Genomic profiling of response to *in vivo* immune perturbations

Benjamin Yu Hang Bai

Churchill College
University of Cambridge
Wellcome Sanger Institute

November 2020

This thesis is submitted for the degree of Doctor of Philosophy

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. 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 except as declared in the Preface and specified in the text. I further state that no substantial part of my thesis 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 or similar institution except as declared in the Preface and specified in the text. It does not exceed the prescribed word limit of 60 000 words for the Biology Degree Committee.

Abstract

Genomic profiling of response to *in vivo* immune perturbations

Benjamin Yu Hang Bai

The human immune system plays a central role in defense against infection, but its dysregulation is implicated in immune-mediated diseases. The past decade has seen increasing application of high-throughput technologies to profile, predict, and understand immune response to perturbation. The ability to measure immune gene expression at scale has led to the identification of transcriptomic signatures that predict clinical phenotypes such as antibody response to vaccines. It has also been recognised that both expression and phenotypic responses are traits with complex genetic architectures. This thesis examines the longitudinal transcriptomic response to immune perturbations, and its association with clinical response phenotypes and common genetic variation.

Chapter 2 explores transcriptomic response to pandemic influenza vaccine in a multi-ethnic cohort of healthy adults: the Human Immune Response Dynamics (HIRD) cohort. The success of vaccination in controlling influenza is indisputable, but it is incompletely understood why some individuals fail to mount protective antibody responses. I meta-analysed blood microarray and RNA sequencing (RNA-seq) datasets, identifying a distinct transition from innate immune response at day 1 after vaccination to adaptive immune response at day 7. Heterogeneity between measurement platforms made it difficult to identify single-gene transcriptomic associations with antibody response. Using a gene set approach, I found expression modules related to the inflammatory response, the cell cycle, $CD4^+$ T cells, and plasma cells to be associated with vaccine-induced antibody response.

In Chapter 3, I map response expression quantitative trait loci (reQTLs) in the HIRD cohort to investigate regulation of transcriptomic response by common genetic variants. Rather than driving differential expression post-vaccination, the strongest reQTLs appeared to be explained by changes in cell composition revealing cell type-specific expression quantitative trait locus (eQTL) effects. For example, a reQTL identified for *ADCY3* specific to day 1 may be explained largely by high monocyte proportions at day 1 compared to other timepoints. Changes in cell composition present a significant challenge to interpreting reQTLs found through bulk sequencing of heterogeneous tissues.

Finally, Chapter 4 applies an analogous longitudinal study design to explore drug response in the Personalised Anti-TNF Therapy in Crohn's Disease (PANTS) cohort: a cohort of Crohn's disease (CD) patients treated with the anti-tumour necrosis factor (TNF) drugs, infliximab

and adalimumab. Anti-TNF treatment has revolutionised patient care for CD, but 20–40 % of patients show primary non-response soon after starting treatment. I identified baseline expression modules associated with primary non-response, but also found significant heterogeneity of associations between the two drugs. Expression changes post-treatment in non-responders were largely magnified in responders, suggesting there may be a continuum of response. Distinct expression trajectories identified for responders and non-responders revealed sustained expression differences during the first year of treatment. Sets of interferon-related genes were regulated in opposing directions in responders and non-responders, presenting an attractive target for future studies of the biological mechanisms underlying non-response.

A little learning is a dangerous thing;
Drink deep, or taste not the Pierian spring:
There shallow draughts intoxicate the brain,
And drinking largely sobers us again.

Alexander Pope, *An Essay on Criticism*

Acknowledgements

"The more you know, the more you know you don't know". No other aphorism may be more apt at describing my perspective on these past four years, but what I do know—without a doubt—is that I owe a great deal to a great many people.

First and foremost, I must thank my supervisor, Carl Anderson. I greatly admire your commitment to embodying a culture of rigorous first-class science, and equally your commitment to the prosperity and well-being of everyone under your wing. I am indebted for all of your support and guidance, and your empathy and optimism during interesting times. Thank you for the privilege of the past four years—it has been an honour and a pleasure.

I must also take the opportunity to address the members of the Anderson team, past and present. The team has greatly evolved these past few years both in scientific direction and in membership, but it has always remained a friendly and enriching place to be. Whether over the lunch table in Murray's, during our legendary team retreats, or in our traditionally lively team meetings, there was never a lack of supportive and productive discussion. In particular, I must thank Aleksejs Sazonovs, Carla Jones-Bell, Elizabeth Goode, Laura Fachal, Leland Taylor, Loukas Moutsianas, Nikolaos Panousis, and Velislava Petrova, each of whom lent their ear to me on a multitude of occasions.

