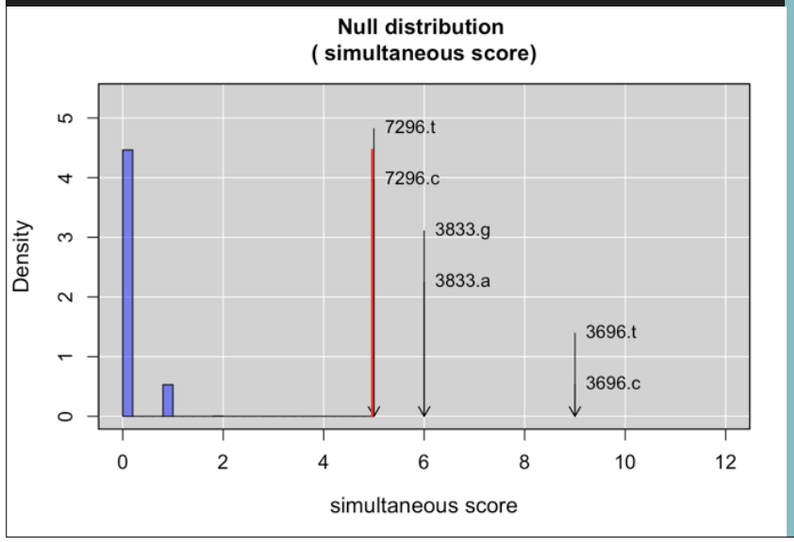
manhattan plot



# output



## null distribution



# output



manhattan plot