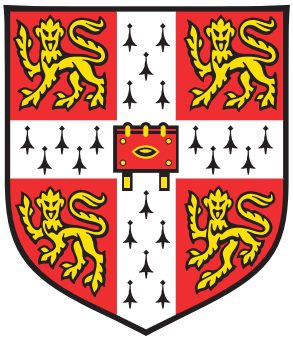
The Genetics of Cellular Phenotypes



Zhihao Ding

Wellcome Trust Sanger Institute

University of Cambridge

This dissertation is submitted for the degree of

Doctor of Philosophy

Darwin College

August 2014

To my grandmother, my parents and my wife

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original. 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. This thesis does not exceed the length limit of 60,000 words specified by the Biology Degree Committee.

Zhihao Ding
August 2014

Acknowledgements

It has been a lengthy journey for me to reach my PhD. My interest and motivation in science is mostly influenced by my grandmother and my parents to whom I owe the most gratitude. I thank my friends and lecturers in Wuhan University. The undergraduate program in the College of Life Science led me into the door of biology. I thank Doctor Andrew Coulson and Professor Mark Blaxter in the University of Edinburgh, who gave me strong support and guidance during my master study, which eventually led me to the bioinformatics field that I thoroughly enjoy. Before my PhD, I worked in the lab of Professor Carlos Caldas, who offered me great opportunities to participate in cancer projects and supported me for my application for a PhD.

During my PhD, I owe most thanks to my supervisor Richard Durbin. His patience in guidance, precision as a mathematician, and outstanding creativity in research have profoundly helped me in my journey of research and influenced my views on science in general. I mostly enjoy the environment of the Durbin group, which has always been a rich resource for discussions and advices. Strong numerical skills in people around me have been tremendously helpful in my PhD education. I particularly thank Andrew Brown, Leopold Parts, Jared Simpson, Kees Albers, Kimmo Palin, Thomas Kean, Shane McCarthy, Aylwyn Scally, Stephan Schiffels, Vladimir Shchur and Andreas Leha for their help on many questions from me during my work. I thank the Wellcome Trust Graduate Program that funded this study. I thank Annabel Smith, Christina Hedberg-Delouka, Alex Bateman and Julian Rayner who supervised the PhD program and offered me great help both on my study and on my status as an international student. The program has been a wonderful training framework that offered precious training opportunities both on science and on personal skills that I

feel greatly benefited.

In the past four years, I worked with excellent collaborators on many aspects of my projects. I would like to thank John Winn in Microsoft Research for his insightful input on the work in chapter 3; Ewan Birney for his leadership in the CTCF project; and Oliver Stegle for his advice on statistical models. I also appreciate the time I worked as a rotation student in the laboratories of Mike Stratton and David Adams. It has been a great pleasure to work with these and many other people.

In the end, this thesis is dedicated to my family, who supported me the entire way.

Abstract

Waves of genome wide association studies (GWAS) have identified a large number of loci associated with disease predisposition and natural traits in the past decade. A number of identified variants have revealed potential causal mechanisms for the associated diseases. However, despite the early success, much of the phenotypic variation is not explained by the GWAS variants and the effect sizes tend to be very small. The real challenge in advancing our understanding, and subsequently making it relevant for clinical application, is deciphering the biological functions of these loci, which remain largely uncertain. Compared to the whole organism phenotypes that are distal to the genetic variants, cellular phenotypes are closer to genetic regulation, thus not only tend to offer effect size, as shown in expression QTL studies, but also are likely to mediate between genotypes and whole organism phenotypes, supporting biological functions.

In chapter 2, I describe a genetic association study on binding of a primary transcription factor CCCTC binding factor (CTCF) in human populations. We search for quantitative trait loci (QTL) for tens of thousands of CTCF binding sites in a group of 51 individuals, making this the first well powered QTL study on a major transcription factor in humans. We discovered a large number of QTLs and revealed a strong genetic component that contributes to binding variation. We found the associated variants are often located near to predicted binding sites, some perturbing the binding motif directly, and others affecting indirectly. We observed allele specific effect (intra-individual) consistent with QTL signals (inter-individuals), supporting a strong genetic component in CTCF binding variation.

In chapter 3, I address the problem of low power in associations between gene

expression levels and phenotypes. This is largely driven by the high degree of stochasticity in the measured gene expression levels. We showed that by applying factor analysis both to remove global confounding effects and to create summarizing factors for biological pathways, the heritability and association strength can be substantially elevated as a result. We applied this idea to a cohort with skin expression data with ageing phenotypes, and discovered heritable ageing pathways.

It is also of great interest to develop new methods for obtaining measurements of cellular phenotypes. In chapter 4 I describe a novel computational method to estimate telomere length from whole genome or exome sequencing data. Using data from the TwinsUK cohort that has both DNA sequencing data and experimental telomere length measurements available, I show that our method can effectively extract telomere length information. The method has been applied to a few cancer studies in collaboration and achieved early success in confirming experimental findings.

Contents

List of Figures

List of Tables