The research presented in this thesis would not have been possible without the help of numerous individuals and organisations. The projects in this thesis are highly collaborative, extending from the Wellcome Sanger Institute to London, Exeter, and the US. To Adrian Hayday, Nicholas Kennedy, Tariq Ahmad, and the team at AbbVie, thank you for granting me the opportunity to be part of these amazing collaborations. I owe a sincere gratitude to the many clinicians, scientists, and administrators who laid the foundation for this thesis, and even more so to the patients who contributed their samples. I would also like to thank the Wellcome Trust and other funding agencies that contributed to the projects in my thesis for their generous financial support.

Thank you to my thesis committee, Daniel Gaffney and Michael Inouye, who provided valuable feedback throughout the course of my PhD. I also extend my gratitude to the amazing research administrators I've worked with at Sanger: Carol Dunbar, Sally Bygraves, Eloise Stapleton, Paris Litterick, Rachel Henry, and Sophie Leggett. I'm sure there has been many a time you've known more about my projects than me. To the Sanger Sample Management, Pipelines, and Informatics teams, the incredible work you do to keep the sequencing and computing resources running smoothly was critical for this thesis, and for all science done on the Campus.

On the other side of the balance, I would like thank all the friends I have made during my time in Cambridge. To my fellow PhD cohort, I'll remember all the punting trips, the pub

quizzes, the eSCAMPS dinners, and our shared commiserations on the Campus bus. To the Churchill College badminton team, thank you for having me, and best wishes for all the training sessions and league matches in the years to come. I would like to give a special mention to my friends from the Cambridge University Anime and Manga Society: many good times were had; many good memories were made. Looking back on it all, my "campus life" has been profoundly multicoloured.

Finally, to my family—to my brother, my mother, my father, my grandparents—I could always depend on your unconditional love and support. Despite the thousands of miles between us, you were always held close to my heart.

Contents

Declaration iii

Abstract v

Acknowledgements ix

Contents xi

List of Figures xv

List of Tables xix

1 Introduction 1

1.1 Genetic association studies for complex traits . . . 1

1.1.1 Structure and variation of the human genome . . . 1

1.1.2 Lessons from the past fifteen years . . . 2

1.1.3 From complex trait to locus . . . 3

1.1.4 From locus to causal variant . . . 5

1.1.5 From causal variant to target gene . . . 5

1.2 Gene expression as an intermediate molecular phenotype . . . 5

1.2.1 Regulation of gene expression . . . 5

1.2.2 Expression quantitative trait loci (eQTLs) . . . 6

1.2.3 Context-dependent eQTLs . . . 7

1.2.4 Response eQTLs (reQTLs) . . . 7

1.2.5 Gene prioritisation using eQTLs . . . 10

1.3 Phenotypes of immune response . . . 10

1.3.1 An overview of the immune system . . . 10

1.3.2 High-throughput immunology . . . 11

1.4 Thesis outline . . . 12

2 Transcriptomic response to Pandemrix vaccine 15

2.1 Introduction . . . 15

2.1.1 Influenza . . . 15

2.1.2 Seasonal influenza vaccines . . . 16

2.1.3 Quantifying immune response to influenza vaccines . . . 17

2.1.4 Systems vaccinology of seasonal influenza vaccines . . . 18
2.1.5 The Human Immune Response Dynamics (HIRD) cohort . . . 20
2.1.6 Chapter summary . . . 21
2.2 Methods . . . 22
2.2.1 Existing HIRD data and additional data generation . . . 22
2.2.2 Computing baseline-adjusted measures of antibody response . . . 22
2.2.3 Genotype data generation . . . 25
2.2.4 Genotype data preprocessing . . . 25
2.2.5 Computing genotype principal components as covariates for ancestry . . . 25
2.2.6 RNA-seq data generation . . . 29
2.2.7 RNA-seq quantification and preprocessing . . . 29
2.2.8 Array data preprocessing . . . 32
2.2.9 Differential gene expression (DGE) . . . 35
2.2.9.1 Platform and batch effects . . . 35
2.2.9.2 Per-platform DGE model . . . 38
2.2.9.3 Choice of DGE meta-analysis method . . . 38
2.2.9.4 Prior for between-study heterogeneity . . . 40
2.2.9.5 Prior for effect size . . . 40
2.2.9.6 Example of priors . . . 40
2.2.9.7 Multiple testing correction . . . 41
2.2.10 Ranked gene set enrichment using blood transcription modules . . . 41
2.3 Results . . . 43
2.3.1 Extensive global changes in expression after vaccination . . . 43
2.3.1.1 Innate immune response at day 1 post-vaccination . . . 43
2.3.1.2 Adaptive immune response at day 7 post-vaccination . . . 45
2.3.2 Expression associations with antibody response . . . 45
2.3.2.1 Between-platform heterogeneity hinders detection of gene-level associations . . . 45
2.3.2.2 Module-level associations with antibody response . . . 47
2.4 Discussion . . . 52

