ABERYSTWYTH UNIVERSITY
COMPUTER SCIENCE AND STATISTICS (GG34)
CS396: MINOR PROJECT

Application of Machine Learning Techniques to Next Generation Sequencing Quality Control

Author:
Sam Nicholls msn

Supervisor:
Dr. Amanda Clare afc

Draft
May 4, 2014

Declaration

I certify that except where indicated, all material in this thesis is the result of my own investigation and references used in preparation of the text have been cited. The work has not previously been submitted as part of any other assessed module, or submitted for any other degree or diploma.

Sam Nicholls
2014

Contents

Chapter 1

Introduction

Over the past few years advances in genetic sequencing hardware have introduced the concept of massively parallel DNA sequencing, allowing potentially billions of chemical reactions to occur simultaneously, reducing both time and cost required to perform genetic analysis[1]. However, these "next-generation" processes are complex and open to error[2], thus quality control is an essential step to assure confidence in any downstream analyses performed.

1.1 Project Aims

The project consists of two sub-projects:

- Analysis of a current quality control system in place
- Identification of quantifiable sample properties that affect downstream analysis

1.1.1 Analysis of Current System

With the support of the Wellcome Trust Sanger Institute in Cambridge, this project works with the Human Genetics Informatics team to investigate **auto_qc**[3], the institute's current automated quality control tool.

During genetic sequencing a large number of metrics are generated to determine the quality of the data read from the sequencing hardware itself. As part of the institute's vertebrate sequencing pipeline[4], **auto_qc** is responsible for applying quality control to samples within the pipeline by comparing a modest subset of these metrics to simple hard-coded thresholds; determining whether a particular sample has reached a level that requires a warning, or has exceeded the threshold and failed entirely. Whilst this catches most of the very poor quality outputs, a large number of samples are flagged for manual inspection at the warning level; a time consuming task which invites both inefficiency and error[5].

In practice most of these manual decisions are based on inspecting a range of diagnostic plots, which suggest that a machine learning classifier could potentially be trained on the combinations of quality control statistics available to make these conclusions without the need for much human intervention.

The first part of the project aims to apply machine learning techniques to replicate the current **auto_qc** rule set by training a decision tree classifier on a large set of these quality metrics. The idea is to investigate whether these simple threshold based rules can be recovered from such data, or whether a new classifier would produce different rules entirely. During this analysis it is hoped the classifier may be able to identify currently unused quality metrics that improve labelling accuracy. An investigation on the possibility of aggregating or otherwise reducing the dimensions of some of the more detailed quality statistics to create new parameters will also be conducted.

The goal is to improve efficiency of quality control classification, whether by improving accuracy of pass and fail predictions over the current system or merely being able to provide additional information to a lab technician inspecting samples labelled with a warning to reduce arbitrary decisions.

1.1.2 Identification of Properties that affect Downstream Analysis

The other half of this project is motivated by the question "What *is* good and bad in terms of quality?"

To be able to classify samples as a pass or a fail with understanding, we need an idea of what actually constitutes a good quality sample and must look at the effects quality has on analysis performed downstream from sequencing. An example of such is **variant calling** – the process of identifying differences between a DNA sample (such as your own) and a known reference sequence.

Given two high quality data sources where DNA sequences from individuals were identified in two different ways (one of which being next-generation sequencing) it would be possible to measure the difference between each corresponding pair. Using this, we could investigate the effect of leaving out part of the next-generation sample during the variant calling process. If we were to leave a part of a sample out of the variant calling pipeline would the variants found be more (or less) accurate than if it had been included? Would they agree more (or less) with the variants called after using the non next-generation sequencing method?

Having identified such sub-samples, can quality control metrics from the previous part be found in common? If so, such parameters would identify "good" or "bad" samples straight out of the machine. Samples that exhibit these quality variables will go on to improve or detriment analysis.

1.2 Project Method

1.2.1 Methodology

Clearly some team-based practices invited by agile methodologies – pair programming immediately comes to mind – are not applicable in a solo project. It is also unreasonable to expect an "on-site" customer for this particular project. In *The Case Against Extreme Programming*, Matt Stephens describes a "self

referential safety net" where the perceived traps in each practice are supported and "made safe" by other extreme programming (XP) practices. This would rule out XP as a viable methodology for a solo project as cutting out some of the processes that allow this form of evolutionary design to work (and flatten that cost-of-change curve) can introduce serious flaws to the management of a project and potentially result in failure. In the same breath it is important to remember that not all agile processes need be discarded just because XP seems incompatible. Indeed, some processes are common sense, for example: frequent refactoring, simple design, continuous integration and version control. Test driven development could also prove a useful process to consider as part of a methodology for this project as setting up a framework that allows for quick and frequent testing (before coding) and ensuring that any refactoring has a positive (or at least non-negative) effect on the system could be a worthwhile contribution to efficiency.

Could a more plan driven approach or form of agile-plan hybrid be considered appropriate here? In *Balancing Agility and Discipline*, Boehm and Turner introduce the idea of "homegrounds" for both agile and plan driven approaches; noting here that for projects that require high reliability and feature a non-collocated "CRACK customer" in fact align with some of these homegrounds for plan driven development. Combined with the thought that the project requirements will also be relatively stable it would seem that there may be no reason to switch to a more agile methodology as its primary feature is the welcoming of change that is not even needed? Perhaps this is the naivety of an optimist.

Personally I think I would approach this with a form of agile-plan hybrid; I like the idea of quick iterations and getting feedback as opposed to leaving acceptance testing until the end of the project, but I also want a somewhat detailed feedback process. In Neil Taylor's *Agile Methodologies* course it was suggested that it is dangerous to pick and choose processes (don't anger the Ring of Snakes) and also merely paying "lip service" to agile must be avoided (otherwise what's the point?), I feel that on this occasion it can be justified by the size of the project itself.

This project will consist of many research steps, each requiring some form of computational process to prepare the data for the next step. Whilst the implementations of the algorithms themselves pose computational complexity, there appears to be little challenge from a planning perspective and in fact a looser overall plan should be considered as we must account for unforeseen and unexpected outcomes from each research step.

The most important part of ensuring this project stays on track will be the development of a sensible testing methodology to ensure we are not only moving in the right direction in terms of which algorithm and parameters to use but also in terms of reliably measuring performance over time in a way that allows justification of such design choices.

Despite this trail of thought, given the research grounding this project entails it might be required to look beyond traditional and even modern software development methodologies and investigate a more scientific approach. A simple scientific method would involve establishing a null hypothesis that can be proven false by testing (*e.g.* "**auto_qc** classifier is more accurate than the new classifier") and executing experiments that attempt to prove this null hypothesis false in favour of an alternative hypothesis (typically the opposite, *e.g.* "The new classifier is more accurate than the old classifier"). This form of hypothesis testing could essentially become the project's acceptance tests (providing we have an empirical definition of what "more accurate" means in terms of this system) and any modification can be classed as an experiment ("Do these parameters allow us to reject the null hypothesis?"). Although care must be taken not to let this descend into unstructured

cycles of mere hack-and-test, code-and-fix style programming.

Overall it is rather difficult to select a methodology for a project such as this as the research element makes it almost impossible to draw on previous personal experience for ideas of what development processes would be effective.

1.2.2 Task Management

It is useful to be able to keep track of current tasks preferably via a medium that would allow some method of sorting and filtering. For various personal projects and the second year group project I have used **Redmine**, a Ruby-based web application designed for bug tracking. However over time I have come to find keeping the task information stored in Redmine a task in itself. Attempts to extend the platform to implement additional functionality have been fruitless.

Out of frustration with thus and other alternatives – including **TaskWarrior** whose simple but effective command line interface was overshadowed by its occasional storage corruptions, I wrote my own open source web based task management application; **Triage**[6]. This will be useful in keeping track of current objectives and allow prioritisation in a quick and simple manner.

The list used to organise my project is also publicly available[7] for transparent progress tracking by my supervisor and those interested at the Sanger Institute.

1.2.3 Time Considerations

It must be remembered that this project needs to meet the requirements for a *minor* project and is to be completed alongside the study of several other modules which each have their own assignments and obligations. It would be easy to become overly ambitious and thus aims and goals will need to be revised as both obstacles and breakthroughs are encountered over the lifetime of the project.

Part I

Analysis of Current System

Chapter 2

Introduction and Background

2.1 Introduction

This part of the project can be outlined as follows:

- Collect data sets on which a machine learning classifier is to be trained
- Construct a program capable of processing and storing such data sets such that required subsets of the data can be quickly and easily returned for further analysis
- Select a suitable machine learning framework to handle the training and validation of a classifier
- Ensure a robust validation methodology exists for assuring quality of our own results
- Set up an environment capable of allowing results from such a classifier to be stored and compared
- Training a suitable classifier on the collected data sets
- Perform experiments by selecting subsets of the variables and observations and measure whether classification accuracy is improved

2.2 Concepts and Terminology

2.2.1 DNA and Bases

...DNA is made up of four bases; G, C, T and A...

2.2.2 Next Generation Sequencing

...the process of recovering or extracting the sequence of nucleotide bases...

2.2.3 Variants and SNPs

...for the purpose of this study we are only interested in single nucleotide polymorphisms (SNPs) "snips" ...

...their *position* refers to the index of the base of the chromosome in which the polymorphism exists... ...commonly these are 1-indexed (although this is not always the case) and thus care must be taken to ensure this is accounted for at any step where positions are considered...

...SNPs...

variant calling - the process of identifying bases in a sample that differ from the reference genome (this is a little simple as you also need to discern as whether the difference is actually a polymorphism or not).

2.2.4 Samples, Lanes and Lanelets

A **sample** is a distinct DNA specimen extracted from a particular person. For the purpose of sequencing, samples are pipetted in to a *flowcell* such as the one in Figure 2.1 – a glass slide containing a series of very thin tubules known as **lanes**. It is throughout these lanes that the chemical reactions involved in sequencing take place.

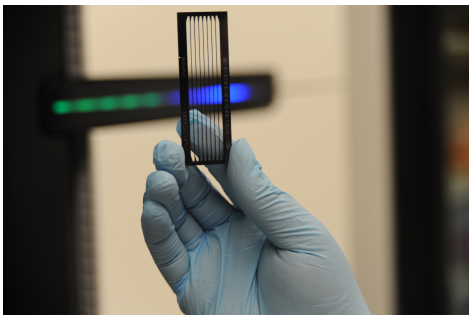

Fig. 2.1 An Illumina HiSeq Flowcell[8]

Once inserted, samples are amplified *in situ*, in the flowcell itself. A process in which the genetic material of each sample is caused to multiply in magnitude to form a dense cluster of the sample around the original. Millions of clusters will be created throughout each lane of the flowcell.

Note that a lane can contain more than one sample and a sample can appear in more than one lane; this is known as *sample multiplexing* and helps to ensure that the failure of a particular lane does not hinder analysis of a sample (as it will still be sequenced as part of another lane).

The more abstract of the definitions, a **lanelet** is the aggregate read of all clusters of a particular sample in a single lane. Figure 2.2 attempts to highlight examples of a this (circled in blue – not all lanelets are highlighted). For example Lane 5 shows the four clusters (in reality there would be millions) of Sample A combine to represent a lanelet. A lane will have as many lanelets as it does samples.

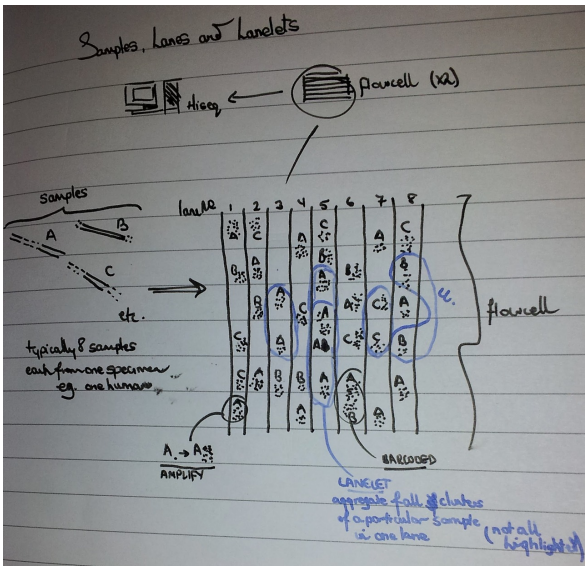

Fig. 2.2 Example of flowcell with some lanelets highlighted

Chapter 3

Materials and Methods

3.1 Input Data and Format

3.1.1 "BAMcheckR'd" Data

As part of the project I have been granted access to significant data sets at the Sanger Institute, unlocking quality control data for two of the largest studies currently undergoing analysis. A wide array of quality metrics are available for each and every lanelet that forms part of either of the two studies, totalling 13,455 files.

The files are created by **samtools stats** – part of a collection of widely used open-source utilities for post-processing and manipulation of large alignments such as those produced by next-generation sequencers that are released under the umbrella name of "SAMtools"[9] (Sequence Alignment and Map Tools). **samtools stats** collects statistics from sequence data files and produces key-value summary numbers as well as more complex tab delimited dataframes tabulating several metrics over time.

The output of **samtools stats** is then parsed by an in-house tool called **bamcheckr**[1] which supplements the summary numbers section of the **samtools stats** output with additional metrics that are later used by **auto_qc** for classification. This process appends additional key-value pairs in the summary numbers section. A truncated example of a "bamcheckr'd" file can be found in Appendix A.1.

It is these summary numbers that will be the main focus of our learning task.

3.1.2 auto_qc Decision Data

To use these "bamcheckr'd" files for training and testing a machine learning classifier, it is necessary to map each file to a classification result from **auto_qc**. The one-to-one mapping between each input file and its label

[1] Named such as **samtools stats** now incorporates **bamcheck** and the tool is written in R

are provided by the Sanger Institute in a separate file hereafter referred to as the *AQC Decision Matrix* or *AQC (Decision) File*.

A truncated example of such a file can be found in Appendix A.2. Only the first few columns are included – indeed we are only interested in the *lanelet* and *aqc* fields which provide an identifier that maps the row to a given input file and its classification by **auto_qc** respectively. Latter columns pertain to a breakdown of decisions made by **auto_qc** which are not included in the example for confidentiality (and brevity).

3.2 Development Environment

3.2.1 Language

Python was selected for the language of the program designed to handle this vast array of input data, more out of personal taste rather than a detailed analysis of required performance and features. From previous experience I was happy with the performance of Python when processing large datasets in terms of both file handling operations and storing the data in memory for later use. Python's generous choice of both built-in and third-party libraries have proven useful. Due to its concise and flexible nature it is possible to rapidly develop applications and its readability eases ongoing maintenance; useful given the short time-span allocated for this project and the possibility of others wishing to contribute to the project codebase after completion.

Whilst the choice was made primarily on preference, this is not to say other options were not considered: a highly popular Java-based collection of data mining tools, **WEKA**[10] would certainly have provided a framework for building decision tree classifiers but did not appear to offer any significant features that were unavailable elsewhere, whilst Java itself has the added constraint of requiring a virtual machine to be installed which could be undesirable from a performance or security standpoint when the application is deployed to servers at the Sanger Institute.

Difficulty was also encountered finding example implementations for **WEKA** with most documentation and tutorials providing information for performing analysis via the graphical "Explorer" interface instead, which would not be appropriate for quickly setting up and repeating experiments automatically.

Given the quality data we'll be using to train a machine learning classifier is output from the previously mentioned R script, **bamcheckr**, it was worth briefly investigating the options available for R itself as the potential of integrating the learning and predicting functions right in to the same process that outputs the data seemed convenient.

Whilst the **tree**[11] and **rpart**[12] packages are available for constructing decision trees in R (and actually **RWeka** provides an R interface to **WEKA**) neither appeared to be as robust as other more well-known frameworks. Also putting it politely, the programming paradigm of R[13] is rather different to other languages and can significantly increase development time (and frustration[14]) if one is not well versed in the patterns and grammar of the language and it seemed best to stick to one's comfort zone given the brief timescale for the project.

