*Pseudomonas aeruginosa* genomics project

Overview

--------

Within this directory are Python tools used to conduct genotype-phenotype studies in *P. aeruginosa*, which are designed to be adaptable to new data sets (e.g. strain sequences, phenotypes).

There are several presentations included that give context for various aspects of the project and readers are encouraged to refer to them for more background on this project and the statistical methods used.

All of these files will be available on https://github.com/justin-torok/Pseudomonas-GWAS

Python Files

------------

pseudomonas\_gwas.py:

I have tried to thoroughly annotate my classes and methods within the file itself, so I will just give a brief overview here. The goal of this project was to correlate the swarming behavior and biofilm formation in clinical isolates to genetic features (accessory genome content and SNPs), anticipating additional phenotypes and genomes as they became available. To those ends, the PseudomonasDataframes class creates a pandas (http://pandas.pydata.org/) Dataframe (among other objects) containing all differential genomic markers. It also outputs several plots of interest, including rarefaction curves. There are three linear regression methods used here: ‘simple’, PCA, and covariance-corrected. PCA\_Regression and GLS\_Regression perform these analyses and can calculate p-values; PCA\_Plotter and GLS\_Plotter create QQ and Manhattan plots for each method.

GenomeExporter creates the two kinds of sequence files within the PA\_Genome\_Files folder (please refer to its README for additional information).

SpringDb.py and SpringDb\_local.py:

These scripts were created by Dr. Joao Xavier. The only method relevant to pseudomonas\_gwas.py is a simple PSQL (http://www.postgresql.org/docs/9.2/static/app-psql.html) database accession using psycopg2 (http://initd.org/psycopg/), which can (perhaps should) be wrapped within that script to eliminate this dependency. There are two corresponding configuration files containing the necessary accession information.

All of the raw data and sequence information is kept within the database **pa2**, which is maintained by Dr. Weigang Qiu and his lab at Hunter College. To eliminate the bottleneck of communicating with the remote server, I downloaded local copies of the necessary tables and views, which is highly recommended for any future users as well. The most recent copies of these files, which can be imported into a local PSQL database, are provided in the PSQL\_Files directory. Please refer to the psycopg2 and PostgreS documentation for further information on how Python interfaces with PostgreS.

PA Genomes

----------

There is a separate README within this directory describing the format of the genome files therein.

Graph Lasso

-----------

This folder contains the complete covariance and precision matrices for the current batch of 30 clinical isolates, as well as several related slides and figures, which were calculated using the Graphical Lasso: http://statweb.stanford.edu/~tibs/ftp/graph.pdf. All calculations were done using the complete genotype Dataframe from pseudomonas\_gwas.py and the GraphLassoCV tool from scikit learn package: http://scikit-learn.org/stable/modules/generated/sklearn.covariance.GraphLassoCV.html.

Author

------------

Justin Torok

Acknowledgements

----------------

Dr. Joao Xavier

Rayees Rahman

Dr. Weigang Qiu

Contact

-----------------

jlt46@cornell.edu

Date

----

2/15/16