3 Genetic architecture of transcriptomic response to Pandemrix vaccine 55
3.1 Introduction . . . 55
3.1.1 Host genetic factors affecting influenza vaccine response . . . 55
3.1.2 reQTLs induced by influenza vaccination . . . 56
3.1.3 Chapter summary . . . 57
3.2 Methods . . . 58
3.2.1 Overall strategy for reQTL mapping . . . 58
3.2.1.1 Adjusting for population structure using linear mixed models . . . 58
3.2.1.2 Multi-condition models . . . 59
3.2.1.3 Additional expression preprocessing . . . 60
3.2.2 Genotype phasing and imputation . . . 61
3.2.3 Estimation of kinship matrices . . . 63

3.2.4 Estimation of cell type abundance from expression . . . 63
3.2.5 Finding unmeasured covariates using factor analysis . . . 67
3.2.6 eQTL mapping per timepoint . . . 70
3.2.7 Joint eQTL analysis across timepoints . . . 70
3.2.8 Defining shared eQTLs and reQTLs . . . 71
3.2.9 Replication of eQTLs in a reference dataset . . . 72
3.2.10 Genotype interactions with cell type abundance . . . 73
3.2.11 Gene set enrichment analyses . . . 75
3.2.12 Statistical colocalisation . . . 75
3.3 Results . . . 76
3.3.1 Mapping reQTLs in the HIRD cohort . . . 76
3.3.2 Characterising reQTLs post-vaccination . . . 77
3.3.3 Exploring possible mechanisms generating reQTLs . . . 81
3.3.3.1 Differential expression of genes with reQTLs . . . 81
3.3.3.2 Genotype by cell type abundance interaction effects . . . 81
3.3.3.3 Colocalisation with external QTL datasets at the *ADCY3* locus 83
3.4 Discussion . . . 91

4 Transcriptomic associations with anti-TNF drug response in Crohn's disease patients 95
4.1 Introduction . . . 95
4.1.1 Crohn's disease and inflammatory bowel disease . . . 95
4.1.2 Anti-TNF therapies for Crohn's disease . . . 96
4.1.3 Anti-TNF treatment failure . . . 97
4.1.4 Predicting patient response to anti-TNFs . . . 97
4.1.5 Chapter summary . . . 98
4.2 Methods . . . 99
4.2.1 The Personalised Anti-TNF Therapy in Crohn's Disease (PANTS) cohort 99
4.2.2 Definition of timepoints . . . 99
4.2.3 Definition of primary response and primary non-response . . . 100
4.2.4 Library preparation and RNA-seq . . . 101
4.2.5 RNA-seq quantification and preprocessing . . . 102
4.2.6 Differential gene expression . . . 102
4.2.6.1 Variable selection by variance components analysis . . . 102
4.2.6.2 Contrasts for pairwise group comparisons . . . 106
4.2.6.3 Spline model of expression over time . . . 107
4.2.6.4 Clustering expression over all timepoints . . . 108
4.2.6.5 Gene set enrichment analyses . . . 108
4.2.7 Genotyping and genotype data preprocessing . . . 108
4.2.8 reQTL mapping . . . 109
4.2.8.1 Computing genotype principal components . . . 109
4.2.8.2 Finding hidden confounders in expression data . . . 109
4.2.8.3 Computing kinship matrices . . . 109

4.2.8.4 Mapping eQTLs per timepoint . . . 111
4.2.8.5 Joint reQTL mapping over all timepoints . . . 111
4.3 Results . . . 112
4.3.1 Longitudinal RNA-seq data from the PANTS cohort . . . 112
4.3.2 Baseline gene expression associated with primary response . . . 112
4.3.3 Assessing previously reported baseline predictors of primary response . . . 114
4.3.4 Post-induction gene expression associated with primary response . . . 118
4.3.5 Magnification of expression changes from baseline to post-induction in responders . . . 119
4.3.6 Interferon modules with opposing expression changes in responders and non-responders . . . 119
4.3.7 Sustained expression differences between primary responders and non-responders during maintenance . . . 123
4.3.8 Limited evidence for changes in genetic architecture of gene expression over time . . . 127
4.4 Discussion . . . 133

