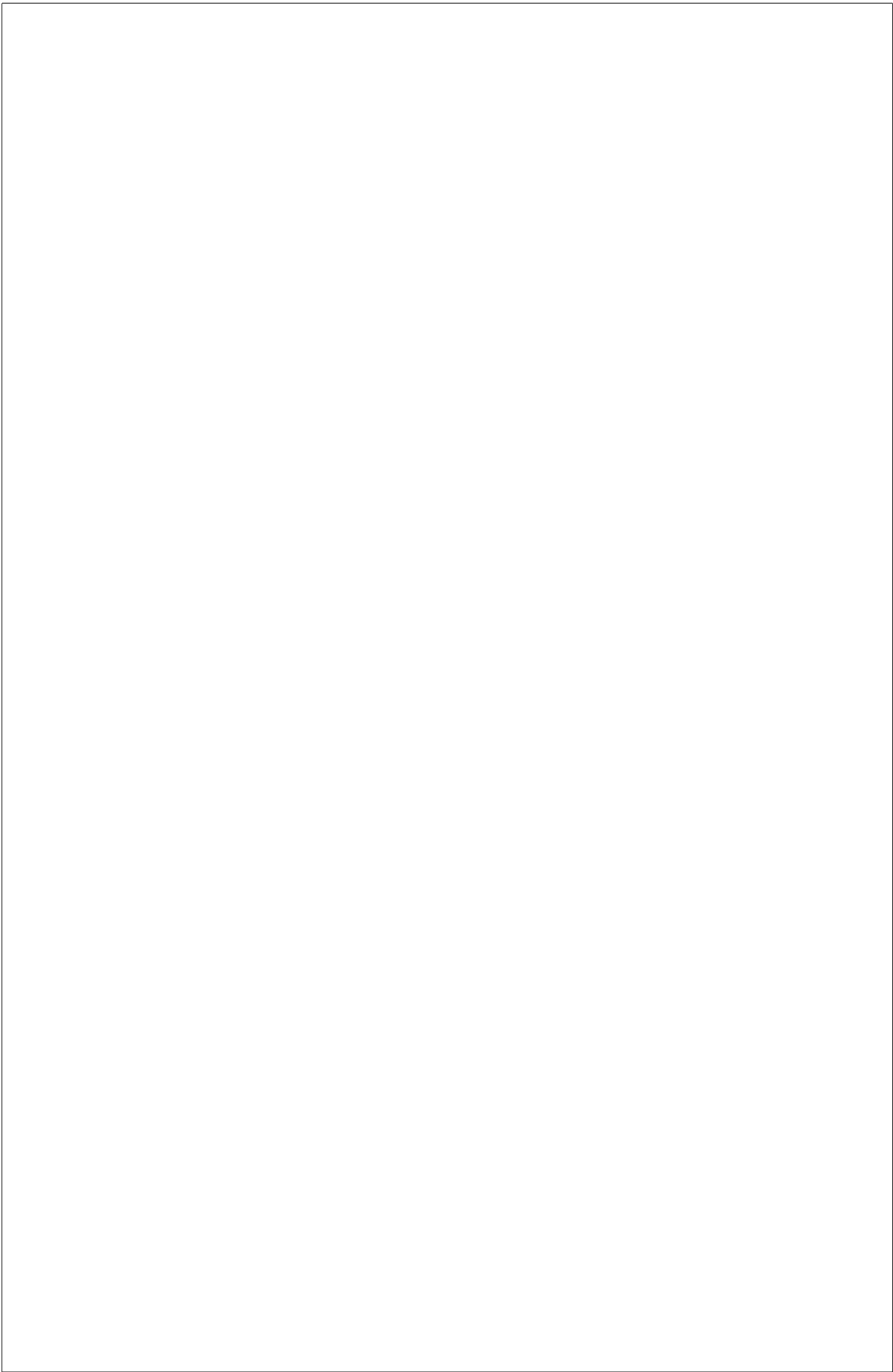

Genetic mapping of transcriptomic response to *in vivo* immune
perturbations

Benjamin Yu Hang Bai

2020-11-10 00:12:12Z



Abstract

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 success of vaccination in controlling influenza is indisputable, but it is not completely 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 same cohort to investigate regulation of transcriptomic response by common genetic variants. Rather than driving differential expression post-vaccination, the strongest reQTL appear 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 was driven 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 Crohn's disease (CD) patient response to 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 significant heterogeneity in 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 up to week

I'm trying to say that the response to immune perturbation is heritable/influenced by genetics.

I prefer associations over signatures. Signatures should have had their predictive power quantified.

54. A set 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.

Contents

| | |
|---|-----------|
| List of Figures | ix |
| List of Tables | xv |
| 1 Introduction | 1 |
| 1.1 Genetic association studies of complex traits | 1 |
| 1.1.1 Structure and variation of the genome | 1 |
| 1.1.2 Lessons from the past fifteen years | 2 |
| 1.1.3 From complex trait to locus | 4 |
| 1.1.4 From locus to causal variant | 4 |
| 1.1.5 From causal variant to target gene | 6 |
| 1.2 Gene expression as an intermediate molecular phenotype | 6 |
| 1.2.1 Regulation of gene expression | 6 |
| 1.2.2 Expression is a complex trait | 7 |
| 1.2.3 Context-dependent eQTL | 7 |
| 1.2.4 Immune Response expression quantitative trait loci | 8 |
| 1.3 Phenotypes of immune response | 10 |
| 1.3.1 Intro to immunology | 10 |
| 1.3.2 High-throughput immunology | 10 |
| 1.3.3 Immune response to vaccination | 11 |
| 1.3.4 Immune response to anti-TNF biologics | 12 |
| 1.4 Thesis overview | 12 |
| 2 Transcriptomic response to Pandemrix vaccine | 15 |
| 2.1 Introduction | 15 |
| 2.1.1 Seasonal and pandemic influenza | 15 |
| 2.1.2 Quantifying immune response to influenza vaccines | 15 |
| 2.1.3 Systems vaccinology of influenza vaccines | 16 |
| 2.1.4 The Human Immune Response Dynamics (HIRD) study | 16 |
| 2.1.5 Chapter summary | 17 |
| 2.2 Methods | 17 |
| 2.2.1 Existing HIRD study data and additional data | 17 |
| 2.2.2 Computing baseline-adjusted measures of antibody response | 18 |
| 2.2.3 Genotype data generation | 21 |

| | | |
|----------|---|-----------|
| 2.2.4 | Genotype data preprocessing | 21 |
| 2.2.5 | Computing genotype principal components as covariates for ancestry . . . | 22 |
| 2.2.6 | RNA-seq data generation | 22 |
| 2.2.7 | RNA-seq quantification and filtering | 22 |
| 2.2.8 | Array data preprocessing | 26 |
| 2.2.9 | Differential gene expression | 30 |
| 2.2.9.1 | Per-platform differential gene expression model | 32 |
| 2.2.9.2 | Choice of differential gene expression meta-analysis method . . . | 32 |
| 2.2.9.3 | Prior for between-studies heterogeneity | 33 |
| 2.2.9.4 | Prior for effect size | 33 |
| 2.2.9.5 | Evaluation of priors | 33 |
| 2.2.9.6 | Multiple testing correction | 35 |
| 2.2.10 | Gene set enrichment analysis using blood transcription modules | 35 |
| 2.3 | Results | 35 |
| 2.3.1 | Extensive global changes in expression after vaccination | 35 |
| 2.3.2 | Innate immune response at day 1 post-vaccination | 35 |
| 2.3.3 | Adaptive immune response at day 7 post-vaccination | 37 |
| 2.3.4 | Expression signatures associated with antibody response | 37 |
| 2.3.5 | Identifying expression signatures for predicting antibody response [prob- ably cut this section and just add to discussion | 37 |
| 2.4 | Discussion | 41 |
| 3 | Genetic architecture of transcriptomic response to Pandemrix vaccine | 43 |
| 3.1 | Introduction | 43 |
| 3.1.1 | Genetic factors affecting influenza vaccine response | 43 |
| 3.1.2 | Response expression quantitative trait loci for seasonal influenza vaccination | 44 |
| 3.1.3 | Chapter summary | 44 |
| 3.2 | Methods | 44 |
| 3.2.1 | Genotype phasing and imputation | 44 |
| 3.2.2 | Overall strategy for detecting reQTLs | 45 |
| 3.2.3 | Controlling for population structure with linear mixed models | 46 |
| 3.2.3.1 | Estimation of kinship matrices | 46 |
| 3.2.4 | Additional eQTL-specific expression preprocessing | 46 |
| 3.2.5 | Estimation of cell type abundance from expression | 47 |
| 3.2.6 | Finding hidden covariates using factor analysis | 54 |
| 3.2.7 | eQTL mapping per timepoint | 54 |
| 3.2.8 | Joint eQTL analysis across timepoints | 56 |
| 3.2.9 | Defining shared and response eQTLs | 56 |
| 3.2.10 | Replication of eQTLs in a reference dataset | 57 |
| 3.2.11 | Genotype interactions with cell type abundance | 59 |
| 3.2.12 | TODO Statistical colocalisation | 59 |
| 3.3 | Results | 59 |
| 3.3.1 | Mapping reQTLs to Pandemrix vaccination | 59 |

| | | |
|----------|---|-----------|
| 3.3.2 | Characterising reQTLs post-vaccination | 61 |
| 3.3.3 | Genotype by cell type interaction effects | 61 |
| 3.3.4 | TODO Genotype by platform interaction effects | 65 |
| 3.3.5 | TODO Colocalisation of reQTLs with known <i>in vitro</i> condition-specific immune eQTLs | 65 |
| 3.4 | Discussion | 66 |
| 4 | Transcriptomic associations with anti-TNF drug response in Crohn's disease patients | 71 |
| 4.1 | Introduction | 71 |
| 4.1.1 | Crohn's disease and inflammatory bowel disease | 71 |
| 4.1.2 | Anti-TNF therapies for Crohn's disease | 72 |
| 4.1.3 | Anti-TNF treatment failure | 72 |
| 4.1.4 | Predicting response to anti-TNFs | 73 |
| 4.1.5 | Chapter summary | 74 |
| 4.2 | Methods | 74 |
| 4.2.1 | The PANTS cohort | 74 |
| 4.2.2 | Definition of timepoints | 75 |
| 4.2.3 | Definition of primary response and non-response | 75 |
| 4.2.4 | Library preparation and RNA-seq | 76 |
| 4.2.5 | RNAseq quantification and preprocessing | 76 |
| 4.2.6 | Differential gene expression | 77 |
| 4.2.6.1 | Variable selection by variance components analysis | 77 |
| 4.2.6.2 | Contrasts for pairwise group comparisons | 80 |
| 4.2.6.3 | Spline model of expression over time | 81 |
| 4.2.6.4 | Clustering expression over all timepoints | 81 |
| 4.2.6.5 | Rank-based gene set enrichment on gene modules | 82 |
| 4.2.7 | Genotyping and genotype data preprocessing | 82 |
| 4.2.8 | Response expression quantitative trait locus mapping | 83 |
| 4.2.8.1 | Computing genotype PCs | 83 |
| 4.2.8.2 | Finding hidden confounders in expression data | 83 |
| 4.2.8.3 | Computing kinship matrices | 83 |
| 4.2.8.4 | Mapping cis-eQTLs in each timepoint separately | 83 |
| 4.2.8.5 | Joint reQTLs mapping over all timepoints | 85 |
| 4.3 | Results | 86 |
| 4.3.1 | Longitudinal RNA-seq data from the Personalised Anti-TNF Therapy in Crohn's Disease (PANTS) cohort | 86 |
| 4.3.2 | Baseline gene expression associated with primary response | 86 |
| 4.3.3 | Assessing previously reported baseline predictors of primary response | 88 |
| 4.3.4 | Post-induction gene expression associated with primary response | 92 |
| 4.3.5 | Magnification of expression change from baseline to post-induction in re- sponders | 92 |

| | | |
|----------|---|------------|
| 4.3.6 | Interferon modules with opposing differential expression in responders and non-responders | 96 |
| 4.3.7 | Sustained expression differences between primary responders and non-responders during maintenance | 96 |
| 4.3.8 | Limited evidence for change in genetic architecture of gene expression over time | 104 |
| 4.4 | Discussion | 104 |
| 5 | Discussion | 115 |
| 5.1 | Increasing the power to detect associations | 115 |
| 5.2 | Responder analysis | 117 |
| 5.3 | Challenges in the interpretation of bulk expression data | 118 |
| 5.4 | From association to prediction | 121 |
| 5.5 | From association to causality | 122 |
| 5.6 | Triangulation | 124 |
| | Bibliography | 127 |
| | List of Abbreviations | 147 |

List of Figures

| | | |
|-----|--|----|
| 1.1 | Information flow through the biological system. | 2 |
| 1.2 | The genomic mosaic: block-like linkage disequilibrium (LD) structure of the genome | 3 |
| 1.3 | The reach of GWAS. OR vs MAF ala tam2019BenefitsLimitationsGenomewide, extended by imputation, sample size, WGS based genotypes, but may be indistinguishable from noise at the limits | 5 |
| 1.4 | eqtl mech models: magnify, dampen, flip | 9 |
| 2.1 | Data types, timepoints, and sample sizes. Individuals were vaccinated after day 0 sampling. Antibodies to the vaccine strain were measured by haemagglutination inhibition (HAI) and microneutralisation (MN) assays. Array and RNA-sequencing (RNA-seq) gene expression measured in the peripheral blood mononuclear cell (PBMC) compartment. | 18 |
| 2.2 | Comparison of titre response index (TRI) to HAI (left column) and MN (right column) titres and binary responder/non-responder status (colored) in 166 Human Immune Response Dynamics (HIRD) individuals. Row 1: baseline titres are positively correlated to post-vaccination titres. Row 2: baseline titres are negatively correlated to fold change. Row 3: TRI regresses out the correlation between baseline titre and response. Row 4: TRI is still comparable in ordering to binary response status. | 19 |
| 2.3 | Distribution of TRI, stratified by platform used to measure expression. | 20 |
| 2.4 | Sample filters for missingness and heterozygosity rate. Samples outside the central rectangle were excluded. | 23 |
| 2.5 | HIRD samples (cyan) projected onto principal component (PC)1 and PC2 axes defined by principal component analysis (PCA) of HapMap 3 samples. The first two PCs separate European (CEU, upper-right) from Asian (CHB and JPT, lower-right) and African (YRI, lower-left) individuals. | 24 |
| 2.6 | FastQC sequence quality versus read position for HIRD RNA-seq samples. | 25 |
| 2.7 | FastQC sequence duplication levels for HIRD RNA-seq samples. | 25 |
| 2.8 | FastQC GC profile for HIRD RNA-seq samples. | 25 |
| 2.9 | Distributions of removed short ncRNA and globin counts as a proportion of total counts in RNA-seq samples. | 27 |

| | | |
|------|--|----|
| 2.10 | Distribution of the proportion of samples in which genes were detected (non-zero expression). Many genes are not detected in any samples. Vertical line shows 5% threshold below which genes were discarded. | 27 |
| 2.11 | Distribution of gene expressions for RNA-seq samples before and after filtering no expression and low expression genes. Vertical line shown at counts per million (CPM) = 0.5 threshold. | 28 |
| 2.12 | Raw foreground intensities for 173 HIRD array samples. Colored by array processing batch. | 28 |
| 2.13 | Array intensity estimates after VSN normalisation and collapsing of probes to genes. Colored by array processing batch. | 29 |
| 2.14 | First four PCs in the HIRD expression data, colored by platform and batch (left), and timepoint (right). | 31 |
| 2.15 | Gamma prior for τ used for bayesmeta (blue), compared to the empirical distribution of per-gene frequentist metafor::rma estimates for τ , for the day 1 vs. baseline effect (small estimates of $\tau < 0.01$ excluded). Empirical log-normal fit also shown (red). | 34 |
| 2.16 | Normal prior for μ used for bayesmeta (blue), compared to the empirical distribution of per-gene frequentist metafor::rma estimates for τ , for the day 1 vs. baseline effect. The non-scaled normal fit is shown (black), as well as a Cauchy fit (red). | 34 |
| 2.17 | Normalised gene expression for genes differentially expressed between any pair of timepoints ($\text{lfsr} < 0.05$, absolute fold change > 1.5) across HIRD samples, clustered by gene (Manhattan distance metric). | 36 |
| 2.18 | Transcriptomic modules significantly up or downregulated post-vaccination. Size of circle indicates effect size. Color of circle indicates significance and direction of effect (red = upregulation, blue = downregulation). | 38 |
| 2.19 | DGE effect sizes estimated in array vs. RNA-seq. Significance colored by frequentist random effects meta-analysis $\text{FDR} < 0.05$. Genes with day 7 expression associated with responder/non-responder status in [106] are circled for that contrast. | 39 |
| 2.20 | DGE effect sizes estimated in array vs RNA-seq. Significance colored by Bayesian random effects meta-analysis $\text{lfsr} < 0.05$. Genes with day 7 expression associated with responder/non-responder status in [106] are circled for that contrast. | 39 |
| 2.21 | Transcriptomic modules enriched in genes with expression associated with antibody response (TRI) at each day. Size of circle indicates effect size. Color of circle indicates significance and direction of effect (red = expression positively correlated with TRI, blue = negative). | 40 |

| | | |
|------|--|----|
| 3.1 | Simulated log scale expression in two conditions for six genes (columns) representing six different scenarios: Scenario 0 has no expression quantitative trait locus (eQTL), scenario 1 is a shared eQTL ($\beta = 1$), scenario 2 is a response expression quantitative trait locus (reQTL) where β increases from 0 to 1, scenario 3 is a reQTL where β increases from 0 to 2, scenario 4 is a reQTL where β increases from 1 to 2, and scenario 6 is a reQTL where β increases from 1 to 4. Rows represent the effect of different expression transformations across samples, conducted both within condition, and including both conditions. | 48 |
| 3.2 | Standardised xCell enrichment scores for seven PBMC cell types in array samples. | 50 |
| 3.3 | Standardised xCell enrichment scores for seven PBMC cell types in RNA-seq samples. | 51 |
| 3.4 | Quality of representation (\cos^2) for each input variable in each PC dimension after PCA of xCell scores. Higher \cos^2 represents higher contribution of that variable to that dimension. | 52 |
| 3.5 | Correlation between standardised xCell scores and normalised fluorescence-activated cell sorting (FACS) measurements for a similar immune subset, in the subset of individuals with FACS data. | 53 |
| 3.6 | Correlation of PEER factors to known factors and other possible covariates. Note that PEER factors are not constrained to be orthogonal, so correlations to known factors are expected. | 55 |
| 3.7 | Number of significant eGenes detected on chromosome 1 (hierarchical Bonferroni-Benjamini-Hochberg (BH)[197] $FDR < 0.05$) as a function of the number of PEER factors included as covariates k | 57 |
| 3.8 | Clustering of within-timepoint Z scores in the strong mashr subset (random sample of 10000/45962 tests), confirming the presence of strong condition-specific effects. | 58 |
| 3.9 | Effect of HIRD lfsr threshold on GTEx whole blood replication rate (π_1), number of p -values used to compute π_1 , and maximum p -value among those p -values; for shared and reQTL called from the array-only, RNA-seq-only and mega-analysis pipelines. Shaded region for π_1 represents the 5th-95th percentile range of 1000 bootstraps. | 60 |
| 3.10 | Summary of eQTL mapping results at 13570 genes-lead eQTL pairs, with intersections based on significance ($lfsr < 0.05$). Counts of shared eQTLs and reQTLs; and distribution of INFO score, min MAF across timepoints, and max PVE across timepoints for those lead variants are shown above each intersection. | 62 |
| 3.11 | Z score for difference in effect vs. day 0, of lead eQTLs for all eGenes significant at either day 1 or day 7; versus distance of the lead SNP to the TSS. Direction of effect is aligned so that the β at day 0 is positive. Points with positive z score are magnified effects post-vaccination, points with negative z scores are dampening and opposite sign effects. | 63 |
| 3.12 | <i>ADCY3</i> , strongest reQTL at day 1. | 64 |

| | | |
|------|---|----|
| 3.13 | <i>SH2D4A</i> , strongest reQTL at day 7. Top: Array and RNA-seq expression before merging with ComBat for mega-analysis. Bottom: eQTL effects at each timepoint condition in the mega-analysis. | 64 |
| 3.14 | Multi-trait colocalisation of HIRD reQTL signal at ADCY3 (500 Kb window), with QTL studies from IHEC, BLUEPRINT, eQTL Catalogue, and GWAS Catalogue. Plots are colored by colocalised cluster. Black indicates non-colocalised datasets. | 67 |
| 4.1 | Correlation matrix of variables measured in Personalised Anti-TNF Therapy in Crohn's Disease (PANTS) that were considered as potential independent variables. | 78 |
| 4.2 | Variance partition analysis showing distribution of gene-wise percentage of variance in expression explained by each variable. "PANTS.ID" is patient ID. | 80 |
| 4.3 | 1000G samples and PANTS samples projected onto 1000G genotype PC1 and PC2 axes, colored by superpopulation (a) and population (b). | 84 |
| 4.4 | Number of eGenes on chr1 vs number of PEER factors included in eQTL mapping. The number of PEER factors for each timepoint is chosen after the plateaus. | 85 |
| 4.5 | Number and distribution of RNA-seq samples in experimental groups over time. Windows for the four major visits are colored in grey. Samples mostly come from major visits, but a small number of loss of response (LOR) and exit visit samples were included according to subsection 4.2.2. | 87 |
| 4.6 | Distribution of RNA-seq samples from each patient among timepoints. | 87 |
| 4.7 | Volcano plots of differential gene expression (DGE) between primary responders 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. Annotated genes include significant associations from this study and previously reported associations from subsection 4.1.4. Dashed line shows significance threshold at FDR = 0.05. | 89 |
| 4.8 | Top modules differentially expressed between primary responders and non-responders at week 0, unadjusted for cell composition. Columns correspond to results from infliximab (IFX), adalimumab (ADA), infliximab-adalimumab difference and pooled analyses. The top 30 modules ranked by minimum false discovery rate (FDR) in any column are shown. Vertical dashed line shows significance threshold at FDR = 0.05. | 90 |
| 4.9 | Top modules differentially expressed between primary responders and non-responders at week 0, adjusted for cell composition. | 91 |
| 4.10 | Volcano plots of DGE between primary responders 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. Annotated genes include significant associations from this study and previously reported associations from subsection 4.1.4. Dashed line shows significance threshold at FDR = 0.05. | 93 |

| | | |
|------|---|-----|
| 4.11 | Top modules differentially expressed between primary responders and non-responders at week 14, unadjusted for cell composition. | 94 |
| 4.12 | Top modules differentially expressed between primary responders and non-responders at week 14, adjusted for cell composition. | 95 |
| 4.13 | Expression logFC from week 0 to week 14 in primary responders versus non-responders, shown for all genes differentially expressed from week 0 to week 14 in both responders and non-responders, with a significantly different effect size in responders and non-responders. Identity line shown by dashed line. | 97 |
| 4.14 | Top modules differentially expressed between week 14 and week 0, unadjusted for cell composition. Columns show effects in primary responders (PR), non-responders (PNR), and the primary responder minus non-responder difference. The top 30 modules ranked by minimum FDR in any column are shown. Vertical dashed line shows significance threshold at $FDR = 0.05$ | 98 |
| 4.15 | Top modules differentially expressed between week 14 and week 0, adjusted for cell composition. Columns show effects in primary responders (PR), non-responders (PNR), and the primary responder minus non-responder difference. The top 30 modules ranked by minimum FDR in any column are shown. Vertical dashed line shows significance threshold at $FDR = 0.05$ | 99 |
| 4.16 | tmod evidence plots showing interferon-related modules specifically upregulated from week 0 to week 14 in primary non-responders. Genes were ranked in ascending order by week 14 versus week 0 DGE z-statistic. The ranks of genes in interferon-related modules are indicated by colored rug plots. Colored curves show the cumulative fraction of genes in each module. For non-responders, these modules are enriched for large ranks (large, positive z-statistics). The area between the colored curves and the null of randomly-distributed ranks (grey diagonal line) is the effect size (AUC). | 100 |
| 4.17 | Gap statistic versus cluster number k. Error bars derived from 500 bootstraps. The optimal number of clusters is defined as the smallest k after which the gap statistic does not significantly increase at k+1. | 101 |
| 4.18 | Normalised expression over the timepoints for genes in the six identified clusters. Error bars are the expression mean and standard deviation for the genes at each timepoint in primary responders and non-responders. | 102 |
| 4.19 | Venn diagrams showing which genes in each spline cluster were also significant in the week 14 responder versus non-responder contrast, or the interaction between week 0 to week 14 change and response status. | 103 |
| 4.20 | Week 30 and week 54 eQTL effect sizes vs baseline. Significant reQTLs at FDR 0.05 are labelled. | 105 |

List of Tables

| | | |
|-----|--|-----|
| 2.1 | Sample descriptive statistics. | 21 |
| 2.2 | HIRD batch balance | 30 |
| 4.1 | Table caption. | 113 |

Chapter 1

Introduction

Observable human characteristics or traits are called phenotypes. Variation in phenotype emerges from the interplay of genetics, environment and pure chance. The contributions of each vary from phenotype to phenotype. Traits for which genetic variation explains a non-zero fraction of phenotypic variation are heritable. Virtually all phenotypic traits are heritable to some degree, and twin studies provide upper bounds on this heritability by partitioning phenotypic variation into genetic and environmental components [1].

Genetic variation presents a unique opportunity to probe the causal molecular mechanisms underlying phenotypes. A mainstay of the field of human genetics is uncovering the specific genetic variants that contribute to the heritability of phenotypes through statistical association of variants and phenotypes. Information encoded in the genome can have phenotypic consequences only after flowing through multiple molecular layers (Fig. 1.1). This guiding principle is the central dogma, whereby the flow is directed from DNA to RNA to protein via transcription and translation. Barring somatic mutation, an individual's genome is fixed at conception, thus providing a causally upstream anchor that can be measured with relatively little error. Although not immune to population-level confounders like stratification, genetic association has intrinsic resistance to reverse causality, an issue that permeates observational studies on the causes of human phenotypes.

1.1 Genetic association studies of complex traits

1.1.1 Structure and variation of the genome

The human genome is almost three billion bp (base pairs) in length, containing 20 000–25 000 protein-coding genes that span 1–3 % of its length, with the remaining sequence being non-coding [2, 3]. Each diploid cell contains two copies of the genome, organised into 46 chromosomes comprised of 23 maternal-parental pairs: 22 pairs of homologous autosomes and one pair of sex chromosomes. Variation in the genome between individuals in a population exists in the form of single nucleotide polymorphisms (SNPs), short indels, and structural variants. For common population variants with minor allele frequency (MAF) >1–5 %, the vast majority (>99.9 %) are SNPs and short indels [3]. On average, a pair of genomes differs by one SNP per 1000–2000 bp

I agree the intro is built around genetics. I have tried to keep the genetics parts concise while elaborating on gene expression more in the later two sections.

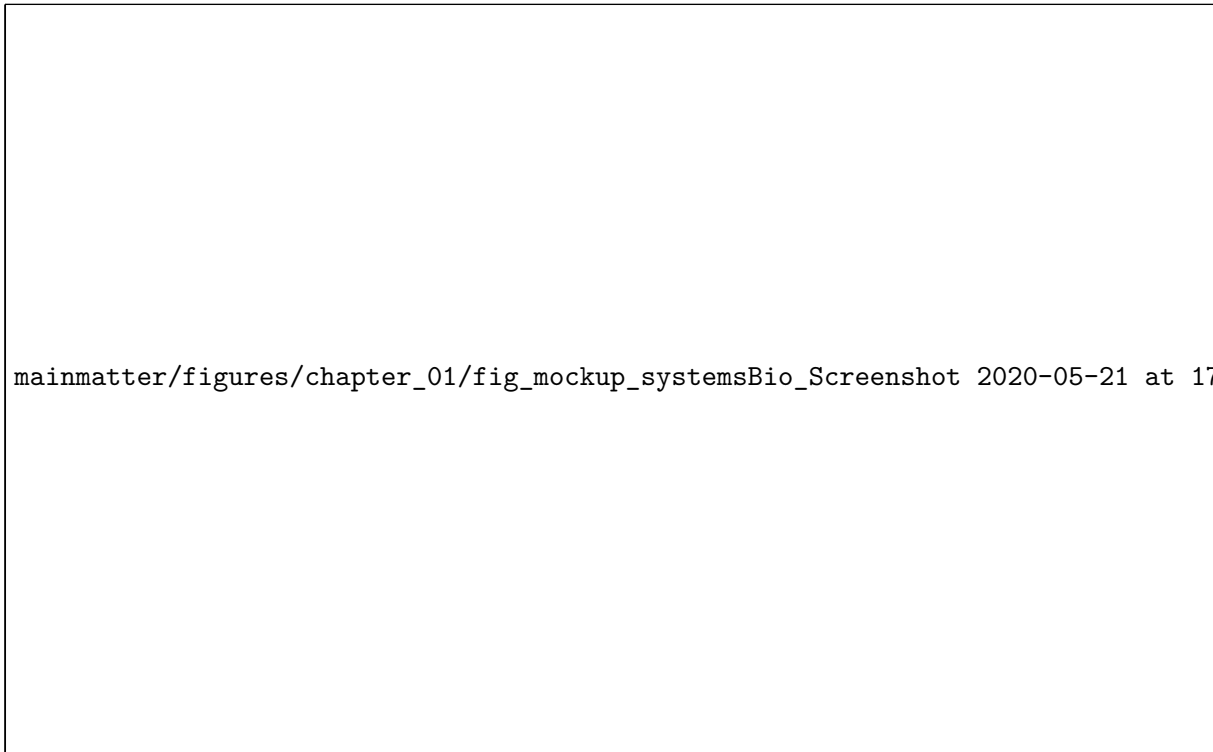


Figure 1.1: Information flow through the biological system.

[4]. Each version of a variant is called an allele; each individual has a maternal and parental allele at each variant.

The large number of variants in a population are inherited in a smaller number of haplotypes: contiguous stretches of the genome passed through generations via meiotic segregation. The fundamental sources of genetic diversity are mutation and meiotic recombination, generating new alleles and breaking apart haplotypes into shorter ones over evolutionary time. Variants that are physically close on a chromosome are less likely to flank a recombination event, hence more likely to cosegregate on the same haplotype (genetic linkage). Genetic linkage is one source of linkage disequilibrium (LD): the non-random association of alleles at two variants, differing from expectation based on their population frequencies and the law of independent assortment. LD can be quantified by r^2 , the squared correlation coefficient between alleles in that specific population [5].

Recombination events are not distributed uniformly throughout the genome. The genome is a mosaic of haplotype blocks delimited by recombination hotspots, characterised by strong LD within blocks, and little LD between blocks [6, 7] (Fig. 1.2). The structure of correlated haplotypes reflects a population’s unique evolutionary history, and can be used to trace the demography of populations back through time [8].

1.1.2 Lessons from the past fifteen years

Genetic variants affect heritable traits by impacting the function or regulation of target genes. How genetic variation contributes to a particular trait defines it’s genetic architecture: the number of genes affecting that trait; and the frequencies, effect sizes, and interactions of trait-



mainmatter/figures/chapter_01/fig_mockup_haplotypeBlocks_Screenshot 2020-05-21 at 17.08.33.png

Figure 1.2: The genomic mosaic: block-like LD structure of the genome

associated alleles [9]. The number of genes defines a spectrum of traits from monogenic (where inheritance follows simple Mendelian patterns) to polygenic (where inheritance is complex). Proposed architectures differ greatly among complex traits, but all have in common that the number of genes affecting complex trait is large (ranging from dozens to many thousands), thus the average effect size of each associated variant is small [10–12].

Since the 1980s, linkage analysis has been used to map the chromosomal position or region (loci) affecting traits by tracing the cosegregation of markers (variants with known position) with the trait in family pedigrees [13–15]. They were complemented by early genetic association studies, which largely focused on variants in or near candidate genes selected on the basis of prior biological knowledge [16]. These methods saw much success for Mendelian traits, but application to most complex traits proved challenging. Small average effect sizes meant penetrance was too low to reliably observe cosegregation problems in pedigrees [15]. Early candidate gene studies were also underpowered to detect small effects [17].

The past fifteen years have seen the rise of genome-wide association studies (GWAS) that systematically test common variants selected in a comparatively hypothesis-free manner across the whole genome (Fig. 1.3). Using large sample sizes to overcome small effects and large multiple testing burden, thousands of associations have been discovered for complex traits and diseases, many robustly replicated across populations [15, 18]. A number of take-home messages have emerged. Most genetic variance is additive, the contribution of dominance and epistatic interactions is small [9], and variants with effects on multiple phenotypes (pleiotropy) are widespread [15]. Even traits that are molecular rather than whole-organism phenotypes can be remarkably

polygenic, with hundreds to thousands of associated loci [19]. GWAS sample sizes in the millions are increasingly commonplace, and the discovery of new associations with ever smaller effects as sample sizes increase shows no sign of plateauing [20, 21].

1.1.3 From complex trait to locus

GWAS rely on the tendency of common variants on the same haplotype to be in strong LD. As the number of haplotypes is comparatively few, it is possible to select a subset of tag variants such that all other known common variants are within a certain LD threshold of that subset. In practice, there is enough redundancy that the number of variants measured on a modern genotyping array (in the order of 10^5 – 10^6) is sufficient to tag almost all common variants [22, 23]. Associations with unmeasured variants are indirectly detected through their strong correlation with a tag variant. Furthermore, as unrelated individuals still share short ancestral haplotypes, study samples can be assigned haplotypes from a panel of haplotypes derived from reference samples by matching on the directly genotyped variants. This process of genotype imputation allows ascertainment of many more variants not directly genotyped [24], but helps to recover rarer variants that are poorly-tagged [18]. Modern imputation panels enable cost-effective GWAS testing tens of millions of variants as rare as 0.0100–0.100 % in diverse populations [25].

Testing large numbers of variants incurs a massive multiple testing burden, but acknowledging the correlation between variants due to LD, there are only the equivalent of $\sim 10^6$ independent tests in the European genome, regardless of the number of tests actually performed [26]. The field has thus converged on a fixed discovery threshold of $0.05/10^6 = 5 \times 10^{-8}$ for genome-wide significance in European populations [27], akin to controlling the type I family-wise error rate (FWER) at $\alpha = 0.05$ using the Bonferroni correction*.

1.1.4 From locus to causal variant

By design, a significantly-associated variant from a GWAS needs not be a variant that causally affects the trait, and may only tag a causal variant. The resolution of the associated locus depends on the local LD structure. Fine-mapping is the process of determining which of the many correlated variants in an associated locus are most likely to be causal. The causal variants in a locus are not necessarily the ones with the strongest associations. State-of-the-art Bayesian fine-mapping methods take a variable selection approach, assigning each variant a posterior probability of causality. A credible set of variants likely to contain all causal variants in the locus with some probability can then be determined [29, 30]. The ability to separate causal and tag variants depends on factors like LD, sample size, and the effect size and number of causal variants [18, 29]. It is important that the causal variant is observed, by directly genotyped or confident imputation.

*The Bonferroni correction makes no assumptions about the dependence structure of the p -values, and is conservative (i.e. controls the FWER at a stricter level than the chosen α) even for independent tests. In fact it is always conservative unless the p -values have strong negative correlations [28].



1.1.5 From causal variant to target gene

Most causal variants for Mendelian traits are coding variants (nonsense, missense, or frameshift) that impact protein sequence [31]. In contrast, over 90 % of GWAS loci fall in non-coding regions [32], and often too far from the nearest coding region to be in LD [33]. Even if the causal variants at a GWAS locus are fine-mapped, it may not be obvious how to prioritise the target genes through which those variants affect the trait. A reasonable heuristic is to assign the nearest gene transcription start site (TSS) or body as the target, particularly for metabolite traits [34]. For improved accuracy across a variety of complex traits, integrative methods for gene prioritisation combine variant-to-gene distance with other metrics and data types drawn from numerous external sources [34–36].

1.2 Gene expression as an intermediate molecular phenotype

1.2.1 Regulation of gene expression

Gene regulation data are indispensable for gene prioritisation. Rather than directly impacting the coding sequence of a gene, many non-coding GWAS loci are hypothesised to affect traits by affecting the regulation of target gene expression [32, 37]. Unlike genotype, expression is context-dependent and dynamic across time and space. Diverse expression programs are responsible for the myriad of cell and tissue types generated during development, and enables adaptation in response to environmental stimuli.

Eukaryotic transcription is an intricate multi-step process involving interactions between DNA, RNA, and hundreds of proteins [38]. Transcription is initiated when RNA polymerase and transcription factors (TFs) form part of a protein complex around the promoter region and TSS at the 5' end of a gene. TFs can also bind to *cis*-regulatory elements such as enhancers, which can be distant from the gene, interacting with the promoter region via DNA looping. Transcription can only happen in regions of open chromatin, where the packing of DNA-histone complexes (nucleosomes) is loose enough that the DNA is physically-accessible to the transcriptional machinery. Chromatin accessibility is partially determined by histone modifications such as methylation, acetylation, phosphorylation, and ubiquitination [39]. During and post-transcription, introns are removed, and the remaining exons spliced together by spliceosomes that cut and rejoin the RNA at splice sites. Further post-transcriptional regulation of mature messenger RNAs (mRNAs) can occur through RNA editing and regulatory elements in 5' and 3' untranslated regions (UTRs).

In line with the regulatory hypothesis, GWAS variants are heavily enriched in regulatory elements annotated by functional genomics studies, including regions of open chromatin, histone binding sites, TF binding sites, enhancers, splice sites, and UTRs [40–44]. Furthermore, enrichment is often observed in particular contexts: tissues, cell types, or cell states [18, 32, 37]. An example is the enrichment of fine-mapped immune-mediated inflammatory disease (IMID) SNPs in CD4⁺ T cell enhancers, particularly to enhancers activated after stimulation [42]. These results put forth expression as an important intermediate linking non-coding GWAS variants to their associated traits, and help nominate trait-relevant contexts.

1.2.2 Expression is a complex trait

- Expression in itself, is a molecular phenotype that is heritable and complex architecture [45]
 - Expression can be assayed by high-throughput technologies e.g. expression microarray or RNA-sequencing (RNA-seq)
 - The variants associated with assayed expression are called expression quantitative trait loci (eQTLs).
 - eQTLs can also be *cis*- or *trans*- to their target gene [46].
 - Their effect size declines with distance to the TSS, so the most readily detectable eQTLs are *cis*, and within 1Mb [47]
- GWAS variants are enriched for eQTLs <https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000888>
 - So GWAS variants that are also eQTL naturally prioritise target genes.
 - Is it a narrow view to assume that the effect of GWAS loci on complex traits not only act through a target gene, but are specifically mediated by eQTL effects?
 - * Over many complex traits, a median of 11% heritability could be explained by mediation of GWAS loci by common (MAF > 0.01) *cis*-eQTL, and this proportion does not include *trans* or post-transcriptional effects.
- <How to use eQTLs to prioritise target genes>
 - With increasing sample size, most genes (>60-80%) have a detectable eQTL [47, 48]. Assuming that a locus on the genome is associated with both a complex trait and an eQTL, how can we separate the scenario where one variant affects both trait and expression (pleiotropy), from coincidental overlap between distinct causal variants that may possibly in LD? Bayesian Colocalisation methods address this by extending fine-mapping to multiple phenotypes [49]. In particular, Bayesian probabilistic colocalisation methods (e.g. eCAVIAR, Sherlock, coloc [50]) estimate the posterior probability that the same causal variants are associated with both phenotypes, distinguishing pleiotropy from linkage, but not vertical pleiotropy (mediation) from horizontal pleiotropy (independent effects on trait and expression) [51]. As colocalisation of a GWAS loci with eQTLs is necessary but not sufficient for mediation, it should be supported by complementary lines of evidence from other methods that integrate intermediate phenotypes (e.g. MR, mediation analysis [51]) to help untangle the multiplex of possible causal pathways from variant to trait.

add uses other vars

1.2.3 Context-dependent eQTL

- Like expression itself, the effects of eQTLs are also context-dependent [46, 47].
 - This represents genotype-environment interactions at those eQTL.

- A non-exhaustive list of environments that eQTLs have been found to interact with:
 - * sex, age <https://academic.oup.com/hmg/article/23/7/1947/655184>
 - * ancestry [52–54]
 - * tissue [55, 56]
 - * cell type composition in bulk samples [57–60]
 - * individual cell type [52, 60–63]
 - * disease status [62],
 - * and experimental stimulation (see section 1.2.4).
- What molecular mechanisms might facilitate genotype-environment interactions at eQTLs?
 - <list of mechanisms related to regulation of gene expression by TFs, enhancers, splice sites etc.>
 - * [64]: defines static, conditional, dynamic eQTLs
 - * Fu *et al.* [65]: proposes TF-based mechanisms for cis-eQTL (here, define mag, damp, flip) (Fig. 1.4)
 - * Gaffney [45] and Rotival [66]: suggests info on more regulatory layers will help break down transcriptional and post-transcriptional mechanisms
- Given the effect of an eQTL can be starkly different between environments, it is critical to use trait-relevant eQTL datasets for target gene prioritisation at GWAS loci.
 - It has already been shown that use of relevant cell-type specific eQTLs increases coloc rates with GWAS hits [60] <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4498151/> <https://www.biorxiv.org/content/10.1101/2020.01.15.907436v1>
 - Conversely, successful colocalisation of GWAS loci with coloc may prioritise not only the target gene, but the specific environments most relevant to a trait.

1.2.4 Immune Response expression quantitative trait loci

- A important subclass of context-dependent eQTL are response expression quantitative trait locus (reQTL), where the interacting environment is experimental stimulation [47, 67].
 - *In vitro*, potential interacting variables such as cell type, and the nature, length, and intensity of stimulation can be precisely controlled.
- A seminal early study was conducted by [68], where eQTLs were mapped separately in monocyte-derived dendritic cells before and after 18h infection with *Mycobacterium tuberculosis*.
 - reQTLs were detected for 198 genes, 102 specific to the uninfected state, and 96 specific to the infected state.



Figure 1.4: eqtl mech models: magnify, dampen, flip

- Since then, *in vitro* immune reQTL studies have been conducted for a variety of cell types (e.g. primary CD14+ monocytes [69]) and stimulations (IFN γ and LPS [69]).
- A complementary approach is *in vivo* reQTL mapping
 - There are pros to *in vivo* stimulation.
 - * the innumerable interactions in the immune system that are absent *in vitro*
 - ability to use whole organism perturbations more suited for in vivo administration
 - * ability to get whole organism phenotypes
 - * ability to get repeated measures: can reason about change in expression over time
 - Major disadvantages:
 - * the choice of stimulation must be ethical *in vivo*,
 - * and many environmental factors (e.g. diet, lifestyle, immune exposures) cannot be controlled, leading to greater experimental noise (?), and more complex interpretations.
 - There are few published *in vivo* reQTL studies.
 - * [71]: seasonal trivalent inactivated influenza vaccine (TIV), whole blood, antigen processing and intracellular trafficking genes, attempted mediation for Ab titres, but concluded they were underpowered
 - * [72]: fold-change expression after inactivated vaccinia vaccine, focus was on pair-wise epistatic interactions, apoptosis pathways

list a few more types and
stims from [69] until [70]

* [73]: whole blood, IFN status and anti-IL6 drug exposure, reQTL driven by ISRE and IRF4 motifs

- <why care about immune reQTLs>
 - Exposes differences in regulatory architecture between conditions.
 - Does not automatically reveal the mechanisms behind those differences, but provides a starting point for forming mechanistic hypotheses e.g. context-dependent expression
 - Nevertheless useful for interpretation of GWAS signals, providing info on likely contexts that mediate the genetic effect
 - Immune *in vitro* reQTL have been shown to be enriched more so than non-reQTL among GWAS loci for immune-related phenotypes such as susceptibility to infectious [68, 74] and immune-mediated diseases [74, 75].
 - Not yet clear whether *in vivo* reQTL have any utility on top of *in vitro* reQTL for interpreting GWAS loci: not that many studies, and complex interpretations.
 - Nevertheless, as the number of cell types systems and stimulations both *in vitro* and *in vivo* increases, the number of known reQTLs continues to grow.

1.3 Phenotypes of immune response

1.3.1 Intro to immunology

- Most reQTL studies to date have been conducted on immune cells *in vitro*, not only due to the abundance of immune cells easily accessible in peripheral blood, amenable to separation (e.g. FACS) and stimulation; but because the immune system is specialised for responding to environmental perturbations in the form of infections.

– The immune system consists of many cell types: lymphoid and myeloid lineages

- * pbmcs
- two major components, innate and adaptive.
- complex interactions through cell surface receptors and signalling molecules

1.3.2 High-throughput immunology

- omics studies of interactions with and between all these layers of the immune system come under the broad umbrella of systems immunology
 - Longitudinal designs and high-throughput technologies to quantify the components of the immune system at baseline, and in response to perturbations.
 - * genomics, epigenomics, transcriptomics, proteomics, cell counts
 - Whatever the perturbation, just like for most phenotypes, immune response varies between individuals
 - Main types of study are descriptive:

specifically introduce cell types mentioned in later chapters

- * finding correlations between the components of the immune system, and with phenotypic response
- predictive:
 - * Use assayed information to predict phenotypic response from molecular assays
- causal:
 - * Mechanisms of response are important to understand e.g.
 - what mechanisms promote effective responses to pathogens and vaccines
 - impede dysregulated responses that lead to immune-mediated disease such as IBD.
- Natural genetic variation can be leveraged, representing small scale perturbations that are causally anchored [76, 77]
 - * Overall estimates of the heritability of many immune parameters, such as cell composition and serum protein levels, lies between 20-40% [78–81]
 - * Genetic regulation is more important for the innate immune system than the adaptive immune system [80].

1.3.3 Immune response to vaccination

- One of the earliest applications of systems immunology was to vaccine response, leading to the sub-discipline of systems vaccinology.
 - <general summary of biology of vaccine response, specific flu vaccine biology goes in ch2>
 - * Vaccines stimulate the immune system with pathogen-derived antigens to induce effector responses (primarily antigen-specific antibodies) and immunological memory against the pathogen itself.
 - * These effector responses are then be rapidly reactivated in cases of future exposure to the pathogen, mediating long-term protection.
 - * Non-Ab mediated responses [...]
 - Vaccination has enormous impact on global health [82]
 - But traditional vaccine dev is empirical (classical "isolate, inactivate, inject" paradigm), often successful vaccine dev does not offer insights into the mechanisms of efficacy
 - The immunological mechanisms that underpin a specific vaccine's success or failure in a given individual are often poorly understood.
 - A vaccine that is highly efficacious in one human population may have significantly lower efficacy in other populations. Particularly challenging populations for vaccination include the infants and elderly, pregnant, immuno compromised patients, ethnically-diverse populations, and developing countries.
 - * e.g. variable vaccine efficacy of rotavirus vaccine

this is a general systems vaccinology review. specific flu studies are reviewed in ch2

* e.g. variable efficacy of flu vaccine <https://www.sciencedirect.com/science/article/pii/S1473309918304900>

- Systems vaccinology is the application of -omics technologies to provide a systems-level characterisation of the human immune system after vaccine-perturbation.
 - Systems vaccinology has been successfully applied to a variety of licensed vaccines [yellow fever, influenza ...], and also to vaccine candidates against [HIV, malaria ...], resulting in the identification of early transcriptomic signatures that predict vaccine-induced antibody responses.
 - * <add more references to lists [...] of what vaccines have been studied>
 - Has revealed many influences on vaccine response (age, sex, dose, adjuvants, expression signatures, microbiome, strain etc.)
 - So far, studies have been mainly descriptive and predictive.
 - Identifying causal relationships can inform more mechanism-based and cost-effective design (rational paradigm), and the move towards personalised vaccinology.
 - Studies of impact of host genetics on response are underrepresented [83]
- Like for other complex traits, from twin studies it’s known that vaccine Ab responses are heritable.
 - Moving out of the candidate gene era (e.g. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3570049/>) into GWAS.
 - Heritability of vaccine response phenotypes [84]
 - Many loci have been implicated by GWAS e.g. HLA [83–88]

– Overall, systems vacc studies that include genetics (sometimes dubbed as vaccinogenomics studies) are nowhere near as mature compared to the trait to gene pipeline described in above e.g. applied to complex disease

1.3.4 Immune response to anti-TNF biologics

<Considering whether I need a section here on anti-TNFs analogous to the above systems vaccinology section. Much anti-TNF response literature is already reviewed in chapter 4 introduction though, so maybe no point duplicating.>

1.4 Thesis overview

- My thesis focuses on two specific instances of in vivo immune response: antibody response to pandemic influenza vaccine in healthy individuals, and clinical response to biologic anti-TNF therapy for CD patients.
- <By chapter context-content-conclusion overview.>
 - <ch 2: systems vaccinology study of Pandemrix>

define what a signature is in terms of correlation and prediction

find best GWAS ref, probably mooney2013SystemsImmunogeneticsVaccines, then prune and reassign these citations

cut this section (???)

- * context: existing Sobolev study of expression differences between pandemic flu vaccine R/NR had small sample size and binary phenotype
- * content: meta-analysis of existing array with new RNAseq data and continuous phenotype
- * conclusion: distinct innate and adaptive expression response at d1 and d7; heterogeneity between array and RNAseq. significant expression differences between R/NR in meta-analysis at the gene set level
- <ch 3: in vivo reQTL study of Pandemrix>
 - * context: relatively few studies have assessed the impact of genetic variation on expression response to flu vaccine
 - * content: reQTL analysis for flu vaccine at d0, d1, d7. many reQTLs including sign flips. no particular gene set enrichments. evidence of cell type interactions at top hits.
 - * conclusion: difficult to separate out modifying effect of cell composition. this may be a fundamental flaw in the study design
- <ch 4: systems immunology and reQTL study of response to anti-TNF treatment in CD>
 - * context: studies on expression signatures of anti-TNF PNR have been small
 - * content: R/NR comparison with larger n, at baseline, w14, and over time. reQTL analysis over 4 timepoints.
 - * conclusion: a few hits for PNR at baseline. much stronger expression differences stronger at w14, then maintained until w54. Weak evidence for reQTLs, probably due to smaller magnitude of cell proportion changes over time vs the previous chapter.
- <discussion: limitations, future outlook>
 - * main themes and parallels tying together the thesis
 - * shared set of limitations permeating all chapters
 - * recommendations for future analyses and study design
 - * future outlook for the fields of systems studies of immune response

Chapter 2

Transcriptomic response to Pandemrix vaccine

2.1 Introduction

2.1.1 Seasonal and pandemic influenza

Influenza is an infectious disease, generally seasonal, caused by the influenza A and influenza B viruses in humans. Influenza A viruses circulate not only in humans, but also in a variety of other birds and mammal hosts. They are classified into antigenically-distinct subtypes by the combination of two surface proteins: haemagglutinin (HA) and neuraminidase (NA)[89].

There are three classes of influenza vaccine against seasonal strains in use: inactivated vaccines, live attenuated influenza vaccines (LAIVs), and recombinant HA vaccines. These vaccines confer a degree of strain-specific protection, primarily by raising serum antibodies against the HA and/or NA proteins. Antigenic drift, the accumulation of mutations in these surface proteins over time, necessitates the annual reformulation of seasonal influenza vaccines to reflect circulating strains[90, 91]. On occasion, a novel subtype against which the majority of the population is immunologically naive can arise suddenly (antigenic shift), often from zoonotic origins. A recent example occurred in 2009, when an outbreak of a novel swine-origin strain, eventually termed influenza A (H1N1)pdm09, resulted in a global pandemic, the fourth to occur in the last 100 years[89].

why? for diff groups of people

2.1.2 Quantifying immune response to influenza vaccines

The 2009 pandemic motivated the rapid development, trialing, and licensing of several novel vaccines[92]. Immune response to influenza vaccines in clinical trials is evaluated by assays that measure levels of antibodies specific to the vaccine strain(s). The haemagglutination inhibition (HAI) assay measures the levels of serum antibodies specific to the HA surface protein. The related microneutralisation (MN) assay measures levels of antibodies (which may or may not be anti-HA) that neutralise the infectivity of the virus in cell culture [93]. Values from these assays can be compared against thresholds for known correlates of protection: markers that associate with whether an individual is protected from the disease. For example, HAI titres are regarded as the primary correlate of protection for inactivated influenza vaccines. Targets that regulatory

add a point that 2009h1n1 is now circulating seasonally, this is a common trend

Add specific section about pandemrix, it's correlates of protection, it's durability? or maybe in methods

Here, add few points about the immunological response to adjuvanted TIVs i.e. what happens after Pandemrix admin? Involve the innate -> B/CD4T response. Goto plotkins

agencies expect a licensed vaccine to meet are based on thresholds such as the proportion of trial individuals achieving HAI titres ≥ 40 and seroconversion (≥ 4 -fold increase in titres)[94, 95].

