Tentative title: Power analysis of Genome-Wide Association Studies (GWAS) in bacterial pathogens using machine-learning

# Abstract:

Methods for GWAS developed for eukaryotic organisms are not easily translated into bacterial pathogens due to high level of clonality. Using simulation, we explored the statistical power of detecting SNP-associated phenotype traits (e.g., antibiotic resistance or MIC) using a machine learning approach. We first designed an algorithm to simulate SNPs associated with a continuous variable. We then applied machine learning techniques to identify the false positive and false negative errors. The result shows a large number of strains are needed to detect causal SNPs. We further developed methods to simulate phylogeny-associated SNPs, which increased the false positive detection rates. The results show the importance of taking phylogenetic relationship and clonality into account when performing GWAS for bacterial pathogens. Our algorithms and tools could be used for designing GWAS studies in pathogenic bacterial.

# Introduction

GWAS inspect dissimilarities from a reference to detect patterns.

# Material & Methods

## Data sets & pre-processing [tentative head]

## CDG measurements, genes, and SNPs

Original data was shared from xyz containing two tables. One of which, is a table log of repeated experimental samples collected from *Pseudomonas aeruginosa* strains at various tissue sites (Table 1). The use of this particular dataset are the expression levels of c-di-GMP, which in brief, is a secondary messenger related to decision making of whether to form biofilm or leave residence in search of a more favorable environment 1. The c-di-GMP levels serve as our continuous variable Y (predictor variable), which we later merge into a bigger matrix once the values have been normalized for the sheer convenience of a scaled number set.

The other table is a binary SNP matrix of 30 rows and 8,532 columns with rows and columns representing, in respective order, number of observed *P. aeruginosa* strains and SNP sites (Table 2). The binary indicates change (1) or no change (0). SNPs are mined and converted into binary using a pipeline that runs through a series of steps connecting the relationship between nucleotides in their corresponding gene and their alleles in each of the 30 strains. These SNPs will be our independent variable X also known as covariates. It is important to note that coherency of this matrix lies in the information that these genes, house-keeping genes, are either directly or indirectly related to inducing c-di-GMP level expression, and building on their possible importance, there seem to be a large number of variants across the strains and with so much information it can unravel a story about how the bacterium makes its decisions.

To hone in on the pathogenic variant(s), we first combine these two tables with respect to their matching strain names. However, due to the repeated process of collecting c-di-GMP levels in a strain, we resorted to averaging the values by the number of experiments performed for that particular strain. It was necessary to proceed with the averaged value in order for the number of row names to match.

In surveying the merged matrix, we quickly realized that the astronomical dimensions would likely exhaust any model or classifier we use, we decided to reduce the number of covariates. By doing so, it would not only use less computer RAM but also allow us to retrieve faster results in our analysis. Subsequently, we shrunk the columns from 8,532 to 1,078 by removing based on column sums equal to 0, 30, and redundant SNP patterns within genes (these are results). The reason being is that if the SNP site is the same across all 30 strains then its effect on the overall analysis will be futile and would in fact, add more redundancy to the matrix. It would also allow the data to have less variance and in this type of prediction analysis, it is not beneficial to have less spread. Though, this might raise concerns of whether it is legal to reduce based on redundant SNP patterns because [insert reason] but the removal is *within* genes and not its neighbors.

While we did successfully reduce the dimensions of our dataset, it was not ready to be fed into our models. If we blindly use this novel data, there would be no way of interpreting the result thus making it as good as a guess. Therefore, we strategized an alternative by using simulations to test the accuracy of our models before we move forward to testing the real data (discussions).

## Simulation of trait-associated SNPs

Simulation modeling sheds light on the uncertainties of randomness that occur even in genomics. Our approach is to first emulate a small dataset laddering onto the actual dimensions of our initial matrix in order to gauge and optimize a working frame. Second, we use a decomposition algorithm, Cholesky decomposition, to artificially inject a correlated variable (X) to the predictor variable (Y). This step conditionally creates a known relationship between the xy variables. We ran the algorithm several times to produce different sized tables with varying numbers of covariates ranging from [insert numerical range]. Lastly, we used a machine learning algorithm to identify such covariates.

**Simulating SNP-associated phenotypes**

**Tools (R packages) & steps**

**Simulate tree-associated phenotypes**

**Tools (R packages) & steps**

## Identification of causal SNPs using xgboost

What is xgboost algorithm with reference;

your customized scripts to run xgboost (description in words & equations & then commands/scripts as supplemental material)

# Results

# Discussion

# Conclusion

# References Cited

1. Yan, J. *et al.* Bow-tie signaling in c-di-GMP: Machine learning in a simple biochemical network. *PLOS Comput. Biol.* **13,** e1005677 (2017).

# Table 1.Two phenotypes: Strains and their normalized c-di-GMP levels & antimicrobial sensitivity indices (Mei – should I make in a separate table)

|  |  |  |  |
| --- | --- | --- | --- |
| Strain | Body site | Cdg-level (mean) | Cdg-level (standard deviation) |
| PA14 | NA | 2.11071438 | 0.11249807 |
| M55212 | Blood | 0.36094968 | 0.14645814 |
| F22031 | Pubic bone | -1.66607404 | 0.20298575 |
| W25637 | Sputum | -0.16582713 | 0.05295834 |

# Table 3. List of c-di-GMP genes (Mei)

|  |  |  |
| --- | --- | --- |
| Gene symbol | Gene annotation | Num of SNPs |
| B136-33\_52\_00507 | HDOD domain || Metal-dependent hydrolase HDOD | 727 |
| B136-33\_52\_01564 | CSS motif domain associated with EAL || Putative cyclic diguanylate phosphodiesterase, CSS motif-containing domain | 1622 |
| B136-33\_52\_01815 | GlnD PII-uridylyltransferase || PII-uridylyltransferase/Glutamine-synthetase adenylyltransferase | 494 |

# Table 4. List of antibiotic-resistant genes (Mei)

|  |  |  |
| --- | --- | --- |
| Gene symbol | Gene annotation | Num of SNPs |
| NalD / PA3574 | NalD is a repressor of MexAB-OprM. Mutations lead to multidrug resistance and MexAB-OprM overexpression. | 5 |
| MexA / PA0425 | MexA is the membrane fusion protein of the MexAB-OprM multidrug efflux complex. | 4 |
| OprD / PA0958 | Basic amino acid, basic peptide and imipenem outer membrane porin OprD precursor |  |

# Figure 1. Flowchart of analytical protocol

# Figure 2. Power analysis using simulated data (show simulated datasets)

# Figure 3. SNPs & genes associated with cyclic-di-GMP expression