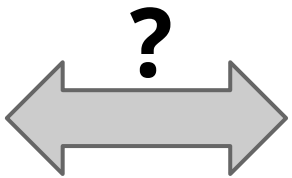
GWAS in *Pseudomonas aeruginosa*

Justin Torok
Xavier Lab
7/6/15

Outline

A. Motivation and GWAS background
B. Results
C. Methodology
D. Future directions
E. Acknowledgements

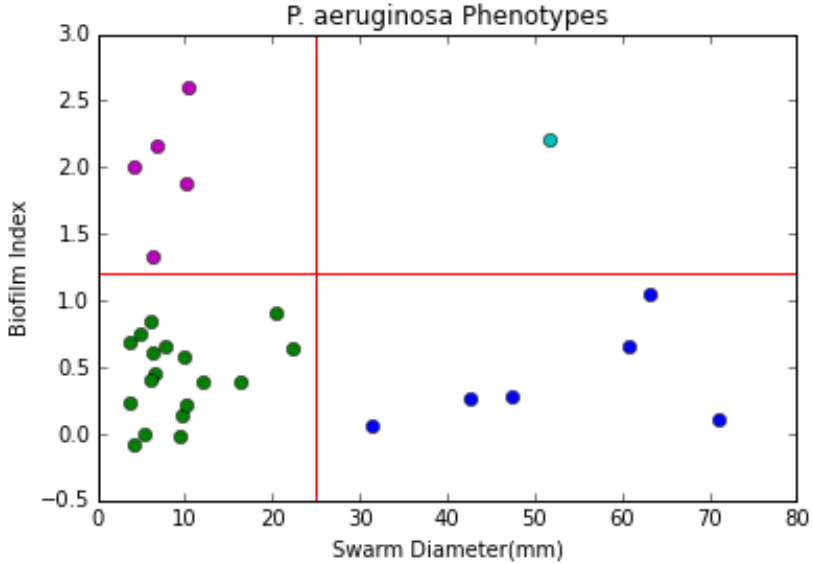
A. Motivation

Biofilm  **Swarming**

- It has been suggested that swarming and biofilm formation in *P. aeruginosa* are inversely regulated
  - Genetic changes promoting the switch from biofilm to swarming?
- Goal: *Use statistical analysis on a genomic scale to explore if these two traits are inversely regulated and identify causal links between genotype and phenotype*

Phenotype clustering

- The 29 clinical isolates + PA14 cluster by phenotype
- No strong trend suggesting a tradeoff or co-occurrence of biofilm and swarming
- For the following analysis, phenotype was encoded in a binary fashion



Non-swarming, non-biofilm-forming - **phen(0,0)**

Swarming, non-biofilm-forming - **phen(1,0)**

Non-swarming, biofilm-forming - **phen(0,1)**

Swarming, biofilm-forming - **phen(1,1)**

Preliminary tests

- First naively tested, using linear regression, whether biofilm formation and swarming are anti-correlated
- Tested alone and with principal components as covariates (more on this later)
- Tested swarming vs. biofilm and biofilm vs. swarming
- Result: *No significant correlations*

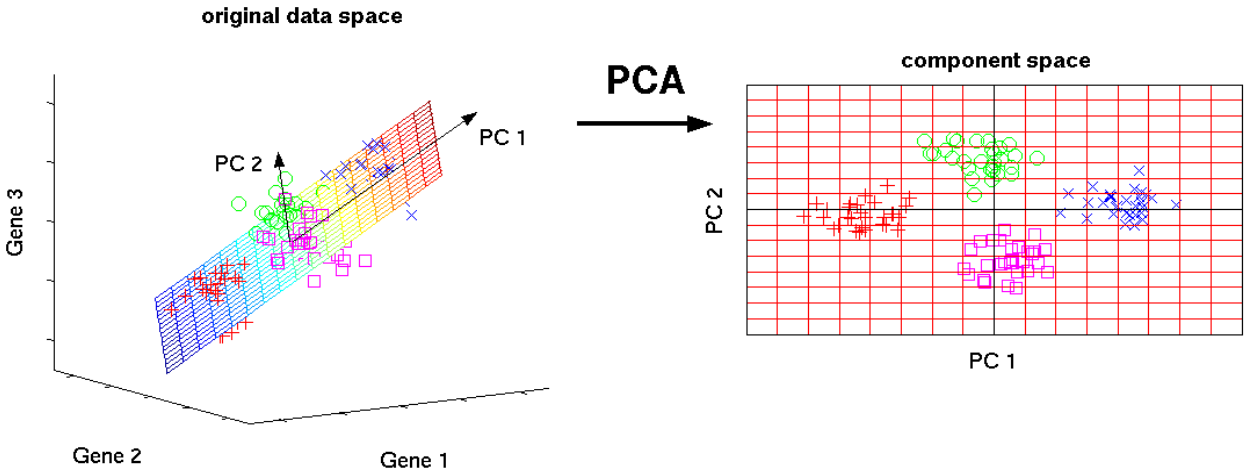
  → Null result prompts further, more detailed analysis