2.1.3 Systems vaccinology of influenza vaccines

Although HAI titres are accepted as established correlates for inactivated seasonal influenza vaccines, they fail to account for alternate mechanisms such as T cell-mediated protection, and correlates for LAIV and pandemic influenza vaccines are less reliable[90]. For novel and emerging diseases, there may be no prior knowledge of robust correlates to use in the vaccine development process. In response, the last decade has seen the rise of systems vaccinology studies: the analysis of high-dimensional data measured using multiple technologies in vaccinated individuals, in order to characterise response to vaccination at multiple levels of the biological system[96].

Such information helps elucidate a vaccine’s mode of action, discover “molecular signatures” predictive of vaccine safety and efficacy, and has become an increasingly important part of the modern vaccine development chain[97, 98].

Various systems vaccinology studies of seasonal influenza vaccines have been conducted, taking longitudinal measurements pre-vaccination, and commonly at some subset of days 1, 3, 7, and 28 post-vaccination. These measurements can be correlated to changes in antibody titres after vaccination to define signatures of antibody response with potential utility as correlates of protection. One of the earliest such studies by Zhu et al.[99] found that expression of type 1 interferon-modulated genes was a signature of response to LAIV. An expression signature including *STAT1*, *CD74*, and *E2F2* correlated with serum antibody titres after vaccination with trivalent inactivated influenza vaccine[100]; kinase *CaMKIV* expression is also a strong predictor [101], as are genes related to B cell proliferation [102].

For these studies of seasonal influenza vaccines in adults, responses tend to be biased by recall from past vaccination or infection[100, 103]. There have also been few studies of adjuvanted influenza vaccines, despite their superior efficacy in comparison to non-adjuvanted counterparts[104, 105].

2.1.4 The Human Immune Response Dynamics (HIRD) study

The Human Immune Response Dynamics (HIRD) study conducted by Sobolev *et al.* [106] was conceived with the above limitations in mind. The study is an observational, prospective cohort design. The vaccine studied was Pandemrix, an AS03-adjuvanted, split-virion, inactivated vaccine against influenza A(H1N1)pdm09 (strain name A/California/07/2009, subtype H1N1), for which the majority of the cohort at the time would be unlikely to have immunological memory. A total of 178 individuals were vaccinated with a single dose of Pandemrix, and longitudinal transcriptomic, cellular, antibody titre, and adverse event phenotypes were collected. Gene expression was profiled using a microarray, and differential gene expression (DGE) analyses detected genes associated with both myeloid and lymphoid effector functions upregulated at day 1, most prominently for genes associated with interferon responses. These early myeloid responses were consistent with studies of unadjuvanted seasonal influenza vaccines, but the interferon gamma-associated lymphoid response was unique to this adjuvanted vaccine.

is there a more recent review?

define 'signature': nvm, remove it

Genes related to plasma cell development and antibody production were more highly expressed in 23 vaccine responders compared to 18 non-responders at day 7 post-vaccination. However, due to high variability among the vaccine non-responders in variables such as baseline antibody titres, a consensus predictive model that segregated the two groups could not be built, even considering other measures such as frequencies of immune cell subsets and serum cytokine levels, suggesting there was no single contributing factor that led to vaccine failure. This is in contrast to several studies of seasonal influenza vaccines, where certain expression signatures are able to predict vaccine response even pre-vaccination[107–110].

high variability, recheck this was the reason, or quote them

make sure gap and how it is filled is emphed enough

2.1.5 Chapter summary

Transcriptomic measurements in the original HIRD study were restricted to a relatively small number (46/178) of individuals, potentially limiting power to detect a expression signatures associated with antibody response. In addition, the responder vs. non-responder phenotype definition used does not account for variation in pre-existing baseline titres, and the binary definition can result in loss of information[111–113].

In this chapter, I integrate the original microarray data from HIRD with RNA-sequencing (RNA-seq) data on a larger subset (75) of newly sequenced individuals from the same cohort using Bayesian random-effects meta-analysis. The overall pattern of expression over time from my meta-analysis agrees with the patterns from the original study [106], with transient innate immune response at day 1 post-vaccination, progressing to adaptive immune response by day 7.

needs 1 more punchline sentence here

From existing HAI and MN data, I compute a baseline-adjusted, continuous measure of antibody response to vaccination, the titre response index (TRI)[100]. Effect sizes of genes with expression that correlated with TRI were very dependent on measurement platform (array or RNA-seq), and no robust hits were detected in the meta-analysis. Leveraging the greater power that rank-based gene set enrichment analyses affords, I find modules of coexpressed genes that correlate with antibody response, with the strongest effects observed for adaptive immune modules at day 7, but also in inflammatory modules at baseline.

2.2 Methods

2.2.1 Existing HIRD study data and additional data

The design of the HIRD study is described in [106]. In brief, the study enrolled 178 healthy adult volunteers in the UK. The vaccine dose was administered after blood sampling on day 0; five other longitudinal blood samples were taken on days -7, 0, 1, 7, 14 and 63. Serological responses were measured on days -7 and 63 using the HAI and MN assays, and various subsets of the cohort were also profiled for serum cytokine levels (Luminex panel, days -7, 0, 1 and 7), immune cell subset counts (fluorescence-activated cell sorting (FACS) panels, all days), and peripheral blood mononuclear cell (PBMC) gene expression (microarray, days -7, 0, 1 and 7). The gene expression microarrays were performed in two batches.

why blood? ready easy supply of immune cells, despite delivery being muscle?

In addition to the existing data, array genotypes were generated for 169 individuals; and RNA-seq data for 75 individuals at days 0, 1, and 7. The sets of individuals with gene expression

assayed by microarray and RNA-seq is disjoint, as no biological material for RNA extraction remained for the microarray individuals. An overview of datasets is shown in Figure 2.1.

2.2.2 Computing baseline-adjusted measures of antibody response

In [106], Pandemrix responders were defined as individuals with ≥ 4 -fold titre increases in either the HAI or MN assays. This is a threshold for seroconversion set out by the U.S. Food and Drug Administration[114], and is used in many studies of seasonal influenza vaccines[97]. The responder status for 166 individuals with both HAI and MN titres available at baseline (day -7) and post-vaccination (day 63) were computed according to this definition. However, [106] noted there was heterogeneity in the baseline titres of non-responders, citing “glass ceiling” non-responders whose high baseline titres made the fixed 4-fold threshold hard to achieve. Dichotomisation of continuous response variables can also result in loss of statistical power [111, 113].

To address these concerns, I computed the TRI as defined in Bucasas *et al.* [100]. For each assay, a linear regression was fit with the \log_2 day 63/day -7 titre fold change as the response, and the \log_2 day -7 baseline titre as the predictor. The residuals from the two regressions were each standardized to zero mean and unit variance, then averaged. The TRI expresses a continuous measure of change in antibody titres across both assays post-vaccination, compared to individuals with a similar baseline titre, and remains comparable to the binary 4-fold change definition (Figure 2.2).

Descriptive statistics for the 114 individuals with both gene expression and antibody titre data are presented in Table 2.1. Although the proportion of responders between array (32/44) and RNA-seq (59/70) individuals is similar ($p = 0.1551$, Fisher’s exact test), the variance of TRI in array individuals is higher ($p = 0.0002098$, Levene’s test), suggesting more extreme antibody response phenotypes are present (Figure 2.3). The cause of this is unknown, there is a possibility that individuals with more extreme phenotypes were prioritised for array transcriptomics in the original HIRD study*.

*Personal communication with authors.

mainmatter/figures/chapter_02/graphics_ashg19/hird_design-crop.pdf

Figure 2.1: Data types, timepoints, and sample sizes. Individuals were vaccinated after day 0 sampling. Antibodies to the vaccine strain were measured by HAI and MN assays. Array and RNA-seq gene expression measured in the PBMC compartment.

atm I'm not using R/NR. wording here implys I am

heterogeneity: well of course there was

cite appropriate subfigures here

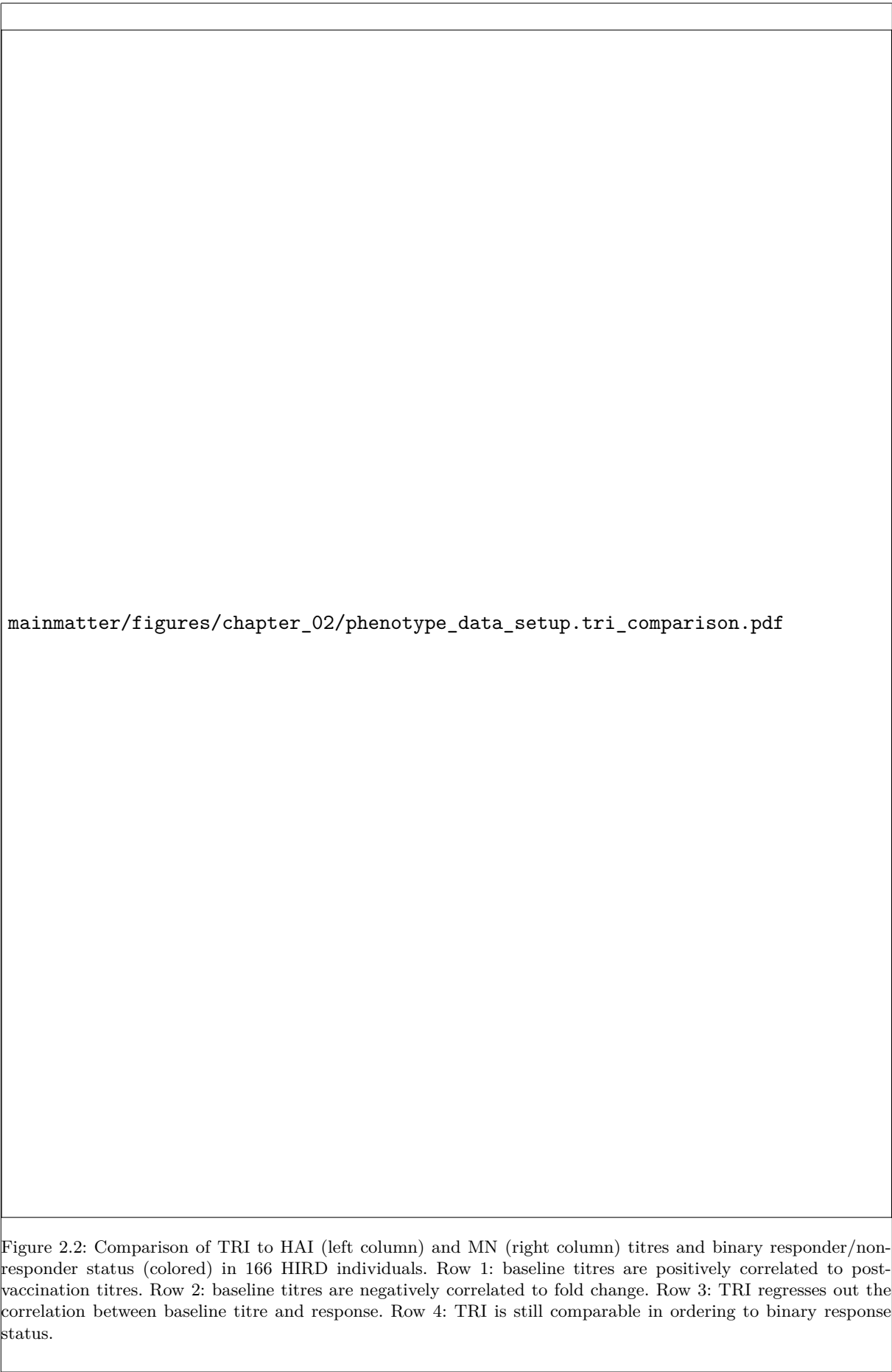
change score is usually negatively correlated to baseline [115]

this is due to within-individual RTM , which is why we residualise

the variable is not used as an inclusion/exclusion criterion for the study, otherwise regression to the mean will be strong

upend change score bit, the only thing we are concerned wth here is clifton2019CorrelationBaseline

cite appropriate subfigures here, after adding proper subfigure labels



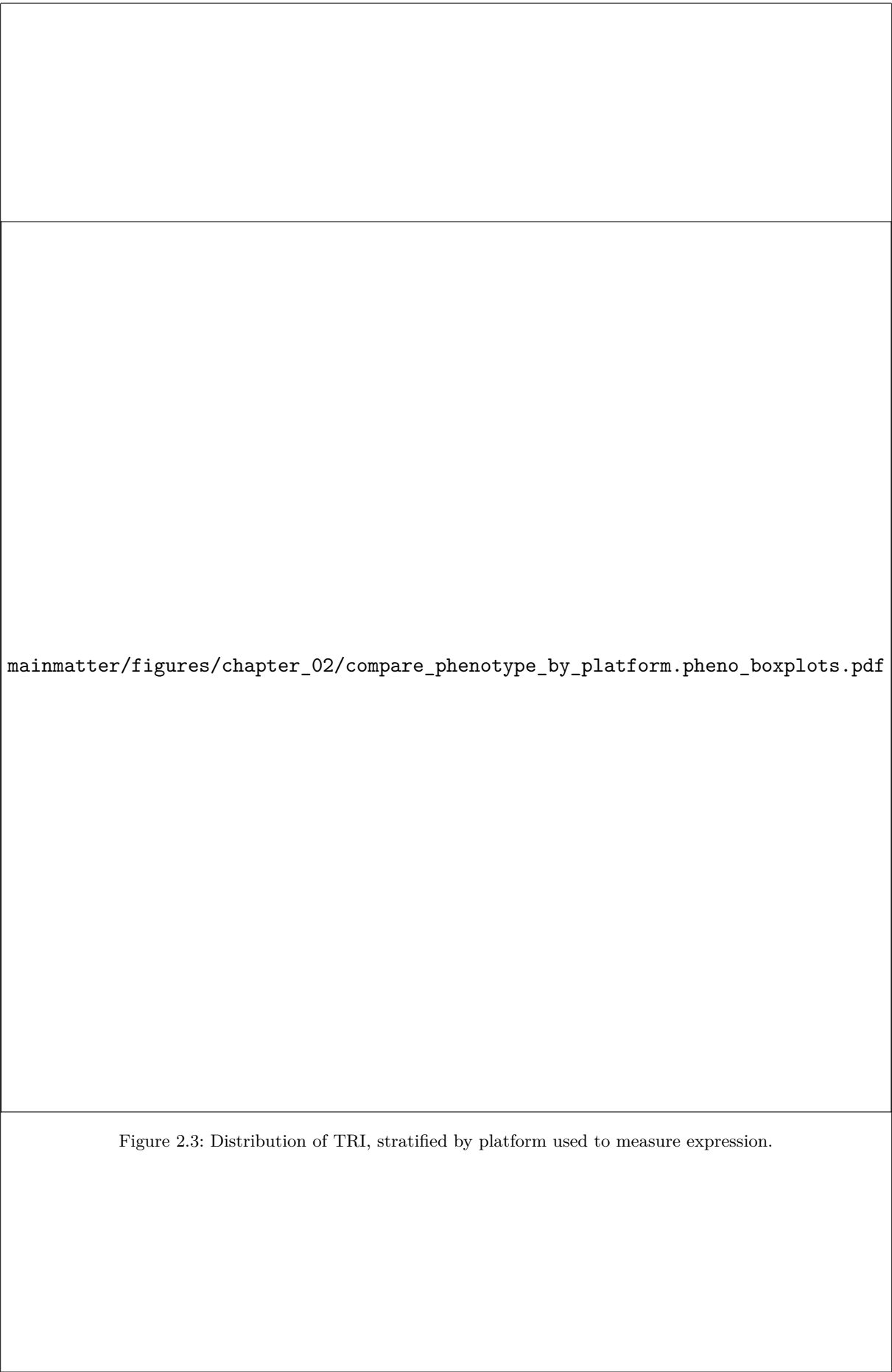


Table 2.1: Sample descriptive statistics.

| | Total | platform | |
|--------------------------|-------------|-----------------|------------------|
| | n = 114 | array n = 44 | rnaseq n = 70 |
| Gender | | | |
| F | 72 (63.2%) | 27 (61.4%) | 45 (64.3%) |
| M | 42 (36.8%) | 17 (38.6%) | 25 (35.7%) |
| Age at vaccination years | 29.2 (11.8) | 32.9 (14.1) | 26.8 (9.4) |
| Ethnic Background | | | |
| Asian | 14 (12.3%) | 5 (11.4%) | 9 (12.9%) |
| Black/African | 9 (7.9%) | 4 (9.1%) | 5 (7.1%) |
| Caucasian | 82 (71.9%) | 33 (75%) | 49 (70%) |
| Latin american | 2 (1.8%) | 1 (2.3%) | 1 (1.4%) |
| Mixed | 5 (4.4%) | 1 (2.3%) | 4 (5.7%) |
| Other - Arab | 1 (0.9%) | 0 (0%) | 1 (1.4%) |
| White Other | 1 (0.9%) | 0 (0%) | 1 (1.4%) |
| log2 HAI 0 | 4.4 (1.8) | 4.2 (1.6) | 4.5 (1.9) |
| log2 HAI 6 | 7.6 (1.8) | 7.4 (2.2) | 7.6 (1.5) |
| log2 HAI ratio | 3.2 (1.9) | 3.2 (2.4) | 3.1 (1.6) |
| log2 MN 0 | 6.2 (2.8) | 5.4 (2.4) | 6.6 (3.0) |
| log2 MN 6 | 10.4 (2.0) | 9.5 (2.2) | 10.9 (1.6) |
| log2 MN ratio | 4.2 (2.3) | 4.1 (2.6) | 4.3 (2.1) |
| responder | | | |
| FALSE | 23 (20.2%) | 12 (27.3%) | 11 (15.7%) |
| TRUE | 91 (79.8%) | 32 (72.7%) | 59 (84.3%) |
| TRI | -0.0 (0.9) | -0.2 (1.2) | 0.1 (0.7) |

2.2.3 Genotype data generation

DNA was extracted from frozen blood using the Blood and Tissue DNeasy kit (Qiagen), and genotyping was performed using on the Infinium CoreExome-24 BeadChip (Illumina). In total, 192 samples from 176 individuals in the HIRD cohort were genotyped at 550601 markers, including replicate samples submitted for individuals where extracted DNA concentrations were low.

2.2.4 Genotype data preprocessing

Using PLINK (v1.90b3w), genotype data underwent the following quality control procedures to remove poorly genotyped samples and markers: max marker missingness across samples < 5%, max sample missingness across markers < 1%, max marker heterozygosity rate within 3 standard

Add to collab note that
extractions were done at
KCL

deviations of the mean (threshold selected visually to exclude outliers, Figure 2.4), removal of markers that deviate from Hardy-Weinberg equilibrium (HWE) (`--hwe` option, $p < 0.00001$).

To exclude highly-related individuals and deduplicate replicate samples, pairwise kinship coefficients were computed on minor allele frequency (MAF) < 0.05 pruned genotypes using KING (v1.4). For each pair of samples with pairwise kinship coefficient > 0.177 (first-degree relatives or closer), the sample with lower marker missingness was selected.

After filtering, 169 samples and 549414 markers remained.

2.2.5 Computing genotype principal components as covariates for ancestry

As shown in Table 2.1, the HIRD cohort is multi-ethnic, hence there is potential for confounding by population structure (sample structure due to genetic background) and genetic association studies [116, 117]. Large-scale population structure explains variation in gene expression [118], so including population structure as covariates can increase power. Treating HapMap 3 samples [119] as a reference population where the major axes of variation in genotypes are likely to be ancestry, principal component analysis (PCA) was performed using smartpca (v8000) on linkage disequilibrium (LD)-pruned genotypes (PLINK `--indep-pairwise 50 5 0.2`). HIRD sample principal components (PCs) were computed by projection onto the HapMap 3 PCA eigenvectors. For non-genotyped individuals, PC values were imputed as the mean value for all genotyped individuals with the same self-reported ancestry. The top PCs separate samples of European, African and Asian ancestry (Figure 2.5), hence these PCs can be used as continuous covariates for ancestry downstream.

2.2.6 RNA-seq data generation

Total RNA was extracted from PBMCs using the Qiagen RNeasy Mini kit, with on-column DNase treatment. RNA integrity was checked on the Agilent Bioanalyzer and mRNA libraries were prepared with the KAPA Stranded mRNA-Seq Kit (KK8421), which uses poly(A) selection. To avoid confounding of timepoint and batch effects from pooling, samples were pooled by library prep plate, ensuring libraries from all timepoints of an individual were in the same pool, and then sequenced across multiple lanes as technical replicates (HiSeq 4000, 75 bp paired-end).

RNA-seq quality metrics were assessed using FASTQC* and Qualimap[120], then visualised with MultiQC[121]. Sequence quality was high (Figure 2.6), and duplication levels were low (Figure 2.7). The unimodal GC-content distribution suggested negligible levels of non-human contamination (Figure 2.8).

2.2.7 RNA-seq quantification and filtering

Reads were quantified against the Ensembl reference transcriptome (GRCh38) using Salmon[122] in quasi-mapping-based mode, which internally corrects for transcript length and GC composition by computing an effective length for each transcript. To combine technical replicates, as the sum of Poisson distributions remains Poisson-distributed, counts for technical replicates

*<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

Add Tracy-Widom statistics for PCs to justify later choice of 4 PCs for covariates

nicer version, copy the peer code, facet the hird and hapmap samples

Can add other fastqc plots e.g. kmers, overrepresented seqs, seq length

add software versions

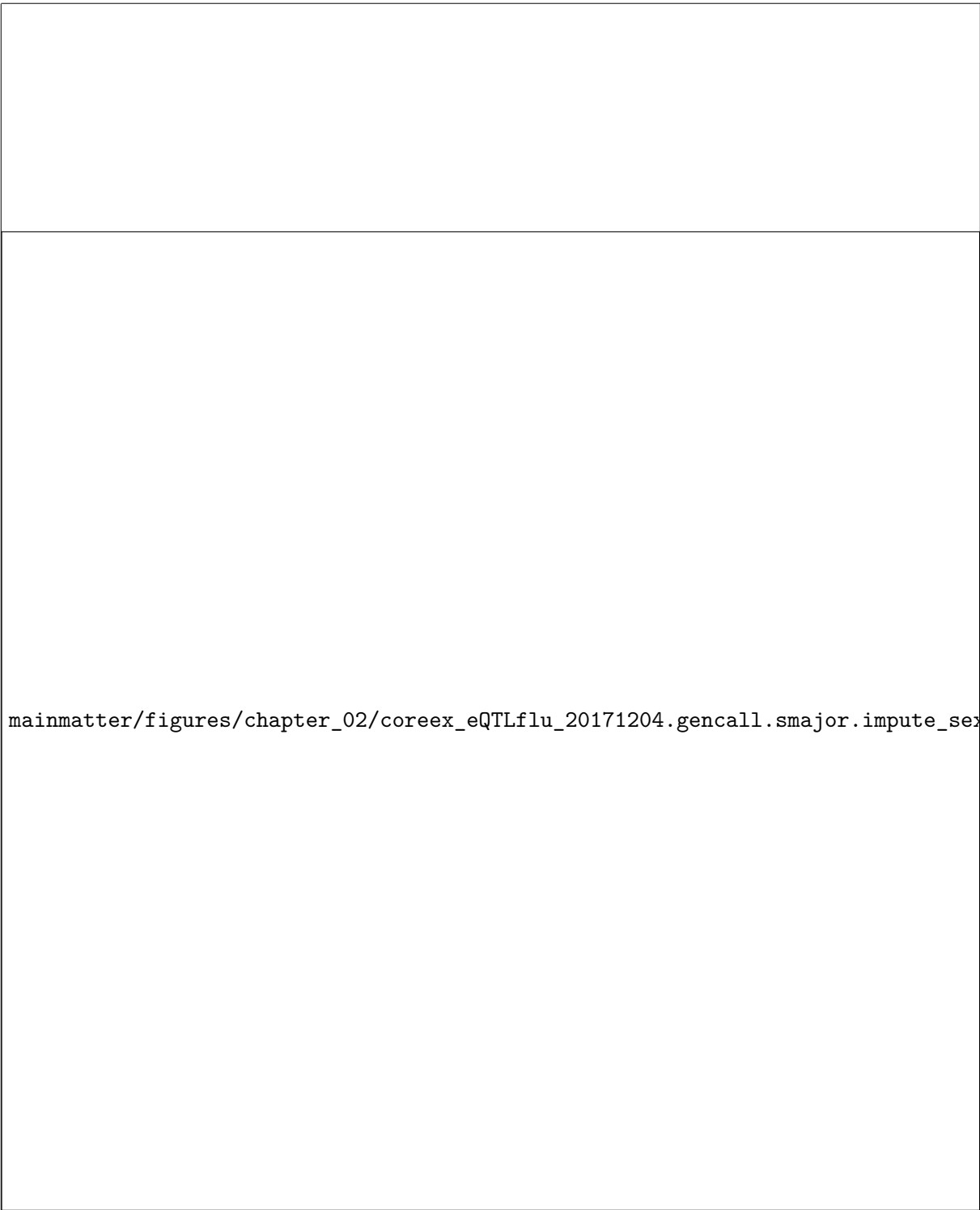


Figure 2.4: Sample filters for missingness and heterozygosity rate. Samples outside the central rectangle were excluded.

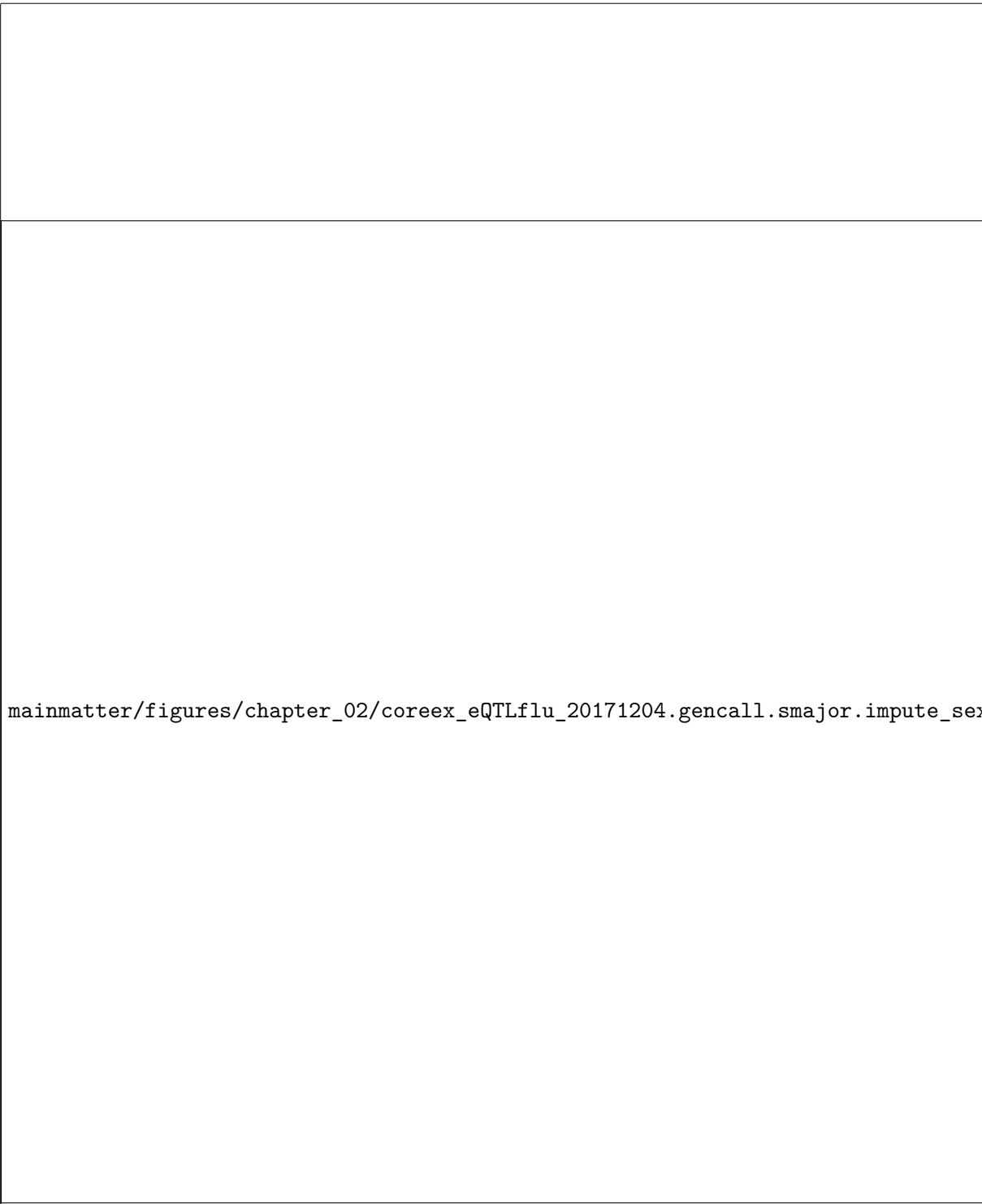
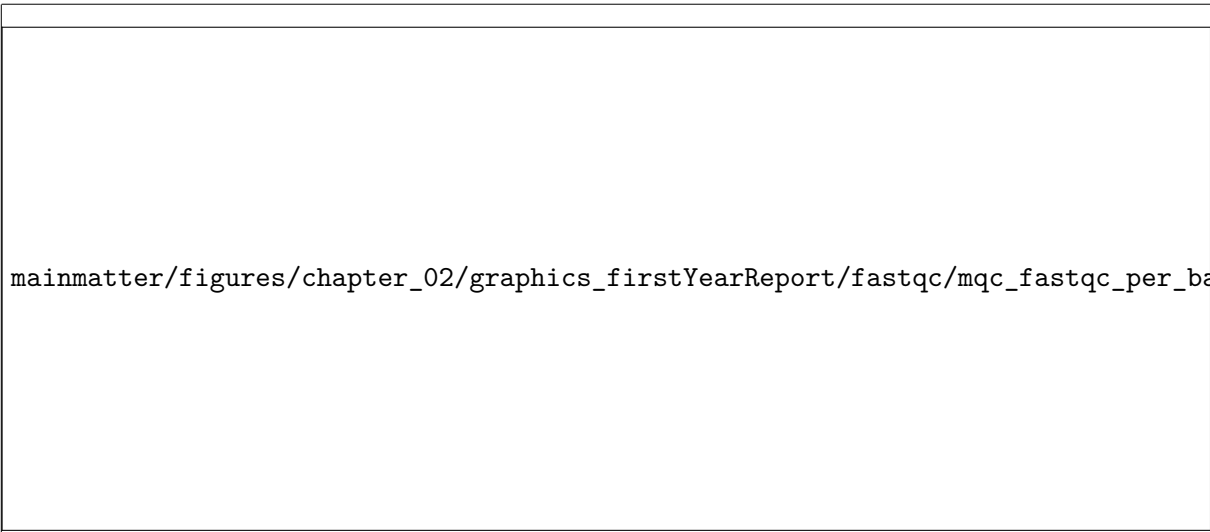


Figure 2.5: HIRD samples (cyan) projected onto PC1 and PC2 axes defined by PCA of HapMap 3 samples. The first two PCs separate European (CEU, upper-right) from Asian (CHB and JPT, lower-right) and African (YRI, lower-left) individuals.

| | | |
|---|--|--|
|  <p>mainmatter/figures/chapter_02/graphics_firstYearReport/fastqc/mqc_fastqc_per_base_sequence_quality_plot_C</p> | | |
| Figure 2.6: FastQC sequence quality versus read position for HIRD RNA-seq samples. | | |
|  <p>mainmatter/figures/chapter_02/graphics_firstYearReport/fastqc/mqc_fastqc_sequence_duplication_levels_pl</p> | | |
| Figure 2.7: FastQC sequence duplication levels for HIRD RNA-seq samples. | | |
|  <p>mainmatter/figures/chapter_02/graphics_firstYearReport/fastqc/mqc_fastqc_per_sequence_gc_content_plot_C</p> | | |
| Figure 2.8: FastQC GC profile for HIRD RNA-seq samples. | | |

were summed for each sample. The mean number of mapped read pairs per sample after summing was 27.09 million read pairs (range 20.24-39.14 million), representing a mean mapping rate of 80.73% (range 75.57-90.10%), comfortably within sequencing depth recommendations for DGE experiments[123]. Relative transcript abundances were summarised to Ensembl gene-level count estimates using tximport (scaledTPM method) to improve statistical robustness and interpretability[124].

Genes with short noncoding RNA biotypes* were removed, as they are generally not polyadenylated, and expression estimates can be biased by misassignment of counts from overlapping protein-coding or lncRNA genes[125]. Globin genes, which are highly expressed in erythrocytes and reticulocytes, cell types expected to be depleted in PBMC [126], were also removed. Given the proportion of removed counts at this stage was low for most samples (Figure 2.9), poly(A) selection and PBMC isolation procedures were deemed to have been efficient.

Many of the genes in the reference transcriptome are not expressed in PBMC (Figure 2.10), and many genes are expressed at counts too low for statistical analysis of DGE. Genes were further filtered to require detection (non-zero expression) in at least 95% of samples, and a minimum of 0.5 counts per million (CPM) in at least 20% of samples. The 0.5 CPM threshold was chosen to correspond to approximately 10 counts in the smallest library, where 10-15 counts is a rule of thumb for considering a gene to be robustly expressed[127]. The change in the distribution of gene expressions among samples before and after filtering shows a substantial number of low expression genes are removed (Figure 2.11).

After the application of all filters, expression values were available for 21626 genes over 223 samples (75/75 individuals on day 0, 73/75 on day 1, and 75/75 on day 7).

2.2.8 Array data preprocessing

Single-channel Agilent 4x44K microarray (G4112F) data for 173 samples from [106] were downloaded from ArrayExpress[†]. These arrays were originally processed in two batches, the effect of which is seen in the raw foreground intensities (Figure 2.12).

VSN[128] was used to perform background correction, between-array normalisation, and variance-stabilisation of intensity values, resulting in expression values on a \log_2 scale.

Most genes are targetted by multiple array probes; 31208 probes were collapsed into 18216 Ensembl genes using by selecting the probe with the highest mean intensity for each gene (`WGCNA::collapseRows(method=MaxMean)`, recommended for probe to gene collapsing[129]). While it would be optimal to select a collapsing method to maximise the concordance between array and RNA-seq expression values, there were no samples assayed by both platforms in the HIRD dataset. The final normalised \log_2 intensity values for these 18216 genes over 173 samples is shown in Figure 2.13.

*miRNA, miRNA_pseudogene, miscRNA, miscRNA_pseudogene, Mt rRNA, Mt tRNA, rRNA, scRNA, snlRNA, snoRNA, snRNA, tRNA, tRNA_pseudogene. List from <https://www.ensembl.org/Help/Faq?id=468>

[†]<https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-2313/>

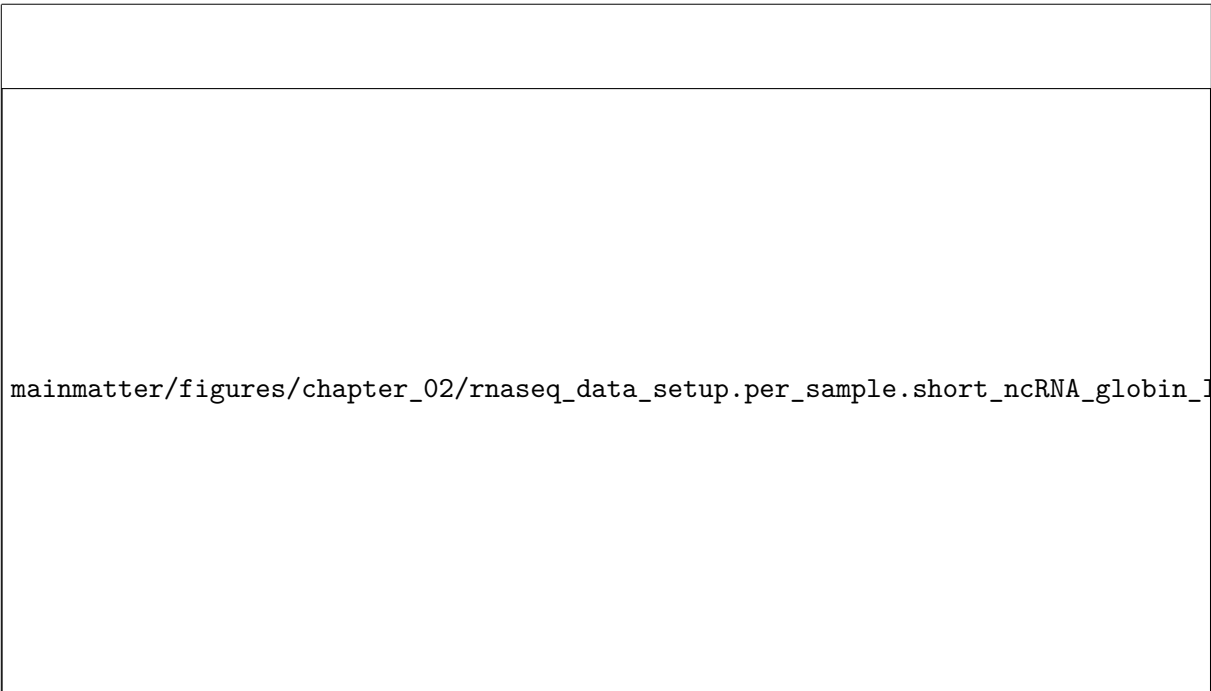


Figure 2.9: Distributions of removed short ncRNA and globin counts as a proportion of total counts in RNA-seq samples.



Figure 2.10: Distribution of the proportion of samples in which genes were detected (non-zero expression). Many genes are not detected in any samples. Vertical line shows 5% threshold below which genes were discarded.

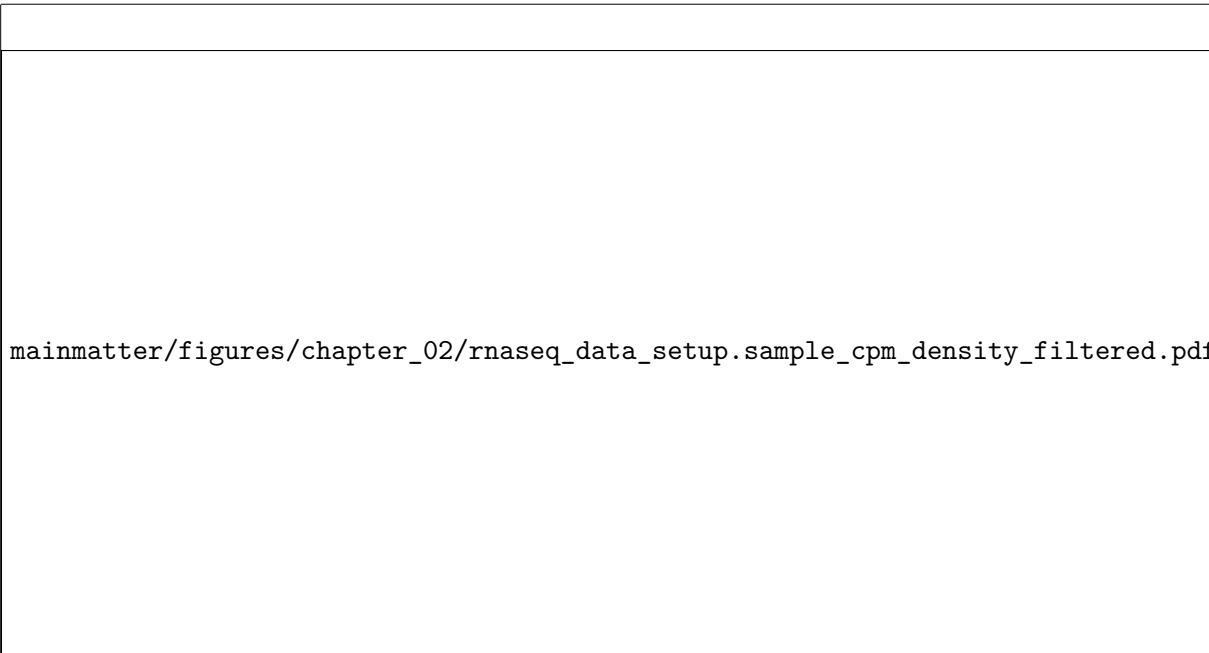


Figure 2.11: Distribution of gene expressions for RNA-seq samples before and after filtering no expression and low expression genes. Vertical line shown at CPM = 0.5 threshold.

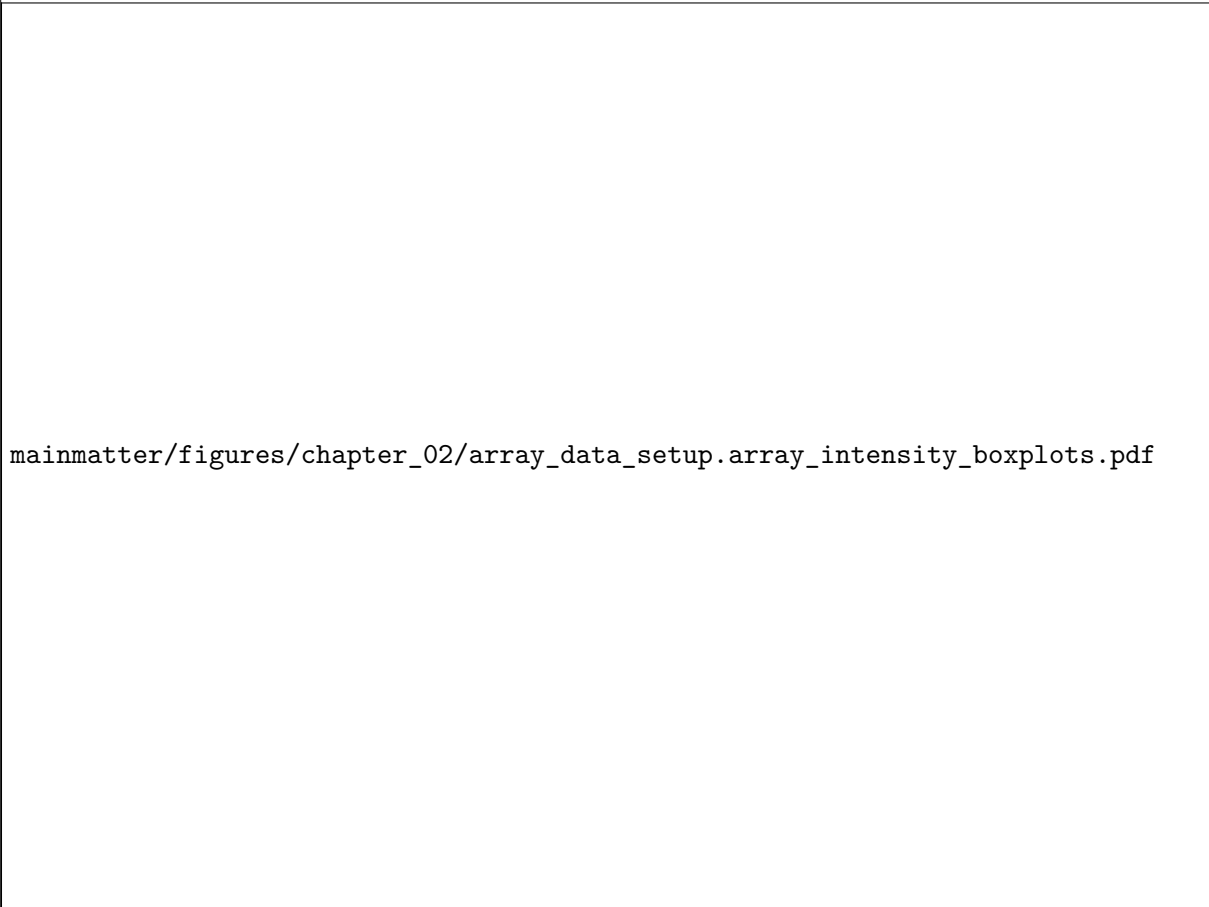


Figure 2.12: Raw foreground intensities for 173 HIRD array samples. Colored by array processing batch.



Figure 2.13: Array intensity estimates after VSN normalisation and collapsing of probes to genes. Colored by array processing batch.

2.2.9 Differential gene expression

PCA of the expression data reveals although samples separate by experimental timepoint along PC3 (Figure 2.14d), measurement platform is by far the largest source of variation. Normalisation was also not able to completely remove the batch effect within the array data (Figure 2.14a). The large platform effect likely stems from systematic technological differences in how each platform measures expression. For example, arrays suffer from ratio compression due to cross-hybridisation[130]. RNA-seq has a higher dynamic range, resulting less bias at low expression levels, but estimates are more sensitive to changes in depth than array estimates are to changes in intensity [131]. There are also differences in the statistical models behind expression quantification and normalisation, as described above.

Despite the shortcomings of array data detailed above, the array dataset tends to contain individuals with more extreme antibody response phenotypes (Figure 2.3), and hence the data should not be excluded. Given the magnitude of the platform effect, I concluded that the appropriate approach should be a two-stage approach that integrates per-platform DGE effect estimates while explicitly accounting for between-platform heterogeneity.

Regarding the batch effect within the array data, a popular adjustment method is ComBat[132], which estimates centering and scaling parameters by pooling information across all genes using empirical Bayes. ComBat is the method used in [106]. In comparisons of microarray batch effect adjustment methods, ComBat performs favourably (vs. five other adjustment packages)[133] or comparably (vs. batch as a fixed or random effect in the linear model)[134]. However, where batches are unbalanced in terms of sample size[135] or distribution of study groups that have an impact on expression[136], ComBat can overcorrect batch differences or bias estimates of group differences respectively. In our data, sample size and timepoint groups are fairly balanced between the two array batches, but the proportion of responders is not- Table 2.2, hence I elect not to use ComBat to pre-adjust the array expression data, and model the batches as fixed effects. In practice, results from the DGE analysis were not substantially affected by the choice of whether to use a ComBat pre-adjustment or a fixed effect.

Table 2.2: HIRD batch balance

| | Total | 1 | 2 | batch | | |
|-----------|-------------|------------|------------|------------|------------|------------|
| | n = 374 | n = 87 | n = 79 | DN500165J | DN500166K | DN500167L |
| | | | | n = 70 | n = 69 | n = 69 |
| visit | | | | | | |
| v1 | 40 (10.7%) | 20 (23%) | 20 (25.3%) | 0 (0%) | 0 (0%) | 0 (0%) |
| v2 | 114 (30.5%) | 24 (27.6%) | 20 (25.3%) | 24 (34.3%) | 23 (33.3%) | 23 (33.3%) |
| v3 | 109 (29.1%) | 21 (24.1%) | 20 (25.3%) | 22 (31.4%) | 23 (33.3%) | 23 (33.3%) |
| v4 | 111 (29.7%) | 22 (25.3%) | 19 (24.1%) | 24 (34.3%) | 23 (33.3%) | 23 (33.3%) |
| responder | | | | | | |
| FALSE | 80 (21.4%) | 12 (13.8%) | 36 (45.6%) | 11 (15.7%) | 9 (13%) | 12 (17.4%) |
| TRUE | 294 (78.6%) | 75 (86.2%) | 43 (54.4%) | 59 (84.3%) | 60 (87%) | 57 (82.6%) |
| TRI | | | | | | |
| | -0.1 (1.0) | -0.1 (1.0) | -0.4 (1.4) | 0.1 (0.6) | -0.0 (0.8) | 0.2 (0.6) |



Figure 2.14: First four PCs in the HIRD expression data, colored by platform and batch (left), and timepoint (right).

2.2.9.1 Per-platform differential gene expression model

For the array data, as [106] demonstrated no significant global differences in expression between day -7 and day 0, I likewise merge these two timepoints into a single “day 0” baseline timepoint in the following DGE models.

For the RNA-seq data, between-sample normalisation factors were computed using the trimmed mean of M-values (TMM) method[137] from `edgeR::calcNormFactors`[138]; then variance-stabilisation was performed using `voom`[139], resulting in expression values with units of \log_2 CPM.

Linear models were fit using `limma`[140], which is computationally fast, and performs well for sufficiently large ($n \geq 3$ per group) sample sizes[141]. For each gene, I fit a model (model 1) with expression as the response variable; with timepoint (baseline, day 1, day 7), TRI, batch, sex, age, and the first 4 genotype PCs as fixed-effect predictors; and individual as a random-effect predictor. Within-individual correlations for the random effect were estimated using `limma::duplicateCorrelation`. A second model (model 2) was also fit, including 3 additional terms for the interactions between each timepoint and TRI. Contrasts were defined, testing if linear combinations of estimated coefficients are different from zero. From model 1, I defined contrasts for day 1 vs. baseline, day 7 vs. baseline, day 7 vs. day 1, TRI, sex, and age. From model 2, I defined contrasts for the TRI specifically at each of the three timepoints. Corresponding coefficients and standard errors for the contrasts were extracted from the linear models, which represent effect size in units of \log_2 expression fold change per unit change in predictor value.

2.2.9.2 Choice of differential gene expression meta-analysis method

In the section , I concluded that a two-stage meta-analysis approach would be appropriate. This meta-analysis is restricted to 13593 genes assayed by both the array and RNA-seq platforms.

Two popular frameworks for effect size meta-analysis are fixed-effect and random-effects[142, 143]. Given k studies, the fixed-effect model assumes a common population effect size shared across all studies, with observed variation explained only by sampling error. The random-effects model assumes the k study-specific effect sizes are drawn from some distribution with variance τ^2 (standard deviation (SD) τ), representing an additional source of variation termed the between-studies heterogeneity, reducing to the fixed-effect model when $\tau = 0$. In the HIRD data, there are $k = 2$ ‘studies’ (array and RNA-seq), where the platform differences described in section contribute to considerable between-studies heterogeneity. The assumption of $\tau = 0$ is unrealistic, hence a random-effects model is more appropriate.

Unfortunately, there is no optimal solution for directly estimating τ in random-effects meta-analyses with small k [145], in the case of $k = 2$ especially[146]. Many estimators are available[147], but lack of information with small k causes estimation to be imprecise, and often results in boundary values of $\tau = 0$ that are incompatible with the assumed positive heterogeneity[148, 149]. In such circumstances, the most sensible choice may be to incorporate prior information about model hyperparameters in a Bayesian random-effects framework[147–150]. For this study, I use the implementation in `bayesmeta` [144], which requires priors for both effect size and between-studies heterogeneity.

this is DGE specific normalisation, which is why it goes here, not in the preprocessing section

link to papers justifying sex, age, ancestry as significant effects on immune gene expression

add equation from ch3. especially justify having TRI in as predictor, by noting equiv of traditional lm to contrasts

add section labels

add label

make all the notation in this section consistent with, and add the equation 2.1. The normal-normal hierarchical model, [144]

2.2.9.3 Prior for between-studies heterogeneity

The choice of prior for between-studies heterogeneity is influential when k is small[150]. Gelman [151] considers the case of $k = 3$, showing that a flat prior places too much weight on implausibly large estimates of τ , and recommends a weakly informative prior that acts to regularise the posterior distribution. Since I assumed zero estimates for τ are unrealistic, I use a weakly-informative gamma prior recommended by [148], which has zero density at $\tau = 0$, increasing gently as τ increases. This constrains τ to be positive, but still permits estimates close to zero if the data support it. This is in contrast to priors used in other studies from the log-normal (e.g. [152, 153]) or inverse-gamma (e.g. [154]) families that have derivatives or zero close to zero, thus ruling out small values of τ no matter what the data suggest; and in contrast to half-t family priors (e.g. [150, 151]), which have their mode at zero, and do not rule out $\tau = 0$.

