bgwas3: a pipeline for kmer based association testing in bacteria

Gregory Leeman

Contents

1 Abstract

Background

What i did

Results

Future

Genome wide association studies are applied extensively to human populations, and aim to to identify the genetic basis of phenotypic traits. Because of clonal reproduction, bacteria experience greater population structure and genetic variation, and the use of single nucleotide polymorphisms is redundant. Only recently have GWAS-like studies started appearing for bacteria. These new 'pangenome wide' association studies facilitate the use of association-free K-mers as units of variation as apposed to SNPs.

I developed, tested, and compared three pipelines for pGWAS using different association models, and using different open source tools to confound for population structure. I also attempted to use these pipelines to test antibiotic resistance traits in a sample of Pseudomonas aeruginosa from the sputum of cystic fibrosis patients; a condition where antibiotic resistance is a significant issue.

I find that performing a mixed linear model with random effects calculated from estimated phylogenies based on either differences in gene content or difference in core genome SNPs seems to confound for population structure better than a linear model where significant principal components are used as fixed effects.

Though I identified significantly associated K-mers, I was not able to map these K-mers to a reference genome; limiting my ability to perform downstream analysis of significant loci. However, I believe that this pipeline may be refined, given more time, into completed into a robust and useful work flow that will simplify pGWAS in the future.

I also hope this report may also serve as a description and evaluation of pGWAS.

2 Acknowledgements

3 Introduction

4 Abreviations

SNPs

4.1 Pangenome wide association studies

TODO make sure not too much like pyseer paper

When used for phenotupes relatng to disease (virulence etc ?) this may help 'clinical interventions'.

There is a recent relative abaundance of whole bacterial genome and phenotype data (from metabolomics)/ rapidly expanding repositories of genomic data for bacteria

Studies that attempt to investigate the genetic causes of traits in bacteria often focus on identifying associated clonal groups as apposed to specific loci.

Genome wide association studies can be used to identify genetetic and phenotype assocations in a hypthoses free manner.

Commonly, SNPS have been used as units of genetc variation of which assocation to phenotypes have been tested.

Becasue bacteria produce clonally, significat regions of the geneome can be in linkage disequilibrium. And though some species experience high rates of recombination, the recombination is not as reliable or consistent as that in, say, humans to reduce LD.

Hard to do this in bactera due to recombination meaning hard to align genes to get snps,

Tools exist for the alignment of the core genome of bacteria to extract SNPs (ref), this is an alignement based method/ requires alignment is limited to the core genome, and does not consider the huge variable accessory genome in bacteria/ will not encapsulate the ful0

Similarly, the presence/ and or absence of known genes have been used.

The problem with both these methods is due to features of bacteria such as clonal reporduction and recombination?, bacterial geneomes vary significantly

As an alternative, some recent studies have instead utlised K-mers. Initially a concept of genome assembly Can caputre multiple geneetic variations SNPs, longer deletions/ insertions and recombination Getting kmers is alignment freee The size of kmers effects the genetic variation captured Longer more specificS Shorter are more sensitive

Seer (se (ref) and its reimplementation

As discussed in my previous project,

4.2 Metabolomics

5 Methods

The foundation of *bgwas3* is an integration of mutltiple open source genomics tools and custom scripts in python and R into a single installable package which facilitates conducting musltiple pangenome wide association studies in a single user step. The tool aims to be comprehensive and robust enough to convert only the most basic required input files into comprehensive and interpretable results including static and interactive visualisations. Similarly, when run on a node within a computer cluster, tasks that are computationaly intensive or long can be run simultaneuosly. TODO better word for long Also, though the additional confugiureation file, the specifics of multile intermiedary steps can be altered, which can have variable effects on the final results. In light of the necessity to peform multpple.. (1)

5.1 CGATcore and Ruffus

TODO how do pipeline refer to cgat core

5.2 Genome Annotation

Prokka (Seemann 2014) TODO write about

6 Kmer minig and counting

bgwas3 integrates the tool fsm-lite to count kmers of user defined variable length kmers in the all samples. fsm lite runs on a single core. Throught the bgwas3 configuration file, the specifics of the kmers mined and counted can be

6.1 Phylogeny estimation and generating covariates

For linear mixed model assocation testing, a bgwas3 constructs a distance matrix based on a phylogenetic tree.

This is different to the standard method of multi-dimensional scaling often used in human gwas and as part of the standard seer workflow.

