Imperial College London

Department of Life Sciences

M.Sc. Bioinformatics and Theoretical Systems Biology

bgwas3: a pipeline for kmer based association testing in bacteria

Author: Gregory Leeman

Supervisor: Prof. Dr. Antonio Berlanga

2nd September 2019

Word count:

Acknowledgements

Abstract

Genome-wide association studies (GWAS) are now applied more often to microbes. However, bacteria genomes are significantly more variable in both content and sequence than eukaryotic genomes, meaning these association studies have naturally differnt, more complex considerations. These studies utilise short unique DNA patterns, reffered to as Kmers, as units of genetic variation as apposed to single nucleotide polymorphisms. Various software solutions are now available which can pefoerm the multiple assocation tests of kmers, notable the python based *pyseer*. However, there is not at present a succint tool which pipelines the necessary pre and post data processing and analysis steps such as associating the Kmers to genes, that make an association study whole.

In metabolomic studies of bacteria there are thousands of possible measures that can be made which can be considered as traits of that bacteria at a that point in time. Coupling genomics and metabolomics may provide a powerful technque to understanding how a bacteria has adapted, and in the scope of disease, may aid in the understanding of the function of drug resistance and other virulence features. However conducting indivudal assocation tests for each trait may prove time and computationally intesnsve.

I developed a tool, *bgwas2*, which wraps the Kmer association function of *pyseer* with other open source software tools into a single installable package. When run, *bgwas3* takes the simplest of input files and a configuration file, and peforms gene annotation, phylogeny estimation, kmer association testing, kmer-gene mapping and finally visualisation with a automatically generated web report.

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 26 traits - 18 relatiing correspjoning to a metabolomic measurement, and 8 corresponding to either a measure of to antibiotic resistance or bacterial motility. In 14 of the traits, between 1 and 16450 significant Kmers were found, and were mapped to Genes.

Contents

Acknowledgements **1**

Abstract **2**

Abbreviations **4**

List of Figures **5**

List of Tables **5**

1 Introduction **6**

1.1 Pangenome wide association studies 6

1.2 The Need for Multiple Assocation Tests - Metabolomics 7

1.3 Scope of work 7

2 Implementation **9**

2.1 Pipelining with CGAT-Core 9

2.2 Genome Annotation 9

2.3 Kmer minig and counting 10

2.4 Phylogeny estimation and covariate generation 10

2.5 Kmer association testing 10

2.6 Bonferoni correction 11

2.7 Significant Kmer mapping 13

2.8 Visualisation 13

2.9 Scienitif computing practices 14

2.9.1 Testing with PyTest and Travis 14

2.9.2 Documentation with Sphinx 14

2.9.3 Packaging with Conda 14

3 Results **14**

3.1 Processing of phenotypes 16

3.2 Annotation 16

3.3 Kmer mining 20

3.4 Association results 20

4 Discussion **21**

4.1 Kmer mining 21

References **22**

Abbreviations

GWAS Genome-wide assocation study

DRMAA Distributed resource management application

List of Figures

1 Pipeline of bgwas 8

2 Graphical representation of algorithm to map Kmers to reference genes 12

3 Correlation matrix of phenotypes 17

4 Density plots 18

5 Density plots 19

6 Gens 20

List of Tables

1 Table of non-metabolite phenotypes 15

1 Introduction

1.1 Pangenome wide association studies

Genome wide association studies have been successfuly used to identify genetetic and phenotype assocations in a hypthoses free manner. Characteristics of microbes suggest they would be ideasl candidates for GWAS: their genomes are small which limits the study size; they exude much less phenotypic flexibility; and, generally, phenotypes that are of interest such as virulence ans drug resisistance tend to be the result of strong selective pressures. There is also good reason to peform assocation studies with bacteria. as identifying the genetic basis of traits that correspond with its pathogenetic ability could lead to better understanding of disease an lead to new targets for pharmaceutical intervention

Commonly, GWAS have focued on single nucelotide polymorphisms (SNPs) as units of genetic variation. SNPs are useful in organsims where recombination is reliable and genomes are relatively consistant among populations. Because of linkage disequilibrium, identigying a highly associated SNP often means that the true associated loci is nearby

However, the nature of bacteria replication makes the use of SNPs in association studies less useful. For one, 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 in eukaryotes to reliably reduce linkage.

Secondly, bacteria experince much higher trans? during recombination. Therfore, bacteria of the same species? can vary consideriably in both order of gene content, and in the accessory genome - regions of the geneome which differ.

Therfore, first finding SNPs becomes a tricky process which involves multiple alignments. Such tools such as mummer (and other?) do meet this problem, but even still, SNPs assocated with ony the core genome can be retrieved and does not consider the huge variable accessory genome in bacteria/ will not encapsulate the ful0

As an alternative, some recent studies have instead utlised K-mers.

Kmers are simply variable length sequences of DNA Initially a concept of genome assembly Can caputre multiple geneetic variations SNPs, longer deletions/ insertions and recombination The size of kmers effects the genetic variation captured Longer more specificS Shorter are more sensitive

Getting kmers is alignment freee, releasignt he burden of ultie alighnemt But more specifically utilisng kmers

means the core ans accesory geneome can be tested.