To estimate the appropriate shape and scale parameters for the gamma empirically, a frequentist random-effects model using the restricted maximum likelihood (REML) estimator for τ (recommended for continuous effects[147]) was first for each gene using `metafor::rma`. Genes with small estimates of $\tau < 0.01$ were excluded, and a gamma distribution was fit to the remaining estimates using `fitdistrplus`.

2.2.9.4 Prior for effect size

While the choice of prior on τ is influential when k is small, there is usually enough data to estimate the effect size μ such that any reasonable non-informative prior can be used [149, 151]. bayesmeta implements both flat and normal priors for μ . Assuming that most genes are not differentially expressed with effect sizes distributed randomly around zero, I selected a normal prior with $N(\mu = 0, \sigma^2)$, over a flat prior. As in the section above, to determine an appropriate scale, a normal distribution with mean $\mu = 0$ was fit to the distribution of effect sizes from the gene-wise frequentist models to empirically estimate σ .

Heavy-tailed Cauchy priors have been proposed for effect size distributions in DGE experiments to avoid over-shrinkage of true large effects in the tails[155]. Since `bayesmeta` does not implement a Cauchy prior, to avoid over-shrinkage, I flatten the normal prior considerably by scaling up the variance to $N(0, 100\sigma^2)$. This is equivalent to assuming placing a 95% prior probability that effects are less extreme than approximately 20σ .

2.2.9.5 Evaluation of priors

An example of the empirically estimated hyperparameters for the priors for the day 1 vs. baseline contrast are shown in Figure 2.15 (for τ) and Figure 2.16 (for μ). For τ , the final prior used was `Gamma(shape = 1.5693, scale = 0.0641)`. This is comparable to [148]’s default recommendation of a `Gamma(shape = 2, scale = λ)` prior where λ is small. For μ , the final prior used was $N(0, (0.3240 * 10)^2)$. The tails of the non-scaled normal fit (black) are light compared to the Cauchy fit (red), which may lead to over-shrinkage, especially since there are many genes with high positive fold changes for the day 1 vs. baseline effect.

why is this? is it having well powered studies? gelman is vague

the derivation here is `qnorm(0.975, mean=0, sd=1*10) = 1*19.59964`, bit iffy, double check this is correct

could also include a table of all sets of parameters here?

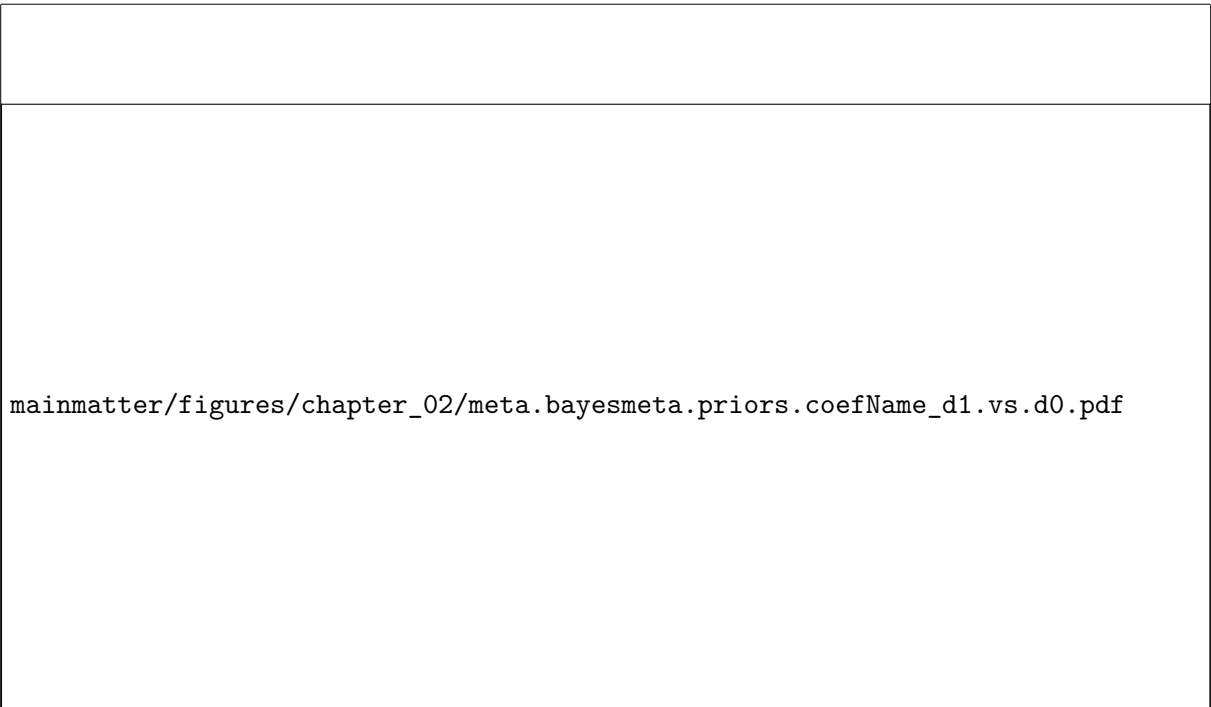


Figure 2.15: Gamma prior for τ used for **bayesmeta** (blue), compared to the empirical distribution of per-gene frequentist **metafor::rma** estimates for τ , for the day 1 vs. baseline effect (small estimates of $\tau < 0.01$ excluded). Empirical log-normal fit also shown (red).

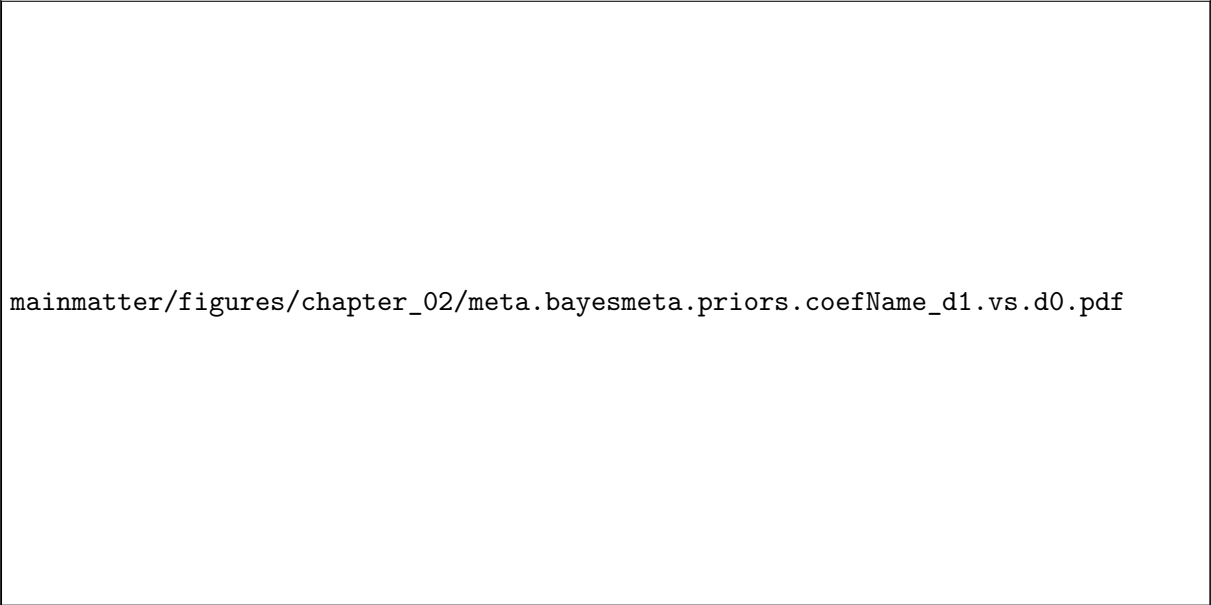


Figure 2.16: Normal prior for μ used for **bayesmeta** (blue), compared to the empirical distribution of per-gene frequentist **metafor::rma** estimates for τ , for the day 1 vs. baseline effect. The non-scaled normal fit is shown (black), as well as a Cauchy fit (red).

2.2.9.6 Multiple testing correction

For the frequentist random-effects meta-analysis, nominal gene-wise p -values are converted to false discovery rate (FDR) estimates using the Benjamini-Hochberg (BH) procedure (`p.adjust` in R). For the Bayesian random-effects meta-analysis, posterior effect sizes and standard errors are supplied to `ashr`, which estimates the local false sign rates (lfsrs), which are analogous to FDR, but quantifies the probability of calling the wrong sign for an effect rather than the confidence of a non-zero effect[156].

add note on ositive regression dependency [28]

add comment on symmetry

2.2.10 Gene set enrichment analysis using blood transcription modules

Gene set enrichment analyses were conducted using `tmod::tmodCERN0test`[157], which assesses the enrichment of small ranks within specific sets of genes compared to all genes, when the genes are ranked by some metric—here I used effect sizes from `bayesmeta`.
The gene sets used were blood transcription modules (BTMs) from[158], which are annotated sets of coexpressed genes mined from publicly available human blood transcriptomic data, and provide sets tailored for enrichment analyses in blood cells.

2.3 Results

more text

2.3.1 Extensive global changes in expression after vaccination

To gain an overview of how the transcriptome changes after vaccination, linear models were fit to identify genes differentially expressed at day 1 or day 7 compared to baseline (day -7 and day 0) in the HIRD array and RNA-seq expression data, accounting for covariates such as batch effects, sex, age, TRI, and ancestry. At 13593 genes with expression measured by both platforms, models were fit within each platform, then effect sizes were combined using Bayesian random-effects meta-analysis.
At a $\text{lfsr} < 0.05$ and absolute $\text{FC} > 1.5$ cutoff, 857/13593 genes were differentially expressed between any pair of timepoints, with their expression clustering into three main clusters (Figure 2.17).

2.3.2 Innate immune response at day 1 post-vaccination

Consistent with global expression at day 1 being markedly different from expression at other timepoints (Figure 2.14), the highest numbers of differentially expressed genes are observed at day 1, with 644 genes differentially expressed vs. baseline. The majority of these (580/644) were upregulated. The gene with the highest FC increase at day 1 compared to baseline was *ANKRD22* ($\log_2 \text{FC} = 4.49$), an interferon-induced gene in monocytes and dendritic cells (DCs) involved in antiviral innate immune pathways[159]. Other key genes in the interferon signalling pathway[160] such as *STAT1* ($\log_2 \text{FC} = 2.1693060$), *STAT2* ($\log_2 \text{FC} = 0.9489341$), and *IRF9* ($\log_2 \text{FC} = 0.8153674$) are also upregulated at day 1. Gene set enrichment analysis using `tmod` revealed that genes with the high FC increases at day 1 were enriched in modules associated with activated DCs, monocytes, toll-like receptor and inflammatory signalling (Figure 2.18).



Figure 2.17: Normalised gene expression for genes differentially expressed between any pair of timepoints ($\text{lfsr} < 0.05$, absolute fold change > 1.5) across HIRD samples, clustered by gene (Manhattan distance metric).

confirming that day 1 responses are dominated by signatures of innate immunity. 64 genes were downregulated at day 1, enriched in modules associated with T cells and natural killer (NK) cells, with the largest absolute fold change observed for *FGFBP2* ($\log_2 \text{FC} = -0.9141547$). For both up and downregulated genes, there was a tendency to return to baseline expression levels by day 7.

2.3.3 Adaptive immune response at day 7 post-vaccination

59 genes were differentially expressed at day 7 vs. baseline, with expression fold changes more modest than those at day 1. The genes with the highest upregulation were the B cell-associated genes *TNFRSF17* ($\log_2 \text{FC} = 1.7538617$) and *MZB1* ($\log_2 \text{FC} = 1.7369668$). Plasma cell-specific genes including *SDC1* (encodes CD138 <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5437827/>) ($\log_2 \text{FC} = 1.3673081$) and *ELL2* (<https://www.nature.com/articles/ni.1786>) ($\log_2 \text{FC} = 0.8679659$) were also prominently upregulated. Strongly enriched modules at day 7 were related to mitosis and cell proliferation, particularly in CD4⁺ T cells (Figure 2.18). Both the CD4⁺ T cell and plasma cell response are indications of an adaptive immune response at day 7.

2.3.4 Expression signatures associated with antibody response

I also looked for genes which have expression associated with baseline-adjusted antibody response, as quantified by TRI. At the initial frequentist meta-analysis stage, with a significance threshold of $\text{FDR} < 0.05$, 6 genes had expression associated with TRI at baseline, 55 at day 7, and 11 pooling samples across timepoints (Figure 2.19). [106] also identified genes with day 7 expression associated with antibody response, where response was defined as a binary phenotype based on 4-fold change (described in section). They reported 62 significant associations at $\text{FDR} < 0.05$, of which 58/62 fall into the 13593 genes considered in my meta-analysis (circled, Figure 2.19), and 15/58 replicated, all with the same positive direction of effect (high expression with high TRI). In the Bayesian meta-analysis, no single gene was detected as significantly associated with TRI at $\text{lfsr} < 0.05$ at any timepoint, or when pooling samples across all timepoints (Figure 2.20).

Significant enrichments were detected at the gene set level; the strongest effects are seen at day 7, where expression of cell cycle, CD4⁺ T cells, and plasma cells are associated with high TRI. At day 0, modules related with inflammatory response in myeloid cells are also associated with high TRI (Figure 2.21).

2.3.5]

Identifying expression signatures for predicting antibody response [probably cut this section and just add to discussion]

can also add MSigDB hallmark sets, which include interferon sets; and of course gene ontology sets

not sure of interpretation at FGFBP2, it is indeed highly expressed in NKs through <https://dice-database.org/genes/FGFBP2>

any point in a table of e.g. top 20 DE genes, or is the gene set analysis already enough?

change x axis labels to baseline, specify top 10 procedure in figure caption

finish citing

add label

figure x labels here should be TRI, not R.vs.NR

mainmatter/figures/chapter_02/compare_dge_eqtl.tmodDotPlot.DGE.timepoint.pdf

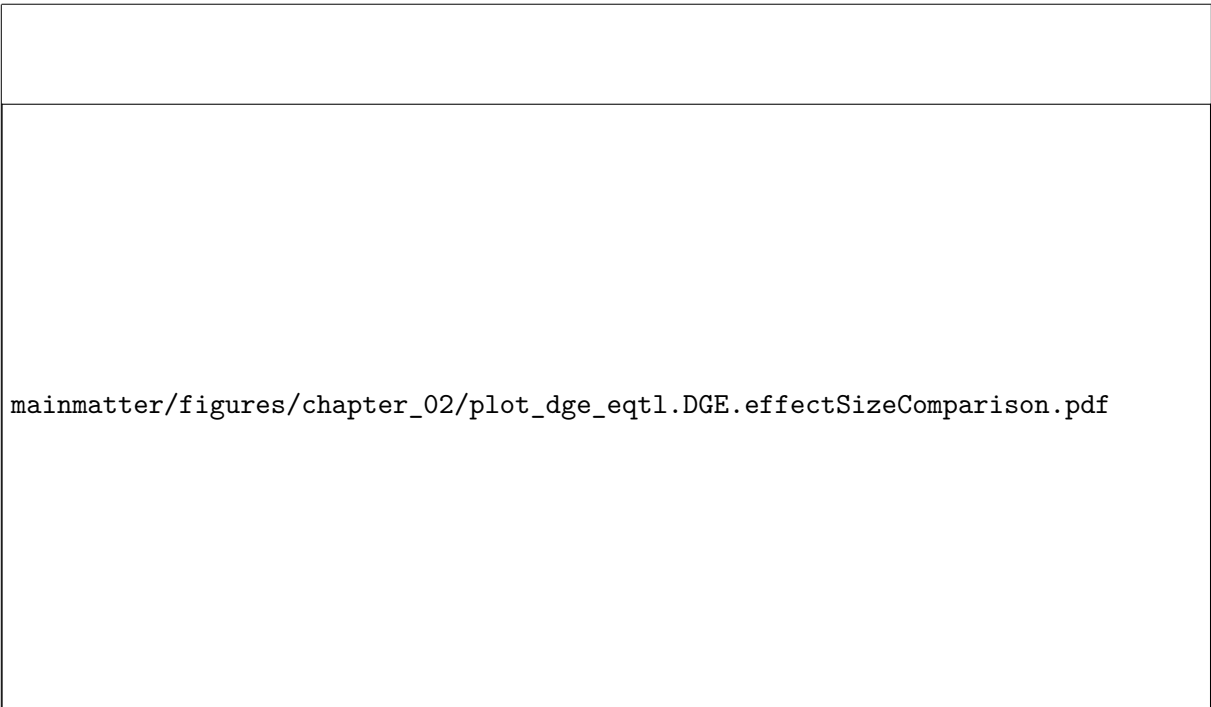


Figure 2.19: DGE effect sizes estimated in array vs. RNA-seq. Significance colored by frequentist random effects meta-analysis $FDR < 0.05$. Genes with day 7 expression associated with responder/non-responder status in [106] are circled for that contrast.

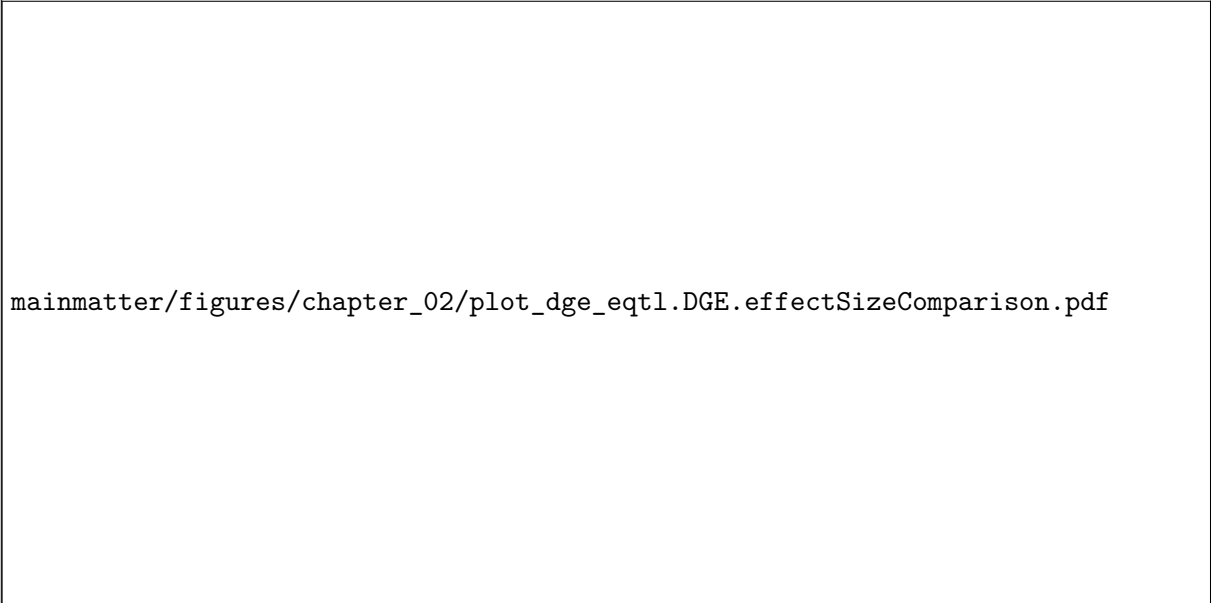


Figure 2.20: DGE effect sizes estimated in array vs RNA-seq. Significance colored by Bayesian random effects meta-analysis $lfsr < 0.05$. Genes with day 7 expression associated with responder/non-responder status in [106] are circled for that contrast.

mainmatter/figures/chapter_02/compare_dge_eqtl.tmodDotPlot.DGE.TRI.pdf

2.4 Discussion

There is extensive transcriptomic response to Pandemrix vaccination in the HIRD cohort. Upregulation of genes and modules related to the interferon signalling pathway, monocytes, inflammatory response, and other aspects of innate immunity were detected at day 1. This response is transient, with most such genes returning to baseline expression by day 7. Upregulation of cell cycle/proliferation, activated CD4⁺ T cell, and B (plasma) cell genes and modules were detected at day 7. This is likely a signature indicating the shift to an adaptive immune response, involving CD4⁺ T cell-supported differentiation and proliferation of antibody-secreting plasmablasts and plasma cells[162]. These patterns of expression change between timepoints in the RNA-seq data are consistent with the patterns in the array data in the original study[106], and with expansions of monocyte and plasma cell populations seen in the FACS data at days 1 and 7 respectively in the original HIRD study[106].

In contrast, I was not able to fully replicate the originally reported single gene-level associations between day 7 expression and antibody response in the RNA-seq data and subsequent and meta-analyses. In [106], 62 genes were reported as differentially expressed between vaccine responders and non-responders. Although [106] encodes responder status as a binary phenotype, whereas my analysis uses TRI, this is not the primary difference, as 51/62 genes replicated (FDR < 0.05) using TRI when considering just the array data. The same analysis using only the RNA-seq data replicated 0/62 genes. The majority of the effects for these genes were simply much stronger in the array dataset than in the RNAseq dataset (Figure 2.19). Given that the range of TRI is higher in the array individuals (Table 2.1), this does not seem unusual that stronger TRI-associated effects are observed there.

58/62 reported hits were measured by both platforms and assessed in the meta-analysis. Only 15/58 signals replicated using frequentist random-effects meta-analysis to combine per-platform estimates. I do not consider these hits as robust, as the REML estimate of between-platform heterogeneity was zero for 8563/13593 for the day 7 TRI contrast overall, and zero for all 15 of these signals. None of these signals replicated in the Bayesian random-effects meta-analysis. The Bayesian meta-analysis is in general more conservative, calling fewer differentially expressed genes compared to the frequentist analysis for all contrasts (Figure 2.20). Prior information about τ is incorporated, discouraging unrealistic estimates of zero heterogeneity. Given the between-platform heterogeneity coming from both platform-specific technical differences and TRI phenotype differences, relative to the modest effect size distributions compared to between-timepoint DGE comparisons, the data are not well-positioned to identify significant single-gene associations with antibody response.

Expression signatures of antibody response were, however, observed at the gene set level, for modules of coexpressed genes that are associated with TRI as a whole. The strongest effects were observed at day 7, where expression of adaptive immune response modules (cell cycle, stimulated CD4⁺ cell, plasma cell modules) were positively associated with TRI. These are the same modules observed to be upregulated at day 7 compared to baseline; it seems that those individuals with the greatest antibody response to vaccination are most able to upregulate these gene sets by day 7 post-vaccination.

Module associations were also observed pre-vaccination (cell adhesion, enriched in B cells,

Not sure if there is a biological interpretation of downreg of T cells and NK cells gene sets at day 1, since it could be due to increase in other cell types in the sample. similar findings in [161] though

lit search for downregulation interpretation paper, and downreg T cell paper

might have to rerun everything using the original binary R/NR if this line of reasoning isn't strong enough

move numbers to results?

proinflammatory cytokines, platelet activation), suggesting baseline immune state has some influence on long-term antibody response to Pandemrix. Over the years, a diverse range of gene sets have been found to be baseline predictors of serological response to influenza vaccination: apoptosis[107]; Fc γ receptor-mediated phagocytosis, TREM1 signaling[108]; enriched in B cells, T cell activation[109]; B cell receptor signalling, inflammatory response, platelet activation [110]; several of which I also observe. It should be noted that comparisons with these signatures from existing influenza systems vaccinology studies should caveated, as most existing studies are for non-adjuvanted influenza vaccines. Adjuvanted influenza vaccines are considerably more immunogenic, and post-vaccination expression patterns differ to those of non-adjuvanted vaccines [104, 106]. Hence, it is particularly important that the robustness of these observed baseline expression signatures be validated in an independent cohort for a comparable AS03-adjuvanted influenza vaccine.

In conclusion, Chapter 2 characterises the expansive changes in PBMC gene expression that follow vaccination with Pandemrix. The dominant trend for all individuals is transient upregulation of the innate immune response at day 1, transitioning into adaptive immunity by day 7. Baseline-adjusted antibody response is correlated with expression of gene sets, particularly adaptive immunity modules at day 7, but also for some modules pre-vaccination. Unfortunately, between-platform variation in expression impedes identification of specific genes that contribute. The fundamental question of why gene expression and antibody responses vary between HIRD individuals remains. Chapter 3 will examine one hypothesis: the impact of common human genetic variation on Pandemrix expression response.

could comment on phenotype differences too, i.e. HIRD measure antibodies at d63, much later than is popular in the field: d28 usually

should probably emph sobolev didn't find pre-vacc signatures, and we did. But it's not exactly fair, as sobolev didn't use gene set enrichment as far as i can tell

At no point in this chapter are we estimating causal effects: add point on why not CIT (lower n than franco)

found signatures, but so what? Feels like chapter lacks a punchline?

Chapter 3

Genetic architecture of transcriptomic response to Pandemrix vaccine

3.1 Introduction

3.1.1 Genetic factors affecting influenza vaccine response

Vaccination is the most effective way by which seasonal influenza is controlled[90], and the mechanism by which influenza vaccines are efficacious is by raising strain-specific antibodies protective against future infection[163]. Humoral responses are influenced by vaccine-associated factors (e.g. type, dose, adjuvants), but are also a complex trait influenced by host genetics[83, 86]. Genetic variants associated with antibody response have been detected for vaccines such as hepatitis B, influenza, measles, rubella, and smallpox[83, 88]. For antibody response to seasonal influenza vaccines, studies have implicated genetic variation within cytokine genes, cytokine receptors[164]; antigen processing and intracellular trafficking genes[71]; immunoglobulin heavy-chain variable region loci[165]; and specific human leukocyte antigen (HLA) alleles [164, 166].

A potential mechanism through which genetic variation can play a causal role in influenza vaccine response is through altering the expression of genes as expression quantitative trait loci (eQTLs). eQTL can have condition-specificity: an interaction between their effect on expression and different environmental contexts such as tissue or cell type[46, 47]. The mechanisms by which eQTLs interact with environment are of great interest; for example, cell type-specificity can inform us about how expression is regulated in a cell type-specific manner[60]. In a vaccination context, an important subset of environment-interacting eQTLs are response expression quantitative trait loci (reQTLs), defined as an eQTL whose effect interacts with external stimulation or perturbation. reQTL have been observed in many human cell types *in vitro*, or in the whole organism *in vivo*. As the pre- and post-stimulation environments are separated in time, a possible mechanism that leads to the observation of reQTL is a genotype-dependent change in gene expression between timepoints, which may underly genotype-dependent differences in antibody phenotypes.

pull in citations from intro

3.1.2 Response expression quantitative trait loci for seasonal influenza vaccination

reQTL can be mapped considering a vaccine as an *in vivo* immune stimulation, looking for genotype-dependent changes in gene expression in immune cells. Little work has been done on vaccine-stimulated reQTLs, except one study conducted for the seasonal trivalent inactivated influenza vaccine (TIV). [71] collected longitudinal data in 247 European adults: peripheral whole blood gene expression measured at four timepoints (day 0, 1, 3, 14), and antibody titres measured at three timepoints (day 0, 14, 28). They identified 20 genes with a cis-eQTL effect, expression correlation with antibody response, and either post-vaccination differential expression *or* a reQTL effect at that cis-eQTL. Genes involved in intracellular antigen transport and processing were enriched among those 20 genes.

3.1.3 Chapter summary

The HIRD cohort represents a unique opportunity for detecting genetic contributions to influenza vaccine response. In chapter 2, we observed global changes in gene expression after Pandemrix vaccination, as well as expression signatures correlated to degree of antibody response. For seasonal influenza vaccines, the contribution is small: antibody responses in adults are largely driven by non-genetic influences such as previous influenza vaccination or infection[167]. As the Pandemrix vaccine is against a pandemic strain that was not in seasonal circulation at the time the Human Immune Response Dynamics (HIRD) cohort was recruited (2010-11), with individuals mounting an expression response that was not recall-dominated [106], the relative contribution of genetic factors to Pandemrix response may be greater.

In this chapter, I model the influence of common host genetic variation on longitudinal *in vivo* expression response to Pandemrix. I map cis-eQTL within each timepoint, accounting for ancestry, cell type abundance and unmeasured covariates, then call shared and reQTL effects from a joint model, looking for genes where the lead eQTL has a different effect size pre- and post-vaccination. Many of the strongest reQTL effects involve opposite signed effects on expression for the same variant at different timepoints. I detect a strong day 1 specific reQTL effect at *ADCY3*. Through modelling interaction of reQTL with cell type abundance estimates and statistical colocalisation with cell type-specific QTL datasets, the reQTL signal was determined to be a monocyte-specific effect likely driven by increase monocyte abundance at day 1.

3.2 Methods

3.2.1 Genotype phasing and imputation

Prior to imputation, 213277 monomorphic variants that provide no information for imputation Variant alleles were aligned such that the reference allele matches the GRCh37 reference, and 358 indels were removed. Imputation for the autosomes and X chromosome was conducted using the Sanger Imputation Service, which involves pre-phasing with EAGLE2 (v2.4) and imputation with PBWT (v3.1) against the Haplotype Reference Consortium (r1.1) panel <https://www.ncbi.nlm.nih.gov/pubmed/27548312>. Imputed variants were lifted-over from GRCh37 to GRCh38

distinction between expression/ab response is blurry here

straighten out tenses

1 more sentence to round off context

coordinates using CrossMap. Poorly-imputed variants with $\text{INFO} < 0.4$ or post-imputation missingness $> 5\%$ were removed, leaving 40290981 variants.

3.2.2 Overall strategy for detecting reQTLs

Since the aim of this chapter is to identify genetic variation that affects expression response to vaccination, it may seem most direct to model the change in each individual's expression after vaccination as the response variable. This approach has been applied for identification of condition-specific eQTL, typically with the response taking units of log fold change between conditions (e.g. [64, 168, 169]). Although potentially powerful if eQTL effects are small and opposite between conditions[64], it is analogous to the “change score” approach, which can suffer from regression to the mean (??), and increased uncertainty from the variance sum law if effects between conditions have positive covariance[115, 170].

Instead, I map eQTLs within each of three timepoint conditions (day 0 pre-vaccination, day 1, and day 7), and find reQTLs by looking for eQTLs that have different effects between conditions. Unlike a test for difference implemented using a genotype-condition interaction term in a joint regression model, homoscedasticity of errors is not assumed for all conditions[171]

Within each timepoint, recall the the HIRD dataset includes expression measured by both array and RNA-sequencing (RNA-seq). As discussed in subsubsection 2.2.9.2, it is difficult to directly estimate the between-studies heterogeneity when the number of studies is small, and Bayesian meta-analysis was preferred for combining array and RNA-seq differential gene expression (DGE) estimates. That method does not scale to eQTL analysis, where the number of tests is large, in the order of thousands of tests per gene, versus the handful DGE contrasts per gene performed in chapter 2. Instead, I perform a mega-analysis within each timepoint, first merging array and RNA-seq expression estimates into a single matrix with ComBat[132]. For comparison purposes, analyses were also run using in the array and RNA-seq samples separately.

Defining whether an eQTL is shared between conditions can be a tricky business. Naively, one can map eQTLs separately in each condition, then assess the overlap of significant associations between conditions. This underestimates sharing due to the difficulty of distinguishing true lack of sharing from missed discoveries from incomplete power within each condition [62, 172]. Condition-by-condition analysis also cannot borrow information across conditions for mapping shared associations[172–174]. Counterintuitively, a joint multivariate analysis may be more powerful even when associations are not shared across all conditions[175].

A variety of models have been employed for joint eQTL mapping, including the use of classical multivariate methods such as multivariate analysis of variance (MANOVA)[176], frequentist meta-analyses (e.g. **Meta-Tissue**[177], **METASOFT**), and Bayesian models (e.g. **eQtlBma**[172], **MT-HESS**, **MT-eQTL**). Joint mapping has been repeatedly been demonstrated to be more powerful than condition-by-condition analysis, and recent methods are now computationally efficient when scaling to large numbers of conditions and variants tested (e.g. **RECOV**[178], **mashr**[173], **HT-eQTL**[174]). In this chapter, I apply **mashr**[173] for the estimation of eQTL effects across my three timepoints. **mashr** learns patterns of correlation among multiple conditions empirically from condition-by-condition summary statistics, then applies shrinkage to provide improved posterior effect size estimates, and compute measures of significance per condition.

upend change score bit, here expression is a y variable. compare all the options as glms etc.

Can this really demonstrate genotype-dependent change in gene expression between time-points? i.e. need understand how the change score/ANCOVA approaches differ from repeated measures ANOVA differ from the interaction/stratified approach I take?

why I didn't just do a mega-analysis in chapter 2 then, given I haven't any evidence if it's better or worse than Bayesian meta-analysis in that context.

add -7 note as with ch2

3.2.3 Controlling for population structure with linear mixed models

There is population structure due to ancestry in the HIRD cohort, which was incorporated in DGE analyses by treating the top principal components (PCs) of the genotype matrix as continuous covariates for large-scale population structure (subsection 2.2.5). In the context of eQTL mapping (and genetic association studies in general), where the aim is to assess the marginal effect of a single genetic variant on expression, population structure can be correlated with both expression (e.g. through polygenic effects) and the tested variant (e.g. through ancestry-dependent frequency differences). This leads to omitted-variable bias (OVV) from confounding, which if not controlled for, leads to genome-wide inflation of test statistics [179]. An useful approach is the linear mixed model (LMM) with a random effect that incorporates genetic correlation between individuals, usually in the form of a kinship matrix, into the covariance of that random effect[117, 179, 180] The LMM approach has the advantage of not only modelling large-scale population structure, but also cryptic relatedness (the presence of closely related individuals in a sample assumed to consist of unrelated individuals[181]) due to finer-scale effects such as family structure[180].

3.2.3.1 Estimation of kinship matrices

When testing a variant for association using LMMs, to avoid loss of power from “proximal contamination”, the kinship matrix used should not include that variant[182]. A simple way to avoid this is to compute a leave-one-chromosome-out (LOCO) kinship matrix using all variants on chromosomes other than the tested variant’s chromosome[183].

I estimated kinship in the HIRD data from common autosomal variants, using LDAK (5.0), which computes kinship matrices adjusted for bias caused by linkage disequilibrium (LD)[184]. Filtered, pre-imputation sample genotypes from subsection 3.2.1 were pruned to MAF > 0.05. A kinship matrix was computed for each autosome, then combined into a single genome-wide matrix using LDAK --join-kins. To obtain a LOCO kinship matrix for each autosome, each autosome’s kinship matrix was then subtracted from this genome-wide matrix (LDAK --sub-grm).

3.2.4 Additional eQTL-specific expression preprocessing

There are a number of transformations often applied to expression data before eQTL mapping, such as the rank-based inverse normal transformation (INT) (e.g. GTEx v8[185]), which converts often non-normal expression data to an approximately normal distribution, and reduces the impact of expression outliers. In the context of genetic association studies, the practice of applying rank-based INT to phenotypes has been criticised for only guaranteeing approximate normality of residuals when effect sizes are small, and potential inflation of type I error, especially in linear models that include interactions[186]. In multi-condition datasets, these transformations are also typically applied within conditions (e.g. within each tissue individually in GTEx v8[185]). Another common transform is standardising (centering and scaling to zero mean and unit variance) (e.g. eQTLGen Consortium[48]), often done so that effects across genes and studies can be comparably interpreted in units of standard deviation expression[187].

add some indication of how much inflation can be reduced by LMMs

add chr1 loco kinship matrix as example, note the estimates for self-relatedness on the diagonals are not constrained to be 1

helps with coloc

emph here that the sims match what my def of reqtl is for rest of chapter

I performed simulations to evaluate the effect of these transformations on reQTL detection between a hypothetical baseline and day 1 post-vaccination condition. Expression values on the log scale were simulated with the eQTL slope (beta) set to specific values corresponding to six scenarios for six gene-variant pairs (Figure 3.1). The simulated scenarios were subjected to rank-based INT (Blom method[186]), standardisation (both centering and scaling), scaling-only, and centering-only transformations. Transformations were applied both within each condition and without separating conditions.

log scale: as interactions depend on the scale at which departure from additivity is detected

The boxed facets in Figure 3.1 represent undesirable effects of transformations on reQTL calls. For example, rank-based INT induces false shared eQTL effects in scenarios 4 and 5. In general, transformations that scale within condition are not appropriate, as different variance between conditions can be what drives a reQTL effect. Scaling without separating conditions can also be problematic, since the total variance also contributes to the reQTL effect size. For example, scenarios 2 and 4 have the same 1 unit increase in slope pre-transformation (the same fold-change between conditions), but after scaling-only the beta increases are $0.75 - 0 = 0.75$ and $0.8 - 0.4 = 0.4$ respectively—eQTL 4 now looks like a weaker effect.

In light of these simulations, I decided that neither rank-based INT nor standardisation were appropriate given my intent of detecting reQTLs between conditions. Only the centering-only transformation avoided both false shared effects and preserves relative reQTL effect sizes between genes. The simple inclusion of an intercept term in the eQTL model already achieves this. Not performing any rank-based transform does lose the advantage of reining in outliers. The expression data have already been preprocessed to remove low-expression outliers in subsection 2.2.7, but automatic outlier exclusion based on standard deviation (SD) thresholds at the eQTL mapping step could be considered in future implementations[48]. Note that many preprocessing steps done prior to this stage in the pipeline (e.g. variance-stabilisation, ComBat batch effect correction) are also expression transformations, but I only consider the preservation of reQTL effects defined from expression values post-adjustment for those technical effects to be important.

add sample sizes and model for expression sim

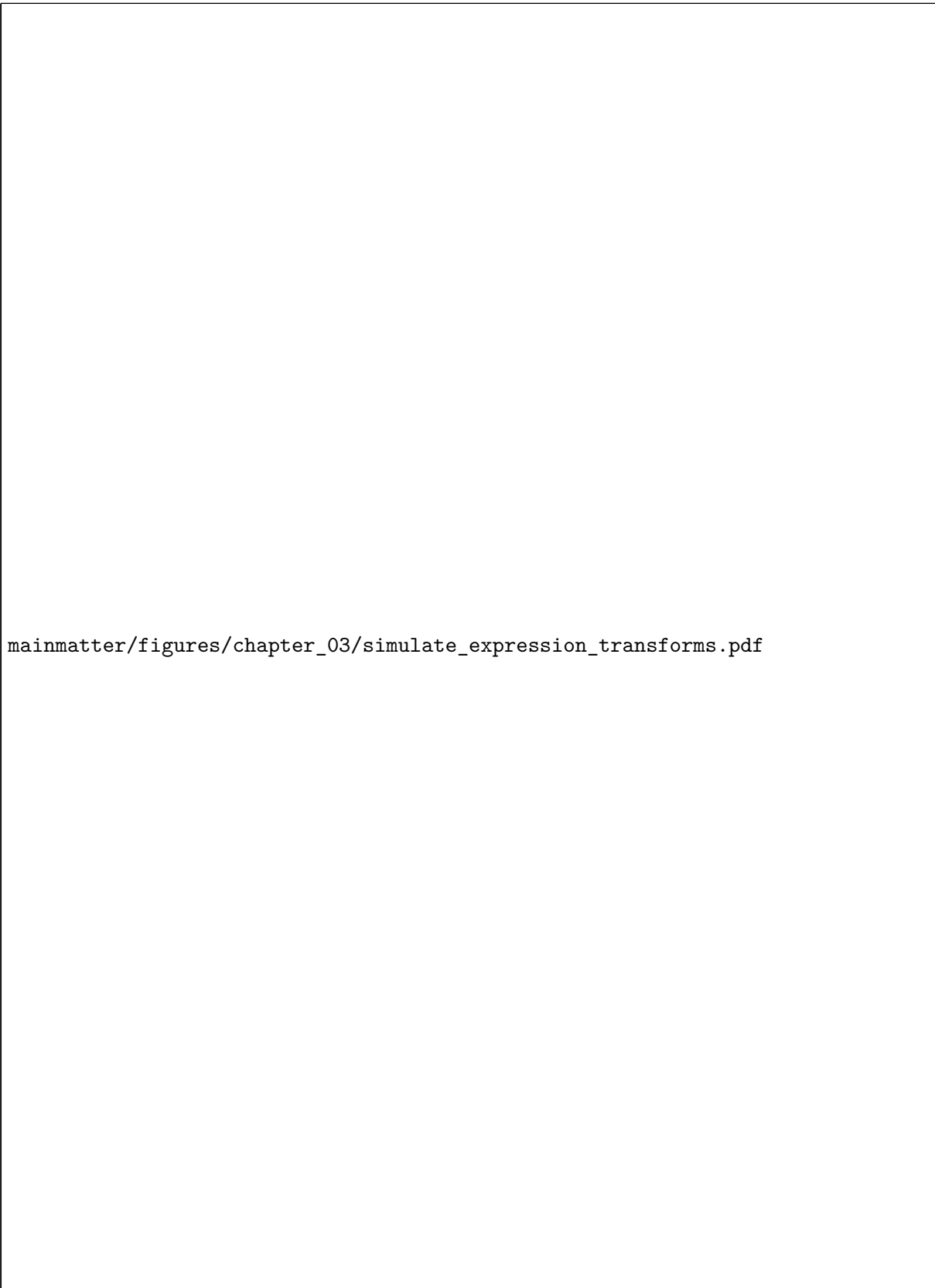
3.2.5 Estimation of cell type abundance from expression

Peripheral blood mononuclear cell (PBMC) samples are a mixture of immune cells, and a fixed input of RNA extracted from that mixture is used to estimate expression, so estimates for genes that have cell type-specific expression depend on the relative proportions of each cell type in each sample. These proportions shift after Pandemrix vaccination[106], and eQTL effects can also be cell type-specific. As genotype can be assumed to stay constant, it is valid to compare the effect of genotype on expression between multiple timepoints to call reQTLs, but changes in cell type abundance influence this by both expression and the effect of genotype on expression. Immune cell abundance also varies naturally between healthy individuals[79, 167], so it is important to model these effects even at baseline.

determine appropriate citations from existing refs in intro

??

Cell type abundance directly measured via fluorescence-activated cell sorting (FACS) are only available for a small subset of HIRD individuals (subsection 2.2.1), so I derived cell type abundance estimates from the expression data as an alternative. Such estimates have previously been used in eQTL analyses from bulk samples where cell type-specific effects are expected[57,



mainmatter/figures/chapter_03/simulate_expression_transforms.pdf

Figure 3.1: Simulated log scale expression in two conditions for six genes (columns) representing six different scenarios: Scenario 0 has no eQTL, scenario 1 is a shared eQTL ($\beta = 1$), scenario 2 is a reQTL where β increases from 0 to 1, scenario 3 is a reQTL where β increases from 0 to 2, scenario 4 is a reQTL where β increases from 1 to 2, and scenario 6 is a reQTL where β increases from 1 to 4. Rows represent the effect of different expression transformations across samples, conducted both within condition, and including both conditions.

60, 73]. As the estimates are based on the expression of multiple genes, is not entirely circular to use them as covariates in this way for genewise eQTL models. I selected `xCell`[188], which previously been shown to outperform other deconvolution methods for cell type-specific eQTL mapping in blood[60]. `xCell` computes enrichment scores based on the expression ranks of approximately 10000 signature genes derived from purified cell types, works for both array and RNA-seq expression data, and implements “spillover compensation” to reduce dependency of estimates between related cell types[188]. `xCell` was originally developed for tumor samples, so many of the built-in cell types are not expected in PBMC. Reviewing the literature to find which broad classes of peripheral blood cell types are commonly-expected in the PBMC compartment[73, 189, 190], I selected 7/64 of the built-in cell types: CD4⁺ T cells, CD8⁺ T cells, B cells, plasma cells, natural killer (NK) cells, monocytes, and dendritic cells (DCs). Array and RNA-seq data from subsection 2.2.8 and subsection 2.2.7 were processed through `xCell` separately. The large batch effect present in the array expression was first removed using ComBat. Finally, enrichment scores were standardised, so that a score of zero estimates the average abundance of that cell type across all timepoints (Figure 3.2 and Figure 3.3).

As with actual cell type abundances, the enrichment scores are correlated. Multicollinearity will be a problem for interpreting effect size estimates when these scores are used as indepedent variables in regression downstream. * To prune the number of scores, I performed a principal component analysis (PCA) of the cell type scores across samples, determined the number of principal components that exceed the eigenvalues-greater-than-one rule of thumb[192], then selected only the one cell type with the highest contribution for each of those components. In both array and RNA-seq datasets, the number of components retained was three, and the selected cell types were monocytes, NK cells, and plasma cells (Figure 3.4). The choice to use the actual cell type scores over principal components directly as covariates is a sacrifice of orthogonality for interpretability.

Scores were validated against FACS measurements in the subset of individuals that had them. Depending on each panel’s gating strategy for each cell subset, the FACS data were in units of either absolute counts, or percentage of the previously gated population. A rank-based INT was applied within each panel and cell subset, so that the transformed measure could be compared between individuals for each subset ([193] takes a similar approach for cell abundance data using a quantile-based INT). Missing values were imputed with `missForest`, a random forest imputation method suitable for high-dimensional data where $p \gg n$. Although the increase in `xCell` score for monocytes at day 1 and plasma cells at day 7 reflect the increases in these cell types observed by[106], overall correlation between `xCell` and FACS was weak (Figure 3.5). Weighing the downside of having imperfect estimates of cell type abundance against the downsides of not accounting for abundance, or excluding samples without FACS measures, I chose to continue the analysis using the `xCell` scores.

*high intercorrelation is not necessary nor sufficient by itself to induce multicollinearity, but multiple correlation does have an inverse relationship with the standard error of coefficient estimates [191]

add comment on existence of chosen cell types in samples, and clustering by visit

does not bias least squares regression, but unstable (vary sample to sample) to changes in data due to sampling var, and more std error of estimates will be high (tending to inf if perfect multi)

no need for both size and color, use one for contribution percent

add info on the markers used for the chosen FACS counterparts

get subset size

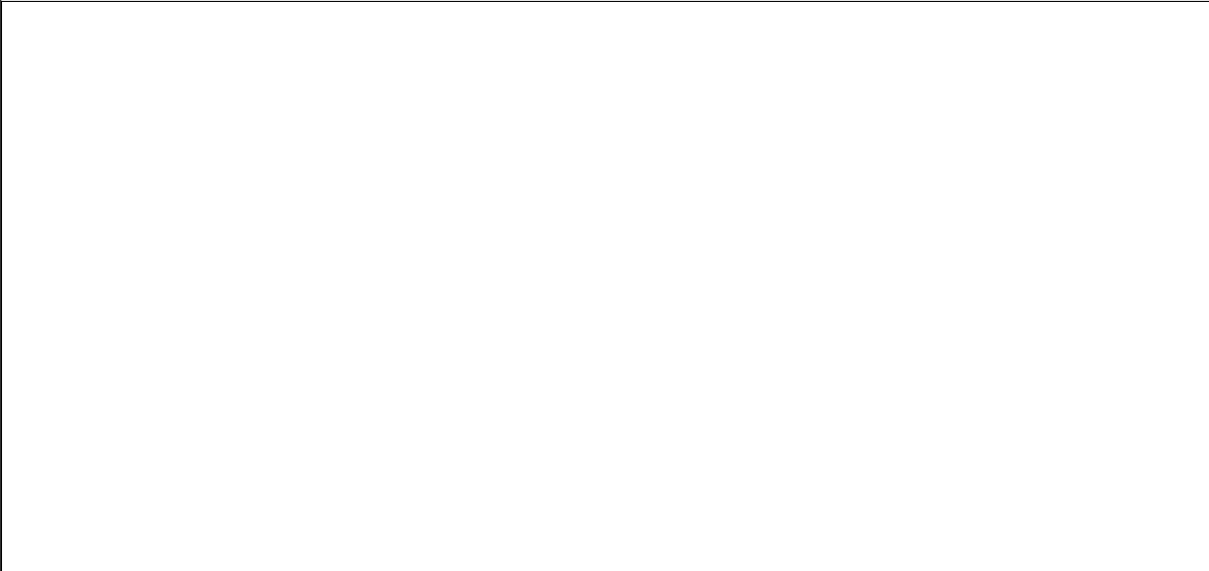
change corr scatterplots to corr matrix







(a) Array estimates.



mainmatter/figures/chapter_03/get_xCell_estimates.dataset_rnaseq.plots.pdf



3.2.6 Finding hidden covariates using factor analysis

Apart from cell type abundance, a myriad of other unmeasured variables contribute to expression variation. Hidden determinants of expression variation were learnt using PEER[194]. As suggested by [194], between-sample normalisation and variance stabilisation on RNA-seq count data was performed using DESeq2::vst. ComBat was applied to first merge array and RNA-seq data into a single log scale expression matrix per timepoint, treating the largest global effects on expression—the two array batches and three RNA-seq library prep pools (Figure 2.14)—as known batch effects. Given selected known covariates (intercept, sex, four genotype PCs from subsection 2.2.5 representing ancestry, and the three xCell scores estimated above), PEER was used to estimate additional hidden factors that explain variation in expression matrix. Factors are assumed to be unmeasured covariates that have global effects on a large fraction of genes, whereas a cis-eQTL will typically only have local effects, so including factors as covariates should not introduce dependence with the genotype term, but should soak up some of residual variation, improving power to detect cis-eQTLs. The analysis was run per timepoint, otherwise global changes in expression between timepoints induced by the vaccine would be recapitulated as factors.

Correlating the estimated factors to a larger set of known covariates reveals many correlations with xCell estimates, indicating that cell type abundance does indeed have substantial global effects on the expression matrix. There is little correlation with known array or RNA-seq batch effects, indicating ComBat did an adequate job of removing batch- and platform-dependent global effects on expression (Figure 3.6). Note that I did not leave this adjustment for PEER to perform, as ComBat estimates centering and scaling factors per gene to adjust for batch effects, whereas the use of PEER factors represent a mean-only adjustment. Given the severity of the batch effect in this dataset, especially between platforms, mean-only adjustment may be insufficient[135].

remake this with only top k factors, and prune the possible covariates

3.2.7 eQTL mapping per timepoint

The performance of various software implementations of LMMs specialised for genetic association studies are highly comparable; the specific choice of implementation can usually be made on the basis of computational efficiency[117]. I map eQTLs within each timepoint using LIMIX[195], which implements univariate and multivariate LMMs with one or more random effects.

