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:3766

Abstract

Genome-wide association studies (GWAS) are now being 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 different, more complex considerations. These studies utilise short unique DNA patterns, referred to as k-mers, as units of genetic variation as opposed to single nucleotide polymorphisms. Various software solutions are now available which can perform the multiple association tests of k-mers, notable the python based *pyseer*. However, there is not at present a succinct tool which pipelines the necessary pre and post data processing and analysis steps that make an association study whole.

In metabolomics, thousands of possible measures that can be taken corresponding to the level of molecules inside or around a cell. These measures can be considered as traits of that bacteria at a that point in time. However conducting individual association tests for each metabolomic trait is time and computationally intensive.

I developed a tool, Bgwas3, which wraps the k-mer 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 performs gene annotation, phylogeny estimation, k-mer association testing, and k-mer-to-gene mapping
while also generating automatic visualisations and a web-based report.

The tool, found at https://github.com/g-r-eg/bgwas3, was built with best practices for scientific computing in mind, and is is installable as a conda package.

In this report I discuss it's implementation.

The tool was also used to test the association of 26 traits of *Pseudomonas auriginosa*; 18 relating corresponding 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 k-mers were found, and were mapped to named genes.

Contents

Abbreviations

CF Cystic Fibrosis

DRMAA Distributed resource management application

GWAS Genome-wide association study

SNP Single Nucleotide Polymorphism

TSV Tab separated value

LMM Linear mixed model

List of Figures

List of Tables

1 Introduction

1.1 Pangenome wide association studies

Genome wide association studies (GWAS) are are popular tool in human genetics due to their ability identify genetic and phenotype associations in a hypotheses free manner. Many of the characteristics of microbes suggest they would be ideal candidates for GWAS. First, their genomes are significantly smaller, which both limits the study size and makes whole genome sequencing easier and cheaper. Secondly, phenotypes of interest such as virulence and drug resistance tend to be the result of strong selective pressures; meaning they are likely to be controlled by fewer, recent mutations. There is also good reason to perform association studies with bacteria, as identifying the genetic basis of traits that correspond with pathogenic ability could lead to better understanding of disease and lead to new targets for pharmaceutical intervention.

In humans, GWAS have generally use single nucleotide polymorphisms (SNPs) as units of genetic variation. This is feasible only due to two defining characteristics of humans genomes. First, the genome experience regular and reliable recombination; meaning only very close loci are in linkage disequilibrium. This means an SNP that is identified as being significantly associated with a phenotype is likely to be very close to the truly causal polymorphism. Secondly, the gene sequence and content between humans remains relatively consistent within the species. This has allowed the development of SNP chip which can interrogate thousands of sites which accurately represent the entire genetic diversity. It also allows for SNPs to be identified though multiple alignment. Bacterial genomes possess neither of these qualities. In bacteria, its very likely that huge regions could be in linkage disequilibrium; obfuscating the true causal loci. It is also likely that the gene sequence is different, making identifying SNPs through multiple alignment an arduous computational problem that will always be limited to the core genome, ignoring completely genes of the accessory genome.

As an alternative, bacterial GWAS studies are taking a pangenomic approach and some recent studies have instead utilised k-mers: all nucleotide sub-strings of length ‘k’ found in the genomes. Initially a concept of genome assembly, k-mers can capture multiple genetic variations such as SNPs, longer deletions/ insertions and recombination sites. The size of k-mers effects the genetic variation captured, with longer being more specific, but shorter are more sensitive. Getting k-mers is alignment free, releasing the burden of multiple aliments, but more importantly, allows both the core and accessory genome to be tested for association.

1.2 Metabolomics and the need for multiple phenotype testing

Metabolomics of bacteria involve the large scale study of small molecules within and immediately adjacent to the cell. Metabolomics can provide a more detailed insight into the underlying biochemical activity and state of a cell than genomics or transcriptomics. Coupling genomics and metabolomics could allow a better understanding of how complex pathways in the cell have adapted. The nature of metabolomics means that multiple molecules abundance may be measured and of interest to investigate.

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 unique, accessible and robust; allowing a user to convert only the most basic required input files into valuable and interpretable results including static and interactive visualisations.

2 Methods

2.1 Pipelining with CGAT-Core

I made Bgwas3 with the Pipelining framework CGATCore, [4]. CGATCore is extremely portable framework that comes as a dependency free Python package, but has been used to construct complex pipelines that are scalable to large amounts of data.

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

Another significant feature of CGATCore which makes it stand out amongst other pipelining tools is an interface to control a distributed resource management application (DRMAA) such as PBS-Pro used by Imperial College’s resource computing service. Tasks which process files in parallel can run in parallel as

