<title>

Benjamin Yu Hang Bai

2020-05-10 04:59:36+01:00

<dedication>

Abstract

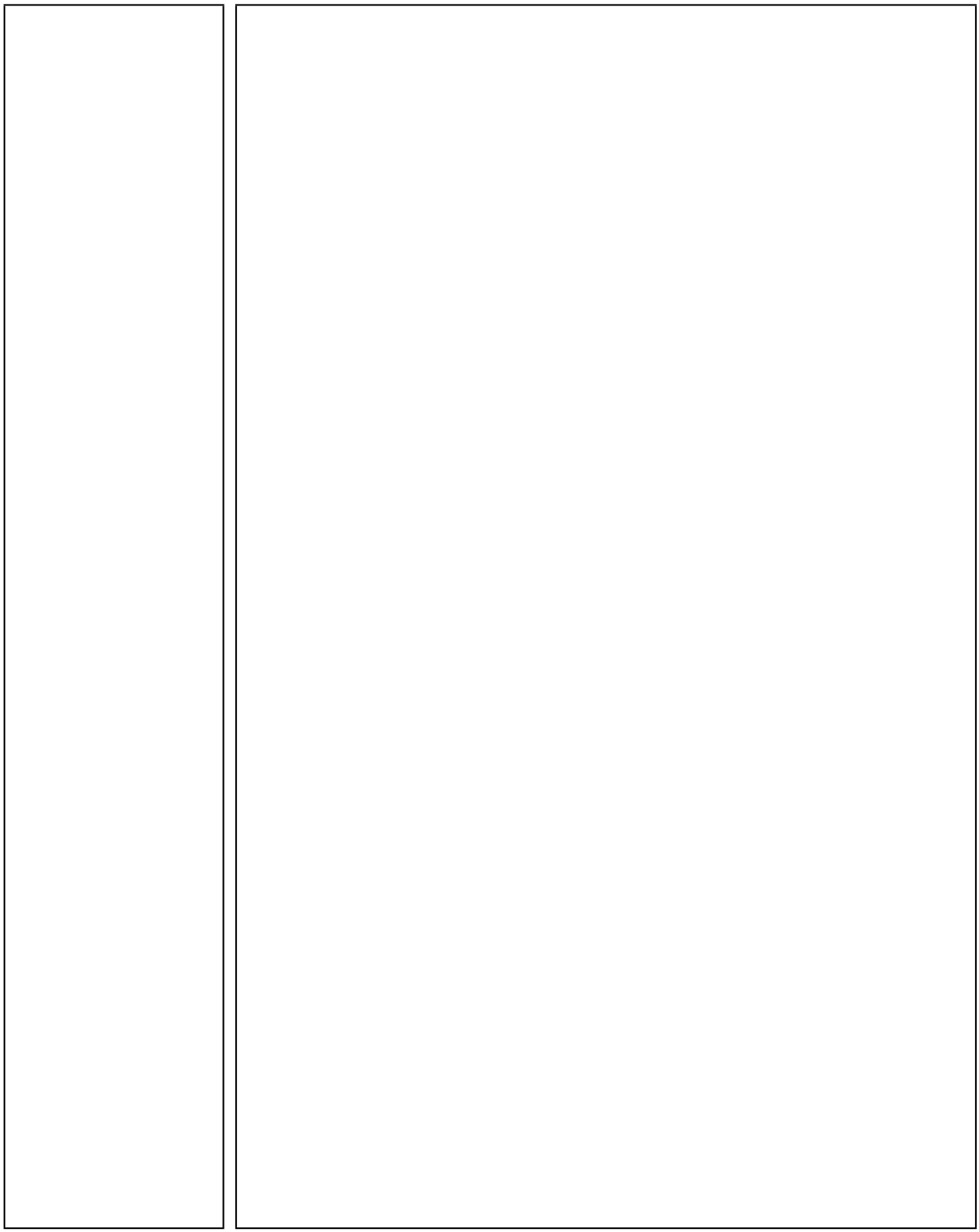
<thesis abstract>

Acknowledgements

<acknowledgements>

Contents

List of Figures



List of Tables

Chapter 1

Introduction

- Variation between humans exists
- The eternal debate: nature vs nurture
- Why study human genetics?
- The structure of the genome and it's variation
- Finding causal anchors
- Leveraging natural G variation.

