Using molecular QTLs to identify cell types and causal variants for complex traits

Jeremy Schwartzentruber

Wellcome Trust Sanger Institute

Clare College

University of Cambridge

This dissertation is submitted for the degree of Doctor of Philosophy

October 2017

Declaration of Originality

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the beginning of each chapter. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University of similar institution. This dissertation does not exceed the word limit set by the Degree Committee for the Faculty of Biology.

Signature:

Date:

Jeremy Schwartzentruber
October 2017

Abstract

Genetic associations have been discovered for many human complex traits, and yet for most associated loci the causal variants and molecular mechanisms remain unknown. Studies mapping quantitative trait loci (QTLs) for molecular phenotypes, such as gene expression, RNA splicing, and chromatin accessibility, provide rich data that can link variant effects in specific cell types with complex traits. These genetic effects can also now be modeled *in vitro* by differentiating human induced pluripotent stem cells (iPSCs) into specific cell types, including inaccessible cell types such as those of the brain. In this thesis, I explore a range of approaches for using QTLs to identify causal variants and to link these with molecular functions and complex traits.

In Chapter 2, I describe QTL mapping in 123 sensory neuronal cell lines differentiated from human iPSCs. I observed that gene expression was highly variable across iPSC-derived neuronal cultures in specific gene categories, and that a portion of this variability was explained by commonly used iPSC culture conditions, which influenced differentiation efficiency. A number of QTLs overlapped with common disease associations; however, using simulations I showed that identifying causal regulatory variants with a recall-by-genotype approach in iPSC-derived neurons is likely to require large sample sizes, even for variants with moderately large effect sizes.

In Chapter 3, I developed a computational model that uses publicly available gene expression QTL data, along with molecular annotations, to generate cell type-specific probability of regulatory function (PRF) scores for each variant. I found that predictive power was improved when the model was modified to use the quantitative value of annotations. PRF scores outperformed other genome-wide scores, including CADD and GWAVA, in identifying likely causal eQTL variants.

In Chapter 4, I used PRF scores to identify relevant cell types and to fine map potential causal variants using summary association statistics in six complex traits. By examining individual loci in detail, I showed how the enrichments contributing to a high PRF score are transparent, which can help to distinguish plausible causal variant predictions from model misspecification.

Acknowledgements

I have been fortunate to spend the past four years at the Wellcome Genome Campus, a place with many inspired scientists whose door is always open. Firstly, I would like to thank my supervisor Dan Gaffney, for allowing me to explore different project and ideas; for being involved every step of the way; for his commitment to think carefully about each claim we make; and for his focus on clearly explaining *why*. I am also grateful to members of our group - Kaur, Natsuhiko, Angela - who so readily shared their thoughts and expertise with me, as well as their code. I benefited greatly from the collaborative and supportive environment at the Sanger Institute, including discussions with Jeff Barrett, Leo Parts, Carl Anderson, Annabel Smith, and Christina Hedberg-Delouka, as well as Chris Wallace in Cambridge. A big thank you Alex Gutteridge, for his cheerfulness, clear code and helpfulness in many analyses.

Thanks also to the other PhD students who always included me and played so well with Maia (junior PhD13 member who grew up from 0 to 4 years old on campus). I wish you all the best Katie, Nicola, Sumana, John, Masha, Li Meng, Veli, Liliana, Alice!

I am grateful to Dr. Jenny Clapham, who was one of the few doctors that seemed to take the chronic pain that I suddenly developed seriously. It was a dark ~1.5 years of my life, and a doctor who believes what you say and trusts you can make such a difference, even when there are no clear answers.

Finally, thank you to my two little girls for the many delightful moments. To Neeltje for supporting me and our girls through both delightful and difficult moments, and for shouldering the lack of sleep! To Kees and Michalien for being so quick to help, so steadfast, and calm. And to my parents, for always listening to me and supporting me no matter what.

Abbreviations and key terms

ATAC	Assay for transposase-accessible chromatin
AUC	Area under the curve (used for ROC or other classifier metrics)
BAM	Binary sequence alignment file format
BF	Bayes factor
CADD	Method predicting deleteriousness of coding and non-coding variants
CAGE	Cap analysis of gene expression
caQTL	Chromatin accessibility QTL
CD	Crohn's disease
ChIP-seq	Chromatin immunoprecipitation followed by sequencing
CNS	Central nervous system
CQN	Conditional quantile normalization for gene expression
CV	Coefficient of variation (standard deviation / mean)
DNase I	Enzyme that preferentially cuts DNA at open chromatin
DRG	Dorsal root ganglion
E8	Essential 8 iPSC culture medium
eGene	eQTL gene
eQTL	Gene expression QTL
ESC	Embryonic stem cell
FANTOM	Dataset of TSS usage based on CAGE
FDR	False discovery rate
fgwas	A fine-mapping method incorporating functional genomic annotations
FPKM	Fragments per kilobase of exon per million mapped reads
GERP	Genomic evolutionary rate profiling, a sequence conservation metric
GTEx	Genome-tissue expression project
GWAS	Genome-wide association study
GWAVA	Method predicting deleteriousness of non-coding variants
HDL	High density lipoprotein
HGMD	Human gene mutation database
HIPSCI	Human induced pluripotent stem cell initiative
IBD	Inflammatory bowel disease
iPSC	Induced pluripotent stem cell
IPSDSN	iPSC-derived sensory neuron
LCL	Lymphoblastoid cell line
LD	Linkage disequilibrium
LDL	Low density lipoprotein
LLK	Log-likelihood (of a model, given certain parameters)
MAF	Minor allele frequency