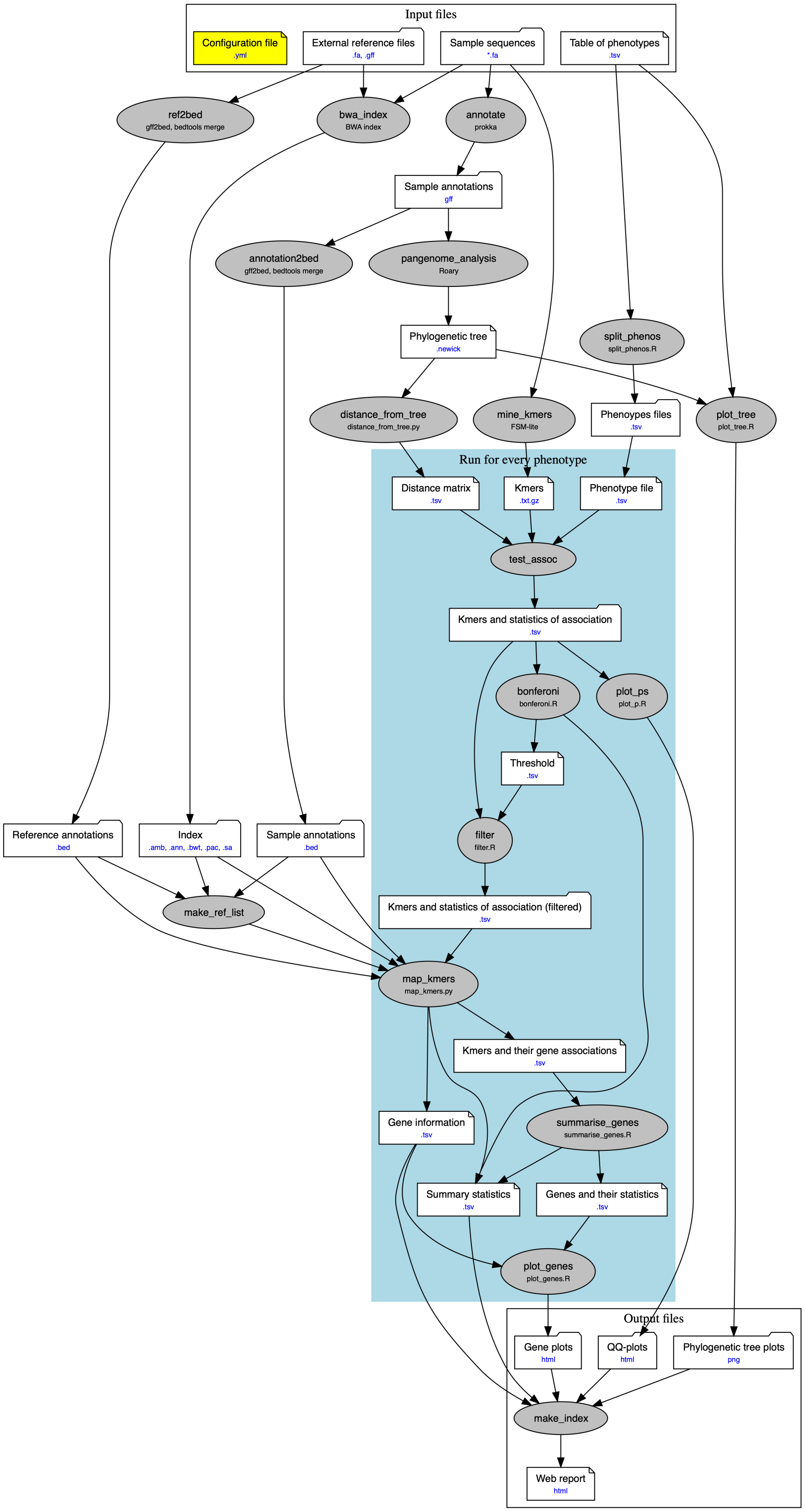

Figure 1: Graphical representation of the 18 tasks in the pipeline of bgwas. Grey ellipses encode individual tasks, and are labelled with the task name and then either the external tool used, or the external script which is run where applicable. White nodes indcate files or groups of files, and are labelled with the file format(s). The yellow node indicates the configuration file which may be optionally included as an input to define some of the pipelines paramters. The light blue box highlight steps which are repeated for each phenotype tested, with all other steps only needing to be run once.

batch processes, and tasks which require either high storage or memory resources it can be distributed to nodes which match these requirements. This makes executing pipeline such as Bgwas3, which involve multiple computationally intensive intermediary steps faster.

Bgwas3 was written as 18 'tasks' which are visualised in figure 1. I wrote the pipeline in python, but individual tasks made calls to external tools or scripts I wrote in either python or R.

As input, the tool requires the following three starting files: 1. A directory containing fasta files uniquely named for each bacteria sample. 2. A tab separated value file (TSV) in which the first column matches the names of sample fasta files, and remaining columns corresponding to trait measures. 3. An optional directory of reference sequences in fasta format accompanied by annotation files in general feature format (GFF). 4. An optional configuration file in .yml format that specifies extra parameters for the running of the pipeline.

2.2 Genome Annotation

In the Bgwas3 task 'annotate' all sample sequences are annotated by the external tool Prokka [23] Prokka first makes predictions on features using a selection of external tools including Prodigal [7] for the annotation of coding region, RNAmmer [9] which identifies for ribosomal RNA genes, Aragorn [10] for transfer RNA genes, and SignalP [1] for signal leader peptides and Infernal for non-coding RNA. After feature prediction, the speculative features are queried against a umber of databases including UniProt [24], RefSeq [17] and Pfam [6].

Draft sequences are annotated in the Bgwas3 pipeline for two reasons: First, gene annotations are later used to estimate the phylogenetic tree of the samples; and secondly, significantly associated k-mers are later mapped to the annotated genomes when attempting identifying the k-mer's genetic identity.

2.3 k-mer minig and counting

Bgwas3 integrates the external tool FSM-lite [26] to 'mine' and count k-mers. A benefit of FSM-lite is, unlike other k-mer mining tools such as DSK [22], FSM-lite allows a range of k-mer-sizes, as defined by the user, to be mined as apposed to a single length. Bgwas3 allows the k-mers length to be changed in the pipeline by editing the Bgwas3 configuration yml file values 'fsm_k-mer_min' and 'fsm_k-mer_max'.

2.4 Phylogeny estimation and covariate generation

Bgwas3 currently takes a pangenomic approach to phylogeny estimation, as in the relative presence and absence of genes are used to calculate the distance between samples. A phylogenetic tree is estimated with the tool Roary [18]. In summary, Roary determines genes which fall within the core genome, and then performs clustering of isolates based in the constitution of the variables accessory genome. One of Roary's outputs, a tree in the common newick format, which is then converted into a distance matrix TSV file.

The single reasoning for predicting phylogeny in the Bgwas3 pipeline is to use the distances defined in the distance matrix as covariates in the later association testing.

