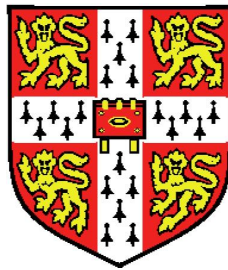


Genetic mapping of cellular traits



Leopold Parts

Wellcome Trust Sanger Institute

Corpus Christi College

University of Cambridge

A thesis submitted for the degree of

Doctor of Philosophy

September 2010

*To people who like swimming at midnight, climbing trees, or
hedgehogs. Or fractal snowflakes.*

Acknowledgements

This will be long - I am very thankful.

There are many people and institutions that have made this document possible. Along the way, I have been given chances that have one after the other, led me to writing these words. I thank Prof. Urmas Varblane for my first job, Prof. Helge Loebler for my first year abroad, Prof. Jaak Vilo for bringing me to the field, Ewan Birney for allowing me to sniff science for the first time, and Manolis Kellis for letting me in on large exciting project. Over the last four years, Richard Durbin, my supervisor, has been a solid source of fair critique, thoughtful comments, and strong writing - I thank him for his time and support. I am grateful to Wellcome Trust for funding my studies.

I have worked with many good collaborators, and slept on the floor of the best ones. Most of the work in this thesis was fought through with Oliver Stegle, often into the late hours at the Cavendish with good coffee. Gianni Liti and Francisco Cubillos have allowed me to think about biology and the experiments - and actually see them happen. There are many others who have been a pleasure to work with.

I have been blessed through time with great company both for enthusiasm for science as well as good times. The people in rdgroup have been a source of advice, knowledge, and banter; other sports-playing, french-speaking, GO-mongering colleagues have made the campus a pleasant environment. Phd06 - Kiki, Mare, Matti, Sergi, Steve - have been great companions for going through the trenches of academia together.

Most of all, I am indebted to my family. Grattitudes in Estonian. Aitähh – emale, kes väikseid lapsi laborisse nuuskima lubas, isale,

kes millestki puudust ei lasknud tunda, vennale, kes hea töö eest jäätiseid jagas, õele, kes teismelisi tööl arvutisse lasi, õele, kes bioloogide toredust näitas, vanaemale, kes bussipeatuses väikelapse 200-ni lugemist kontrollis, ja kõigile teistele suurtele-vanadele, noortele-väikestele, sõpradele-sugulastele, kes ei lase meelest minna, kus kodu on.

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.

Leopold Parts
31 August 2010

Abstract

Many important traits are heritable, and have a strong genetic component. In simple cases, such as Mendelian diseases, the genetic cause can be found with linkage methods, and many trait genes have been mapped to date. More recently, association mapping studies have focused on complex traits that include prevalent human diseases, such as type 2 diabetes, hypertension, and others. Numerous genome-wide association studies have corroborated that no single gene explains all or even a large part of the heritable variability in such traits, and that individual effect sizes due to common variants are small. The effect of a single locus genotype on a global trait has to be mediated by cellular, tissue, and organ phenotypes. Thus, genetics of cellular traits is central to developing an understanding of the genetic basis of complex traits.

In this thesis, we address the problem of mapping cellular traits. First, we develop a statistical model based on Bayesian regression and factor analysis for association mapping with high-dimensional phenotypes. We show how accounting for global, non-genetic variance components in the phenotype data increases power to detect genetic associations. Applying the method on human gene expression variation data, we find that up to 30% of transcripts have a statistically significant association to a proximal locus genotype.

Second, we show how to infer intermediate phenotypes and use them for mapping genetic associations and interactions. We use a sparse factor analysis model to infer hidden factors, which we treat as intermediate cellular phenotypes that in turn affect gene expression in a yeast dataset. We find that the inferred phenotypes are associated

with locus genotypes and environmental conditions, and can explain genetic associations to nearby genes. For the first time, we consider and find interactions between genotype and intermediate phenotypes inferred from gene expression levels, complementing and extending established results.

