bgwas3: a pipeline for kmer based association testing in bacteria

Gregory Leeman

Contents

1 Abstract

Genome-wide association studies (GWAS) are now being applied more frequently to bacterial populations. Specificially, alignment free association studies based on mixed width K-mers as apposed to short nucleotide polymorphisms are being deployed, and packages such as *pyseer* are making it possible. However, there is not at present a succint tool which pipelines the necessary intermediry data processing and analysis steps. Let alone combines all these tests and allows for multiple association studies to be peformed at once, and distrubuted onto multiple nodes in a computer cluster.

which has utility in a time metabolomics data is becoming more common, and so the need to determine the genetic bases of possible hundreds of phenotypes as once is a requreid.

I researched, developed, and wrote and installable pipeline tool, *bgwas* which can test the association of multiple phenotypes at once to genetic loci. The tool integrates various open source software tools for Kmer counting, gene annotation, phylogeny estimation, kmer association testing, kmer-gene mapping and finally visualisation.

In creating the tool, best practices for computational biology where excercised; resulting in a final tool that expresses approprate testing, documentation and packaging; meaning it can easily be utilised by other in the future.

The tool was also used to test the assocation of n phenotypes - some relatiing to antibiotic resistance and others metabolomics measures - and the outputs suggest the tool is usuable, but similary provides insights into future improvements that could be made

2 Acknowledgements

3 Abreviations

SNPs

4 Introduction

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

4.2 The Need for Multiple Assocation Tests - Metabolomics

Metabolomics provides a detailed snapshot of an organism's physiological state through the quantification of hundreds to thousands of small molecules (Zampieri et al., 2017). Genomics on the other hand, can help us understand the function and evolution of an organism with unprecedented power (Marvig et al., 2015). Coupling genomics and metabolomics to study Pseudomonas aeruginosa in unique clinical samples can allow us to understand how bacteria adapt to the lung environment. In order to do so, we will use differential abundance, quantitative trait loci (QTL) and bacterial genome-wide association study (GWAS) methods (Power et al., 2017).

4.3 Results

The foundation of this project was to create a tool,

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 multiple 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

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 Methods

5.1 CGATcore and Ruffus

TODO how do pipeline refer to cgat core

bgwas3 was written primarily in python, utilising the pipline tools ruffus (ref) and cgatcore (ref). Ruffus Cgatcore

In its current iteration, *bgwas3* is comprised of n steps

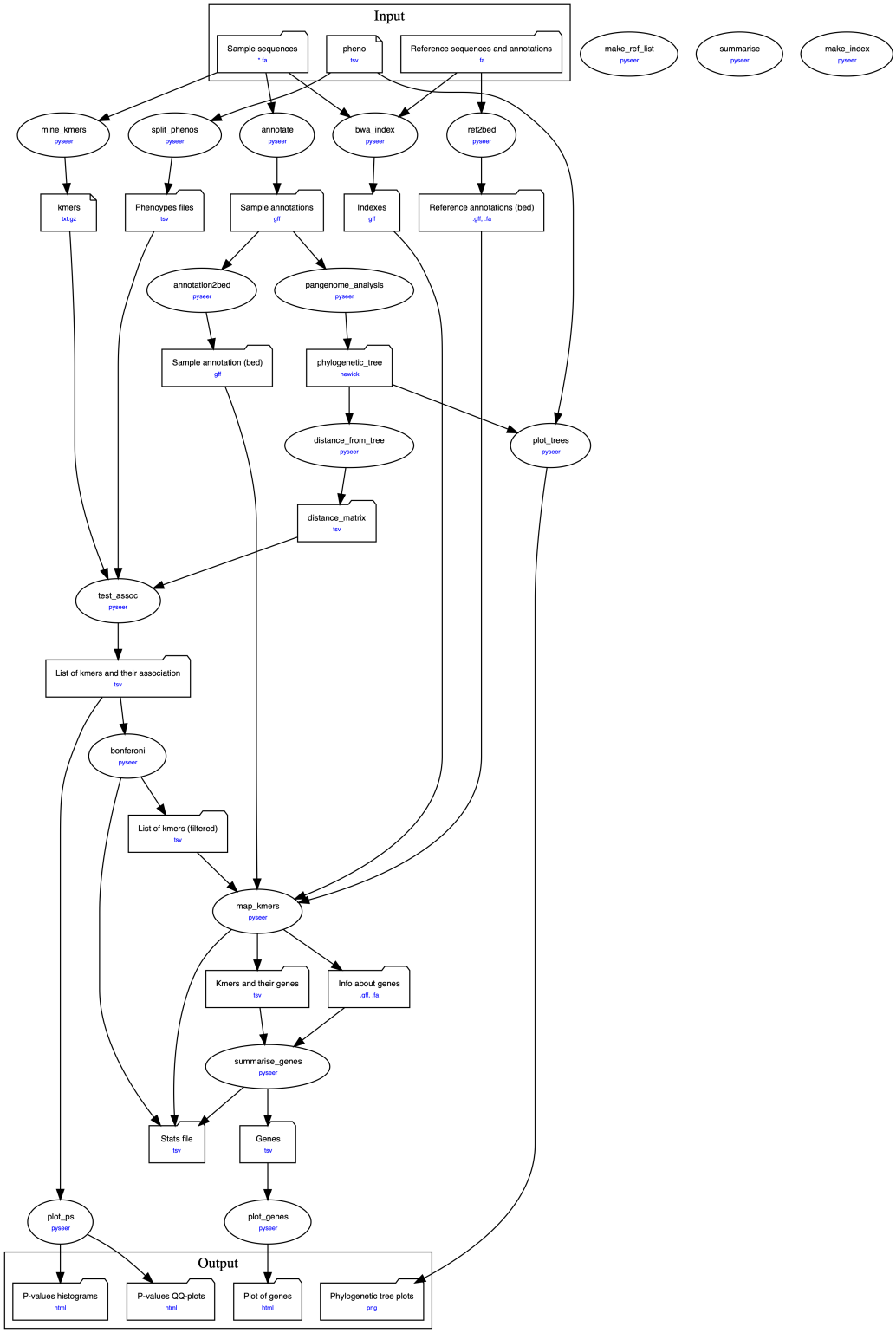

Figure 1: Pipeline of bgwas

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

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 with PyTest and Travis

Pytest and

6.7.2 Documentation with Sphinx

6.7.3 Packaging with Conda

7 Results

Pseudomonas aeruginosa forms chronic infections in the lungs of cystic fibrosis (CF) patients, and is the leading cause of morbidity and mortality in patients with CF. Understanding how this opportunistic pathogen adapts to the CF lung during chronic infections is important to increase the efficacy of treatment and is likely to increase insight into other long-term infections. TODO rewrite Pseudomonas aeruginosa forms intractable infections in around 80% of patients with cystic fibrosis (CF). Once infection is established within the respiratory tract, the bacterium is able to withstand both attacks from the host immune system and prolonged exposure to antibiotics. These chronic infections are associated with decreased lung function and an increased risk of respiratory failure and death (Schaedel et al., 2002; Hauser, 2011). As a result, much research has focused on understanding the factors that enable the opportunistic pathogen P. aeruginosa to play a dominant role during long-term infection of the CF lung. T We used untargeted metabolic profiling (metabolomics) of cell supernatants (exometabolome analysis, or metabolic footprinting) to compare 179 strains, collected over time periods ranging from 4 to 24 years for the individual patients, representing a series of mostly clonal lineages from 18 individual patients. There was clear evidence of metabolic adaptation to the CF lung environment: acetate production was highly significantly negatively associated with length of infection.

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

8 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.