2.5 k-mer association testing

The typical analytical strategy implemented in GWAS is some form of linear regression. Standard regression assumes that data is are identically and independentlly distributed. However, due to population stucture, this is almost always an incorrect assumption in GWAS studies, and if that assumption is made incorrectly, other genetic polymorphisms will be incorecctly identified as associated to the trait of interest.

Most GWAS techniques involve implement a statistical method to to control for the confounding population structure potential. In human studies, for example, its not abnormal to remove highly related individuals, though this will ultimately reduce the power of the study by decreasing sample size. Another popular techique involves peforming multiple dimensional scaling, and then including significant principal compoents as fixed effects in the linear model. Bacteria experience a much stronger population structure as a result of clonal reproduction, and so a need for control becomes more important.

Bgwas3 implements a linear mixed model (LMM) as provided by pyseer to tackle population, specifically factored spectrally transformed LMM [13]. LMMs tackle confounders by using measures of genetic similarity as random effects within the linear model. Given a good measure of hierarchical relatedness between samples, the mixed model is generally preferred, and has been shown in past studies to control the inflation of p-values better [11].

Bgwas3 uses the distance matrix from the estimated phylogeny as covariates. In the Bgwas3 task 'test_assoc' an association test for each phenotype is run using all the k-mers.

The output of 'test_assoc', a list of all k-mers and statistics relating to their association to the given phenotype, are then filtered by their P-value though bonferoni correction. The Bonferroni correction is a multiple-comparison correction used when several statistical tests are performed, to limit the number of false-positive results. It's vital in GWAS studies to apply some filtering, as the number of independent tests are so high that it is very likely that spurious rare events (like chance false positive) will occur. Only significant k-mers are then used in later steps of the pipeline.

2.6 k-mer mapping

The Burrows-Wheeler Alignment Tool (BWA) [12] is primarily used to map significant k-mers to genes. The tool first requires that sequences are converted into an FM-index: a data-structure similar to a suffix array. This is performed with the Bgwas3 task 'bwa_index' (figure 1) on all sample and reference fast files.

For use with bedtools [20], all annotation files in GFF format are converted into BED format, and subsequently filtered with the tasks 'annatation2bed' and 'ref2bed' repsectively. Sample annotations 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-genes. Entries in the reference annotation file which correspond to the same loci are merged, collating the information from both.

The task 'map_k-mers', executes a python script I wrote whose algorithm is visualised in figure 2.

In summary, the algorithm works as follows:

1. Generate an artificial multi-fasta file with an entry for each unmapped k-mer.
2. Choose the next reference file.
3. Attempt to align the artificial fasta file to the reference fasta file with BWA-mem.
4. Generate an artificial BED file where each entry corresponds to a successful alignment with BWA-mem.
5. Use 'bedtools intersect' command line tool to compare and retrieve matches in the query and reference BED files.
6. Harvest all information about each intersect stored in a gene info file,
7. Mark the k-mer as mapped.
8. Repeat until all k-mers mapped or until all references have been used

2.7 Output and Visualisation

Bgwas3 automatically generates multiple figures which are then integrated into a final web based report. Static visualisations are made with external scripts I wrote in R that make use of ggplot2 package [27]. The task 'plot_ps' generated a quantile-quantile plot from all unfiltered p-values (see figure 5), and a phylogenetic