5 Discussion **141**
5.1 Strategies for detecting robust associations . . . 141
5.2 Responder analysis . . . 143
5.3 Challenges in the interpretation of bulk expression data . . . 144
5.4 From association to prediction . . . 146
5.5 From association to causality . . . 147
5.6 Triangulation . . . 149
5.7 Concluding remarks . . . 149

Bibliography **151**

List of Abbreviations **185**

List of Figures

1.1 The mosaic structure of human genetic variation. . . . 3
1.2 Effect size and allele frequency of trait-associated genetic variants. . . . 4
1.3 Types of tissue-dependent *cis*-expression quantitative trait locus (eQTL) effects. 8
1.4 High-throughput technologies for systems immunology. . . . 12

2.1 Overview of Human Immune Response Dynamics (HIRD) study data. . . . 23
2.2 Antibody titre data and responder definitions. . . . 24
2.3 Distribution of patient titre response indexs (TRIs), stratified by expression measurement platform. . . . 27
2.4 Sample filters for marker missingness and marker heterozygosity rate. . . . 28
2.5 HIRD samples projected onto principal component (PC) axes defined by principal component analysis (PCA) of HapMap 3 samples. . . . 30
2.6 FastQC per-base sequence quality (Phred scores) versus read position for RNA sequencing (RNA-seq) samples. . . . 31
2.7 FastQC per-read GC distributions for RNA-seq samples. . . . 31
2.8 Distributions of removed short non-coding RNA (ncRNA) and globin counts as a proportion of total counts in RNA-seq samples. . . . 33
2.9 Distribution of the proportion of samples in which genes were detected (non-zero expression). . . . 33
2.10 Distributions of gene expression for RNA-seq samples before and after filtering low expression and non-detected genes. . . . 34
2.11 Distribution of raw foreground intensities for HIRD array samples ($n = 173$). . . . 34
2.12 Distribution of per-sample expression estimates after normalisation and collapsing of probes to genes. . . . 36
2.13 First four standardised PCs in the expression data, colored by array batch/RNA-seq pool (a, c), timepoint (b, d), and binary response status (e, f). . . . 37
2.14 Gamma prior for τ (blue) used for `bayesmeta` analyses of the day 1 versus baseline effect, compared to the empirical distribution of per-gene frequentist `metafor::rma.uni` estimates for τ. . . . 42
2.15 Normal prior for μ (blue) used for `bayesmeta` analyses of the day 1 versus baseline differential gene expression (DGE) effect, compared to the empirical distribution of per-gene frequentist `metafor::rma.uni` estimates for τ. . . . 42

2.16 Normalised gene expression for 857 genes differentially expressed between any pair of timepoints (lfsr < 0.05, $|\text{FC}| > 1.5$). 44
2.17 Transcriptomic modules up or downregulated between pairs of timepoints. . . . 46
2.18 DGE effect sizes ($\log_2$ FC) estimated in array versus RNA-seq samples, colored by significance in frequentist random effects meta-analysis using `rma.uni` at BH FDR < 0.05. 48
2.19 DGE effect sizes ($\log_2$ FC) estimated in array versus RNA-seq samples, colored by significance in Bayesian random effects meta-analysis using `bayesmeta` at `ashr` LFSR < 0.05. 49
2.20 Estimates of between-platform heterogeneity τ from frequentist and Bayesian meta-analysis, for the 58 genes with a significant association between day 7 expression and binary responder/non-responder status in Sobolev *et al.* [162]. 50
2.21 Gene expression modules associated with antibody response (TRI). 51