Had performance been a critical decision factor, lower level languages such as C, C++ or even Fortran could be

used. Briefly looking at frameworks available for C++ in particular, two popular solutions include: **dlib**[15], which did not support tree-based classifiers but did offer implementations of other potentially useful algorithms such as multi-class support vector machines; and **Shark**[16], which supported both decision tree and random forest classifiers. Both packages also provided a series of utility functions for performing mathematical or statistical operations on vectors of data.

3.2.2 Framework

Having studied the *Machine and Intelligent Learning* module in my final year, the prospect of getting stuck in to the deep of a machine learning algorithm was exciting. However the reality is a lot of cumulative time and effort has gone in to the creation and optimisation of a framework, which is unlikely to be surpassed successfully by a short-term one-person project. Thus utilisation of a third party machine learning library seems a wise investment for the project's codebase.

There are clearly numerous machine learning frameworks available in many languages, some of which were touched upon in the previous section and formed part of the development environment decisions. Whilst it is obviously unnecessary to select a framework which uses the same language as the project, it seemed counter-intuitive to select otherwise, for the establishing of additional arbitrary output and input steps to move data between the two environments could impede prompt experiment repeatability and introduce error.

A mixed bag of machine learning frameworks exist in Python[17], ranging from several highly active general purpose libraries to a multitude of smaller projects that focus on one specific learning task.

I investigate two of the larger libraries; **scikit-learn**[18] and **Orange**[19], primarily as their general purpose nature will allow exploration of various classifier solutions (given time) without the need to integrate many packages together, but also for their built-in functionality that aids the measuring of classifier performance and accuracy. These two packages were also selected for investigation on their recommendation from the project supervisor.

scikit-learn is somewhat the "new kid on the block" in terms of machine learning packages for Python, originally starting as a Google Summer of Code project in 2007 under the name of **scikits.learn**[2] the package was not officially published until 2010 where it quickly gained popularity and built an international team of contributors[20].

Arguably one of the most useful features of **scikit-learn** is its seamless integration with the "big names" of scientific computing in Python: **NumPy**[21] and **SciPy**[22], making heavy use of the optimised data structures and mathematical algorithms these packages offer, not only providing improved performance against packages that implement their own structures but also easing interaction by not requiring users to populate abstruse structures with their data or implement such algorithms themselves.

scikit-learn offers support for a wide range of algorithms covering many different machine learning tasks, including decision trees and random forests. Also of interest are the various utility subpackages, taking particular note of one for measuring the performance of the various classifiers; including functions for executing

[2] Named as a **SciPy** Toolkit for Learning Tasks

cross-validation and generating confusion matrices; and another which assists with the selection of parameters (features) to use in a machine learning model.

The library boasts a growing community of users and a dedicated team of developers[3] who submit hundreds of commits to the project repository each month according to Github Pulse.

Orange, like **scikit-learn** is designed to be a general purpose machine learning framework and provide a wide array of tools to work with many different types of problems. Setting it apart from many of the other packages is the inclusion of a graphical user interface (GUI) that presents drag-and-drop-eqsue access to widgets from which users assemble a workflow to apply to their data. The GUI also allows visualisation of elements such as decision trees or confusion matrix plots without any effort from the end user.

Despite the repository commit history stretching back to mid 90s, only recent improvements including a major redesign of the GUI, significant changes to its object hierarchy and its 2013 academic publication have brought more attention to the framework.

At first glance **Orange** appears to offer a larger collection of functions to work specifically with classification tasks when compared to **scikit-learn**. Indeed it would later be discovered that it ships with features useful to the project that have not been implemented in **scikit-learn** including tree pruning and printing.

However it does not feature the same integration with Python's scientific computing packages; **NumPy** and **SciPy** and also requires that end users input their data in a structure defined by the framework rather than a generic data structure. Thus use of **Orange** would introduce more data wrangling and munging steps to ensure that data is represented in the correct format for a classifier to perform its learning and prediction role and also to interpret returned objects.

Although **Orange** affords a suitable (and in some cases better) range of functions to use for decision trees, I was concerned at the push to use its GUI as opposed to the package API, potentially impeding experiment repeatability by having to reset the graphical environment each time.

Despite the **scikit-learn** package not yet reaching a major version milestone, its feature set, interaction with **NumPy** and **SciPy** and useful built-in subpackages led me to choose **scikit-learn** as the machine learning framework for this part of the project.

3.2.3 Additional External Libraries

As discussed, the project will make use of the open source[4] **SciPy Stack**, consisting of several core packages including **NumPy**[21]: a package for efficient numerical computation which defines generic N-dimensional array and matrix data containers; and **SciPy**[22], which provides a multitude of optimised numerical routines.

Whilst the stack also includes **Matplotlib**, a highly popular graphing package, I typically prefer to use the R package **ggplot2**[25] for creating graphs. However it is useful to note that **Matplotlib** integrates easily with the data structures of **NumPy** (and thus **scikit-learn** too).

[3] Some full time developer positions are funded by *INRIA*, the French Institute for Research in Computer Science and Automation

[4] Components of the **SciPy Stack** are distributed under the 3-clause Modified BSD License[23][24]

argparse is a third-party Python library[5] used for specifying the arguments and options of a command line interface via a simple API. **argparse** spares the developer from having to check the presence and validity of a user's arguments to a Python program themselves whilst also automatically generating a friendly interface to the program for the end user based on the definitions provided by the developer.

3.2.4 Tools

git

Keeping your code base inside some form of version control is common sense,

3.2.5 Testing

As discussed in Chapter 1.2.1, testing forms a critical part of the project given the need to monitor the impact of changes to classification accuracy as well as to ensure the program is working correctly. Ideally, execution of a test suite should be simple and easily repeatable. Results that pertain to accuracy should also be stored for future reference to monitor ongoing performance of the classifier.

Such requirements could be fulfilled by a continuous integration platform – a server dedicated to the building and testing of the code contained in a centralised repository typically to which an entire team will have write access[27]. Whilst in this scenario there will be much less "risk" from integration issues due to the single person team size, the themes of automated building and self-testing code can still be taken on board.

Jenkins[28] is a highly popular[29] example of such a platform with which I am familiar. Although an out-of-the-box **Jenkins** instance is suitable for variety of software engineering projects, it would be necessary to invest some time to install and tweak plugins to perform actions on test results (such as failing a build that causes accuracy to decrease). Previous experience found that more specific tasks will often require a plugin to be authored to overcome limitations in the feature set of a more generic plugin, which given the intricacies of the **Jenkins** package layout could easily turn in to a project of its own. Unfortunately, other features that would be useful to the project including the indexing and searching of logs are somewhat lacking in **Jenkins**.

Online solutions such as **Travis**[30] and **Wercker**[31] are free for open source projects and could potentially offer a quicker set up, as both services merely require a small configuration file in the root of the repository and a hook to be registered (allowing builds to be triggered on a code push for example) before being able to deploy a build in the cloud.

However these online services would not have been able to handle the return of build artifacts (such as logs, graphs or dot files containing a decision tree) without some convoluted solution of uploading them to another service or committing them to a private code repository during execution of the job – the non-persistent nature of these nodes would cause any artifacts to be destroyed when the node is terminated at the end of the build.

[5] Although since Python 2.7, **argparse** has been included as part of the standard library[26]

Also, as these virtual nodes are isolated from any sort of persisting file system it would be necessary to upload large quantities of training data any time a build and test job is to be run. This sounds rather inefficient in terms of both bandwidth and time and would more than likely constitute unreasonable use in view of the platform's "fair use" terms and conditions too.

It would seem that none of the widely available solutions would provide an environment suitable for the testing of this project. I wanted to provide my own platform for this problem but given the time constraints of the project this was simply not practical and in the end we settled for well formatted log files which could be searched and processed with command line tools such as **awk**.

Chapter 4

Pre-Implementation

4.1 Classification Correlation

An important consideration for statistical analysis is the relation between observations. The "bamcheckr'd" input data described in Chapter 3.1.1 is available per lanelet, however as shown in Chapter 2.2.4 a lane may contain more than one lanelet. Herein lies the trouble: if during a sequencing run the flowcell is somehow subjected to abnormal conditions (*e.g.* a temperature increase due to an air conditioning failure) or the device is depleted of reagents then every lane (and thus all lanelets within) will be of considerably poor quality.

Fig. 4.1 **Heatmap of lanelet QC status by lane**: Lanes are vertical bars with each lanelet cell coloured red to represent a failure, yellow for a warning and grey for a pass.

In such a case there would appear to exist a relationship between the respective qualities of each lanelet in

a lane as well as each lane in a sequencing run. To examine this further, an R script utilising **ggplot2** was authored to visually inspect whether correlation existed and if so to what degree that data is affected.

Figure 4.1 displays a plot of **auto_qc** classification for each lanelet in a lane. The plot itself is a dense heatmap where each lane stands as a vertical bar, broken in to horizontal cells, each of which represents a lanelet that was sequenced in that particular lane. These lanelets are colour coded using; red for failures, yellow for warnings and grey for passes (to allow the other two classes to be more easily seen).

Therefore an unbroken vertical red line indicates that all lanelets that comprise of that line failed to pass some aspect of the current **auto_qc** thresholds. In reality there are few conditions under which a lanelet would fail irrespective of the status of the rest of the lanelets in the same lane which typically involve an error during the preparation of the sample (an easy to spot result as it will cause poor quality across all lanelets using that sample).

Overall there are a series of instances appearing to support correlation for whole-lane failures and warnings but despite this there do appear to be occasions where a lanelet has failed where the remainder of the lane has not. Having discussed this with the project supervisor and contacts at the Sanger Institute we decided to continue to the implementation stage, agreeing that whilst some evidence of correlation between lanelets in the same lane has emerged, we will still be able to recover parameters that will be useful to quality control and statistical testing may be required following this analysis to describe how powerful such parameters are taking this possible correlation in to account.

It should be noted that the proportion of failures and warnings is considerably smaller than passes and so care will need to be taken to find a balance; for example it would not be feasible to merely discard lanelets from lanes that have failed entirely as there'd hardly be any data on which to train a classifier. Indeed other solutions may be possible, perhaps weighting observations which exhibit similar behaviour to other lanelets in their lane so as to give their parameter values less priority during the construction of the classifier itself.

It is worth noting that although the plot does not make a particular distinction between lanes in the same flow cell, lanes are sequentially identified so the red bars of thicker-width arguably display some failures across entire flow cells.

As a final note it should be stressed that this plot should be regarded as a diagnostic rather than an experiment with a direct conclusion. Given more time it would be useful to investigate the nature of these possible correlations, given a lane that has failed across all lanelets: do those lanelets actually express similar quality metrics?

4.2 Recovering Ratios

An initial scan of the summary numbers available in the "bamcheckr'd" files introduced in Chapter 3.1.1 revealed 82 different parameters. However, when comparing this parameter set to the threshold rules of **auto_qc**, it appeared that some parameters were "missing".

In the pursuit of replicating the decisions of the existing system, it is ideal to have all the parameters used

in the making of those decisions at hand. In particular, the missing parameters were of a normalised nature, typically in the form of a ratio or percentage which should make them more valuable than a parameter that just represents a raw count of some property.

It was found that these missing parameters are calculated in a step preceeding **auto_qc** as part of the **vr-pipe** pipeline[1].

Whilst I could have attempted to set-up my own instance of **vr-pipe** to recover this data, speaking with contacts at the Sanger Institute, it was decided that this would prove troublesome work; requiring an involved deployment to a cloud based facility such as Amazon's Elastic Compute Cloud and installation of various Perl dependencies as well as requiring access to many controlled databases within the institute.

Most of the functions that calculated these parameters turned out to be straightforward and could easily be ported to another application. At first it was intended to add these functions to the program authored for this part of the project, but the Sanger Institute suggested it would be more useful to implement such functionality in **bamcheckr** directly, removing some of **auto_qc**'s dependence on its position in **vr-pipe**.

```
# Install and include the 'devtools' package
install.packages("devtools")
library(devtools)

# Install package directly from Github repository
install_github("samstudio8/seq_autoqc", subdir="bamcheckr")

# Install package from local directory
install("/home/sam/Projects/seq_autoqc/bamcheckr")
```

Listing 1 : Installing an in-development R package with **devtools**

With the help of **devtools**[32] (see Listing 1) it was simple to develop and test contributions to **bamcheckr** locally without needing to re-publish the package after changes. An additional script was added to **bamcheckr**'s NAMESPACE... to recover the missing ratio and percentage based parameters...

...unfortunately, the performance of the script was poor, taking 5.5s on average and increasing to 16.1s when enabling a complex function containing many vector operations (potentially inefficient due to my lack of R experience)... for overlapping_base_duplicate_percentage

...utilising the program designed and implemented in the following chapter... loaded required information and used the API to calculate the missing parameters...

...Morandat[33]

...with such an intriguing difference in performance with significantly more time it would be appealing to explore the idiosyncrasies of each implementation ...although such an investigation would quite likely sufficiently

[1] https://github.com/wtsi-hgi/vr-pipe/blob/hgi-release/modules/VRPipe/Steps/vrtrack_auto_qc_hgi_3.pm

generate an entire project of its own.

4.3 Contributions to bamcheckr

...R CMD BATCH issue ...Fixed a graph plotting failure.

Chapter 5

Frontier

This chapter introduces **Frontier**: the main programming effort for this part of the project. Frontier is a Python package that serves as a data manager, providing both interfaces to read inputs into structures in memory and to retrieve them in formats acceptable to a machine learning framework.

5.1 Design

5.1.1 Purpose

Frontier's purpose is to supplement analysis with **scikit-learn** by allowing a user to read in and parameterise data in a format that can then be used for analysis by the **scikit-learn** library. Frontier was designed to simplify the process of setting up machine learning tasks and enable experiment repeatability by removing the need for users to spend time constructing classes and functions to parse their input data and to just get on with analysis using **scikit-learn**.

Initially Frontier was to act as a wrapper around **scikit-learn**, essentially removing the end user's interaction with the library and merely providing an interface for data to be passed in and some sort of classifier to be returned. However this quite clearly limited Frontier's audience by tying it to a particular framework and would quickly become unmanageable in the task of providing wrappers for all aspects of an external library.

Instead it was decided that Frontier would be used alongside a user's chosen machine learning framework, providing a useful API to parse and extract observations and variables from input data and arrange them in structures suitable for processing with **scikit-learn**.

If possible, Frontier could also handle any objects returned, displaying or logging any textual or graphical information pertaining to a returned classifier's accuracy to assist a user in the ongoing performance monitoring of changes to used data subsets or parameters.

5.1.2 Format

Frontier is designed as a Python package, allowing a user to import its functionality in to other programs. The result of this project's technical output could almost be considered as two separate entities: Frontier itself, the package designed to ease user interaction with scikit-learn and **Front**, a Python script which implements Frontier's functionality in order to interact with scikit-learn to conduct analysis on the **auto_qc** data.

5.1.3 Method

.. the result of abstracting code and tools created during the use of **scikit-learn** for analysis of the current **auto_qc** system .. initially hard-coded to suit the specific needs of the project but followed an evolutionary design process...

designed specifically for the given learning problem, with hard coded classes and encodings...

5.2 Concepts

In this section we introduce some ideas that will assist understanding of the application's purpose and terminology used in the following implementation section.

5.2.1 Observations and Parameters

An **observation** refers to a distinct object or "thing" from which we will attempt to understand the relationship between its known properties and classification to be able to label future observations with unknown class based on those properties. In the context of this project, this would be a lanelet. Other documentation in the field may use the term *sample* but to avoid confusion with the definition of sample introduced in Chapter 2.2.4 we will use the term observation.

The aforementioned *properties* or *attributes* of an observation will be referred to as the **parameters** of that observation. Traditionally these may be described as *features* but it was felt that this wording may have connotations with discrete data. Early versions of Frontier referred to parameters as *regressors*, using terminology from statistical modelling. This wording was dropped to remove any confusion between regression and classification machine learning problems.

5.2.2 Data and Targets

Data will be used somewhat generically to refer to a matrix in which observations act as rows and their parameters act as columns. It is expected that all observations will have the same set of parameters.

