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
April 2, 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

Abstract

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[3]. 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.

During sample sequencing a large number of quality control metrics are generated to determine the quality of the reads from the sequencing hardware itself. At the Wellcome Trust Sanger Institute, the automated QC system currently relies on hard thresholds to make such quality control decisions with individual hard-coded values on particular metrics determining whether a lane has reached a level that requires a warning, or has exceeded the threshold and failed entirely. Whilst this does catch most of the very poor quality lanes, a large number of lanes are flagged for manual inspection at the warning level; a time consuming task which invites inefficiency and error.

In practise most of these manual decisions are based on inspecting a range of diagnostic plots which suggests 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.

Contents

Contents **iii**

1 Introduction **1**

1.1 Project Aims 1

1.1.1 Analysis of Current System 1

1.1.2 Identification of Properties that affect Downstream Analysis 2

1.1.3 Description 2

1.1.4 Samples, Lanes and Lanelets 2

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[3]. 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**, 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 current vertebrate sequencing pipeline[1] at the institute, **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 hard thresholds; determining whether a particular sample has reached a level that requires a warning, or has exceeded the threshold and failed entirely. Whilst this does catch 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.

In practise most of these manual decisions are based on inspecting a range of diagnostic plots which suggests 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 or bad 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.

1.1.3 Description

1.1.4 Samples, Lanes and Lanelets

A **sample** is a distinct DNA specimen from a particular human such as yourself. Samples are then inserted in to a flowcell, which is essentially a series of very thin glass tubules, each tubule is described as a **lane**. It is throughout these lanes that the chemical reactions involved in sequencing will take place. Before the process

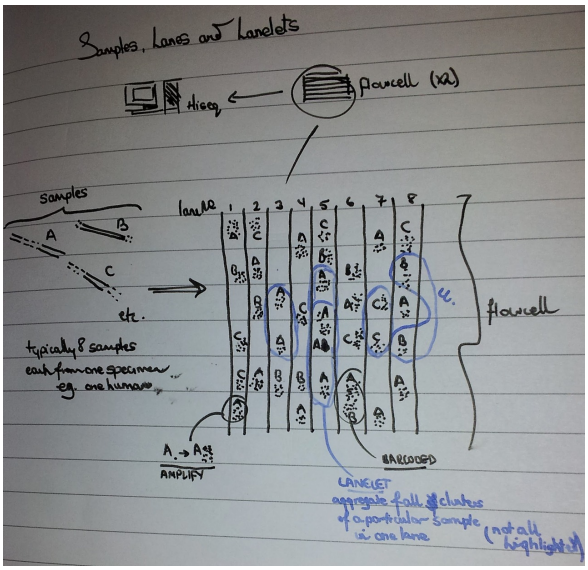

Fig. 1.1 Example of flowcell with some lanelets highlighted

begins, the samples are amplified in situ inside the flowcell itself (see Fig 1.2) whereby samples multiply in magnitude to form millions of dense clusters.

Note that a lane can contain more than one sample and a sample can appear in more than one lane; this is "sample multiplexing" and helps to ensure that the failure of a particular lane does not hinder analysis of a sample.

A **lanelet** is the aggregate read of all clusters of a particular sample in a single lane. Figure 1.1 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 actually be millions) of Sample A combine to represent a lanelet. A lane will have as many lanelets as it does samples.

References

[1] vr-pipe, a generic pipeline system [Github]. [Online]. Available: https://github.com/wtsi-hgi/vr-pipe/

[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] 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.