1.1 A brief overview of genetic association for complex traits

1.1.1 Early days

- Early days, prior to GWAS
- Mendelian genetics, family and linkage studies
- Complex traits and the Common disease, common variant hypothesis
- Twin studies and heritability estimates of complex traits
- Candidate gene studies (Border et al., 2019)
- Appreciation of polygenicity

1.1.2 The advent of GWAS

- 10 years of GWAS
- "The case of the missing heritability"
- genotyping arrays
  - common known variants
  - designed to cover tag variants that represent most genetic variation
  - imputation
- if discovering new var
- WES (about 40Mbp of the genome)
  - covers more of the genome in terms of bp
  - but lower n, so lower power than array genotyping to do single variant associations
  - why 50x? variable coverage due to pulldown
- WGS
  - tradeoff between variant capture (n needed to observe variant) and sequencing depth (gives confidence to call variants)
  - 20x ok to call 90% of singletons
  - rare variants, including in nc regions
    - current discovery biases, finding higher effect size vars first
    - burden tests (e.g. SAIGE)
      - to get gene, aggregate based on variant consequence scores e.g. vep scores
  - structural variants

1.1.3 Narrowing the signal

- PheWAS[1]
- Fine-mapping

  - as sample sizes get larger, and provided that sequencing or imputation can more exhaustively identify all of the candidate SNPs on the haplotype, rare recombination events will pile up, helping to make the causal SNP stand out above the passenger SNPs that usually travel on its haplotype [Huang 2017].
  - tag snps: causal snps may not be directly typed, may need to be imputed

1.1.4 Interpretation of genetic associations with molecular studies

- Locus to gene mapping problem
  - nc snps
    * Genome-wide association studies have successfully identified genetic variants associated with immune-mediated disease, the majority of which are non-coding[10 Years of GWAS Discovery].
- using intermediate/endophenotypes
  - endophenotypes paper
  - expression as an important intermediate
    * measure by array, rnaseq
  - theory is that genetic variants manifest their effects through these phenotypes, central dogma based
- eqtl reviwe: albert2015RoleRegulatoryVariationa

  recent review Vandiedonck
- coloc methods
  - coloc
    * Under the assumption that the mechanism by which non-coding associations affect disease risk is through their effect on gene expression, a successful way to link associations to their target gene is by statistical colocalisation with eQTL datasets, to determine if the GWAS and eQTL signal share

the same causal variant[Co-localization of Conditional eQTL and GWAS Signatures in Schizophrenia].

- TWAS
- MR
- a transcriptional risk score (TRS)

- for eqtls, closest gene is often not the best candidate
  - annotation of nc var is functional genomics
    - e.g. gtex, ENCODE

1.1.5 So what? Translational directions [can cut this whole section]

- Why care?
  - polygenic scores, prs: marker for diagnosis
    - use in the clinic
      - e.g. polygenic background can modify penetrance
    - but challenges from:
      - ancestry effects
      - need expanding into global populations, global biobanks e.g. Gains from Africa H3Africa, japanese biobanks
      - non-ancestry effects
  - pathway analysis: "the great hairball gambit"
  - pathway prs
    - challenge is variant to gene assignment/mapping
      - e.g. restrictions to fine mapped eQTLs
  - Understand mech. of causal genes: molecular pathogenesis
  - Drug target prioritisation for disease traits
  - how to drug a complex disease with no single 'candidate gene'?
    - e.g. of successful GWAS -> drug target
      - drug targets with genetic support are more likely
    - building allelic series

1.2 The effects of genetic variation on expression: context is key

- inthe dreaded GxE interaction
  - "In genetics, context matters"
  - for both gwas, and molQTLs, context is key
- Architecture varies e.g. across cell type and tissues
- emph all these are just interactions with diff things
  - tissue
  - cell type
  - interaction between cells in vivo
  - stimulation conditions

    review: condition/Cell-type specific methods refere to 2019-11-19 Cell-count specific eQTL mapping papers
- QTLs can interact with sex and age
- types of context specific QTL
  - ackerman conditional vs dynamic
- Mechanisms of reQTLs What molecular mechanisms might allow for interaction between Expression quantitative trait locus (eQTL) and different environmental conditions? Four categories of tissue-dependent cis-eQTL effects, and proposed two molecular models.

  coloc of immune mediated traits is enhanced by context-specific eQTLs

1.3 Immunity is a complex trait

Is it even plausible that genetic var is important? Brodin: most env paper.

Immune-mediated diseases Heritability of immune parameters and immune-mediated diseases ranges from

1.3.1 Genetic factors affecting the healthy immune system

Why study health? Factors affecting the healthy immune system.

In healthy populations, ≈50% variation in immune system driven by non-genetic factors, ≈30–40% variation is driven by genetic variation (Liston and Goris 2018).

"Such systems immunology studies in healthy individuals have revealed that human immune systems are incredibly variable among individuals, but very stable within individuals over time (11), and most of this variation is attributed to non-heritable factors (12)."

1.3.2 Genetic factors affecting immune response to challenge

Given the genetic control of the healthy immune system, one can hypothesise that immune response to challenge may also be influenced by genetic factors.

The need for controlled immune challenge in trials. Studies of natural infection are complicated. clinical trials as an opportunity: Vaccines and drugs used as controlled immune challenge.

Posit that eQTls where the genetic effect of

1.3.2.1 Context-specific immune response eQTLs in vitro

The majority of response eQTL mapping experiments to date have been conducted *in vitro*, where one can precisely adjust both the length and intensity of stimulation. Environmental variables including cell type composition or tissue type that are expected to interact with the eQTL effect and may confound the interaction effect with stimulation can be controlled. The choice of experiment system and stimulation can also be hypothesis-driven, for example, if certain tissues are expected to be more relevant for a specific disease. .

add more pros for in vitro reQTLs here, and find citations

One of the first studies to perform response expression quantitative trait locus (reQTL) mapping for an immune stimulation was[2], 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. These reQTLs were enriched for GWAS SNPs associated with host susceptibility to tuberculosis; this was not observed for eQTLs that were not reQTLs.