Genomic analysis and phylogeny

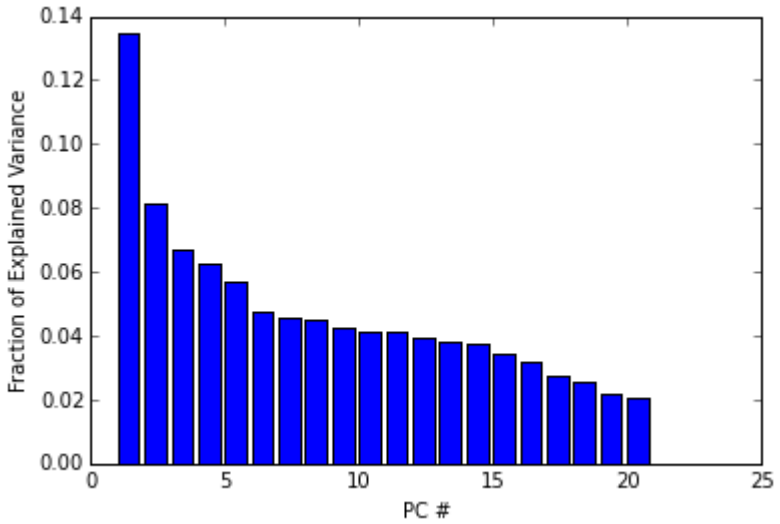
- Goal of genomic analysis - correlate genotype with phenotype (biofilm formation and swarming)
- Problem: Phylogeny
  - Some of the strains of *P. aeruginosa* are highly related to each other
  - Hypothesis testing without accounting for inherent genotype-genotype correlations generates false positives
- Solution: Model variation as principal components

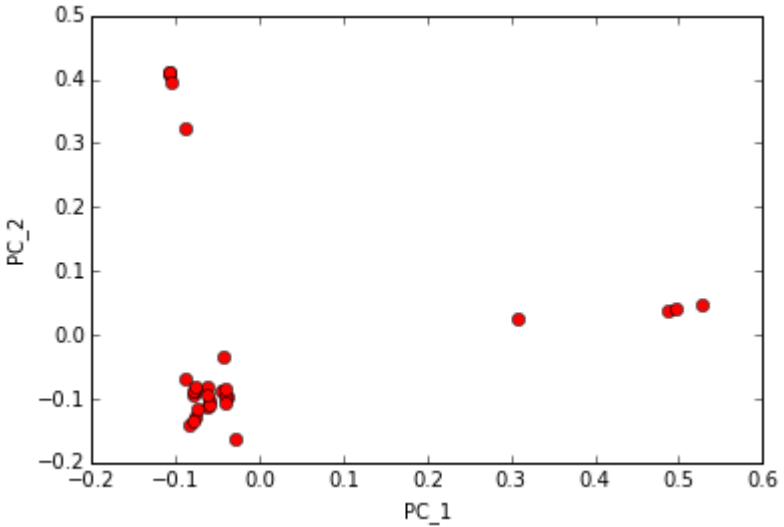
Principal Component Analysis (PCA)

- Goal of PCA - dimensionality reduction through a change of basis
- Recasting high-dimensional data onto the axes of greatest variability preserves the structure of the data



Genotypic variation in *Pseudomonas*





~94% variance in 20 components; clustering apparent from the first two PCs

Linear regression

- We would like to create a linear model against which we can test each genomic feature (N independent tests):

$$\mathbf{y} = \beta_{gen}\mathbf{x} + \boldsymbol{\epsilon};\ \boldsymbol{\epsilon} \sim N(0, \sigma_{\epsilon}^2)$$

- However, a direct test will produce spurious FPs
- Instead, model principal components (3) as covariates:

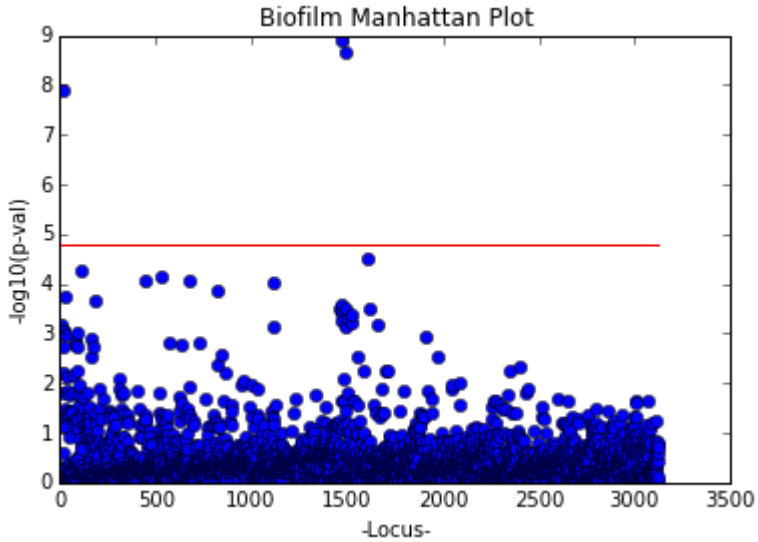
$$\mathbf{y} = \beta_{gen}\mathbf{x} + \boldsymbol{\beta}'\mathbf{z}_{\mathbf{PC}} + \boldsymbol{\epsilon}$$