Seer (Lees et al. 2016) and its python reimplementation *pyseer* (Lees et al. 2018) dbgwas

1.2 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).

Coupling genomics and metabolomics to study Pseudomonas aeruginosa in unique clinical samples can allow us to understand how bacteria adapt to the lung environment.

The nature of metabolimics measn that multiple molecules abundance may be measured and of interest to investigate.

Determint the genetic basis of metabolomics of P ing the context of chronic invextions

1.3 Scope of work

The foundation of this project was to create a tool, *bgwas3*, an integration of mutltiple open source genomics tools and custom scripts written in python, R and bash into a single installable package that 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 valuable and interpretable results including static and interactive visualisations.

When run on a node within a computer cluster, tasks that are computationaly intensive or long are distrubuted among the cluster, and can run simultaneuously.

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)

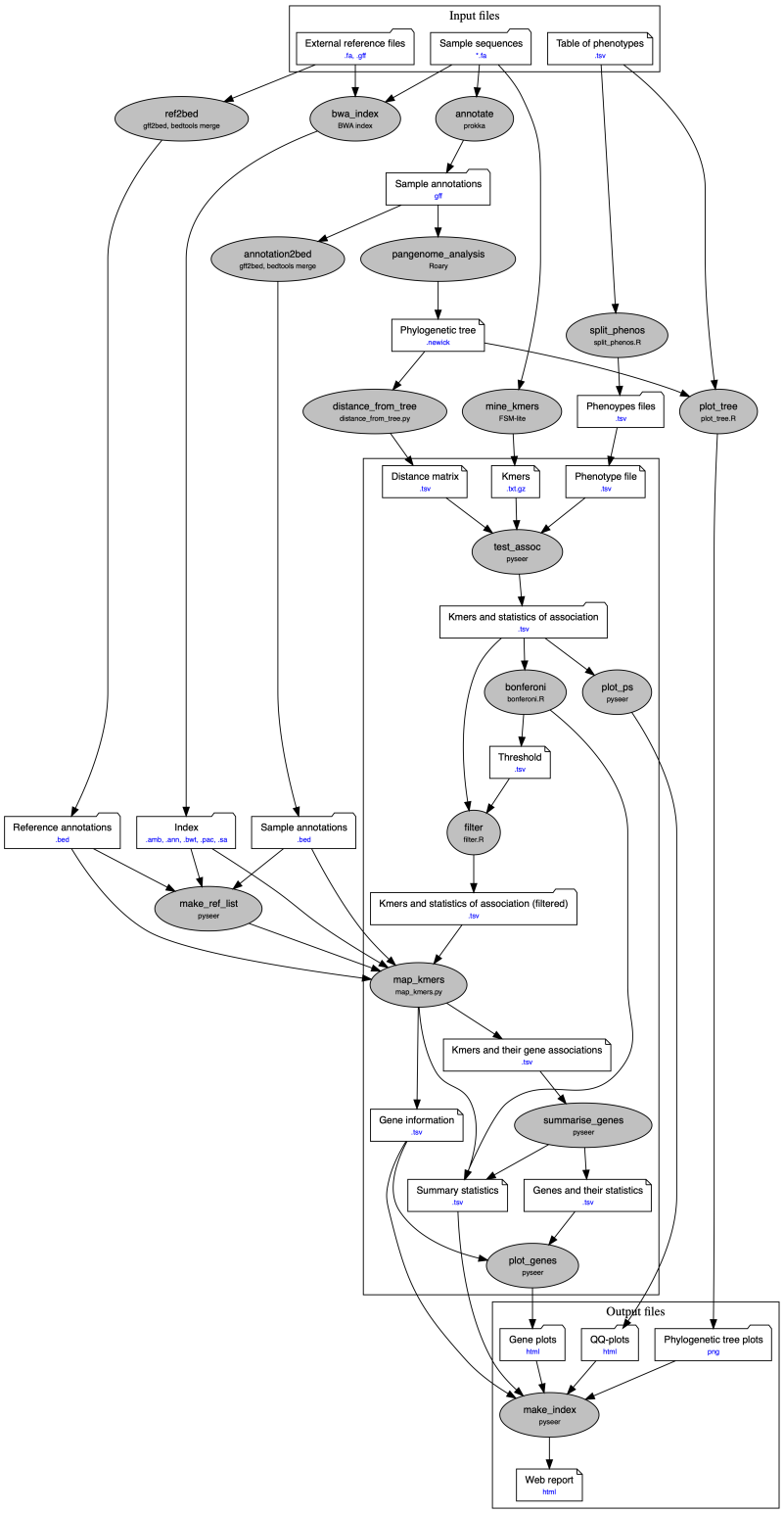

Figure 1: Pipeline of bgwas

2 Implementation

2.1 Pipelining with CGAT-Core

*bgwas3" was made utilising the piplining framework CGATCore, (Cribbs et al. 2019). CGATCore is extremeley portable package coming as a dependency free Python package, but has been used to custruct many complex pipelines that are scalable to large amounts of data ?ref

With the framework, individual pipelining steps, referred to as 'tasks', are defined as python functions whose arguments include the necessary input files and the expected output files.

