Improving SNP-based microbial genotyping software

[Stakeholders 1](#_Toc492115443)

[Documents 1](#_Toc492115444)

[Synopsis 2](#_Toc492115445)

[Purpose: 2](#_Toc492115446)

[Scope: 2](#_Toc492115447)

[Period: 2](#_Toc492115448)

[Deliverable schedule: 2](#_Toc492115449)

[Acceptance criteria: 2](#_Toc492115450)

[Background 3](#_Toc492115451)

[The team and skills 4](#_Toc492115452)

[Milestones 4](#_Toc492115453)

[Methodology/Method 5](#_Toc492115454)

[Risks 5](#_Toc492115455)

[Tools 6](#_Toc492115456)

[Deliverables 6](#_Toc492115457)

# Stakeholders

|  |  |  |  |
| --- | --- | --- | --- |
| Phil Giffard | phil.giffard@menzies.edu.au | 0489468548 | Client |
| Peter Shaw | peter.shaw@cdu.edu.au | 0489466157 | Lecturer |
| Kian Soon Hoon | ldwgkshoon@gmail.com | 0487034025 | Student |
| Wei Zhong Teo | 54081@siswa.unimas.my | 0468396956 | Student |
| Daniel Wilson | drac1223@hotmail.com | 0424537796 | Student |

# Documents

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Document | Started | 50% completed | Signed off | Note |
| Test Documentation |  |  |  |  |
| User Manual |  |  |  |  |
| Technical Documentation |  |  |  |  |
| Project Report |  |  |  |  |

# Synopsis

## Purpose:

To design and create a software that will help the microbiologists to identify significant SNP and well as the position of those SNPs.

## Scope:

The software will make use of open source library and do the followings:

1. Input DNA sequence alignment
2. Allow user to select % or D mode
3. Identify optimised set of SNPs using either: For D mode, the greedy algorithm (i.e. Simpson’s index of diversity): for % mode: simply the maximum percentage of excluded non-target sequences in the alignment.

## Period:

The development of software for the analysis of genome-wide orthologous SNP is expected to last until 23rd September 2017 (week 10, semester 2).

## Deliverable schedule:

1. 1st release: 28th August
2. 2nd release: 4th September
3. 3rd release: 11th September
4. 4th release: 23th September

## Acceptance criteria:

1. Software conforms to the requirements specified and accepted by client.
2. Completing the project before 23 September 2017.
3. Easy to understand documentation that provide enough details for the users.

# Background

In clinical, public health, and research microbiology, there is frequently a need to divide bacterial species into “types”. A bacterial type may be thought of as a classification unit at a finer scale than “species”, and roughly synonymous with the term “strain”. Bacterial typing is used to trace instances and patterns of transmission and dissemination, and may be used at all scales from global to within individual buildings (e.g. hospitals), to indicate virulence or resistance properties of isolates, and in basic research to assist understanding of evolutionary history. While early typing methods were based upon reactions with standard sets of antisera (serotyping), or susceptibility to standard sets of viruses (phage typing), for decades, the great majority of bacterial typing has been performed using genetic methods. As a result, the term “genotyping” will be used in this document. Over the years, considerable ingenuity has been applied to the development of bacterial genotyping methods. Many methods have been described, with most being variations on the theme of electrophoresis of complex mixes of DNA fragments so as to yield banding pattern “fingerprints”, or sequencing of standardised gene fragment(s).

In the early 2000s, Phil Giffard led a project within the QUT node of the Cooperative Research Centre for Diagnostics. This was essentially a re-think of bacterial genotyping methodology. The innovative concept was that the volume of known comparative gene sequence data from within bacterial species was exploding. This constitutes a suburb resource to mine for sets of variable genetic locations (single nucleotide polymorphisms (SNPs)) that are optimised for use in bacterial genotyping methods with predefined performance specifications. In this way, genotyping methods that yield the needed resolving power or other information as efficiently as possible could be easily designed. An added impetus for this area of research was rapid development in technology for interrogating SNPs. The first publication outlining this approach was published in 2004. This reported the construction of the SNP-mining “Minimum SNPs” software package, and the demonstration of two bacterial genotyping methods, using allele specific PCR on a real time PCR platform.