For problems of a classification nature such as ours, an observation's **target** is the known classification of that particular observation. Depending on the context of the learning problem targets may be discovered manually

(*e.g.* counting leaves from images of plants, one will need to manually count all the leaves in the image before being able to use it for learning) or as the output from another system such as **auto_qc**.

In the context of this project where the objective is to begin learning the rules of **auto_qc** the targets refer to whether a particular lanelet observation was labelled as a pass, fail or warn by the current system.

5.2.3 Class Labels and Codes

An observation's target may also be known as its **label**. A labelled observation is said to be a member of that label's **class**. For example a lanelet that has failed **auto_qc** is said to be labelled as a fail and is a member of the class of failures.

String based class labels can pose difficulties when handling data later on, not only taking more memory to store as opposed to a type such as an integer but also introducing accidental subsetting or discarding of data whose labels differ unexpectedly. For example are "fail", ,"fial", "FAIL" and "Failed" supposed to be members of the same class?

Often such labels are **encoded** in to simpler types such as integers.

5.2.4 Training and Testing

...

5.2.5 Python Data Structures

The Python standard library offers many different data structures each with their own properties. I introduce three such structures below to assist the reader in understanding some of the implementation decisions made in the following section:

List

A **list**[34] essentially represents an ordered sequence of elements. *Lists* are incredibly versatile, allowing elements of any data type, including *lists* and other container structures or arbitrary objects. Elements in a *list* are accessed by their index (location) in the sequence. *Lists* permit duplicate values.

Lists are represented in memory as a contiguous block and so great expense is incurred if the *list* grows beyond its bounds and must be relocated elsewhere in memory to resize to a larger contiguous block. Whilst items can be appended to or popped from anywhere in the container, performing those actions on a location that is not at the end of the *list* will require the contents of the structure to be moved in memory to maintain the contiguous layout.

Checking whether an item is in a *list* requires iterating over all n elements in the container until either a match is found or there are no elements left to check.

Dictionary

A **dict**[35] or *dictionary* is a structure containing key-value pairs in which string or numeric keys[1] are mapped to an arbitrary value or object through a hashing function. Keys in a *dictionary* are unique; attempting to add a key-value pair where the key already exists will overwrite the original entry.

Python *dictionaries* offer constant lookup access and more importantly, unlike a list, querying the structure for whether it contains a particular key can be performed in constant time[36]. However it is important to note that a *dictionary* does not maintain order, keys are not sorted alphanumerically nor are they sorted by insertion time. Ordering is determined arbitrarily by the **dict**'s underlying hash table[2].

Set

Sets[37] represent a collection of elements and are similar to a **dict** in that they are allowed to contain any immutable (thus hashable) elements but entries are not mapped to a value or object. *Sets* are forbidden from containing duplicate objects and do not maintain order.

Like **dict**, checking for membership of an item in a *set* completes in constant time, making the structure particularly useful for membership testing and filtering duplicates out of other structures.

5.3 Implementation

This section investigates the implementation of Frontier's major components:

- Informing the package of the problem domain...
- Interfaces provided for reading in data
- Storage of read data in memory
- Retrieval of stored data
- Interaction with scikit-learn

5.3.1 Class Definitions

Frontier was designed to support classification machine learning problems, to adequately support this task, the package must be aware of each of the possible classes in the problem space. Early versions of Frontier were specifically designed for training and testing data from **auto_qc** and could only support encoding and decoding of the pass, fail and warn classes.

[1] Any type can actually be used as a dictionary key providing it is immutable, including tuples[35]
[2] This is probably the question I see most often from users new to Python on StackOverflow

However this implementation was clearly esoteric and held little to no further use outside the domain of the project's learning task. What would happen if a class label were to be added or removed in future? Most likely many lines of code would need to be re-written to handle such cases; the package was inflexible.

To counter this, Frontier was refactored to remove hard coded label definitions enabling its use as a more general purpose tool where users can specify the domain's classes and their associated labels and encodings. Listing 2 shows the definitions used by Front to work with the **auto_qc** data.

```
CLASSES = {
        "pass": {
            "names": ["pass", "passed"],
            "code": 1,
        },
        "fail": {
            "names": ["fail", "failed"],
            "code": -1,
        },
        "warn": {
            "names": ["warn", "warning"],
            "code": 0,
        },
}
```

Listing 2 : Class definitions for **auto_qc** as passed to Frontier

As per Listing 2, to define classes a user must provide a Python dict containing the following for each label:

- **class** *String*
  The dictionary key is used as the canonical name of the class label
- **names** *[String]*
  A list of labels that denote membership of this class
- **code** *Integer*
  The encoded representation of this class (See Chapter 5.2.3)
- **_recode** *Boolean, Private*
  A flag to indicate whether an API action has changed the code of this class, typically used when treating all members of one class as a member of another – A user should never set this manually
- **_count** *Integer, Private*
  The number of observations with this label, typically used when calculating weightings based on the proportions of class sizes and outputting logging information – A user should not set this manually

This simple structure allows Frontier to be compatible with almost any classification learning task with minimum input from the end user. Additional utilities provided by the Frontier utils subpackage use this

structure to automatically classify and encode labels without user intervention with the following functions:

- **classify_label**
  Attempt to classify a label by comparing a given string to each set of *names*, locating an exact match will return the relevant canonical class label
- **encode_class**
  Given a canonical class label, return its *code*
- **decode_class**
  Given a *code*, return the canonical class label unless *_recode* is True
- **count_class**
  Increment the *_count* for a particular class given its canonical label

These functions are put to use throughout Frontier and are essential for reader classes (detailed in the next chapter) to parse, classify and then encode observation targets automatically from relevant files.

5.3.2 Input Handling

Frontier's modular nature allows users to write their own Python classes to read data from any form of input file or stream. Two examples of which are the classes used to read from the "bamcheckr'd files" documented in Appendix A.1 and the **auto_qc** decisions matrix briefly demonstrated in Appendix A.2.

These classes are described as **Readers** and implement a common base class, **AbstractReader** which takes care of setting up the file handler, including functions to both close and iterate over the file's contents. It will also automatically call its own **process_file** function that skips over any header lines before passing each line in the file stripped of any newline characters to **process_line**.

It is expected that derived classes will at least provide their own implementations for **process_line** and **get_data**. Failing to do so will cause Frontier to throw a **NotImplementedError** when attempting to use the class to read data.

process_line defines the line handling operations that extract and store desired data found in a given line. This responsibility includes returning None for lines that contain comments and irrelevant data and conducting any necessary sanitisation (the **BamcheckReader** for example makes use of a utility function to translate spaces, underscores and dots to hyphens).

get_data must return any read in data in a suitable structure for storage by Frontier. Typically this will be a Python dictionary using some unique identifier for each observation as a key, mapping to an arbitrary value or object containing that observation's parameters. Further discussion on Frontier's storage of data and targets is to follow.

The **AbstractReader** class is designed to simplify the process of reading in observations and their targets for end users, however it is still up to the author of the dervied class to set up any structures to store data (which cannot be done automatically without likely enforcing potentially unhelpful constraints) before initialisation of the inherited base class (via the call to **super**) as shown in Listing 3.

```
class BamcheckReader(AbstractReader):
    [...]

    def __init__(self, filepath, CLASSES, auto_close=True):
        self.summary = SummaryNumbers()
        self.indel = IndelDistribution()
        super(BamcheckReader, self).__init__(filepath, CLASSES, auto_close, 0)

    [...]
```

Listing 3 : Extract from **BamcheckReader** class documenting initialisation of necessary data structures and calling for initialisation of its inherited base class

As shown in Listing 3, the initialisation of the **AbstractReader** allows four arguments:

- **filepath** *String*
  A relative or absolute path to a file from which to extract data or targets
- **CLASSES** *Dictionary*
  A dictionary of user specified class labels defined as described in the previous chapter
- **auto_close** *Boolean, Optional*
  Whether or not to close the file handle immediately after executing **process_file**, this is True by default to prevent either memory leaking when users are reading in a large number of files and are perhaps unaware that they require closing or the throwing of an IOError caused by having too many file handles open at once
- **header** *Integer, Optional*
  The number of lines to ignore before the reader should begin passing stripped lines to **process_line**, defaults to 0

Currently it is also the responsibility of the author of a derived class to perform relevant sanity checking of any extracted data. For example the **BamcheckReader** class checks for the presence of multiple entries of a particular metric which will print a notice if found, unless the entries have differing values, upon which an exception is thrown and the process is halted.

Once a reader has been defined for a particular file format, a user need only provide a directory of files to be parsed and the name of the class designed to complete the parsing. Frontier will then take care of executing the parsing process on all files in the directory. After a derived reader has completed file handling, Frontier will call its **get_data** function to "move"[3] the extracted data to its own storage.

At this point Frontier will also check the integrity of the data, primarily that all parameters have a non-zero variance. Users will be warned when this requirement is violated; parameters with no variance cannot provide much information for successful classification as their values are equal for all class labels!

[3] Rather, a pointer to the address of the extracted data's dictionary hashmap will be copied to memory inside a Frontier class

In future it would be useful to investigate whether it would be feasible to perform such sanity checking in a generic manner to ensure it could be applied to a wide enough range of scenarios to make it worth including functionality in the **AbstractReader** directly.

With more time, future improvements could overhaul the reader interfaces to allow users to simply specify the format of a file in a string that can be parsed by Frontier's IO subpackage, rather than having to write their own derived class. Classes could also list file extensions they are capable of processing which could potentially be used to automatically determine which readers to use without requiring the user to explicitly specify.

5.3.3 Storage

Frontier specifies a class called the **Statplexer**[4] which provides users with a single point of access to all read in data. The reader interfaces described in the previous chapter implement their own loading functions to populate the **_data** and **_targets** class members of the **Statplexer** object, both of which are Python dictionaries.

During the parsing of observation data with the relevant reader class, the reader is expected to locate an appropriate unique ID for each observation. In the case of processing of **auto_qc** data, this would be the lanelet's barcode which is collected from the filename of that particular lanelet's "bamcheckr'd" file. This identifier is then used as a key in both the **_data** and **_targets** dictionaries to map to a structure (typically another dictionary or an arbitrary class) that stores that observation's parameters and its known target (encoded class label), respectively.

Although these attributes can be manipulated directly (and indeed they are for testing purposes) the leading underscore follows a popular convention defined in Python's style guideline, PEP8[38], where class members with leading underscores should be treated as non-public. Python doesn't have private variables such as those that may be found in other languages like Java, indeed the Python style guide points out that "no attribute is really private in Python"[38]. In an interesting StackOverflow answer on the subject, a user describes that this is "cultural"[39] and that Python programmers are trusted not to circumvent convention and "mess around with those private members". Users are therefore expected to use the functionality Frontier provides for controlled getting and setting of data stored in these pseudo-private **_data** and **_targets** members.

load_data, for example, should be used exclusively when populating the two dictionaries and is automatically called on construction of the **Statplexer** if the construction arguments are valid. **load_data** will automatically call other necessary (pseudo-private) functions of the class including **_test_variance** which checks that parameter variances are non-zero and also warns users if new observations are overwriting old ones, or if an observation does not appear to have a corresponding target.

The following section details the retrieval of data and targets from the **Statplexer** which are returned to the user in **NumPy** arrays. Why does **Frontier** not just store the read in data in such a structure to begin with?

This is primarily due to the underlying layout of a list structure as introduced in Section 5.2.5. Given the nature of input data, the number of observations and parameters are unknown before the input data is actually read and so memory cannot be reserved to prevent these operations. The potential number of input files renders

[4] A somewhat contrived contraction of 'Statistics Multiplexer'

reading through the data once to reserve the right amount of memory for a second read-through impractical in terms of time.

It is arguable that despite this, the data could be unloaded from the **_data** and **_targets** dictionaries in to some form of array at the end of the call to **load_data**. However as the **Statplexer** allows loading of additional data at any time, which would not only risk increasingly expensive resizing operations as the list continues to expand but the sanity checking that takes place in **load_data** involves checking for membership of a given ID in **_data** and **_targets** which is constant for dictionaries and significantly slower for large lists (see Section 5.2.5).

It is possible that an implementation could use both a dictionary and an array; appending the observations to the array and entering a mapping between the observation ID and its index in the array. Though this would still not avoid the issue of relocating the array once it grows beyond its bound in memory.

Worse still, as results requested from the **Statplexer** API are returned to the user in a sorted matrix[5] (*i.e.* rows and columns are ordered alphanumerically by the observation ID and parameter name, respectively), if observations are not processed in a sorted order (which is not a requirement) then returning results will involve accessing the proposed **_data** and **_targets** arrays in a non-linear fashion, losing any efficiency that could be gained from using an array in the first place.

Whilst not entirely optimal[6], in terms of a compromise between memory complexity and avoiding expensive operations on data input and retrieval, dictionaries offer the best balance.

It should be noted that the **Statplexer** stores a copy of the user defined CLASSES dictionary (as the class member **_classes**), which is an argument to its construction, allowing the **Statplexer** to share this information with any readers or utility functions who require it.

5.3.4 Retrieval

Frontier's **Statplexer** is designed to provide functions to an end user for extracting desired data in a format suitable for passing as an argument to functions of an external framework or library, such as **scikit-learn**. As specified in **Frontier**'s purpose, the package must provide user-friendly functions to extract observations and parameters of interest from the internal **Statplexer** representation.

The previous section suggests that directly accessing elements of the **Statplexer**'s **_data** and **_targets** member dictionaries, whilst possible, would violate the psuedo-private nature of the variables and even so, this would hardly be a user friendly way in which to obtain data from the **Statplexer**.

As described in Section 5.1.3, development of **Frontier** was evolutionary, growing to meet the needs and requirements of the underlying machine learning problem that this project strives to solve. It became clear that in trying many combinations of observation and parameter subsets that **Frontier** would need to provide not just a function to return the contents of **_data** and **_targets** but to assist in retreiving similar subsets automatically with as little effort from the end user as possible.

[5] A two-dimensional **NumPy** array
[6] Dictionaries still require use of **sorted** when creating a result matrix via the API

But how can a user make an informed decision on parameters to subset? Though this is primarily the role of feature selection, which will be discussed later in this chapter, firstly a user will need to have a feeling for what parameters are actually present and ideally be offered functionality to allow for quick elementary exploration of that set. For this, **Frontier** can be used to inspect the parameters extracted from the observations via the functions:

- **list_parameters**
  Return a sorted list of all parameters
- **find_parameters**
  Given a list of input strings, return a list of parameters which contain any of those strings as a substring
- **exclude_parameters**
  Given a list of input strings, return a list of parameters which do not contain any of the input strings as a substring, or if needed an exact match

Once users have an idea of the parameters they wish to extract from each observation, **Frontier** maintains an additional set of functions that act as a form of API, allowing users to retrieve subsets of data based on both parameters and targets:

- **get_data_by_parameters**
  Return data for each observation, but only include columns for each parameter in the given list
- **get_data_by_target**
  Return data for each observation that have been classified in one of the targets specified and additionally only return columns for the parameters in the given list

The prior section first introduces the format in which data is returned, the two-dimensional **NumPy** array – referred to as the *Frontier Results Matrix* – in which a row represents a particular observation and each column represents a parameter (or feature). By default both dimensions are ordered alphanumerically; rows are sorted by their observation ID, columns by the parameter name[7]. Targets are also returned in a **NumPy** array, with each index of the target array mapping 1:1 with the observation row of the same index in the *Results Matrix*.

5.3.5 Interaction with scikit-learn

Submitting Data

...

Cross Validation

...method in which to measure classification accuracy... ...potentially use a weighting to penalise mistakes in smaller classes...

[7] Typically the 'cleaned' or otherwise sanitised parameter name, as noted in Section 5.3.2 for example where the **BamcheckReader** class will automatically convert spaces and other predefined characters to hyphens.

...K fold cross validation ...using stratified K fold cross validation...

Confusion Matrices

"Normal" confusion matrix and "Warnings" confusion matrix...

5.3.6 Logging

Section 3.2.5 concluded that time constraints would not allow significant effort to be invested in building a platform for cataloguing results from test runs of the classifier. I decided to settle for populating fields in a structured text file, an example of which can be found in Appendix B.1.