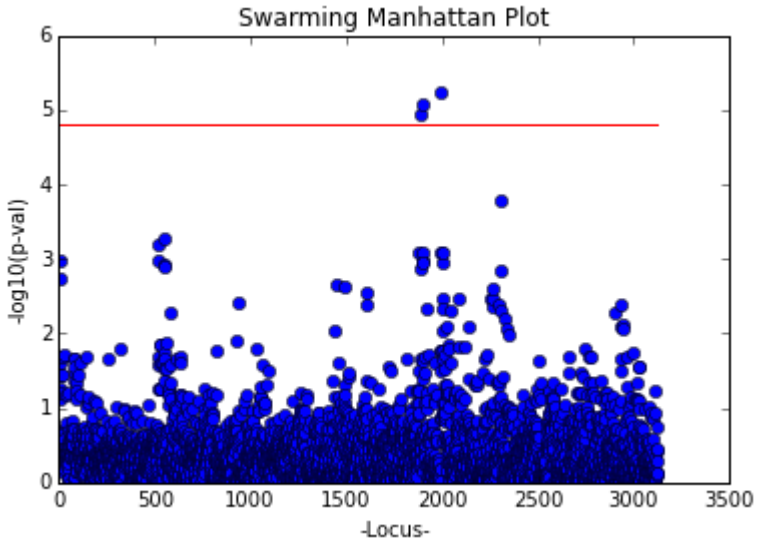
- Assumption 1: Most of the variation in the data is due to phylogeny and not related to phenotype
- Assumption 2: Data normally distributed*

Hypothesis testing

- Hypotheses:
  - Null (H_0): $\mathbf{y} = \boldsymbol{\beta}'\mathbf{z}_{\mathbf{PC}} + \epsilon$
  - Alternative (H_A): $\mathbf{y} = \beta_{gen}\mathbf{x} + \boldsymbol{\beta}'\mathbf{z}_{\mathbf{PC}} + \epsilon$
- Likelihood Ratio Test:
  - $L_0 = P(\boldsymbol{\beta}', \beta_{gen}=0|\mathbf{y})$, $L_A = P(\boldsymbol{\beta}', \beta_{gen}\neq 0|\mathbf{y})$
  - $LRT = 2\ln(L_A) - 2\ln(L_0)$
- The LRT statistic should follow a chi-squared distribution with one degree of freedom