As found in my previous study (ref),

The

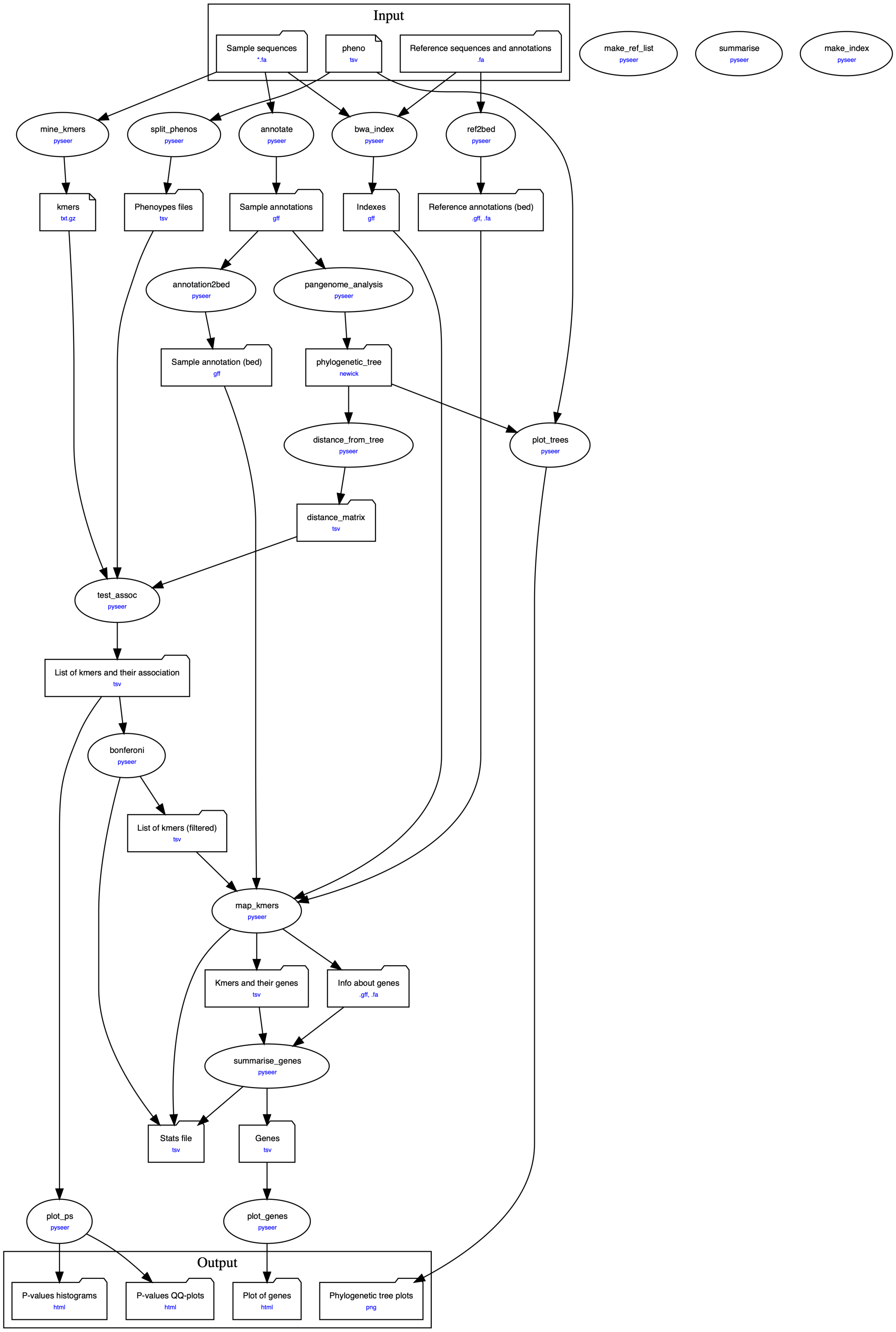

Figure 1: Pipeline of bgwas

6.2 Kmer association testing

Assocation between kmers and phenotypes is implemented with pyseer (Lees et al. 2018).

Specificially, pyseer is used to peform a linear mixed model using the Fast-LMM algorithm (Lippert et al. 2011).

LMM tackle confounders in association tests by using a measure of similarity (in this case a distance defined from a estimated phylogenetic tree) as a random effect in a linear model.

TODO equation

Given a means of determining the hierarchichal relatedness between samples, the mixed model is generally preffered, and has been shown in past studies to control the inflation of p-values better (Lees et al. 2017).

6.3 Bonferoni correction

The output of pyseer, a list of all kmers and statistics relating to their association to the given phenotype, are then filtered by their p-value though bonferoni correction.

6.4 Significant Kmer mapping

Kmers A new script, written in R, was made which itereatively TODO workflow diagram

6.5 Pathway analysis

6.6 Visualisation

6.7 Scienitif computing practices

A significant goal of bgwas3 was to implement a useful and resuable tool that may be used outside the scope of this singlular project.

As such, various software development principles and concepts were applied to the project.

6.7.1 Testing

Pytest and

6.7.2 Documentation

6.7.3 Packaging

6.8 Use case

7 Results

TO TODO describe the data

TODO figures (trees)

TODO

To test the usability and application of bgwas3,

The strong LD caused by the clonal reproduction of bacterial populations means that non-causal k-mers may also appear to be associated.?

Confirmation of known resistance # Discussion # References

Lees, John A, Nicholas J Croucher, David Goldblatt, François Nosten, Julian Parkhill, Claudia Turner, Paul Turner, and Stephen D Bentley. 2017. "Genome-Wide Identification of Lineage and Locus Specific Variation Associated with Pneumococcal Carriage Duration." Edited by Sarah Cobey. *eLife* 6 (July): e26255. https://doi.org/10.7554/eLife.26255.

Lees, John A, Marco Galardini, Stephen D Bentley, Jeffrey N Weiser, and Jukka Corander. 2018. "Pyseer: A Comprehensive Tool for Microbial Pangenome-Wide Association Studies." Edited by Oliver Stegle. *Bioinformatics* 34 (24): 4310–2. https://doi.org/10.1093/bioinformatics/bty539.

Lippert, Christoph, Jennifer Listgarten, Ying Liu, Carl M Kadie, Robert I Davidson, and David Heckerman. 2011. "FaST Linear Mixed Models for Genome-Wide Association Studies." *Nature Methods* 8 (10): 833–35. https://doi.org/10.1038/nmeth.1681.

Seemann, Torsten. 2014. "Prokka: Rapid Prokaryotic Genome Annotation." *Bioinformatics* 30 (14): 2068–9. https://doi.org/10.1093/bioinformatics/btu153.