Imputed genotype probabilities were converted to continuous alternate allele dosages using bcftools (1.7-1-ge07034a). Variants with sample AC < 15 within each timepoint were excluded.

add approximate MAFs, then cite hierarch paper

add note on treating x chrom variants with caution

At each of 13570 genes, at all cis-variants within ± 1 Mbp of the gene transcription start site (TSS), I fit the following model to map eQTL:

$$Y = 1 + sex + \sum_{i=1}^4 PC_i + \sum_{i=1}^3 xCell_i + \sum_{i=1}^k PC_i + \beta G + \mathbf{u} + \epsilon \quad (3.1)$$

lift proper vector notation from limix, then redo this with a timepoint subscript

where the eQTL effect size of interest is the slope of the genotype fixed effect β , the average additive effect of the alternate allele [9]; and \mathbf{u} is a random effect with zero mean and covariance

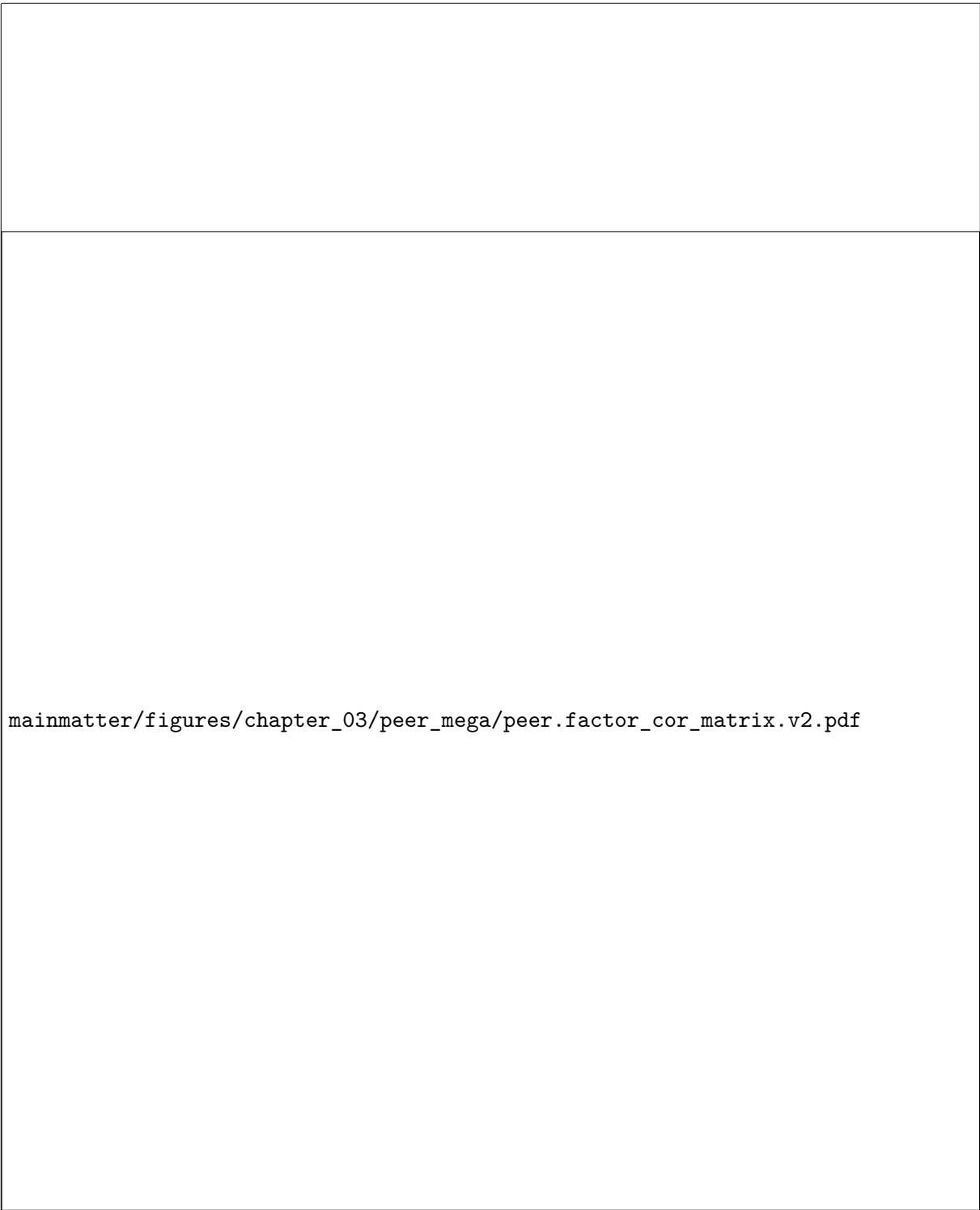


Figure 3.6: Correlation of PEER factors to known factors and other possible covariates. Note that PEER factors are not constrained to be orthogonal, so correlations to known factors are expected.

add formulation of the 0-mean random effect to show exactly how the kinship matrix is used [196]

note stacking of kinship for day -7 repeated measures

i leave the pcs in to guard against unusually differentiated between pop markers, where random effect alone may not be enough [179], <https://www.nature.com/articles/srep06874>

recheck if did I do a SNPs only filter

note this is critical, since we know a priori not independent due to eqtl sharing

move lfsr explanation prior to ashr in dge chapter

matrix proportional to the LOCO kinship matrix*.

PEER factors are automatically weighted such that the variance of factors tends to zero as more factors are estimated, hence continuing to add more and more factors as covariates will not continue to improve eQTL detection power, and eventually the model degrees of freedom will be depleted. To optimise k , the number of factors to include as covariates[†], Per-timepoint eQTL mapping was performed in chromosome 1, iteratively increasing the number of factors until the number of eQTLs detected plateaus. I settled on a final choice of $k = 10$ factors for pre-vaccination, 5 factors for day 1, and 5 factors for day 7 (Figure 3.7).

3.2.8 Joint eQTL analysis across timepoints

Joint analysis was conducted with `mashr`[173], at 40197618 gene-variant pairs (mean of 2962 tests per gene) for which summary statistics from within timepoint mapping were available in all three timepoint conditions. The `mashr` model incorporates multiple canonical (the identity matrix etc.) and data-driven covariance matrices to represent patterns of effects across conditions (in this case, 3×3 matrices). Data-driven covariance matrices are derived by dimension reduction of a strong subset of tests likely to have an effect in at least one condition. I took the most significant variant per gene per condition, which ensures strong condition-specific effects are included (Figure 3.8), then further filtered to only nominally significant tests, resulting in a strong subset of 45962 tests.

The `mash` model was trained on a random subset of 200000 tests, using the Exchangeable Z-scores model[173]. The correlation of null tests between conditions, critical to account for due to the repeated measures structure of the data, was estimated using `mashr::estimate_null_correlation`.■

The fitted model was used as a prior to compute posterior effects and standard errors for all tests through shrinkage. A condition-specific Bayesian measure of significance local false sign rate (`lfsr`) is returned, which can be interpreted as the the probability given the data, that the declared sign of the effect is incorrect.

3.2.9 Defining shared and response eQTLs

Many of the tested variants for each gene will be in high LD. To unambiguously select a lead eQTL variant per gene, I selected the variant with the lowest `lfsr` in any condition, breaking ties by highest imputation INFO, highest MAF, most upstream of the TSS, and genomic coordinate. Sharing was then evaluated for that gene-variant pair across all three conditions.

Thresholding on the `lfsr` is not appropriate for determining sharing, as the difference between significant and non-significant effect estimates in two conditions is not necessarily significant[198, 199]. [173] provides a heuristic that two effects are shared by magnitude if they have the same sign, and are also within a factor of 2 of one another, but this does not consider the posterior standard error of the estimates. Between a pair of effects in two conditions, I compute a z-score for the difference in effects[171, 198]:

*For chromosome X variants, no LOCO matrix is available from LDAK, so the matrix for chromosome 1 is used.

[†]I avoid the commonly-performed two-stage approach of treating PEER residuals as expression phenotypes, as the degrees of freedom seen downstream will be incorrect, which can have a substantial effect on estimates at this modest sample size.

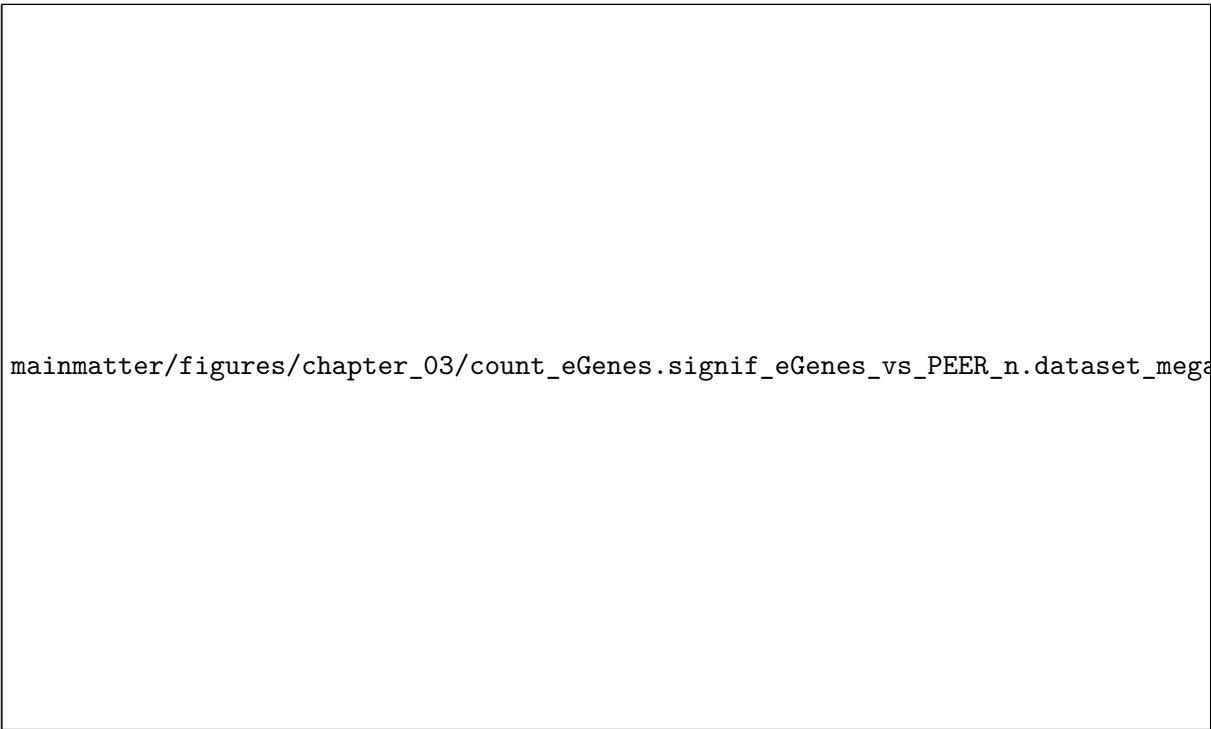


Figure 3.7: Number of significant eGenes detected on chromosome 1 (hierarchical Bonferroni-Benjamini-Hochberg (BH)[197] FDR < 0.05) as a function of the number of PEER factors included as covariates k.

$$z = \frac{\beta_x - \beta_y}{\sqrt{\sigma_x^2 + \sigma_y^2 - 2\sigma^2(x, y)}} \tag{3.2}$$

This strategy has been applied to call reQTLs by [75], assuming posterior pairwise covariance of effects is zero $\sigma^2(x, y)$. A Wald test p -value for the difference can be computed, as under the null hypothesis of zero difference, asymptotically $z \sim \mathcal{N}(0, 1)$. I use nominal p -value < 0.05 as a heuristic threshold (like the mashr recommended 2-fold threshold) to define reQTL effects that are strong, rather than a formal measure of significance. Effects are only compared if at least one of the two effects has $\text{lfsr} < 0.05$, to avoid sharing being driven by null effects.

3.2.10 Replication of eQTLs in a reference dataset

To validate the eQTL mapping approach, I estimate the replication of significant eQTLs in a large independent reference. Due to the lack of large sample size eQTL maps specific to PBMC, I use the GTEx v8 whole blood dataset as my reference dataset (n=670, 51.2% eGene rate). For lead variants called as significant in the HIRD dataset at a given lfsr threshold, I lookup the nominal p -value for that variant in GTEx (where the variant exists in both datasets). I applied `qvalue::qvalue_truncp` to estimate the proportion of those GTEx nominal p -values that are null (π_0), the compute a measure of replication $\pi_1 = 1 - \pi_0$.

The mega-analysis has comparable replication rate to RNA-seq-only analysis for shared eQTLs at moderately stringent lfsr thresholds up to 10^{-5} (Figure 3.9). Past this, as the π_1 procedure assumes a well-behaved p -valuedistribution in $[0, 1]$, reliability declines due to the

not sure whether this is conservative or anti-conservative

mashr does not provide by default



Figure 3.8: Clustering of within-timepoint Z scores in the strong mashr subset (random sample of 10000/45962 tests), confirming the presence of strong condition-specific effects.

number of p -values being too small*, or the maximum p -value being too far from 1. The numbers of reQTLs were too low to assess replication using this method, and one might not expect them to replicate in a baseline dataset such as GTEx whole blood, especially for those reQTLs significant only at post-vaccination timepoints. As the mega-analysis has a higher eGene rate (50.8 % vs. 29.9 %) compared to the RNA-seq-only analysis, with similar replication, I assume this represents a power advantage from having larger a sample size, rather than technical effects from merging the expression data.

RNAseq does test about 7000 more genes though...

3.2.11 Genotype interactions with cell type abundance

be more specific: "moderator", 'modify'????

If the abundance of a particular cell type does truly modify the eQTL effect, then an interaction term between genotype and cell type abundance is required, otherwise the regression slope of the eQTL term will be biased; one can not adjust for this modification just by including the main effect for cell type abundance. Given the modest sample size, I use the two-step approach used by others[57, 62, 73, 75], where tests for interaction are only performed at a subset of tests, often the lead eQTL variant for each gene. The key to the two-stage approach is that if the estimates for the interaction effect are sufficiently independent from the estimates of the main effect from main-effect only models, the type I error can be controlled based on the number of interactions that are actually tested, rather the number of interactions that could have been tested for[62, 200]. It is unclear whether this assumption holds, as the size of the main effect may contribute to power for detecting interaction effects. As the main purpose of the interaction analyses is scanning for cell type effects at detected reQTLs, I chose to test for interactions only at the lead eQTL variant for each gene with a significant main eQTL, then apply the BH false discovery rate (FDR), as used by others[62, 75].

point is, doesn't make sense to assume the genotype effect is the same at all levels of cell type abundance

Models in interactions between genotype and other predictors were fit using `lme4qt1`. The model specification identical to Equation 3.1, with the addition of three interaction terms between genotype and each xCell score. Significance is assessed using the likelihood-ratio test versus the nested model with no interaction terms.

can we interpret with peer in? add note of CLAIM here that although peer is correlated with xcell, interactions are only formed with xcell, so the interaction term can be interpreted per unit of genotype increase when xcell=0

3.2.12 TODO Statistical colocalisation

- if adding coloc analysis, add <https://github.com/jrs95/hyprcoloc> methods here - cannot deal with multiple causal within each group

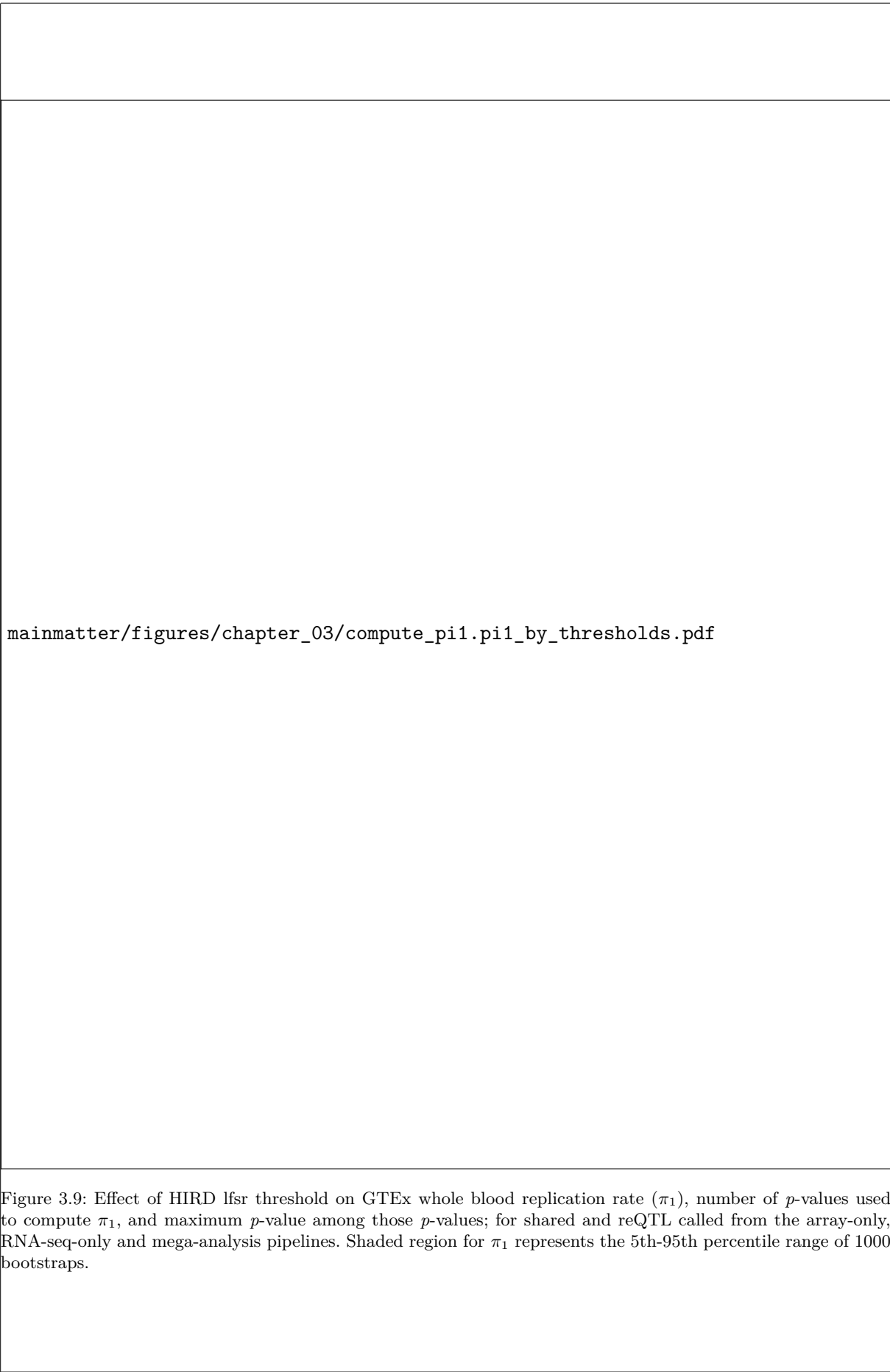
3.3 Results

this analysis is incomplete, and is one of the things I would suggest to round off this chapter

3.3.1 Mapping reQTLs to Pandemix vaccination

Within each timepoint condition (day 0 pre-vaccination, day 1, and day 7), cis-eQTLs ($\pm 1\text{Mb}$ of the TSS) were mapped using LIMIX, then joint analysis of effects was done using `mashr` to obtain posterior effect size and standard errors. At $\text{lfsr} < 0.05$, 6887/13570 genes (50.8 %) were eGenes (genes with a significant eQTL) in at least one timepoint. To sidestep the issue of multiple tested variants per gene being in LD, the most significant eQTL variant across all timepoints

*<https://github.com/StoreyLab/qvalue/pull/6#commitcomment-26277751>



was selected as the lead variant for each eGene, then reQTLs were defined by comparing the effect size of this lead eQTL between each pair of timepoints. Most eQTLs were shared across timepoints; 1154/6887 (16.8 %) eQTLs were classified as reQTLs between any pair of timepoints (nominal p difference < 0.05).

Figure 3.10 illustrates the difference between calling sharing using a significance threshold versus difference in betas approach. For instance, day 0 was the timepoint with the largest number of eGenes, reflecting the larger sample size compared to other timepoints. Although there are 1427 eGenes significant at only day 0, there are only 646/1427 reQTL among them, as the effect size at day 0 does not differ significantly when compared to day 1 or day 7 for the remainder. The strongest eQTLs with the highest proportion of variance explained (PVE)* are shared between timepoints, highlighting the power advantage for mapping shared effects granted by joint analysis.

3.3.2 Characterising reQTLs post-vaccination

As detection power is greatest at day 0, I focus on eQTLs that are reQTLs between day 0 and either day 1 or day 7 post-vaccination, and are significant at the corresponding timepoint: 819 reQTL between day 1 and day 0, and 1002 reQTL between day 7 and day 0 (Figure 3.11). Gene set enrichment analysis on the eGenes targets for these sets of reQTLs did not detect any significant enrichments (gprofiler2, g:SCS adjusted p < 0.05). Many of the reQTL that satisfy this criteria have opposite effects pre- and post-vaccination—as lfsr quantifies uncertainty in the sign of the effect, I do not compare the sign unless the reQTL is also significant at day 0. Shared eQTLs are enriched close to the TSS, whereas reQTLs are distributed across the cis- window.

The strongest reQTL at day 1 was for *ADCY3* (p difference = 8.68×10^{-6} , BH FDR = 0.118), where the reQTL variant explained approximately 1.9 % of expression variation at day 0, increasing to 14.1 % at day 1 (Figure 3.12). At day 7 the strongest reQTL was at *SH2D4A* (p difference = 1.37×10^{-6} , BH FDR = 0.0175). Here, the reQTL variant explained similar amounts of expression variation at day 0 (8.2 %) and day 7 (9.0 %), with opposite directions of effect (Figure 3.13). Both *ADCY3* and *SH2D4A* have moderately high percentile expression at all timepoints, and are not differentially expressed post-vaccination. Overall, compared to genes without reQTL, reQTLs were less likely be differentially expressed post-vaccination at day 1 (26.5 % for reQTL vs. 42.3 %, Fisher’s test p < 2.20×10^{-16}), and no significant difference was observed at day 7 (2.2 % for reQTL vs. 1.4 %, Fisher’s test p = 0.0509). Only 5/68 (13.2 %) genes with reQTLs that explain more variation at day 1 were upregulated at day 1 vs. day 0; 5/226 (2.2 %) for day 7 vs. day 0.

3.3.3 Genotype by cell type interaction effects

Given that many reQTLs are not explained by differential expression post-vaccination, the presence of cell type-specific eQTL effects was considered as an alternate explanation. As described in subsection 3.2.5, xCell enrichment scores were used to approximate abundance of seven PBMC cell types from the expression data. After pruning highly correlated cell types to avoid

*TODO: add to methods <https://journals.plos.org/plosone/article/file?type=supplementary&id=info:doi/10.1371/journal.pone.0120758.s001>

if it would be interesting to compare the sharing estimate condition by condition approach to mashr, then redo and pull in eigenmt-bh values

actually, i’ve found that my PVE approximation is basically rescaled abs(Z), so pve is a bit pointless if we already have z, and doesn’t really help with comparability between genes with diff var/MAF

requiring signif post-vaccination may not be correct, as it excludes many dampening effects

the lack of any positional enrichment makes me concerned for false positives? check with ASE?

expand this to plot 1, list top 5 damp, flip, amp at each timepoint

note anything in lit about any of the 30

reword not significant

double check denoms

convert to subfigures

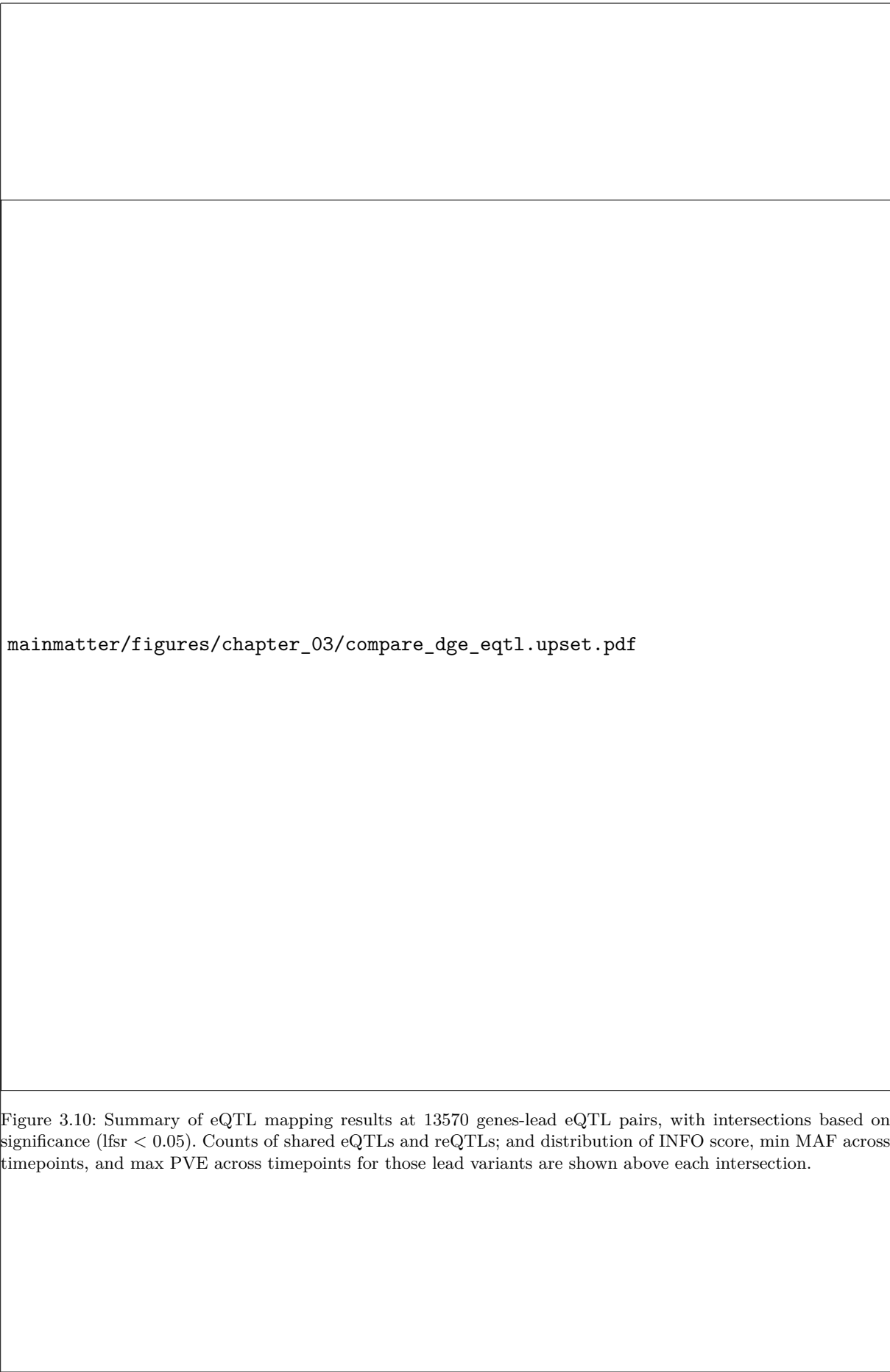




Figure 3.11: Z score for difference in effect vs. day 0, of lead eQTLs for all eGenes significant at either day 1 or day 7; versus distance of the lead SNP to the TSS. Direction of effect is aligned so that the beta at day 0 is positive. Points with positive z score are magnified effects post-vaccination, points with negative z scores are dampening and opposite sign effects.



Figure 3.12: *ADCY3*, strongest reQTL at day 1.

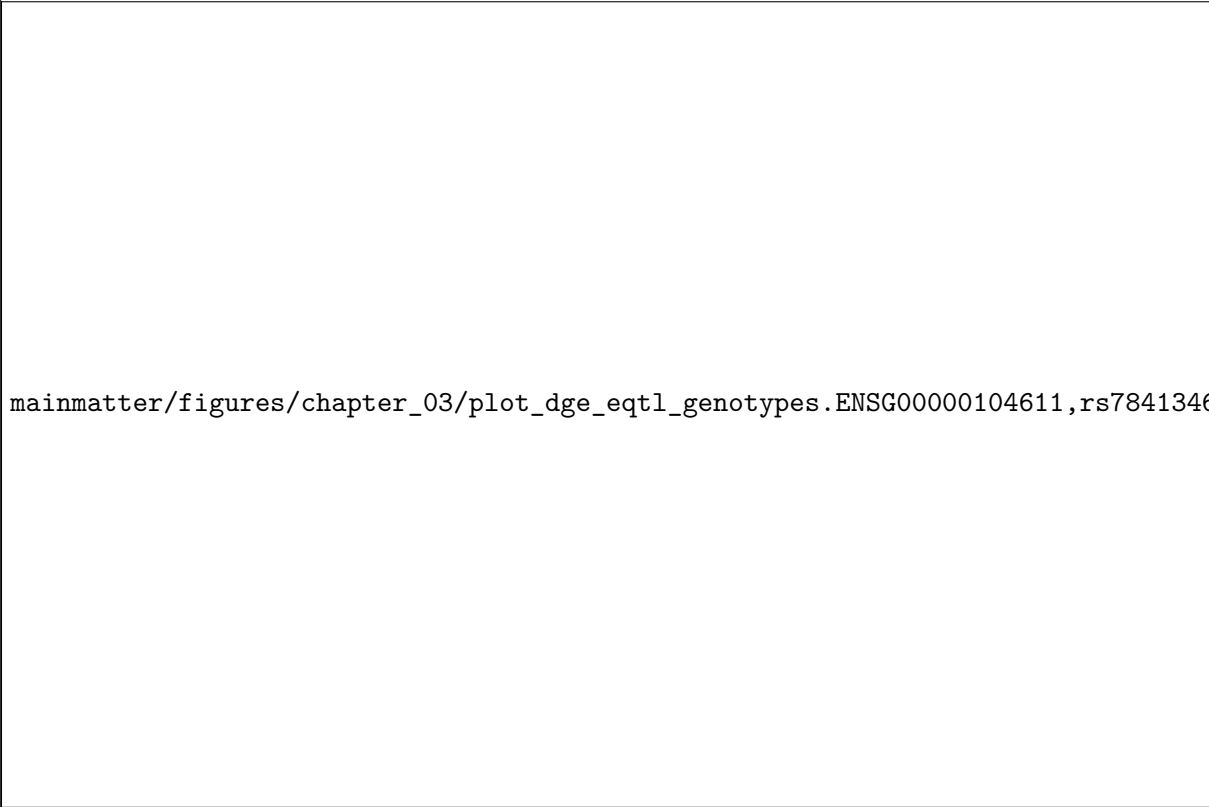
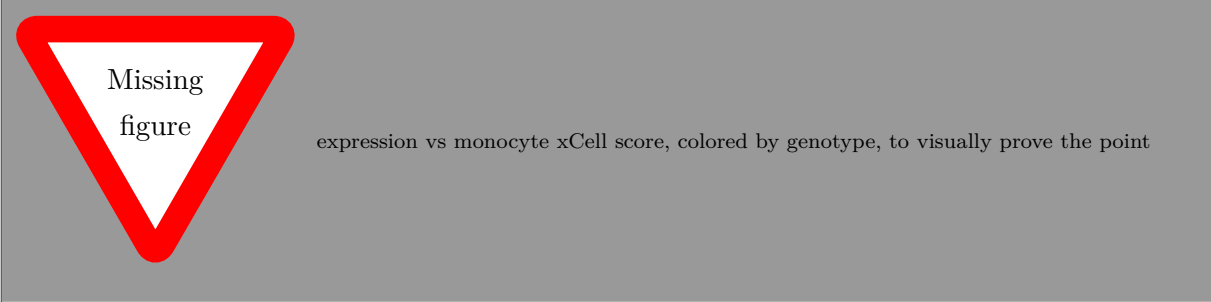


Figure 3.13: *SH2D4A*, strongest reQTL at day 7. Top: Array and RNA-seq expression before merging with ComBat for mega-analysis. Bottom: eQTL effects at each timepoint condition in the mega-analysis.

multicollinearity, standardised scores for monocytes, NK cells and plasma cells were tested for genotype interactions. Within-timepoint full eQTL models including the genotype main effect, the three cell type abundance main effects, and three cell type-genotype interaction terms, were fit using `lme4qt1`, then compared to a nested model excluding the three interaction terms.

Significant cell type interactions were detected at 16/1154 reQTLs (BH FDR < 0.05) in any timepoint, including *ADCY3* at day 1 ($\chi^2(3) = 26.3$, likelihood ratio test (LRT) BH FDR = 9.54×10^{-5}). Although the genotype effect size was 0.256 (SE = 0.0334) in the nested model, the estimate in the full model was $-0.007\,22$ (0.0666); with the three cell type-genotype interaction term estimates being: monocyte=0.213 (0.0490), NK cells= $-0.009\,20$ (0.0447), and plasma cells=0.0162 (0.0663). The small magnitude of the genotype main effect in the full model vs. the nested model indicates the eQTL effect is driven largely by the monocyte score (or a cell type that is highly correlated with monocyte score, see Figure 3.4). In the case where the monocyte score is zero (representing an average abundance across all samples, as scores are standardised), the effect of increasing genotype dosage on *ADCY3* expression is minimal.



3.3.4 TODO Genotype by platform interaction effects

- Perhaps using platform specific effects as a filter for reQTLs.

3.3.5 TODO Colocalisation of reQTLs with known *in vitro* condition-specific immune eQTLs

- Colocalisation is used to understand the molecular basis of GWAS associations (of a variety of human disease traits) (Giambartolome, 2014)
- Here the inverse: coloc is used to understand the biological relevance of observed reQTL by coloc with known immune QTL
- In a 500 Mb window around the lead *ADCY3* variant rs916485, HyPrColoc to colocalise with existing datasets and fine map.
- Day 1 HIRD colocs with BLUEPRINT and Fairfax monocytes (both stim and non stim), but not with Quach or Schmiedel monocytes (Figure 3.14)
 - Biases from ethnicity-derived differences in LD?
 - Also, priors need tuning?

siunitx permits uncertainties

gene set enrichment for cell type interacting genes to further validate xCell score usefulness

Need to consider Nikos' comment that there are too many (1069/13570 significant BH FDR) genotype-platform interactions to use mega-analysis. Consider filtering.

this analysis is incomplete, and is one of the things I would suggest to round off this chapter

- HyPrColoc fine maps the signal to rs13407913 (credible set size=1, PP = 1), an intronic variant 45064 bp downstream of the TSS.

– note not accurate due to MAF issues

3.4 Discussion

In the HIRD cohort, eQTL were detected for 50.8 % of genes in at least one timepoint, day 0, day 1, or day 7. Even in a joint mapping framework, defining reQTL by set significance thresholds, or change in the amount of expression variation explained, will miss classifying equal but opposite effect sizes. I account for the direction and magnitude of effect sizes, defining reQTL strength as the difference in effect size between timepoints. Most eQTL are shared between conditions and replicate well in GTEx whole blood; 16.8 % of lead eQTL for each eGene were reQTL that differed in effect size between timepoints.

Multiple independent eQTLs are present for a large fraction of eGenes[201]. As the lead variant for reQTL assessment for each eGene was chosen based on significance across all conditions, I can not detect reQTL that are masked by a stronger shared eQTL at that gene. This is not expected to be uncommon, as the effective sample size for shared eQTLs is usually large due to borrowing of information across conditions. Secondary eQTL signals tend to be weaker, more distal to the TSS, more likely to be enriched in enhancers rather than promoters, and importantly, more context-specific[47, 202]. The proportion of genes with reQTL I detect based on only the lead signal likely represents a lower bound.

Given the larger global changes in expression vs. baseline at day 1 compared to day 7 described in chapter 2, the larger number of reQTLs detected at day 7 was unexpected (819 vs. 1002). Opposite sign effects among reQTL post-vaccination were common. Prevalance of opposite sign effects between pairs of conditions has been previously described in multi-tissue studies. In [203], the proportion of opposite sign effects as a percentage of all eGenes was 7.4 % (48 tissues); in HIRD, I find 39/6887 (0.6 %) at day 1, and 211/6887 (3.1 %) at day 7. In [65], the proportion of opposite sign effects as percentage of all reQTLs was 4.4 % (5 tissues); in HIRD, I find 39/819 (4.8 %) at day 1, and 211/1002 (21.1 %) at day 7. The enrichment of opposite sign effects in HIRD is also apparent at day 7. The strongest reQTL at day 7 is one such opposite sign effect; *SH2D4A* has constitutive expression in T cells, B cells, macrophages, and DCs, encoding a adapter protein involved in intracellular signal transduction*. An approach for validating these opposite sign reQTL using the existing HIRD RNA-seq data is allele-specific expression (ASE) (e.g. [204]), where one would expect true opposite sign reQTL effects would also be recapitulated as opposite directions of expression imbalance.

The strongest reQTL detected at day 1 was *ADCY3*, a membrane-bound enzyme that catalyses the conversion of ATP to the second messenger cAMP[205]. Genome-wide association studies (GWAS) have identified *ADCY3* as a candidate gene for diseases such as obesity[205] and IBD[206]. *ADCY3* has been identified as a target for reQTLs in multiple studies involving stimulated blood immune cells: in PBMC 24h post-infection with rhinovirus[207], in whole blood in

*<https://doi.org/10.1111/j.1600-065X.2009.00829.x>

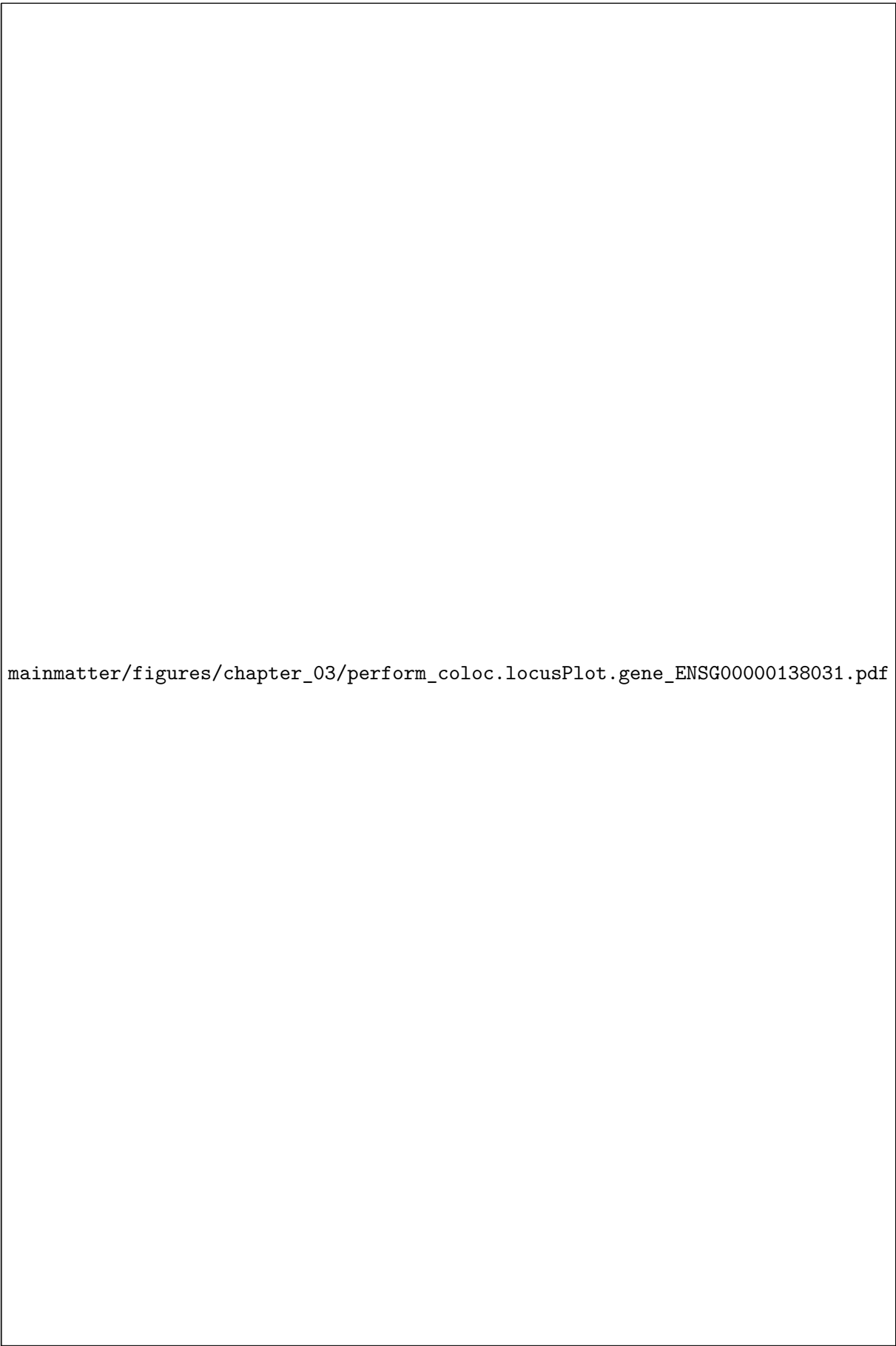
FYI the IBD/T cell coloc fine maps to chr2:24935139 T C (rs713586) with PP=1

add obesity GWAS

compare sharing with mashr and on-gen2017EstimatingCausalIssues

I'm not exactly sure why at the moment. Enrichment analyses so far have not turned up much. Up regulation of cell cycle TFs is a possibility.

replace mcgov-ern2015GeneticsInflammatoryBowel with more recent



mainmatter/figures/chapter_03/perform_coloc.locusPlot.gene_ENSG00000138031.pdf

Figure 3.14: Multi-trait colocalisation of HIRD reQTL signal at ADCY3 (500 Kb window), with QTL studies from IHEC, BLUEPRINT, eQTL Catalogue, and GWAS Catalogue. Plots are colored by colocalised cluster. Black indicates non-colocalised datasets.

vivo day 1 after vaccination with seasonal TIV[71], and in whole blood after stimulation with *M. leprae* antigen for 26-32 h[74]. Given the diversity of stimulations and tissue types, the effect is likely a consequence of general immune activation, rather than a Pandemrix-specific response.

Statistical colocalisation suggests that the day 1 reQTL signal identified here is likely to be a monocyte-specific effect—and independent to the IBD signal, which colocalises with T cell and macrophage datasets. The proportion of monocytes in the PBMC increase at day 1, supported by both FACS[106] measurements, and an increase in monocyte xCell score. Expression of *ADCY3* is not monocyte-specific, as despite the increase in monocyte proportion, no upregulation is observed at day 1. Colocalisation is also not restricted to stimulated monocytes, hence the signal could be hypothesised to result simply from the increased proportion of the bulk sample taken up by monocytes, rather than a upregulation-driven increase in detection power, or a vaccine-induced activation of the locus at day 1.

Changes in relative abundances for many cell types occur in the bulk PBMC samples after vaccination. I accounted for the effect of abundance on mean expression including xCell scores and PEER factors as fixed effects in the model, and also considered the effect of abundance on the genotype effect using interaction terms between xCell scores and genotype. Due to the modest sample size, and computational requirements for `lme4qt1`, I focused only whether reQTLs that have a detectable main effect may be driven by cell type interactions, testing only for interactions at significant lead reQTL. Compared to FACS measurements in a cohort subset, the xCell scores used above were only weakly correlated. Some discrepancy is expected, as the cell types as defined in the xCell signatures do not directly correspond to the combinations of surface markers used for FACS. The FACS gating strategy also meant that for some cell populations, the only available FACS measure was a proportion of the previously gated population, whereas xCell attempts to estimate scores that represent proportions of the whole mixture. The accuracy of the built-in signatures is lower when applied to the expression matrix for a stimulated state, likely because the enrichment-based method can not distinguish differential expression of signature genes due to stimulation from actual changes in cell abundance. Nevertheless, as assuming a single genotype where cell-type specific slopes are likely is inappropriate, so xCell scores were used as a best approximation. At 16/1154 reQTLs, the genotype effect was detected to interact with abundance of one or more of the tested cell types (or a correlated cell type). At the day 1 *ADCY3* reQTL, the genetic effect can be mainly attributed to the monocyte score-genotype interaction term, further supporting the hypothesis that it is monocyte-specific.

A pressing question remains: what molecular mechanisms underlie the *ADCY3* reQTL, and indeed the remainder of the reQTLs? Power differences due to condition-specific expression are unlikely to explain a large proportion of reQTLs. As in [73, 75], the overlap between differentially expressed genes and genes with reQTL was poor, and reQTL were not more likely to be differentially expressed compared to genes without reQTL. One mechanism by which cis-eQTL affect expression is through their impact on transcription factor (TF) binding affinity to motifs in promoters and enhancers[208]. Immune cells, including monocytes, are regulated by cell type-specific TFs[209]. cell type-specific expression of different TFs have been proposed as a model for explaining magnifying, dampening and opposite reQTL effects; for example, opposite effects can result from TFs regulating the same gene, that are activating in one cell type and suppressive

add lfsr.dge

need to consider: if this kind of thing is what bulk in vivo reQTL can find, they what is the additional value over FACS?

dge is coupled to reqtl, if you do an enrichment of dge+reqtl overlap genes, likely their enrichment is driven by DGE signal

harmonise terminology for 'opposite'

check "rs2223286 is associated with profound directional effects in the expression of *SELL* dependent upon genotype, with the minor C allele

in another[65]. There is evidence that TF activity is important for *in vivo* immune reQTL: [207] found rhinovirus reQTLs in PBMCs were enriched in ENCODE ChIP-seq peaks for the TFs *STAT1* and *STAT2*, and [73] found interferon and anti-IL6 drug reQTLs likely disrupt *ISRE* and *IRF4* binding motifs. Rather than condition-specific expression of the eGene, what may be condition-specific is the expression of TFs whose activity is affected by the reQTL*.

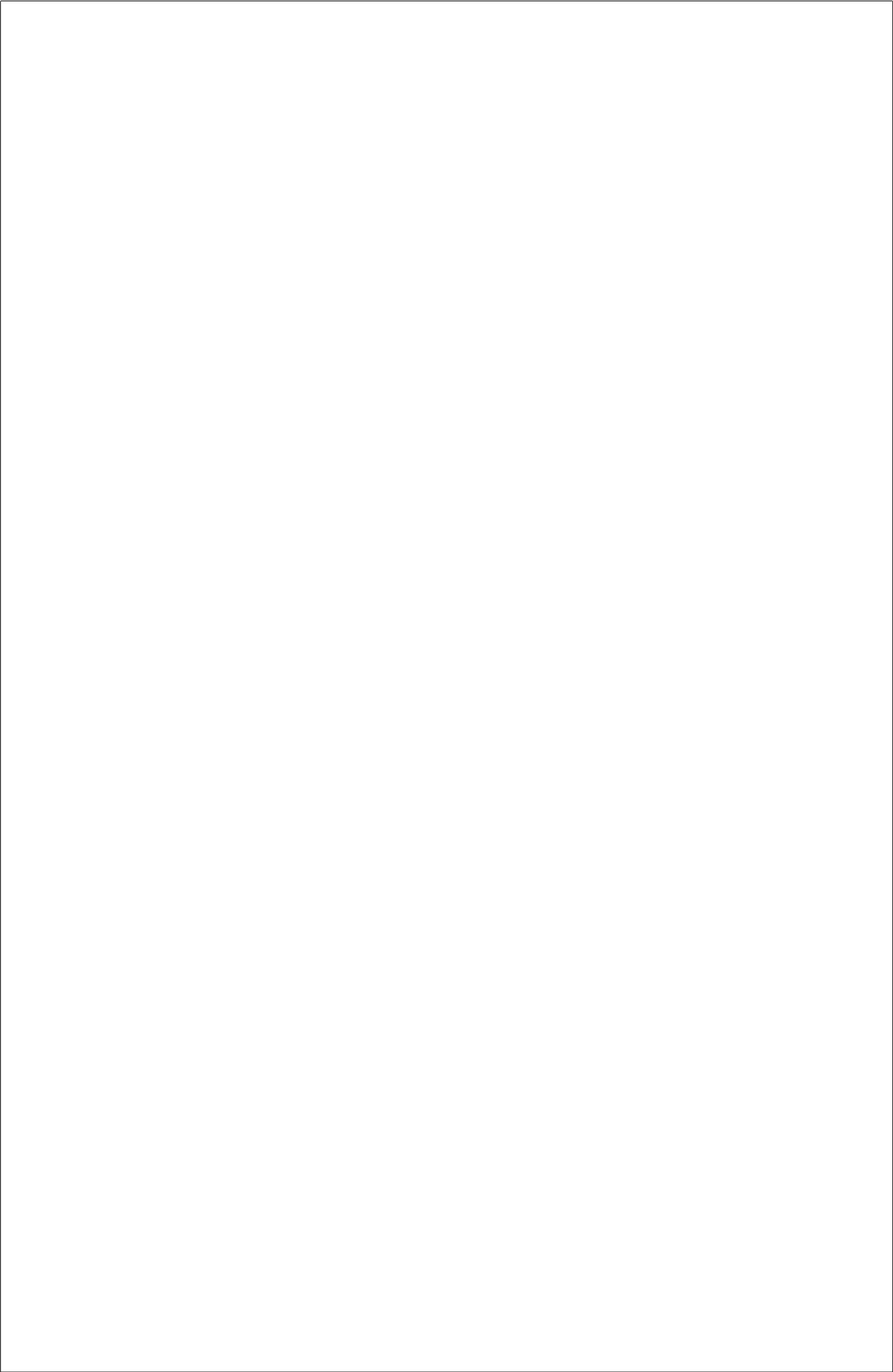
Finally, I address the prospect that common genetic variation may explain some variation in antibody response to Pandemrix. I have indirectly demonstrated genotype-dependent effects on expression response by identifying reQTLs with differing effect size between timepoints, but have not yet to determined resulting genotype-dependent differences in antibody phenotypes. Some of the identified reQTLs will undoubtedly affect genes whose expression or post-vaccination expression change correlates with antibody response, but correlation is not transitive[210], and a formal tests such as the causal inference test (CIT)[211] are required to distinguish mediation of genotype-antibody associations through gene expression from competing models. [71] realised this, but concluded that they had insufficient power with a greater sample size and comparable study design to HIRD. The HIRD cohort is also too small for a direct GWAS of Pandemrix antibody response. A suitable approach for prioritising reQTL that contribute to the antibody response to Pandemrix will be to leverage external genetic associations to similar phenotypes, for example, colocalisation with existing GWAS summary statistics for antibody response to a similar type of adjuvanted, inactivated vaccine.

note coloc doesn't distinguish pleiotropy from mediation?

add 1 concluding line

Overall I feel like the chapter is too descriptive, and falls short of making biological insights into Pandemrix response. Any additional analyses would hope to address that.

*A cursory scan of TF motifs disrupted by the location of the fine-mapped *ADCY3* reQTL intronic variant rs13407913 on <https://ccg.epfl.ch/snp2tfbs/snpviewer.php>, does indeed show several motifs (for NR2C2, HNF4A, HNF4G, NR2F1) where the PWM score is higher for the ALT allele, consistent with the direction of effect for the day 1 reQTL.



Chapter 4

Transcriptomic associations with anti-TNF drug response in Crohn's disease patients

4.1 Introduction

4.1.1 Crohn's disease and inflammatory bowel disease

Crohn's disease (CD) is a chronic, inflammatory disease of the gastrointestinal tract. It is characterised by patchy inflammation, where lesions are interspersed with regions of normal mucosa. The lesions can be distributed anywhere in the gastrointestinal tract, and tend to be transmural, affecting all layers of the gut wall.

CD is one of the two main forms of inflammatory bowel disease (IBD). The second form, ulcerative colitis (UC), is characterised by continuous inflammation, with lesions that are superficial rather than transmural, and restricted to the colon [212]. Whilst the two are distinct forms of IBD, similarities in clinical presentation, available therapies, and genetic architecture mean CD and UC have often been studied together. Both are immune-mediated inflammatory diseases (IMIDs), a group of related diseases involving immune dysregulation of common inflammatory pathways. Other IMIDs include type 1 diabetes (T1D), systemic lupus erythematosus (SLE), rheumatoid arthritis (RA), multiple sclerosis (MS), and psoriasis [213, 214].

Pathogenesis of CD is not completely understood, but involves interaction of the immune system, environmental factors (e.g. smoking, stress, diet [212, 215]), and gut microbial factors in a genetically-susceptible individual [216]. Since the seminal discovery by linkage analysis in 2001 that genetic variation in *NOD2* was linked to CD risk [217], much progress has been made in establishing the disease's genetic architecture. The most recent genome-wide association study (GWAS) studies catalogue over 240 risk loci for IBD [218]. Most associations are shared between CD and UC, with some heterogeneity of effects, such as *NOD2* being more strongly linked to CD risk [219, 220].

CD has historically been considered a disease of the Western world. The highest prevalence and incidence of new cases of CD are in North America and western Europe [212], although disease burden is now rising in newly industrialised countries in Asia, Africa and South America

difficult to separate out CD from IBD in literature, section probably needs to decide to focus on CD only or IBD.

[221, 222]. The modal age of onset is typically between late adolescence and early adulthood. The disease is progressive: within 20 years of diagnosis, 50% of patients with CD will develop gastrointestinal tract complications and approximately 15% require surgical intervention [212]. Given the rising prevalence and large impact on quality of life, there is active research into developing treatment regimes with the goal of inducing complete mucosal healing [212, 223].

4.1.2 Anti-TNF therapies for Crohn’s disease

Tumour necrosis factor (TNF), also known by the archaic name TNF- α , is a proinflammatory cytokine produced mainly by immune cells such as monocytes, macrophages, natural killer (NK) cells, T cells, and B cells. It is synthesised in transmembrane form, then enzymatically cleaved into its soluble form. TNF binds to receptors TNFR1 and TNFR2; most cells in the body express one receptor or the other. Binding triggers a signalling cascade that in different contexts, regulates inflammation, apoptosis, cell proliferation, and cell survival [224–226]. In the context of IBD pathogenesis, current models suggest high TNF levels promote apoptosis of monocytes, macrophages, and gut epithelial cells via TNFR1, while inhibiting apoptosis of mucosal CD4⁺ T cells via TNFR2 [223, 226, 227], contributing to maintained gut inflammation.