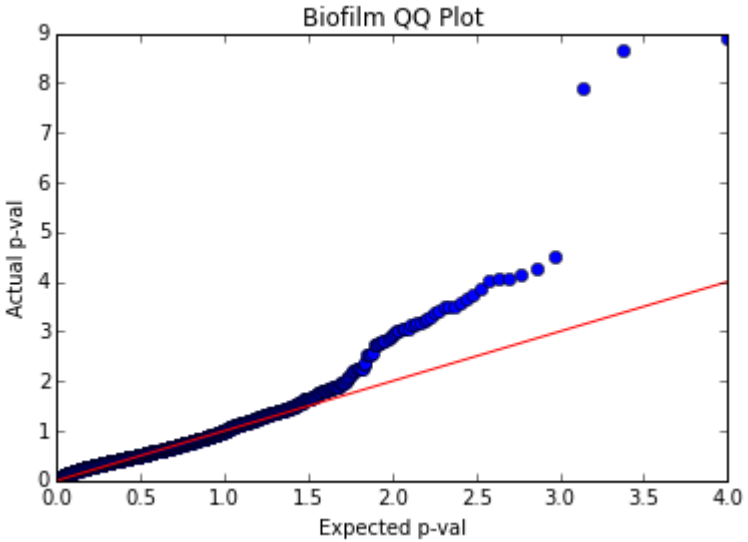
B. Manhattan Plots

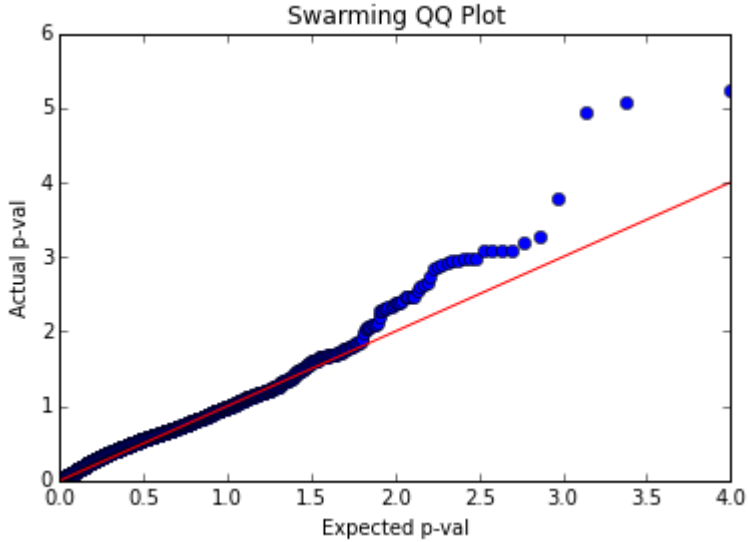




- There are 3 strong hits for biofilm and 3 weak hits for swarming
- Bonferroni cutoff = 4.795, significance $p<0.05$
  - Bonferroni criterion = $-\log_{10}(\alpha_{desired}/N)$

QQ Plots - Normality test





The QQ plots are both mostly linear, suggesting the data are normally distributed

Significant (p_{corr}<0.05) hits

Biofilm Formation		Swarming	
Feature	-log10(p-value)	Feature	-log10(p-value)
cdhit 2305	7.89	cdhit 5678, T54N	5.25
cdhit 487, A359T	8.91	cdhit 2626, E71V	5.08
cdhit 2590, A58V	8.91	cdhit 955, A160V	4.93
cdhit 4403, D206G	8.66		

Have yet to be evaluated...

C. Encoding binary genotype 1

- Query the database (maintained by collaborators at HC)
  - Contains genomic sequence data for 39 *Pseudomonas* strains
  - >1900 fully-aligned, high-quality open reading frames for sequence comparison
- Example:
  - ‘SELECT genome_id, swarm_diameter FROM phenotype;’
  - Outputs a table containing the strain id and the swarm diameter from the ‘phenotype’ table

Encoding binary genotype 2

- Convert table of gene, genome id pairs to a binary matrix indicating the presence or absence of that gene in all strains
  - Ex. if gene 1 occurs in strains 1 and 3 but not 2, 4, and 5, the row [1 0 1 0 0] is added to the matrix
- Identify core genes (i.e. present in all strains)
- For core genes with in-frame, aligned sequences in the database, identify all non-synonymous point mutations and encode their presence/absence

Encoding binary genotype 2

- Convert table of gene, genome id pairs to a binary matrix indicating the presence or absence of that gene in all strains
  - Ex. if gene 1 occurs in strains 1 and 3 but not 2, 4, and 5, the row [1 0 1 0 0] is added to the matrix
- Identify core genes (i.e. present in all strains)
- For core genes with in-frame, aligned sequences in the database, identify all non-synonymous point mutations and encode their presence/absence

Encoding binary genotype 3

- In all there are 17244 genomic features (pan genes and core gene mutations) for the 30 strains for which phenotype information is available
  - Database is continuing to expand as sequence information becomes available
  - Additional phenotypes and strains may be added
- General platform for generating the binary matrix is coded in Python

Hypothesis testing specifics

- General PCA packages (in Python) fail to calculate PCs when samples << features
  - Used scikit.learn ‘Randomized PCA’ package; algorithm for approximating PCs
- Packages patsy and StatsMethods used for regression
  - Facilitates regression using R-like syntax and unique, more convenient operators
- Pandas used throughout; contains Series and DataFrame objects that function as indexed numpy arrays

D. Future directions - Investigate hits

- Genes may or may not be annotated in global data
  - What is their function if annotated?
  - Do they have orthologs in other strains/species?
- For mutations, examine sequences microscopically and evaluate whether these hits are high-quality
- Test hits experimentally in *P. aeruginosa* and evaluate phenotype!

Future directions - Further analysis

- Supplement PCA-based regression with simpler methods
  - Naive linear regression with no covariates
  - Binary correlations; find features with the fewest number of differences from biofilm pattern
- Use other sophisticated models - Maximum Entropy

Future directions - Infrastructure

- Code for generate genotype matrix packaged
  - ~2-3 minutes to run
  - Additional functionality: use MUSCLE to align imperfect sequences for added information
- Build package for downstream analysis to generate results quickly and easily
- Build Max. Entropy package in Python

E. Acknowledgements

- Joao Xavier
- Xavier Lab Members
  - Jinyuan Yan
  - Maxime Deforet
- Weigang Qiu and the Qiu Lab
- Cornell University, Tri-I CBM Program
- MSKCC