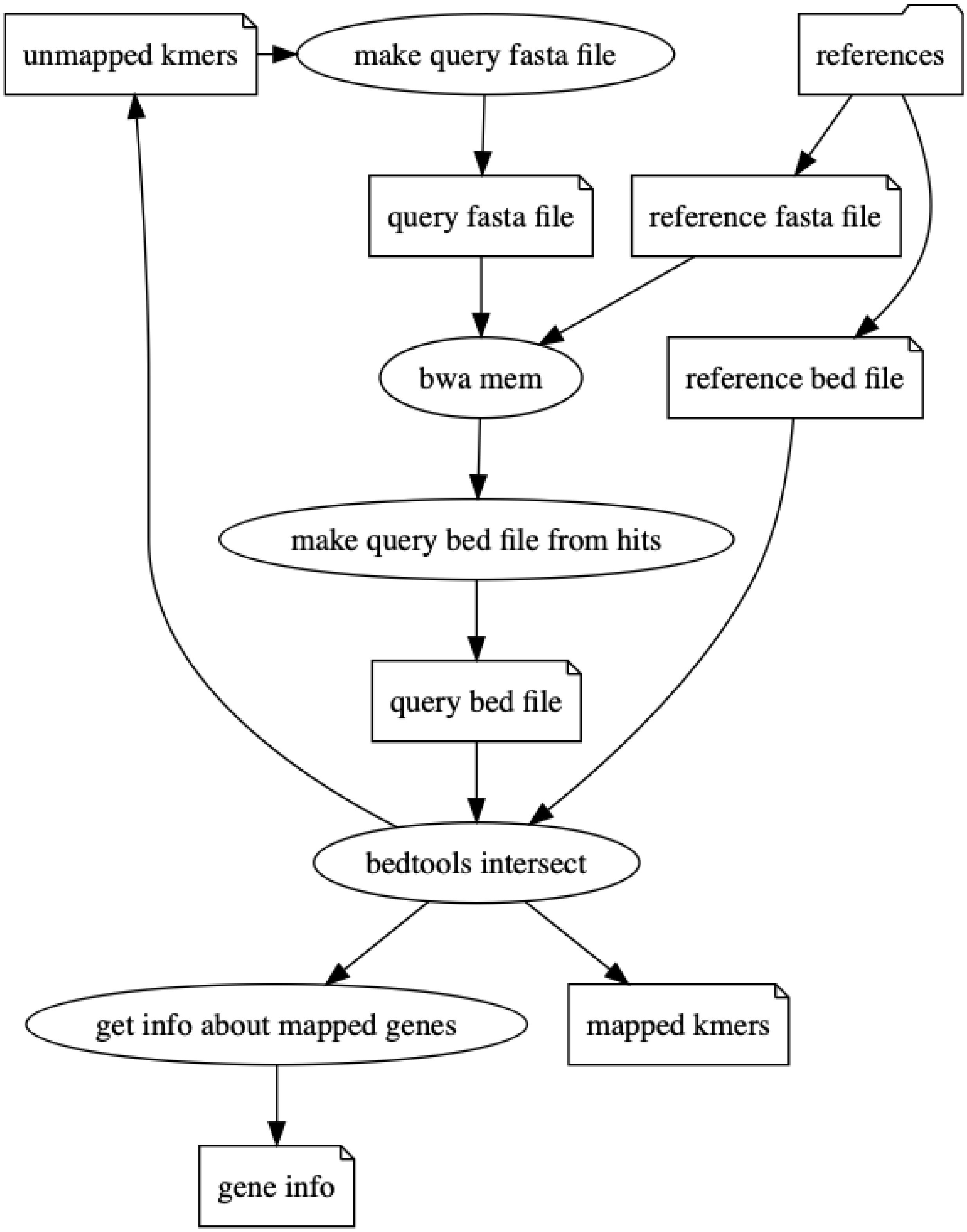

Figure 2: Graphical representation of algorithm to map k-mers to reference genes

tree visualisation is built from the newick file from Roary, and the input phenotype file utiliseing external R package ggtree [28]. (see figure 4).

Finally, for each phenotype, a plot of genes is made in which the following features are visually encoded (see figure 5):

- Maximum $-log_{10}(Pvalue)$ of k-mers mapped to that gene
- Average beta value (effect size)
- Average allele frequency
- Number of k-mers 'hits;

Finally, a web based report is generated with the task 'make_index'. The report includes an interactive table of all phenotypes that can be sorted by the number of significant k-mers found and the number of genes., such as the protein product. Links then take the user to an interactive version of the gene plot (see figure 5) in which all information that was originally stored in the annotation files can be retrieved, such as the protein product.

3 Results with test data

Dataset For the development, testing, and evaluation of Bgwas3 a dataset corresponding to genomes and phenotype measurements of *Pseudomonas aeruginosa* was used.

Cystic fibrosis (CF) is an autosomal recessive disorder that, due to a single gene mutation, causes a patient to have defective trans-membrane regulator protein. This protein is situated in epithelial cells that make up the mucus membranes of the body, and is primarily responsible for transporting chloride ions and bicarbonate [21]. The dysfunctional form of this protein ultimately limits the osmotic movement of water, and the mucus at these membranes remain viscous and immotile. In the lungs of healthy individuals, a less viscous mucus is able to be transported by cilia out of the lungs, but in patients with CF, the stagnant mucus instead becomes an ideal environment for bacteria to propagate, and so patients experience at first episodic, but then chronic infection of the lungs. With prolonged antibiotic use, the fast generation time of bacteria mean resistant strains soon develop, and the prologned inflammation leads to respiratory failure and death.

Pseaudomonas auruginosa is a gram negative bacteria which can easily integrate exogenous DNA into its own genome, making it able to adapt to antibiotic pressures rapidly. For this reason, it is a common hosptial-aquired infection [19]. It is also one of the primary bacteria present in the lungs of late stage and terminal patients with cystic fibrosis when the lungs function starts to decrease. For this reason, undesrstanding the genetic adaptations pseudomonas experience when in chronic state is significantly important in talking this disease

In a previous study [2] 91 strains were collected over a period of 24 years from 18 patients suffering from Cystic Fribrosis. Assays of antibiotic resistance, bacterial motility, and later metabolomic measurements were taken.

Bgwas3 was used to test the genetic association of 26 traits (see tables 1 and 2). 18 traits correspond to metabolomic measurement and 8 corresponding to either a measure of to antibiotic resistance or bacterial motility.

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

Table 2: Table of metabolite phenotypes

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

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

Phenotype data preprocessing

When association testing a qualitative variable with linear regression, it is generally assumed that the variable follows a normal distribution. When this assumption proves not to be true, and the continuous trait displays severe skewness, the regression may fail to control for type-1 errors (false positives). A popular statistical technique in GWAS involes transforming the data into a normal form. Often, a simple log transformation can be sufficient. Recently, rank based inverse normal transformations (INT) have become popular among genetics researches [16].

Prior to running Bgwas3, the 26 phenotypes were separately log transformed and transformed using INT, essentially tripling the number of phenotypes tested to 78. Density plots of the unadjusted and transformed traits are visualised in figure 3.

Phylogeny estimation

All 91 genomes where annotated and 14643 unique genes were identified with Prokka. As identified by Roary, these genes were found in >99% of the genomes, and constitute the core genome, leaving 10005 in the accessory.

From these genes, a phylogenetic tree was estimated, and visualised (figure 4).

