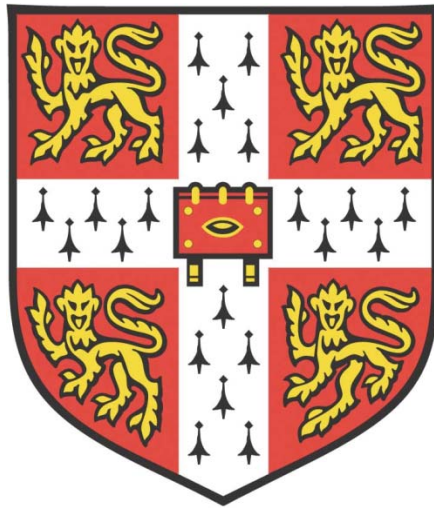


# **The Genetic Architecture of Immune-Mediated Complex Diseases**



**Jimmy Zhenli Liu**

**Darwin College  
University of Cambridge**

**This dissertation is submitted for the degree of  
Doctor of Philosophy  
September 2014**



## **Declaration**

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 Contributions section within each chapter. It does not exceed the word limit set by the Degree Committee for the Faculty of Biology, and is not substantially the same as any work that has been, or is being, submitted to any other university degree, diploma or any other qualification.

Jimmy Liu

18 September, 2014

That's the whole problem with science. You've got a bunch of empiricists trying to describe things of unimaginable wonder.

— Bill Watterson, Calvin and Hobbes

# Abstract

Complex disease risk is characterised by a combination of multiple genetic factors along with the environment. Since 2005, genome-wide association studies have discovered thousands of genetic variants associated with hundreds of such diseases. Following on from these types of studies, custom genotyping arrays with dense SNP content have allowed for greater refinement across risk loci, while their low cost has enabled powerful locus discovery projects and cross-phenotype comparisons in very large sample sizes. Combining risk loci with disease-relevant functional genomic data allows for insights into the biology of disease. In this dissertation, I explore locus discovery, cross-phenotype comparisons and functional data integration across four immune-mediated complex diseases – primary biliary cirrhosis, primary sclerosing cholangitis, and the two forms of inflammatory bowel disease – Crohn’s disease and ulcerative colitis.

In Chapter 1, I provide a historical background of our understanding of how genetic variation contributes to phenotypic variation, and the technological and theoretical advances in the last twenty years that have led to the large-scale high-throughput locus discovery projects of today.

In Chapter 2, I describe a locus discovery project using the ImmunoChip custom genotyping array for primary biliary cirrhosis. In addition to identifying three new risk loci and refining associated variants within known risk loci, I explore how integrating association results with functional genomic annotations across various cell lines from the ENCODE Project can provide insights into the cell types and genomic features most relevant to disease.

In Chapter 3, I describe a similar locus discovery project using the ImmunoChip for primary sclerosing cholangitis (PSC), where nine novel risk loci were identified. Over 80% of PSC patients are also diagnosed with inflammatory bowel disease, the majority of which is ulcerative colitis. I explore genetic factors

that may explain this overlap, and show that despite this high comorbidity, around half of PSC risk loci appear unique to PSC.

In Chapter 4, I describe a trans-ethnic genome-wide association meta-analysis for inflammatory bowel disease (IBD) comprising individuals of European, East Asian, Indian and Iranian ancestry genotyped on a combination of genome-wide arrays and the ImmunoChip. Forty new IBD loci were discovered associated with Crohn's disease, ulcerative colitis or both. I show that there exists pervasive sharing of IBD risk loci between European and non-European populations, while also noting specific loci where effect sizes differ between populations. The study demonstrates the utility of performing large-scale GWAS meta-analyses across different populations to identify novel susceptibility loci.

I then move beyond locus discovery in Chapter 5, where I describe a simple method for integrating differential gene expression datasets with disease risk loci. I applied the method to two gene expression datasets reflecting the genes that are involved in maintaining intestinal T cell homeostasis, and those triggered in the gut in response to infection. I find that in both cases, genes that are differentially expressed between these conditions are significantly overrepresented among risk loci for a range of autoimmune disorders, allowing for the identification of additional candidate genes at these loci and the generation of hypotheses about the mechanism through which they mediate disease.

Finally, in Chapter 6, I discuss the major themes of the preceding chapters on unravelling the genetic architecture of complex diseases. I then look to the types locus discovery projects that will shape the field in the coming years, and the potential for these to be ultimately translated into better treatment outcomes for patients.