Frontier provides functions such as **write_log** to dump **Statplexer** parameters to a given file path along with cross-validation performance associated with a particular training and prediction run of a decision tree classifier. Such files can easily be manipulated with common command line tools such as **grep** and the text processing language, **awk**.

Given the time to create specialised software for storing and comparing different runs of the classifier in future, it would be trivial to import old **Frontier** log files.

5.4 Usage Example

```
from Frontier import frontier
from Frontier.IO import DataReader, TargetReader

data_dir = "/home/sam/Projects/owl_classifier/data/"
target_path = "/home/sam/Projects/owl_classifier/targets.txt"

CLASSES = {
        "hoot": {
            "names": ["owl", "owls"],
            "code": 1,
        },
        "unhoot": {
            "names": ["cat", "dog", "pancake"],
            "code": 0,
        },
}

statplexer = frontier.Statplexer(data_dir,
                                 target_path,
                                 CLASSES,
                                 DataReader,
                                 TargetReader)
```

Listing 4 : Example usage of Frontier

Constructing the Statplexer requires the following arguments:

- **data_dir** *String*
  Root data directory under which all files will be parsed for observation data
- **target_path** *String*
  Path to file to be parsed for observation targets
- **CLASSES** *Dictionary*
  A dictionary of user specified class labels defined as described in Chapter 5.3.1
- **DataReader** *Module*
  Module containing the class (of the same name) to be used to parse each file in the *data_dir* tree for observation data
- **TargetReader** *Module*
  Module containing the class (of the same name) to be used to parse the *target_path* file for target data

5.5 Testing

Frontier is bundled with a small testing suite designed to flex the functionality of the two included readers (**AQCReader** and **BamcheckReader**) as well as both the **Frontier** utilities and the **Frontier** API itself. The suite consists of three Python scripts which utilise the **unittest**[40] package in the Python standard library. The scripts are located in the *tests* directory of the **Frontier** package.

Listing 5 displays the commands necessary to execute each test script. Note the `-m` option to the `Python` command instructs the interpreter to load the module named as a script.

```
python -m tests/frontier
python -m tests/aqcreader
python -m tests/bamcheckreader
```

Listing 5 **Execution of the Frontier Testing Suite**

5.5.1 AQCReader and BamcheckReader

Both of **Frontier**'s included readers are distributed with their own modest test suite, designed to test the particular functionality of that reader.

* Test cleaning of data * Test duplicates warning * Test duplicates exception * BC Tests that some hardcoded expected values are returned * AQC Tests that number of labels were as expected...

5.5.2 Frontier Utilities

* Test both encoding and decoding of class labels and class codes * Ensures that exceptions are thrown when unknown class labels or codes are encountered (as this should not happen!)

5.5.3 Frontier API

* More complex * Sets up a series of test observations and parameters * Ensures each API function returns the data expected * ...including when parameters are incorrect or unknown * ...ensures parameters should be excluded or included... * actually found a problem with excluding! * Ensures targets map correctly to data

Chapter 6

Results

6.1 Introduction

6.1.1 Why Decision Trees?

6.1.2 CART

6.2 Parameter Selection

...important to find the "best" parameters ...what is best? scikit-learn uses total gini information

...frontier uses two methods:

- Backward elimination; pruning parameters with the lowest total gini
- Call scikit-learn's SelectKBest

* incorrect degrees of freedom * warnings: /usr/lib64/python2.7/site-packages/sklearn/feature_selection/univariate_selection.py:
RuntimeWarning: invalid value encountered in divide, causing NaN * Replaced univariate_selection with
version from master * needed use force np.float64 * ...actually data was 0... gg

6.3 Trees

6.3.1 Initial Trees

6.3.2 Parameter Sets

6.3.3 Ignoring Warnings

Part II

Identification of Qualitative Sample Properties

Chapter 7

Introduction and Motivation

7.1 Introduction

The second part of the project can be outlined as follows:

- Collation of variant locations from Sanger data sets
- Construction of script to select a 'representative' genomic region
- Extraction of regions from whole-genome sequence data
- Establishment of leave-one-out analysis pipeline
- Comparison of concordance between pipeline results and SNP chip

7.2 Background

7.2.1 GWAS and iCHIP Data Sets

As briefly explained in Section 1.1.2 this part of the project focuses on measuring the similarity between called variants in pairs of samples which have been sequenced in different ways. The Sanger Institute has granted access to a set of human samples which have been sequenced twice, once with next-generation sequencing (NGS) and the other with SNP genotyping. The resultant data sets will hereafter be referred to as:

- **GWAS**[1] *NGS*
  Every base in the genome of the sample has been sequenced with some level of accuracy.
- **iCHIP** *SNP Genotyping*
  Only particular bases of interest in the genome of the sample has been sequenced with high accuracy.

[1] Genome-Wide Association Study

Each sample will thus have a corresponding result from each of the two studies. A result in this context refers to the output from a variant calling pipeline at the Sanger Institute – a **VCF** file (introduced in Chapter 8.1) detailing variant sites that were located during analysis of the sequenced sample.

7.2.2 The 'Goldilocks' Region

For the next step of the project we are looking to document what effects lanelet quality has on analysis that occurs downstream from whole-genome sequencing. To investigate this I'll be performing a leave-one-out analysis on the **GWAS** data set: consisting of leaving a particular lanelet out, re-performing variant calling and comparing the resulting variants called for each sample to their corresponding **iCHIP** variants.

The basic idea is to answer the question of what constitutes "good" or "bad" lanelet quality. Does leaving out a lanelet from the full sequence data lead to variant calls that better match the SNP chip data, or cause the correspondence between the two sets to decrease? In which case, having identified such lanelets, can we look back to the quality parameters we've been analysing so far and find whether they have something in common in those cases?

If we can, these parameters can be used to identify "good" or "bad" lanelets straight out of the machine. We know that lanelets that exhibit these quality variables will go on to improve or detriment analysis.

However, variant calling is both computationally and time intensive. Whilst the Sanger Institute has significant computing resources available, the project time scale is too small to conduct the analysis on thousands of full sequence samples and thus we must focus on a smaller region of the human genome.

It is for this reason we are looking for what Sanger described as a "representative autosomal region". The region must not have too many variants, or too few: a "Goldilocks genome".

7.3 Concepts and Terminology

In the search for the aforementioned "Goldilocks genome", we define the following terminology:

7.3.1 Group

Rather than considering the location of all variants in our samples as a whole, we may wish to consider how variants are distributed across the two different data sets we have, namely the **GWAS** and **iCHIP** sets. Thus variants will be considered as members of a **group**. The group may be arbitrarily defined but for the purpose of our own analysis the set in which a variant appears will be considered as its group.

A particular variant location may appear in more than one group.

7.3.2 Length and Stride

The **length** of a region represents the number of bases included in the region and the **stride** refers to the number of bases to be used as an offset between the first base of one region and the beginning of the next.

Sensible choices for these parameters will need to be selected after considering both time and memory.

7.3.3 Density

The **density** of a region represents the number of variants located within that region. **Density** will be recorded for each *group* of variants.

7.3.4 Candidate Regions

A **candidate** is a region of *length* bases whose first base falls on a *stride*. A **candidate** region will contain some count (described as the *density*) of variants for each group. In the context of this project we consider the number of variant locations that fall within this region for both the **GWAS** and **iCHIP** data sets.

Candidates must meet all (any any additional) criteria to be considered. For example, regions whose first position falls on a base in such a way that the chromosome ends before a region of *length* can be constructed will violate the size criteria.

Chapter 8

Materials and Methods

8.1 Input Data and Format

8.1.1 Variant Call Format

Variant Call Format (**VCF**) is a (typically compressed) text file... ...containing records ...developed for the 1000 Genomes Project

8.1.2 Variant Query File

Named such for the command that creates it, the **Variant Query File** (or *Query File*) is an input format for populating the candidate searching script with variants. The file is formed of line-delimited, tab-seperated entries where the first field consists of a colon-delimited pair representing a chromosome and position respectively (*e.g.* 1:5000 is the base at position 5000 on the first chromosome). Other fields may be provided but are currently ignored.

The **Query File** is generated by the **vcf-query** command (demonstrated in Listing 6) which is a member of the **vcftools** collection – a tool set designed... specifically... for handling **VCF** files.

...although having found this tool, it turns out that **bcftools** implements a more performance efficient **query** function, thanks to the performance gains as a result of using **htslib**... ...the command uses exactly the same syntax, as shown in Listing 7.

Note that information on downloading and installing both tool sets can be found in Appendix C.

```
vcf-query cd.ichip.vcf.gz -f '%CHROM:%POS\t%REF\t%ALT\n'
```

Listing 6 : Extracting variant positions from a **VCF** file with **vcftools**

http://vcftools.sourceforge.net/perl_module.html http://www.biostars.org/p/51076/ http://vcftools.sourceforge.net/htslib.html#qu

```
bcftools query -f '%CHROM:%POS\t%REF\t%ALT\n' cd-seq.vcf.gz > cd-seq.vcf.gz.q
# ! A faster alternative to using vcftools exists in bcftools ! #
# Complains about lack of tabix index but still makes the file...
```

Listing 7 : Extracting variant positions from a **VCF** file with **bcftools**

8.1.3 Paths File

The **Paths File** is a trivial and esoteric file type used to handle grouping of **Variant Query Files**. The format is simple, using lines beginning with a '#' to delimit desired file groups. Entries following that line (until the next group definition) are considered to be a member of that group. Entries are line-seperated, tab-delimited and consist of only two fields: a unique slug that serves only for simple user identification of a file, followed by the the actual file path (relative or absolute).

It is an error to provide a file entry before the first group definition. An exception will also be raised if a user defines a group more than once or uses a duplicate slug due to the ambiguity this introduces.

An example can be found in Appendix A.3.

8.2 Development Environment

Many of the features of the development environment for Part I are just as applicable to Part II. Differences and additions to the environment described in Chapter 3.2 are outlined below.

8.2.1 Language

Once again, Python was the language of choice for many of the reasons previously discussed in Chapter 3.2.1. Additionally, Python is useful for rapid prototyping, thanks partly to its emphasis on readability; concise syntax, its dynamic type system and extensive standard library; making it a suitable choice to author a working script in a short time span.

Although performance was more of a concern, with the use of **NumPy** (whose mathematical functions are typically wrappers around C or Fortran implementations) reasonable efficiency was expected to be obtainable.

8.2.2 Testing

Given the more traditional development methodology, testing need not follow a continuous integration style roadmap. A test suite would need to be written to ensure the components of the script work as intended, simulating input variants and ensuring the correct candidate region densities are found.

Chapter 9

Goldilocks

This chapter introduces **Goldilocks**; the first programming hurdle for this part of the project. **Goldilocks** is a Python script designed specifically to read in and store locations of user-selected variants on the human genome and to then find regions of a given size on the genome where variants are of a desired density.

9.1 Design

9.1.1 Purpose

Expanding on the introduction, **Goldilocks** is responsible for parsing input files that contain chromosome and base position entries (in the format described in Section 8.1.2). These files can be grouped to pool the set of unique variants between each file group to treat them as one search space.

For each group, a list of positions for each chromosome is stored in memory. Once all variants have been loaded **Goldilocks** is designed to iterate over each chromosome[1] and load all of the variants from the lists in to an array and essentially count the number of variants present (in each group) between a given start and end position, determined by the arguments specified by the user as described in Section 7.3.2.

This combination of a start and end position and the counts of variants present inbetween forms the structure of a **candidate**. Having performed this search over all the autosomes, candidates can then be ranked by user-specified criteria related to the density of the counted variants.

9.1.2 Format

Goldilocks is a modestly sized but elegantly complex Python script which can also serve as a stand-alone module to be imported in to other programs. Whilst the use case is certainly more esoteric than **Frontier**, I

[1] Strictly speaking we are only handling the 22 *autosomes* and ignoring the remaining 2 sex chromsomes (*allosomes*)

made the choice to modularise the functionality, in the hope that if an end user were to find some function of the script particularly useful they may call it as part of their own program.

9.1.3 Method

Goldilocks had to be constructed quickly but carefully. Although avoiding a quick and dirty solution, the early codebase was arguably cluttered with blocks of repeated code to perform similar actions for different hard coded groups. Initial testing was not automated and utilised eyeballing of resulting data and occasional checking of input files to ensure results were as expected.

Development was akin to a more traditional methodology; requirements were simple and available up front and formal testing was scheduled to be completed following the completion of a working prototype.

9.1.4 Pitfalls

Python lists and other sequence based structures such as **NumPy** arrays are 0-indexed, however base positions in a genome are commonly 1-indexed. These differing indices must be normalised in some fashion; for example subtracting 1 from all positions or simply ignoring the presence of element 0 in the structure. The method that is used does not matter so much as consistency; it is trivially easy to make 1-base-out mistakes when handling genomic data.

Of note also; after a discussion with the Sanger Institute it has been decided to exclude candidates located on chromosome 6 from being utilised as a "Goldilocks region". Chromosome 6 contains a system of genes called the Human Leukocyte Antigen (HLA)[2] which encode for Major Histocompatibility Complex (MHC)[43]: a critical component of the body's immune system and of particular importance to anyone requiring an organ transplant. The HLA can cause difficulties during the variant calling process due to the high level of variability expressed between samples in these areas and could thus give less accurate results if we were to measure concordance using candidates that include these genes.

9.2 Implementation

This section considers the implementation of the major components of **Goldilocks**:

- Handling user defined Paths File
- Loading variants from specified Variant Query Files
- Storage and recall of variant groups in memory
- Searching for candidates with user specified length and stride

[2] A much younger version of me was introduced to this concept via the partwork *How Your Body Works*[41], which also distributed English dubs of a 1987 French television show titled *Il était une fois... la vie (Once upon a time... Life)[42]* a relevant episode can be viewed here: http://www.youtube.com/watch?v=yvhU2UyRv-Q

- Ranking suitable candidates based on variant density

9.2.1 Input Sanity...?

Having recovered from the miscommunication that led me to attempt to find reverse strands in VCF files with already known errors, I've been making performance improvements to the script that finds candidate genomic regions. The task poses an interesting problem in terms of complexity and memory, as the human genome is over three billion bases long in total which can easily lead to data handling impracticalities.

9.2.2 Parsing Paths File

Goldilocks requires the user provide their own valid **Paths File**, meeting the specification introduced in Section 8.1.3 (an actual example of such a file can be found in Appendix A.3).

As previously described the file merely documents the locations of the **Variant Query Files** to be loaded, along with the desired grouping of those files. Any variant location that appears in at least one file of a group will be added to the pooled set of variants for that group.

Goldilocks provides the simple **load_variant_files** to parse the user's **Paths File**, where for each group, the following structures are initialized and added to a dictionaries encapsulated by **Goldilocks** itself, using the group name as a key:

- **groups** *Dictionary*
  Keys refer to a particular allosome number, mapped to a list structure which will later be populated with the positions of all variants that appear on that chromosome in at least one **Query File** of the group. It is from this structure that chromosome **NumPy** arrays will be loaded with data.

- **group_buckets** *Dictionary*
  Keys refer to a specific number of variants counted (considering only variants from the group), mapped to a list of all i'th candidate regions which contained that density. This structure eases locating groups which match desired density criteria once summary statistics have been calculated.

- **group_counts** *List*
  The number of variants which appear in each candidate, for this variant group will be appended to this list. This structure simplifies the gathering of summary data (such as mean, median and variance) of all the candidate regions of the group once the search has completed.

A **Paths File** which violates the specification will raise an exception during execution of this function, informing the end user of their error. This was decided to ensure ambiguous input did not cause confusion later, for example should defining a file group twice merge or override the first definition?

9.2.3 Loading Variant Query Files

load_variants_from_file is responsible for loading a given **Variant Query File** and adding the position of each record contained within to the appropriate group-chromosome list pair in the **groups** structure.

It is important to note that this function is not responsible for ensuring the positions added to the list are unique to avoid performance penalties in checking whether a variant is already a member of the list (which increase proportionally for the size of the list). However as examined in Section 5.2.5, appending a significant number of items to a list will likely induce costly memory operations as the list is resized. Thus use of a set might have been more appropriate (especially considering the container forbids duplicate entries) but a set must be constructed from an existing iterable such as a list anyway!