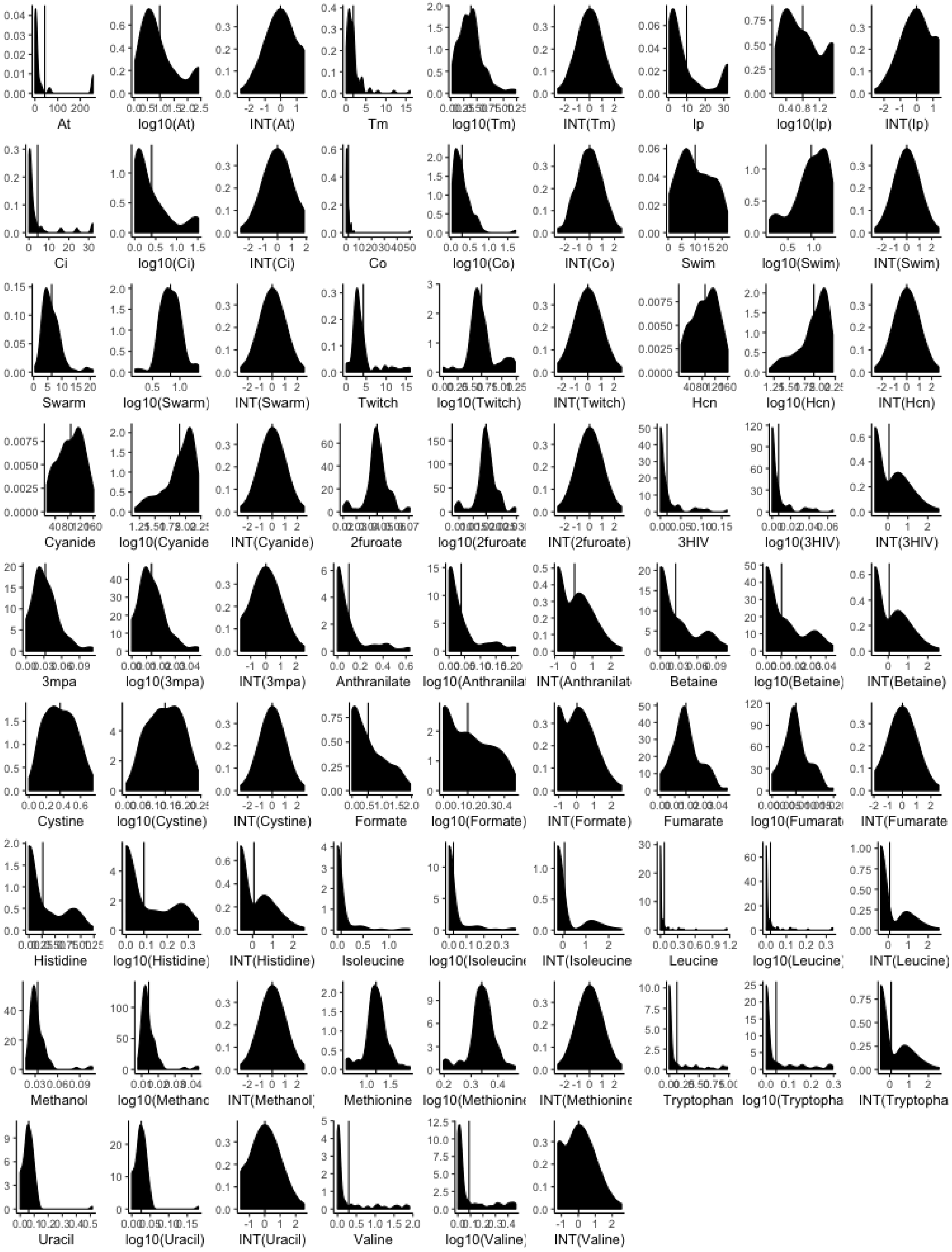

Figure 3: Density plots

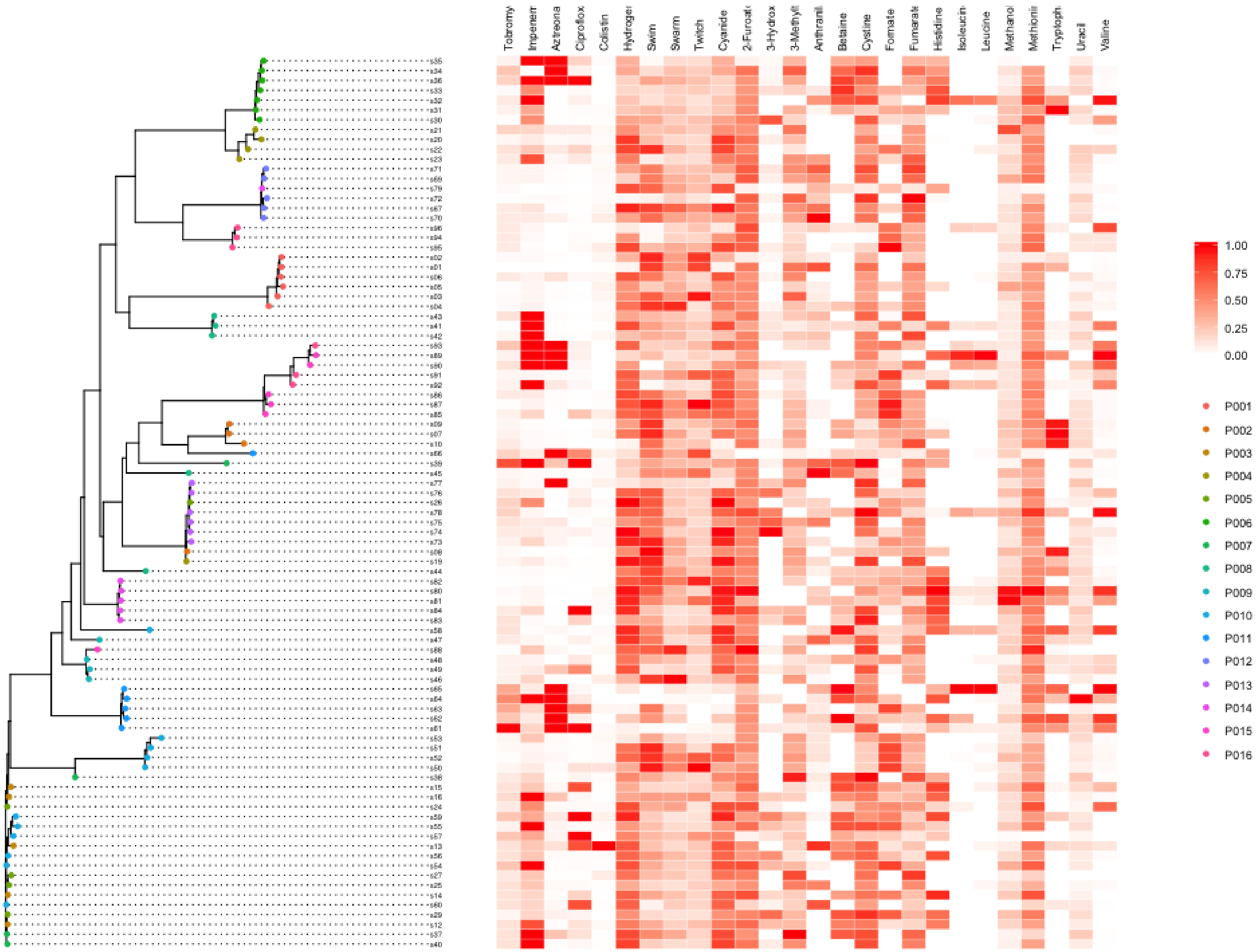

Figure 4: Density plots

An inspection of the tree shows that, in general, samples from the same patient are generally clustered together, which leads me to believe the phylogeny estimate is somewhat accurate. It is not unexpected that the dividing between patients is not perfect, due to possible cross contamination and horizontal gene transfer, as many of the patients originated from the same hospital. Traits of antibiotic resistance seem to be best associated with phylogeny, seen as red clusters in figure 4.

k-mer mining

497827 unique k-mers were mined of sizes between 9 and 100 base pairs in length.

Association results

Phenotype	Significant Kmers	Genes
Leucine	16450	289
Colistin	9811	233
Isoleucine	5097	132
Uracil	2707	75
Tobromycin	1848	46
Aztreonam	900	19
Ciprofloxacin	285	14
3-Hydroxyisovalerate	1814	13
Anthranilate	210	7
Methanol	216	7
Swarm	229	7
Swim	2	1
Tryptophan	20	1

13 of the 26 traits, either in their untransformed state or log transformed state had significant k-mers that passed the bonferoni threshold of p < 1.004*10-7, and were mapped to one or more genes. The trait with the most number of signifiant k-mers was the one corresponging to Leucine, in which 16450 significant k-mers were identified and mapped to 338 genes. By inspecting the QQ-plots, it becomes apparent that the statudyhas mixed results in regard to accounting for population structure. A well controlled study would present a smooth QQ plot wtith little or no large ridges. However, judging from the QQ-plots, most of the studues were well controlled for low p-values.