# Acknowledgements

First and foremost, I thank my supervisor, Carl Anderson, for giving me the opportunity to pursue this PhD. It has been an absolute privilege. This dissertation would not have been possible without your continued guidance, enthusiasm and ceaseless faith in me throughout these years. With your stubborn attention to detail and unrelenting loyalty to rigour, you stand as a role model scientist for myself and no doubt many others to come.

I also thank my secondary supervisor, Jeff Barrett, and my degree committee, Stephen Sawcer and Ines Barroso for their guidance over these years. Thank you also to Christina Hedberg-Delouka, Annabel Smith, Alex Bateman and Julian Rayner for keeping the Sanger PhD program such a well-oiled machine.

To members of the Anderson Group (both past and present) - Tejas Shah, Eva Serra, Sun-Gou Ji and Jamie Floyd - I could not have asked for nicer folks to share an office with. It's been an absolute pleasure working with you all; thank you for putting up with me.

The work presented in this dissertation would not have been possible without the efforts of collaborators both at Sanger and around the world, of which there are far too many to list here. But for their hard work, dedication and willingness to share the spoils of research, I am especially indebted to Mohammed Al Marri, Luke Jostins, Daniel Gaffney, Richard Sandford, Trine Folseraas, Tom Hemming Karlsen, Johannes Roksund Hov, Eva Ellinghaus, Andre Franke, Tim Raine, Adam Reid, Suzanne van Sommeren, Rinse Weersma and Hailiang Huang. Thank you also to legions of doctors, nurses, researchers and administrators of the UK PBC Consortium, the International PSC Genetics Consortium and the International IBD Genetics Consortium for their tireless efforts in bringing together groups around the world towards the common noble goal of advancing disease research. None of this of course would have been possible without the >100,000 donors whose DNA were used in these projects, for which I will be forever grateful.

I would also like to thank the many friends and colleagues I've gotten to know during my time at Sanger. Chris Franklin, Yang Luo, James Morris, Scott Shooter, Isabelle Cleyman, Mari Niemi and everyone in Morgan N333 and N309 - cheers for the lunchtime banter, post-lunchtime strolls, post-stroll tea breaks, mingles and thoughts. To everyone I've enjoyed playing with (and competing against) in the Genome Campus volleyball, football and cricket leagues, I tried my best and can only apologise. Thank you to the fellow PhD students and Sanger and those I've gotten to know in Cambridge for making my time here so enjoyable.

I also thank the Wellcome Trust for generously funding my time at Sanger, as well as the various funding bodies for making these projects possible.

Lastly, I want to thank my family, especially my parents, Hua and Xiaoyu, for the unconditional love, support, understanding and encouragement that they have shown me in everything that I do.



# Table of Contents

<b>Abstract</b> .....	<b>iv</b>
<b>Acknowledgements</b> .....	<b>vi</b>
<b>Publications</b> .....	<b>xii</b>
<b>From this dissertation</b> .....	<b>xii</b>
<b>Arising elsewhere</b> .....	<b>xii</b>
<b>List of tables</b> .....	<b>xiii</b>
<b>List of figures</b> .....	<b>xiv</b>
<b>Chapter 1. Introduction and historical perspective</b> .....	<b>1</b>
<b>1.1 Immune-mediated diseases</b> .....	<b>2</b>
1.1.1 The immune system .....	2
1.1.2 Epidemiology .....	3
<b>1.2 Genetic studies of complex autoimmune disorders</b> .....	<b>4</b>
1.2.1 Mendelian inheritance, multifactorial traits and heritability .....	4
1.2.2 Twin studies .....	8
1.2.3 The major histocompatibility complex.....	9
1.2.4 Linkage.....	9
1.2.5 Candidate genes .....	11
1.2.6 Genome-wide association studies.....	12
<b>1.3 Insights from GWAS</b> .....	<b>16</b>
1.3.1 Biology.....	16
1.3.2 Genetic overlap between immune-mediated disorders.....	17
<b>1.4 Locus discovery beyond GWAS</b> .....	<b>18</b>
1.4.1 Dense genotyping.....	18
1.4.2 Finemapping and inferring causality.....	19
1.4.3 Sequencing and rare variant associations .....	20
<b>1.5 Conclusions</b> .....	<b>23</b>
<b>1.6 Outline of dissertation</b> .....	<b>23</b>
<b>Chapter 2. Discovery, refinement and functional genomics integration of primary biliary cirrhosis risk loci using the ImmunoChip</b> .....	<b>26</b>