Third, we develop a novel approach to map trait loci rapidly and in narrow intervals using massively parallel sequencing. We created advanced intercross lines between two phenotypically different wild isolates of baker's yeast with sequenced reference genomes. We then applied selective pressure on the intercross pool by growing it in a restrictive condition to enrich for individuals with protective alleles. Sequencing DNA from the pool before and after selection pinpoints genes responsible for the increased fitness. This novel method provides a rapid and fine scale QTL mapping strategy improving resolution and power.

Finally, we conclude the thesis by exploring mapping cellular traits in a series of short studies in different organisms.

Contents

Nomenclature	x <i>i</i>
List of Figures	x <i>iii</i>
List of Tables	x <i>iv</i>
1 Introduction	1
1.1 Mapping the genetic basis of heritable traits	2
1.1.1 Linkage mapping	4
1.1.2 Association studies	6
1.1.3 Other approaches	8
1.2 Genetic structure of traits	10
1.2.1 Independent locus effects	10
1.2.2 Context dependent locus effects	11
1.2.3 Missing heritability in human traits	13
1.3 Genetics of cellular traits	13
1.3.1 Assaying cellular traits	15
1.3.2 Genetics of gene expression	18
1.3.3 Genetics of other high-dimensional cellular traits	20
1.4 Quantitative genetic models	21
1.4.1 Single trait	22
1.4.2 High-dimensional traits	26
1.5 Inference for trait models	27
1.5.1 Frequentist inference	27
1.5.2 Bayesian inference	29

1.6	Major open questions	31
1.7	Contributions of this Thesis	31
2	Association mapping with high-dimensional traits	33
2.1	Expression QTLs	34
2.2	The PEER framework	36
2.2.1	Model	37
2.2.2	Inference	41
2.2.3	Alternatives	42
2.3	Method comparison	44
2.3.1	Comparison on simulated data	44
2.3.2	Comparison on real data	48
2.4	Expression QTL mapping in large human populations	53
2.4.1	HapMap phase 2 dataset.	53
2.4.2	The MuTHER study	56
2.4.3	The 1000 Genomes low coverage pilot	59
2.5	Interpretation of learned hidden factors.	61
2.6	Discussion	62
3	Genetic mapping with inferred traits	65
3.1	Expression analysis with cellular traits	66
3.2	Model of expression with unmeasured traits	69
3.2.1	Statistical model	69
3.2.2	Trait inference	70
3.2.3	Association and interaction testing	72
3.3	Phenotype inference	75
3.3.1	Factor analysis model performance	76
3.3.2	Datasets used	76
3.4	Association analysis with inferred phenotypes	77
3.4.1	Genotype and environment	77
3.4.2	mRNA and protein levels	80
3.4.3	eQTL hotspots	82
3.5	Interaction analysis with inferred phenotypes	82
3.5.1	Discovering interactions	82

3.5.2	Recovering interactions	85
3.6	Discussion	86
4	Genetic mapping using artificial selection	90
4.1	Trait mapping with natural genetic variation	91
4.1.1	Shortcomings of existing approaches	91
4.1.2	Leveraging artificial selection	92
4.2	Very large segregating yeast population	93
4.2.1	Recombination rate	93
4.2.2	Parental allele frequency	94
4.3	Mapping trait loci using selection	95
4.3.1	QTLs in haploid pool	95
4.3.2	QTL validation	95
4.3.3	Mapping resolution	98
4.3.4	Lack of fixation	98
4.3.5	Negative epistatic interactions	98
4.3.6	QTLs in diploid pool	99
4.3.7	Comparison with F1 segregant analysis	99
4.4	Data analysis	100
4.4.1	Sequencing data handling	100
4.4.2	Segregant analysis	100
4.4.3	Copy number and missing sequence.	101
4.4.4	Allele frequency inference	101
4.4.5	QTL inference	102
4.4.6	Linkage analysis in F1 segregants.	102
4.4.7	Epistatic interaction tests	103
4.5	Simulation experiments	103
4.5.1	Adaptive mutations.	104
4.5.2	Effects of selection on allele frequency.	105
4.6	Discussion	106