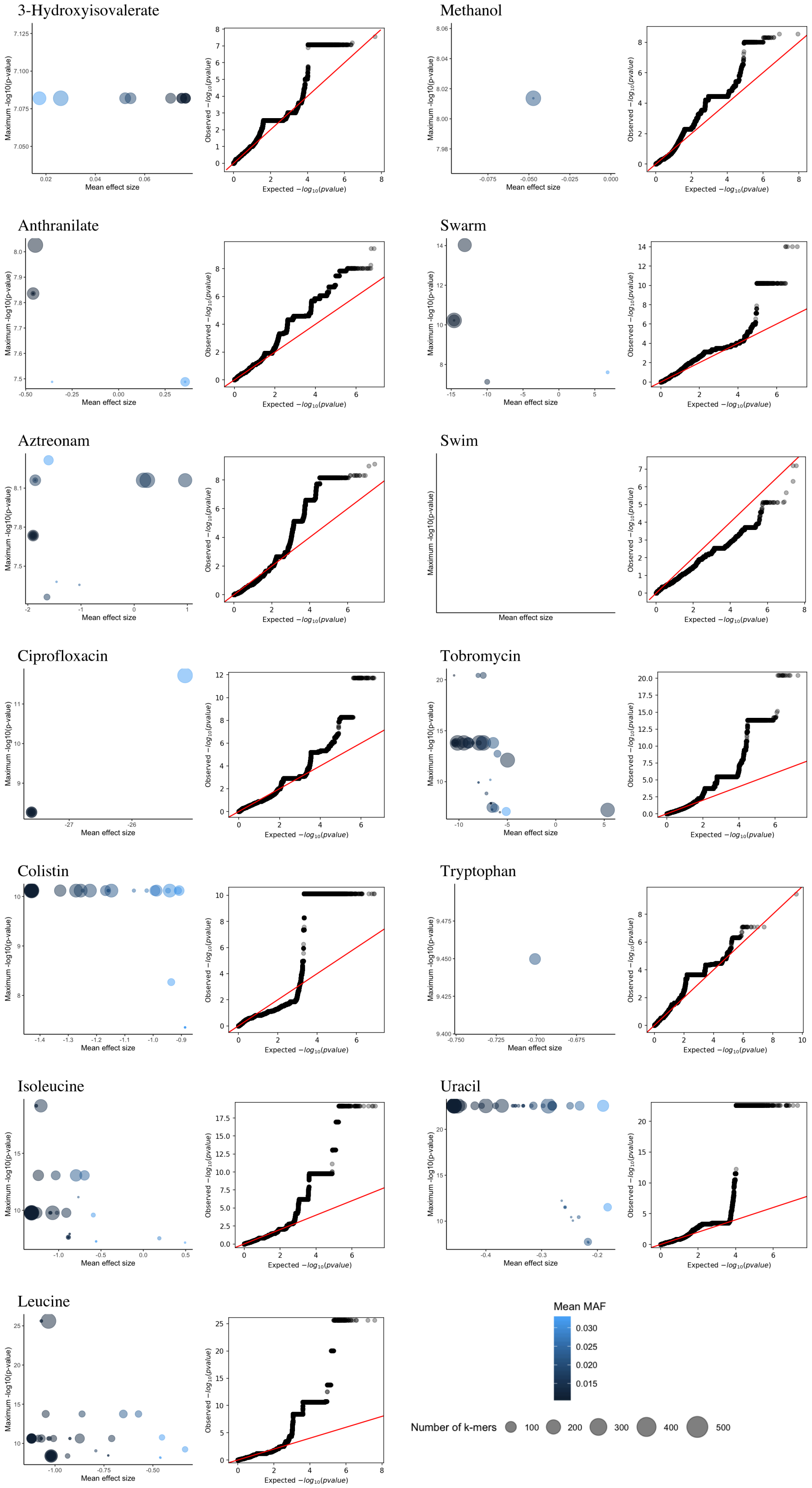

Figure 5: Genes

4 Discussion

A significant goal of bgwas3 was to implement a useful and reusable tool that may be used outside the scope of this project. As such, Bgwas3 was built with the best scientific computing practices in mind to aid reproducability and ease of use. Bgwas3 is packaged and is currently installable as a conda recipe (https://anaconda.org/g-r-eg/bgwas3) whereby all external dependecies are included, or through pypi (https://pypi.org/project/bgwas3/). It also has some testing using continuous integration (https://travis-ci) and has been success2ully deployed on OSX and Ubuntu. The project resides on a github repository, and is under a GNU General Public License v3.0, and external additions and development is encouraged using the standard git protocols.

4.1 Future work

Through the development of Bgwas3 it's use on test data, many limitations and possible areas for addition became apparent. However, because Bgwas3 was written using a pipelining framework, it is possible to add additional methods of individual steps and add more post analysis tasks without compromising what has already been built. It is hoped that development of bgwas3 will continue, and some of these new additions will be included.

Phylogeny estmation Currently, Bgwas3 implements only a pangenomic approach approach of distance estimation. This approach, gene presence and absence has not been critically evalulated or compared with other methods, though many other possbible methods exist. For example, a common approach to estimating phylogeny involves peforming a multiple alignment of the core geneome, and using the difference in SNP content as distance estimates. Tools such the general multiple aligner Mummer4 [14] and the bacteria-focused tools ParSnp [25] and ClonalML [5] suite attempt this mutliple alignement, while peforming inference of recombination.

Interpretation of k-mer results K-mer-based GWAS is relatively new, but already some tools are finding new and possibly better ways to interpret significant K-mers in a manner besides mapping them to genes. This is the primary feature of bacteria GWAS tool DBGWAS [8]. DBGWAS uses all the k-mers from the study to rebuild a de-bruijn graph in a manner similar to genome assembly. After k-mers are then tested for assoctiion, the associated kmers are then visualised as sub-graph of the full de-bruijn graph, where shape and colour encode the statititics of significance and p-value. This allows clusters of genes which are both close to one another and significicantly associated to be quickly identified. Another possible addition to Bgwas3 which would significantly advance its usefulness as a tool is pathway enrichment analysis. For example, in paper by Marvig et al. [15] a sample of pseudomonas aeruginosa were analysed using a clone based methods and identified the convergent eveolution of 52 genes. As a final analysis of the genes, they were grouped together according to their function as defined by PseudoCap, and through enrichment analysis, ovr-represented classes were identfifed.