Indeed the container does specify an **update** function but this could in itself be a source of overhead as the structure ensures that the no duplicate condition has not been violated on every call. Given more time it would be interesting to compare performance differences between container implementations.

load_variants_from_file has another responsibility in keeping track of the "furthest along" variant position for each chromosome across all groups, this is used in the search step later, firstly to ensure that candidates are not constructed beyond where the last variant on a chromosome lies (as by definition they would be empty and searching over them would waste resources) and secondly to normalize the candidate regions across all groups to allow for easy comparison between them.

9.2.4 Storage and Recall of Variants

As I discussed previously, on the hunt for my Goldilocks genomic region, it is important to consider both time and memory as it is simple to deliver a poor performing solution.

Searching for candidates regions over the entire genome at once is probably unwise. Luckily, since our candidate must not span chromosomes (telomeres – the ends of chromosomes are not very good for reads) then we can easily yield a great improvement from processing chromosomes individually.

The process is to extract the locations of all the variants across the genome from the SNP chip VCF files (these files list the alleles for each detected variant for each sample), load them in to some sort of structure, “stride” over each chromosome (with some stride offset) and finally list the variants present between the stride start and stride start plus the desired length of the candidate. These are our regions.

Due to the use of striding, one does not simply walk the chromosome. A quick and dirty solution would be to just look up variants in a list:

```
for chromosome in autosomes:
    for start in range(1, len(chromosome), STRIDE):
        for position in range(start, start+LENGTH):
            if position in variant_position_list:
                # Do something...
```

This is of course rather dumb. Looking up a list has $\mathcal{O}$(n) performance where n is the number of variants (and there are a lot of variants). Combining this with the sheer number of lookups performed, with the default parameters the list would be queried half a billion times for the first chromosome alone.

You could improve this somewhat by dividing the variant_position_list in to a list of variants for each chromosome, so at least n is smaller. It is pretty doubtful that this would make any useful impact given the number of lookups. I'm happy to say I bypassed this option but thought it would be fun to consider its performance.

A far more sensible solution is to replace the list with a dictionary whose lookup is amortized to a constant time. The number of lookups is still incredibly substantial but a constant lookup is starting to make this sound viable. Using the variant positions as keys of a dictionary (in fact let us use the previous minor suggestion and have a dictionary for each chromosome) we have:

```
for chromosome in autosomes:
    for start in range(1, len(chromosome), STRIDE):
        for position in range(start, start+LENGTH):
            if position in variant_position_dict[chromosome]:
                # Do something...
```

This still takes half an hour to run on my laptop though, surely there is a better way. I wondered if whether dropping the lookup would improve things. Instead, how about an area of memory is allocated to house the current chromosome and act as a variant "mask" – essentially an array where each element is a base of the current chromosome and the value of 1 represents a variant at that position and a 0 represents no variant.

```
for chromosome in autosomes:
    chro = np.zeros(len(chromosome), np.int8)
    for variant_loc in snps_by_chromosome[chromosome]:
        chro[variant_loc] = 1

    for start in range(1, len(chromosome), STRIDE):
        for position in range(start, start+LENGTH):
            if chro[position] == 1:
                # Do something...
```

We of course have to initially load the variants in to the chromosome array, but this need only be done once (per chromosome) and is nothing compared to the billions of lookups of the previous implementation.

Rather to my surprise this was slow (maybe if I have time I should check how it compared to looking up with a list). Was it the allocation of memory? Perhaps I had run out of RAM and most of the work was going in to fetching and storing data in swap?

I switched out the comparison (== 1) step for the previous dictionary based lookup and the performance improved considerably. What was going on? There must be more to looking at a given element in the numpy

chromosome array, but what?

After a brief spell of playing with Python profilers and crashing my laptop by allocating considerably more memory than I had with numpy.zeroes, I read the manual (!) and discovered that numpy.zeroes returns an ndarray which uses "advanced indexing" even for single element access, effectively copying the element to a Python scalar for comparison.

That's a lot of memory action.

It then occurred to me that we're interested in just how many variants are inside each region and our chromosome is handily encapsulated in a numpy array. Why don't we just sum together the elements in each region? Remember variant positive base positions are 1 and 0 otherwise, useful. So the work boils down to some clever vector mathematics calculated by numpy.

We don't even lose any detail because the actual list of variants inside a region can be recovered in a small amount of time just given the start and end position of the region.

```
for chromosome in autosomes:
    chro = np.zeros(len(chromosome), np.int8)
    for variant_loc in snps_by_chromosome[chromosome]:
        chro[variant_loc] = 1

    for start in range(1, len(chromosome), STRIDE):
        num_variants = np.sum(chro[start:start+LENGTH])
        # Do something...
```

With conservative testing this runs at least 60 times faster than before, the entirety of the human genome can be analysed for candidate regions in less than twenty seconds (with the first few seconds taken reading in all the variant locations in the first place).

...Ignoring empty regions etc.

```
def load_chromosome(self, size, locations):
    chro = np.zeros(size+1, np.int8)

    # Populate the chromosome array with 1 for each position a variant exists
    for variant_loc in locations:
        chro[variant_loc] = 1

    return chro
```

load_chromosome returns a **numpy** array representing the chromosome, given its desired length and a list of variants...

...supervisor suggested bloom filters... ...would be a useful experiment to measure what further performance gains could be acheived with such a method...

9.2.5 Searching for Candidate Regions

search_regions is responsible for performing the main task of the script, walking over the genome and creating objects who represent candidate regions. The function is possibly a misnomer as it conducts a candidate census rather than a search, indiscriminately adding all valid candidates to the **regions** dictionary.

search_regions will iterate over each chromosome and load the listed variants in each group-chromosome pair from the **groups** structure (by calling **load_chromosome**) into a **NumPy** array whose length is set to the last position from which a full region can be extracted based on the maximum position seen by **load_variants_from_file** for this chromosome, plus one to account for the 1-indexed nature of genome positions. The function will then take slices of these arrays of the user defined *length* and count the number of variants contained within by summing across the slice (elements with variants contain a 1, otherwise 0). The next region is sliced by moving the first base of the current region forward by the user defined *stride*.

Each valid region will be entered in to the **regions** dictionary using the current region counter (**region_i**) as the key, mapping to an arbitrary object (a dictionary) as defined in Listing 8 below:

```
regions[region_i] = {
    "ichr": i,               # The i'th region on the current chromosome
    "group_counts": {},      # A dictionary mapping variant group names to
                             # the number of variants counted for that group
    "chr": chrno,            # The allosome number
    "pos_start": region_s,   # The first base of this region (inclusive)
    "pos_end": region_e      # The last base on this region (inclusive)
}
```

Listing 8 : Candidate Region Data Structure

Regions will also have their respective group densities appended to the appropriate **group_counts** list and its **region_i** added to the relevant **group_buckets** structure as established in Section 9.2.2.

9.2.6 Filtering and Ranking Candidates for Selection

For the purpose of this project we're looking to filter candidates by their absolute distance from the median density of variants located in the **GWAS** variant group (the median density will give the most representative regions) and then rank those by their corresponding **iCHIP** variant group densities, descending (enriching the number of comparisons that can be made between the **GWAS** data and **iCHIP** data).

Two functions in the **Goldilocks** API provide a process to rank candidate regions by a desired group's density:

- **initial_filter**
  Uses the relevant filtering group's **group_counts** list (in this case, **GWAS**) to calculate a window around the median density. The function will then return all candidates inside the group's **group_buckets** dictionary whose density fall inside the upper and lower bounds of the window around the median.

- **enrich**
  Sorts the output of **initial_filter** by the desired ranking variant group (in our case, **iCHIP**). The function will prioritise candidates with the closest absolute value to the filter group's median and ensure that the density of the ranking group is higher than the filtering group. Returns a ranked list of candidates in the format displayed in Listing 8 for manual inspection.

Currently ranking and filtering criteria are still hard coded to meet the needs of our analysis. In future it would be useful to generalise the **Goldilocks** API to allow more operations such as filtering by a given percentile rather than only the median and to provide more 'enrichment' options such as ranking by minimum.

It would be sensible to consider re-naming the two functions to **filter** and **rank** respectively and allow users to chain together these methods to compound filtering and ranking operations as necessary.

Whilst in this state the script proposes little re-use value, we at least have a basis for repeating results in future with different data sets or additional variant groups.

9.3 Testing

The testing of **Goldilocks** presented a challenge; we'd not only like to simply ensure that the correct query files are located and loaded in to memory but also that the expected candidates are found, filtered and selected. It is thus not practical to conduct testing on a large data set such as the two we have for analysis, we need a smaller, well defined set with known results.

To acheive this the testing suite is designed to generate valid **Variant Query Files** from a **TEST_DATA** dictionary, which defines test groups and variant locations on test chromosomes and can easily be altered for future testing. Test files are re-generated from this dictionary each time the test suite is instantiated using the **setUpClass** classmethod provided by the **unittest** library.

Checking both the **Paths File** and generated **Variant Query Files** were loaded successfully is trivial and the suite goes on to test whether the input files and all generated variants within are mapped to the specified chromosome of the correct group, both in memory and also when loaded in to the **NumPy** array structure. Once regions have been searched (or more accurately, once a a candidate census has populated the **regions** dictionary via **search_regions**), additional tests ensure that all candidate regions are of the correct *length* and begin on a valid *stride*.

```
GROUP0
CHR 1                                                          Expected
           4     10 13      21  25                              Size    i
      1 |===*======*==*========*===*========================*| 50
         |        |                                             3      0
              |        |                                       2      1
                   |        |                                  1      2
                        |        |                             2      3
                             |        |                        2      4
                                  |        |                   0      5
                                       |        |              0      6
                                            |        |         0      7
                                                 |        |    1      8
                                                      |        |    x

GROUP0
CHR 2                                                          Expected
           5                                    91                Size    i
      1 |====*=================\/...\/==========*********| 100
         |        |                                             1      9
              |        |                                       0      10
                                  ...                          .      .
                                            |        |         5      26
                                                 |        |    10     27
                                                      |        |    x
```

Listing 9 **Testing Expected Candidates with Goldilocks**: The test suite defines the locations of variants across many files in two groups, the figure represents the expected regions for *GROUP0*. The horizontal line represents a particular chromosome on which a * signals the location of a variant (= otherwise). The expected size column tabulates the number of variants that should appear as a member of region i (start-end inclusive).

As discussed, testing the discovery of correct regions is more complex. Listing 9 details the regions expected for the variants specified in **TEST_DATA**. The test suite checks that the number of regions located and their content is correct for each chromosome in each variant group. Tests also ensure that the previously mentioned **group_counts** and **group_buckets** structures are populated appropriately.

Finally the **initial_filter** and **enrich** functions can be tested (as we are able to manually calculate the expected candidates given the small data set) and ensure those candidates are returned.

Admittedly the suite could still be described as modest and with more time it would be appropriate to further improve coverage, possibly generating larger (or even random) data sets. Results however, have also undergone manual inspection both by myself and the Sanger Institute and the regions described in the following chapter are considered legitimate.

Chapter 10

Results

Listing 10 lists the top 25 candidate regions selected by **Goldilocks** for inspection. Regions were constructed using a *length* of 1Mnt[1](mega nucleotides) and a *stride* of 0.5Mnt[2]. As detailed in Section 9.2.6, candidates for this study were filtered to include regions with a **GWAS** density within ±12.5% of the median **GWAS** density (*i.e.* in the middle 25%) before being ranked firstly by absolute difference from the median **GWAS**, then by **iCHIP** variant density, descending.

Empty regions were excluded from a variant group's analysis, including the calculation of any percentiles. This was to ensure that we located the most "representative" density from the regions and avoid potential skewing from 0 density candidates. However repeating the experiment including empty regions leads to highly similar results and the candidate that was actually selected as the Goldilocks region appears in both sets of results.

[1] 1,000,000 bases
[2] 500,000 bases

```
i        GWAS       iCHIP       CHR       POSITIONS
0234       297         470       1        117000001 -  118000000
1074*       294        1540         3         46000001 -   47000000
5222       294         336       21          16500001 -   17500000
3125       298         310       10          60000001 -   61000000
0880       293         344       2         191500001 -  192500000
3560       299         772       12           9000001 -   10000000
4407       299         512       15          78500001 -   79500000
1036       292         300       3          27000001 -   28000000
2734       300         515       9           5000001 -    6000000
3426       300         486       11          76000001 -   77000000
0015       291        1029        1           7500001 -    8500000
0365       301         487       1         182500001 -  183500000
3415       301         419       11          70500001 -   71500000
1581       290         802       4         102500001 -  103500000
3554       290         403       12           6000001 -    7000000
3184       302         449       10          89500001 -   90500000
1580       289         603       4         102000001 -  103000000
1948       288        1297        5          96000001 -   97000000
2215       288         622       7          49500001 -   50500000
0414       288         346       1         207000001 -  208000000
2055       304        1377        5         149500001 -  150500000
0384       287         827       1         192000001 -  193000000
0959       306         406       2         231000001 -  232000000
4214       286         393       14          88500001 -   89500000
0320       307         620       1         160000001 -  161000000
```

Listing 10 **Goldilocks Results**: "Top 25" Candidate Regions using a *length* of 1Mnt and a *stride* of 0.5Mnt. Candidates are filtered by median **GWAS** density and ranked by maximum **iCHIP** density.

Candidate 1074 (marked by *) was selected as the region with the closest absolute difference between the median **GWAS** density of 297, whilst also offering the highest **iCHIP** density of the top 25 returned candidates.

The implementation of the filtering and ranking functions in the previous chapter noted that during final ranking, the **enrich** function will discard candidates if the number of variants in the ranking group is less than the filtering group – in this case, if the **iCHIP** density is lower than the **GWAS**. This was a crude rule implemented to prevent candidates with a significantly low **iCHIP** density ($\leq$100) from being put forward following 'enrichment'; these regions would give us little to compare when it comes to measuring the concordance between variants called in our leave-one-out study and the original **iCHIP** data set.

Curious about the wide range and presence of so many low resolution **iCHIP** candidates, I authored an R script (making use of **ggplot2**[25] for graphing) to plot figures 10.1 and 10.2 as a means of visual inspection to confirm such results should be anticipated.

Figure 10.1 shows a large range in the densities of variants in the **iCHIP** regions, between 0 and almost 3000, with a significant baseline below 100. Figure 10.2 plots the same data set but also facetting by chromosome, again it can be seen that the **iCHIP** variant group counts are well spread when compared to the rather constant variance of their **GWAS** counterparts.

This isn't a surprise when you consider that the **GWAS** data set represents the variants sequenced from whole-genome analysis of our samples, whereas the **iCHIP** set consists of the variants output from SNP genotyping, a process that focuses on particular genomic locations of interest.

Fig. 10.1 **Goldilocks Candidate Plot A**: Candidate regions of one megabase length are arranged along the x axis (coloured by chromosome) against the number of variants contained within that region.

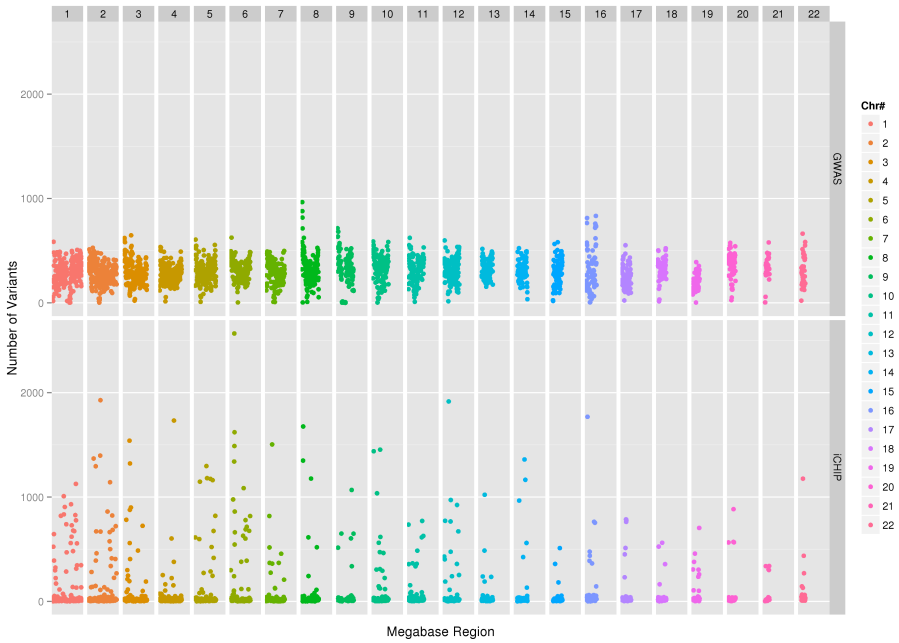

Fig. 10.2 **Goldilocks Candidate Plot B**: Candidates of one megabase length are arranged along the x axis (coloured and facetted by allosome) and plotted against the number of variants contained within that region.