With the use of python decoraters, the output files of one task can be used as the input arguments of later tasks. This allows multiple data anlysis to be strung together and run in the correct order. The framework utilises a system of checking the modifcation date of files, so only runs tasks in which their are either new input files, or input files have been modified. This checking system means a directory of project files may be moved from one location to another, even between computer,s and the framework will still recognise which tasks and steps are out of date.

CGATCore introduces another significant feature which is an interface to control a distributed resource management management application API (DRMAA) such as PBS-Pro used by Imperial College Londons resource computing service. Therefore, tasks which are deemed large are need to be run at batch, parallel processes can be distributed to computers in a cluster, and the system resource (eg RAM) can be set on a per task-basis. This makes executing pipeline that involve cimputationally intesive intermediary steps faster.

bgwas3 was written as 18 'tasks' which are visualised in 1.

While the tasks were written in python, many individual tasks made calls to external tools or scripts I wrote in either pyton, R and bashscript.

The program requires a directory named 'contigs' with fasta files uniquely named for each bacteria sample, an optional directory of reference sequences with accompanying annotations in gff format, and finally a single tab seperated value file (TSV) in which the first column matches the names of sample fasta files, and remaining columns correspong to phenotype values.

2.2 Genome Annotation

In the *bgwas* task "annotatate" all sample sequences are annotated by the external tool Prokka (Seemann 2014) Prokka first makes predictions on features using a number of external tools including Prodigal [@]

for the anotation of coding region, RNAmmer for ribosomal RNA genes, Aragorn for transfer RNA genes, SignalP for signal leader peptides and Infernal for non-coding RNA. After feature prediction, the speculative features are queryed against a umber of databes suchs as UniProt [@], Refseq [@] and Pfam.

The function of annotating input sequences in the *bgwas3* pipeline is twofold. First, gene annotations are later used to estimate the phylogenetic tree of the sample, and secondly, significantly associated Kmers are later mapped to the annatated genomes when identifying the Kmer's genetic identity.

2.3 Kmer minig and counting

bgwas3 integrates the external tool FSM-lite (Välimäki 2018) to 'mine' and count Kmers. A benefit of FSM-lite is, unlike other Kmer mining tools such as DSK (Rizk, Lavenier, and Chikhi 2013) is FSM-lite allows a range of kmer-sizes, as defined by the user, to be mined as apposed to a single length. *bgwas3* allows the Kmers length to be changed in the pipeline by editing the configuration yml file values 'fsm_kmer_min' and 'fsm_kmer_max'.

2.4 Phylogeny estimation and covariate generation

bgwas3 currently takes a pangenomic approach to phylogeny estimation; as in it consideres the relative prescence and absence of genes in the samples as factors in which to judge distance. An estimated phylogenetic tree is estimated by utilisng the tool Roary (Page et al. 2015). In summary, Roary determnes genes which fall within the coregeneome, and then peforms clustering of isolates based in the constetution of the variable accessory geneome. One of roary's outputs is a newick tree, which is then converted into a distance matrix using an adapted python script from pyseer.

The single reasoning for prediciting phylogeny in the *bgwas3* pipeine is to use the distances defined in the distance matrix as covariates int the assocation testing.

This is different to the standard method of correcting for population structrure of used in humsn studies

In these studies, multi-dimensional scaling is often used on the ...

2.5 Kmer association testing

bgwas implements only one of the three available tests of assocation availabe in the pyseer (Lees et al. 2018) suite: linear mixed model (LMM), specifically factored spectrally transformedas LMM (Lippert et al. 2011).

LMM is considered to be more robust method of association testing, especially when a source of phylogeny is available ?ref

basic linear regression linear regression, assume that the correlation between a phenotype and a genotypic marker exists because of the marker itself or a strong linkage disequilibrium with the causative locus. f the analysis is not accounted for confounding by population substructure, the test statistic is inflated [1], which makes its statistical interpretation difficult and may lead to false-positive findings.

LMMs can capture all of these confounders simultaneously, without knowledge of which are present and without the need to tease them apar

Roughly speaking, LMMs tackle confounders by using mea sures of genetic similarity to capture the probabilities that pairs of individuals have causative alleles in common. Such measures include those based on identity by descent 10

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).

TODO equation

In the *bgwas* task 'test_assoc' and association test for each phenotype on all the Kmers. In a computer cluster, these individual tests may be run simultaneuously.

2.6 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.

This is done in two steps. First the bonferoni threshold is calculated from the number of kmers minded.

Bgwas3 allows the user to change the alpha level

The Bonferroni correction is a multiple-comparison correction used when several dependent or independent statistical tests, utilised to limit the number of spurious positive tests. f multiple hypotheses are tested, the chance of a rare event increases, and therefore, the likelihood of incorrectly rejecting a null hypothesis (i.e., making a Type I error) increases

Only significant kmers are then used in later steps of the pipeline.

