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. Gratitudes 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

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

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

List of Tables