Since then, *in vitro* immune reQTL studies have been conducted for a variety of experimental systems (e.g. primary CD14+ monocytes[3]) and stimulations (IFNγ and LPS[4]).

Take home messages: - reQTLs develop trans-effects on stimulation[3] Overall, as the number of experimental systems and stimulations increases, large number of eQTLs are only detected.

moost recent are very high thrgouhpug

1.3.2.2 *in vivo* response QTL mapping

less popular A complementary approach.

in vivo pros choice of context whole organism phenotypes more likely to be repeated measures

Review of in vivo mapping. What we learn on top of in vitro (Franco et al., 2013)

Large cohorts:

1.4 Immune response to vaccination

Vaccination has enormous impact on global health [10.1098/rstb.2013.0433].

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.

1.4.1 Systems vaccinology: from empirical to rational vaccinology

History of vaccine dev [summary of low-throughput immunology e.g. animal models]

- Vaccination coverage in vunerable populations is below optimal

However, a vaccine that is highly efficacious in one human population may have significantly lower efficacy in other populations. [1 statistic on vaccine efficiacy differences e.g. rotavirus] Particularly challenging populations for vaccination include the infants and elderly, pregnant, immuno comprimised patients, ethnically-diverse populations, and developing countries. For

the majority of licensed vaccines, there is a lack of understanding regarding the molecular mechanisms that underpin this variation in host immune response. Immunological mechanisms that underpin a specific vaccine's success or failure in a given individual are often poorly understood[Immunological mechanisms of vaccination].

rational vacc, where the key is sys vacc

Review of systems vaccinology (pull out of self_viva_copypasta) These systems vaccinology studies often consider longitudinal measurements of the transcriptomic, cellular, cytokine, and antibody immune responses following vaccination[Vaccinology in the era of high-throughput biology.].

Systems vaccinology is the application of -omics technologies to provide a systems-level characterisation of the human immune system after vaccine-perturbation. Measurements are taken at multiple molecular levels (e.g. genome, transcriptome, proteome), and molecular signatures that correlate with and predict vaccine-induced immunity are identified [http://dx.doi.org/10.1098/rstb.2014.0

define what a signature is

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.

Cotugno - dna meth: DNA methylation [52, 53, 54] events

How to use sysvacc to inform better design (A systems framework for vaccine design Mooney2013), and how to move towards personalised vaccinology (https://doi.org/10.1016/j.vaccine.2017.07.062).

Overview, including pathogen-side factors

1.4.2 Genetic factors affecting vaccine response

Read this `Vaccine.2018August28;36(36):5350\T1\textendash5357.doi:10.1016/j.vaccine.2017.07.062.` Search for "variation in vaccine response genetics GA Poland" in google scholar

measles

Relatively few studies have assessed the impact of human genetic variation on responses[Franco, Lareau 2016].

This is despite evidence from genome-wide association studies suggesting such genetic variation influences immune response to vaccines and susceptibility to disease[Systems immunogenetics of vaccines.].

Results from vaccine-related twin studies e.g. in "TWIN STUDIES ON GENETIC VARIATIONS IN RESISTANCE TO TUBERCULOSIS", and (Defective T Memory Cell Differentiation after Varicella Zoster Vaccination in Older Individuals)

Nice summary table for gwas Review paper on GWAS for vaccines mooney2013SystemsImmunogenetics

Genetics of adverse events e.g. `https://www.ncbi.nlm.nih.gov/pubmed/18454680`

1.5 Immune response to biologic therapies

1.5.1 Genetic factors affecting biologic responses

e.g. PANTS immunogenicity

1.6 Thesis overview

Chapters 1 and 2. Chapter 3. Chapter 4. Chapter 5.

Chapter 2

Transcriptomic response to influenza A (H1N1)pdm09 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)[5].

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[6,7]. On occasion, a novel subtype against which the majority of the population is immunogically 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,

why? for diff groups of people

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

the fourth to occur in the last 100 years[5].

2.1.2 Quantifying immune response to influenza vaccines

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

The 2009 pandemic motivated the rapid development, trialing, and licensing of several novel vaccines[8]. 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[9]. 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 agencies expect a licensed vaccine to meet are based on thresholds such as the proportion of trial individuals achieving HAI titres ≥ 40 and seroconversion ($\geq$ 4-fold increase in titres)[10,11].

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

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[6]. 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[12]. 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[13,14].

is there a more recent review?

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.[15] 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[16]; kinase CaMKIV expression is also a strong predictor[17], as are genes related to B cell proliferation[18].

For these studies of seasonal influenza vaccines in adults, responses tend to be biased by recall from past vaccination or infection[16,19]. There have also been few studies of adjuvanted influenza vaccines, despite their superior efficacy in comparison to non-adjuvanted counterparts[20,21].