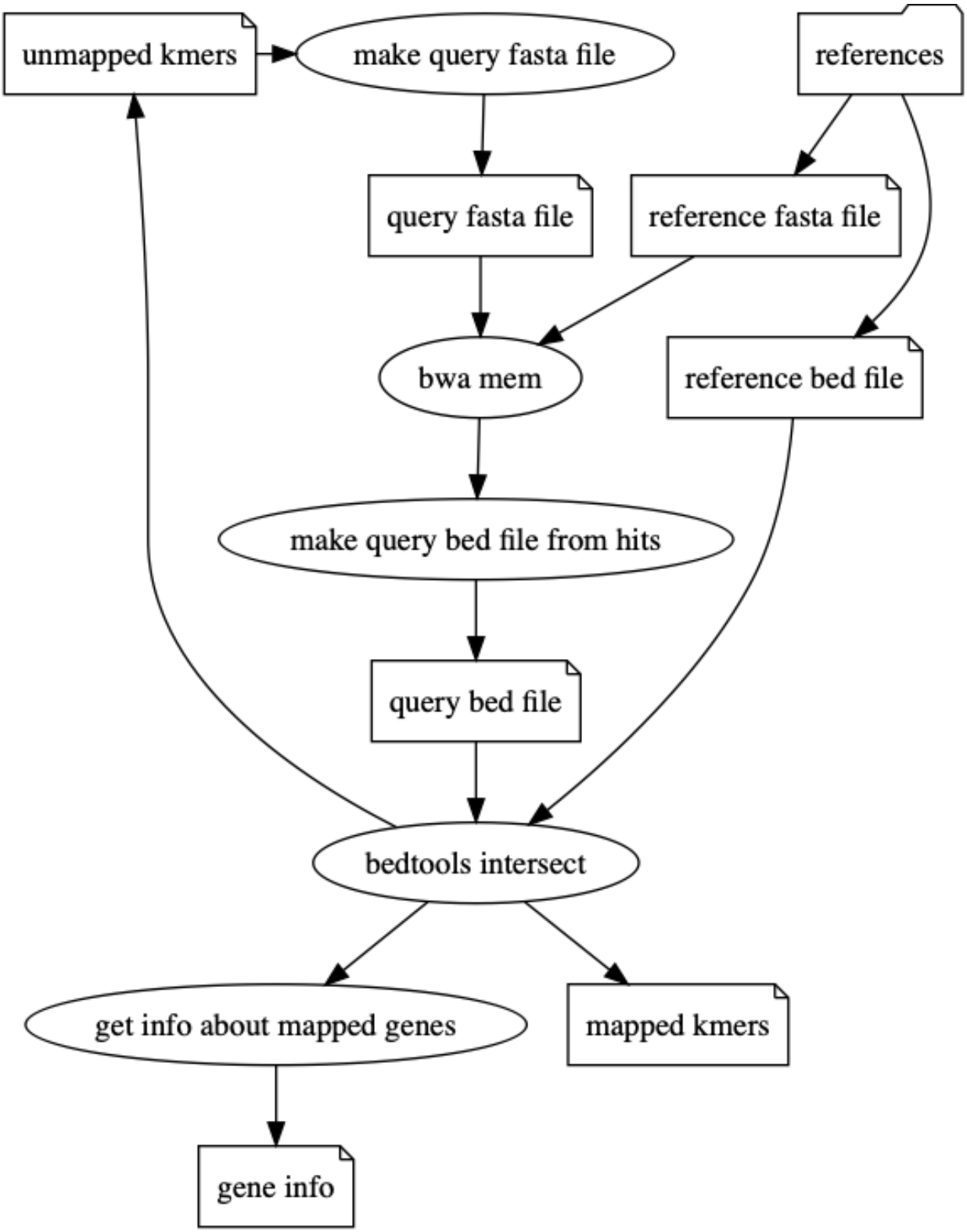

Figure 2: Graphical representation of algorithm to map Kmers to reference genes

2.7 Significant Kmer mapping

The Burrows-Wheeler Alignment Tool (BWA) is primarily used to map significant Kmers to genes.

The tool first required that sequences are first converted into an FM-index, an data-structure similar to a suffix array. This is peformed with the *bgwas* task 'bwa_index', on all sample and reference fast files.

For use with bedtools, all annotation files in gff are similaryl converted into bed format, and then subsequently filtered with the tasks 'annatation2bed' and 'ref2bed' repsectively. Sample annotation from prokka are filtered to include only those annotation which correspond to a named gene, and so exclude ones which relate to hypothetical proteins or non-gene loci. Reference annotations which correspond to the same region but are present as two sperate entries are merged, keeping the information from both, as this was found to be a common nomenclature in well annotated references.

The task 'map_kmers', executes a python script I wrote whose algorithm is visualised in **??**.

In summuary, the algorithm works as follows: - A fake multi-fasta file is generated with an entry for each unmapped, significant kmer - The next reference is chosen - The command-line tool 'bwa-mem (Li 2013) attempts to align to the reference - A fake bed file is generated with an entry correspoinding to each successful alignement with bwa mem - The query bed file is compared to the reference bed file using the command line tool 'bed tools intersect' - The information about each intersection is harvested and stored in a gene info file, while the kmer is marked as mapped Repeat until all kmers mapped or until all references have been used

2.8 Visualisation

An important feature of *bgwas3* is the automated generation of multiple figures, integrated into a final wbe based report.

Static visualisations are made with external scripts I wrote in R that make use of ggplot2 package (**???**).

