

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

Contents	i
List of Figures	iv
List of Tables	vii
1 Introduction	8
1.1 Hunting for genetic determinants of phenotypes	9
1.1.1 Mendelian traits	9
1.1.2 Quantitative traits	10
1.1.3 Genetic variation and markers	10
1.2 Mapping quantitative traits	12
1.2.1 Linkage analysis and its limitation	12
1.2.2 Population association analysis	13
1.2.2.1 Mapping disease variants with case control phenotypes	14
1.2.2.2 Mapping QTLs using a simple <i>t</i> -test	14
1.2.2.3 Mapping QTLs using linear regression models	15
1.2.2.4 Mapping QTLs using linear mixed models	17
1.2.3 Multiple testing correction	18
1.2.4 Statistical power in genetic associations	19
1.2.5 GWAS results and interpretation	21
1.3 The promise of cellular phenotypes	22
1.3.1 Moving towards cellular phenotypes	22

1.3.2	The measurement of cellular phenotypes	23
1.3.3	Latent variables in analyzing high dimensional genotypes and cellular phenotypes	25
1.3.4	The genetics of gene expression	27
1.3.5	The genetics of transcription factor binding	30
1.3.6	The genetics of other epigenetic variation	33
1.3.7	Tissue and environment effect in QTL mapping	35
1.3.8	Resolving the causative relationship	36
1.4	Overview of the remainder of this thesis	37
2	The genetics of CCCTC binding factor	38
2.1	Overview	38
2.2	Measuring CTCF binding in HapMap cell lines	41
2.3	Quantification of CTCF binding	46
2.4	Imputing missing genotypes	50
2.5	Association testing	52
2.6	Allele specific analysis	53
2.7	Results	53
2.8	Discussion	81
3	Using latent factors to enhance power in mapping expression QTLs for ageing	86
3.1	Overview	86
3.2	Expression profiling	88
3.3	Gene expression pathway factors	91
3.4	Pathway factor and phenotype association	92
3.5	Heritability analysis	93
3.6	Single-gene based pathway enrichment analysis	93
3.7	Results	94
3.8	Discussion	102

4	Measuring telomere length from sequence data	105
4.1	Overview	105
4.2	Study samples and data	107
4.3	Estimating telomere length from whole genome sequence data.	107
4.3.1	Estimator	107
4.3.2	Simulation	112
4.3.3	Results	112
4.4	Estimating telomere length from exome sequence data.	119
4.5	Applications of the method	119
4.6	Conclusion	121
4.7	Software implementation	121
5	Conclusions and Future Work	123
5.1	Conclusions	123
5.2	Future Work	125
	References	129
A	Supplementary Tables	156

List of Figures

2.1	ChIP-seq production	42
2.2	Number of merged binding regions plotted as a function of $-\log(\text{BH-adjusted binomial P-value})$	44
2.3	Number of merged binding regions as a function of number of calling cell lines, at three adjusted P-values.	44
2.4	Proportion of merged binding regions as a function of number of calling cell lines, at three adjusted P-values.	45
2.5	Quality control by raw signal intensity and inter cell line variability. . .	48
2.6	Proportion of phenotypic variance explained by each principal component (PC)	51
2.7	Overall design of the experiment and overview of an example binding site.	55
2.8	Correlation of binding intensities between samples.	56
2.9	Power of association with normalized data sets.	57
2.10	An example CTCF QTL. Here shows all associations for all variants in the region of the binding region at chr3:108125397-108125829. SNPs are shown as solid circles and INDELS are shown as triangles, colored by R^2 . Inset is boxplot showing the normalized adjusted binding intensity (NABI) for the different possible genotypes of SNP rs936266. Genotype is strongly associated with the binding intensity of the binding region ($P=1.69\text{E-}19$), with the C allele favoring binding.	58

2.11	QQ plot for all associations between CTCF binding intensities and genotypes of variants within 50kb to the centre of binding sites.	59
2.12	The distributions of the effect size (β) and proportion of phenotypic variance explained of the QTL variants.	61
2.13	Effect sizes and proportion of variance explained of QTLs discovered at 1%FDR and 10%FDR.	61
2.14	CTCF binding QTLs.	63
2.15	The QTL effect size correlates with the information content and the GERP score for the variants present in the motif	64
2.16	P value distribution of the proximal variants.	66
2.17	Distribution of the proximal variants that are on motif and in LD with the distal lead QTL variants.	67
2.18	Evidence for indirect effects when a second binding region is present in the distal QTL window.	68
2.19	The interaction between QTL binding region and neighboring binding region correlates with regulatory events.	69
2.20	Change of histone modifications depending on the interaction models between the QTL binding region and the neighboring binding region (see Figure 2.18 and Figure 2.19 for explanations about the models).	70
2.21	Effect size versus derived allele frequency for all CTCF QTLs identified at 1 % FDR.	71
2.22	Example of CTCF peak shape QTL.	72
2.23	Summary of allele-specific analysis.	78
2.24	Allele-specificity correlates with QTL effect size (β).	79
2.25	Effect of the reference allele. Even when the reference allele is the derived allele (Derived), the binding bias remained towards the ancestral allele.	80
2.26	Effect of alignment to allele specific analysis.	82
2.27	QTLs with strong effect size in binding regions tend to show strong allele specificity.	83

3.1	QQ plot for factor analysis and single-gene based methods.	95
3.2	Network of connected factor phenotypes.	98
3.3	Histograms showing the proportion of environmental variation explained by age, heritability, and the proportion of variance explained by the unique environment for pathway factors and the individual gene measurements. The calculations correspond to equations in Box 1. Note that the proportions are not sum to one as they are not normalized by a same denominator: for age the variance explained by the genetic factors is removed.	100
3.4	The relative importance of sources of variation to global, pathway and gene phenotypes.	101
4.1	The effect of duplication rate and coverage to TelSeq performance. . . .	108
4.2	Identification of telomeric reads.	110
4.3	Normalising by reads with similar GC improves the performance of TelSeq	111
4.4	The effect of sequencing coverage on TelSeq measurement, assessed by simulation.	113
4.5	Comparison of TelSeq with experimental measure and age in TwinsUK samples.	114
4.6	Compare correlation coefficient obtained from mTRF and TelSeq. . . .	116
4.7	Sequencing lane variation in TelSeq measures.	117
4.8	The mTRF measurement is longer than TelSeq estimates across a range of values for the choices of TelSeq threshold (k).	118
4.9	TelSeq estimates from exome data are highly correlated with those from whole genome data.	120
4.10	Measuring telomere lengths in melanoma cases.	122

List of Tables

2.1	Correlations between PC1, PC2 and the experimental variables. In association tests PC1 was removed.	50
2.2	Summary statistics of the CTCF QTL scan.	60
2.3	CTCF QTLs with associated variants in different distance ranges.	62
2.4	The overlap between CTCF QTL variants and GWAS variants.	73
3.1	List of 20 pathways most significantly associated with age.	96
A.1	Sites with Random Allelic Bias.	156
A.1	Sites with Random Allelic Bias.	157
A.2	List of all pathways significantly associated with age.	157
A.3	List of the seven pathways which were significantly associated with age, discovered by looking for enrichment of single gene age associations.	160
A.4	Key showing which pathways correspond to which nodes in Figure 3.2, and the maximum Spearman correlation of that phenotype with any of the others representing pathways.	161
A.5	Heritability and proportion of variance explained by age for all pathways.	162