A major development in very recent years has been the advent of low cost whole genome sequencing of bacteria. The work described above made extensive use of multilocus sequence typing (MLST) databases as the input data for SNP mining. MLST was developed in the UK (e.g. see http://saureus.mlst.net/) , as a major and very successful initiative to standardize bacterial genotyping methods and terminology. For each bacterial species for which it has been implemented, seven standardised gene fragments are defined. These are sequenced to generate the “sequence type” (ST). The MLST database curators maintain comprehensive internet accessible information regarding sequencing variants at the loci (alleles) and the alleles found together in bacterial isolates (STs). Often several thousand STs are known. Our SNP based typing methods have been based on mining alignments of these STs. The MLST loci only cover ~0.1% of the bacterial genome. Lost cost whole genome sequence make it possible to genotype a bacteria by whole genome sequencing and causes MLST obsolete. However, despite the decreasing cost and time needed, sequencing an entire genome still requires a substantive amount of money and time. On the other hand, the increasing data whole genome sequences available openly online provides the opportunity of identifying useful Single Nucleotide Polymorphism (SNP). Accordingly Giffard, work is shifting to SNP mining from alignments of entire genomes rather than alignments of small fragments of genomes such as MLST databases. This means that ~1000 more SNPs are available for selecting optimal combinations, resulting in much higher performance of the genotyping methods and hence called for the improvement of the existing software.

# The team and skills

Kian Soon Hoon

* Experience in Python, C & C++

Wei Zhong Teo

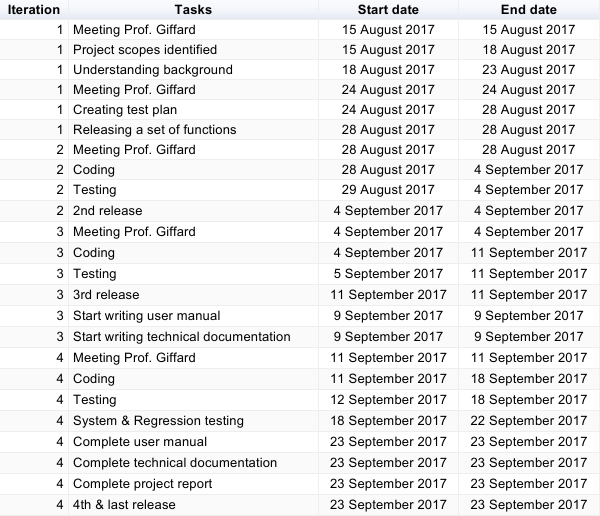
* Experience in web programming, Java, C & C++

Daniel Wilson

* Experience in web programming, Java & Documentation

# Milestones

1. Project scopes identified
2. Test plan
3. First release
4. Second release
5. Third release
6. Fourth/final release
7. Complete user manual
8. Complete technical documentation
9. Complete project report



# Methodology/Method

The project will utilise the agile methodology (Kanban). 4 iterations are incorporated into the kanban practice. At each iteration, there are few things that have to be done, which include: (1) Meeting with client, (2) Extracting features that can be integrated in the iteration, (3) Coding the selected features using Test Driven Development, and (4) releasing the set of features.

Kanban board is used to ease the communication between team members. Asana is chosen as the online platform which consist of a Kanban board. Besides, pair programming is also used to ensure that all team members have understand the code and can modify it in future.

# Risks

A list of risks as well as the mitigation strategies has been listed in the table below.

|  |  |
| --- | --- |
| Risks | Mitigation strategies |
| Lack of knowledge in the area of microbiology | 1. Background research (Internet, journal articles). 2. Consult with Phil Giffard. 3. Consult with microbiology master/PhD student under Phil’s supervision (in the event that Phil is unavailable). |
| Lack of experience programming in R | 1. Pair programming. 2. Doing MOOC online (coursera/edx). 3. Using online tutorial resources (tutorialspoint, etc.). 4. Utilising online discussion forum (stackoverflow, etc.). 5. Youtube. |
| Technical aspects of the algorithm | 1. Utilising open source library as much as possible. 2. Consult with Peter/Phil depending on the situation. 3. Ask question in online forum. |
| Weekly physical meeting | 1. Online collaboration tools (google docs, whatsapp, etc.). 2. Whenever possible, decide on a day when everyone is free. |

# Tools

-Git repository (Accessible at <https://github.com/ludwigHoon/PRT452-BINFO->)

-Git Desktop

-R programming environment

-Asana (Accessible at <https://app.asana.com/0/420615199430406/420615199430406>)

-Whatsapp for communication

-Email to schedule meeting & share files with client .

# Deliverables

1. Functional prototype of the improved software for SNP-based microbial typing in R programming language
2. User documentations on how to use the software
3. Test documentation
4. Project report