2.1.4 The Human Immune Response Dynamics (HIRD) study

The Human Immune Response Dynamics (HIRD) study conducted by Sobolev *et al.* [22] was conceived with the above limitations in mind. The vaccine studied was Pandemrix, an AS03-adjuvanted, split-virion, inactivated vaccine against the influenza A (H1N1)pdm09 strain, 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.

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

make sure gap and how it is filled is emphed enough

vaccine response even pre-vaccination[23–26].

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 statistical power[27,28].

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[22], 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)[16]. 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 co-expressed 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[22]. 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 medasured 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.

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 Fig. 2.1.

2.2.2 Computing baseline-adjusted measures of antibody response

In[22], 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[29], and is used in many studies of seasonal influenza vaccines[13]. 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,[22] 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[27,28].

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

cite appropriate subfigures here

To address these concerns, I computed the TRI as defined in Bucasas *et al.* [16]. For each assay, a linear regression was fit with the $\log_2$ day 63/day -7

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.

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 (Fig. 2.2).

cite appropriate subfigures here, after adding proper subfigure labels

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 (Fig. 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*.

2.2.3 Genotype data generation

Add to collab note that extractions were done at KCL

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 deviations of the mean (threshold selected visually to exlude outliers, Fig. 2.4), removal of markers that deviate from Hardy–Weinberg equilibrium (`--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

*Personal communication with authors.

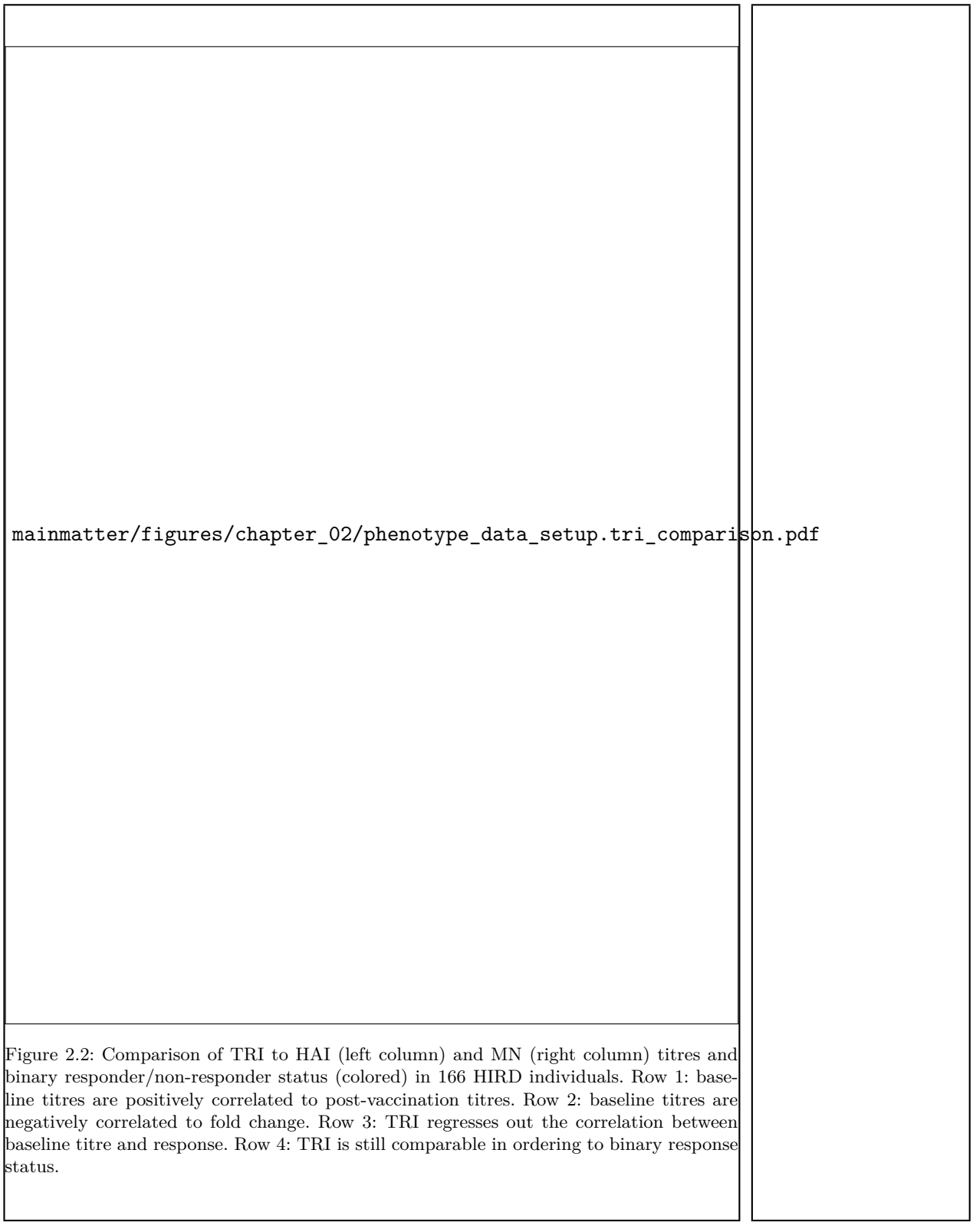


Figure 2.2: Comparison of TRI to HAI (left column) and MN (right column) titres and binary responder/non-responder status (colored) in 166 HIRD individuals. Row 1: baseline titres are positively correlated to post-vaccination titres. Row 2: baseline titres are negatively correlated to fold change. Row 3: TRI regresses out the correlation between baseline titre and response. Row 4: TRI is still comparable in ordering to binary response status.