The task 'plot_ps' generated a quantile-quantile plot from all p-values unfilterd. This may give See fig

A single phylogenetic tree is built from the newick file and phenotype file utiliseing external R package (**???**). See figs

Finally, for each phenotype, a plot of genes is made in which the following features are visually encoded

- Maximum -log10(p-value) of Kmers mapped to that gene
- Average beta value (effect size)

- Average allele frequency
- Number of kmers 'hits'

Finally, a web based report which incorporates all the static visualisations, and remakes an interactive visualisation anaologous tot he gene plot, is buil with the task "make_index".

A datatable which can be filtered by number of genes

The interactive visualisation for each gene Shows

2.9 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

CGAT-core is implemented in Python 3 and installable via Conda and PyPI with minimal dependencies. We have successfully deployed and tested the code on OSX, Red Hat and Ubuntu. We have made CGAT-core and associated repositories open-source under the MIT licence, allowing full and free use for both commercial and non-commercial purposes. Our software is fully documented (https://pypi.org), version controlled and has extensive testing using continuous integration (https://travis-ci. org/cgat-developers.) We welcome community participation in code development and issue reporting through GitHub

2.9.1 Testing with PyTest and Travis

Pytest and

2.9.2 Documentation with Sphinx

2.9.3 Packaging with Conda

3 Results

For the development, testing, and evaluation of *bgwas3* a dataset of genomes and phenotypes related to *Pseudomonas aeruginosa* was used.

92 strains were collected over a period of 24 years from 18 patients suffering from Cystic Fribrosis. ref

Cystic fibrosiis is a an autosomal recessive disorder that, due to a single gene mutation, limits the ings ability to Pseaudomonas euruginosa is one of the prmary bacteria present in the lungs of late stage and terminal patients with cystic fibrosis. ref(Schaedel et al., 2002; Hauser, 2011) (becomes chronic). when the lung begins to exhibit decreased function and signs of failure For this reason, undesrstanding the genetic adaptations pseudomonas experience when in chrnic state is significantly important in talking this disease

In paper () it was shown that there was 'clear evidence' of the metabolic lung environment surrounding the bacteria. Specifically acetate production was negatively associated with length of infection and others?

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

Table 1: Table of non-metabolite phenotypes

Name	Description
Tobromycin	Resistance to inhalant antibiotic Tobromycin
Impenem	Resistance to intravenous antibiotic Imipenem
Aztreonam	Resistance to intravenous/intramuscular antibiotic Aztreonam
Ciprofloxacin	Resistance to oral antibiotic Ciprofloxacin
Colistin	Resistance to 'last-resort' antibioti Colistin
Swim	Measure of cell surface bactera movement by flagella
Swarm	Mesaure of rapid surface movement by multiple bacteria with rotating flagella
Twitch	Measure of slow baceria movement powered by pili

Chemical
Hydrogen Cyanide
Cyanide
2-Furoate
3-Hydroxyisovalerate
3-Methylthiopropionic acid
Anthranilate

Chemical
Betaine
Cystine
Formate
Fumarate
Histidine
Isoleucine
Leucine
Methanol
Methionine
Tryptophan
Uracil
Valine

3.1 Processing of phenotypes

When assocation testing a qualitatitve variable with linear regression, it is generally assumed that the variable follows a normal distrubution. Non-normal variables may fail to control for type-1 errors.

Quantitative (continuous) traits are preferred because they contain more information. the linear regression analysis requires all variables to be multivariate normal.

A popular statistical technique to bring normality to a continuous trait in association studies involves peforming log transformation, or a rank based normal transformation.

Prior to running *bgwas*, the n phenotypes were seperateley log transformed and INT, essesintially trippling the number of phenotypes tested.

3.2 Annotation

All 91 geneomes where annotated and 14643 unique genes were identified using Prokka.

