Analysis of short tandem repeat variation in large scale resequencing data

Weldon W. Whitener

Wellcome Trust Sanger Institute

St. Edmunds College

University of Cambridge

A thesis submitted for the degree of

Doctor of Philosophy

September 2011

I would like to dedicate this thesis to my loving parents...

Acknowledgements

None of this would of been possible without the love, support and guidance from the best parents a son could ever hope to have. Thank you Mom and Dad for always believing in me.

Now, where to begin?

There have been so many people who have contributed in making this document possible. Along the way, I've had more support than anyone could ask for, as well as, plenty of lucky breaks! I want to begin by thanking all my professors back at my *alma mater*, North Carolina State University. Specifically, I would like to thank Dr. Frank Abrams for always allowing me to use the 'big words', as well as, Dr. Elizabeth Loboa for giving me my first shot at scientific exploration.

I want to thank Dr. Mark van Dyke for allowing a scrappy biomedical engineer into one of the most amazing labs I've ever had the good fortune of being a part of. You are one of the most inspiring and friendly persons I've ever met and I am forever grateful for your support and guidance along the way. None of my subsequent success would of been possible had I not met you by chance at a career event my third year at North Carolina State University.

The next person I'd like to thank is someone whose guidance and support cannot be fully conveyed in writing. Thank you Jennie LaMonte for always going above and beyond. More than anything, you have been a great friend and mentor to me for the past six years and

I hope that we remain friends for many years to come.

The past four years have been some of the most challenging, yet thought provoking, years of my life. Richard Durbin accepted me into his group and helped me cultivate a probing and scientific outlook that has helped me answer some of the most difficult questions I've ever been posed. Richard's continued support, guidance and occasional knock on the head has given me a true confidence in myself that I will be forever grateful for. Thank you for teaching me to stand on my own two feet and giving me the tools to help me be successful no matter where life takes me.

This thesis would not of been possible without the help of countless people at the Sanger Institute and University of Cambridge. Thank you Leopold Parts and Aylwyn Scally for helping me come to grips with statistical modeling. You were the best labmates I could of asked for and the help you offered me will never be forgotten. Thank you Jim Stalker, Thomas Keane and the entire vertebrate sequencing group for making my life so much easier by doing a great job at curating all the sequencing data – an almost insurmountable task! Thank you Avril Coghlan for your continued collaboration throughout the duration of my thesis. Thank you Stijn van Dongen, Sergei Manakov, David Adams, Theo Whipp and Steve Russen for your support and friendship throughout my time at the Sanger Institute. I also wish to send a huge thank you to my favorite collaborator, David Knowles. Without David's masterful grasp of machine learning and diligence, most of the higher level analysis conducted in the latter chapters of this thesis would not of been possible. David even managed to make it a fun process – through his energy and good nature – which by no means is no small feat!

Lastly, I want to thank a person who, as much as my parents, has helped me become the the person I am today. Without my sister

Witnee's guidance, I would not of been able to achieve all that I have. She taught me how to conduct myself in a manner that has opened doors for me both socially and professionally. She also offered the best advice while keeping my spirits high. I could not of done this without her, and for that, I am forever grateful.

Declaration

This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text. This thesis does not exceed the length limit set by the Biology Degree Committee.

Weldon W. Whitener
30 September 2011

Abstract

The average eukaryotic genome contains many types of variation; from single nucleotide polymorphisms, small, medium and large insertions and deletions to copy number variation, translocations and inversions to name a few. The genome is also highly non-uniform, with some regions more variable than others. Tandem repeats are stretches of DNA comprised of a short motif repeated end-to-end multiple times. They are of interests to geneticists because they exhibit a high rate of length variation and are relatively frequent in the genome. However, until now they have been hard to assay using new sequencing technologies, which have revolutionized the study of other types of genetic variation. In this thesis, we address this deficit by developing methods to genotype short tandem repeats from shotgun short sequencing reads and applying them to human genome data.

To begin, I present a statistical model based on a Bayesian framework which uses Illumina paired end sequencing reads to determine the genotype of a diploid individual at a given short tandem repeat locus. This method is applied to all triplet tandem repeats (repeat motifs three bases in length) in the human genome for an individual sequenced deeply from multiple libraries as part of the 1000 Genomes project. We show that our method has good sensitivity and specificity for both homozygous and heterozygous indel genotypes measuring over three bp in length.

Next, we build upon the previous chapter by utilizing our model for genotyping across nine deeply sequenced individuals. We use the putative indel calls made in this data set to gain an understanding of

what factors of a tandem repeat have the largest effect on observing an indel at a given locus. We look at the effect that various measures of repeat length, repeat purity, GC content and tandem repeat motif have on triplet repeat variation. This analysis furthers our understanding of tandem repeat variation.

Lastly, we reformulate our individual genotyping model to take sequencing data from multiple, low sequence depth individuals in a population to understand the population distributions of variants at tandem repeat loci. This uses machine learning approaches including the expectation-maximization algorithm and Gibbs sampler, that help elucidate which loci show evidence of variation in the sample population, and allow us to explore the distribution of alternate alleles at a locus. As well as cataloguing variation efficiently, this allows us to examine a broader picture of the contribution the previously described factors have in influencing variation at a tandem repeat locus.

Contents

List of Figures