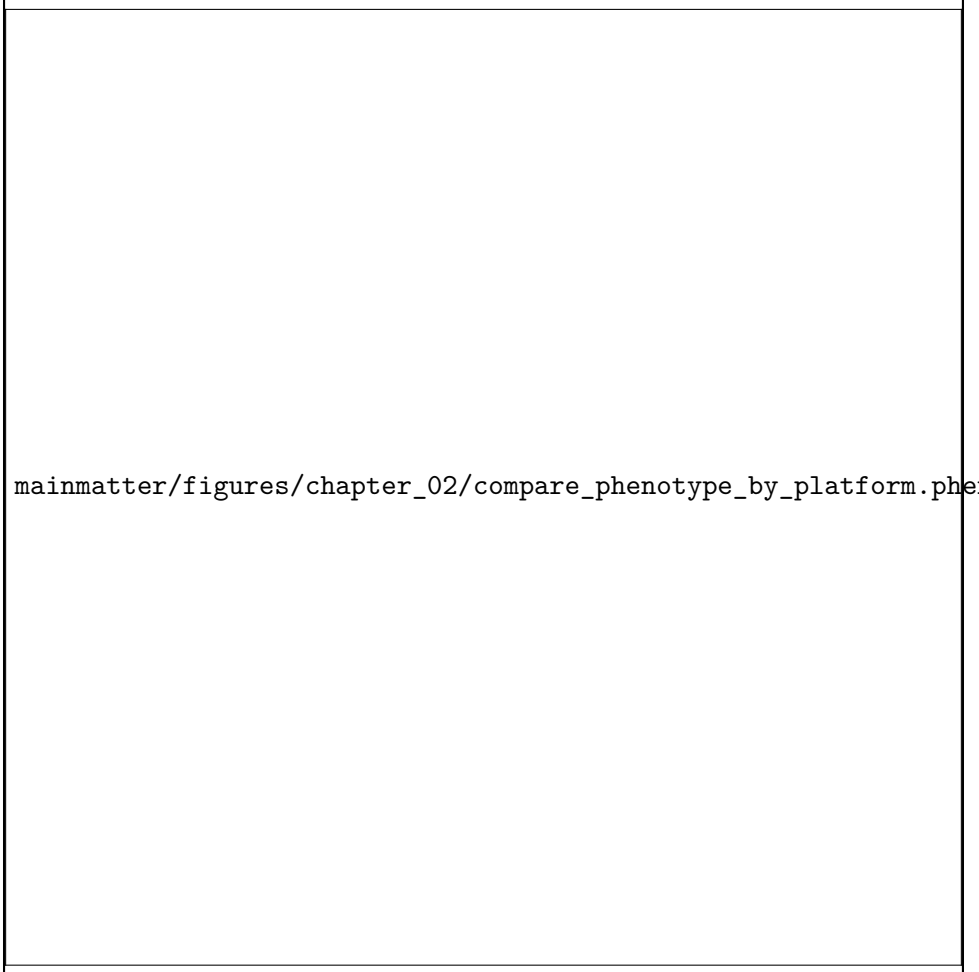


Figure 2.3: Distribution of TRI, stratified by platform used to measure expression.

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

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 ancestry) in expression and genetic association studies[30–32]. Treating HapMap 3 samples 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 (Fig. 2.5), hence these PCs can be used as covariates for ancestry downstream.

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

2.2.6 RNA-seq data generation

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

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, 75bp paired-end).

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

RNA-seq quality metrics were assessed using FASTQC* and Qualimap[33], then visualised with MultiQC[34]. Sequence quality was high (Fig. 2.6), and duplication levels were low (Fig. 2.7). The unimodal GC-content distribution suggested negligible levels of non-human contamination (Fig. 2.8).