Control for population structure Other software packages exist for performing GWAS experiments with bacteria that don't rely on the LMM approach of population control. For example, the programme as treeWas [3] instead generates a simulation of a null genetic dataset using parameters of the emperial dataset phylogeney. From the simulated dataset it can then peform association tesing, and build a null distrubution of score statistics under the null hypotheses of no association. This provides a strict control over the false popsitveity rate.

References

[1] José Juan Almagro Armenteros et al. "SignalP 5.0 improves signal peptide predictions using deep neural networks". en. In: *Nature Biotechnology* 37.4 (Apr. 2019), pp. 420–423. ISSN: 1546-1696. DOI: 10.1038/s41587-019-0036-z. URL: https://www.nature.com/articles/s41587-019-0036-z (visited on 08/31/2019).

[2] V. Behrends et al. "Metabolic adaptations of *Pseudomonas aeruginosa* during cystic fibrosis chronic lung infections: Metabolomic adaptations of cystic fibrosis isolates". en. In: *Environmental Microbiology* 15.2 (Feb. 2013), pp. 398–408. ISSN: 14622912. DOI: 10.1111/j.1462-2920.2012.02840.x. URL: http://doi.wiley.com/10.1111/j.1462-2920.2012.02840.x (visited on 06/03/2019).

[3] Caitlin Collins and Xavier Didelot. "A phylogenetic method to perform genome-wide association studies in microbes that accounts for population structure and recombination". en. In: *PLOS Computational Biology* 14.2 (Feb. 2018), e1005958. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1005958. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005958 (visited on 05/06/2019).

[4] Adam P. Cribbs et al. "CGAT-core: a python framework for building scalable, reproducible computational biology workflows". en. In: *F1000Research* 8 (Apr. 2019), p. 377. ISSN: 2046-1402. DOI: 10.12688/f1000research.18674.1. URL: https://f1000research.com/articles/8-377/v1 (visited on 06/28/2019).

[5] Xavier Didelot and Daniel J. Wilson. "ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial Genomes". en. In: *PLOS Computational Biology* 11.2 (Feb. 2015). Ed. by Andreas Prlic, e1004041. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1004041. URL: http://dx.plos.org/10.1371/journal.pcbi.1004041 (visited on 05/06/2019).

[6] Sara El-Gebali et al. "The Pfam protein families database in 2019". en. In: *Nucleic Acids Research* 47.D1 (Jan. 2019), pp. D427–D432. ISSN: 0305-1048. DOI: 10.1093/nar/gky995. URL: https://academic.oup.com/nar/article/47/D1/D427/5144153 (visited on 09/01/2019).

[7] Doug Hyatt et al. "Prodigal: prokaryotic gene recognition and translation initiation site identification". eng. In: *BMC bioinformatics* 11 (Mar. 2010), p. 119. ISSN: 1471-2105. DOI: 10.1186/1471-2105-11-119.