The development of anti-TNF biologic therapy has revolutionised CD patient care in the last two decades. Infliximab and adalimumab are the two major drugs approved for treatment of CD, UC, and a number of other IMIDs. Both are IgG1 monoclonal antibodies that bind soluble and transmembrane TNF to inhibit interactions with its receptors [227, 228]. Two main mechanisms for their action have been proposed: induction of CD4⁺ cell apoptosis in the gut mucosa by inhibiting the TNF-TNFR2 interaction; and binding of the antibody tail (Fc) region of the antibody to Fc receptors on monocytes, inducing their differentiation into wound-healing M2 macrophages [223].

Adalimumab is a human antibody, typically administered subcutaneously via auto-injector pen, with two initial doses aimed to induce remission, then a dose every two weeks to maintain remission. Infliximab is a chimeric mouse-human antibody administered via intravenous infusion, with a three-dose induction and a dose every eight weeks for maintenance [227]. Anti-TNF biologics are among the drugs with the highest global spendings; adalimumab and infliximab ranked first and fourth respectively in 2017, with an estimated combined spending of 29 billion USD [229].

4.1.3 Anti-TNF treatment failure

Despite the expense, anti-TNF therapy is not always effective at treating CD. Various types of treatment failure can occur: primary non-response (PNR) within the induction period (the first 12-14 weeks for adalimumab and infliximab), developing secondary loss of response (LOR) during maintenance after an initial response, failure to achieve remission after the treatment course, and adverse events that lead to treatment stoppage [230]. For IBD patients, the incidence of PNR is 10-40%, and the incidence of secondary LOR is 24-46% in the first year of treatment [231–233]. Another factor affecting treatment outcome is immunogenicity, the generation of antibodies against the drug, thought to increase the probability of treatment failure by increasing drug clearance rate [228, 233]. As a chimeric antibody, infliximab is more immunogenic than

the subsection on anti-TNFs in the intro chapter may be merged here

I assume dosing follows the induction and maintenance schedule from [227]. Can’t find anything about it in the PANTS protocol, although Sim confirmed the 2w/8w frequencies.

adalimumab [233, 234]. Although remission with complete mucosal healing is the gold standard for successful treatment [212], PNR and LOR can be defined much earlier in the treatment course, and help guide changes in treatment such as dose intensification or switching to a drug class with a different mechanism of action [228, 231].

Anti-TNF biologics are near the top of the therapeutic pyramid for CD, among the treatment options with the highest toxicity and costs [235]. The traditional approach to disease management is “step-up”, beginning at the bottom of the pyramid with steroids [227]. This may undertreat patients that require more aggressive therapy, allowing the disease time to progress. An inverted approach begins at biologic therapies, then steps-down the pyramid if possible. This risks exposing patients to aggressive therapies they may not have needed [232]. The best approach would be to predict whether a particular treatment will be required and effective for a patient, especially given the costliness and patient risks associated with therapies near the top of the pyramid. Baseline prediction would be especially valuable for stratifying patients to specific therapies.

4.1.4 Predicting response to anti-TNFs

Clinical variables reported to have association to anti-TNF response include age, disease duration, body mass index (BMI), smoking, C-reactive protein (CRP), faecal calprotectin, serum drug concentrations, and anti-drug antibody concentrations. These have mostly been found in small retrospective cohorts, and have rarely been independently validated [226, 232, 236–239]. In the Personalised Anti-TNF Therapy in Crohn’s Disease (PANTS) study, the largest study of infliximab and adalimumab response in CD patients to date (n=1610), baseline obesity, smoking, and greater disease activity were associated with low serum drug concentration after induction. Low drug concentration was in turn associated with PNR and non-remission, suggesting immunogenicity may be mediating treatment failure [233].

Studies have looked for transcriptomic predictors for anti-TNF response [226, 239]. In gut biopsies, expression of sets of “signature” genes were found to be predictive of mucosal healing after infliximab treatment in cohorts of UC (*TNFRSF11B*, *STC1*, *PTGS2*, *IL13RA2*, *IL11*; n=46 [240]) and CD patients (*TNFAIP6*, *S100A8*, *IL11*, *G0S2*, *S100A9*; n=19 [241]). Expression of *OSM* was associated with anti-TNF response defined by improved Mayo score, a multiparameter clinical score of UC activity (n=227) [242]. Most recently, single-cell RNA-sequencing (RNA-seq) identified a module of IgG plasma cells, inflammatory mononuclear phagocytes, activated T cells, and stromal cells associated with clinical remission after anti-TNF therapy in two separate CD cohorts (total n=340) [243].

As obtaining blood samples is non-invasive, there has been great interest in finding transcriptomic predictors of response in blood. While blood is not the disease-relevant tissue for CD, many genes in gut biopsy signatures have high expression in infiltrating immune cells, and blood gene expression may capture the precursors of those cells [244]. Blood *TREM1* expression has been identified as a marker of anti-TNF response in two studies with inconsistent directions of effect. Gaujoux *et al.* [244] defined response based on “clinical and/or endoscopic improvement”. *TREM1* expression was lower in infliximab responders in gut biopsies (total n=72), but higher in responders in a separate cohort measuring whole blood expression (n=22). Verstockt *et al.*

[245] defined response based on endoscopic remission, reporting *TREM1* to be a marker of response with lower expression in responders to infliximab and adalimumab in both gut biopsies and blood. Proposed reasons for the discrepancy include small sample sizes, patient ethnicity, and different definitions of response <https://pubmed.ncbi.nlm.nih.gov/30007919/> [226].

Attempts have been made to find genetic markers for response, as they are baseline by definition, and may also help form mechanistic hypotheses of treatment response. Anti-TNF response does not necessarily share the same genetic architecture as disease risk. Variants in TNF-regulated genes that are also associated with IBD risk (*NOD2*, *TNFR1*, *TNFR2*) are not associated with response to infliximab [226, 239]. A number of candidate gene studies found single nucleotide polymorphism (SNP) associations with response in genes such as apoptosis-related Fas ligand and caspase-9 that have yet to be validated [232, 246]. Recently, larger cohorts have enabled GWAS of anti-TNF response in IBD. In PANTS, although no associations to PNR were genome-wide significant*, HLA-DQA1*05 carriage was associated with higher anti-drug antibody levels [247], which may mediate related phenotypes such as LOR.

Overall, small sample sizes and variation in analysis methods, anti-TNF drug, response definition, tissues sampled, and disease among studies make a consensus hard to establish. No algorithms using clinical, transcriptomic, or genetic markers for predicting IBD patient response to anti-TNF therapy have yet been translated to clinical practice, although several are currently undergoing validation [239].

4.1.5 Chapter summary

This chapter focuses on identifying transcriptomic associations with anti-TNF primary response in a subset of the PANTS cohort with longitudinal RNA-seq data from the first year of follow-up. I model differential gene expression (DGE) between primary responders and non-responders at the gene and module-level at baseline (week 0), post-induction (week 14), and during maintenance (week 30 and week 54). As this is one of the largest datasets currently available for assessing transcriptomic associations with anti-TNF response in IBD, I attempt to validate and resolve conflicts in the literature for previously identified transcriptomic markers such as *TREM1*. Finally, I integrate existing genotype data to map response expression quantitative trait locus (reQTL) between timepoints, with the aim of identifying common genetic variants controlling expression response to anti-TNF drugs.

4.2 Methods

4.2.1 The PANTS cohort

PANTS is a prospective, observational, UK-wide, cohort study of response to anti-TNF therapy in CD patients, described in detail by Kennedy *et al.* [233]. The study was registered with ClinicalTrials.gov identifier NCT03088449, and the protocol is available at <https://www.ibdrresearch.co.uk/pants/>. Total enrollment was 1610 patients, who were at least 6 years old, had

*

add ref to Alex’s ppt

trying not to ‘spoil’ any results in the intro, I guess

active luminal CD, and were naive to anti-TNF therapy. Patients were invited to attend up to ten major study visits over a maximum follow-up period of three years, or until drug withdrawal.

The anti-TNF drugs evaluated were adalimumab and infliximab. The study also evaluated infliximab biosimilars; data from patients who received a biosimilar is not included in this chapter. All major visits were scheduled immediately prior to a drug dose. Adalimumab and infliximab have 2-week and 8-week dosing intervals respectively, so the timing of major visits was chosen such that the same visit structure could be used for patients on either drug. Additional visits were scheduled in case of secondary LOR, or premature exit due to drug withdrawal, usually replacing the next scheduled major visit.

The overall rate of primary non-response by week 14 was 21.9% for infliximab, and 26.8% for adalimumab. The rate of secondary LOR by week 54 among primary responders was 36.9% for infliximab and 34.1% for adalimumab. Remission rate by week 54 was 39.1% for infliximab and 33.1% for adalimumab.

4.2.2 Definition of timepoints

The RNA-seq data for this chapter comes from a subset of the cohort sampled around four timepoints: week 0, week 14, week 30, and week 54. These are the target timings for four major visits in the first year of follow-up. Whole blood samples were taken prior to the scheduled drug dose and preserved for RNA-seq in Tempus Blood RNA Tubes. The study day that sampling occurred relative to the first drug dose was recorded.

To measure the transcriptome at trough drug levels, I mapped samples from major and additional visits to four timepoints centered around the four major visits. As it could not be guaranteed that visits occurred on the exact day specified in the protocol, I considered the visit windows defined by Kennedy *et al.* [233]: week 0 (week -4-0), week 14 (week 10-20), week 30 (week 22-38), and week 54 (week 42-66) Samples were mapped according to the following criteria:

- Major visit samples were mapped to the corresponding timepoint, regardless of whether they fell within the corresponding window i.e. an available week 0 sample is always mapped to the week 0 timepoint.
- Samples taken at additional LOR or exit visits falling within one of the windows were mapped to that timepoint, unless the patient also had a major visit sample inside that window.

Only a small minority of major visit samples fell outside their corresponding windows. Inclusion of samples from additional visits was important as they often replaced major visits for patients with primary non-response or LOR. Samples included under both criteria should be representative of trough drug levels, as major visits and LOR visits were always scheduled prior to a drug dose, and exit visits were scheduled for when the next drug dose would have been.

4.2.3 Definition of primary response and non-response

The definition of primary response and non-response was based on the clinical decision tree from Kennedy *et al.* [233]. Primary response status was assessed at week 12, prior to the scheduled

compute maximum deviation?

Still discussing with Sim on the exact def of LOR and exit visits to decide whether this is sensible.

week 14 visit. The criteria for primary non-response was *either* of the following:

- exit for treatment failure before week 14 (e.g. as decided by physician global assessment), *or* corticosteroid use at week 14 (a continuing or new prescription);
- compared to week 0, a decrease in CRP by less than 50% or to $>3 \text{ mg l}^{-1}$, *and* a decrease in Harvey Bradshaw index (HBI) by less than 3 points or to >4 .

As PANTS was an observational study that continued until drug withdrawal, a patient's clinician may have decided to continue anti-TNF therapy even if a patient had primary non-response, so it was possible for non-responders to be sampled past week 14.

The criteria for primary response was *all* of the following:

- not classified as a primary non-responder;
- $\text{CRP} \leq 3 \text{ mg l}^{-1}$ by week 14;
- $\text{HBI} < 4$ by week 14.

Grey zone patients that only met a subset of criteria for either primary response or non-response were excluded.

There were additional inclusion criteria applicable to just the RNA-seq subcohort. Patients were required to be at least 16 years old, and have an available baseline serum sample. Primary non-responders were filtered to exclude patients in remission at week 54.

4.2.4 Library preparation and RNA-seq

Total RNA was extracted following the Qiagen QIAasympyphony instrument protocol (RNA Isolation PAX RNA CR22332 ID 2915). RNA was quantified with the ThermoFisher QuBit BR RNA (Q10211), and RNA integrity assessed with the Agilent RNA ScreenTape assay (5067-5579, 5067-5577, 5067-5576) on the Agilent 4200 TapeStation.

Library preparation was done using the Kapa mRNA HyperPrep Kit, including enrichment for mRNA using magnetic oligo-dT beads, depletion of rRNA and globin mRNA using the QIAseq FastSelect RNA Removal Kit, and adapter ligation with IDT xGEN Dual Index UMI adapters. Libraries were sequenced on the Illumina HiSeq 4000 with 75bp paired-end reads.

4.2.5 RNAseq quantification and preprocessing

A total of 1141 samples were initially sequenced. Sequencing data was demultiplexed with Picard. Reads were mapped to GRCh38 using STAR (2.6.1d) and deduplicated to unique reads using UMI-tools. Gene expression was quantified against the Ensembl 96 gene annotation with featureCounts (1.6.4).

Total number of read pairs, sequence quality, overrepresented sequences, adapter content and sequence duplication rates were checked using FastQC. Samples were filtered to remove outliers ($>2\text{SD}$ from the mean) according to percentage of aligned reads in coding regions reported by Picard, percentage of unique reads, and number of unique reads. Samples that could not be mapped to a timepoint according to subsection 4.2.2 were removed. Samples that came from

patients with sex mismatch grey zone primary response, or missing data for variables considered in the variable selection process (subsubsection 4.2.6.1), were removed. A total of 814 samples remained after filtering.

The Ensembl 96 gene annotation contains 58 900 genes, many of which are not expressed in whole blood. Normalised library sizes were computed using `edgeR::calcNormFactors(method='TMM')`. Between-sample normalisation for library size was done using `edgeR::cpm`, converting counts to counts per million (CPM). Genes with low expression were filtered, requiring >1.25 CPM in $>10\%$ of samples (1.25 CPM was approximately 10 counts at the median library size of 8 million assigned read pairs), and non-zero expression in $>90\%$ of samples. Globin genes and short ncRNAs were removed. A total of 15511 genes remained after filtering. Finally, CPMs were converted to the \log_2 scale, and precision weights to account for the expression mean-variance relationship were computed for each gene and sample using `limma::voomWithDreamWeights`.

4.2.6 Differential gene expression

4.2.6.1 Variable selection by variance components analysis

For each gene, the DGE model was a regression expressing the response variable, gene expression, as a linear function of predictor variables of interest (primary response status, drug, timepoint), and other selected predictor variables. In estimating the association $X \rightarrow Y$, of predictor X to response Y by regression, adjustment for a third variable Z can increase, decrease, or even reverse the effect estimate (the regression coefficient). I aimed to select third variables for inclusion into the DGE model that were covariates, defined here as a Z that is associated with Y , but not with X . Such variables are also known as neutral controls [248], precision variables, or prognostic variables. At the cost of 1 degree of freedom (df), Z explains variation in Y that would otherwise be considered residual, so conditioning on Z increases the efficiency of estimating $X \rightarrow Y$, but does not change the effect estimate.

Many variables were available for selection; Figure 4.1 shows their correlation matrix. These included three variables associated with primary response in Kennedy *et al.* [233]: baseline immunomodulator use, smoking and BMI. Also available were proportions of six common cell types in whole blood ($CD4^+$ T cells, $CD8^+$ T cells, B cells, NK cells, monocytes, granulocytes), estimated using the Houseman method (`minfi::estimateCellCounts` <https://academic.oup.com/bioinformatics/article/30/10/1363/267584>) from whole blood Illumina MethylationEPIC data collected from the same patients and timepoints. The Houseman method uses differentially methylated regions between immune cell types as cell type markers <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-13-86>.

A variance components analysis was done to quantify the proportion of expression variance explained by each variable for each gene using `variancePartition` [249]. Variables that do not explain much variation in the response are unlikely to improve efficiency if conditioned on. The model was a mixed effects regression model with variables in Figure 4.1 included as predictors. Additional categorical variables were included for patient and RNA-seq library preparation plate. An additional continuous variable consisting of random numbers drawn from the standard normal distribution was also included as a null. Granulocyte proportion estimates were dropped

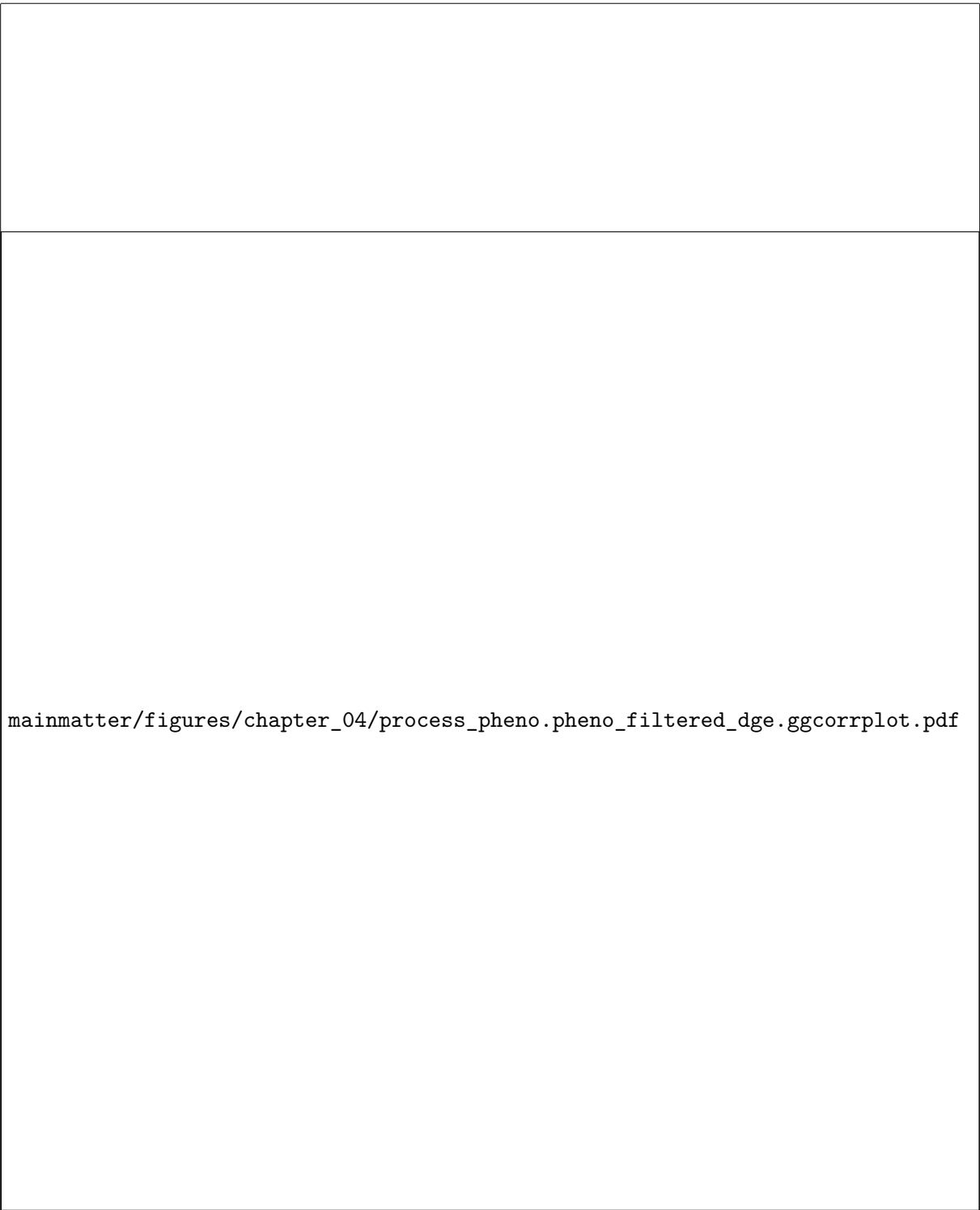


Figure 4.1: Correlation matrix of variables measured in PANTS that were considered as potential independent variables.

to relieve perfect multicollinearity. Categorical variables were coded as random intecepts, and continuous variables as fixed effects. Surprisingly, simulations from Hoffman *et al.* [249] showed variance proportion estimates were unbiased even when coding categorical variables with as few as two categories as random, as long as model parameters were estimated using maximum likelihood (ML) rather than restricted maximum likelihood (REML), which is commonly used for variance components analysis. It was also shown this approach avoids overestimates of variance proportions that occur if categorical variables with many levels are treated as fixed.

As downstream DGE methods require the same set of predictors for all genes, the aim was to select variables that explain a lot of variance for many genes. Variables were ordered by the median of gene-wise variance proportion estimates (Figure 4.2). Variables that explained the most variance on average were patient, cell proportions and RNA-seq plate. Some variables that did not explain more variance on average than the null nevertheless had high maximum values, indicating their importance for a relatively small number of genes. These included sex, library preparation protocol version, and smoking status, but also primary response status, a variable of interest, so it was difficult to justify excluding all variables with lower median variance explained than the null. Thus all non-null variables in Figure 4.2 were selected as predictors in downstream models apart from “Ever_Immunomodulator” (whether the patient had ever had immunomodulator treatment), as that had both low median variance explained and was correlated with baseline immunomodulator use. This is a crude approach, but the sample size is large compared to number of df lost by including predictors that may not be relevant for some genes.

How are interpretations of effect sizes of interest be affected by including this suite of other variables, all of which can be considered as third variables? If a third variable Z is not a precision variable, but is also associated with X , conditioning changes the effect estimate. The regression model is mathematically agnostic to causal relationships between variables, but distinct types of third variable can be distinguished conceptually by assuming the direction of causal relationships <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2819361/>. Conditioning on a confounder ($X \leftarrow Z \rightarrow Y$) reduces bias of the effect estimate, conditioning on a collider ($X \rightarrow Z \leftarrow Y$) induces bias, and conditioning on a mediator in the causal path ($X \rightarrow Z \rightarrow Y$) changes the effect estimated by removing the indirect effect mediated by Z , usually biasing the effect estimate towards zero*. The major influence I consider in this chapter is adjusting for cell proportions. From the variance partition analysis (Figure 4.2), cell proportions were among the biological factors that explained the most variance on average. I have not systematically assessed what types of third variable other selected predictors may act as.

The assumed causal relationship is response \rightarrow cell proportion \rightarrow expression, where cell proportion mediates the association between response and expression[†]. I fit separate DGE models including and excluding cell proportions as predictors, but otherwise identical. These have complementary interpretations. In models without cell proportions, the effect size of response status could represent up or downregulation on a per cell basis, but could also come from differences

*It is not easy to determine the direction of bias (positive or negative) for any of these cases in general <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7064555/>.

[†]The assumption in the DGE model that response status is an intrinsic property of a patient that is causally upstream will be discussed in chapter 5.

The observational study design does not allow straightforward estimation of causal effects, so this is only assumed



Figure 4.2: Variance partition analysis showing distribution of gene-wise percentage of variance in expression explained by each variable. “PANTS.ID” is patient ID.

in cell proportions in the bulk samples associated with response. The estimates from models adjusted for cell proportions are more likely to reflect up or downregulation on a per cell basis.

Could cell proportions act as colliders?

4.2.6.2 Contrasts for pairwise group comparisons

gene-wise DGE models were fit in `dream` [250]. Like the variance partition model models were linear mixed models:

where the response variable is expression at a particular visit; 0 indicates there is no intercept term; and `concat(Visit, Response, Drug)` is experimental group defined by combinations of the predictors of interest: visit (week 0, 14, 30, 54), response (responder, non-responder), and drug (infliximab, adalimumab). This is equivalent to a model with an intercept term and a three-way interaction between visit, response, and drug, including all lower order terms ($1 + \text{Visit} * \text{Response} * \text{Drug}$), but is more convenient for testing pairwise expression differences between groups, as the coefficient for each term is the estimate of mean expression for that group. As the interest was in estimating a single coefficient for each predictor’s effect size on expression (rather than estimating variance components), most predictors are modelled as fixed effects. RNA_Plate and Patient are nuisance variables with a large number of levels, so they are modelled as random intercepts. To avoid small-sample bias in estimates of fixed effect standard errors, estimation was done with REML [251].

Alter and move this to ch2

Two sets of gene-wise models were fit with and without the cell proportion terms in square brackets. Two more sets of gene-wise models were fit with drugs pooled (i.e. `concat(Visit, Response, Drug)` replaced with `concat(Visit, Response) + Drug`), but otherwise identical to the first two sets.

Specific hypotheses were tested using sum-to-zero contrasts, which are linear combinations of

the model coefficients with weights summing to zero. For example, to test for DGE between responders and non-responders to infliximab at baseline in the non-pooled model, I used a contrast where the weight for the (week 0, responder, infliximab) group coefficient was 1, the weight for the (week 0, non-responder, infliximab) group coefficient was -1, and all other coefficient weights were 0.

To get p -values, the contrast divided by its standard error was compared to the t distribution using the Satterthwaite approximation for df. False discovery rate (FDR) control was done with the Benjamini-Hochberg (BH) method, computed separately for each contrast*.

4.2.6.3 Spline model of expression over time

The aim was to use expression data from all four timepoints to find genes associated with response, while avoiding a large number of pairwise comparisons. I fit a natural cubic spline (`splines::ns`) to the study day to allow for non-linear trajectories of expression over time. The natural cubic spline is a continuous function defined piecewise in each successive interval between a set of k knots in the range of the input variable. The $k - 1$ pieces between knots are polynomials of degree 3. As a natural spline, the function is constrained to be linear outside of the boundary (first and last) knots to avoid unpredictable behaviour at the boundaries. The model form used was:

Two inner knots set at week 14 and week 30, as expression is expected to change after each drug dose. To include all data to within the boundaries, the two boundary knots were set at the minimum and maximum values of study day rather than week 0 and week 54. The function `ns(Study_Day, knots=7*c(14, 30))` returns a basis matrix for the spline with 3 df (columns). The basis used by `ns()` is a B-spline basis which transforms the input [TODO...] [252]. Separate

sets of gene-wise models were again fit with and without cell proportions. To test for response-associated differences in the spline parameters, the predictors of interest are the three second-order interaction terms between response and the basis matrix columns. The three terms are tested jointly by F-test, and FDR correction was done with the BH method separately for each test. A significant result indicates a significant difference in the trajectory of expression over study day between responders and non-responders.

4.2.6.4 Clustering expression over all timepoints

To aid in interpretation of significant spline hits from the cell proportion adjusted model, I clustered genes by their expression trajectories.

Expression data was converted to the CPM scale using trimmed mean of M-values (TMM) normalisation factors, then regressed against cell proportions. Residuals were centered and scaled per gene, A distance matrix was computed using 1 - Pearson correlation as the distance metric. Hierarchical clustering was done with complete agglomeration for inter-cluster distance (`fastcluster::hclust(method='complete')`). The optimal number of clusters was determined using the gap statistic (`factoextra::fviz_nbclust(method='gap_stat', nboot=500)`), which determines

*It could also have been computed globally over all contrasts if it were necessary to have the same t -statistic threshold for statistical significance in all contrasts.

trying to understand how exactly the B-spline basis for the natural spline works will take a few more youtube lectures...

when the change in within-cluster dispersions are no longer significantly improved by increasing the number of clusters <https://rss.onlinelibrary.wiley.com/doi/epdf/10.1111/1467-9868.00293>. The hierarchical clustering tree was then cut into that number of clusters.

4.2.6.5 Rank-based gene set enrichment on gene modules

Rank-based gene set enrichment analyses were done using `tmod::tmodCERNOtest` [157], which assesses the enrichment of small ranks within predefined sets of genes compared to all measured genes after ranking the genes by some metric. For each contrast, as the t-statistics are not comparable between genes due to the use of approximate df, I ranked genes by the signed Z score reported by `dream`, which is a monotonic transformation of the *p*-value. Similarly, moderated F-statistics from the spline are not comparable between genes, so I used the signed F-statistic from the transformation of the *p*-value.

The CERNO statistic for a gene set is:

$$-2 \sum_{i=1}^n \ln \frac{r_i}{N} \sim \chi^2(2N) \quad (4.1)$$

where *n* is the number of genes in the set, *r* is the rank of the *i*th gene, and *N* is the number of measured genes. CERNO is relatively robust to the ranking metric [253]. Since CERNO only considered enrichment of small ranks but genes can be down or upregulated, separate tests were done sorting in ascending and descending order, and the more significant result was used to determine the overall direction of effect for each gene set. FDR control for the number of gene sets tested was controlled using BH separately for each model.

The gene sets used were blood transcription modules (BTMs) from Chaussabel *et al.* [254] (prefixed 'DC') and S. Li *et al.* [158] (prefixed 'LI'). Modules are sets of genes with transcriptional and functional similarities across a variety of healthy, diseased and stimulated conditions. The 260 modules from Chaussabel *et al.* [254] were constructed by unsupervised clustering of 239 peripheral blood mononuclear cell (PBMC) transcriptomes from multiple disease datasets, then annotated by data mining of gene names in PubMed abstracts. The 334 modules from S. Li *et al.* [158] were constructed from coexpression analysis of approximately 30 000 blood transcriptomes, then annotated making use of Gene Ontology terms, cell type-specific markers, pathway databases, and manual literature searches. S. Li *et al.* [158] modules are better annotated in general, and were used for the majority of gene set enrichments in this chapter.

4.2.7 Genotyping and genotype data preprocessing

Genotype data were subsetted from the post-quality control PANTS cohort genotypes used in Sazonovs *et al.* [247], where the preprocessing pipeline is described in full detail. These data are from whole blood samples collected into EDTA tubes at week 0 and genotyped on the Illumina CoreExome microarray. Pre-imputation quality control was performed as described in de Lange *et al.* [218]. Imputation was done using the Sanger Imputation Service with the Haplotype Reference Consortium panel. Post-imputation, samples that were non-European, related (proportion identity-by-descent > 0.1875), or were outliers in genotype missingness or heterozygosity rate

move this to ch2, then refer to it

collab note: QC done by Alex

were removed; SNPs that poorly imputed (INFO score < 0.4), deviated from Hardy-Weinberg equilibrium (HWE) ($p < 1e-10$), had high missingness (> 5%), or low minor allele frequency (MAF) (< 1% before subsetting) were removed. 7 500 000 SNPs remained after filtering. Genotypes were converted to dosages of the non-reference allele.

4.2.8 Response expression quantitative trait locus mapping

The overall strategy and methods used were largely identical to chapter 3. For each timepoint, covariates for the expression quantitative trait locus (eQTL) models were computed, eQTLs were mapped separately for each timepoint, followed by joint mapping using summary statistics from all timepoints to call reQTLs.

4.2.8.1 Computing genotype PCs

Samples were projected onto principal components (PCs) defined by 1000 Genomes Project samples using SNP weights from akt, confirming that samples were of European ancestry (Figure 4.3). The first four PCs in the 1000G genotype data were found to be significant according to the Tracy-Widom test in subsection 2.2.5. Here I chose the first five PCs for use downstream; the specific number is not important, as long as a sufficient number of PCs are included to capture large-scale population structure [116]. PCs were centered and scaled before downstream use to improve model convergence.

4.2.8.2 Finding hidden confounders in expression data

Between-sample normalisation and variance stabilisation was applied to the counts matrix (DESeq2::vst), resulting in log2 scale expression estimates. Given known factors (response, drug, five scaled genotype PCs, cell proportions), PEER was used to find additional hidden factors that explain variance in the expression matrix for a large fraction of genes. This is similar to the process undertaken in subsubsection 4.2.6.1, but these hidden factors can be unmeasured. To maximise efficiency for cis-eQTL mapping, the number of PEER factors retained for each timepoint was selected to maximise the number of genes with at least one significant eQTL detected on chromosome 1 (Figure 4.4). The selected numbers were 25, 20, 15, and 5 for weeks 0, 14, 30, and 54.

4.2.8.3 Computing kinship matrices

Leave-one-chromosome-out (LOCO) kinship matrices were computed on typed SNPs for each chromosome using LDAK as described in subsubsection 3.2.3.1. The kinship matrix will be incorporated into the eQTL model to adjust for fine-scale population structure.

4.2.8.4 Mapping cis-eQTLs in each timepoint separately

eQTLs were mapped in each timepoint using a linear mixed model in limix. Within each timepoint, patients with multiple samples (taken on different study days) had the non-major visit sample removed. Although the eQTL model form does not prohibit duplicate samples, it led to strange behaviour in limix where eQTL betas and standard errors were comparable to

I've collected the QC details here from Alex's thesis

precision for integers

much of this section is brief, since the same pipeline as ch3

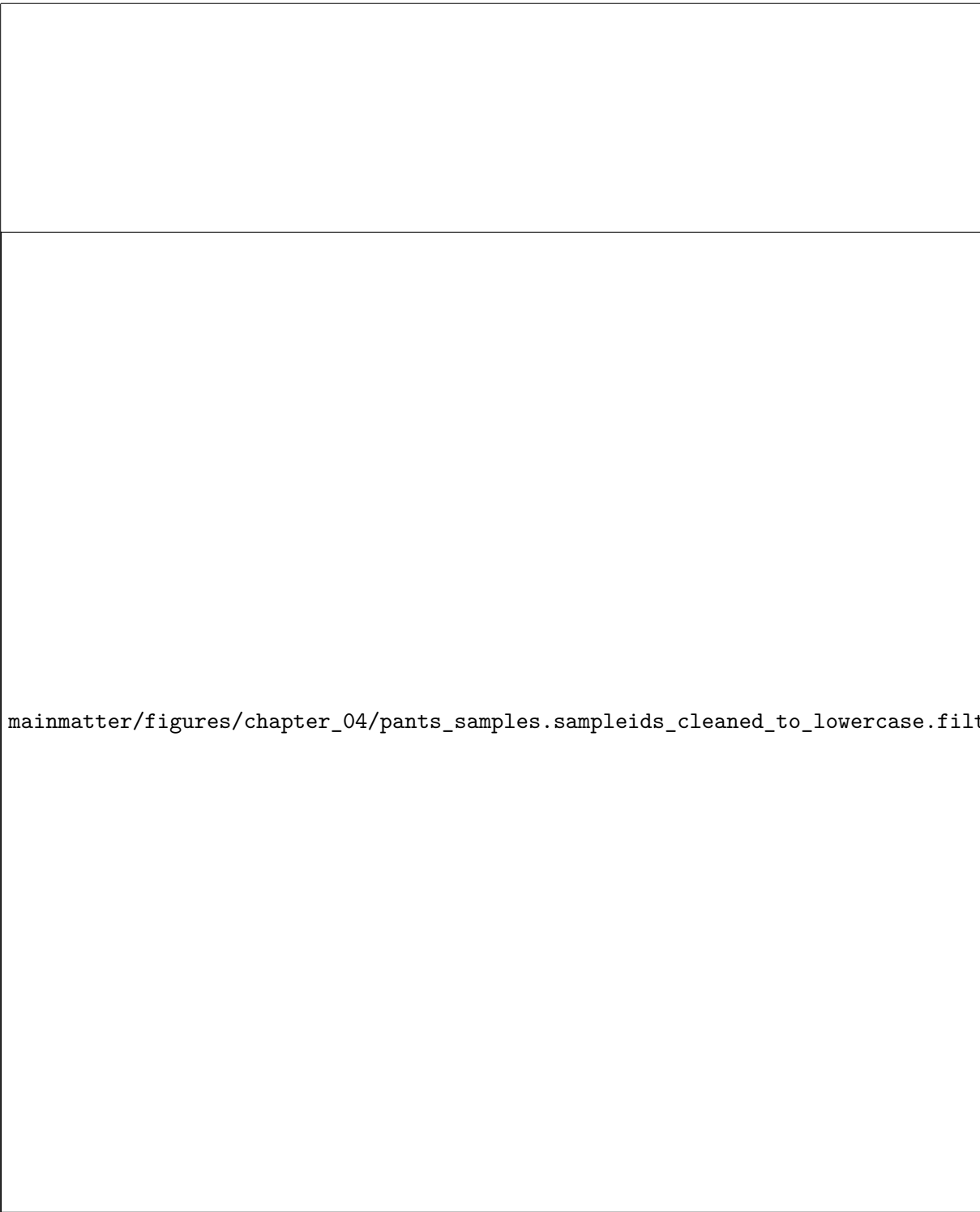


Figure 4.3: 1000G samples and PANTS samples projected onto 1000G genotype PC1 and PC2 axes, colored by superpopulation (a) and population (b).

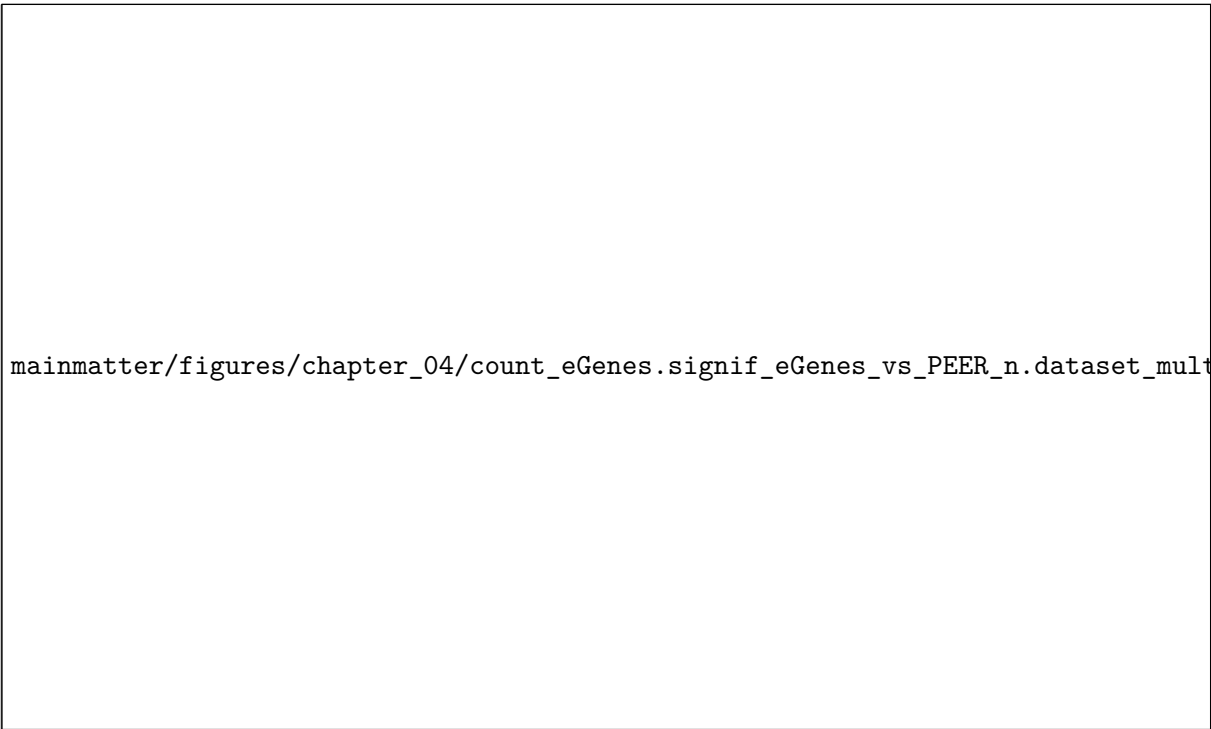


Figure 4.4: Number of eGenes on chr1 vs number of PEER factors included in eQTL mapping. The number of PEER factors for each timepoint is chosen after the plateaus.

the deduplicated results, but log-likelihood estimates were abnormally high. The deduplicated sample sizes at weeks 0, 14, 30 and 54 were 223, 205, 167, and 84.

For each autosomal gene, cis-SNPs within 1 Mb of the Ensembl gene start (gene end on the minus strand), were filtered to keep SNPs where the number of samples in the smallest dosage group was at least 5. Small group numbers lead to data points with high leverage that may be unduly influential on the genotype beta. This is done in place of a per-timepoint MAF or minor allele count filters.

The limix model for each SNP-gene pair had log2 expression as the response variable and genotype dosage as the predictor of interest. Other fixed effect predictors were the intercept, known factors (response, drug, five scaled genotype PCs, cell proportions), and PEER hidden factors (timepoint-specific number selected above). A random intercept term was also included with mean zero and covariance matrix proportional to the LOCO kinship matrix for the SNP’s chromosome.

4.2.8.5 Joint reQTLs mapping over all timepoints

Summary statistics from per-timepoint mapping were input to `mashr` [173]. A total of 25 900 000 SNPs were tested in all timepoints. The null correlation structure of the timepoints was estimated using null tests within a random subset of 200 000. Data-driven covariance matrices representing patterns of effects across timepoints were estimated using a strong subset of 129 000 eQTLs. The strong subset should contain eQTL that are likely to have an effect in at least one timepoint. For each gene and timepoint, I selected the eQTL with the smallest p -value, if that p -value was < 0.05 . The mash model was fit on the full random subset, accounting for the

as with most of this section, this is more explained in ch3

computed null correlation and covariance matrices, in exchangeable Z-scores mode. Finally, posterior betas and standard errors were computed for all tests using the fitted model parameters. A corresponding local false sign rate (lfsr) is also returned, controlling for multiple testing.

The lead eQTL for each gene was chosen as the eQTL with the lowest lfsr in any condition, breaking ties by highest INFO, highest MAF, shortest dist to gene start (or end), and smallest genomic coordinate. Each lead eQTL was assessed for being a significant reQTL by a Z test for whether the difference in betas was zero, between the week 0 beta and each of the other three weeks. Multiple testing for the number of genes was controlled using the BH FDR for each of the three comparisons separately.

4.3 Results

4.3.1 Longitudinal RNA-seq data from the PANTS cohort

To define transcriptomic differences between primary responders and non-responders to anti-TNF therapy in the PANTS cohort, I analysed whole blood RNA-seq gene expression measured at up to four timepoints per patient: week 0 baseline before commencing anti-TNF therapy, and weeks 14, 30 and 54 after commencing anti-TNF therapy. After quality control, expression data was available for 15584 genes and 814 samples (Figure 4.5). These samples came from 324 patients, with a median of three samples per patient (Figure 4.6).

Patient characteristics are shown in Table 4.1. The proportion of primary non-responders is high (43.8%) compared to the overall proportion in the PANTS cohort (23.8%, [233]). This is due to sample selection for RNA-seq to balance the sample size for each combination of drug and primary response status.

4.3.2 Baseline gene expression associated with primary response

Patient primary response to anti-TNF was defined after the induction period (week 12–14) according to the clinical decision algorithm from Kennedy *et al.* [233] described in subsection 4.2.3, which integrates clinician assessment with change in CRP level and HBI score. To identify differences in baseline gene expression associated with future primary response, I fit gene-wise linear models at 15511 genes, comparing week 0 gene expression in primary responders with primary non-responders. Comparisons were performed both within infliximab-only and adalimumab-only subgroups, and with both drugs pooled. Models were run both adjusting for cell composition estimates of six immune cell types, and without adjustment. Throughout this section, the significance threshold was set at $FDR < 0.05$ for each comparison, and positive logFCs indicate increased expression in responders versus non-responders.

Without adjusting for cell composition, the largest effects were infliximab-only, with 859 genes differentially expressed. Only *KCNN3* ($\log_2FC = -0.843$) was significant for the adalimumab-only comparison, and only *SIGLEC10* ($\log_2FC = 0.346$) was significant in the pooled analysis (Figure 4.7).

After adjustment for cell composition, there were no longer any significant genes in the infliximab-only analysis, with 856/859 genes that were significant before the comparison having

yes, there could be ties in lfsr

what to put in results vs discussion. going with the pattern of providing enough info for the reader to intepret the data in the results, then doing a summary and my own interpretation in the discussion

Some primary non-responders have loss of response samples. Not sure why.

add caption and label after finalising table data

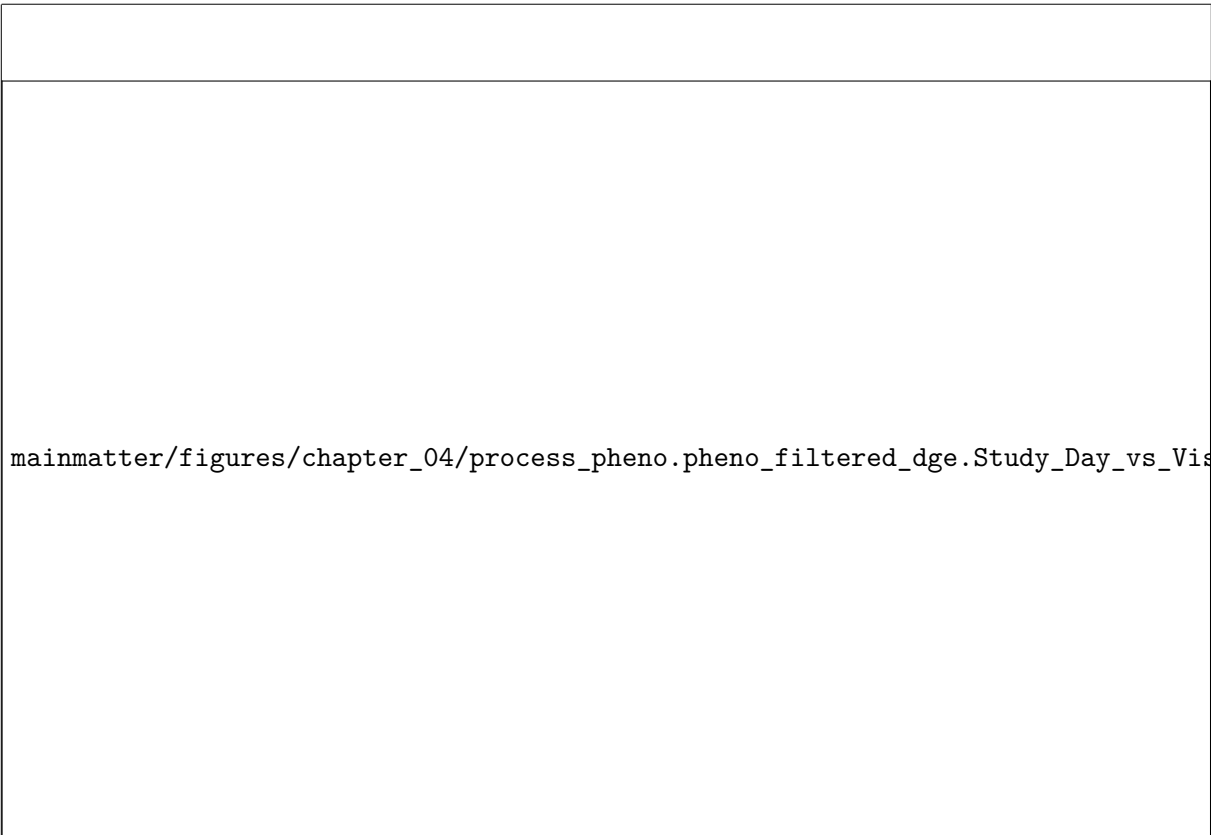


Figure 4.5: Number and distribution of RNA-seq samples in experimental groups over time. Windows for the four major visits are colored in grey. Samples mostly come from major visits, but a small number of LOR and exit visit samples were included according to subsection 4.2.2.

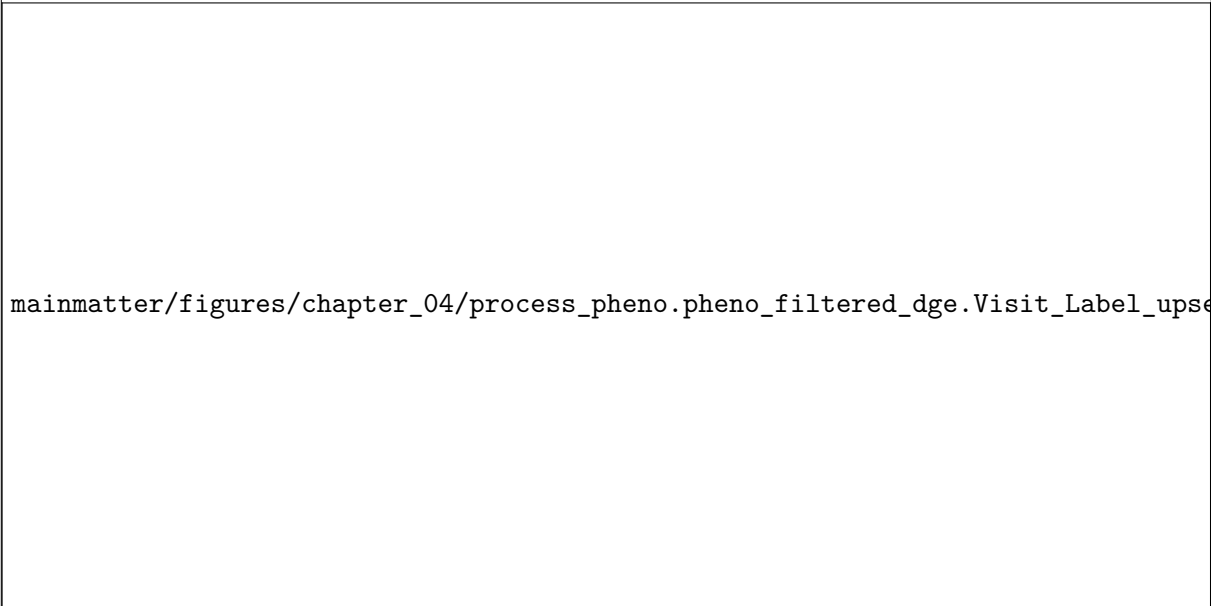


Figure 4.6: Distribution of RNA-seq samples from each patient among timepoints.

a dampened effect size after correction (smaller absolute effect and same sign), suggesting many effects may be mediated by cell composition. *SIGLEC10* in the combined analysis was also non-significant after adjustment (adjusted log2FC=0.314, FDR=0.0503). Conversely, at three genes downregulated in the adalimumab-only analysis that were the only significant genes post-adjustment, I observed increased significance: *PDIA5* (unadjusted log2FC=−0.335, adjusted log2FC=−0.351), *KCNN3* (−0.843, −0.880), and *IGKV1-9* (−1.15, −1.22).

To identify coordinately up and downregulated gene sets and increase sensitivity for detecting differences between responders and non-responders, I performed rank-based gene set analysis on the gene-wise z-statistics using blood transcriptomic modules: annotated sets of coexpressed genes in peripheral whole blood from S. Li *et al.* [158] (prefixed “LI”). This module-level analysis was also run unadjusted (Figure 4.8) and adjusted for cell composition (Figure 4.9).

Despite only *SAMD10* having a significantly different effect between drugs at the gene-level (a significant interaction between drug and response at week 0), the large global differences observable in Figure 4.7 were detected in the module-level analysis*. Without adjusting for cell composition, many of the most significantly upregulated modules in the pooled analysis, including upregulation of monocyte (LI.M11.0, LI.S4), neutrophil (LI.M37.1, LI.M11.2), and dendritic cell (LI.M165, LI.S11), appear to be driven by infliximab. These modules have heavily reduced significance after adjusting for cell composition. The new modules that are most upregulated in the pooled analysis after adjustment have more consistent effects between drugs, such as MHC-TLR7-TLR8 cluster (LI.M146), antigen presentation (LI.M71, LI.M95.0), and myeloid cell enriched receptors and transporters (LI.M4.3).

For downregulated modules before adjustment, I observed infliximab-specific effects for NK cell (LI.M7.2) and T cell (LI.M7.0, LI.M7.1) modules. Adalimumab-specific effects were observed for plasma cell, B cell and immunoglobulin modules (LI.M156.0, LI.M156.0, LI.S3); and cell cycle and transcription modules (LI.M4.0, LI.M4.1). After adjustment, the significance of infliximab-specific modules was reduced, but the significance of adalimumab-specific modules and the corresponding interaction effects was increased. In both gene-level and module-level results, there is striking heterogeneity between infliximab and adalimumab that is only partially reduced by cell proportion adjustment.

4.3.3 Assessing previously reported baseline predictors of primary response

In addition to hits from this study, Figure 4.7 is annotated with genes whose expression in gut biopsies or blood has been previously evaluated for baseline prediction of primary response [240, 241, 245, 255]. Some genes expressed in gut mucosa (e.g. *IL13RA2*) were not appreciably expressed in this whole blood dataset, and most other genes that were expressed were not significantly differentially expressed. Only *TNFRSF1B* and *PTGS2* were associated with primary response, specifically in the infliximab-only comparison, unadjusted for cell composition.

A previously identified marker in blood, *TREM1*, found to have opposing in two studies [244, 245] was not significantly associated with response in this study, neither before (logFC=0.293,

*It is likely the study is not powered to detect gene-level three-way interaction effects between timepoint, drug and response. I am not aware of which subgroup analyses may have been prespecified during the study design and sample size calculations for the PANTS RNA-seq cohort.



Figure 4.7: Volcano plots of DGE between primary responders 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. Annotated genes include significant associations from this study and previously reported associations from subsection 4.1.4. Dashed line shows significance threshold at $FDR = 0.05$.