<b>2.1</b>	<b>Introduction .....</b>	<b>26</b>
2.1.1	Chapter overview .....	27
2.1.2	Contributions .....	28
<b>2.2</b>	<b>Methods.....</b>	<b>28</b>
2.2.1	Samples, DNA extraction and genotyping.....	28
2.2.2	Quality control .....	29
2.2.3	Imputation.....	31
2.2.4	Association analysis .....	31
2.2.5	HLA Imputation.....	31
2.2.6	Variance in disease risk explained.....	32
2.2.7	eQTL analysis .....	32
2.2.8	Enrichment of open chromatin regions.....	33
<b>2.3</b>	<b>Results and discussion.....</b>	<b>34</b>
2.3.1	Replicating known PBC risk loci .....	34
2.3.2	Multiple independent signals.....	36
2.3.3	Novel PBC risk loci .....	39
2.3.4	Associations with HLA haplotypes .....	40
2.3.5	Functional annotations and enrichment of open chromatin regions among risk loci.....	41
<b>2.4</b>	<b>Conclusion.....</b>	<b>46</b>
<b>Chapter 3.</b>	<b>Discovery of primary sclerosing cholangitis risk loci and the genetic relationship with inflammatory bowel disease.....</b>	<b>48</b>
<b>3.1</b>	<b>Introduction .....</b>	<b>48</b>
3.1.1	Chapter overview .....	49
3.1.2	Contributions .....	49
<b>3.2</b>	<b>Methods.....</b>	<b>49</b>
3.2.1	Samples, DNA extraction and genotyping.....	49
3.2.2	Quality control .....	50
3.2.3	Imputation.....	51
3.2.4	Association analysis .....	53
3.2.5	Functional annotation of risk loci .....	54
3.2.6	GRAIL and DAPPLE analyses.....	55

3.2.7	HLA imputation and association analysis.....	55
3.2.8	Heritability explained.....	55
3.2.9	Prediction of PSC using IBD risk loci.....	56
3.2.10	Genetic correlation between PSC and IBD.....	56
<b>3.3</b>	<b>Results and discussion.....</b>	<b>57</b>
3.3.1	Locus discovery.....	57
3.3.2	Associations at previously reported non-HLA PSC risk loci.....	61
3.3.3	Candidate gene prioritisation.....	62
3.3.4	HLA association.....	63
3.3.5	Genetic overlap with IBD.....	66
<b>3.4</b>	<b>Conclusion.....</b>	<b>73</b>

<b>Chapter 4.</b>	<b>Trans-ethnic meta-analysis for inflammatory bowel disease risk loci and population comparisons.....</b>	<b>75</b>
<b>4.1</b>	<b>Introduction.....</b>	<b>75</b>
4.1.1	Contributions.....	76
<b>4.2</b>	<b>Methods.....</b>	<b>77</b>
4.2.1	Sample collection and genotyping.....	77
4.2.2	ImmunoChip quality control.....	77
4.2.3	Per-population association analysis.....	80
4.2.4	Transethnic meta-analysis.....	80
4.2.5	Gene prioritisation.....	82
4.2.6	Variance explained.....	82
4.2.7	Heterogeneity of effect sizes and allele frequencies between populations.....	82
4.2.8	Genetic correlation.....	83
4.2.9	Gene-based likelihood ratio test.....	83
<b>4.3</b>	<b>Results and discussion.....</b>	<b>85</b>
4.3.1	Per-population association and transethnic meta-analysis.....	85
4.3.2	Candidate genes.....	87
4.3.3	Validation of known loci.....	90
4.3.4	Population comparisons.....	92
4.3.5	Gene-based likelihood ratio test.....	98

4.3.6	Conclusions .....	100
<b>Chapter 5.</b>	<b>Immune-mediated disease risk loci are enriched for differentially expressed genes from tissue-relevant functional genomic datasets .....</b>	<b>102</b>
<b>5.1</b>	<b>Introduction .....</b>	<b>102</b>
5.1.1	Contributions .....	105
<b>5.2</b>	<b>Methods.....</b>	<b>105</b>
5.2.1	Human T cell transcripts.....	105
5.2.2	Mouse cecum transcripts .....	106
5.2.3	GWAS enrichment.....	106
<b>5.3</b>	<b>Results .....</b>	<b>107</b>
5.3.1	Human T cell transcripts.....	107
5.3.2	Mouse cecum transcripts .....	109
<b>5.4</b>	<b>Discussion .....</b>	<b>111</b>
5.4.1	Conclusions .....	115
<b>Chapter 6.</b>	<b>Conclusions and future prospects .....</b>	<b>117</b>
<b>6.1</b>	<b>Effect sizes, power and the genetic architecture of complex traits.....</b>	<b>118</b>
<b>6.2</b>	<b>Future prospects for complex disease genetics .....</b>	<b>123</b>
6.2.1	Array-based approaches .....	123
6.2.2	Sequencing approaches for rare variant studies.....	124
6.2.3	Genetic studies in non-European populations .....	129
6.2.4	Genetic prediction.....	130
<b>6.3</b>	<b>From causal variants to treatment outcomes .....</b>	<b>131</b>
<b>6.4</b>	<b>Concluding remarks.....</b>	<b>135</b>
<b>Bibliography</b>		<b>136</b>

