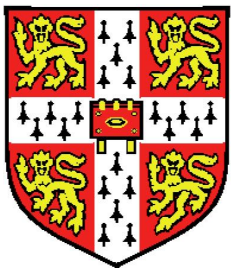
Efficient sequence assembly and variant calling using compressed data structures



Jared Thomas Simpson

Queens' College

Wellcome Trust Sanger Institute

University of Cambridge

This dissertation is submitted for the degree of

Doctor of Philosophy

September 2012

I took a lengthy path to reach this point and my family supported me the entire way. I dedicate this work to them - thank you Mom, Dad, Calley and Kim.

Declaration

This dissertation describes work carried out from May 2009 to July 2012 under the supervision of Dr Richard Durbin at the Wellcome Trust Sanger Institute, while a member of Queens' College, Cambridge. This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text. The content in Chapter 2 was published in Simpson and Durbin [2010]. The content of Chapter 3 was published in Simpson and Durbin [2012].

This thesis does not exceed the length limit of 60,000 words as specified by the Biology Degree Committee.

Jared Thomas Simpson
August 29, 2012

Acknowledgements

My decision to undertake a Ph.D. was influenced by many of my friends and colleagues. I owe thanks to my close friends Gwynn Elfring and Scott Drader, who have been a constant source of support and advice. My friend Christian Steidl introduced me to Computational Biology. This introduction led me to work on the sequence assembly problem, which directly led to this work.

My Ph.D. benefitted from many people at the Sanger Institute, most importantly my supervisor Richard Durbin. Without Richard's guidance and insight this work would be greatly diminished. I am very appreciative of the opportunity to work with Richard and proud of the work we accomplished over the last four years. I had many great discussions within Richard's group in journal clubs, group meetings, lunches and over coffee. These discussions introduced me to many areas of genetics and provided a great environment for developing as a scientist. A strength of the Ph.D. program at Sanger is the opportunity to build collaborations across many different research areas. Within Richard's group I enjoyed collaborating with Leopold Parts, Aylwyn Scally and Kees Albers. I also appreciate the time I spent working in David Adams' lab and the Cancer Genome Project under Peter Campbell. I owe thanks to the Wellcome Trust for financially supporting my Ph.D. and Annabel Smith and Christina Hedberg-Delouka for making sure the graduate program at Sanger runs smoothly.

Summary

De novo genome assembly is one of the most computationally demanding problems in genomics. In this thesis, I describe a collection of novel algorithms for performing *de novo* assembly using compressed data structures. First, I describe an algorithm to directly construct the assembly string graph - a model of overlap-based sequence assembly - using the compressed FM-index data structure. Previous algorithms for constructing the string graph required the intermediate step of building a full overlap graph, then removing transitive edges from the graph. My novel FM-index based algorithm does not require this time-consuming intermediate step. This algorithm allows fast and memory efficient overlap-based assembly. In Chapter 3, I extend my FM-index algorithms to build a space-efficient assembler for real sequencing data by designing error correction, read merging and scaffolding algorithms. Using these efficient algorithms I am able to reduce the memory requirement for assembling a human genome to 54GB.

In Chapter 4, I address the problem of detecting DNA sequence differences between two related genomes - the *variant calling* problem. Traditional approaches to variant calling align short sequence reads to a reference genome. While this approach is effective for simple differences, like isolated SNPs, it is more difficult to find complex changes like the insertion or deletion of sequence. My approach is based on analyzing the structure of an assembly graph built from the sequence data from multiple individuals. In Chapter 5, I apply this approach to real sequencing problems, including finding *de novo* mutations in the child of two parents, somatically acquired mutations in cancer and polymorphic variants present in a large human population.

Contents

List of Figures