5	Additional gene mapping studies	109
5.1	Gene mapping with one million bacterial transposon mutants . . .	109
5.1.1	Transposon insertion library for <i>Salmonella Typhi</i>	110
5.1.2	Mapping essential genes	110
5.1.3	Mapping condition-specific essential genes	112
5.2	Gene mapping with grid of yeast crosses	112
5.2.1	Grid of yeast crosses	114
5.2.2	Recombination analysis	114
5.2.3	Linkage mapping	116
 6	 Conclusions	 118
6.1	Conclusions and discussion	118
6.1.1	Abundance and importance of eQTLs	118
6.1.2	Abundance and importance of interactions	119
6.1.3	Trait mapping using artificial selection	120
6.2	Future work	121
 A	 Variational inference in PEER	 123
 B	 Supplementary Tables	 135
 References		 171

Nomenclature

Acronyms

CEU	HapMap 2 'European' population - U.S. residents with Northern and Western European ancestry
CHB	HapMap 2 Chinese population - individuals from Beijing
EBV	Epstein-Barr virus
eQTL	Expression QTL
FDR	False discovery rate
FNR	False negative rate
FPR	False positive rate
fVBQTL	Fast VBQTL
GEO	Gene Expression Omnibus
GO	Gene Ontology
GWAS	Genome-wide association study
GxE interaction	Gene-environment interaction
iVBQTL	Iterative VBQTL

CONTENTS

JPT	HapMap 2 Japanese population - individuals from the Tokyo area
KEGG	Kyoto Encyclopedia of Genes and Genomes
KL	Kullback-Leibler
LCL	Lymphoblastoid cell line
LOD	Log-odds
MCMC	Markov chain Monte Carlo
mRNA	Messenger RNA
MS	Mass spectrometry
NA	North American strain
PCA	Principal Components Analysis
PCAsig	PCA with significance testing
PEER	Probabilistic estimation of expression residuals
QTL	Quantitative Trait Locus
RIN	RNA integrity number
SNP	Single nucleotide polymorphism
SVA	Surrogate Variable Analysis
VBeQTL	eQTL on residuals of fVBQTL
VBQTL	Variational Bayesian QTL mapper
WA	West African strain
YRI	HapMap 2 Nigerian population - Yoruba people of Ibadan

List of Figures

2.1	Additive model of sources of gene expression variability	35
2.2	Bayesian network of VBQTL	38
2.3	VBQTL simulation comparison	46
2.4	VBQTL real data comparison	49
2.5	eQTL findings in HapMap CEU population	54
2.6	VBQTL association validation	55
2.7	MuTHER eQTL discoveries	58
2.8	1000 Genomes eQTL discoveries	60
3.1	Gene expression analysis with inferred factors	68
3.2	Correlation of transcription factor activation and its mRNA level	81
3.3	Example interactions	84
3.4	Density of interacting genes	86
4.1	QTL mapping strategy	92
4.2	F6 intercross pool recombination landscape	94
4.3	Allele frequency change upon selection	96
4.4	Validation of <i>IRA1</i> and <i>IRA2</i>	97
4.5	Simulation allele frequency changes	107
5.1	Transposon insertion frequency distribution	111
5.2	Change in transposon insertions upon selection	113
5.3	Recombination rate in grid of crosses	115
5.4	Paraquat growth QTLs	117

List of Tables

B.1	Number of eQTLs in HapMap populations	135
B.2	Overlap of eQTLs between HapMap populations	136
B.3	Overlap of VBQTLs and standard eQTLs	136
B.4	HapMap trans eQTL counts and overlaps	137
B.5	Correlations of latent factors and known covariates	137
B.6	Method comparison statistics for human chromosome 19	138
B.7	Method comparison statistics for yeast	139
B.8	Method comparison statistics for mouse	140
B.9	Yeast factor associations	143
B.10	KEGG factor associations	143
B.11	Freeform factor associations	144
B.12	Locus associations for yeast factors	145
B.13	Locus associations for KEGG factors	146
B.14	Locus associations for freeform factors	147
B.15	eQTL peak associations to Yeast factors	148
B.16	Sequencing coverage of samples	148
B.17	Regions selected during intercross rounds 6 to 12	149
B.18	Heat QTLs	150
B.19	Genes changing in copy number in response to selection	151