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

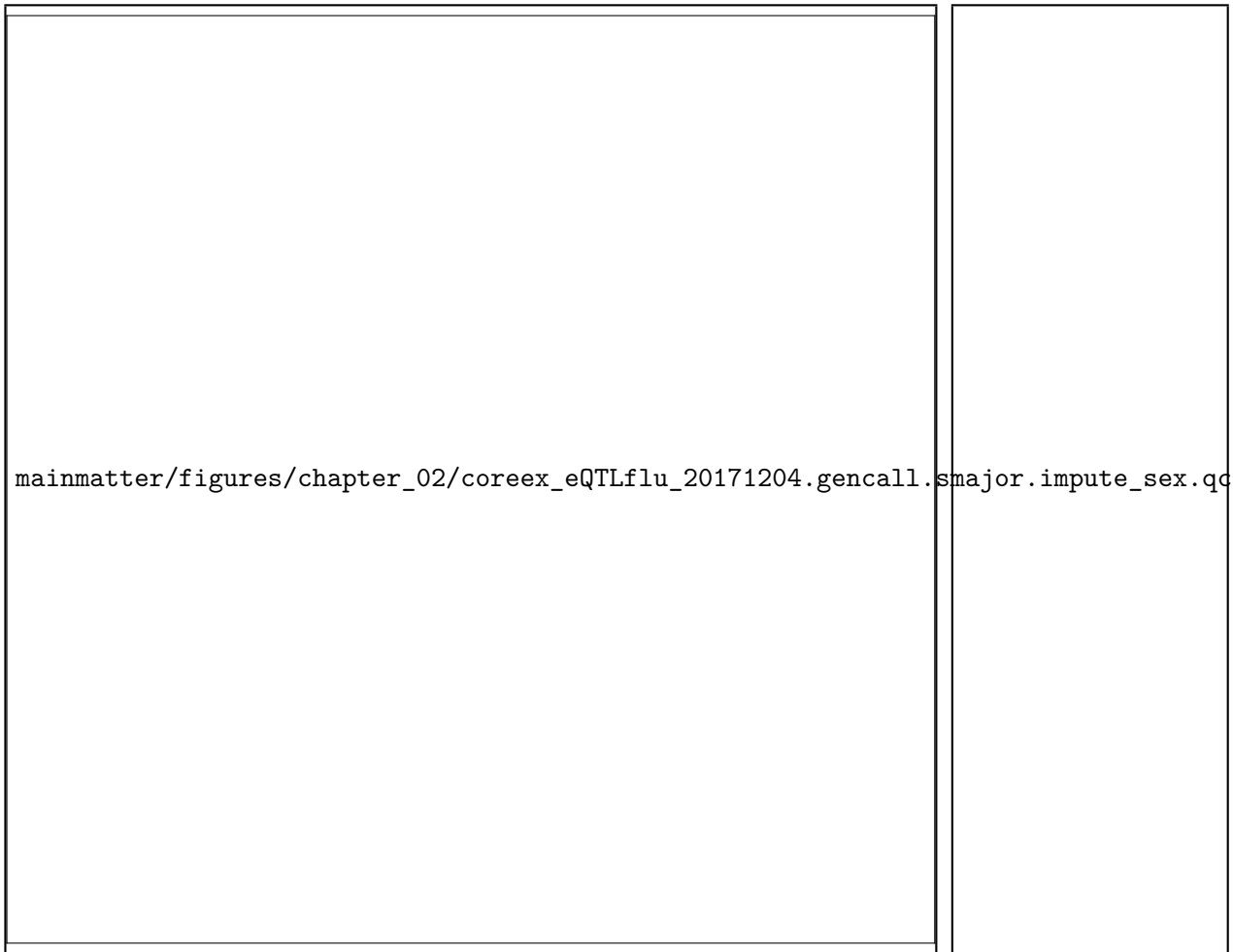


Figure 2.5: HIRD samples (cyan) projected onto PC1 and PC2 axes defined by PCA of HapMap 3 samples. The first two PCs separate European (CEU, upper-right) from Asian (CHB and JPT, lower-right) and African (YRI, lower-left) individuals.

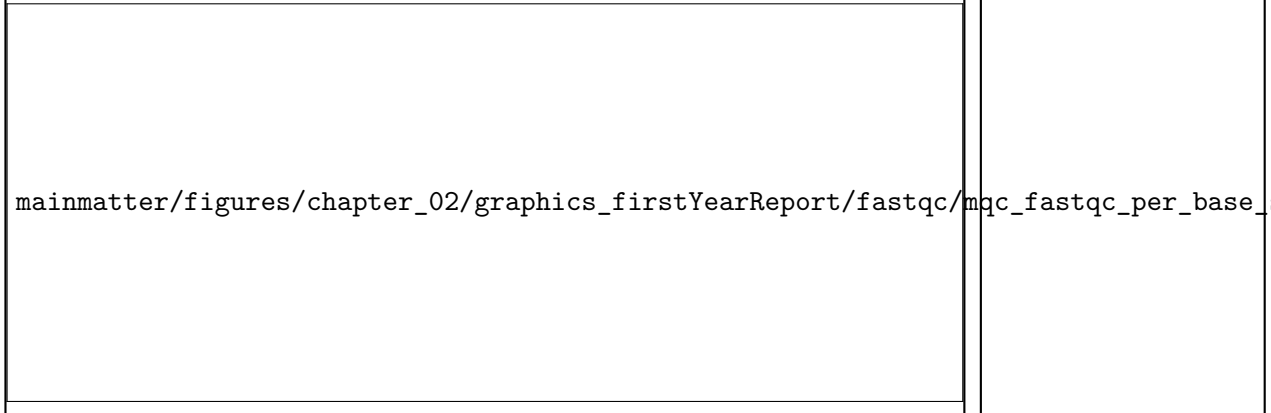


Figure 2.6: FastQC sequence quality versus read position for HIRD RNA-seq samples.

mainmatter/figures/chapter_02/graphics_firstYearReport/fastqc/mqc_fastqc_sequer

Figure 2.7: FastQC sequence duplication levels for HIRD RNA-seq samples.

mainmatter/figures/chapter_02/graphics_firstYearReport/fastqc/mqc_fastqc_per_se

Figure 2.8: FastQC GC profile for HIRD RNA-seq samples.