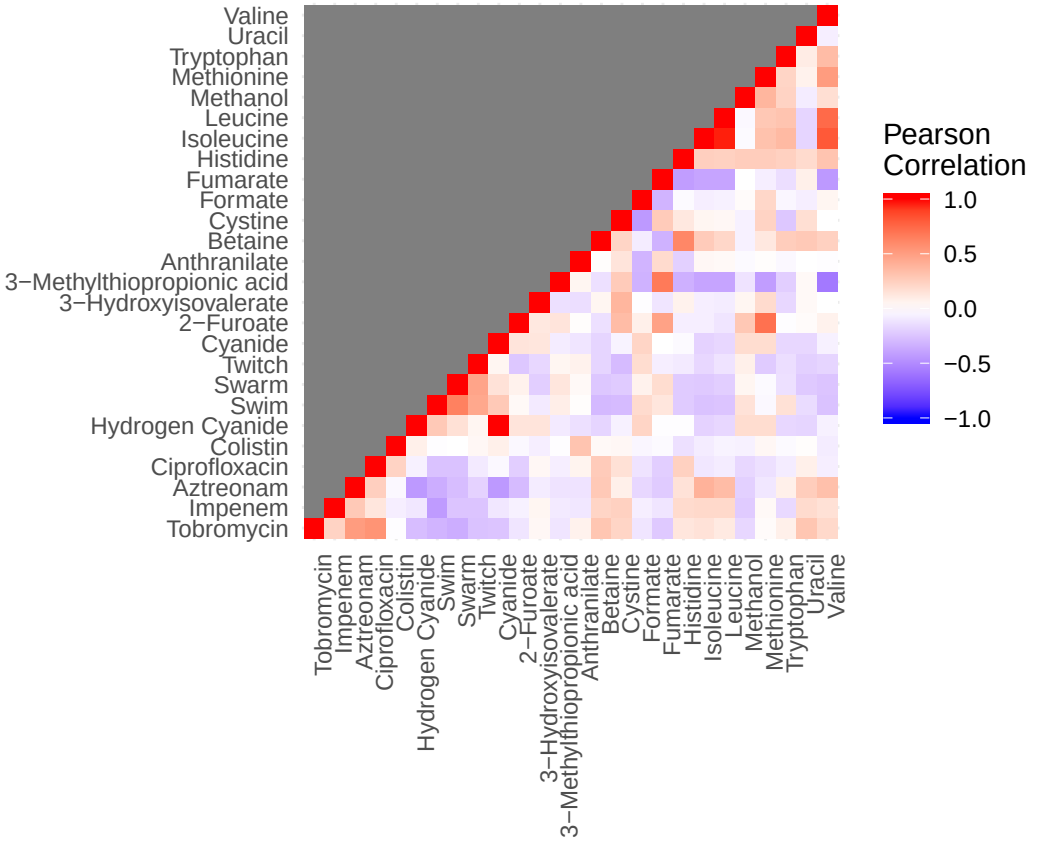

Figure 3: Correlation matrix of phenotypes

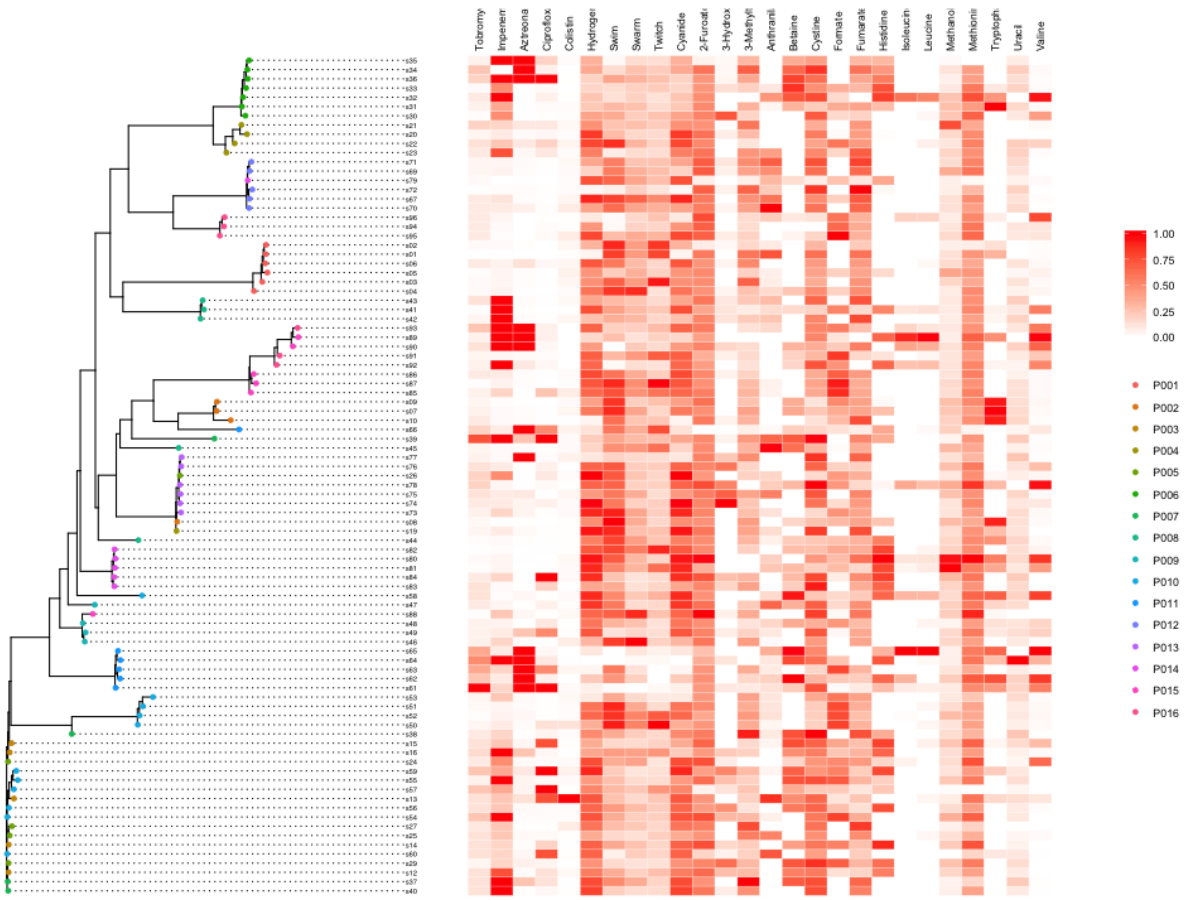

Figure 4: Density plots

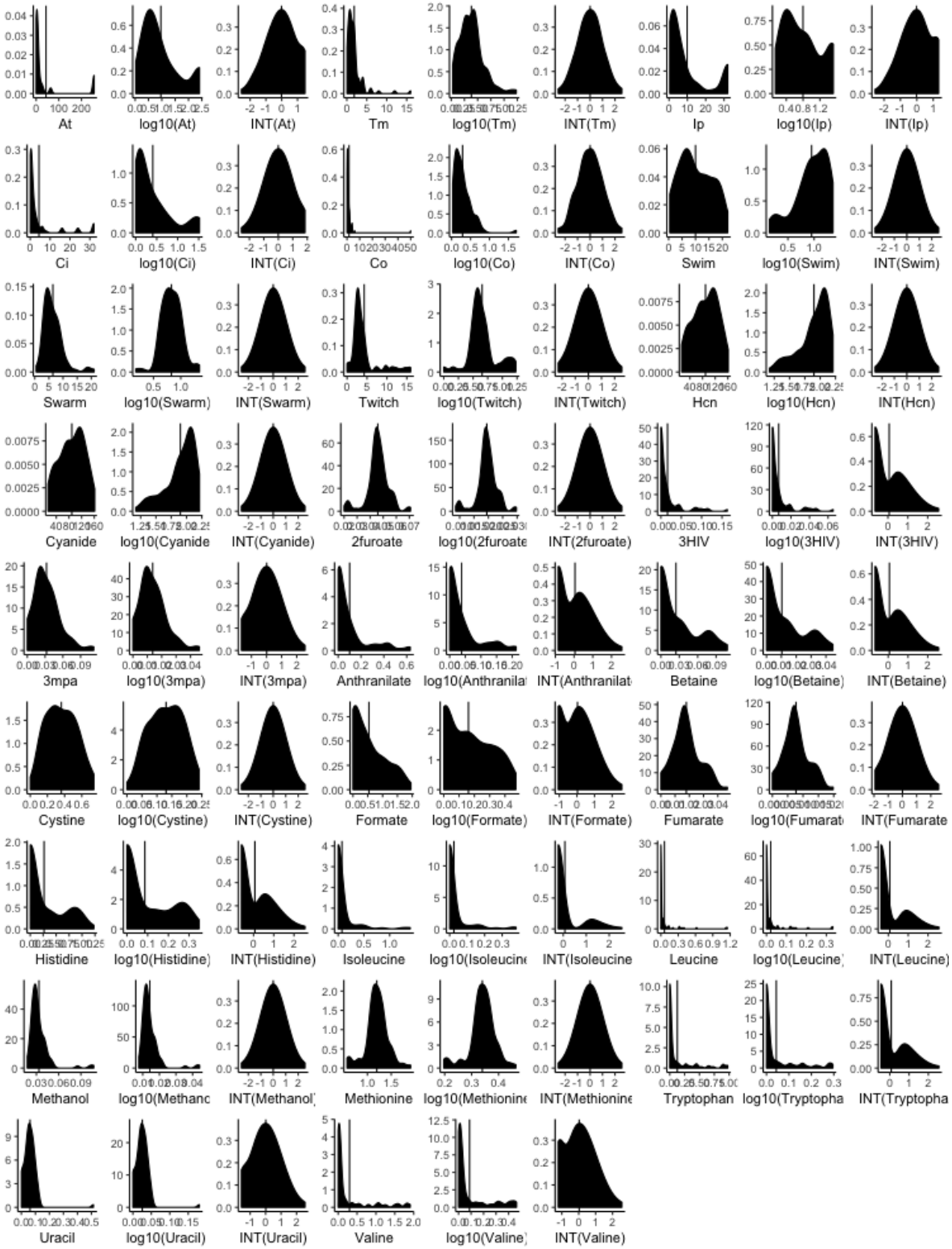

Figure 5: Density plots

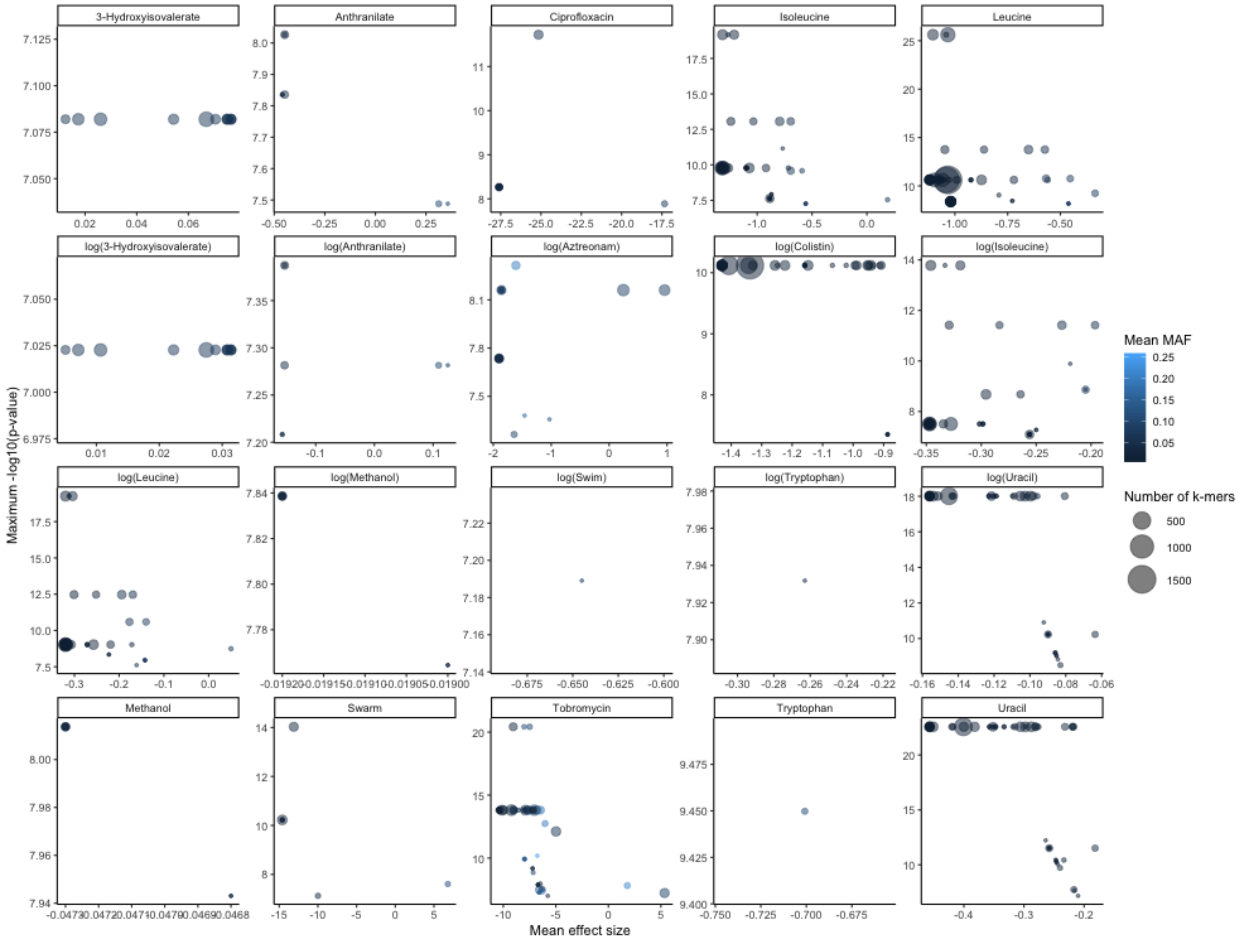

Figure 6: Gens

3.3 Kmer mining

Number etc

3.4 Association results

Phenotype	Significant Kmers	Genes
Leucine	16450	289
log(Colistin)	9811	233
Isoleucine	5097	132
log(Isoleucine)	4985	128
log(Leucine)	4913	125
Uracil	2707	75

Phenotype	Significant Kmers	Genes
log(Uracil)	2470	72
Tobromycin	1848	46
log(Aztreonam)	900	19
Ciprofloxacin	285	14
3-Hydroxyisovalerate	1814	13
log(3-Hydroxyisovalerate)	1814	13
Anthranilate	210	7
log(Anthranilate)	210	7
Methanol	216	7
log(Methanol)	216	7
Swarm	229	7
log(Swim)	2	1
Tryptophan	20	1
log(Tryptophan)	1	1
Aztreonam	26891	0
log(Ciprofloxacin)	1	0
log(Hydrogen Cyanide)	1	0
isoleucine_int	1	0

4 Discussion

In this work... ## Genome annotation