Chapter 11

Analysis Pipeline

Having selected the Goldilocks region (Chromosome 3: 46,000,001—47,000,000) via **Goldilocks**, we must now begin constructing the pipeline that will conduct the leave-one-out analysis over all the **GWAS** lanelets and compare the called variants to the appropriate **iCHIP** sample.

This chapter introduces the various components of the pipeline, their purpose and any difficulties encountered.

11.1 Overview

The pipeline can be divided up in to the following analysis steps:

- **Extraction**
  For every whole-genome lanelet from the **GWAS** study, extract the Goldilocks region.
- **Indexing**
  Create indexes for the extracted Goldilocks regions, improving efficiency for tools which require random access to variants inside the region.
- **Merge**
  Merge the data from each extracted region in to one file to reduce file handling overhead in later analysis.
- **Pileup**
  Calculate genotype likelihoods based on the reads seen across all the extracted Goldilocks regions.
- **Call**
  Use the genotype likelihood scores to actually call the variants for each position of interest in each of the Goldilocks regions.
- **Compare**
  For each pair of **GWAS** and **iCHIP** samples, measure the concordance of called variants.

11.2 Concepts and Terminology

11.2.1 htslib

...high-throughput sequencing library...

11.2.2 samtools

11.2.3 bcftools

...spun off from the **samtools** repository... ...now using **htslib**...

11.2.4 SAM and BAM Files

11.2.5 "The Farm"

...Sanger Institute cluster... ..."the farm"... ...jobs submitted and managed through **LSF**

```
bsub -R"select[mem>4000] rusage[mem=4000]" -M4000 ...
#                  |                 |        | Raise maximum job memory to 4000mb
#                  |                 | Pre-reserve 4000mb for job before execution
#                  | Only run on a node with more than 4000mb memory
```

Listing 11 **LSF Resource Syntax**: **bsub** flags required to raise memory allocation for a job.

11.2.6 GWAS Data Set

...the **GWAS** is made up two studies... ...one of which used 2x, the other 4x ...due to time constraints we had to discard the smaller 4x study...

11.3 Region Extraction and Indexing

11.3.1 Extraction

Now the Goldilocks region has been found, it must be extracted for each **GWAS** sample in our study. The Sanger Institute currently stores the electronic samples in the **BAM** format at one of their in-house data centres, storage and retrieval is managed by **iRODS**[44]: the *Integrated Rule-Oriented Data System*.

iRODS is a BSD licensed, C++ based data management application, describing itself as "middleware for your critical data"[45]. The system was deployed at the institute in 2011 to cope with the data deluge brought on by constantly improving next-generation sequencing technologies. "[T]raditional data management methods [became] saturated" and other more advanced file systems did not provide the desired support for customised metadata such as project information or sequencing specific fields like sample and lanelet ID to be automatically added to files[46] which **iRODS** does support, along with improved data replication and access control policy management.

Once I had been added to the relevant access group, it was possible to extract the regions from **iRODS** via **samtools_irods**[47] which provides wrappers around both **samtools** and **iRODS** functionality to allow **samtools** commands to directly support interaction with samples stored in **iRODS**.

```
samtools_irods view -bh irods:\${file} 3:46000001-47000000
    > \$(basename \${file}).goldilocks.bam
```

Listing 12 **BAM Extraction**: Retrieve Goldilocks region for a particular sample ($file) from **iRODS**.

11.3.2 Indexing

An index file provides a mapping between genomic positions and the compressed blocks of a **BAM** file that contain them, affording programs such as members of the **samtools** collection efficient access to the relevant blocks of a **BAM** file without conducting costly searches.

Whilst indexing of **BAM** files is a somewhat optional step, many tools will emit warnings or errors if one is not available. The process itself requires little resources in terms of time and memory and can significantly improve the efficiency of later commands which are capable of reading and using the index.

Both the commands in Listing 12 and 13 can be executed in parallel for extracting and indexing many samples.

```
samtools index \$(basename \${file}).goldilocks.bam
```

Listing 13 **BAM Indexing**: Creation of a **BAM Index** (BAI) for a sample ($file). with **samtools index**[48]

11.4 Pileup

Following extraction and indexing of the Goldilocks region for each whole-genome sample from the **GWAS** study, the next step is to "pileup" the samples, a process which stacks the bases seen at each variant location from all the indexed samples up in a structure before calculating the statistical likelihood of those bases actually appearing in each sample given the data overall.

```
bsub -o ~/goldilocks/joblog/samtools_mpileup.%J.o
    -e ~/goldilocks/joblog/samtools_mpileup.%J.e
    -G hgi -J "samtools_mpileup"
    -M1000 -R "select[mem>1000] rusage[mem=1000]"
    bash -c 'samtools mpileup -b ../goldilocks-3:46000001-47000000.fofn -g -I -f
        /lustre/scratch113/resources/ref/Homo_sapiens/1000Genomes_hs37d5/hs37d5.fa
            > ~/goldilocks/all.withref.bcf'
```

Listing 14 **Pileup**: Submission of **samtools mpileup**[49] task to **LSF** for execution on the **Farm**. Note the request to increase the memory allocation and how the pileup command is passed as a string to `bash -c`.

However, performing a pileup is a rather intensive process, with initial test runs requiring 6.5 hours of CPU time, with an average memory usage of 905.85MB (920MB max), outputting a 13GB **BCF** file. In a scenario such as ours where there are thousands of input files, one for each of the extracted regions, we place considerable strain on the cluster's file system and introduce incredibly inefficiency given the overhead of file handling. As this pipeline will need to be executed once for each leave-one-out experiment, it would be required to reduce the strain on the file server before the job could be submitted. Usefully another member of the **samtools** collection provides functionality to merge a list of given input files such as those extracted Goldilocks regions.

11.5 Merge

samtools merge[48] is a utility for merging two or more sorted **BAM** files in to one sorted output. After viewing the source of the command I found an undocumented feature that allowed an end user to provide a file of filenames to merge rather than being required to list all inputs on the command line.

Various issues with use of **samtools merge** at scale were encountered during the assembly of the analysis pipeline. Ultimately I ran out of time attempting to overcome these problems before being able to put the pipeline to use. The difficulties experienced with **samtools merge** specifically are detailed in Appendix **??**.

11.6 Calling

Variant calling was briefly introduced as a concept in Part I (See Section 2.2.3). Calling for variants is the process of finding differences between a reference sequence[1] and a given sample[2] or set of different samples.

The genotype likelihoods calculated by **samtools mpileup** in the previous step of the pipeline represent a probability distribution of the bases seen across all piledup samples at a particular location (site). This allows

[1] Our analysis maps to the **hs37d5** reference sequence, used by the *1000 Genomes Project Phase II*[50]. **hs37d5** extends the **b37** genomic reference with "decoy sequences" that are designed to reduce false positives (incorrectly mapping to poorly referenced areas around centromeres for example[51]), optimising it for variant calling[52].

[2] Technically to be confident you have an actual SNP as opposed to a variant (or incorrect read!), you'll need to see whether such variation occurs through a population of different samples rather than just one.

variant calling algorithms to make an informed decision as to whether particular sites actually contain a SNP or if a differing base is statistically unlikely to be a real variant[3] with the current data.

bcftools call[53][54] is the successor to the 'consensus' variant caller included with the **samtools** package and utilises **htslib** for improved file handling efficiency with common genomic data formats.

Unfortunately during the initial testing of this step, using a pileup of all the Goldilocks regions, the output contained only the standard header information generated by **bcftools** and not a single record detailing a called variant. I discovered this was due to the **samtools mpileup** job being completed without including a reference sequence, causing the *REF* (reference) column of the resulting **BCF** file to be populated with 'N': an ambiguity code which translates to 'any base'[55][56].

bcftools call does provide the `-M` flag for this scenario, where the reference is described as a "masked reference". However attempting to repeat calling with the flag set merely caused the command to produce a segmentation fault instead of an empty output file. This appeared to be due to a compatibility issue between the versions of **htslib** and **bcftools** I was using on the **Farm**.

Finally this initial test of the calling component of our pipeline was able to be completed in 66 minutes with an average memory usage of 11.97MB, resulting in a 114MB compressed VCF file.

A similar compressed VCF file will be output from each iteration of the leave-one-out analysis, generating thousands of files. Whilst the Sanger Institute manages petabytes of storage, it is sensible to minimise memory and storage resource use if at all possible. In future runs, output size could be reduced by generating a list of variants that appear in both the **GWAS** and **iCHIP** Goldilocks regions (as these are the only variants we are interested in anyway), **bcftools call** will then discard records pertaining to variants that do not appear in the targets list. This "targets" file actually uses the same format as the **Variant Query Files** described in Section 8.1.2 and should be provided to **bcftools call**'s `-T` (`--targets-file`) option.

11.7 Measuring Concordance

The terminal step of the Goldilocks pipeline consists of analysing the differences between variants called from the left-one-out **GWAS** data and the **iCHIP** study. The two data sets will form matched pairs where an individual's DNA has been sequenced and processed by our pipeline (possibly with one of that sample's lanelets ignored) and alternatively via SNP genotyping. The pipeline will measure the concordance between the variants called for each individual, this measure will then be used to determine the effect of the lanelet that was ignored on that run.

For this step I retrofitted a Python script I had authored earlier to sanity check the bases seen at the known variant sites were sensible (*i.e.* they agreed on the same genotype – the same sets of bases appeared). Due to the difficulties encountered in assembling the pipeline (particularly with **samtools merge**), it was not possible to adequately test the pipeline and thus the concordance script with the time that we had. Considering this and for the sake of brevity, discussion on the design and implementation is considered outside the scope of this write-up.

[3] Although it could also be a very low frequency variant...

Another option for this component was **SnpSift Concordance**[57], a Java based package which can also calculate concordances between two **VCF** files. However I was concerned with the possible performance of using Java to process thousands of concordances.

Part III

Discussions and Conclusions

Chapter 12

Current Status

Chapter 13

Critical Evaluation

...Section 3.2.3 The **SciPy Stack** also includes the "Python Data Analysis Library" (**pandas**), which is designed to provide tools to supplement the Python environment, allowing users to perform analysis in Python instead of having to switch to a more analysis focused language such as R.

...at first glance appears to be a far more developed Frontier... ...however it should be considered that Frontier was designed to complement the machine learning experiments and provides

...whilst also providing the **AbstractReader** framework to allow users to quickly define methods to read in their own data regardless of how esoteric or cryptic the file format is...

Whilst a 2003 paper[58][1] that analysed the spread of SNPs across the human genome to assert whether variants could be modelled with a statistical distribution, divided chromosomes in to equally sized bins (a concept akin to candidates of uniform length)

[1] Titled *An Exponential Dispersion Model for the Distribution of Human Single Nucleotide Polymorphisms*[58]

Chapter 14

Conclusions

References

[1] T. Strachan and A. Read, *Human Molecular Genetics*, 4th ed. Garland Science, 2011, pp. 214–254.
A concise introduction to the processes involved in massively parallel DNA sequencing.

[2] M. Kircher, U. Stenzel and J. Kelso, "Improved base calling for the Illumina Genome Analyzer using machine learning strategies," *Genome Biology*, vol. 10, no. 8, p. R83, 2009.
Useful introduction to relevant Illumina hardware and the errors that can occur during sequencing.

[3] auto_qc: Additional high-throughput sequencing autoQC steps [Github]. [Online]. Available: https://github.com/wtsi-hgi/seq_autoqc

[4] vr-pipe: A generic pipeline system [Github]. [Online]. Available: https://github.com/wtsi-hgi/vr-pipe/

[5] M. Niemi, "Internal QC Summary Report," 09 2012, unpublished.

[6] Triage: A todo manager for the disaster that is your day [Github]. [Online]. Available: https://github.com/SamStudio8/triage

[7] sam/Dissertation | Triage. [Online]. Available: https://triage.ironowl.io/sam/dissertation/

[8] Maggie Bartlett. (2011) Illumina HiSeq Flow Cell. National Human Genome Research Institute. Image. [Online]. Available: http://www.genome.gov/dmd/img.cfm?node=Photos/Technology/Genome%20analysis%20technology&id=80102

[9] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and G. Subgroup, " The Sequence Alignment/Map format and SAMtools. ," *Bioinformatics*, vol. 25, p. 2078, 2009.

[10] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA Data Mining Software: An Update," *SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, Nov. 2009. [Online]. Available: http://doi.acm.org/10.1145/1656274.1656278

[11] B. Ripley, *tree: Classification and regression trees*, 2014, r package version 1.0-35. [Online]. Available: http://CRAN.R-project.org/package=tree

[12] T. Therneau, B. Atkinson, and B. Ripley, *rpart: Recursive Partitioning*, 2013, r package version 4.1-3. [Online]. Available: http://CRAN.R-project.org/package=rpart

[13] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2013. [Online]. Available: http://www.R-project.org/

[14] T. Smith. (2012–2013) aRrgh: a newcomer's (angry) guide to R. [Online]. Available: http://tim-smith.us/arrgh/

[15] D. E. King, "Dlib-ml: A Machine Learning Toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[16] C. Igel, V. Heidrich-Meisner, and T. Glasmachers, "Shark," *Journal of Machine Learning Research*, vol. 9, pp. 993–996, 2008.

[17] Python Wiki. (2014) Python for Artificial Intelligence.

[18] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
A machine learning framework for Python.

[19] J. Demšar, T. Curk, A. Erjavec, Črt Gorup, T. Hočevar, M. Milutinovič, M. Možina, M. Polajnar, M. Toplak, A. Starič, M. Štajdohar, L. Umek, L. Žagar, J. Žbontar, M. Žitnik, and B. Zupan, "Orange: Data Mining Toolbox in Python," *Journal of Machine Learning Research*, vol. 14, pp. 2349–2353, 2013. [Online]. Available: http://jmlr.org/papers/v14/demsar13a.html

[20] scikit-learn. About Us. [Online]. Available: http://scikit-learn.org/stable/about.html

[21] T. E. Oliphant, "Python for Scientific Computing," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 10–20, 2007.

[22] E. Jones, T. Oliphant, P. Peterson, *et al.*, "SciPy: Open source scientific tools for Python," 2001–. [Online]. Available: http://www.scipy.org/

[23] SciPy.org. Scipy license. [Online]. Available: http://scipy.org/scipylib/license.html

[24] NumPy.org. Numpy license. [Online]. Available: http://www.numpy.org/license.html

[25] H. Wickham, *ggplot2: elegant graphics for data analysis*. Springer New York, 2009. [Online]. Available: http://had.co.nz/ggplot2/book 978-0-387-98140-6.

