
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
March 28, 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

## 1.1 Introduction and Background

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[2]. However, these "next-generation" processes are complex and open to error[1], thus quality control is an essential step to assure confidence in any downstream analyses performed.

## 1.2 Description

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.

# References

[1] M. Kircher, U. Stenzel and J. Kelso (2009). Improved base calling for the Illumina Genome Analyzer using machine learning strategies. *Genome Biology*, 10(8):R83.

[2] Strachan, T. and Read, A. (2011). *Human Molecular Genetics*. Garland Science, 4th edition.