4.1 Kmer mining

The tool fsm

Prokka is a good tool ?Settings ## Phylogeny prediction Currently, *bgwas3* implements only a pangenomic approach approach of distance estimation. There are other tools which involve alignment of the core genome and snps... May or may not be bettter Reintroduce the problem of a large multiple alignment.

References

Cribbs, Adam P., Sebastian Luna-Valero, Charlotte George, Ian M. Sudbery, Antonio J. Berlanga-Taylor, Stephen N. Sansom, Tom Smith, et al. 2019. "CGAT-Core: A Python Framework for Building Scalable, Reproducible Computational Biology Workflows." *F1000Research* 8 (April): 377. https://doi.org/10.12688/f1000research.18674.1.

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.

Lees, John A., Minna Vehkala, Niko Välimäki, Simon R. Harris, Claire Chewapreecha, Nicholas J. Croucher, Pekka Marttinen, et al. 2016. "Sequence Element Enrichment Analysis to Determine the Genetic Basis of Bacterial Phenotypes." *Nature Communications* 7 (1): 12797. https://doi.org/10.1038/ncomms12797.

Li, Heng. 2013. "Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM." *arXiv:1303.3997 [Q-Bio]*, March. http://arxiv.org/abs/1303.3997.

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.

Page, Andrew J., Carla A. Cummins, Martin Hunt, Vanessa K. Wong, Sandra Reuter, Matthew T. G. Holden, Maria Fookes, Daniel Falush, Jacqueline A. Keane, and Julian Parkhill. 2015. "Roary: Rapid Large-Scale Prokaryote Pan Genome Analysis." *Bioinformatics* 31 (22): 3691–3. https://doi.org/10.1093/bioinformatics/btv421.

Rizk, Guillaume, Dominique Lavenier, and Rayan Chikhi. 2013. "DSK: K-Mer Counting with Very Low Memory Usage." *Bioinformatics (Oxford, England)* 29 (5): 652–53. https://doi.org/10.1093/bioinformatics/btt020.

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

Välimäki, Niko. 2018. "Fsm-Lite." https://github.com/nvalimak/fsm-lite.