[26] S. Bethard. argparse 1.2.1: Python command-line parsing library. [Online]. Available: https://pypi.python.org/pypi/argparse/1.2.1

[27] M. Fowler, "Continuous Integration," 2006. [Online]. Available: http://www.martinfowler.com/articles/continuousIntegration.html

[28] Kohsuke Kawaguchi. Meet jenkins. [Online]. Available: https://wiki.jenkins-ci.org/display/JENKINS/Meet+Jenkins

[29] Jenkins Usage Statistics. [Online]. Available: http://stats.jenkins-ci.org/jenkins-stats/svg/svgs.html

[30] Getting Started: Travis CI Overview. [Online]. Available: http://docs.travis-ci.com/user/getting-started/

[31] Introduction to wercker. [Online]. Available: http://devcenter.wercker.com/articles/introduction/

[32] Hadley Wickham and Winston Chang, *devtools: Tools to make developing R code easier*, 2013, r package version 1.4.1. [Online]. Available: http://CRAN.R-project.org/package=devtools

[33] Morandat, Floréal and Hill, Brandon and Osvald, Leo and Vitek, Jan, "Evaluating the design of the R language," in *ECOOP 2012–Object-Oriented Programming*. Springer, 2012, pp. 104–131.

[34] Python Software Foundation. More on Lists. [Online]. Available: https://docs.python.org/2/tutorial/datastructures.html#more-on-lists

[35] ——. Dictionaries. [Online]. Available: https://docs.python.org/2/tutorial/datastructures.html#dictionaries

[36] Python Wiki. (2012) Time Complexity. [Online]. Available: https://wiki.python.org/moin/TimeComplexity

[37] Python Software Foundation. Sets. [Online]. Available: https://docs.python.org/2/tutorial/datastructures.html#sets

[38] W. B. Van Rossum, Guido and N. Coghlan, "PEP 8 – Style Guide for Python Code," *Python Enhancement Proposals*, 2001. [Online]. Available: http://legacy.python.org/dev/peps/pep-0008/

[39] Kirk Strauser (http://stackoverflow.com/users/32538/kirk-strauser), "Does python have 'private' variables in classes?" [Online]. Available: http://stackoverflow.com/questions/1641219/does-python-have-private-variables-in-classes

[40] Python Software Foundation. unittest — Unit testing framework. [Online]. Available: https://docs.python.org/2/library/unittest.html

[41] Orbis Play and Learn, Albert Barillé, "How My Body Works," 1992–1994.

[42] Procidis and Albert Barillé, "Once upon a time... Life," 1987, [Television broadcast].

[43] U.S. National Library of Medicine. (2009) HLA Gene Family. [Online]. Available: http://ghr.nlm.nih.gov/geneFamily/hla

[44] iRODS Consortium. iRODS: About. [Online]. Available: http://irods.org/about/

[45] irods: Middleware for your critical data [Github].

[46] G.-T. Chiang, P. Clapham, G. Qi, K. Sale, and G. Coates, "Implementing a genomic data management system using iRODS in the Wellcome Trust Sanger Institute," *BMC Bioinformatics*, vol. 12, no. 1, p. 361, 2011. [Online]. Available: http://www.biomedcentral.com/1471-2105/12/361

[47] K. Lewis. samtools_irods [Github]. WTSI. [Online]. Available: https://github.com/wtsi-npg/samtools_irods

[48] *Manual Reference Pages - samtools (1): Samtools Commands And Options*, SAM tools Project. [Online]. Available: http://samtools.sourceforge.net/samtools.shtml#3

[49] Calling SNPs/INDELs with SAMtools/BCFtools. SAM tools Project. [Online]. Available: http://samtools.sourceforge.net/mpileup.shtml

[50] 1000 Genomes. 1000 Genomes FAQ: Which reference assembly do you use? [Online]. Available: http://www.1000genomes.org/faq/which-reference-assembly-do-you-use

[51] zam.iqbal.genome (https://www.biostars.org/u/5152/). (2013) Decoy In Reference Assembly. [Online]. Available: https://www.biostars.org/p/73100/#73120

[52] Heng Li. (2011) The missing human sequences (version 5). [Online]. Available: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/phase2_reference_assembly_sequence/hs37d5.slides.pdf

[53] *The bcftools/htslib VCF commands: bcftools call*, VCFTools Project. [Online]. Available: http://vcftools.sourceforge.net/htslib.html#call

[54] *bcftools call*, Genome Research Ltd. [Online]. Available: http://samtools.github.io/bcftools/bcftools.html#call

[55] A. Cornish-Bowden, "Nomenclature for Incompletely Specified Bases in Nucleic Acid Sequences: Recommendations 1984." *Nucleic Acids Research*, vol. 13, no. 9, p. 3021, 1985.

[56] C. Liébecq, *Biochemical Nomenclature and Related Documents*. Portland Press, 1992.

[57] Pablo Cingolani, *SnpSift Concordance*, 2013. [Online]. Available: http://snpeff.sourceforge.net/SnpSift.html#concordance

[58] W. S. Kendal, "An Exponential Dispersion Model for the Distribution of Human Single Nucleotide Polymorphisms," *Molecular Biology and Evolution*, vol. 20, no. 4, pp. 579–590, 2003.

Appendix A

Input Examples

A.1 BAMcheckR'd Example Output

```
# Summary Numbers. Use `grep ^SN | cut -f 2-` to extract this part.
SN	raw total sequences:	41400090
SN	filtered sequences:	0
SN	sequences:	41400090
SN	is paired:	1
SN	is sorted:	1
SN	1st fragments:	20700045
SN	last fragments:	20700045
SN	reads mapped:	41291484
SN	reads unmapped:	108606
SN	reads unpaired:	60000
SN	reads paired:	41231484
SN	reads duplicated:	5756822
SN	reads MQ0:	1038644
SN	reads QC failed:	0
SN	non-primary alignments:	0
SN	total length:	3105006750
SN	bases mapped:	3096861300
SN	bases mapped (cigar):	3090885143
SN	bases trimmed:	0
SN	bases duplicated:	431761650
SN	mismatches:	9107833
SN	error rate:	0.002946675
SN	average length:	75
SN	maximum length:	75
SN	average quality:	36
SN	insert size average:	178.7
SN	insert size standard deviation:	44.1
SN	inward oriented pairs:	20577242
SN	outward oriented pairs:	3140
SN	pairs with other orientation:	3711
SN	pairs on different chromosomes:	31535
SN	fwd percent insertions above baseline:	1.43135383851191
SN	fwd percent insertions below baseline:	0.686265539012562
SN	fwd percent deletions above baseline:	1.38326380878871
SN	fwd percent deletions below baseline:	0.44923551909251
SN	rev percent insertions above baseline:	1.08264446659241
SN	rev percent insertions below baseline:	0.457290262062496
```

```
SN	rev percent deletions above baseline:	1.15931214598243
SN	rev percent deletions below baseline:	0.413119424753248
SN	contiguous cycle dropoff count:	36
SN	fwd.percent.insertions.above.baseline:	1.43135383851191
SN	fwd.percent.insertions.below.baseline:	0.686265539012562
SN	fwd.percent.deletions.above.baseline:	1.38326380878871
SN	fwd.percent.deletions.below.baseline:	0.44923551909251
SN	rev.percent.insertions.above.baseline:	1.08264446659241
SN	rev.percent.insertions.below.baseline:	0.457290262062496
SN	rev.percent.deletions.above.baseline:	1.15931214598243
SN	rev.percent.deletions.below.baseline:	0.413119424753248
SN	quality.dropoff.fwd.high.iqr.start.read.cycle:	0
SN	quality.dropoff.fwd.high.iqr.end.read.cycle:	0
SN	quality.dropoff.fwd.high.iqr.max.contiguous.read.cycles:	0
SN	quality.dropoff.fwd.mean.runmed.decline.start.read.cycle:	20
SN	quality.dropoff.fwd.mean.runmed.decline.end.read.cycle:	51
SN	quality.dropoff.fwd.mean.runmed.decline.max.contiguous.read.cycles:	32
SN	quality.dropoff.fwd.mean.runmed.decline.high.value:	36.9775883578997
SN	quality.dropoff.fwd.mean.runmed.decline.low.value:	36.301749247405
SN	quality.dropoff.rev.high.iqr.start.read.cycle:	0
SN	quality.dropoff.rev.high.iqr.end.read.cycle:	0
SN	quality.dropoff.rev.high.iqr.max.contiguous.read.cycles:	0
SN	quality.dropoff.rev.mean.runmed.decline.start.read.cycle:	18
SN	quality.dropoff.rev.mean.runmed.decline.end.read.cycle:	56
SN	quality.dropoff.rev.mean.runmed.decline.max.contiguous.read.cycles:	39
SN	quality.dropoff.rev.mean.runmed.decline.high.value:	36.1517621338504
SN	quality.dropoff.rev.mean.runmed.decline.low.value:	35.3152133727245
SN	quality.dropoff.high.iqr.threshold:	10
SN	quality.dropoff.runmed.k:	25
SN	quality.dropoff.ignore.edge.cycles:	3
SN	A.percent.mean.above.baseline:	0.0991164444444441
SN	C.percent.mean.above.baseline:	0.127379555555556
SN	G.percent.mean.above.baseline:	0.0603679999999997
SN	T.percent.mean.above.baseline:	0.0868000000000005
SN	A.percent.mean.below.baseline:	0.0991164444444451
SN	C.percent.mean.below.baseline:	0.127379555555555
SN	G.percent.mean.below.baseline:	0.0603680000000002
SN	T.percent.mean.below.baseline:	0.0867999999999993
SN	A.percent.max.above.baseline:	0.601733333333332
SN	C.percent.max.above.baseline:	0.394266666666667
SN	G.percent.max.above.baseline:	0.2956
SN	T.percent.max.above.baseline:	0.768000000000001
SN	A.percent.max.below.baseline:	0.318266666666666
SN	C.percent.max.below.baseline:	0.825733333333332
SN	G.percent.max.below.baseline:	0.554400000000001
SN	T.percent.max.below.baseline:	0.251999999999999
SN	A.percent.max.baseline.deviation:	0.601733333333332
SN	C.percent.max.baseline.deviation:	0.825733333333332
SN	G.percent.max.baseline.deviation:	0.554400000000001
SN	T.percent.max.baseline.deviation:	0.768000000000001
SN	A.percent.total.mean.baseline.deviation:	0.198232888888889
SN	C.percent.total.mean.baseline.deviation:	0.254759111111111
SN	G.percent.total.mean.baseline.deviation:	0.120736
SN	T.percent.total.mean.baseline.deviation:	0.1736
# First Fragment Qualitites. Use 'grep ^FFQ | cut -f 2-' to extract this part.
# Columns correspond to qualities and rows to cycles. First column is the cycle number.
FFQ	1	8968	3619	9863	747	5094	0	6642	1609	4673	...
FFQ	2	21676	0	0	0	0	0	0	0	43	1885	...
FFQ	3	7	0	177	0	0	0	0	0	0	0	...
[...]
FFQ	74	3697	39	0	919	4933	0	0	56866	1524	...
FFQ	75	4542	0	0	0	0	0	4634	77822	0	...
FFQ	76	0	0	0	0	0	0	0	0	0	...
```

```
# Last Fragment Qualitites. Use ‘grep ^LFQ | cut -f 2-‘ to extract this part.
# Columns correspond to qualities and rows to cycles. First column is the cycle number.
LFQ	1	8869	0	0	0	0	0	63	0	0	1156	...
LFQ	2	3300	0	0	0	0	0	0	0	0	0	...
LFQ	3	6816	0	0	0	573	0	83	0	7011	...
[...]
LFQ	74	5980	3	91	0	0	0	1340	9696	72939	...
LFQ	75	4314	0	0	168	0	848	8591	0	70358	...
LFQ	76	0	0	0	0	0	0	0	0	0	0	...
# Mismatches per cycle and quality. Use ‘grep ^MPC | cut -f 2-‘ to extract this part.
# Columns correspond to qualities, rows to cycles. First column is the cycle number, second
# is the number of N’s and the rest is the number of mismatches
MPC	1	14078	0	2594	6777	416	1919	0	2222	...
MPC	2	21407	0	0	0	0	0	0	0	5	...
MPC	3	3205	0	0	37	0	43	0	12	0	...
[...]
MPC	74	779	0	3	0	131	440	0	93	7485	...
MPC	75	136	0	0	0	3	0	47	704	9302	...
MPC	76	0	0	0	0	0	0	0	0	0	...
# GC Content of first fragments. Use ‘grep ^GCF | cut -f 2-‘ to extract this part.
GCF	0.5	56
GCF	1.76	60
GCF	3.02	126
GCF	4.27	212
GCF	5.78	347
[...]
GCF	93.72	378
GCF	95.23	186
GCF	96.48	87
GCF	97.74	55
GCF	99.25	17
# GC Content of last fragments. Use ‘grep ^GCL | cut -f 2-‘ to extract this part.
GCL	0.5	118
GCL	1.76	175
GCL	3.02	230
GCL	4.27	354
GCL	5.78	525
[...]
GCL	93.72	613
GCL	95.23	430
GCL	96.48	274
GCL	97.74	185
GCL	99.25	110
# ACGT content per cycle. Use ‘grep ^GCC | cut -f 2-‘ to extract this part. The columns are: cycle, and A,C,G,T counts [%]
GCC	1	26.93	23.09	22.77	27.2
GCC	2	26.78	23.24	22.97	27.02
GCC	3	26.46	23.59	23.3	26.66
GCC	4	26.29	23.79	23.45	26.46
GCC	5	26.47	23.61	23.3	26.62
[...]
GCC	70	26.09	24.26	23.45	26.2
GCC	71	26.07	24.25	23.46	26.22
GCC	72	26.04	24.27	23.49	26.2
GCC	73	26.07	24.25	23.47	26.22
GCC	74	26.08	24.24	23.45	26.23
GCC	75	26.01	24.31	23.51	26.18
# Insert sizes. Use ‘grep ^IS | cut -f 2-‘ to extract this part.
# The columns are: pairs total, inward oriented pairs, outward oriented pairs, other pairs
IS	0	10	0	1	9
IS	1	3	0	3	0
IS	2	4	0	4	0
IS	3	5	0	5	0
IS	4	2	0	2	0
```

```
IS	5	3	0	3	0
[...]
IS	110	33952	33952	0	0
IS	111	38433	38433	0	0
IS	112	43373	43370	0	3
IS	113	48160	48159	0	1
IS	114	53175	53171	0	4
IS	115	59504	59502	0	2
IS	116	64668	64668	0	0
IS	117	71107	71105	0	2
IS	118	77157	77156	0	1
IS	119	84044	84044	0	0
IS	120	90116	90110	3	3
[...]
IS	327	6546	6546	0	0
IS	328	6483	6483	0	0
IS	329	6201	6201	0	0
IS	330	6228	6228	0	0
IS	331	5852	5852	0	0
# Read lengths. Use 'grep ^RL | cut -f 2-' to extract this part. The columns are: read length, count
RL	75	41400090
# Indel distribution. Use 'grep ^ID | cut -f 2-' to extract this part.
# The columns are: length, number of insertions, number of deletions
ID	1	128650	183418
ID	2	26409	39770
ID	3	10213	16046
ID	4	7756	11444
ID	5	1746	3455
[...]
ID	35	0	8
ID	36	0	1
ID	37	0	1
ID	38	0	1
ID	40	0	2
# Indels per cycle. Use 'grep ^IC | cut -f 2-' to extract this part.
# The columns are: cycle, number of insertions (fwd), .. (rev) , number of deletions (fwd), .. (rev)
IC	1	0	0	105	97
IC	2	24	15	150	179
IC	3	129	138	441	509
IC	4	253	310	623	829
IC	5	557	724	786	1164
[...]
IC	70	571	710	638	761
IC	71	350	428	309	434
IC	72	154	150	38	45
IC	73	60	61	15	23
IC	74	20	19	11	12
# Coverage distribution. Use 'grep ^COV | cut -f 2-' to extract this part.
COV	[1-1]	1	332980694
COV	[2-2]	2	105004580
COV	[3-3]	3	29112182
COV	[4-4]	4	13415014
COV	[5-5]	5	6716815
[...]
COV	[996-996]	996	2
COV	[997-997]	997	2
COV	[998-998]	998	2
COV	[1000-1000]	1000	4
COV	[1000<]	1000	116
# GC-depth. Use 'grep ^GCD | cut -f 2-' to extract this part.
# The columns are: GC%, unique sequence percentiles, 10th, 25th, 50th, 75th and 90th depth percentile
GCD	0	0.001	0	0	0	0	0
GCD	0.4	0.002	0.101	0.101	0.101	0.101	0.101
```

```
GCD     19      0.003     0.049     0.049     0.049     0.049     0.049
GCD     20      0.004     0.06      0.06      0.06      0.06      0.06
GCD     21      0.004     0.045     0.045     0.045     0.045     0.045
[...]
GCD     66      99.99     0.244     2.693     6.746     11.794    15.885
GCD     67      99.994    1.279     1.279     4.305     9.667     11.483
GCD     68      99.997    4.148     4.148     4.463     5.741     7.354
GCD     69      99.999    0.499     0.499     0.499     1.935     1.935
GCD     72      100       0.476     0.476     0.476     1.219     1.219
```