[8] Magali Jaillard et al. "A fast and agnostic method for bacterial genome-wide association studies: Bridging the gap between k-mers and genetic events". en. In: (Nov. 2018), p. 28.

[9] Karin Lagesen et al. "RNAmmer: consistent and rapid annotation of ribosomal RNA genes". eng. In: *Nucleic Acids Research* 35.9 (2007), pp. 3100–3108. ISSN: 1362-4962. DOI: 10.1093/nar/gkm160.

[10] Dean Laslett and Bjorn Canback. "ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences". In: *Nucleic Acids Research* 32.1 (2004), pp. 11–16. ISSN: 0305-1048. DOI: 10.1093/nar/gkh152. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC373265/ (visited on 08/31/2019).

[11] John A Lees et al. "Genome-wide identification of lineage and locus specific variation associated with pneumococcal carriage duration". In: *eLife* 6 (July 2017). Ed. by Sarah Cobey, e26255. ISSN: 2050-084X. DOI: 10.7554/eLife.26255. URL: https://doi.org/10.7554/eLife.26255 (visited on 08/24/2019).

[12] Heng Li. "Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM". In: *arXiv:1303.3997 [q-bio]* (Mar. 2013). arXiv: 1303.3997. URL: http://arxiv.org/abs/1303.3997 (visited on 08/30/2019).

[13] Christoph Lippert et al. "FaST linear mixed models for genome-wide association studies". en. In: *Nature Methods* 8.10 (Oct. 2011), pp. 833–835. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/nmeth.1681. URL: http://www.nature.com/articles/nmeth.1681 (visited on 08/24/2019).

[14] Guillaume Marçais et al. "MUMmer4: A fast and versatile genome alignment system". en. In: *PLOS Computational Biology* 14.1 (Jan. 2018), e1005944. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1005944. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005944 (visited on 05/07/2019).

[15] Rasmus Lykke Marvig et al. "Convergent evolution and adaptation of Pseudomonas aeruginosa within patients with cystic fibrosis". en. In: *Nature Genetics* 47.1 (Jan. 2015), pp. 57–64. ISSN: 1061-4036, 1546-1718. DOI: 10.1038/ng.3148. URL: http://www.nature.com/articles/ng.3148 (visited on 05/09/2019).

[16] Zachary R. McCaw et al. "Operating Characteristics of the Rank-Based Inverse Normal Transformation for Quantitative Trait Analysis in Genome-Wide Association Studies". en. In: *bioRxiv* (June 2019), p. 635706. DOI: 10.1101/635706. URL: https://www.biorxiv.org/content/10.1101/635706v2 (visited on 09/01/2019).

[17] Nuala A. O'Leary et al. "Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation". eng. In: *Nucleic Acids Research* 44.D1 (Jan. 2016), pp. D733–745. ISSN: 1362-4962. DOI: 10.1093/nar/gkv1189.

[18] Andrew J. Page et al. "Roary: rapid large-scale prokaryote pan genome analysis". en. In: *Bioinformatics* 31.22 (Nov. 2015), pp. 3691–3693. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/btv421. URL: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btv421 (visited on 05/04/2019).

[19] Joshua Quick et al. "Seeking the source of Pseudomonas aeruginosa infections in a recently opened hospital: an observational study using whole-genome sequencing". en. In: *BMJ Open* 4.11 (Nov. 2014), e006278. ISSN: 2044-6055, 2044-6055. DOI: 10.1136/bmjopen-2014-006278. URL: https://bmjopen.bmj.com/content/4/11/e006278 (visited on 09/01/2019).

[20] Aaron R. Quinlan and Ira M. Hall. "BEDTools: a flexible suite of utilities for comparing genomic features". en. In: *Bioinformatics* 26.6 (Mar. 2010), pp. 841–842. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btq033. URL: https://academic.oup.com/bioinformatics/article/26/6/841/244688 (visited on 08/30/2019).

[21] J. R. Riordan et al. "Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA". en. In: *Science* 245.4922 (Sept. 1989), pp. 1066–1073. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.2475911. URL: https://science.sciencemag.org/content/245/4922/1066 (visited on 05/09/2019).

[22] Guillaume Rizk, Dominique Lavenier, and Rayan Chikhi. "DSK: k-mer counting with very low memory usage". eng. In: *Bioinformatics (Oxford, England)* 29.5 (Mar. 2013), pp. 652–653. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btt020.

[23] Torsten Seemann. "Prokka: rapid prokaryotic genome annotation". en. In: *Bioinformatics* 30.14 (July 2014), pp. 2068–2069. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btu153. URL: https://academic.oup.com/bioinformatics/article/30/14/2068/2390517 (visited on 06/04/2019).

[24] The UniProt Consortium. "UniProt: a worldwide hub of protein knowledge". en. In: *Nucleic Acids Research* 47.D1 (Jan. 2019), pp. D506–D515. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gky1049. URL: https://academic.oup.com/nar/article/47/D1/D506/5160987 (visited on 09/01/2019).

[25] Todd J. Treangen et al. *Rapid Core-Genome Alignment and Visualization for Thousands of Intraspecific Microbial Genomes*. en. preprint. Bioinformatics, July 2014. DOI: 10.1101/007351. URL: http://biorxiv.org/lookup/doi/10.1101/007351 (visited on 06/04/2019).

[26] Niko Välimäki. *fsm-lite*. original-date: 2015-10-25T11:52:01Z. Nov. 2018. URL: https://github.com/nvalimak/fsm-lite (visited on 08/30/2019).

[27] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN: 978-3-319-24277-4. URL: https://ggplot2.tidyverse.org.

[28] Guangchuang Yu et al. "ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data". en. In: *Methods in Ecology and Evolution* 8.1 (2017), pp. 28–36. ISSN: 2041-210X. DOI: 10.1111/2041-210X.12628. URL: https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.12628 (visited on 09/01/2019).