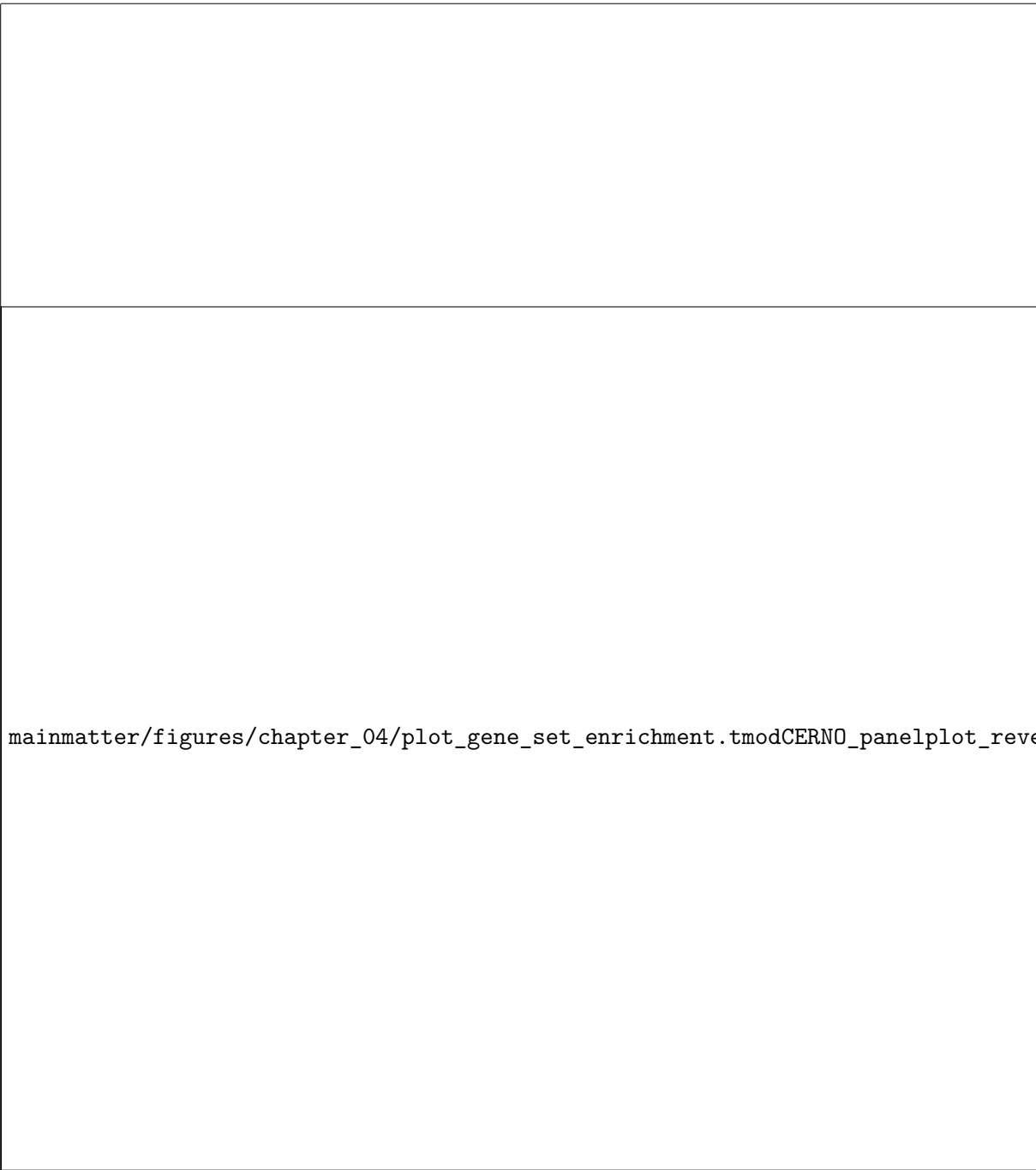


Figure 4.8: Top modules differentially expressed between primary responders and non-responders at week 0, unadjusted for cell composition. Columns correspond to results from infliximab (IFX), adalimumab (ADA), infliximab-adalimumab difference and pooled analyses. The top 30 modules ranked by minimum FDR in any column are shown. Vertical dashed line shows significance threshold at $FDR = 0.05$.

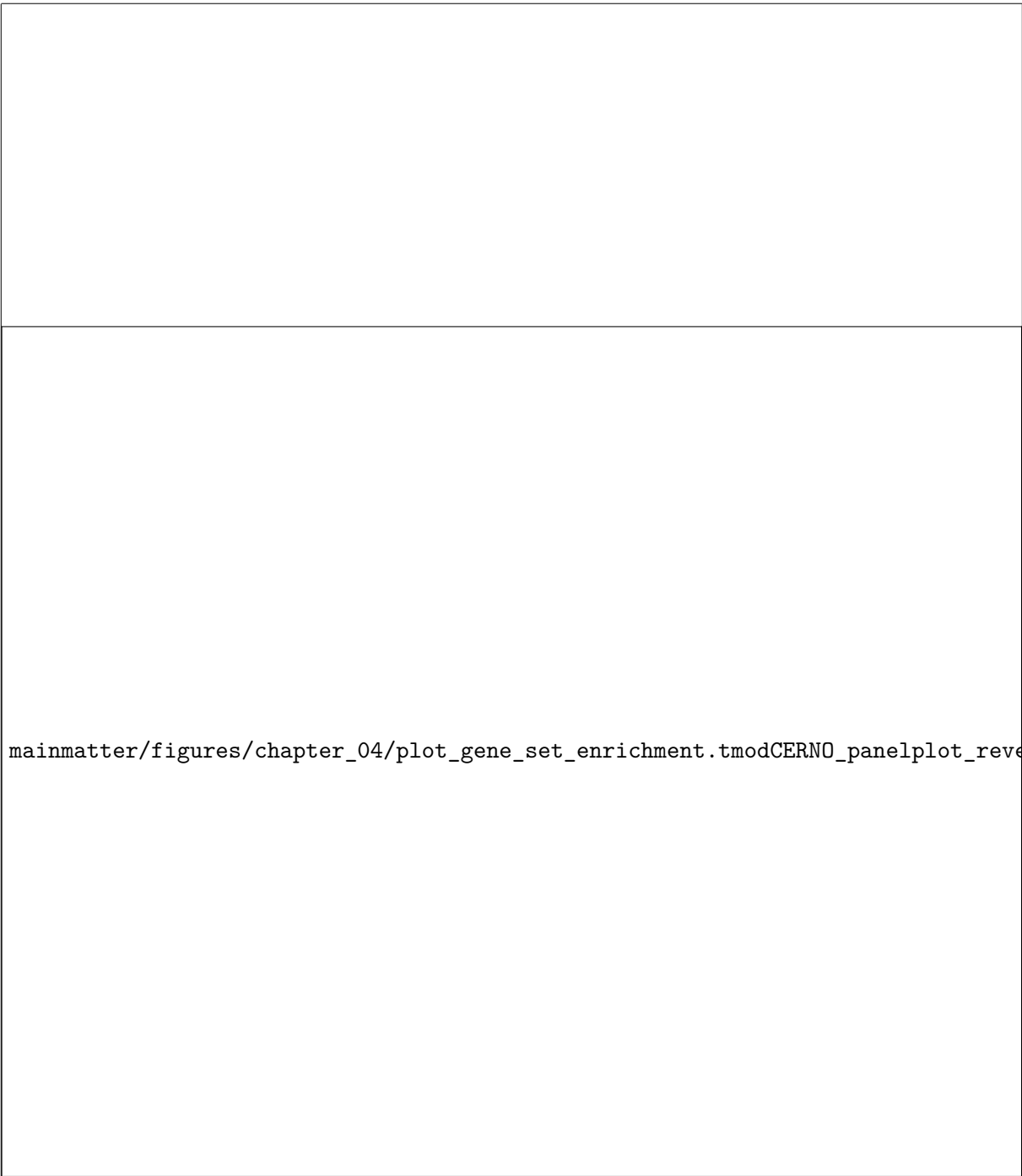


Figure 4.9: Top modules differentially expressed between primary responders and non-responders at week 0, adjusted for cell composition.

FDR=0.0595) nor after adjusting for cell composition (logFC=0.0463, FDR=0.986).

4.3.4 Post-induction gene expression associated with primary response

The same methodology applied at week 0 was applied at week 14 to identify differences in post-induction expression associated with primary response. A larger proportion of the transcriptome is differentially expressed at week 14: 1364 for the infliximab-only comparison, 1544 for the adalimumab-only comparison, and 4841 pooling both drugs (Figure 4.10). No significant interactions between drug and response were detected at the gene-wise level. Given that sample sizes at week 0 and week 14 are comparable (Figure 4.5), the overall signal-to-noise ratio is much stronger than at baseline.

Adjusting for cell composition, 1320/1367, 1515/1544 and 4653/4841 genes have dampened effects; and the numbers of significant genes drop to 379, 177, and 1302 for infliximab, adalimumab, and pooled analyses respectively. This again suggests many effects are mediated by differences in immune cell composition between responders and non-responders.

Modules including generic immune activation, monocytes, TLR and inflammatory signalling, and neutrophils were downregulated in responders; whereas B cell and plasma cell modules were upregulated (Figure 4.11). These modules remained differentially expressed with the same direction of effect after adjusting for cell composition (Figure 4.12), suggesting there is per-cell up or downregulation on top of abundance changes of the cell types expressing these modules. Modules related to antigen presentation (LI.M71, LI.M97.0, LI.M5.0), interferon (LI.M75, LI.M127, LI.M111.1), and dendritic cells (LI.M64, LI.M165) also appear among significantly downregulated modules after cell composition adjustment. Directions of effect for the most significant modules were largely consistent between drugs. The significance of drug by response interaction effects compared to main effects is less prominent than in the baseline analysis, where many of the strongest effects were driven by one drug.

SIGLEC10 from the baseline analysis retains its significant association with primary response post-induction, with the same direction of effect (adjusted logFC=0.366). Some genes previously proposed as baseline markers of response in gut mucosa: *G0S2*, *TNFAIP6*, *S100A8* and *S100A9* by [241]; and *OSM* by [242], were differentially expressed in post-induction blood in this study. The direction of effect in both cases, downregulation of markers in primary responders, also matches this study.

4.3.5 Magnification of expression change from baseline to post-induction in responders

Given the stronger differences in expression between primary responders and non-responders at week 14 versus week 0, I estimated the change in expression from week 0 to week 14 within the two groups, and also estimated the timepoint by response interaction. I perform only the pooled comparison here both to simplify the analysis, and because similarly to the within week 14 comparison, change from week 0 to week 14 was relatively consistent between drugs, with exceptions noted.

Without adjusting for cell composition, 12862 genes were differentially expressed in primary responders at week 14 vs week 0 in the pooled analysis, 8310 genes in primary non-responders,

highlight a few more individual genes specific to this analysis too? not sure how to pick them at the moment.



Figure 4.10: Volcano plots of DGE between primary responders 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. Annotated genes include significant associations from this study and previously reported associations from subsection 4.1.4. Dashed line shows significance threshold at $FDR = 0.05$.

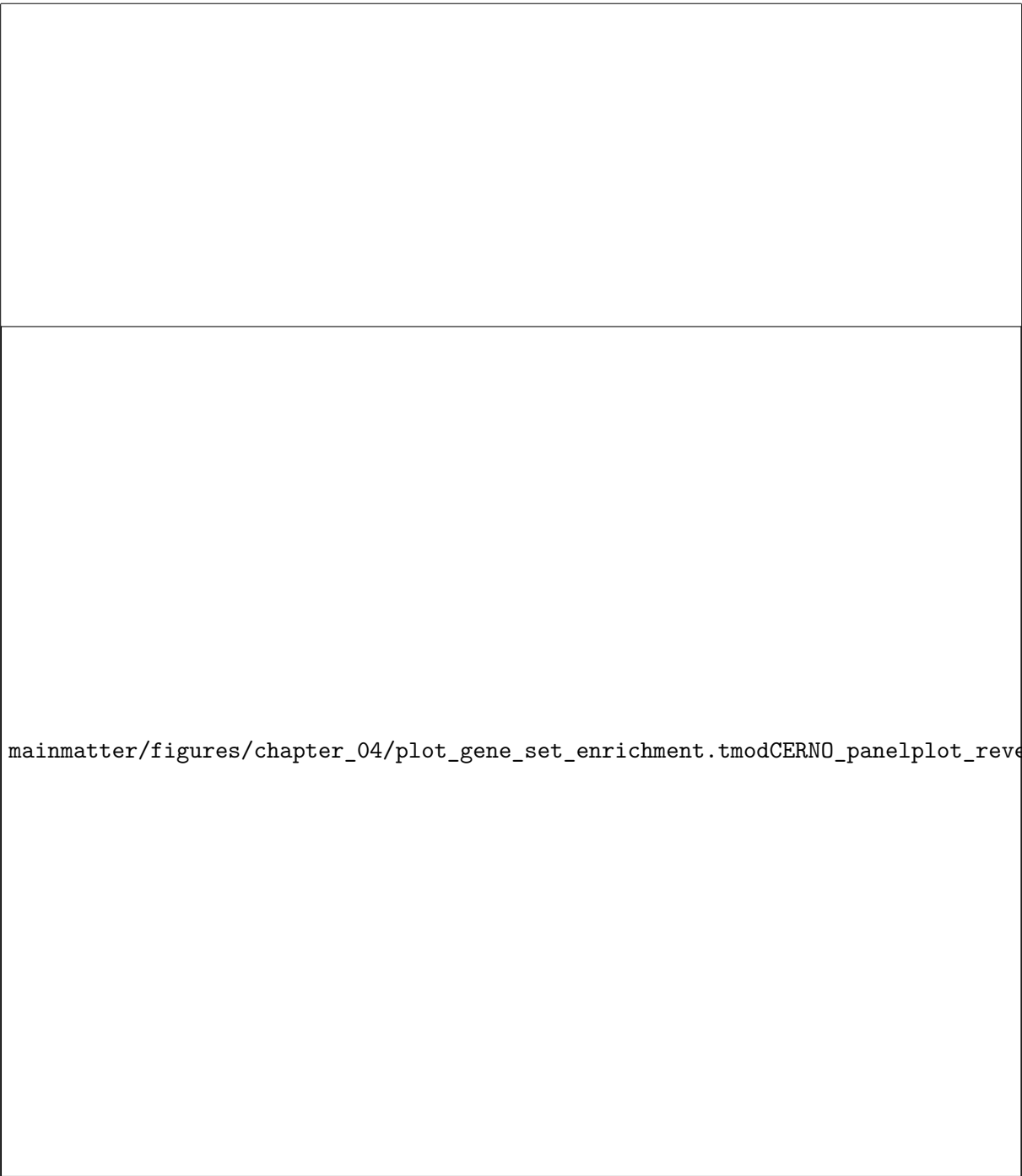


Figure 4.11: Top modules differentially expressed between primary responders and non-responders at week 14, unadjusted for cell composition.



mainmatter/figures/chapter_04/plot_gene_set_enrichment.tmodCERNO_panelplot_reversed_C_3RI_3NI,C_3RA_3NA

Figure 4.12: Top modules differentially expressed between primary responders and non-responders at week 14, adjusted for cell composition.

and 6320 genes had a significant interaction. After adjusting for cell composition, 5572 genes were differentially expressed in primary responders, 626 genes in primary non-responders, and 179 genes had a significant interaction. Of the genes differentially expressed between week 14 and week 0 in both primary responders and non-responders, and with a significant interaction between timepoint and response, nearly all (4885/4891 unadjusted for cell composition, 31/32 adjusted) were magnified by primary response, such that the same genes have larger fold-changes in the same direction for primary responders (Figure 4.13).

The most significant modules that change from week 0 to week 14 in responders included upregulation of B cell (LI.M47.0), plasma cell (LI.M156.0), and T cell activation (LI.M7.1); and downregulation of immune activation (LI.M37.0), monocyte (LI.M11.0), neutrophil (LI.M37.1) and TLR and inflammatory signalling (LI.M16) modules (Figure 4.14). Many of these are the same modules associated with response within the week 0 and week 14 timepoints.

Adjusting for cell composition decreases the significance of a majority of modules (Figure 4.15), with T cell modules in the adalimumab-only analysis especially decreased. Magnification is also observed at the module level, with nearly all module effects aligned in the same direction in responders and non-responders, with significant interactions also in the same direction.

4.3.6 Interferon modules with opposing differential expression in responders and non-responders

Figure 4.13 also contains genes that were downregulated from week 0 to week 14 in responders, but upregulated in non-responders (“flipped”). At the module level these opposing effects were most apparent after cell composition adjustment for antiviral interferon signature (LI.M175), type I interferon response (LI.M127), and antigen presentation (LI.M95.0) (Figure 4.15). I extended my module analysis to include modules from [254] (prefixed “DC”). Although these modules are on the whole poorly annotated compared to modules from [158], interferon modules are annotated. *STAT2*, *GBP5*, and *PARP14* from Figure 4.13 are annotated into an interferon module (DC.M3.4). *IFIT3* and *GBP2* are also annotated into separate interferon modules (DC.M1.2, DC.M5.12). Adjusted for cell composition, these modules are all significantly upregulated at week 14 in non-responders only: DC.M3.4 $\text{FDR}=3.45 \times 10^{-21}$, DC.M1.2 $\text{FDR}=9.49 \times 10^{-16}$, and DC.M5.12 $\text{FDR}=1.36 \times 10^{-13}$ (Figure 4.16).

4.3.7 Sustained expression differences between primary responders and non-responders during maintenance

As PANTS is an observational study, it was able to include some patients who continued with anti-TNF therapy even after meeting the definition of primary non-response at week 14. For both responders and non-responders, expression data could also be available from blood samples around week 30 and week 54, and at additional visits scheduled in the event of secondary LOR.

I fit a natural cubic spline to the expression of each gene as a function of study day, and tested for general differences in expression over time between responders and non-responders. This analysis was done only with drugs pooled due to lower sample sizes for later timepoints. Without

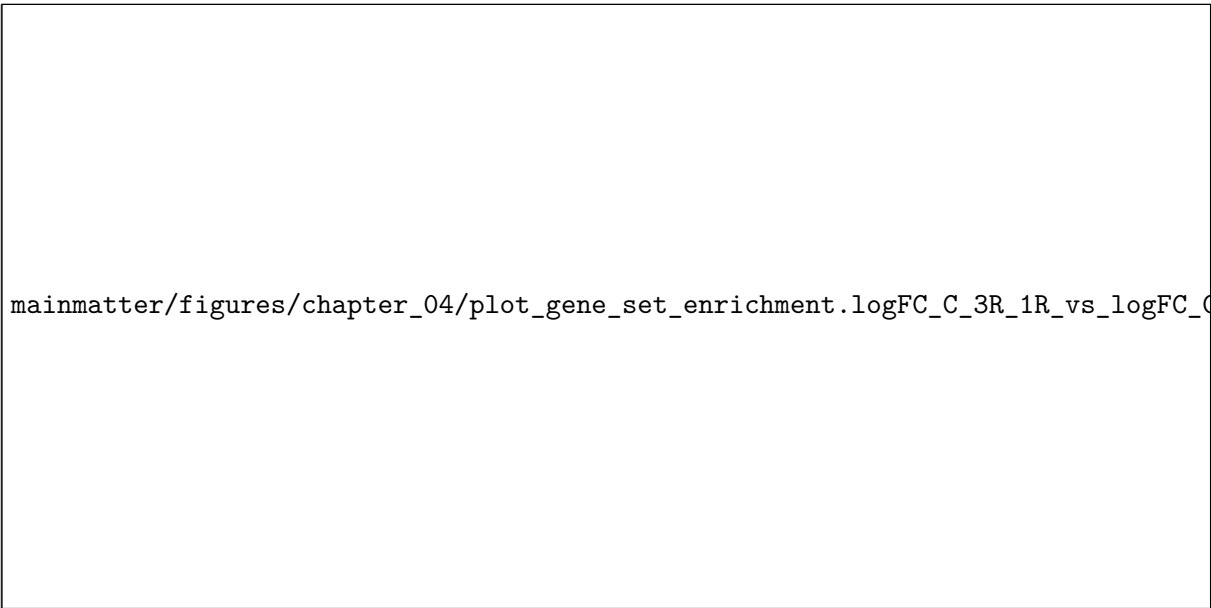


Figure 4.13: Expression logFC from week 0 to week 14 in primary responders versus non-responders, shown for all genes differentially expressed from week 0 to week 14 in both responders and non-responders, with a significantly different effect size in responders and non-responders. Identity line shown by dashed line.

adjusting for cell composition, 4426 genes were differentially expressed between responders and non-responders; 210 genes were differentially expressed after adjustment. To identify distinct trajectories of expression over time, I hierarchically clustered those 210 genes by their mean expression in responders and non-responders at each timepoint, and determined the optimal number of clusters by the gap statistic method (Figure 4.17). Six distinct clusters were proposed (Figure 4.18).

Many of these genes had previously been identified as having significant differences in expression between responders and non-responders either within week 14, or for change in expression from week 0 to week 14. Cluster 1 contained mainly previously identified genes (Figure 4.19), and was enriched for modules including myeloid cells and monocytes (LI.M81, hypergeometric test, $FDR=2.11 \times 10^{-6}$), platelet activation (LI.M196, $FDR=1.35 \times 10^{-5}$), immune activation (LI.M37.0, $FDR=1.44 \times 10^{-4}$), and TLR and inflammatory signalling (LI.M16, $FDR=2.36 \times 10^{-3}$). The spline analysis highlighted that expression differences at week 14 are maintained at week 30 and week 54.

The highest proportion of genes uniquely identified as significant by the spline analysis were in cluster 2 (26/31) and cluster (15/20). Cluster 2 was enriched in [158] B cell modules (LI.M47.0, $FDR=1.53 \times 10^{-6}$; LI.M47.1, $FDR=4.53 \times 10^{-5}$) previously identified as having a greater increase from week 0 to week 14 in primary responders versus primary non-responders (Figure 4.15), matching the observed cluster trajectory. Cluster 4 was not enriched in any modules from S. Li *et al.* [158], but is enriched for a B cell module (DC.M4.10, $FDR=0.00137$) from Chaussabel *et al.* [254]. Although no genes were significantly associated with response at week 0 (Figure 4.7), the genes in cluster 4 are coordinately downregulated as a set in primary responders (CERNO test, $p=6.18 \times 10^{-25}$).

Cluster 3 is also of interest, enriched for type I interferon response (LI.M127, $FDR=0.00568$) and interferon (DC.M3.4, $FDR=0.000527$) modules, as well as genes that contain putative

TODO I forget to include the non-ranked hypergeometric test and gprofiler tests in the gene set enrichment section in the methods.

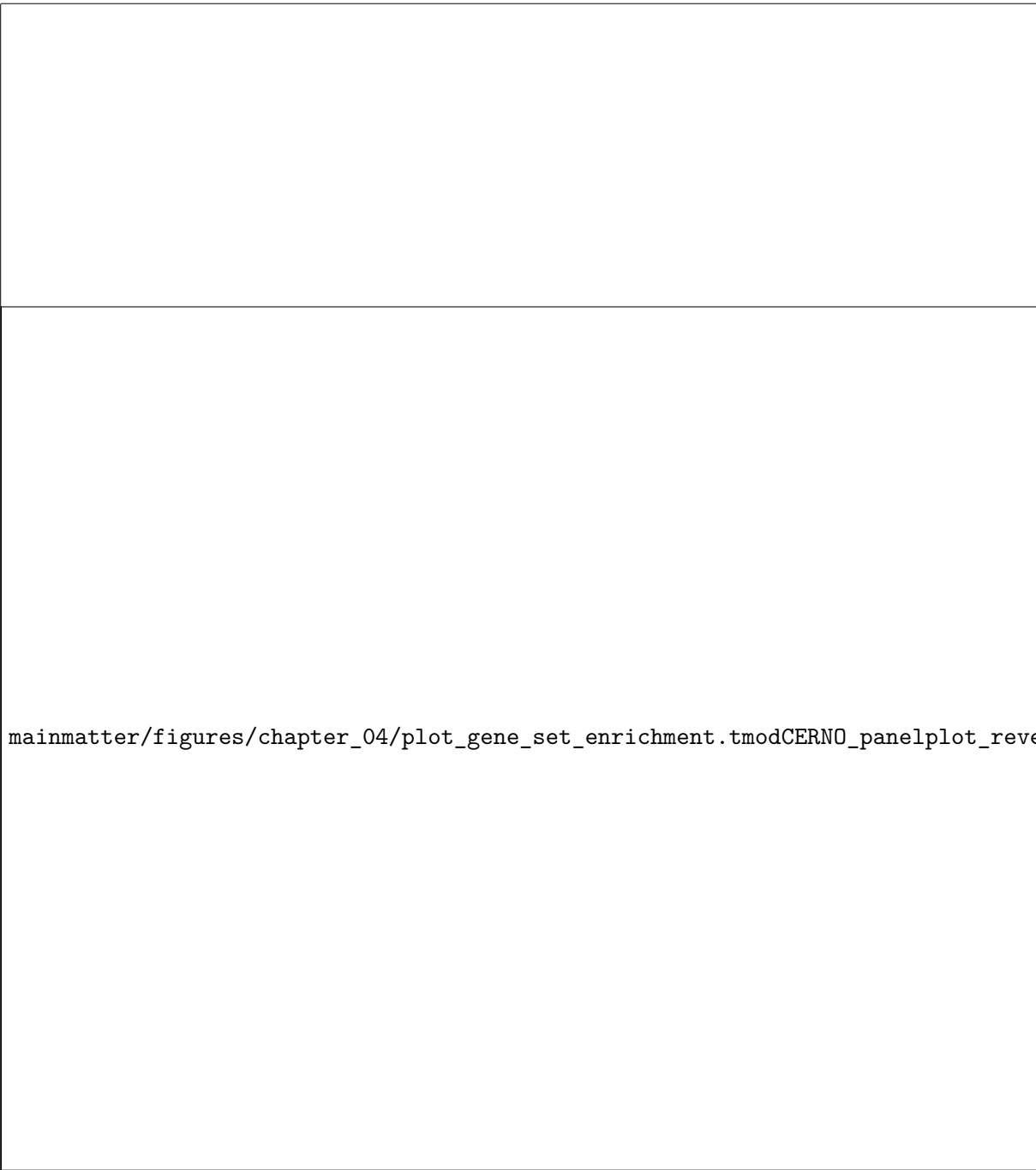


Figure 4.14: Top modules differentially expressed between week 14 and week 0, unadjusted for cell composition. Columns show effects in primary responders (PR), non-responders (PNR), and the primary responder minus non-responder difference. The top 30 modules ranked by minimum FDR in any column are shown. Vertical dashed line shows significance threshold at $FDR = 0.05$.



Figure 4.15: Top modules differentially expressed between week 14 and week 0, adjusted for cell composition. Columns show effects in primary responders (PR), non-responders (PNR), and the primary responder minus non-responder difference. The top 30 modules ranked by minimum FDR in any column are shown. Vertical dashed line shows significance threshold at $FDR = 0.05$.



Figure 4.16: `tmod` evidence plots showing interferon-related modules specifically upregulated from week 0 to week 14 in primary non-responders. Genes were ranked in ascending order by week 14 versus week 0 DGE z-statistic. The ranks of genes in interferon-related modules are indicated by colored rug plots. Colored curves show the cumulative fraction of genes in each module. For non-responders, these modules are enriched for large ranks (large, positive z-statistics). The area between the colored curves and the null of randomly-distributed ranks (grey diagonal line) is the effect size (AUC).



Figure 4.17: Gap statistic versus cluster number k . Error bars derived from 500 bootstraps. The optimal number of clusters is defined as the smallest k after which the gap statistic does not significantly increase at $k+1$.



Figure 4.18: Normalised expression over the timepoints for genes in the six identified clusters. Error bars are the expression mean and standard deviation for the genes at each timepoint in primary responders and non-responders.



Figure 4.19: Venn diagrams showing which genes in each spline cluster were also significant in the week 14 responder versus non-responder contrast, or the interaction between week 0 to week 14 change and response status.

transcription factor binding motifs for interferon regulatory factors *IRF7* (g:Profiler term ID TF:M00453_1, adj. p value=0.005 05) and *IRF8* (TF:M11684_1, adj. p value=0.007 77; TF:M11685_1, adj. p=0.0103). The cluster trajectory shows direction of expression change is opposing in responders and non-responders from week 0 to week 14, followed by sustained differences at week 30 and week 54. The trajectory and interferon-related gene set enrichments are consistent with those identified in subsection 4.3.6. Of the 9 genes in this cluster, 8 genes (*STAT1*, *BATF2*, *GBP1*, *GBP5*, *IRF1*, *TAP1*, *APOL1*, *APOL2*) have significant interaction between week 0 to week 14 expression change and response status, whether or not correcting for cell composition. However, only *GBP5* was differentially expressed from week 0 to week 14 in both responders and non-responders, and only when unadjusted for cell composition (Figure 4.13). This indicates that such small and opposite effects in responders and non-responders are best detected at a single-gene level in the interaction analysis that tests the difference, and in the spline analysis, with the support of additional data from week 30 and week 54.

4.3.8 Limited evidence for change in genetic architecture of gene expression over time

Given the substantial changes in expression from baseline to post-induction after starting the drug, and the differing trajectories observed in responders and non-responders, I performed eQTL mapping to identity common genetic variants associated with expression that may contribute to these differences. Variants cis (within 1 Mb of the TSS) to 15040 genes were tested for association. Mapping was done within each timepoint (weeks 0, 14, 30, and 54), followed by joint analysis of per-timepoint eQTL summary statistics and control for multiple testing using *mashr*.

The majority 11156/15040 (74.2 %) of genes were eGenes (a gene with at least one significant cis-eQTL) in at least 1 timepoint (lfsr < 0.05). The variant with the lowest lfsr in any timepoint for each gene was chosen as the lead variant (eSNP) for that gene. Most eSNPs were significant in multiple timepoints: 999 significant in 1 timepoint, 381 significant in 2 timepoints, 526 significant in 3 timepoints, 9250 significant in all 4 timepoints. I compared eSNP effect sizes between week 0 and each of weeks 14, 30 and 54, to identify reQTLs with significant difference in effect versus baseline, as they may also explain changes in expression from baseline. Most eSNPs were shared across timepoints; only six eSNPs-eGene pairs were significant at BH FDR < 0.05: with 1/6 for week 30 versus week 0 and 5/6 for week 54 versus week 0 (Figure 4.20). Of the six eGenes, *NMT1* and *EPSTI1* both had their respective eSNPs having magnified effects on expression at week 54 compared to week 0, and both are annotated to contain putative binding motifs for *IRF8* and *IRF2* (g:Profiler term IDs TF:M11685_1 and TF:M11665_1). However, direct interpretation of these reQTLs is complicated by confounding by cell composition in bulk expression data (discussed in subsection 3.2.11).

4.4 Discussion

In PANTS, a cohort of CD patients receiving infliximab or adalimumab anti-TNF therapy for the first time, there were substantial differences in whole blood gene expression between primary

right word???



So many modules associations here, maybe try to Google some of them...

responders and non-responders. At baseline, the greatest differences in expression were observed between future responders and non-responders to infliximab, with increased expression of monocyte, neutrophil and dendritic cell gene modules in responders, and decreased expression of T cell and NK cell modules. These effects appear to be infliximab-specific, and are attenuated after adjusting for the proportions of six major immune cell types, suggesting expression differences may be driven by mediation via the proportions of these cell types.

In contrast, future responders to adalimumab had lower baseline expression of plasma cell and cell division modules. The module-level results line up with the three gene-wise hits for the adalimumab-only analysis: *IGKV1-9* encodes the immunoglobulin light chain variable region that forms part of antibodies produced by plasma cells, *KCNN3* is annotated to plasma cell surface signature module (LI.S3 [158]), and the expression of both *KCNN3* and *PDIA5* are correlated with blood plasmablast frequencies [108]. It was reported by Gaujoux *et al.* [244] that baseline plasma cell abundances are lower for infliximab responders, hypothesising that plasma cell survival is supported by increased TNF levels in non-responders. Plasma cells also formed a part of a correlated module of cell populations identified by [243], where lower module expression was associated with better response to anti-TNF in a cohort with patients taking both infliximab and adalimumab. However, both these studies were done in gut biopsy samples, and there was no mention of strong between-drug heterogeneity.

The adalimumab-specific associations I found were more significant after cell proportion adjustment, which may indicate per-cell downregulation rather than cell abundance being associated with response. However, cell composition differences mediated by rarer cell types that have abundances poorly captured by the six major types used in the model will be poorly adjusted for. For example, plasma cell proportions are only weakly correlated with other immune cell types in the healthy immune system [256], although the relationship may differ for CD patients. It has also been shown that differentially expressed genes in blood with correction for only common cell types will identify associations that are proxies for rare cell types <https://www.biorxiv.org/content/10.1101/2020.05.28.120600v1>. If this is the case, the role of the cell composition estimates for adalimumab-specific effects may be more akin to precision variables, which would be consistent with increased significance after adjusting.

The differences between drugs are puzzling, especially the greater effect of cell composition adjustment for infliximab. Baseline patient differences between drugs may offer a partial explanation. There may be characteristics not listed in Table 4.1 that differ between patients on different drugs [233]. In the full PANTS cohort, lower albumin, higher CRP, and higher faecal calprotectin in infliximab patients suggest that they may have had greater disease severity. Differences may be driven by patient or physician preference, for example, patients with more severe disease are often given infliximab rather than adalimumab*. I have not yet been able to access clinical variables such as CRP and faecal calprotectin levels to consider as variables to adjust for in my modelling. A richer phenotype dataset containing some of these variables has been requested from collaborators.

The strongest single-gene association in the pooled analysis was *SIGLEC10*, which had reduced significance post-adjustment with a comparable effect size, where baseline expression was

*Kennedy, N. A., personal communication, 4 June (2020).

approximately 25% higher in responders. Direction of effect was consistent between drugs, but most significant in infliximab without cell composition adjustment. In IBD, small molecules called damage-associated molecular patterns (DAMPs) are released due to tissue damage and cell death, and further promote inflammation through pathogen sensing pattern recognition receptor (PRR) pathways that include Toll-like receptors (TLR) family receptors [216, 257]. For instance, faecal calprotectin, a marker for IBD activity, is a complex of two DAMPs, S100A8 and S100A9 [216]. SIGLEC10 has been shown to repress DAMP-mediated inflammation through binding CD24 [257]. *SIGLEC10* is expressed on B cells, monocytes and eosinophils <https://www.nature.com/articles/nri2056>, and of these cell types, module level results posit monocytes as the most likely candidate cell type to have increased module expression in responders. In monocytes, *SIGLEC10* expression is more specific to the CD16+ monocytes [258], and in particular the CD14+CD16++ non-classical monocytes rather than the classical CD14++CD16- or intermediate CD14++CD16+ subsets [259]. In PANTS, it was suggested by Kennedy *et al.* [233] that higher inflammatory load as indicated by low baseline albumin levels may result in low week 14 drug levels due to faster drug clearance, and low drug levels at week 14 were in turn associated with non-response. A hypothetical model might be high baseline *SIGLEC10* expression reflecting higher proportions of CD16+ monocytes (or lower proportions of CD16- monocytes), decreased DAMP-mediated inflammation, and increased chance of primary response, possibly by affecting drug clearance rate. This is an extremely tentative model: both the cell proportion estimates and module definitions used thus far only represent monocytes as a whole, lacking the resolution to properly explore shifts in the three monocyte subsets. It may be possible to use expression of monocyte subset marker genes such as those identified by Villani *et al.* [259] to improve the resolution of the cell proportion estimates.

Despite the strong heterogeneity in effects between drugs, one consistent effect that emerged after adjusting for cell composition was baseline upregulation of MHC-TLR7-TLR8, antigen presentation, and interferon modules in responders. As mentioned above, TLR receptors are involved in pathogen sensing, and TLR7 and TLR8 are endosomal proteins primarily expressed in monocytes, macrophages and dendritic cells (DCs), part of an antigen presentation pathway that senses bacterial DNA and activates downstream innate immune pathways including type I interferon response <https://www.nature.com/articles/cmi201238>. Type I interferons have pathogenic or protective roles in many IMIDs <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4084561/>. It has been suggested that type I interferon responses induced via TLR7 and TLR8 can suppress colitis in mouse models, and play a role in maintaining gut homeostasis <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5797585/> <https://www.frontiersin.org/articles/10.3389/fmed.2018.00032/full>, so upregulation here may again represent a less severe baseline disease in future responders.

Most previously reported baseline markers in blood and gut biopsies were non-significant in this study. For gut markers, this may not be unexpected. Although a subset of gut infiltrating immune cells and their precursors may also be circulating, genes specific to epithelium and immune cell types that differentiate after they migrate into tissues (e.g. monocyte-derived macrophages), will be difficult to observe in blood. For blood markers, I sought to clarify the conflicting results in the literature about the association of *TREM1* expression in blood with anti-TNF response

How does the MHC come in here?

[244, 245]. I did not find *TREM1* to be significantly differentially expressed in PANTS, although the direction of effect is increased expression in responders, matching the Gaujoux *et al.* [244] direction of effect in blood. *TREM1* is expressed on myeloid lineage cells such as monocytes and macrophages; Villani *et al.* [259] reported that *TREM1* expression is most specific to classical monocytes and a newly identified subtype within the intermediate monocytes (“Mono3”). The *TREM1* effect is one of the infliximab-specific differentially expressed genes that is much stronger without cell proportion adjustment, so it may reflect association of baseline monocyte cell proportions with response.

There are many factors could explain failures to replicate reported markers, or identification of different markers from study to study. Many existing studies pool cohorts with different anti-TNF biologics due to the scarcity of large datasets, yet even within this study, there is heterogeneity between drugs. There are between-study differences in the definition of primary response, such as endoscopic healing [244] versus scoring on clinical parameters [245]. Any two studies are unlikely to have adjusted for the same combinations of covariates in modelling, and some covariates like cell composition are very influential for bulk expression data. Finally, small sample sizes have considerable sampling error. Set-based association tests that draw on changes in multiple genes, such as the expression module associations for blood in this study, may be more reproducible compared to single-gene markers.

Although this study is purely descriptive, a future aim will be to see if the identified baseline module associations also imply that response status can be predicted from baseline expression. Because modules associated with response appear to be mediated by cell proportions, much of the predictive ability may also lie in differences in cell proportions between responders and non-responders. Indeed, Gaujoux *et al.* [244] noted that adjusting expression for cell composition resulted in gut gene signatures that were worse at discriminating responders from non-responders. Testing specific subpopulations such as CD16+ monocyte or plasma cell abundance for association with response can also be viewed as a type of set-based test that represents a set of cell-type specific genes, and thus may also be more reproducible than single-gene markers.

Much larger proportions of the transcriptome are associated with response after the induction period at week 14. Module associations showed downregulation of immune activation, TLR, inflammatory, monocyte and neutrophil modules in responders; and upregulation of B and T cell modules. Similar module associations were also found when considering modules differentially expressed from week 0 to week 14. The differences between responders and non-responders at week 14 were qualitatively similar to the differences pre and post anti-TNF induction, suggesting there may be relatively little change in the transcriptome of non-responders after induction. Associations were generally consistent between drugs for both the within week 14 and change from week 0 to week 14 analyses, perhaps because the effect of baseline differences between patients taking different drugs on the transcriptome is diluted by the large transcriptomic perturbation caused by taking an anti-TNF drug. Many of the same modules were also significant regardless of cell proportion correction.

There appears to be a general reduction in immune activation in responders at week 14, presumably due to successful inhibition of TNF, which is consistent with reduced neutrophil activation and reduced monocyte recruitment <https://www.ncbi.nlm.nih.gov/pmc/article>

s/PMC5820413/. Apoptosis of monocytes induced by anti-TNF in CD patients has also been previously observed [https://www.gastrojournal.org/article/S0016-5085\(01\)32104-2/fulltext](https://www.gastrojournal.org/article/S0016-5085(01)32104-2/fulltext). Certain B cell subsets are reduced in the blood IBD patients compared to controls <https://www.frontiersin.org/articles/10.3389/fimmu.2019.00361/>, so upregulation of B cell modules after treatment may represent a shift towards health. Another potential explanation would be increased immunogenicity due to higher drug levels in responders [233]. Although lack of between-drug heterogeneity does not support the greater immunogenicity of infliximab versus adalimumab. Overall, it is difficult to determine exact mechanisms in this observational study design, with bulk expression data, using such broad module definitions.

Some previously identified baseline gut markers of response that were not differentially expressed in blood at week 0, were differentially expressed at week 14. *S100A8* and *S100A9*, identified as markers by Arijs *et al.* [241], which encode components of the inflammatory marker CRP, were downregulated in week 14 responders. The cytokine *OSM*, which promotes inflammation in gut stromal cells [242], was similarly downregulated. Although it is pointless to use a week 14 marker to predict a response that is defined at week 14, this does demonstrate that gut markers can coincide with blood markers if expressed in immune cells present in both tissues.

When considering the interaction between change from week 0 to week 14 and response, the general pattern is magnification in responders, where the same expression changes occurring in both responders and non-responders tend to have greater magnitude in responders. A potential hypothesis is a continuum of response from non-response to response. Gaujoux *et al.* [244] found changes in cell proportions in response to anti-TNF treatment were magnified in responders, also supporting response as continuous phenotype. This study confirms a similar trend at the transcriptional level.

There were some rare exceptions to magnification for genes and modules in the type I interferon pathway. These showed upregulation in non-responders from week 0 to week 14, yet were either downregulated or not significantly different for responders. Single-gene examples include the interferon-induced guanylate-binding proteins *GBP2* and *GBP5* <https://rupress.org/jem/article/216/3/482/120360/Interferon-induced-guanylate-binding-proteins>, and *STAT2*, a key transcription factor for interferon-stimulated genes [160]. Genes such as *IFIT3* and *STAT2* are more strongly induced by type I interferons compared to type II <https://www.pnas.org/content/109/11/4239>. A study of RA, an IMID also treated with anti-TNF drugs, also found increases in type I interferon-regulated gene expression in blood after infliximab treatment associated with poor clinical response <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2875639/>.

A spline model of expression over all four timepoints confirmed the above observations made in week 0 and week 14 samples. Two main clusters of genes (clusters 1 and 5) contained mostly genes significantly associated with response in the two pairwise comparisons: within week 14, and change from week 0 to week 14. An example is the most significant single-gene association from the cluster 1 in spline model, *KREMEN1*, which was also among most significant associations in the pairwise comparisons. *KREMEN1* is part of an inflammatory apoptotic pathway in gut epithelium https://academic.oup.com/ibdjournal/article/14/suppl_1/S4/4653822, and is downregulated in responders post-induction. The trajectories of expression for genes in

move some of this to results section

these cluster 1 and 5 confirmed changes in expression post-induction were generally greater for responders. and in addition demonstrates that post-induction expression differences between responders and non-responders are sustained in samples taken around week 30 and week 54 during the anti-TNF maintenance period. In PANTS, “continuing standard dosing regimens after primary non-response was rarely helpful” for inducing remission by week 54 [233]. This phenomenon may have a transcriptomic basis.

Making use of data from later timepoints allowed more subtle effects to be detected in the spline analysis. Clusters 2 and 4 are enriched for B cell genes that were not significantly different at the single-gene level in the within week 0 comparison, although some downregulation of B cell and plasma cell modules were detected. Cluster 3 reproduced the observation that interferon-induced genes have opposing trajectories of expression in responders and non-responders. Expression of these genes was higher in responders at week 0 and lower at all post-treatment timepoints. Here, the cluster contains genes such as *STAT1*, *IRF1* and *TAP1* that are induced by both type I and type II interferons <https://www.pnas.org/content/109/11/4239>. I propose that blood expression of interferon-related genes is an attractive target for future studies of the biological basis of anti-TNF response, or prediction of primary response status. If one assumes the model where response and non-response lie along a phenotypic continuum, it is less likely that the opposing directions of regulation can be explained by unmodelled variables such as serum drug level being higher in responders. Since the difference is maintained until week 54, by which time patients would have received many doses of drug, it is also more likely that response is due to some biological property of an individual patient. Studies of anti-TNF response in RA patients have also found high baseline interferon activity in blood to be associated with good clinical response <https://academic.oup.com/rheumatology/article/54/1/188/1840617> <https://onlinelibrary.wiley.com/doi/full/10.1002/art.27226>.

It should be noted that the number of clusters is only the optimal number determined in this dataset, and does not imply that genes in different clusters represent biologically distinct pathways. Clusters 2 and 4 have similar trajectories and enrichments for B cell genes, and interferon pathway genes appear in both clusters 1 and 3.

Finally, I also attempted to determine if there were changes in genetic architecture of expression over time, which could indicate that expression response to anti-TNF has a genetic component. Out of all significant lead eQTLs for 11156 genes, only six reQTLs were detected with significantly different effect sizes at baseline versus one of the three post-treatment timepoints. Although no enrichment analyses are reported due to the small number of associations, *NMI* and *EPSTI1* are both interferon-induced genes with significant reQTLs that have their strongest effect size on expression at week 54. Given the issues with doing this reQTL analysis in bulk expression data are similar to those encountered in chapter 3, I did not place emphasis on interpreting these small numbers of associations. Before proceeding, I would also like to verify that these significant reQTLs are not artifacts from shrinkage of effect sizes in the joint eQTL model, as their posterior effect sizes from *mashr* were very different from the inputs from the timepoint-stratified models. If these hits are indeed reproducible by complementary methods such as allele-specific expression (ASE) [260], it may then be worth introducing genotype-response interaction terms in the eQTL models to prioritise eQTLs with differing ef-

not entirely sure this this is statistically rigorous due to third var effects?

fects in responders and non-responders. Given there is prior interest in the interferon pathway from DGE analyses, a more statistically powerful approach may be to generate a continuous interferon pathway score for each sample, which would then act as the interacting variable, similar to the approach of Davenport *et al.* [73].

Several threats to the validity of the study remain to be discussed. The most pressing may be the meaning of time in the study. For pairwise DGE comparisons, expression trajectory clustering, and reQTL mapping, samples were divided into four timepoints that corresponded to the major visits in PANTS. The DGE spline model was fit to study day directly. Study day has substantial variation around the target for later timepoints. The particular targets (weeks 14, 30, 54) for post-baseline samples were chosen so that patients on infliximab (8 weeks between doses) and adalimumab (2 weeks between doses) could both be sampled with the same visit structure. Drug levels peak sharply after each dose and decline exponentially over time. Visits were scheduled to be as close as possible (within a week) to the next scheduled drug dose to capture trough drug levels. Neither approach is perfect, as matching patients by timepoint and study day are only attempts to gather samples matched by trough drug level.

A further complication is the inclusion of LOR samples in analyses. One treatment option after LOR is dose escalation, which may raise trough drug levels for all subsequent visits for those patients. However, since the PANTS protocol allows for LOR visits that coincide with major visits to be labelled as a major visit, there is no guarantee that simply excluding samples labelled as LOR would resolve this.

The best solution may be to explicitly model measured serum drug levels as a covariate, where like cell proportions, it would likely act as a mediator of some associations with response. I did not do this as data missingness would reduce the sample size by about 40% in this study. Finding a suitable normalisation of drug level for use in pooled drug analyses is also challenging. Infliximab and adalimumab have differing pharmacokinetics: infliximab has higher peak concentrations, higher peak-trough ratios, and shorter half-life. The same serum concentrations of infliximab and adalimumab also have different biological effects due to differing therapeutic windows [228, 261] <https://onlinelibrary.wiley.com/doi/full/10.1111/apt.15643>.

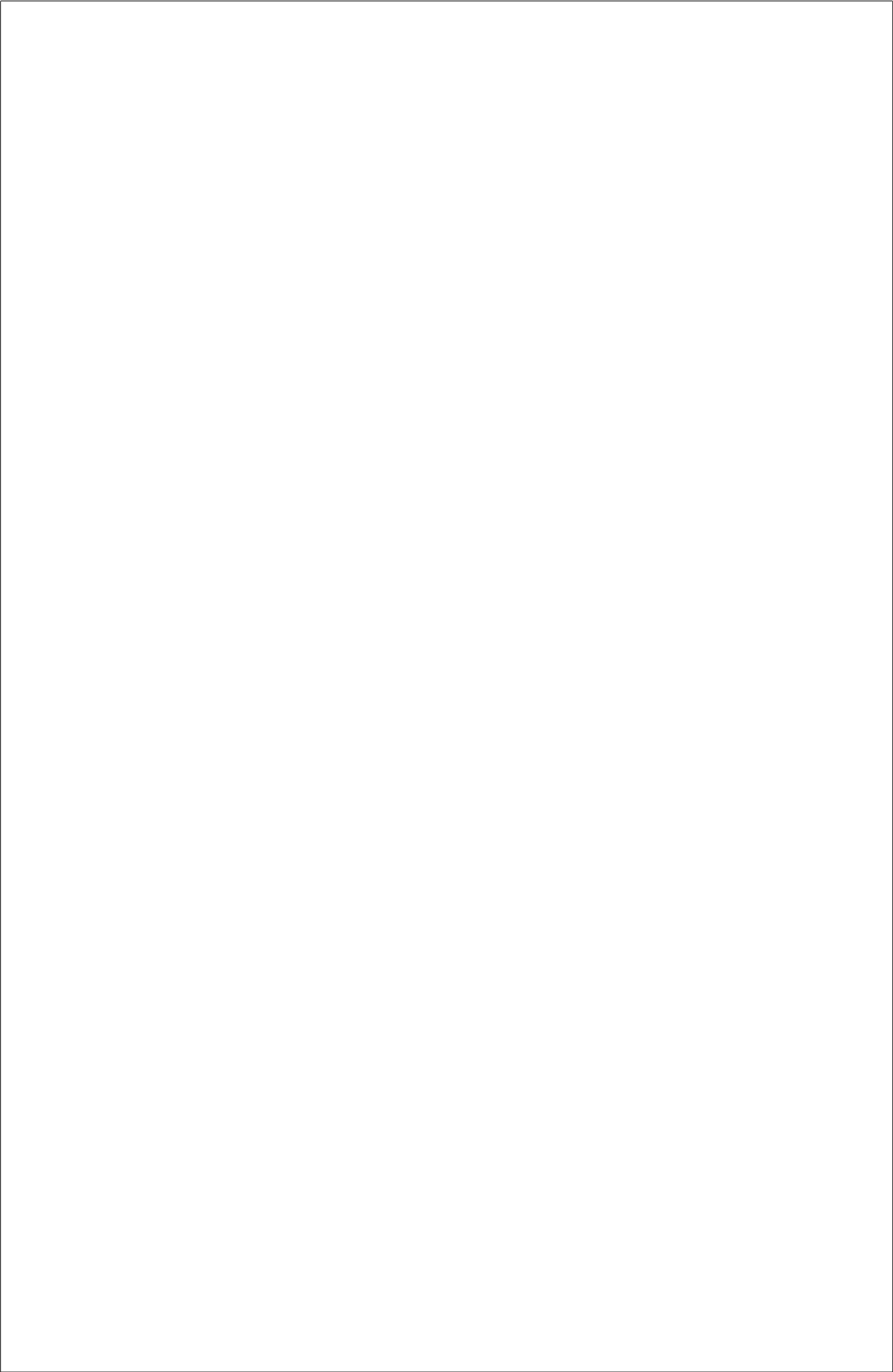
Finally, the effects of differential drop out in responders and non-responders has not been explored. The three main mechanisms of missing data are: missing completely at random (MCAR), probability of data being missing is independent of both observed and missing data; missing at random (MAR), probability of data being missing conditional on observed data is independent of missing data; and missing not at random (MNAR), probability of data being missing depends on missing data <https://doi.org/10.1016/B978-0-12-801342-7.00014-9>. Even conditional on response status, it is more likely that expression data from more extreme non-responders is missing for later timepoints, so the likely mechanism here is MNAR, and the linear mixed models used in this study may be biased. If it is indeed the most extreme non-responders dropping out, the estimation of responder versus non-responder effects may be conservative. Note there is no sidestepping a MNAR mechanism by analysing only the complete cases, since they will differ systematically from the sample as a whole <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3016756/>.