# Publications

## From this dissertation

Liu J.Z., van Sommeran S., Huang H., Ng S.C. *et al.*, Association study discovers 38 susceptibility loci for inflammatory bowel disease and shows pervasive sharing of genetic risk across diverse populations. *Under review*.

Raine T., Liu J.Z., Anderson C.A., Parkes M. and Kaser A., Generation of primary human intestinal T cell transcriptomes reveals differential expression at genetic risk loci for immune-mediated disease. *Gut. In press*.

Liu J.Z. and Anderson C.A., Genetic studies of Crohn's disease: past, present and future. *Best Practice & Research: Clinical Gastroenterology*, 28:373-386, 2014.

Foth B.J., Isheng J.T., Reid A.J., Bancroft A., *et al.*, Whipworm genomes and dual-species transcriptome analysis provide molecular insights into an intimate host-parasite interaction. *Nature Genetics*, 46:693-700, 2014.

Liu J.Z., Hov J.R., Folseraas T., Ellinghaus E., *et al.*, Dense genotyping of immune-related disease regions identifies nine new risk loci for primary sclerosing cholangitis. *Nature Genetics* 45:670-675, 2013.

Liu, J.Z., Almarri, M.A., Gaffney, D.J., Mells, G.F., Jostins, L., Cordell, H.J., Ducker, S.J., Day, D.B., Heneghan, M.A., Neuberger, J.M., *et al.* Dense fine-mapping study identifies new susceptibility loci for primary biliary cirrhosis. *Nature Genetics*, 44:1137-1141, 2012.

## Arising elsewhere

Curtis J., Luo Y., Zenner H.L., Cucho-Lourenco D., *et al.*, Susceptibility to tuberculosis is associated with the ASAP1 gene that regulates dendritic cell migration. *Under Review*.

Houldcroft C.J., Petrova V., Liu J.Z., Frampton D., *et al.*, Host genetic variants and gene expression patterns associated with Epstein-Barr virus copy number in lymphoblastoid cell lines. *PLoS One* 9:e108384, 2014.

Robles-Espinoza C.D., Harland M., Ramsay A.J., Aoude L.G., *et al.*, POT1 loss-of-function variants predispose to familial melanoma. *Nature Genetics*, 46:478-481, 2014.

Shah T.S., Liu J.Z., Floyd J.A.B., Morris J.A., *et al.*, optiCall: A robust genotyping-calling algorithm for rare, low frequency and common variants. *Bioinformatics* 28:1598-1603, 2012.

## List of tables

Table 2.1. Sample quality control.....	30
Table 2.2. SNP quality control .....	30
Table 2.3. Unconditioned and conditioned association results for the four independent signals at 3q25 .....	37
Table 2.4. Genome-wide significant HLA-type associations.....	41
Table 3.1. Post-QC patient and control panels .....	50
Table 3.2. Association results of twelve non-HLA genome-wide significant risk loci for PSC .....	58
Table 3.3. Association of genome-wide significant PSC risk loci with other diseases .....	61
Table 3.4. Candidate functional annotations and genes among genome-wide significant PSC risk loci.....	63
Table 3.5. Stepwise conditional analyses of the classical HLA genes .....	64
Table 3.6. IBD Subphenotypes among PSC cases .....	66
Table 4.1. Post-QC patient and control panels genotyped on the ImmunoChip....	78
Table 4.2. Post-QC case and control panels used in the transethnic meta-analysis .....	81
Table 4.4. Candidate genes implicated by coding variants, eQTLs, GRAIL and DAPPLE in 28 of the 40 novel IBD risk loci.....	88
Table 4.5. New genes in known IBD risk loci implicated from GRAIL and DAPPLE network analyses .....	89
Table 4.6. Pairwise genetic correlation ( $r_G$ ) tagged by ImmunoChip SNPs.....	94
Table 4.7. Genes that exceeded $P < 5 \times 10^{-5}$ in at least one non-European cohort in the likelihood ratio locus-based test.....	99
Table 5.1. Enrichment of genes that are upregulated in gut T cells compared with blood T cells in loci associated with six phenotypes .....	108
Table 5.2. Enrichment of genes that are differentially expressed between infected and uninfected cecum tissue among loci associated with six phenotypes. ....	109
Table 5.3. Annotation of disease-associated loci that are show nominal levels of enrichment ( $P < 0.05$ ) for genes that show differential expression in healthy gut vs. blood T cells and in infected vs. uninfected mouse cecum tissue.....	110