2.2.7 RNA-seq quantification and filtering

add software versions

Reads were quantified against the Ensembl reference transcriptome (GRCh38) using Salmon[35] in quasi-mapping-based mode, which internally accounts for transcript length and GC composition. To combine technical replicates, as the sum of Poisson distributions remains Poisson-distributed, counts for technical replicates 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[36]. Relative transcript abundances were summarised to Ensembl gene-level count estimates using tximport (scaledTPM method) to improve statistical robustness and interpretability[37].

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[38]. Globin genes, which are highly expressed in erythrocytes and reticulocytes, cell types expected to be depleted in PBMC[39], were also removed. Given the proportion of removed counts at this stage was low for most samples (Fig. 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 (Fig. 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[40]. The change in the distribution of gene expressions among samples before and after filtering shows a substantial number of low expression genes are removed (Fig. 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).

*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

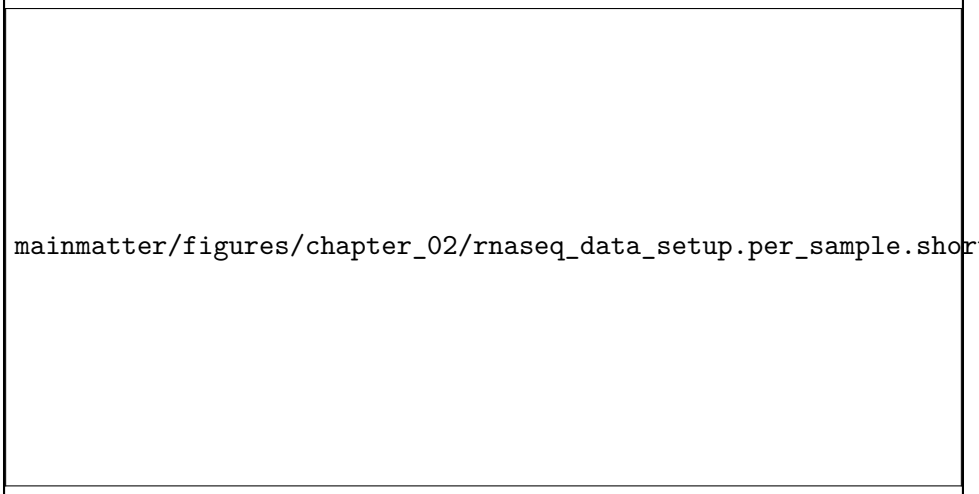


Figure 2.9: Distributions of removed short ncRNA and globin counts as a proportion of total counts in RNA-seq samples.

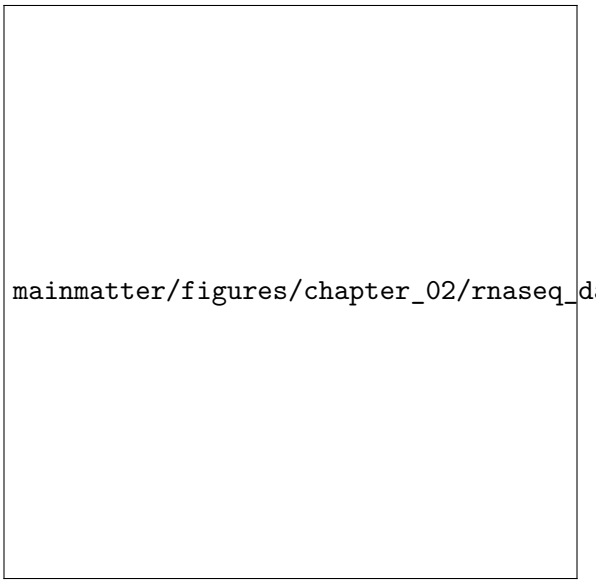


Figure 2.10: Distribution of the proportion of samples in which genes were detected (non-zero expression). Many genes are not detected in any samples. Vertical line shows 5% threshold below which genes were discarded.

mainmatter/figures/chapter_02/rnaseq_data_setup.sample_cpm_density_filtered.pdf

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.

2.2.8 Array data preprocessing

Single-channel Agilent 4x44K microarray (G4112F) data for 173 samples from[22] were downloaded from ArrayExpress*. These arrays were originally processed in two batches, the effect of which is seen in the raw foreground intensities (Fig. 2.12).

VSN[41] 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[42]). 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 Fig. 2.13.