In conclusion, it remains unclear whether there are any robust single-gene markers for anti-

TNF response in baseline whole blood of CD patients. Baseline module associations were observed, but there was unexpected heterogeneity between infliximab and adalimumab patients, so it remains to be seen if such associations will be replicated in separate cohorts. Large upcoming datasets with drug response phenotypes such as the 1000IBD project [262] will be invaluable for attempted replication of the associations found in PANTS. Expression differences between responders and non-responders were more distinct at timepoints after the induction period. I found type I interferon genes went against the general trend of greater transcriptomic change from baseline in responders, being more upregulated post-treatment in non-responders. Given the type I interferon expression in blood has also been associated with anti-TNF response in RA patients, there may be an opportunity to consider the shared biology of anti-TNF response in IBD and RA. Much work has been done generating and validating signatures for anti-TNF response in RA <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0033199>; but not much work on validating RA signatures in IBD cohorts and vice versa.

This chapter has been purely descriptive. I have deliberately avoided the term “signature” in my own results, as I have not yet had a chance to assess their predictive capability. I also did not find strong evidence for strong reQTL effects over time in whole blood, so was unable to form hypotheses on the genetic mechanisms influencing anti-TNF response. However, the presence of eQTLs for most genes and the presence of strong differences in expression post-induction may allow testing for causal mechanisms where genotype affects drug response via expression. Strategies for moving onto both prediction and causal inference will be discussed in chapter 5.

| Table 4.1: Table caption. | | | | |
|---------------------------------------|-------------|-------------|-------------|-------------------|
| | ADA | IFX | pooled | p-value |
| Sex | | | | 0.317 |
| (Col %) | | | | Fisher exact |
| FEMALE | 78 (48.4%) | 89 (54.6%) | 167 (51.5%) | |
| MALE | 83 (51.6%) | 74 (45.4%) | 157 (48.5%) | |
| Age of onset (years) | | | | 0.774 |
| Mean (SD) | 33.3 (15.4) | 32.8 (15.3) | 33.1 (15.3) | Wilcoxon rank-sum |
| Missing | 0 | 0 | 0 | |
| Disease duration (years) | | | | 0.546 |
| Mean (SD) | 6.1 (8.1) | 5.9 (7.7) | 6.0 (7.9) | Wilcoxon rank-sum |
| Missing | 0 | 0 | 0 | |
| Smoking status | | | | 0.263 |
| (Col %) | | | | Fisher exact |
| Current | 28 (17.4%) | 36 (22.1%) | 64 (19.8%) | |
| Ex | 55 (34.2%) | 43 (26.4%) | 98 (30.2%) | |
| Never | 78 (48.4%) | 84 (51.5%) | 162 (50.0%) | |
| Crohn’s-related surgery | | | | 0.549 |
| (Col %) | | | | Fisher exact |
| FALSE | 114 (70.8%) | 110 (67.5%) | 224 (69.1%) | |
| TRUE | 47 (29.2%) | 53 (32.5%) | 100 (30.9%) | |
| On immunomodulator ever | | | | 0.543 |
| (Col %) | | | | Fisher exact |
| FALSE | 23 (14.3%) | 28 (17.2%) | 51 (15.7%) | |
| TRUE | 138 (85.7%) | 135 (82.8%) | 273 (84.3%) | |
| On immunomodulator at baseline | | | | 0.912 |
| (Col %) | | | | Fisher exact |
| FALSE | 79 (49.1%) | 81 (49.7%) | 160 (49.4%) | |
| TRUE | 82 (50.9%) | 82 (50.3%) | 164 (50.6%) | |
| On corticosteroids at baseline | | | | 0.011 |
| (Col %) | | | | Fisher exact |
| FALSE | 113 (70.2%) | 92 (56.4%) | 205 (63.3%) | |
| TRUE | 48 (29.8%) | 71 (43.6%) | 119 (36.7%) | |
| Baseline BMI | | | | 0.237 |
| Mean (SD) | 25.2 (6.2) | 24.3 (5.5) | 24.8 (5.9) | Wilcoxon rank-sum |
| Missing | 0 | 0 | 0 | |
| Primary response status | | | | 0.263 |
| (Col %) | | | | Fisher exact |
| Primary non-response | 76 (47.2%) | 66 (40.5%) | 142 (43.8%) | |
| Primary response | 85 (52.8%) | 97 (59.5%) | 182 (56.2%) | |
| CD8+ T cell (%) | | | | 0.380 |
| Mean (SD) | 2.8 (4.2) | 2.8 (5.2) | 2.8 (4.7) | Wilcoxon rank-sum |
| Missing | 38 | 18 | 56 | |
| CD4+ T cell (%s) | | | | 0.752 |
| Mean (SD) | 9.2 (6.3) | 9.2 (6.8) | 9.2 (6.5) | Wilcoxon rank-sum |
| Missing | 38 | 18 | 56 | |
| B cell (%s) | | | | 0.094 |
| Mean (SD) | 1.9 (2.0) | 1.5 (1.9) | 1.7 (1.9) | Wilcoxon rank-sum |
| Missing | 38 | 18 | 56 | |
| Monocyte (%s) | | | | 0.497 |
| Mean (SD) | 8.9 (3.5) | 9.2 (3.7) | 9.0 (3.6) | Wilcoxon rank-sum |
| Missing | 38 | 18 | 56 | |
| NK cell (%s) | | | | 0.683 |
| Mean (SD) | 1.9 (3.2) | 1.9 (3.8) | 1.9 (3.5) | Wilcoxon rank-sum |
| Missing | 38 | 18 | 56 | |
| Granulocyte (%s) | | | | 0.911 |
| Mean (SD) | 74.3 (9.7) | 74.3 (10.8) | 74.3 (10.3) | Wilcoxon rank-sum |
| Missing | 38 | 18 | 56 | |



Chapter 5

Discussion

- In this thesis, ... <a summary of the thesis projects>
 - Chapters 2 and 4 focus on describing the transcriptomic response of circulating immune cells to perturbation with uncontrolled longitudinal designs
 - * looking at how expression changes over time, and differences in transcriptomic response associated with phenotypic definitions of response
 - * associations are identified at a gene and gene set level
 - Chapter 3 (and the last part of chapter 4) focus on finding common genetic variants associated with transcriptomic response to perturbation
 - * reQTLs are identified at lead eQTL variants for each gene
 - * this is a first foray into defining causal molecular mechanisms that explain individual variation in response to these perturbations
 - Each chapter discusses the results and limitations of each study, but being very similar studies in design and analysis, they also suffer from a core set of limitations.
 - This chapter discusses these, and considerations for design and analysis of similar longitudinal studies in the future to better understand the biology of immune response to drugs and vaccines.

5.1 Increasing the power to detect associations

- It has been challenging to find robust single gene-level associations between expression and response, especially at baseline.
 - The biological signal may be small compared to other sources of variation: measurement platform (ch2); drug, response definition (ch4)
 - The obvious thing to do is increase sample size.
 - * Under a fixed resource constraints, longitudinal studies tradeoff more samples over time for fewer individuals
 - * Multiomic designs further tradeoff against richness/number of layers sampled

- there are distinct advantages to these designs
 - * e.g. repeated measures allows modelling of within patient covariance between timepoints with mixed model DGE (ch2 and ch4)
 - * e.g. spline analysis and clustering in ch4 using multiple timepoints to separate responders and nonresponders
 - * e.g. ability to test for association between layers (eQTLs), to explore molecular mechanisms
- but especially if the intent is to consider the small effect of genetic variants, a lesson learnt in the field of complex diseases from years of underpowered candidate gene studies, is that even a single complex phenotype requires large samples
- with rich datasets, be careful about subgroup analyses, subgrouping quickly attenuates your effective sample size
 - e.g. looking for timepoint, responder, cell type interactions
 - be wary of rules of thumb for sample size,
 - * e.g. in ch3, the intention was to map eQTLs at each timepoint, look for timepoint x G interaction, but it became apparent that cell proportion interactions were necessary for interpretation
 - prespecify interactions of interest at the power analysis stage in design of new experiments
 - for existing datasets, a workaround like the two stage strategy in ch3, assuming that interactions are only interesting at significant main effects
- Rather than increasing the sample size, can change the unit of analysis from single gene to multiple genes
 - use information from coordinated changes in multiple genes to detect subtle effects
 - use prior biological knowledge to define set of genes
 - this is the basis for gene set analyses used throughout this thesis
 - tradeoff between better interpretability but worse resolution
- <reproducibility of gene set vs single gene associations>
 - reports about any single gene associations e.g. SIGLEC10 in ch4 of this thesis, should be treated with caution pending replication
 - while the rank of a single gene may vary from sample to sample due to sampling noise, a summary measure computed from multiple genes should be more robust
 - it is hoped that the replicability of gene sets will be greater

5.2 Responder analysis

- <how well these models correspond to reality>
 - What does it mean to be a responder?
 - this has been discussed on this in the field of personalised medicine in the context of RCTs, but similar issues pertain to response in observational studies. [263]
 - Is being a responder is a stable personal characteristic of an individual over time, or would the same individual vary in response status over time?
 - Observing between individual variation in response for a single measure is not sufficient to imply response is personal.
 - e.g. a hypothetical agent with a 60% response rate in a group of individuals: “60% of the time, it works every time”
 - what does this mean? working 60% of the time on 100% of individuals, 100% of the time on 60% of individuals, or anything in between would be compatible
 - cannot distinguish with a single measure, yet the assumption is often made at the study design stage that it is the 2nd option
 - 1st option means no associations to detect
- Need to establish how consistent response is over time
 - to determine within-individual variation, needs replication at the level of the individual, i.e. need to perturb and measure the same person more than once [264]
 - * in ch4, repeated anti-TNF dosing showed expression differences between responders and NR, at least, were stable over time
 - * although probably partially due to selection of non-remission by w54
 - note for the specific case of genetic factors, the field of complex trait genetics has long had a solution for this: twin studies are analogous to measuring the same individual twice
 - * For phenotypes like vaccine response in ch2 that can be, twin studies have already demonstrated heritability of certain response phenos like influenza vaccine Ab titres [167]
 - * In pharmacogenetics, twin studies are less common <https://www.sciencedirect.com/science/article/abs/pii/S0928098719300326> and have not been done for response to anti-TNFs (the ch4 setting)
 - * repeated crossover trials can be a feasible alternative to twin studies https://www.researchgate.net/profile/Laszlo_Endrenyi/publication/13553486_Hypothesis_Comparisons_of_inter-_and_intra-individual_variations_can_substitute_for_twin_studies_in_drug_research/links/59e97c2e458515c36370e6a7/Hypothesis-Comparisons-of-inter-and-intra-individual-variations-can-substitute-for-twin-studies-in-drug-research.pdf [264]

might be worth subbing predictor for independent and response with dependent in all chapters, but then that gets confused with statistical independence...

- <How a responder to a perturbation was modelled in this thesis>
 - As discussed in subsection 2.2.2 in ch2, a binary definition of response based on dichotomisation is inefficient and assumes a sharp change in biological relevance at the threshold
 - In ch2, used titre response index (TRI), a continuous change score residualised on the baseline, which gets around the dichotomisation and dependence of change score on baseline
 - * two-stage approach here, so df is wrong
 - In multiPANTS, subsection 4.2.3 in ch4 sets out the binary definition of response, which is complex and based on many inputs
 - * having response dichotomies based on multiple inputs can lead to large discontinuities and non-monotonicity in response probabilities under small changes in input [112]
 - neither approach perfect, but the driving force was pragmatism
 - * for DGE, the model traditionally requires the response variable to be gene expression, and a predictor variable to be response status
 - * thus I needed an approach that generates a single predictor variable to represent response status
 - * In ch2, needed to combine pre-post titre data from two assays
 - * In ch4, the clinical algorithm already combines many measures into one.
 - having complicated response definitions can hinder replication by adding to interstudy variation e.g. the failure to replicate the TREM1 association in ch4
 - it may be better to directly with the constituent phenotypes: ch2 (log titre levels), ch4 (CRP level, HBI score), but then the fundamental modelling approach for both DGE/eQTL would need to change to not be restricted to a single response x variable
 - also applies to continuous measures of response, as what is being considered is that it only makes sense to talk about response if the response measure is tightly correlated within an individual over time

5.3 Challenges in the interpretation of bulk expression data

move and merge this paragraph into ch2 section on composition bias in RNAseq, it's out of place here

- NGS data is compositional, only quantifies relative abundance
 - raw abundance estimates are often called “counts”, but actually dimensionless proportions, interpretable only compared to other genes in the same sample (composition)
 - a fixed number of reads is sequenced: if a gene's counts go up, another will go down
 - naive normalisations such as CPM, TPM, RPKM are still rates (per transcriptome), and cannot be compared across samples, as they still depend on composition

- normalisation such as TMM tries to account for composition by normalising on non DE genes, to allow for valid DGE analysis
- getting absolute measures is difficult without spike-ins
- for valid DGE, need to decide on the units e.g. mRNA/cell? mRNA/transcriptome? [141]
- bulk data is a mixture of cell types with different expression patterns
 - one of the largest sources of variation in bulk expression data is cell proportions
 - the more cell type-specific a gene's expression, the more it is affected by cell proportions [265]
 - * perfectly cell type-specific genes are often called marker genes, and are used in deconvolution methods to estimate cell proportions from bulk data when they are not directly measured
 - both variation from in true variation in proportions, and sampling noise
 - in this thesis, estimated by:
 - * In HIRD ch 3, xcell deconvolution from expression data
 - * In PANTS ch 4, deconvolution from methylation data
 - these are fit as covariates in models
 - * could act as precision variables (for sampling noise in non cell type-specific genes),
 - * but also mediators for cell type-specific genes, where the interpretation of DGE effects changes (perturbation -> cell prop -> expression)
 - but coarse grained (just the main cell types) adjustment is not always sufficient, you'll still find associations that are proxies for rare cell types <https://www.biorxiv.org/content/10.1101/2020.05.28.120600v1>
 - * Recommend the use of e.g. PEER, as in ch3/4 eQTL analysis, some of those factors are likely to be rarer cell types
- <eQTL analysis>
 - analogous to DGE where independent variable of interest is now genotype
 - as discussed in subsection 3.2.11, it is model misspecification to not have the genotype x cell proportion interaction if the effect of genotype changes depending on cell proportions
 - * in fact, it is popular to use genotype x cell proportion (or a proxy of) interactions to call cell type-specific eQTLs e.g. [57, 60]
 - * but you cannot distinguish between correlated cell types this way
- <in vivo reQTL analysis>
 - here the cell props are causally affected by the perturbation (active recruitment, differentiation, proliferation of immune cells)

would an in vitro stimulation of a mixture of cells be immune to this? differentiation and proliferation could still occur

assumes eQTL is
finemapped

I cannot figure out what
sort of conditioning would
need to be done to block
this sort of effect even
if you had perfect cell
proportion measurements

- consider the contrived case, where vaccine perturbation causes active increase of a rare cell type that is near absent at baseline, but is a greatly increased proportion of the bulk mixture after perturbation.
- further assume this is true for all individuals (no variation in cell proportions).
- an eQTL specific to this cell type will show up as a reQTL in the post-perturbation timepoint because expression of that cell type contributes more to the bulk mixture.
 - * this will happen even if you adjust for proportion of the cell type, since there is no variation between individuals, it will only offset the regression line
 - * this does not require the gene to be upregulated on average, as the effect of interest is $G \rightarrow \Delta E$, not ΔE .
- if the aim of a reQTL study is to identify variants that causally affect transcriptomic response to perturbation, such that if we changed the genotype of an individual (observe the counterfactual), the transcriptomic response would be different.
- does this achieve that aim? yes, as the change in expression after perturbation differs when you change someone's genotype, as there is no difference between genotypes at baseline, but there is after perturbation.
- for genotype to explain more variance in bulk expression after perturbation, there are multiple compatible mechanisms
 - * a cell type with a cell type-specific eQTL increases in proportion (recruitment)
 - * a gene with an eQTL increases in expression (activation or recruitment depending on if you condition by cell proportions)
 - * the effect of a cell type-specific eQTL increases within that cell type (activation)
- one could imagine crude ways to falsify each hypothetical model e.g. interaction/coloc with each cell type (ch3), but always at risk of missing a cell type
- the challenge is distinguishing these scenarios in bulk in vivo data to form mechanistic hypotheses for expression response
- future studies should consider alternatives to bulk before they can fully take advantage of benefits of in vivo design
 - rather than conditioning on cell proportion, effectively looking at the subset of data with same cell proportions, just control them
 - If you want a mechanistic interpretation, need to control
 - single-cell RNAseq has the advantage of getting expression and composition simultaneously, and can be comparable in cost to bulk on FACS sorted cells <https://arxiv.org/ftp/arxiv/papers/1909/1909.12550.pdf>
 - do DGE/map eQTLs within each cell type
 - * combines advantage of in vivo stimulation with the controlled cell composition of in vitro studies

- can also be used to annotate bulk reQTLs such as those found in ch3 e.g. find the likely cell types for bulk reQTL effects by looking at eGene expression in single cell data [266]

5.4 From association to prediction

- I have now discussed both the independent and dependent variables used in descriptive models of expression vs response
 - so far, always put expression as dependent, and response status as independent, and estimated the effect size of response status on expression
 - did this even when response is measured after expression (baseline expression)
 - * this is the assumption that response is a stable individual characteristic
 - * also assumes it is measured without error, otherwise would stray into the realm of error in variables models
 - but might also want to consider if given expression, you could predict response (probability) as a dependent variable
 - this is also directly relevant to clinical goal of predicting patient response from baseline expression
- a large $p \gg n$ prediction problem
 - given the few gene-level associations with response is it futile to try?
 - Efron [267] provides an encouraging counterexample on $n=102$ (52 prostate cancer, 50 controls), $p=6033$ genes
 - * random forest to predict cancer status from gene expression: 2% test set error
 - * removed the 348 genes with positive importance scores, these are all the genes that were ever involved in a splitting rule
 - * rerun on resulting 102×5685 matrix, still a similar error rate
 - * repeat that process, removing 364 genes. rerun on resulting 102×5321 matrix, still a similar error rate
 - performance of such purely predictive models is dominated by combining many weak predictors
 - pure prediction is an easier problem than the traditional approach trying to attribute statistical significance to a few strong predictors
 - this may be why there is sometimes little overlap in the gene signatures from different studies on the same trait
- Need to consider what it takes for a gene expression prediction model to be clinically useful?
 - Useful to look at examples from cancer, one of the earliest fields where the approach generated interest

- Despite the early gene expression signatures being reported in the early 2000s, only a handful in use [268] <https://clincancerres.aacrjournals.org/content/21/21/4743>
- many hurdles: need to be accurate enough to base changes in therapy on, but incrementally more accurate than existing clinical markers, and cost effective. <https://doi.org/10.1093/annonc/mdw307>
- measuring a whole transcriptome is likely not cost effective. what are potential methods that might be?
- for predicting vaccine response from transcriptomic data, popular methods are:
 - * DAMIP e.g. [269], which gives rulesets composed of small sets of genes
 - * elastic net
 - * sparse partial least squares e.g. [108]
- Note the sparsity assumption for these methods (most genes are not predictive) is in opposition to the idea of combining many weak predictors to maximise predictive power, but will be less costly to translate to clinical practice

5.5 From association to causality

- uncovering “mechanism”
 - Knowing the molecular mechanisms of response are crucial for conceiving of possible interventions.
 - * e.g. ch4 baseline SIGLEC10 expression is associated with anti-TNF response, but would changing baseline SIGLEC10 affect response?
 - In observational data, a causally upstream anchor is required: genetic variation
 - the regression analyses in this thesis:
 - * DGE describes expression response association with phenotypic response (such as Ab titre)
 - * reQTLs describe genotype association with expression response
 - Need a model that encodes causal assumptions integrating multiple data layers: genotype -> gene expression -> phenotypic response
 - there are several families of complementary methods that integrate multiple data layers
- <Mendelian randomisation>
 - <explain the analogy to RCTs: segregation should be independent of env factors>
 - is a type of IV analysis that uses genetic instruments: can use eQTLs as instrumental variables (IV) to estimate causal effect between expression (exposure) and phenotype (outcome)

- three assumptions that define a valid IV [51, 270, 271]
 - * IV1 SNP is associated with expression i.e. an eQTL (no need to be fine mapped [49, 270])
 - * IV2 SNP is not associated with confounders of the expression-phenotype association
 - * IV3 SNP has no association with phenotype except through expression
- combined, these assumptions mean that expression is a complete mediator of the causal effect of the SNP on phenotype (a vertical pleiotropy is assumed)
- two-sample MR is where $G \rightarrow E$ and $E \rightarrow \text{pheno}$ are estimated in two non-overlapping samples [51, 271], and can come from large existing QTL datasets
 - * a related family of methods, TWAS <https://www.nature.com/articles/ng.3506>, tests association of genetically predicted expression to phenotypes, have their methodological basis in two-sample MR <https://link.springer.com/article/10.1007/s40484-020-0207-4>
- assumption violations will lead to bias in estimating causal effect of E on pheno
 - IV1 is satisfied by definition of an eQTL
 - IV2 violations are not likely, as genetic variants are unlikely to be associated with environmental confounders
 - IV3 violation is most likely
- <coloc>
 - IV3 violated by linkage: if the instrument eQTL is actually not causal, but simply correlated with another variant that causally affects phenotype
 - * use colocalisation methods to test whether the same variants are associated with E and phenotype [51]
- <mediation analysis>
 - IV3 violated by horizontal pleiotropy: instrument eQTL causes expression and phenotype independently (e.g. by also being an eQTL for another gene whose expression causally affects pheno), such that expression and phenotype are not causally related
 - use mediation analysis e.g. (CIT[211], findr[272]) that model comparison to distinguish between models of horizontal from vertical pleiotropy
 - * [211] used by Franco *et al.* [71] in a vaccine response reQTL setting, although they concluded they were underpowered
 - however, requires individual level data, and more susceptible to measurement error than MR [51]
- IV3 violated by reverse causation: assumed expression \rightarrow response, but the causal direction might actually be the reverse: $G \rightarrow \text{phenotype} \rightarrow \text{expression}$

citation needed,
try davey-smith2020MendelLawsMendelian

- if suspected, recommendations to distinguish [51, 270, 271] <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5036871/>
 - * perform MR in the reverse direction if there are instruments for the phenotype (bi-directional MR)
 - * extensions that test the directionality based on variance explained (MR Steiger) <https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1007081>
 - * incorporate biological knowledge, e.g. if eQTL variant is in a TF motif for a gene, more likely G → E than G → pheno → E

5.6 Triangulation

- Triangulation is the use of methods with different assumptions, biases, and limitations that address the same question [273]
 - discussed above is one example of triangulation, using complementary methods MR, colocalisation, mediation analysis to tease apart causal mechanisms from G → E → pheno
 - * e.g. Taylor *et al.* [274]: an example combining MR, coloc (HEIDI), MR Steiger, CIT, to disentangle causal effect of SNPs on DNA Me and expression
 - * [275] Predicted associations between proteins and phenotypes may indicate four explanations: causality, reverse causality, confounding by LD between the leading SNPs for proteins and phenotypes or horizontal pleiotropy (Supplementary Fig. 3). Given these alternative explanations, we conducted a set of sensitivity analyses to evaluate whether each MR association reflected a causal effect of protein on phenotype: tests of reverse causality using bidirectional MR²² and MR Steiger filtering;^{23,24} heterogeneity analyses for proteins with multiple instruments²⁵; and colocalization analyses²⁶ to investigate whether the genetic associations with both protein and phenotype shared the same causal variant (Fig. 1).
 - another example, how G → E effect changes in response to perturbation
 - * combine reQTL mapping (between individuals), dynamic ASE (within heterozygous individuals), (and maybe variance QTLs)
 - * dynamic ASE example for multiple T cell activation states <https://www.nature.com/articles/s41588-020-0579-4>
- another example, multiomics integration
 - different layers have non-redundant readouts
 - e.g. pQTLs
 - e.g. foreshadowing of eQTL by caQTLs
 - network approaches, MOFA
- many limitations discussed cannot be solved by increasing n

- triangulation will be key in moving from a descriptive to mechanistic understanding of immune response to perturbation

Bibliography

1. Polderman, T. J. C. *et al.* Meta-Analysis of the Heritability of Human Traits Based on Fifty Years of Twin Studies. *Nature Genetics* **47**, 702–709. doi:10.1038/ng.3285 (2015).
2. The ENCODE Project Consortium. An Integrated Encyclopedia of DNA Elements in the Human Genome. *Nature* **489**, 57–74. doi:10.1038/nature11247 (2012).
3. 1000 Genomes Project Consortium *et al.* A Global Reference for Human Genetic Variation. *Nature* **526**, 68–74. doi:10.1038/nature15393 (2015).
4. The International SNP Map Working Group. A Map of Human Genome Sequence Variation Containing 1.42 Million Single Nucleotide Polymorphisms. *Nature* **409**, 928–933. doi:10.1038/35057149 (2001).
5. Slatkin, M. Linkage Disequilibrium — Understanding the Evolutionary Past and Mapping the Medical Future. *Nature Reviews Genetics* **9**, 477–485. doi:10.1038/nrg2361 (2008).
6. Wall, J. D. & Pritchard, J. K. Haplotype Blocks and Linkage Disequilibrium in the Human Genome. *Nature Reviews Genetics* **4**, 587–597. doi:10.1038/nrg1123 (2003).
7. The International HapMap Consortium. A Second Generation Human Haplotype Map of over 3.1 Million SNPs. *Nature* **449**, 851–861. doi:10.1038/nature06258 (2007).
8. Karczewski, K. J. & Martin, A. R. Analytic and Translational Genetics. *Annual Review of Biomedical Data Science* **3**. doi:10.1146/annurev-biodatasci-072018-021148 (2020).
9. Visscher, P. M. & Goddard, M. E. From R.A. Fisher’s 1918 Paper to GWAS a Century Later. *Genetics* **211**, 1125–1130. doi:10.1534/genetics.118.301594 (2019).
10. Hindorff, L. A. *et al.* Potential Etiologic and Functional Implications of Genome-Wide Association Loci for Human Diseases and Traits. *Proceedings of the National Academy of Sciences* **106**, 9362–9367. doi:10.1073/pnas.0903103106 (2009).
11. Gibson, G. Rare and Common Variants: Twenty Arguments. *Nature reviews. Genetics* **13**, 135–145. doi:10.1038/nrg3118 (2011).
12. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169**, 1177–1186. doi:10.1016/j.cell.2017.05.038 (2017).
13. Altshuler, D., Daly, M. J. & Lander, E. S. Genetic Mapping in Human Disease. *Science* **322**, 881–8. doi:10.1126/science.1156409 (2008).
14. Ott, J., Kamatani, Y. & Lathrop, M. Family-Based Designs for Genome-Wide Association Studies. *Nature Reviews Genetics* **12**, 465–474. doi:10.1038/nrg2989 (2011).

15. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five Years of GWAS Discovery. *The American Journal of Human Genetics* **90**, 7–24. doi:10.1016/j.ajhg.2011.11.029 (2012).
16. Hirschhorn, J. N., Lohmueller, K., Byrne, E. & Hirschhorn, K. A Comprehensive Review of Genetic Association Studies. *Genetics in Medicine* **4**, 45–61. doi:10.1097/00125817-200203000-00002 (2002).
17. Border, R. *et al.* No Support for Historical Candidate Gene or Candidate Gene-by-Interaction Hypotheses for Major Depression Across Multiple Large Samples. *American Journal of Psychiatry* **176**, 376–387. doi:10.1176/appi.ajp.2018.18070881 (2019).
18. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *The American Journal of Human Genetics* **101**, 5–22. doi:10.1016/j.ajhg.2017.06.005 (2017).
19. Sinnott-Armstrong, N., Naqvi, S., Rivas, M. & Pritchard, J. K. GWAS of Three Molecular Traits Highlights Core Genes and Pathways alongside a Highly Polygenic Background. *bioRxiv*. doi:10.1101/2020.04.20.051631 (2020).
20. Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G. & Meyre, D. Benefits and Limitations of Genome-Wide Association Studies. *Nature Reviews Genetics*. doi:10.1038/s41576-019-0127-1 (2019).
21. Crouch, D. J. M. & Bodmer, W. F. Polygenic Inheritance, GWAS, Polygenic Risk Scores, and the Search for Functional Variants. *Proceedings of the National Academy of Sciences* **117**, 18924–18933. doi:10.1073/pnas.2005634117 (2020).
22. The International HapMap Consortium. A Haplotype Map of the Human Genome. *Nature* **437**, 1299–1320. doi:10.1038/nature04226 (2005).
23. Barrett, J. C. & Cardon, L. R. Evaluating Coverage of Genome-Wide Association Studies. *Nature Genetics* **38**, 659–662. doi:10.1038/ng1801 (2006).
24. Das, S., Abecasis, G. R. & Browning, B. L. Genotype Imputation from Large Reference Panels. *Annual Review of Genomics and Human Genetics* **19**, 73–96. doi:10.1146/annurev-genom-083117-021602 (2018).
25. Taliun, D. *et al.* Sequencing of 53,831 Diverse Genomes from the NHLBI TOPMed Program. *bioRxiv*. doi:10.1101/563866 (2019).
26. Pe'er, I., Yelensky, R., Altshuler, D. & Daly, M. J. Estimation of the Multiple Testing Burden for Genomewide Association Studies of Nearly All Common Variants. *Genetic Epidemiology* **32**, 381–385. doi:10.1002/gepi.20303 (2008).
27. Jannot, A.-S., Ehret, G. & Perneger, T. $P < 5 \times 10^{-8}$ Has Emerged as a Standard of Statistical Significance for Genome-Wide Association Studies. *Journal of Clinical Epidemiology* **68**, 460–465. doi:10.1016/j.jclinepi.2015.01.001 (2015).
28. Goeman, J. J. & Solari, A. Multiple Hypothesis Testing in Genomics. *Statistics in Medicine* **33**, 1946–1978. doi:10.1002/sim.6082 (2014).

29. Schaid, D. J., Chen, W. & Larson, N. B. From Genome-Wide Associations to Candidate Causal Variants by Statistical Fine-Mapping. *Nature Reviews Genetics* **19**, 491–504. doi:10.1038/s41576-018-0016-z (2018).
30. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A Simple New Approach to Variable Selection in Regression, with Application to Genetic Fine Mapping. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. doi:10.1111/rssb.12388 (2020).
31. Chong, J. X. *et al.* The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. *The American Journal of Human Genetics* **97**, 199–215. doi:10.1016/j.ajhg.2015.06.009 (2015).
32. Gallagher, M. D. & Chen-Plotkin, A. S. The Post-GWAS Era: From Association to Function. *The American Journal of Human Genetics* **102**, 717–730. doi:10.1016/j.ajhg.2018.04.002 (2018).
33. Brodie, A., Azaria, J. R. & Ofra, Y. How Far from the SNP May the Causative Genes Be? *Nucleic Acids Research* **44**, 6046–6054. doi:10.1093/nar/gkw500 (2016).
34. Stacey, D. *et al.* ProGeM: A Framework for the Prioritization of Candidate Causal Genes at Molecular Quantitative Trait Loci. *Nucleic Acids Research* **47**, e3–e3. doi:10.1093/nar/gky837 (2019).
35. Forgetta, V. *et al.* An Effector Index to Predict Causal Genes at GWAS Loci. *bioRxiv*. doi:10.1101/2020.06.28.171561 (2020).
36. Ghoussaini, M. *et al.* Open Targets Genetics: Systematic Identification of Trait-Associated Genes Using Large-Scale Genetics and Functional Genomics. *Nucleic Acids Research*. doi:10.1093/nar/gkaa840 (2020).
37. Cano-Gamez, E. & Trynka, G. From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. *Frontiers in Genetics* **11**. doi:10.3389/fgene.2020.00424 (2020).
38. Cramer, P. Organization and Regulation of Gene Transcription. *Nature* **573**, 45–54. doi:10.1038/s41586-019-1517-4 (2019).
39. Bannister, A. J. & Kouzarides, T. Regulation of Chromatin by Histone Modifications. *Cell Research* **21**, 381–395. doi:10.1038/cr.2011.22 (2011).
40. Schaub, M. A., Boyle, A. P., Kundaje, A., Batzoglou, S. & Snyder, M. Linking Disease Associations with Regulatory Information in the Human Genome. *Genome Research* **22**, 1748–1759. doi:10.1101/gr.136127.111 (2012).
41. Maurano, M. T. *et al.* Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science* **337**, 1190–1195. doi:10.1126/science.1222794 (2012).
42. Farh, K. K.-H. *et al.* Genetic and Epigenetic Fine Mapping of Causal Autoimmune Disease Variants. *Nature* **518**, 337–343. doi:10.1038/nature13835 (2015).
43. Trynka, G. *et al.* Disentangling the Effects of Colocalizing Genomic Annotations to Functionally Prioritize Non-Coding Variants within Complex-Trait Loci. *The American Journal of Human Genetics* **97**, 139–152. doi:10.1016/j.ajhg.2015.05.016 (2015).

44. Nasser, J. *et al.* Genome-Wide Maps of Enhancer Regulation Connect Risk Variants to Disease Genes. *bioRxiv*. doi:10.1101/2020.09.01.278093 (2020).
45. Gaffney, D. J. Global Properties and Functional Complexity of Human Gene Regulatory Variation. *PLoS Genetics* **9** (ed Abecasis, G. R.) e1003501. doi:10.1371/journal.pgen.1003501 (2013).
46. Albert, F. W. & Kruglyak, L. The Role of Regulatory Variation in Complex Traits and Disease. *Nature Reviews Genetics* **16**, 197–212. doi:10.1038/nrg3891 (2015).
47. Vandiedonck, C. Genetic Association of Molecular Traits: A Help to Identify Causative Variants in Complex Diseases. *Clinical Genetics*. doi:10.1111/cge.13187 (2017).
48. Vösa, U. *et al.* Unraveling the Polygenic Architecture of Complex Traits Using Blood eQTL Meta-Analysis. *bioRxiv*. doi:10.1101/447367 (2018).
49. Burgess, S., Foley, C. N. & Zuber, V. Inferring Causal Relationships Between Risk Factors and Outcomes from Genome-Wide Association Study Data. *Annual Review of Genomics and Human Genetics* **19**, 303–327. doi:10.1146/annurev-genom-083117-021731 (2018).
50. Wallace, C. Eliciting Priors and Relaxing the Single Causal Variant Assumption in Colocalisation Analyses. *PLOS Genetics* **16** (ed Epstein, M. P.) e1008720. doi:10.1371/journal.pgen.1008720 (2020).
51. Hemani, G., Bowden, J. & Davey Smith, G. Evaluating the Potential Role of Pleiotropy in Mendelian Randomization Studies. *Human Molecular Genetics* **27**, R195–R208. doi:10.1093/hmg/ddy163 (2018).
52. De Jager, P. L., Hacohen, N., Mathis, D., Regev, A., Stranger, B. E. & Benoist, C. ImmVar Project: Insights and Design Considerations for Future Studies of “Healthy” Immune Variation. *Seminars in Immunology* **27**, 51–57. doi:10.1016/j.smim.2015.03.003 (2015).
53. Nédélec, Y. *et al.* Genetic Ancestry and Natural Selection Drive Population Differences in Immune Responses to Pathogens. *Cell* **167**, 657–669.e21. doi:10.1016/j.cell.2016.09.025 (2016).
54. Quach, H. & Quintana-Murci, L. Living in an Adaptive World: Genomic Dissection of the Genus Homo and Its Immune Response. *Journal of Experimental Medicine* **214**, 877–894. doi:10.1084/jem.20161942 (2017).
55. Nica, A. C. *et al.* The Architecture of Gene Regulatory Variation across Multiple Human Tissues: The MuTHER Study. *PLoS Genetics* **7** (ed Barsh, G.) e1002003. doi:10.1371/journal.pgen.1002003 (2011).
56. Aguet, F. *et al.* Genetic Effects on Gene Expression across Human Tissues. *Nature* **550**, 204–213. doi:10.1038/nature24277 (2017).
57. Westra, H.-J. *et al.* Cell Specific eQTL Analysis without Sorting Cells. *PLOS Genetics* **11** (ed Pastinen, T.) e1005223. doi:10.1371/journal.pgen.1005223 (2015).
58. Zhernakova, D. V. *et al.* Identification of Context-Dependent Expression Quantitative Trait Loci in Whole Blood. *Nature Genetics* **49**, 139–145. doi:10.1038/ng.3737 (2017).

59. Glastonbury, C. A., Couto Alves, A., El-Sayed Moustafa, J. S. & Small, K. S. Cell-Type Heterogeneity in Adipose Tissue Is Associated with Complex Traits and Reveals Disease-Relevant Cell-Specific eQTLs. *The American Journal of Human Genetics* **104**, 1013–1024. doi:10.1016/j.ajhg.2019.03.025 (2019).
60. Kim-Hellmuth, S. *et al.* Cell Type-Specific Genetic Regulation of Gene Expression across Human Tissues. *Science* **369**, eaaz8528. doi:10.1126/science.aaz8528 (2020).
61. Dimas, A. S. *et al.* Common Regulatory Variation Impacts Gene Expression in a Cell Type-Dependent Manner. *Science* **325**, 1246–1250. doi:10.1126/science.1174148 (2009).
62. Peters, J. E. *et al.* Insight into Genotype-Phenotype Associations through eQTL Mapping in Multiple Cell Types in Health and Immune-Mediated Disease. *PLOS Genetics* **12** (ed Plagnol, V.) e1005908. doi:10.1371/journal.pgen.1005908 (2016).
63. Chen, L. *et al.* Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. *Cell* **167**, 1398–1414.e24. doi:10.1016/j.cell.2016.10.026 (2016).
64. Ackermann, M., Sikora-Wohlfeld, W. & Beyer, A. Impact of Natural Genetic Variation on Gene Expression Dynamics. *PLoS Genetics* **9** (ed Wells, C. A.) e1003514. doi:10.1371/journal.pgen.1003514 (2013).
65. Fu, J. *et al.* Unraveling the Regulatory Mechanisms Underlying Tissue-Dependent Genetic Variation of Gene Expression. *PLoS Genetics* **8** (ed Gibson, G.) e1002431. doi:10.1371/journal.pgen.1002431 (2012).
66. Rotival, M. Characterising the Genetic Basis of Immune Response Variation to Identify Causal Mechanisms Underlying Disease Susceptibility. *HLA* **94**, 275–284. doi:10.1111/tan.13598 (2019).
67. Huang, Q. *The Genetics of Gene Expression: From Simulations to the Early-Life Origins of Immune Diseases* (2019).
68. Barreiro, L. B., Tailleux, L., Pai, A. A., Gicquel, B., Marioni, J. C. & Gilad, Y. Deciphering the Genetic Architecture of Variation in the Immune Response to Mycobacterium Tuberculosis Infection. *Proceedings of the National Academy of Sciences* **109**, 1204–1209. doi:10.1073/pnas.1115761109 (2012).
69. Fairfax, B. P. *et al.* Innate Immune Activity Conditions the Effect of Regulatory Variants upon Monocyte Gene Expression. *Science* **343**, 1246949–1246949. doi:10.1126/science.1246949 (2014).
70. Alasoo, K., Rodrigues, J., Danesh, J., Freitag, D. F., Paul, D. S. & Gaffney, D. J. Genetic Effects on Promoter Usage Are Highly Context-Specific and Contribute to Complex Traits. *eLife* **8**. doi:10.7554/eLife.41673 (2019).
71. Franco, L. M. *et al.* Integrative Genomic Analysis of the Human Immune Response to Influenza Vaccination. *eLife* **2**, e00299. doi:10.7554/eLife.00299 (2013).

72. Lareau, C. A., White, B. C., Oberg, A. L., Kennedy, R. B., Poland, G. A. & McKinney, B. A. An Interaction Quantitative Trait Loci Tool Implicates Epistatic Functional Variants in an Apoptosis Pathway in Smallpox Vaccine eQTL Data. *Genes & Immunity* **17**, 244–250. doi:10.1038/gene.2016.15 (2016).
73. Davenport, E. E. *et al.* Discovering in Vivo Cytokine-eQTL Interactions from a Lupus Clinical Trial. *Genome Biology* **19**. doi:10.1186/s13059-018-1560-8 (2018).
74. Manry, J. *et al.* Deciphering the Genetic Control of Gene Expression Following Mycobacterium Leprae Antigen Stimulation. *PLOS Genetics* **13** (ed Sirugo, G.) e1006952. doi:10.1371/journal.pgen.1006952 (2017).
75. Kim-Hellmuth, S. *et al.* Genetic Regulatory Effects Modified by Immune Activation Contribute to Autoimmune Disease Associations. *Nature Communications* **8**. doi:10.1038/s41467-017-00366-1 (2017).
76. Tsang, J. S. Utilizing Population Variation, Vaccination, and Systems Biology to Study Human Immunology. *Trends in Immunology* **36**, 479–493. doi:10.1016/j.it.2015.06.005 (2015).
77. Villani, A.-C., Sarkizova, S. & Hacohen, N. Systems Immunology: Learning the Rules of the Immune System. *Annual Review of Immunology* **36**, 813–842. doi:10.1146/annurev-immunol-042617-053035 (2018).
78. Liston, A., Carr, E. J. & Linterman, M. A. Shaping Variation in the Human Immune System. *Trends in Immunology* **37**, 637–646. doi:10.1016/j.it.2016.08.002 (2016).
79. Brodin, P. & Davis, M. M. Human Immune System Variation. *Nature Reviews Immunology* **17**, 21–29. doi:10.1038/nri.2016.125 (2017).
80. Patin, E. *et al.* Natural Variation in the Parameters of Innate Immune Cells Is Preferentially Driven by Genetic Factors. *Nature Immunology*. doi:10.1038/s41590-018-0049-7 (2018).
81. Liston, A. & Goris, A. The Origins of Diversity in Human Immunity. *Nature Immunology* **19**, 209–210. doi:10.1038/s41590-018-0047-9 (2018).
82. Greenwood, B. The Contribution of Vaccination to Global Health: Past, Present and Future. *Philosophical Transactions of the Royal Society B: Biological Sciences* **369**, 20130433. doi:10.1098/rstb.2013.0433 (2014).
83. Linnik, J. E. & Egli, A. Impact of Host Genetic Polymorphisms on Vaccine Induced Antibody Response. *Human Vaccines & Immunotherapeutics* **12**, 907–915. doi:10.1080/21645515.2015.1119345 (2016).
84. O'Connor, D. & Pollard, A. J. Characterizing Vaccine Responses Using Host Genomic and Transcriptomic Analysis. *Clinical Infectious Diseases* **57**, 860–869. doi:10.1093/cid/cit373 (2013).
85. Mooney, M., McWeeney, S. & Sékaly, R.-P. Systems Immunogenetics of Vaccines. *Seminars in Immunology* **25**, 124–129. doi:10.1016/j.smim.2013.06.003 (2013).

86. Mentzer, A. J., O'Connor, D., Pollard, A. J. & Hill, A. V. S. Searching for the Human Genetic Factors Standing in the Way of Universally Effective Vaccines. *Philosophical Transactions of the Royal Society B: Biological Sciences* **370**, 20140341–20140341. doi:10.1098/rstb.2014.0341 (2015).
87. Scepanovic, P. *et al.* Human Genetic Variants and Age Are the Strongest Predictors of Humoral Immune Responses to Common Pathogens and Vaccines. *Genome Medicine* **10**. doi:10.1186/s13073-018-0568-8 (2018).
88. Dhakal, S. & Klein, S. L. Host Factors Impact Vaccine Efficacy: Implications for Seasonal and Universal Influenza Vaccine Programs. *Journal of Virology* **93** (ed Coyne, C. B.) doi:10.1128/JVI.00797-19 (2019).
89. Krammer, F. *et al.* Influenza. *Nature Reviews Disease Primers* **4**. doi:10.1038/s41572-018-0002-y (2018).
90. Houser, K. & Subbarao, K. Influenza Vaccines: Challenges and Solutions. *Cell Host & Microbe* **17**, 295–300. doi:10.1016/j.chom.2015.02.012 (2015).
91. Sautto, G. A., Kirchenbaum, G. A. & Ross, T. M. Towards a Universal Influenza Vaccine: Different Approaches for One Goal. *Virology Journal* **15**. doi:10.1186/s12985-017-0918-y (2018).
92. Broadbent, A. J. & Subbarao, K. Influenza Virus Vaccines: Lessons from the 2009 H1N1 Pandemic. *Current Opinion in Virology* **1**, 254–262. doi:10.1016/j.coviro.2011.08.002 (2011).
93. Klimov, A. *et al.* in *Influenza Virus* (eds Kawaoka, Y. & Neumann, G.) 25–51 (Humana Press, Totowa, NJ, 2012). doi:10.1007/978-1-61779-621-0_3.
94. Plotkin, S. A. Correlates of Protection Induced by Vaccination. *Clinical and Vaccine Immunology* **17**, 1055–1065. doi:10.1128/CVI.00131-10 (2010).
95. Cox, R. Correlates of Protection to Influenza Virus, Where Do We Go from Here? *Human Vaccines & Immunotherapeutics* **9**, 405–408. doi:10.4161/hv.22908 (2013).
96. Pulendran, B. Systems Vaccinology: Probing Humanity's Diverse Immune Systems with Vaccines. *Proceedings of the National Academy of Sciences* **111**, 12300–12306. doi:10.1073/pnas.1400476111 (2014).
97. Hagan, T., Nakaya, H. I., Subramaniam, S. & Pulendran, B. Systems Vaccinology: Enabling Rational Vaccine Design with Systems Biological Approaches. *Vaccine* **33**, 5294–5301. doi:10.1016/j.vaccine.2015.03.072 (2015).
98. Raeven, R. H. M., van Riet, E., Meiring, H. D., Metz, B. & Kersten, G. F. A. Systems Vaccinology and Big Data in the Vaccine Development Chain. *Immunology* **156**, 33–46. doi:10.1111/imm.13012 (2019).
99. Zhu, W. *et al.* A Whole Genome Transcriptional Analysis of the Early Immune Response Induced by Live Attenuated and Inactivated Influenza Vaccines in Young Children. *Vaccine* **28**, 2865–2876. doi:10.1016/j.vaccine.2010.01.060 (2010).

100. Bucasas, K. L. *et al.* Early Patterns of Gene Expression Correlate With the Humoral Immune Response to Influenza Vaccination in Humans. *The Journal of Infectious Diseases* **203**, 921–929. doi:10.1093/infdis/jiq156 (2011).
101. Nakaya, H. I. *et al.* Systems Biology of Vaccination for Seasonal Influenza in Humans. *Nature Immunology* **12**, 786–795. doi:10.1038/ni.2067 (2011).
102. Tan, Y., Tamayo, P., Nakaya, H., Pulendran, B., Mesirov, J. P. & Haining, W. N. Gene Signatures Related to B-Cell Proliferation Predict Influenza Vaccine-Induced Antibody Response. *European Journal of Immunology* **44**, 285–295. doi:10.1002/eji.201343657 (2014).
103. Nakaya, H. I., Li, S. & Pulendran, B. Systems Vaccinology: Learning to Compute the Behavior of Vaccine Induced Immunity. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* **4**, 193–205. doi:10.1002/wsbm.163 (2012).
104. Wilkins, A. L. *et al.* AS03- and MF59-Adjuvanted Influenza Vaccines in Children. *Frontiers in Immunology* **8**. doi:10.3389/fimmu.2017.01760 (2017).
105. Tregoning, J. S., Russell, R. F. & Kinnear, E. Adjuvanted Influenza Vaccines. *Human Vaccines & Immunotherapeutics* **14**, 550–564. doi:10.1080/21645515.2017.1415684 (2018).
106. Sobolev, O. *et al.* Adjuvanted Influenza-H1N1 Vaccination Reveals Lymphoid Signatures of Age-Dependent Early Responses and of Clinical Adverse Events. *Nature Immunology* **17**, 204–213. doi:10.1038/ni.3328 (2016).
107. Furman, D. *et al.* Apoptosis and Other Immune Biomarkers Predict Influenza Vaccine Responsiveness. *Molecular Systems Biology* **9**, 659. doi:10.1038/msb.2013.15 (2013).
108. Tsang, J. S. *et al.* Global Analyses of Human Immune Variation Reveal Baseline Predictors of Postvaccination Responses. *Cell* **157**, 499–513. doi:10.1016/j.cell.2014.03.031 (2014).
109. Nakaya, H. I. *et al.* Systems Analysis of Immunity to Influenza Vaccination across Multiple Years and in Diverse Populations Reveals Shared Molecular Signatures. *Immunity* **43**, 1186–1198. doi:10.1016/j.immuni.2015.11.012 (2015).
110. HIPC-CHI Signatures Project Team & HIPC-I Consortium. Multicohort Analysis Reveals Baseline Transcriptional Predictors of Influenza Vaccination Responses. *Science Immunology* **2**, eaal4656. doi:10.1126/sciimmunol.aal4656 (2017).
111. Cohen, J. The Cost of Dichotomization. *Applied Psychological Measurement* **7**, 249–253. doi:10.1177/014662168300700301 (1983).
112. Senn, S. *Dichotomania: An Obsessive Compulsive Disorder That Is Badly Affecting the Quality of Analysis of Pharmaceutical Trials* in *Proceedings of the International Statistical Institute* 55th World Statistics Congress (International Statistical Institute, Sydney, 2005), 14.
113. Fedorov, V., Mannino, F. & Zhang, R. Consequences of Dichotomization. *Pharmaceutical Statistics* **8**, 50–61. doi:10.1002/pst.331 (2009).

114. Food and Drug Administration. *Guidance for Industry: Clinical Data Needed to Support the Licensure of Pandemic Influenza Vaccines* (2007), 20.
115. Clifton, L. & Clifton, D. A. The Correlation between Baseline Score and Post-Intervention Score, and Its Implications for Statistical Analysis. *Trials* **20**. doi:10.1186/s13063-018-3108-3 (2019).
116. Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A. & Reich, D. Principal Components Analysis Corrects for Stratification in Genome-Wide Association Studies. *Nature Genetics* **38**, 904–909. doi:10.1038/ng1847 (2006).
117. Eu-ahsunthornwattana, J. *et al.* Comparison of Methods to Account for Relatedness in Genome-Wide Association Studies with Family-Based Data. *PLoS Genetics* **10** (ed Abecasis, G. R.) e1004445. doi:10.1371/journal.pgen.1004445 (2014).
118. Brown, B. C., Bray, N. L. & Pachter, L. Expression Reflects Population Structure. doi:10.1101/364448 (2018).
119. The International HapMap 3 Consortium. Integrating Common and Rare Genetic Variation in Diverse Human Populations. *Nature* **467**, 52–58. doi:10.1038/nature09298 (2010).
120. Okonechnikov, K., Conesa, A. & García-Alcalde, F. Qualimap 2: Advanced Multi-Sample Quality Control for High-Throughput Sequencing Data. *Bioinformatics* **32**, btv566. doi:10.1093/bioinformatics/btv566 (2015).
121. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: Summarize Analysis Results for Multiple Tools and Samples in a Single Report. *Bioinformatics* **32**, 3047–3048. doi:10.1093/bioinformatics/btw354 (2016).
122. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon Provides Fast and Bias-Aware Quantification of Transcript Expression. *Nature Methods* **14**, 417–419. doi:10.1038/nmeth.4197 (2017).
123. Liu, Y., Zhou, J. & White, K. P. RNA-Seq Differential Expression Studies: More Sequence or More Replication? *Bioinformatics* **30**, 301–304. doi:10.1093/bioinformatics/btt688 (2014).
124. Soneson, C., Love, M. I. & Robinson, M. D. Differential Analyses for RNA-Seq: Transcript-Level Estimates Improve Gene-Level Inferences. *F1000Research* **4**, 1521. doi:10.12688/f1000research.7563.2 (2016).
125. Zhao, S., Zhang, Y., Gamini, R., Zhang, B. & von Schack, D. Evaluation of Two Main RNA-Seq Approaches for Gene Quantification in Clinical RNA Sequencing: polyA+ Selection versus rRNA Depletion. *Scientific Reports* **8**. doi:10.1038/s41598-018-23226-4 (2018).
126. Min, J. L. *et al.* Variability of Gene Expression Profiles in Human Blood and Lymphoblastoid Cell Lines. *BMC Genomics* **11**, 96. doi:10.1186/1471-2164-11-96 (2010).