3.1 Simulating the effect of data transformation on response expression quantitative trait locus (reQTL) effects. 62
3.2 Correlation matrix of standardised `xCell` cell type enrichment scores in HIRD array and RNA-seq datasets. 65
3.3 Contribution of each cell type score to each PC dimension after PCA of standardised `xCell` cell type enrichment scores. 66
3.4 Comparison of standardised `xCell` scores with normalised HIRD fluorescence-activated cell sorting (FACS) measurements, for monocytes, natural killer (NK) cells, and plasma cells. 68
3.5 Correlation of known variables to the first 25 PEER factors estimated from the array and RNA-seq mega-analysis expression data at baseline. 69
3.6 Number of significant genes with an eQTL detected on chromosome 1 as a function of the number of PEER factors included as covariates. 71
3.7 Replication rate π_1 of HIRD eQTLs in GTEx whole blood eQTL reference data. 74
3.8 Summary of HIRD eQTL mapping from mega-analysis of array and RNA-seq expression data, binned by patterns of lead variant significance over the three timepoints. 78
3.9 eGenes where the lead eQTL was a reQTL between a pair of timepoints. 79
3.10 z-statistic for difference in beta post-vaccination versus baseline for shared and reQTLs, against distance from the eGene transcription start site (TSS). 80
3.11 Expression and lead eQTL of *ADCY3* over study timepoints. 82
3.12 Expression and lead eQTL of *SH2D4A* over study timepoints. 83
3.13 Effect of estimated monocyte abundance on *ADCY3* expression at baseline, stratified by genotype at a day 1 *ADCY3* reQTL. 84
3.14 Effect of estimated monocyte abundance on *ADCY3* expression at day 1, stratified by genotype at a day 1 *ADCY3* reQTL. 85
3.15 Expression of *ADCY3* in sorted immune cell subsets. 87
3.16 Sensitivity analysis for multi-trait colocalisation at the *ADCY3* locus. 89
3.17 Multi-trait colocalisation at the *ADCY3* locus. 90

4.1 Sample size and study day distribution for Personalised Anti-TNF Therapy in Crohn's Disease (PANTS) study RNA-seq samples, stratified by timepoint and study group. . . . 101
4.2 Distribution of RNA-seq samples from each patient among timepoints. . . . 103
4.3 Correlation matrix of variables measured in PANTS that were considered as potential predictor variables. . . . 104
4.4 Variance components analysis showing the distribution of per-gene percentage of variance in expression explained by each variable. . . . 106
4.5 1000 Genomes Project (1000G) samples and PANTS samples projected onto 1000G genotype PC1 and PC2 axes, colored by (a) superpopulation and (b) population. 110
4.6 Number of eGenes on chromosome 1 vs. number of PEER factors included in eQTL mapping as covariates. . . . 111
4.7 Volcano plots of DGE between primary responders (PR) and non-responders at week 0; unadjusted (top row) and adjusted (bottom row) for cell composition; for infliximab (IFX), adalimumab (ADA), or with both drugs pooled. . . . 115
4.8 Top modules differentially expressed between primary responders (PR) and non-responders at week 0, unadjusted for cell composition. . . . 116
4.9 Top modules differentially expressed between primary responders (PR) and non-responders at week 0, adjusted for cell composition. . . . 117
4.10 Volcano plots of DGE between primary responders (PR) and non-responders at week 14; unadjusted (top row) and adjusted (bottom row) for cell composition; for infliximab (IFX), adalimumab (ADA), or with both drugs pooled. . . . 120
4.11 Top modules differentially expressed between primary responders (PR) and non-responders at week 14, unadjusted for cell composition. . . . 121
4.12 Top modules differentially expressed between primary responders (PR) and non-responders at week 14, adjusted for cell composition. . . . 122
4.13 Expression $\log_2$ FC from week 0 to week 14 in primary responders (PR) versus non-responders (PNR), for genes that differentially expressed from week 0 to week 14 in both responders and non-responders, with a significantly different effect size between responders and non-responders. . . . 123
4.14 Top modules differentially expressed between week 14 and week 0, unadjusted for cell composition. . . . 124
4.15 Top modules differentially expressed between week 14 and week 0, adjusted for cell composition. . . . 125
4.16 `tmod` evidence plots showing interferon-related modules specifically upregulated from week 0 to week 14 in primary non-responders (PNR), but not in primary responders (PR). . . . 126
4.17 Gap statistic versus cluster number k. . . . 128
4.18 Normalised expression over the timepoints for genes in the six identified clusters. 129

4.19 Overlap of genes differentially expressed between responders and non-responders from the spline model, the week 14 responder versus non-responder contrast (w14), and the interaction between week 0 to week 14 change and response status contrast (w14 – w0). 130
4.20 Post-treatment eQTL effect sizes at week 30 and week 54 compared to baseline effect sizes at week 0. 132

5.1 The three assumptions of Mendelian randomisation (MR). 149

List of Tables

2.1 Descriptive statistics for Human Immune Response Dynamics (HIRD) individuals with both expression and antibody data. 26
2.2 Distribution of HIRD samples among timepoint and responder groups in the array batches and RNA sequencing (RNA-seq) pools. 39

4.1 Patient characteristics for the Personalised Anti-TNF Therapy in Crohn's Disease (PANTS) RNA-seq subcohort. 113