## List of figures

Figure 1.1. Mendel’s laws of inheritance.....	5
Figure 1.2. Polygenic inheritance in a normally distributed trait: height. ....	6
Figure 1.3. The liability threshold model.....	7
Figure 1.4. Power of linkage vs. association outlined in Risch and Merikengas (1996) .....	14
Figure 1.5. Number of publications indexed in PubMed with the terms “autophagy” and “Crohn’s” in the abstract since 2006 .....	17
Figure 2.1. Principal component analysis of PBC cases and controls.....	30
Figure 2.3. PBC risk loci odds ratios from this study vs. those from Mells <i>et al.</i> (2011) .....	36
Figure 2.4. Multiple independent signals at 3q25 from stepwise conditional regression.....	38
Figure 2.5. Enrichment of DNase-seq peaks among PBC risk loci in Gm12878 compared to other ENCODE cell lines.....	43
Figure 2.6. Enrichment of DNase-seq peaks among PBC risk loci calculated from P-value bins. ....	44
Figure 3.1. Heterozygosity rate and proportion of missing genotypes for PSC cases and controls.....	52
Figure 3.2. Principal components analysis of PSC cases and controls with 1000 Genomes Omni2.5-8 data .....	52
Figure 3.3. Quantile-quantile plots and genomic inflation factors of observed vs. expected P-values .....	54
Figure 3.4. Regional association plots for genome-wide significant associations at previously established PSC risk loci .....	59
Figure 3.5. Regional association plots of nine newly associated PSC risk loci .....	60
Figure 3.6. Regional association plots from stepwise conditional regression in the HLA complex in PSC .....	65
Figure 3.7. Odds ratio comparisons for PSC risk loci in IBD. IBD ORs and designation of loci as UC, CD or both (IBD) were obtained from Jostins <i>et al.</i> (2012) .....	67
Figure 3.8. Venn diagram of directions of effect in PSC of SNPs associated with either CD, UC or both (IBD) .....	67
Figure 3.9. Predicting PSC using OR estimates from CD and UC risk loci.....	68
Figure 3.10. Predicting the IBD subphenotypes of PSC patients using OR estimates from CD and UC risk loci .....	68

Figure 3.11. Genetic correlation ( $r_G$ ) estimates using genome-wide SNP data between CD/UC and PSC subphenotypes.....	70
Figure 3.12. Two models of pleiotropy.....	72
Figure 3.13. Odds ratios of PSC risk loci calculated using all PSC cases compared with odds ratios calculated using PSC+UC and PSC+no IBD subphenotypes .....	73
Figure 4.1. Principal components analysis of non-European IBD patients and controls.....	79
Figure 4.2. Comparison of samples used in this study with those from Jostins et al. (2012).....	80
Figure 4.3. GRAIL network for all genes with GRAIL $P < 0.05$ .....	90
Figure 4.4. Comparison of association P-values reported in Jostins <i>et al.</i> (2012) and Europeans in this present study.....	91
Figure 4.5. Odds ratio comparison between European and non-European populations at 233 SNPs associated with CD, UC other both.....	93
Figure 4.6. Belgravia plot of (A) CD and (B) UC risk variants in Europeans and East Asians .....	96
Figure 5.1. Number of upregulated genes that overlap among CD4+ LPL, CD8+ LPL, CD4+ IEL and CD8+ IEL T cells vs. counterparts in blood.....	107
Figure 5.2. Quantile-quantile plots of gene length of differentially expressed genes in (A) gut T cells vs. blood and (B) infected vs. uninfected cecum tissue.....	114
Figure 6.1. Effective sample size vs. number of genome-wide significant risk loci across GWAS and Immunochip studies of nine immune-mediated disorders....	119
Figure 6.2. Cumulative proportion of variance in disease liability explained by the genome-wide significant loci identified in Chapters 2-4 .....	120
Figure 6.3. The genetic architecture of inflammatory bowel disease .....	126