127. Chen, Y., Lun, A. T. L. & Smyth, G. K. From Reads to Genes to Pathways: Differential Expression Analysis of RNA-Seq Experiments Using Rsubread and the edgeR Quasi-Likelihood Pipeline. *F1000Research* **5**, 1438. doi:10.12688/f1000research.8987.2 (2016).
128. Huber, W., von Heydebreck, A., Sultmann, H., Poustka, A. & Vingron, M. Variance Stabilization Applied to Microarray Data Calibration and to the Quantification of Differential Expression. *Bioinformatics* **18**, S96–S104. doi:10.1093/bioinformatics/18.suppl_1.S96 (Suppl 1 2002).
129. Miller, J. A. *et al.* Strategies for Aggregating Gene Expression Data: The collapseRows R Function. *BMC Bioinformatics* **12**, 322. doi:10.1186/1471-2105-12-322 (2011).
130. Draghici, S., Khatri, P., Eklund, A. & Szallasi, Z. Reliability and Reproducibility Issues in DNA Microarray Measurements. *Trends in Genetics* **22**, 101–109. doi:10.1016/j.tig.2005.12.005 (2006).
131. Robinson, D. G., Wang, J. Y. & Storey, J. D. A Nested Parallel Experiment Demonstrates Differences in Intensity-Dependence between RNA-Seq and Microarrays. *Nucleic Acids Research*, gkv636. doi:10.1093/nar/gkv636 (2015).
132. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting Batch Effects in Microarray Expression Data Using Empirical Bayes Methods. *Biostatistics* **8**, 118–127. doi:10.1093/biostatistics/kxj037 (2007).
133. Chen, C. *et al.* Removing Batch Effects in Analysis of Expression Microarray Data: An Evaluation of Six Batch Adjustment Methods. *PLoS ONE* **6** (ed Kliebenstein, D.) e17238. doi:10.1371/journal.pone.0017238 (2011).
134. Espín-Pérez, A., Portier, C., Chadeau-Hyam, M., van Veldhoven, K., Kleijnans, J. C. S. & de Kok, T. M. C. M. Comparison of Statistical Methods and the Use of Quality Control Samples for Batch Effect Correction in Human Transcriptome Data. *PLOS ONE* **13** (ed Krishnan, V. V.) e0202947. doi:10.1371/journal.pone.0202947 (2018).
135. Zhang, Y., Jenkins, D. F., Manimaran, S. & Johnson, W. E. Alternative Empirical Bayes Models for Adjusting for Batch Effects in Genomic Studies. *BMC Bioinformatics* **19**. doi:10.1186/s12859-018-2263-6 (2018).
136. Nygaard, V., Rødland, E. A. & Hovig, E. Methods That Remove Batch Effects While Retaining Group Differences May Lead to Exaggerated Confidence in Downstream Analyses. *Biostatistics*, kxv027. doi:10.1093/biostatistics/kxv027 (January 2015).
137. Evans, C., Hardin, J. & Stoebe, D. M. Selecting Between-Sample RNA-Seq Normalization Methods from the Perspective of Their Assumptions. *Briefings in Bioinformatics* **19**, 776–792. doi:10.1093/bib/bbx008 (2018).
138. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data. *Bioinformatics* **26**, 139–140. doi:10.1093/bioinformatics/btp616 (2010).
139. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. Voom: Precision Weights Unlock Linear Model Analysis Tools for RNA-Seq Read Counts. *Genome Biology* **15**, 1–17 (2014).

140. Ritchie, M. E. *et al.* Limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies. *Nucleic Acids Research* **43**, e47–e47. doi:10.1093/nar/gkv007 (2015).
141. Soneson, C. & Delorenzi, M. A Comparison of Methods for Differential Expression Analysis of RNA-Seq Data. *BMC Bioinformatics* **14**. doi:10.1186/1471-2105-14-91 (2013).
142. Cohn, L. D. & Becker, B. J. How Meta-Analysis Increases Statistical Power. *Psychological Methods* **8**, 243–253. doi:10.1037/1082-989X.8.3.243 (2003).
143. Borenstein, M., Hedges, L. V., Higgins, J. P. & Rothstein, H. R. A Basic Introduction to Fixed-Effect and Random-Effects Models for Meta-Analysis. *Research Synthesis Methods* **1**, 97–111. doi:10.1002/jrsm.12 (2010).
144. Röver, C. *Bayesian Random-Effects Meta-Analysis Using the Bayesmeta R Package* <http://arxiv.org/abs/1711.08683> (2018).
145. Bender, R. *et al.* Methods for Evidence Synthesis in the Case of Very Few Studies. *Research Synthesis Methods*. doi:10.1002/jrsm.1297 (2018).
146. Gonnermann, A., Framke, T., Großhennig, A. & Koch, A. No Solution yet for Combining Two Independent Studies in the Presence of Heterogeneity. *Statistics in Medicine* **34**, 2476–2480. doi:10.1002/sim.6473 (2015).
147. Veroniki, A. A. *et al.* Methods to Estimate the Between-Study Variance and Its Uncertainty in Meta-Analysis. *Research Synthesis Methods* **7**, 55–79. doi:10.1002/jrsm.1164 (2016).
148. Chung, Y., Rabe-Hesketh, S., Dorie, V., Gelman, A. & Liu, J. A Nondegenerate Penalized Likelihood Estimator for Variance Parameters in Multilevel Models. *Psychometrika* **78**, 685–709. doi:10.1007/s11336-013-9328-2 (2013).
149. Friede, T., Röver, C., Wandel, S. & Neuenschwander, B. Meta-Analysis of Few Small Studies in Orphan Diseases. *Research Synthesis Methods* **8**, 79–91. doi:10.1002/jrsm.1217 (2017).
150. Seide, S. E., Röver, C. & Friede, T. Likelihood-Based Random-Effects Meta-Analysis with Few Studies: Empirical and Simulation Studies. *BMC Medical Research Methodology* **19**. doi:10.1186/s12874-018-0618-3 (2019).
151. Gelman, A. Prior Distributions for Variance Parameters in Hierarchical Models (Comment on Article by Browne and Draper). *Bayesian Analysis* **1**, 515–534. doi:10.1214/06-BA117A (2006).
152. Pullenayegum, E. M. An Informed Reference Prior for Between-Study Heterogeneity in Meta-Analyses of Binary Outcomes: Prior for between-Study Heterogeneity. *Statistics in Medicine* **30**, 3082–3094. doi:10.1002/sim.4326 (2011).
153. Turner, R. M., Jackson, D., Wei, Y., Thompson, S. G. & Higgins, J. P. T. Predictive Distributions for Between-Study Heterogeneity and Simple Methods for Their Application in Bayesian Meta-Analysis. *Statistics in Medicine* **34**, 984–998. doi:10.1002/sim.6381 (2015).

154. Higgins, J. P. T. & Whitehead, A. Borrowing Strength from External Trials in a Meta-Analysis. *Statistics in Medicine* **15**, 2733–2749. doi:10.1002/(SICI)1097-0258(19961230)15:24<2733::AID-SIM562>3.0.CO;2-0 (1996).
155. Zhu, A., Ibrahim, J. G. & Love, M. I. Heavy-Tailed Prior Distributions for Sequence Count Data: Removing the Noise and Preserving Large Differences. *Bioinformatics* **35** (ed Stegle, O.) 2084–2092. doi:10.1093/bioinformatics/bty895 (2019).
156. Stephens, M. False Discovery Rates: A New Deal. *Biostatistics*, kxw041. doi:10.1093/biostatistics/kxw041 (2016).
157. Weiner 3rd, J. & Domaszewska, T. Tmod: An R Package for General and Multivariate Enrichment Analysis. doi:10.7287/peerj.preprints.2420v1 (2016).
158. Li, S. *et al.* Molecular Signatures of Antibody Responses Derived from a Systems Biology Study of Five Human Vaccines. *Nature Immunology* **15**, 195–204. doi:10.1038/ni.2789 (2013).
159. Bin, L., Li, X., Feng, J., Richers, B. & Leung, D. Y. M. Ankyrin Repeat Domain 22 Mediates Host Defense Against Viral Infection Through STING Signaling Pathway. *The Journal of Immunology* **196**, 201.4 LP –201.4 (1 Supplement 2016).
160. Schneider, W. M., Chevillotte, M. D. & Rice, C. M. Interferon-Stimulated Genes: A Complex Web of Host Defenses. *Annual Review of Immunology* **32**, 513–545. doi:10.1146/annurev-immunol-032713-120231 (2014).
161. Nakaya, H. I. *et al.* Systems Biology of Immunity to MF59-Adjuvanted versus Nonadjuvanted Trivalent Seasonal Influenza Vaccines in Early Childhood. *Proceedings of the National Academy of Sciences* **113**, 1853–1858. doi:10.1073/pnas.1519690113 (2016).
162. Murphy, K. & Weaver, C. *Janeway's Immunobiology* 9th edition. 904 pp. (Garland Science/Taylor & Francis Group, LLC, New York, NY, 2016).
163. Nauta, J. J., Beyer, W. E. & Osterhaus, A. D. On the Relationship between Mean Antibody Level, Seroprotection and Clinical Protection from Influenza. *Biologicals* **37**, 216–221. doi:10.1016/j.biologicals.2009.02.002 (2009).
164. Poland, G. A., Ovsyannikova, I. G. & Jacobson, R. M. Immunogenetics of Seasonal Influenza Vaccine Response. *Vaccine* **26**, D35–D40. doi:10.1016/j.vaccine.2008.07.065 (2008).
165. Avnir, Y. *et al.* IGHV1-69 Polymorphism Modulates Anti-Influenza Antibody Repertoires, Correlates with IGHV Utilization Shifts and Varies by Ethnicity. *Scientific Reports* **6**, 20842. doi:10.1038/srep20842 (2016).
166. Moss, A. J. *et al.* Correlation between Human Leukocyte Antigen Class II Alleles and HAI Titers Detected Post-Influenza Vaccination. *PLoS ONE* **8** (ed Sambhara, S.) e71376. doi:10.1371/journal.pone.0071376 (2013).
167. Brodin, P. *et al.* Variation in the Human Immune System Is Largely Driven by Non-Heritable Influences. *Cell* **160**, 37–47. doi:10.1016/j.cell.2014.12.020 (2015).

168. Maranville, J. C. *et al.* Interactions between Glucocorticoid Treatment and Cis-Regulatory Polymorphisms Contribute to Cellular Response Phenotypes. *PLoS Genetics* **7** (ed Gibson, G.) e1002162. doi:10.1371/journal.pgen.1002162 (2011).
169. Shpak, M. *et al.* An eQTL Analysis of the Human Glioblastoma Multiforme Genome. *Genomics* **103**, 252–263. doi:10.1016/j.ygeno.2014.02.005 (2014).
170. Allison, P. D. Change Scores as Dependent Variables in Regression Analysis. *Sociological Methodology* **20**, 93. doi:10.2307/271083 (1990).
171. Clogg, C. C., Petkova, E. & Haritou, A. Statistical Methods for Comparing Regression Coefficients Between Models. *The American Journal of Sociology* **100**, 1261–1293 (1995).
172. Flutre, T., Wen, X., Pritchard, J. & Stephens, M. A Statistical Framework for Joint eQTL Analysis in Multiple Tissues. *PLOS Genet* **9**, e1003486. doi:10.1371/journal.pgen.1003486 (2013).
173. Urbut, S. M., Wang, G., Carbonetto, P. & Stephens, M. Flexible Statistical Methods for Estimating and Testing Effects in Genomic Studies with Multiple Conditions. *Nature Genetics*. doi:10.1038/s41588-018-0268-8 (2018).
174. Li, G., Jima, D., Wright, F. A. & Nobel, A. B. HT-eQTL: Integrative Expression Quantitative Trait Loci Analysis in a Large Number of Human Tissues. *BMC Bioinformatics* **19**. doi:10.1186/s12859-018-2088-3 (2018).
175. Stephens, M. A Unified Framework for Association Analysis with Multiple Related Phenotypes. *PLoS ONE* **8** (ed Emmert-Streib, F.) e65245. doi:10.1371/journal.pone.0065245 (2013).
176. Kim, S. *et al.* Characterizing the Genetic Basis of Innate Immune Response in TLR4-Activated Human Monocytes. *Nature Communications* **5**. doi:10.1038/ncomms6236 (2014).
177. Sul, J. H., Han, B., Ye, C., Choi, T. & Eskin, E. Effectively Identifying eQTLs from Multiple Tissues by Combining Mixed Model and Meta-Analytic Approaches. *PLoS Genetics* **9** (ed Schork, N. J.) e1003491. doi:10.1371/journal.pgen.1003491 (2013).
178. Duong, D. *et al.* Applying Meta-Analysis to Genotype-Tissue Expression Data from Multiple Tissues to Identify eQTLs and Increase the Number of eGenes. *Bioinformatics* **33**, i67–i74. doi:10.1093/bioinformatics/btx227 (2017).
179. Price, A. L., Zaitlen, N. A., Reich, D. & Patterson, N. New Approaches to Population Stratification in Genome-Wide Association Studies. *Nature Reviews Genetics* **11**, 459–463. doi:10.1038/nrg2813 (2010).
180. Golan, D., Rosset, S. & Lin, D.-Y. in Borgan, Ø., Breslow, N. E., Chatterjee, N., Gail, M. H., Scott, A. & Wild, C. J. *Handbook of Statistical Methods for Case-Control Studies* (eds Borgan, Ø., Breslow, N., Chatterjee, N., Gail, M. H., Scott, A. & Wild, C. J.) 1st ed., 495–514 (Chapman and Hall/CRC, 2018). doi:10.1201/9781315154084-27.
181. Astle, W. & Balding, D. J. Population Structure and Cryptic Relatedness in Genetic Association Studies. *Statistical Science* **24**, 451–471. doi:10.1214/09-STS307 (2009).

182. Listgarten, J., Lippert, C., Kadie, C. M., Davidson, R. I., Eskin, E. & Heckerman, D. Improved Linear Mixed Models for Genome-Wide Association Studies. *Nature Methods* **9**, 525–526. doi:10.1038/nmeth.2037 (2012).
183. Lippert, C., Listgarten, J., Liu, Y., Kadie, C. M., Davidson, R. I. & Heckerman, D. FaST Linear Mixed Models for Genome-Wide Association Studies. *Nature Methods* **8**, 833–835. doi:10.1038/nmeth.1681 (2011).
184. Speed, D., Hemani, G., Johnson, M. R. & Balding, D. J. Improved Heritability Estimation from Genome-Wide SNPs. *The American Journal of Human Genetics* **91**, 1011–1021. doi:10.1016/j.ajhg.2012.10.010 (2012).
185. Aguet, F. *et al.* The GTEx Consortium Atlas of Genetic Regulatory Effects across Human Tissues. *bioRxiv*. doi:10.1101/787903 (2019).
186. Beasley, T. M., Erickson, S. & Allison, D. B. Rank-Based Inverse Normal Transformations Are Increasingly Used, But Are They Merited? *Behavior Genetics* **39**, 580–595. doi:10.1007/s10519-009-9281-0 (2009).
187. Qi, T. *et al.* Identifying Gene Targets for Brain-Related Traits Using Transcriptomic and Methyloomic Data from Blood. *Nature Communications* **9**. doi:10.1038/s41467-018-04558-1 (2018).
188. Aran, D., Hu, Z. & Butte, A. J. xCell: Digitally Portraying the Tissue Cellular Heterogeneity Landscape. *Genome Biology* **18**. doi:10.1186/s13059-017-1349-1 (2017).
189. Kleiveland, C. R. in *The Impact of Food Bioactives on Health* (eds Verhoeckx, K. *et al.*) 161–167 (Springer International Publishing, Cham, 2015). doi:10.1007/978-3-319-16104-4_15.
190. Van der Wijst, M. G. P. *et al.* Single-Cell RNA Sequencing Identifies Celltype-Specific Cis-eQTLs and Co-Expression QTLs. *Nature Genetics* **50**, 493–497. doi:10.1038/s41588-018-0089-9 (2018).
191. Maddala, G. S. *Introduction to Econometrics* 2nd ed. 631 pp. (Macmillan Pub. Co. ; Maxwell Macmillan Canada ; Maxwell Macmillan International, New York : Toronto : New York, 1992).
192. Kanyongo, G. Y. The Influence of Reliability on Four Rules for Determining the Number of Components to Retain. *Journal of Modern Applied Statistical Methods* **5**, 332–343. doi:10.22237/jmasm/1162353960 (2005).
193. Astle, W. J. *et al.* The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell* **167**, 1415–1429.e19. doi:10.1016/j.cell.2016.10.042 (2016).
194. Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using Probabilistic Estimation of Expression Residuals (PEER) to Obtain Increased Power and Interpretability of Gene Expression Analyses. *Nature protocols* **7**, 500–507. doi:10.1038/nprot.2011.457 (2012).
195. Lippert, C., Casale, F. P., Rakitsch, B. & Stegle, O. LIMIX: Genetic Analysis of Multiple Traits. doi:10.1101/003905 (2014).

196. Sul, J. H., Martin, L. S. & Eskin, E. Population Structure in Genetic Studies: Confounding Factors and Mixed Models. *PLOS Genetics* **14** (ed Barsh, G. S.) e1007309. doi:10.1371/journal.pgen.1007309 (2018).
197. Huang, Q. Q., Ritchie, S. C., Brozynska, M. & Inouye, M. Power, False Discovery Rate and Winner's Curse in eQTL Studies. *Nucleic Acids Research* **46**, e133–e133. doi:10.1093/nar/gky780 (2018).
198. Schenker, N. & Gentleman, J. F. On Judging the Significance of Differences by Examining the Overlap Between Confidence Intervals. *The American Statistician* **55**, 182–186 (2001).
199. Gelman, A. & Stern, H. The Difference Between “Significant” and “Not Significant” Is Not Itself Statistically Significant. *The American Statistician* **60**, 328–331. doi:10.1198/000313006X152649 (2006).
200. Kooperberg, C. & LeBlanc, M. Increasing the Power of Identifying Gene \times Gene Interactions in Genome-Wide Association Studies. *Genetic Epidemiology* **32**, 255–263. doi:10.1002/gepi.20300 (2008).
201. Zeng, B. *et al.* Comprehensive Multiple eQTL Detection and Its Application to GWAS Interpretation. *Genetics* **212**, 905–918. doi:10.1534/genetics.119.302091 (2019).
202. Dobbyn, A. *et al.* Landscape of Conditional eQTL in Dorsolateral Prefrontal Cortex and Co-Localization with Schizophrenia GWAS. *The American Journal of Human Genetics* **102**, 1169–1184. doi:10.1016/j.ajhg.2018.04.011 (2018).
203. Mizuno, A. & Okada, Y. Biological Characterization of Expression Quantitative Trait Loci (eQTLs) Showing Tissue-Specific Opposite Directional Effects. *European Journal of Human Genetics* **27**, 1745–1756. doi:10.1038/s41431-019-0468-4 (2019).
204. Kumasaka, N., Knights, A. J. & Gaffney, D. J. Fine-Mapping Cellular QTLs with RASQUAL and ATAC-Seq. *Nature Genetics* **48**, 206–213. doi:10.1038/ng.3467 (2016).
205. Wu, L., Shen, C., Seed Ahmed, M., Östenson, C.-G. & Gu, H. F. Adenylate Cyclase 3: A New Target for Anti-Obesity Drug Development: ADCY3 and Anti-Obesity. *Obesity Reviews* **17**, 907–914. doi:10.1111/obr.12430 (2016).
206. McGovern, D. P., Kugathasan, S. & Cho, J. H. Genetics of Inflammatory Bowel Diseases. *Gastroenterology* **149**, 1163–1176.e2. doi:10.1053/j.gastro.2015.08.001 (2015).
207. Çalışkan, M., Baker, S. W., Gilad, Y. & Ober, C. Host Genetic Variation Influences Gene Expression Response to Rhinovirus Infection. *PLOS Genetics* **11** (ed Gibson, G.) e1005111. doi:10.1371/journal.pgen.1005111 (2015).
208. Pai, A. A., Pritchard, J. K. & Gilad, Y. The Genetic and Mechanistic Basis for Variation in Gene Regulation. *PLoS Genetics* **11** (ed Lappalainen, T.) e1004857. doi:10.1371/journal.pgen.1004857 (2015).
209. Choudhury, M. & Ramsey, S. A. Identifying Cell Type-Specific Transcription Factors by Integrating ChIP-Seq and eQTL Data-Application to Monocyte Gene Regulation. *Gene Regulation and Systems Biology* **10**, GRSB.S40768. doi:10.4137/GRSB.S40768 (2016).

210. Langford, E., Schwertman, N. & Owens, M. Is the Property of Being Positively Correlated Transitive? *The American Statistician* **55**, 322–325 (2001).
211. Millstein, J., Zhang, B., Zhu, J. & Schadt, E. E. Disentangling Molecular Relationships with a Causal Inference Test. *BMC Genetics* **10**. doi:10.1186/1471-2156-10-23 (2009).
212. Roda, G. *et al.* Crohn's Disease. *Nature Reviews Disease Primers* **6**. doi:10.1038/s41572-020-0156-2 (2020).
213. Cotsapas, C. & Hafler, D. A. Immune-Mediated Disease Genetics: The Shared Basis of Pathogenesis. *Trends in Immunology* **34**, 22–26. doi:10.1016/j.it.2012.09.001 (2013).
214. David, T., Ling, S. F. & Barton, A. Genetics of Immune-Mediated Inflammatory Diseases. *Clinical & Experimental Immunology* **193**, 3–12. doi:10.1111/cei.13101 (2018).
215. Ananthakrishnan, A. N. Epidemiology and Risk Factors for IBD. *Nature Reviews Gastroenterology & Hepatology* **12**, 205–217. doi:10.1038/nrgastro.2015.34 (2015).
216. De Souza, H. S. P. & Fiocchi, C. Immunopathogenesis of IBD: Current State of the Art. *Nature Reviews Gastroenterology & Hepatology* **13**, 13–27. doi:10.1038/nrgastro.2015.186 (2016).
217. Todd, J. A. Tackling Common Disease. *Nature* **411**, 537–539. doi:10.1038/35079223 (2001).
218. De Lange, K. M. *et al.* Genome-Wide Association Study Implicates Immune Activation of Multiple Integrin Genes in Inflammatory Bowel Disease. *Nature Genetics* **49**, 256–261. doi:10.1038/ng.3760 (2017).
219. Jostins, L. *et al.* Host–Microbe Interactions Have Shaped the Genetic Architecture of Inflammatory Bowel Disease. *Nature* **491**, 119–24. doi:10.1038/nature11582 (2012).
220. Liu, J. Z. *et al.* Association Analyses Identify 38 Susceptibility Loci for Inflammatory Bowel Disease and Highlight Shared Genetic Risk across Populations. *Nature Genetics* **47**, 979–986. doi:10.1038/ng.3359 (2015).
221. Kaplan, G. G. The Global Burden of IBD: From 2015 to 2025. *Nature Reviews Gastroenterology & Hepatology* **12**, 720–727. doi:10.1038/nrgastro.2015.150 (2015).
222. Alatab, S. *et al.* The Global, Regional, and National Burden of Inflammatory Bowel Disease in 195 Countries and Territories, 1990–2017: A Systematic Analysis for the Global Burden of Disease Study 2017. *The Lancet Gastroenterology & Hepatology* **5**, 17–30. doi:10.1016/S2468-1253(19)30333-4 (2020).
223. Levin, A. D., Wildenberg, M. E. & van den Brink, G. R. Mechanism of Action of Anti-TNF Therapy in Inflammatory Bowel Disease. *Journal of Crohn's and Colitis* **10**, 989–997. doi:10.1093/ecco-jcc/jjw053 (2016).
224. Aggarwal, B. B. Signalling Pathways of the TNF Superfamily: A Double-Edged Sword. *Nature Reviews Immunology* **3**, 745–756. doi:10.1038/nri1184 (2003).
225. Kalliolias, G. D. & Ivashkiv, L. B. TNF Biology, Pathogenic Mechanisms and Emerging Therapeutic Strategies. *Nature Reviews Rheumatology* **12**, 49–62. doi:10.1038/nrrheum.2015.169 (2016).

226. Digby-Bell, J. L., Atreya, R., Monteleone, G. & Powell, N. Interrogating Host Immunity to Predict Treatment Response in Inflammatory Bowel Disease. *Nature Reviews Gastroenterology & Hepatology*. doi:10.1038/s41575-019-0228-5 (2019).
227. Adegbola, S. O., Sahnan, K., Warusavitarne, J., Hart, A. & Tozer, P. Anti-TNF Therapy in Crohn's Disease. *International Journal of Molecular Sciences* **19**, 2244. doi:10.3390/ijms19082244 (2018).
228. Lichtenstein, G. R. Comprehensive Review: Antitumor Necrosis Factor Agents in Inflammatory Bowel Disease and Factors Implicated in Treatment Response. *Therapeutic Advances in Gastroenterology* **6**, 269–293. doi:10.1177/1756283X13479826 (2013).
229. Aitken, M., Kleinrock, M., Simorellis, A. & Nass, D. *The Global Use of Medicine in 2019 and Outlook to 2023: Forecasts and Areas to Watch* (IQVIA Institute, NC, USA, 2019).
230. Roda, G., Jharap, B., Neeraj, N. & Colombel, J.-F. Loss of Response to Anti-TNFs: Definition, Epidemiology, and Management: *Clinical and Translational Gastroenterology* **7**, e135. doi:10.1038/ctg.2015.63 (2016).
231. Ben-Horin, S., Kopylov, U. & Chowers, Y. Optimizing Anti-TNF Treatments in Inflammatory Bowel Disease. *Autoimmunity Reviews* **13**, 24–30. doi:10.1016/j.autrev.2013.06.002 (2014).
232. Flamant, M. & Roblin, X. Inflammatory Bowel Disease: Towards a Personalized Medicine. *Therapeutic Advances in Gastroenterology* **11**, 1756283X1774502. doi:10.1177/1756283X17745029 (2018).
233. Kennedy, N. A. *et al.* Predictors of Anti-TNF Treatment Failure in Anti-TNF-Naive Patients with Active Luminal Crohn's Disease: A Prospective, Multicentre, Cohort Study. *The Lancet Gastroenterology & Hepatology* **4**, 341–353. doi:10.1016/S2468-1253(19)30012-3 (2019).
234. Vermeire, S., Gils, A., Accossato, P., Lula, S. & Marren, A. Immunogenicity of Biologics in Inflammatory Bowel Disease. *Therapeutic Advances in Gastroenterology* **11**, 1756283X1775035. doi:10.1177/1756283X17750355 (2018).
235. Rogler, G. Where Are We Heading to in Pharmacological IBD Therapy? *Pharmacological Research* **100**, 220–227. doi:10.1016/j.phrs.2015.07.005 (2015).
236. D'Haens, G. R. *et al.* The London Position Statement of the World Congress of Gastroenterology on Biological Therapy for IBD With the European Crohn's and Colitis Organization: When to Start, When to Stop, Which Drug to Choose, and How to Predict Response?: *American Journal of Gastroenterology* **106**, 199–212. doi:10.1038/ajg.2010.392 (2011).
237. Ding, N. S., Hart, A. & De Cruz, P. Systematic Review: Predicting and Optimising Response to Anti-TNF Therapy in Crohn's Disease - Algorithm for Practical Management. *Alimentary Pharmacology & Therapeutics* **43**, 30–51. doi:10.1111/apt.13445 (2016).
238. Kopylov, U. & Seidman, E. Predicting Durable Response or Resistance to Antitumor Necrosis Factor Therapy in Inflammatory Bowel Disease. *Therapeutic Advances in Gastroenterology* **9**, 513–526. doi:10.1177/1756283X16638833 (2016).

239. Noor, N. M., Verstockt, B., Parkes, M. & Lee, J. C. Personalised Medicine in Crohn's Disease. *The Lancet Gastroenterology & Hepatology* **5**, 80–92. doi:10.1016/S2468-1253(19)30340-1 (2020).
240. Arijis, I. *et al.* Mucosal Gene Signatures to Predict Response to Infliximab in Patients with Ulcerative Colitis. *Gut* **58**, 1612–1619. doi:10.1136/gut.2009.178665 (2009).
241. Arijis, I. *et al.* Predictive Value of Epithelial Gene Expression Profiles for Response to Infliximab in Crohn's Disease. *Inflammatory Bowel Diseases* **16**, 2090–2098. doi:10.1002/ibd.21301 (2010).
242. West, N. R. *et al.* Oncostatin M Drives Intestinal Inflammation and Predicts Response to Tumor Necrosis Factor–Neutralizing Therapy in Patients with Inflammatory Bowel Disease. *Nature Medicine* **23**, 579–589. doi:10.1038/nm.4307 (2017).
243. Martin, J. C. *et al.* Single-Cell Analysis of Crohn's Disease Lesions Identifies a Pathogenic Cellular Module Associated with Resistance to Anti-TNF Therapy. *Cell* **178**, 1493–1508.e20. doi:10.1016/j.cell.2019.08.008 (2019).
244. Gaujoux, R. *et al.* Cell-Centred Meta-Analysis Reveals Baseline Predictors of Anti-TNF α Non-Response in Biopsy and Blood of Patients with IBD. *Gut* **68**, 604–614. doi:10.1136/gutjnl-2017-315494 (2019).
245. Verstockt, B. *et al.* Low TREM1 Expression in Whole Blood Predicts Anti-TNF Response in Inflammatory Bowel Disease. *EBioMedicine* **40**, 733–742. doi:10.1016/j.ebiom.2019.01.027 (2019).
246. Burke, K. E. *et al.* Genetic Markers Predict Primary Nonresponse and Durable Response to Anti-Tumor Necrosis Factor Therapy in Ulcerative Colitis. *Inflammatory Bowel Diseases* **24**, 1840–1848. doi:10.1093/ibd/izy083 (2018).
247. Sazonovs, A. *et al.* HLA-DQA1*05 Carriage Associated With Development of Anti-Drug Antibodies to Infliximab and Adalimumab in Patients With Crohn's Disease. *Gastroenterology*. doi:10.1053/j.gastro.2019.09.041 (2019).
248. Cinelli, C., Forney, A. & Pearl, J. *A Crash Course in Good and Bad Controls* R-493 (2020), 10.
249. Hoffman, G. E. & Schadt, E. E. variancePartition: Interpreting Drivers of Variation in Complex Gene Expression Studies. *BMC Bioinformatics* **17**. doi:10.1186/s12859-016-1323-z (2016).
250. Hoffman, G. E. & Roussos, P. Dream: Powerful Differential Expression Analysis for Repeated Measures Designs. *Bioinformatics* (ed Gorodkin, J.) doi:10.1093/bioinformatics/btaa687 (2020).
251. McNeish, D. Small Sample Methods for Multilevel Modeling: A Colloquial Elucidation of REML and the Kenward-Roger Correction. *Multivariate Behavioral Research* **52**, 661–670. doi:10.1080/00273171.2017.1344538 (2017).

252. Perperoglou, A., Sauerbrei, W., Abrahamowicz, M. & Schmid, M. A Review of Spline Function Procedures in R. *BMC Medical Research Methodology* **19**. doi:10.1186/s12874-019-0666-3 (2019).
253. Zyla, J., Marczyk, M., Domaszewska, T., Kaufmann, S. H. E., Polanska, J. & Weiner, J. Gene Set Enrichment for Reproducible Science: Comparison of CERNO and Eight Other Algorithms. *Bioinformatics* **35** (ed Wren, J.) 5146–5154. doi:10.1093/bioinformatics/btz447 (2019).
254. Chaussabel, D. *et al.* A Modular Analysis Framework for Blood Genomics Studies: Application to Systemic Lupus Erythematosus. *Immunity* **29**, 150–164. doi:10.1016/j.immuni.2008.05.012 (2008).
255. Salvador-Martín, S. *et al.* Gene Signatures of Early Response to Anti-TNF Drugs in Pediatric Inflammatory Bowel Disease. *International Journal of Molecular Sciences* **21**, 3364. doi:10.3390/ijms21093364 (2020).
256. Zalocusky, K. A. *et al.* The 10,000 Immunomes Project: Building a Resource for Human Immunology. *Cell Reports* **25**, 513–522.e3. doi:10.1016/j.celrep.2018.09.021 (2018).
257. Boyapati, R. K., Rossi, A. G., Satsangi, J. & Ho, G.-T. Gut Mucosal DAMPs in IBD: From Mechanisms to Therapeutic Implications. *Mucosal Immunology* **9**, 567–582. doi:10.1038/mi.2016.14 (2016).
258. Martinez, F. O. The Transcriptome of Human Monocyte Subsets Begins to Emerge. *Journal of Biology* **8**, 99. doi:10.1186/jbiol206 (2009).
259. Villani, A.-C. *et al.* Single-Cell RNA-Seq Reveals New Types of Human Blood Dendritic Cells, Monocytes, and Progenitors. *Science* **356**, eaah4573. doi:10.1126/science.aah4573 (2017).
260. Gutierrez-Arcelus, M. *et al.* Allele-Specific Expression Changes Dynamically during T Cell Activation in HLA and Other Autoimmune Loci. *Nature Genetics* **52**, 247–253. doi:10.1038/s41588-020-0579-4 (2020).
261. Tracey, D., Klareskog, L., Sasso, E. H., Salfeld, J. G. & Tak, P. P. Tumor Necrosis Factor Antagonist Mechanisms of Action: A Comprehensive Review. *Pharmacology & Therapeutics* **117**, 244–279. doi:10.1016/j.pharmthera.2007.10.001 (2008).
262. Imhann, F. *et al.* The 1000IBD Project: Multi-Omics Data of 1000 Inflammatory Bowel Disease Patients; Data Release 1. *BMC Gastroenterology* **19**. doi:10.1186/s12876-018-0917-5 (2019).
263. Senn, S. Statistical Pitfalls of Personalized Medicine. *Nature* **563**, 619–621. doi:10.1038/d41586-018-07535-2 (2018).
264. Senn, S. Mastering Variation: Variance Components and Personalised Medicine. *Statistics in Medicine* **35**, 966–977. doi:10.1002/sim.6739 (2016).
265. Farahbod, M. & Pavlidis, P. Untangling the Effects of Cellular Composition on Coexpression Analysis. *Genome Research*, gr.256735.119. doi:10.1101/gr.256735.119 (2020).

266. De Vries, D. H. *et al.* Integrating GWAS with Bulk and Single-Cell RNA-Sequencing Reveals a Role for LY86 in the Anti-Candida Host Response. *PLOS Pathogens* **16** (ed May, R. C.) e1008408. doi:10.1371/journal.ppat.1008408 (2020).
267. Efron, B. Prediction, Estimation, and Attribution. *Journal of the American Statistical Association* **115**, 636–655. doi:10.1080/01621459.2020.1762613 (2020).
268. Chibon, F. Cancer Gene Expression Signatures – The Rise and Fall? *European Journal of Cancer* **49**, 2000–2009. doi:10.1016/j.ejca.2013.02.021 (2013).
269. Querec, T. D. *et al.* Systems Biology Approach Predicts Immunogenicity of the Yellow Fever Vaccine in Humans. *Nature Immunology* **10**, 116–125. doi:10.1038/ni.1688 (2009).
270. Davey Smith, G. & Hemani, G. Mendelian Randomization: Genetic Anchors for Causal Inference in Epidemiological Studies. *Human Molecular Genetics* **23**, R89–R98. doi:10.1093/hmg/ddu328 (2014).
271. Neumeyer, S., Hemani, G. & Zeggini, E. Strengthening Causal Inference for Complex Disease Using Molecular Quantitative Trait Loci. *Trends in Molecular Medicine* **26**, 232–241. doi:10.1016/j.molmed.2019.10.004 (2020).
272. Wang, L. & Michoel, T. Efficient and Accurate Causal Inference with Hidden Confounders from Genome-Transcriptome Variation Data. *PLOS Computational Biology* **13** (ed Listgarten, J.) e1005703. doi:10.1371/journal.pcbi.1005703 (2017).
273. Munafò, M. R. & Davey Smith, G. Robust Research Needs Many Lines of Evidence. *Nature* **553**, 399–401. doi:10.1038/d41586-018-01023-3 (2018).
274. Taylor, D. L. *et al.* Integrative Analysis of Gene Expression, DNA Methylation, Physiological Traits, and Genetic Variation in Human Skeletal Muscle. *Proceedings of the National Academy of Sciences* **116**, 10883–10888. doi:10.1073/pnas.1814263116 (2019).
275. Zheng, J. *et al.* Phenome-Wide Mendelian Randomization Mapping the Influence of the Plasma Proteome on Complex Diseases. *Nature Genetics* **52**, 1122–1131. doi:10.1038/s41588-020-0682-6 (2020).

List of Abbreviations

AC allele count

ASE allele-specific expression

BH Benjamini-Hochberg

BMI body mass index

BTM blood transcription module

CD Crohn's disease

CIT causal inference test

CPM counts per million

CRP C-reactive protein

DAMP damage-associated molecular pattern

DC dendritic cell

df degree of freedom

DGE differential gene expression

eQTL expression quantitative trait locus

FACS fluorescence-activated cell sorting

FC fold change

FDR false discovery rate

FWER family-wise error rate

GWAS genome-wide association study

HA haemagglutinin

HAI haemagglutination inhibition

| | |
|---------------|--|
| HBI | Harvey Bradshaw index |
| HIRD | Human Immune Response Dynamics |
| HLA | human leukocyte antigen |
| HWE | Hardy-Weinberg equilibrium |
| IBD | inflammatory bowel disease |
| IMID | immune-mediated inflammatory disease |
| INT | inverse normal transformation |
| LAIV | live attenuated influenza vaccine |
| LD | linkage disequilibrium |
| lfsr | local false sign rate |
| LMM | linear mixed model |
| LOCO | leave-one-chromosome-out |
| LOR | loss of response |
| LRT | likelihood ratio test |
| MAF | minor allele frequency |
| MANOVA | multivariate analysis of variance |
| MAR | missing at random |
| MCAR | missing completely at random |
| ML | maximum likelihood |
| MN | microneutralisation |
| MNAR | missing not at random |
| mRNA | messenger RNA |
| MS | multiple sclerosis |
| NA | neuraminidase |
| NK | natural killer |
| OVB | omitted-variable bias |
| PANTS | Personalised Anti-TNF Therapy in Crohn's Disease |

PBMC peripheral blood mononuclear cell

PC principal component

PCA principal component analysis

PNR primary non-response

PRR pattern recognition receptor

PVE proportion of variance explained

RA rheumatoid arthritis

REML restricted maximum likelihood

reQTL response expression quantitative trait locus

RNA-seq RNA-sequencing

SD standard deviation

SLE systemic lupus erythematosus

SNP single nucleotide polymorphism

T1D type 1 diabetes

TF transcription factor

TIV trivalent inactivated influenza vaccine

TLR Toll-like receptors

TMM trimmed mean of M-values

TNF tumour necrosis factor

TRI titre response index

TSS transcription start site

UC ulcerative colitis

UTR untranslated region

- spell-check
- make sure package versions are in, and package names are monospace
- add automatic rounding to x decimal places using num and sisetup
- collaboration note in italics at start of each chapter
- fncychap

Todo list

| | |
|--|-----|
| I'm trying to say that the response to immune perturbation is heritable/influenced by genetics. | iii |
| I prefer associations over signatures. Signatures should have had their predictive power quantified. | iii |
| I agree the intro is built around genetics. I have tried to keep the genetics parts concise while elaborating on gene expression more in the later two sections. | 1 |
| add uses other vars | 7 |
| list a few more types and stims from [69] until [70] | 9 |
| specifically introduce cell types mentioned in later chapters | 10 |
| this is a general systems vaccinology review. specific flu studies are reviewed in ch2 | 11 |
| define what a signature is in terms of correlation and prediction | 12 |
| find best GWAS ref, probably mooney2013SystemsImmunogeneticsVaccines, then prune and reassign these citations | 12 |
| cut this section (???) | 12 |
| why? for diff groups of people | 15 |
| add a point that 2009h1n1 is now circulating seasonally, this is a common trend | 15 |
| Add specific section about pandemrix, it's correlates of protection, it's durability? or maybe in methods | 15 |
| Here, add few points about the immunological response to adjuvanted TIVs i.e. what happens after Pandemrix admin? Involve the innate -> B/CD4T response. Goto plotkins | 15 |
| is there a more recent review? | 16 |
| define 'signature': nvm, remove it | 16 |
| high variability, recheck this was the reason, or quote them | 17 |
| make sure gap and how it is filled is emphed enough | 17 |
| needs 1 more punchline sentence here | 17 |
| why blood? ready easy supply of immune cells, despite delivery being muscle? | 17 |
| atm I'm not using R/NR. wording here implys I am | 18 |
| heterogeneity: well of course there was | 18 |
| cite appropriate subfigures here | 18 |
| change score is usually negatively correlated to baseline [115] | 18 |
| this is due to within-individual RTM , which is why we residualise | 18 |
| the variable is not used as an inclusion/exclusion criterion for the study, otherwise regression to the mean will be strong | 18 |

| | |
|---|----|
| upend change score bit, the only thing we are concerned with here is clifton2019CorrelationBaselineScore | 18 |
| cite appropriate subfigures here, after adding proper subfigure labels | 18 |
| Add to collab note that extractions were done at KCL | 21 |
| Add Tracy-Widom statistics for PCs to justify later choice of 4 PCs for covariates | 22 |
| nicer version, copy the peer code, facet the hird and hapmap samples | 22 |
| Can add other fastqc plots e.g. kmers, overrepresented seqs, seq length | 22 |
| add software versions | 22 |
| cite relevant preprocessing sections | 30 |
| combat does have a pro in that it can do per gene scaling, that fixed fx won't do | 30 |
| this is not a very precise justification. actually, if I were to color R/NR in the PCA plot, R/NR doesn't really explain a lot of var in global gene expression. that's probably why the results don't change much. | 30 |
| weaken this. combat is used multiple times in ch3 | 30 |
| be more specific about how combat works i.e. estimates factors per gene per batch? | 30 |
| this is DGE specific normalisation, which is why it goes here, not in the preprocessing section | 32 |
| link to papers justifying sex, age, ancestry as significant effects on immune gene expression | 32 |
| add equation from ch3. especially justify having TRI in as predictor, by noting equiv of traditional lm to contrasts | 32 |
| add section labels | 32 |
| add label | 32 |
| make all the notation in this section consistent with, and add the equation 2.1. The normal- normal hierarchical model, [144] | 32 |
| why is this? is it having well powered studies? gelman is vague | 33 |
| the derivation here is $qnorm(0.975, mean=0, sd=1*10) = 1*19.59964$, bit iffy, double check this is correct | 33 |
| could also include a table of all sets of parameters here? | 33 |
| add note on ositive regression dependency [28] | 35 |
| add comment on symmetry | 35 |
| more text | 35 |
| can also add MSigDB hallmark sets, which include interferon sets; and of course gene ontology sets | 37 |
| not sure of interpretation at FGFBP2, it is indeed highly expressed in NKs through https://dice-database.org/genes/FGFBP2 | 37 |
| any point in a table of e.g. top 20 DE genes, or is the gene set analysis already enough? . | 37 |
| change x axis labels to baseline, specify top 10 procedure in figure caption | 37 |
| finish citing | 37 |
| add label | 37 |
| figure x labels here should be TRI, not R.vs.NR | 37 |
| Not sure if there is a biological interpretation of downreg of T cells and NK cells gene sets at day 1, since it could be due to increase in other cell types in the sample. similar findings in [161] though | 41 |

| | |
|--|----|
| lit search for downregulation interpretation paper, and downreg T cell paper | 41 |
| might have to rerun everything using the original binary R/NR if this line of reasoning isn't strong enough | 41 |
| move numbers to results? | 41 |
| could comment on phenotype differences too, i.e. HIRD measure antibodies at d63, much later than is popular in the field: d28 usually | 42 |
| should probably emph sobolev didn't find prevacc signatures, and we did. But it's not exactly fair, as sobolev didn't use gene set enrichment as far as i can tell | 42 |
| At no point in this chapter are we estimating causal effects: add point on why not CIT (lower n than franco) | 42 |
| found signatures, but so what? Feels like chapter lacks a punchline? | 42 |
| pull in citations from intro | 43 |
| distinction between expression/ab response is blurry here | 44 |
| straighten out tenses | 44 |
| 1 more sentence to round off context | 44 |
| upend change score bit, here expression is a y variable. compare all the options as glms etc. | 45 |
| Can this really demonstrate genotype-dependent change in gene expression between time- points? i.e. need understand how the change score/ANCOVA approaches differ from repeated measures ANOVA differ from the interaction/stratified approach I take? . . | 45 |
| why I didn't just do a mega-analysis in chapter 2 then, given I haven't any evidence if it's better or worse than Bayesian meta-analysis in that context. | 45 |
| add -7 note as with ch2 | 45 |
| add some indication of how much inflation can be reduced by LMMs | 46 |
| add chr1 loco kinship matrix as example, note the estimates for self-relatedness on the diagonals are not constrained to be 1 | 46 |
| helps with coloc | 46 |
| emph here that the sims match what my def of reqtl is for rest of chapter | 46 |
| log scale: as interactions depend on the scale at which departure from additivity is detected | 47 |
| add sample sizes and model for expression sim | 47 |
| determine appropriate citations from existing refs in intro | 47 |
| ?? | 47 |
| add comment on existence of chosen cell types in samples, and clustering by visit | 49 |
| does not bias least squares regression, but unstable (vary sample to sample) to changes in data due to sampling var, and more std error of estimates will be high (tending to inf if perfect multi) | 49 |
| no need for both size and color, use one for contribution percent | 49 |
| add info on the markers used for the chosen FACS counterparts | 49 |
| get subset size | 49 |
| change corr scatterplots to corr matrix | 49 |
| remake this with only top k factors, and prune the possible covariates | 54 |
| add approximate MAFs, then cite hierarch paper | 54 |
| add note on treating x chrom variants with caution | 54 |

| | |
|---|----|
| lift proper vector notation from limix, then redo this with a timepoint subscript | 54 |
| add formulation of the 0-mean random effect to show exactly how the kinship matrix is used [196] | 56 |
| note stacking of kinship for day -7 repeated measures | 56 |
| i leave the pcs in to guard against unusually differentiated between pop markers, where random effect alone may not be enough [179], https://www.nature.com/articles/srep06874 | 56 |
| recheck if did I do a SNPs only filter | 56 |
| note this is critical, since we know a priori not independent due to eqtl sharing | 56 |
| move lfsr explanation prior to ashr in dge chapter | 56 |
| not sure whether this is conservative or anti-conservative | 57 |
| mashr does not provide by default | 57 |
| RNAseq does test about 7000 more genes though... | 59 |
| be more specific: "moderator", 'modify'????? | 59 |
| point is, doesn't make sense to assume the genotype effect is the same at all levels of cell type abundance | 59 |
| can we interpret with peer in? add note of CLAIM here that although peer is correlated with xcell, interactions are only formed with xcell, so the interaction term can be interpreted per unit of genotype increase when xcell=0 | 59 |
| this analysis is incomplete, and is one of the things I would suggest to round off this chapter | 59 |
| if it would be interesting to compare the sharing estimate condition by condition approach to mashr, then redo and pull in eigenmt-bh values | 61 |
| actually, i've found that my PVE approximation is basically rescaled abs(Z), so pve is a bit pointless if we already have z, and doesn't really help with comparability between genes with diff var/MAF | 61 |
| requiring signif post-vaccination may not be correct, as it excludes many dampening effects | 61 |
| the lack of any positional enrichment makes me concerned for false positives? check with ASE? | 61 |
| expand this to plot 1, list top 5 damp, flip, amp at each timepoint | 61 |
| note anything in lit about any of the 30 | 61 |
| reword not significant | 61 |
| double check denoms | 61 |
| convert to subfigures | 61 |
| siunitx permits uncertainties | 65 |
| gene set enrichment for cell type interacting genes to further validate xCell score usefulness | 65 |
| Figure: expression vs monocyte xCell score, colored by genotype, to visually prove the point | 65 |
| Need to consider Nikos' comment that there are too many (1069/13570 significant BH FDR) genotype-platform interactions to use mega-analysis. Consider filtering. | 65 |
| this analysis is incomplete, and is one of the things I would suggest to round off this chapter | 65 |
| FYI the IBD/T cell coloc fine maps to chr2:24935139 T C (rs713586) with PP=1 | 66 |
| add obesity GWAS | 66 |
| compare sharing with mashr and ongen2017EstimatingCausalTissues | 66 |

| | |
|---|----|
| I'm not exactly sure why at the moment. Enrichment analyses so far have not turned up | |
| much. Up regulation of cell cycle TFs is a possibility. | 66 |
| replace mcgovern2015GeneticsInflammatoryBowel with more recent | 66 |
| add lfsr.dge | 68 |
| need to consider: if this kind of thing is what bulk in vivo reQTL can find, they what is | |
| the additional value over FACS? | 68 |
| dge is coupled to reqtl, if you do an enrichment of dge+reqtl overlap genes, likely their | |
| enrichment is driven by DGE signal | 68 |
| harmonise terminology for 'opposite' | 68 |
| check "rs2223286 is associated with profound directional effects in the expression of SELL | |
| dependent upon genotype, with the minor C allele associated with increased expres- | |
| sion of SELL in B-cells and reduced expression of SELL in monocytes " | 68 |
| note coloc doesn't distinguish pleiotropy from mediation? | 69 |
| add 1 concluding line | 69 |
| Overall I feel like the chapter is too descriptive, and falls short of making biological insights | |
| into Pandemrix response. Any additional analyses would hope to address that. . . . | 69 |
| difficult to separate out CD from IBD in literature, section probably needs to decide to | |
| focus on CD only or IBD. | 71 |
| the subsection on anti-TNFs in the intro chapter may be merged here | 72 |
| I assume dosing follows the induction and maintenance schedule from [227]. Can't find | |
| anything about it in the PANTS protocol, although Sim confirmed the 2w/8w fre- | |
| quencies. | 72 |
| add ref to Alex's ppt | 74 |
| trying not to 'spoil' any results in the intro, I guess | 74 |
| compute maximum deviation? | 75 |
| Still discussing with Sim on the exact def of LOR and exit visits to decide whether this is | |
| sensible. | 75 |
| The observational study design does not allow straightforward estimation of causal effects, | |
| so this is only assumed | 79 |
| Could cell proportions act as colliders? | 80 |
| Alter and move this to ch2 | 80 |
| trying to understand how exactly the B-spline basis for the natural spline works will take | |
| a few more youtube lectures... | 81 |
| move this to ch2, then refer to it | 82 |
| collab note: QC done by Alex | 82 |
| I've collected the QC details here from Alex's thesis | 83 |
| precision for integers | 83 |
| much of this section is brief, since the same pipeline as ch3 | 83 |
| as with most of this section, this is more explained in ch3 | 85 |
| yes, there could be ties in lfsr | 86 |

| | |
|--|-----|
| what to put in results vs discussion. going with the pattern of providing enough info for the reader to intepret the data in the results, then doing a summary and my own interpretation in the discussion | 86 |
| Some primary non-responders have loss of response samples. Not sure why. | 86 |
| add caption and label after finalising table data | 86 |
| highlight a few more individual genes specific to this analysis too? not sure how to pick them at the moment. | 92 |
| TODO I forget to include the non-ranked hypergeometric test and gprofiler tests in the gene set enrichment section in the methods. | 97 |
| right word??? | 104 |
| So many modules associations here, maybe try to Google some of them... | 106 |
| How does the MHC come in here? | 107 |
| move some of this to results section | 109 |
| not entirely sure this this is statistically rigourous due to third var effects? | 110 |
| might be worth subbing predictor for independent and response with dependent in all chapters, but then that gets confused with statistical independence... | 118 |
| move and merge this paragraph into ch2 section on composition bias in RNAseq, it's out of place here | 118 |
| would an in vitro stimulation of a mixture of cells be immune to this? differentiation and proliferation could still occur | 119 |
| assumes eQTL is finemapped | 120 |
| I cannot figure out what sort of conditioning would need to be done to block this sort of effect even if you had perfect cell proportion measurements | 120 |
| citation needed, try daveysmith2020MendelLawsMendelian | 123 |
| spell-check | 150 |
| make sure package versions are in, and package names are monospace | 150 |
| add automatic rounding to x decimal places using num and sisetup | 150 |
| collaboration note in italics at start of each chapter | 150 |
| fncychap | 150 |