A.2 auto_qc Decision Matrix

```
lanelet     sample      study     npg      aqc      ...
9999_9#1    AQCTest000000     AQCTest     pass     passed     ...
9999_9#3    AQCTest000002     AQCTest     fail     failed     ...
9999_9#2    AQCTest000001     AQCTest     pass     pass     ...
9999_9#5    AQCTest000004     AQCTest     warn     warn     ...
9999_9#4    AQCTest000003     AQCTest     warn     warning     ...
9999_9#7    AQCTest000006     AQCTest     fail     fail     ...
9999_9#6    AQCTest000005     AQCTest     pass     passed     ...
9999_9#9    AQCTest000008     AQCTest     warn     warning     ...
9999_9#8    AQCTest000007     AQCTest     pass     passed     ...
```

A.3 Goldilocks Paths File

```
# gwas
cd-gwas     /pools/encrypted/sanger/vcf/cd-seq.vcf.gz.q
uc-gwas     /pools/encrypted/sanger/vcf/uc-seq.vcf.gz.q
# ichip
cd-ichip     /pools/encrypted/sanger/vcf/cd.ichip.vcf.gz.q
uc-ichip     /pools/encrypted/sanger/vcf/uc.ichip.vcf.gz.q
```

Appendix B

Output Examples

B.1 Frontier Log Example

```
Frontier
********
Data Dir        /pools/encypted/sanger/frontier/data/bamcheck_2013dec25/uc
AQC File        /pools/encypted/sanger/frontier/data/crohns-uc-table-a.2013dec25.manual_qc_update.txt

Class Def        fail        pass        warn
Class Read        1542        9154        2759
Class Used        234        3137        576

Total Read        3947
Total Used        3371

Param Set        ALL
Param Count        82
Param List:
                1st-fragments
                A-percent-max-above-baseline
                A-percent-max-baseline-deviation
                A-percent-max-below-baseline
                A-percent-mean-above-baseline
                A-percent-mean-below-baseline
                A-percent-total-mean-baseline-deviation
                C-percent-max-above-baseline
                C-percent-max-baseline-deviation
                C-percent-max-below-baseline
                C-percent-mean-above-baseline
                C-percent-mean-below-baseline
                C-percent-total-mean-baseline-deviation
                G-percent-max-above-baseline
                G-percent-max-baseline-deviation
                G-percent-max-below-baseline
                G-percent-mean-above-baseline
                G-percent-mean-below-baseline
                G-percent-total-mean-baseline-deviation
                T-percent-max-above-baseline
                T-percent-max-baseline-deviation
                T-percent-max-below-baseline
                T-percent-mean-above-baseline
```

```
                T-percent-mean-below-baseline
                T-percent-total-mean-baseline-deviation
                average-length
                average-quality
                bases-duplicated
                bases-mapped
                bases-mapped-(cigar)
                bases-trimmed
                error-rate
                filtered-sequences
                fwd-percent-deletions-above-baseline
                fwd-percent-deletions-below-baseline
                fwd-percent-insertions-above-baseline
                fwd-percent-insertions-below-baseline
                insert-size-average
                insert-size-standard-deviation
                inward-oriented-pairs
                is-paired
                is-sorted
                last-fragments
                maximum-length
                mismatches
                non-primary-alignments
                outward-oriented-pairs
                pairs-on-different-chromosomes
                pairs-with-other-orientation
                quality-dropoff-fwd-high-iqr-end-read-cycle
                quality-dropoff-fwd-high-iqr-max-contiguous-read-cycles
                quality-dropoff-fwd-high-iqr-start-read-cycle
                quality-dropoff-fwd-mean-runmed-decline-end-read-cycle
                quality-dropoff-fwd-mean-runmed-decline-high-value
                quality-dropoff-fwd-mean-runmed-decline-low-value
                quality-dropoff-fwd-mean-runmed-decline-max-contiguous-read-cycles
                quality-dropoff-fwd-mean-runmed-decline-start-read-cycle
                quality-dropoff-high-iqr-threshold
                quality-dropoff-ignore-edge-cycles
                quality-dropoff-rev-high-iqr-end-read-cycle
                quality-dropoff-rev-high-iqr-max-contiguous-read-cycles
                quality-dropoff-rev-high-iqr-start-read-cycle
                quality-dropoff-rev-mean-runmed-decline-end-read-cycle
                quality-dropoff-rev-mean-runmed-decline-high-value
                quality-dropoff-rev-mean-runmed-decline-low-value
                quality-dropoff-rev-mean-runmed-decline-max-contiguous-read-cycles
                quality-dropoff-rev-mean-runmed-decline-start-read-cycle
                quality-dropoff-runmed-k
                raw-total-sequences
                reads-MQ0
                reads-QC-failed
                reads-duplicated
                reads-mapped
                reads-paired
                reads-unmapped
                reads-unpaired
                rev-percent-deletions-above-baseline
                rev-percent-deletions-below-baseline
                rev-percent-insertions-above-baseline
                rev-percent-insertions-below-baseline
                sequences
                total-length

Feature Importances:
                0.32        T-percent-mean-above-baseline
                0.20        T-percent-mean-below-baseline
```

```
               0.14        rev-percent-insertions-above-baseline
               0.14        T-percent-total-mean-baseline-deviation
               0.05        bases-duplicated
               0.05        fwd-percent-insertions-above-baseline
               0.03        reads-duplicated
               0.02        insert-size-average
               0.01        insert-size-standard-deviation

CV Score (Fld)        1.00 +/- 0.01 (10)

Decision PDF        pdf/2014-03-23_2151__IGNWARN_ALL_/
********
```

Appendix C

Tool Installation

C.1 htslib

```
# Download repositories from Github
git clone htslib

cd htslib/
make
```

C.2 bcftools

Prerequisites: **htslib**(Appendix C.1)

```
# Download repositories from Github
git clone bcftools

cd bcftools/

# Build bcftools
make #(requires htslib to be above samtools/bcftools dir)
sudo cp bcftools usr/local/bin
```

C.3 samtools

Prerequisites: **htslib**(Appendix C.1)

```
# Download repositories from Github
git clone samtools

cd samtools/

# Build samtools
make #(requires htslib to be above samtools/samtools dir)
sudo cp samtools usr/local/bin
```

C.4 tabix

```
# Tabix
# Download from sourceforge
# http://sourceforge.net/projects/samtools/files/tabix/
make
sudo cp tabix /usr/local/bin

# Generate an index (vcfidx) (not necessary)
vcftools --gzvcf cd.ichip.vcf.gz

# Index with tabix
# "The input data file must be position sorted and compressed by bgzip
# which has a gzip(1) like interface"
tabix -p vcf file.vcf.gz
diff cd-seq.vcf.gz.tbi diff cd-seq.vcf.gz.tbi.sanger
# Shows no difference :)
# http://samtools.sourceforge.net/tabix.shtml
```

Appendix D

Scaling Difficulties with samtools merge

D.1 Memory Leak

By default the **LSF** job scheduler at the Sanger Institute will issue 100MB to any submitted job. Given both the number of input files and their total size it was anticipated that this merge job would require more memory. At the very least an estimated 35kB for a 64-bit pointer to each input file and approximately 150MB to house a 32kB buffer to read data from each file also. Yet executing the job with 500MB of memory caused the scheduler to forceably terminate the process for exceeding the maximum allocated memory limit.

Assuming that the intermediate structures for storing and sorting the input data must have been greater than expected, the job's memory limit was generously increased with the syntax previously demonstrated by Listing 11 only to meet the same fate, even when reserving 16GB of memory. Even considering the thousands of input files this is an absurd amount of memory, I'd discovered a memory leak, a leak that appeared to drastically increase in severity in proportion to the number of input files.

Once reported, the leak was fixed by the author; several large variables constructed when translating the input format were not freed when no longer needed. Despite this, during execution of the **samtools** test suite, a series of leaks were still being reported.

D.2 Memory Leak in Test Harnesses

Wanting to brush up on finding memory leaks with **valgrind**, I volunteered to find and patch the remaining reported leaks. The results of this investigation are detailed in Appendix **??**.

...getline ...regcomp...

```
==30464== 416 bytes in 1 blocks are indirectly lost in loss record 85 of 103
==30464==    at 0x4A082F7: realloc (in /usr/lib64/valgrind/vgpreload_memcheck-amd64-linux.so)
==30464==    by 0x3FBCCCA725: duplicate_node (in /usr/lib64/libc-2.17.so)
==30464==    by 0x3FBCCD3ADA: duplicate_node_closure (in /usr/lib64/libc-2.17.so)
==30464==    by 0x3FBCCD415A: calc_eclosure_iter (in /usr/lib64/libc-2.17.so)
==30464==    by 0x3FBCCD79F6: re_compile_internal (in /usr/lib64/libc-2.17.so)
==30464==    at 0x4A08121: calloc (in /usr/lib64/valgrind/vgpreload_memcheck-amd64-linux.so)
==30464==    by 0x3FBCCCCD88: create_cd_newstate (in /usr/lib64/libc-2.17.so)
==30464==    by 0x3FBCCCD506: re_acquire_state_context.constprop.41 (in /usr/lib64/libc-2.17.so)
==30464==    by 0x3FBCC2132E: build_trtable (in /usr/lib64/libc-2.17.so)
==30464==    by 0x3FBCCD3792: re_search_internal (in /usr/lib64/libc-2.17.so)
==30464==    by 0x3FBCCD8E94: regexec@@GLIBC_2.3.4 (in /usr/lib64/libc-2.17.so)
==30464==    by 0x40C6FA: check_test_2 (test_trans_tbl_init.c:124)
==30464==    by 0x402CC4: main (test_trans_tbl_init.c:348)
```

Listing 15 : Example of **valgrind** locating a memory leak in one of the **samtools** test harnesses following the failure to release memory allocated to a compiled regular expression

...following this, further **samtools merge** jobs were submitted only to also be repeatedly terminated by the LSF scheduler , this time for exceeding the maximum execution time limit for the queue...

D.3 Poor Time Performance

...submitting the same job to the "long" (48hr) and "basement" (essentially unlimited) queues, it is clear that the job is taking an extraordinary length of time to complete...

...during this time I took the opportunity to patch various memory leaks in the test harnesses of both merge and split...

valgrind, the tool I used to track down memory leaks in the test harnesses of both samtools merge and samtools split actually consists of more than just memcheck.

callgrind is a profiling tool that keeps track of a program's call stack history, a handy feature built in to some development environments such as QtCreator.

Having constructed a modest test set of files to merge, I called samtools merge from the command line, attaching callgrind and later imported the resulting text file to QtCreator's profiling tool to interpret the results (I usually use KCacheGrind for this but I've been investigating QtCreator's feature set for reasons unrelated to this project), the result is immediately obvious - millions of calls to functions in zlib; a free compression library.

Further investigations using the QtCreator Analyze interface revealed that these calls all boiled down to one line called not during the process of deflating the input files (as I had expected) but actually during the compression of the output!

Looking at a brief explanation of the deflate algorithm, it seems reasonable to conclude the computational

Function	Location	Called	Self Cost: Ir	Incl. Cost: Ir
0x0000003fbdc032d0	/usr/lib64/libz.so.1.2.7	1097	2,904,064,736	10,226,990,234
0x0000003fbdc0b4e0	/usr/lib64/libz.so.1.2.7	1112	944,594,568	944,594,568
crc32	/usr/lib64/libz.so.1.2.7	2202	220,011,172	220,011,172
re_search_internal	/usr/lib64/libc-2.17.so	690	118,935,704	283,936,905
bam_aux_get	/home/sam/git/htslib/sam.c in /usr/local/bin/samtools	232074	112,513,788	112,513,788
0x0000003fbdc0a770	/usr/lib64/libz.so.1.2.7	756420	95,768,664	95,768,664
__memcpy_ssse3_back	/usr/lib64/libc-2.17.so	938957	79,871,763	79,871,763
0x0000003fbdc0a870	/usr/lib64/libz.so.1.2.7	3336	60,220,888	155,989,552
bam_translate	/home/sam/git/samtools/bam_sort.c:(387) in /usr/local/bin/samtools	116037	36,996,141	210,497,579
check_halt_state_context.isra.20	/usr/lib64/libc-2.17.so	276477	33,089,554	40,556,431

Fig. D.1 callgrind output following merge with default output compression

Function	Location	Called	Self Cost: Ir	Incl. Cost: Ir
sam_format1	/home/sam/git/htslib/sam.c:(127) in /usr/local/bin/samtools	116037	1,465,277,216	1,470,263,575
0x0000003fbdc07370	/usr/lib64/libz.so.1.2.7	1188	1,075,724,775	1,075,724,775
re_search_internal	/usr/lib64/libc-2.17.so	690	118,935,056	283,738,997
bam_aux_get	/home/sam/git/htslib/sam.c in /usr/local/bin/samtools	232074	112,513,788	112,513,788
__memcpy_ssse3_back	/usr/lib64/libc-2.17.so	953054	50,203,177	50,203,177
bam_translate	/home/sam/git/samtools/bam_sort.c:(387) in /usr/local/bin/samtools	116037	36,996,141	209,790,146
check_halt_state_context.isra.20	/usr/lib64/libc-2.17.so	276470	33,088,917	40,555,605
0x0000003fbdc0a230	/usr/lib64/libz.so.1.2.7	3516	29,249,135	29,249,135
build_trtable	/usr/lib64/libc-2.17.so	3381	24,214,798	38,420,667
bam_aux_del	/home/sam/git/htslib/sam.c in /usr/local/bin/samtools	232074	18,673,215	36,202,474

Fig. D.2 callgrind output following merge with uncompressed output

cost is rather asymmetric between compressing and uncompressing - in that the effort is locating blocks to compress and in comparison uncompressing is a reversible function on the known blocks.

Indeed, samtools merge specifies a -u option for uncompressed output and the callgrind output (second image) indicates significantly less calls to zlib functionality.

It remains to be seen whether this option will cut down the time needed for the large merge job, perhaps this is merely a red herring and we're yet to discover the true speed trouble. In the meantime let's see if sending this job to the farm will work.

D.4 The Red Herring

... might be interesting to use gprof which is more geared towards finding functions that spend all your execution time as opposed to callgrind which I believe counts CPU instructions.

After re-compiling htslib and samtools with the -pg flag to enable such profiling and executing the same previous merge command on the modest test set, the output as parsed by gprof seems to indicate that the trouble lies with bam_aux_get in htslib, with almost 50% of the execution time being spent in this particular function.

```
uint8_t *bam_aux_get(const bam1_t *b, const char tag[2])
{
    uint8_t *s;
    int y = tag[0]<<8 | tag[1];
    s = bam_get_aux(b);
    while (s < b->data + b->l_data) {
        int x = (int)s[0]<<8 | s[1];
        s += 2;
        if (x == y) return s;
        s = skip_aux(s);
```

```
        }
        return 0;
    }
```

At a glance it seems that bam_aux_get receives a pointer to a BAM record and a "tag", an array of two characters representing an optional field as defined in Section 1.5 of the SAM file spec.

The function then appears to fetch all these auxiliary tags and iterates over each, comparing a transformation of that tag (x) to a pre-computed transformation on the input tag (y).

This would of course be inherently slow for files with many such tags; especially given that the function is called twice for potentially each line in a BAM file.

D.5 The Plot Thickens

...as we increase the number of input files, the time taken to read them in becomes non-linearly slower. Currently my money is on the seemingly inefficient trans_tbl_init that appears to be called for each file, with the current table of all previous files as an input...