2.2.9 Differential gene expression

PCA of the expression data reveals although samples separate by experimental timepoint along PC3 (Fig. 2.14d), measurement platform is by far the

*https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-2313/

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

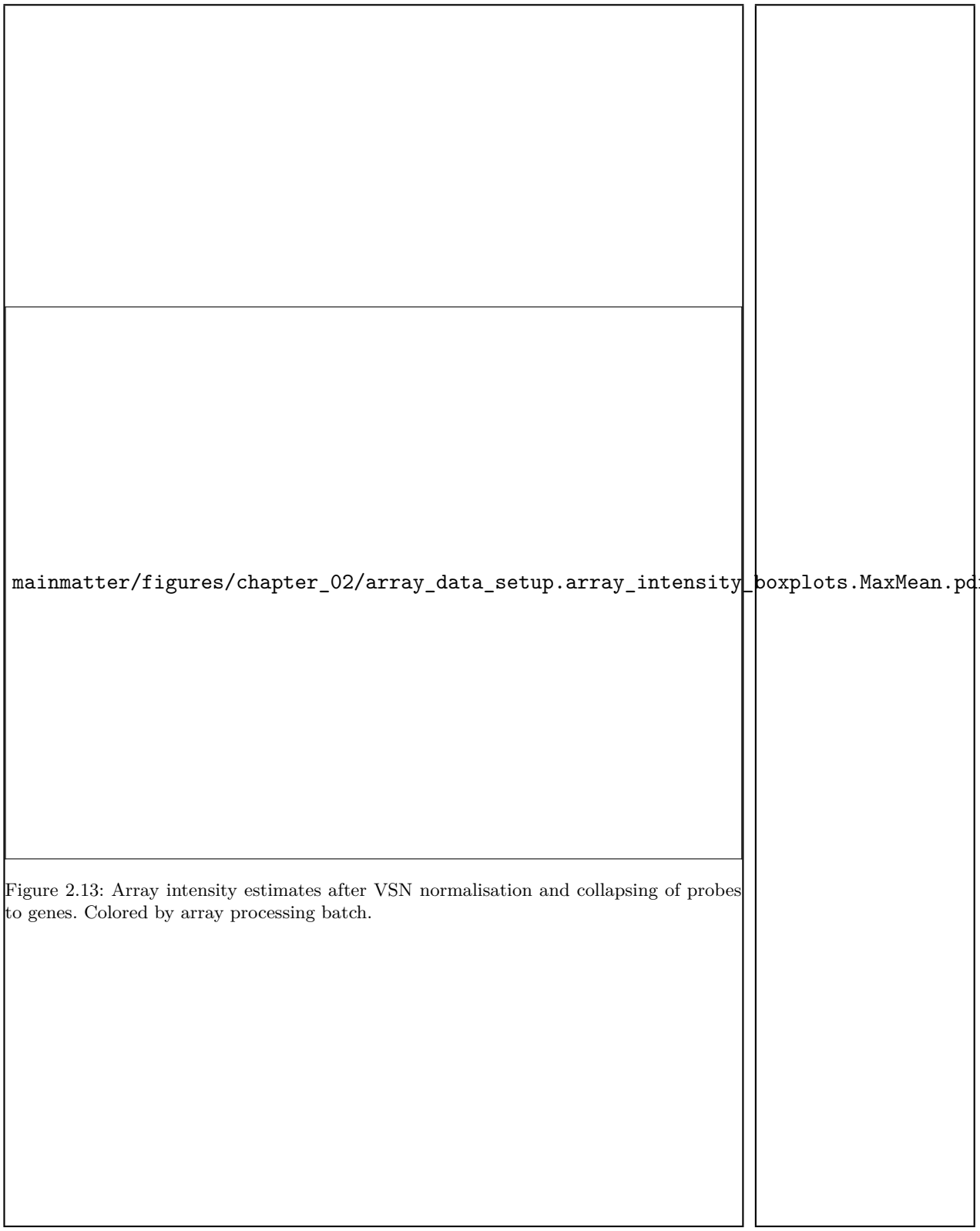


Figure 2.13: Array intensity estimates after VSN normalisation and collapsing of probes to genes. Colored by array processing batch.

largest source of variation. Normalisation was also not able to completely remove the batch effect within the array data (Fig. 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[43]. 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[44]. There are also differences in the statistical models behind expression quantification and normalisation, as described above.

cite relevant preprocessing sections

Despite the shortcomings of array data detailed above, the array dataset tends to contain individuals with more extreme antibody response phenotypes (Fig. 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[45], which estimates centering and scaling parameters by pooling information across all genes using empirical Bayes. ComBat is the method used in[22]. In comparisons of microarray batch effect adjustment methods, ComBat performs favourably (vs. five other adjustment packages)[46] or comparably (vs. batch as a fixed or random effect in the linear model)[47]. However, where batches are unbalanced in terms of sample size[48] or distribution of study groups that have an impact on expression[49], 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 notTable 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.

combat does have a pro in that it can do per gene scaling, that fixed fx won't do

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.

weaken this. combat is used multiple times in ch3

be more specific about how combat works i.e. estimates factors per gene per batch?

2.2.9.1 Per-platform differential gene expression model

For the array data, as[22] 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 was performed

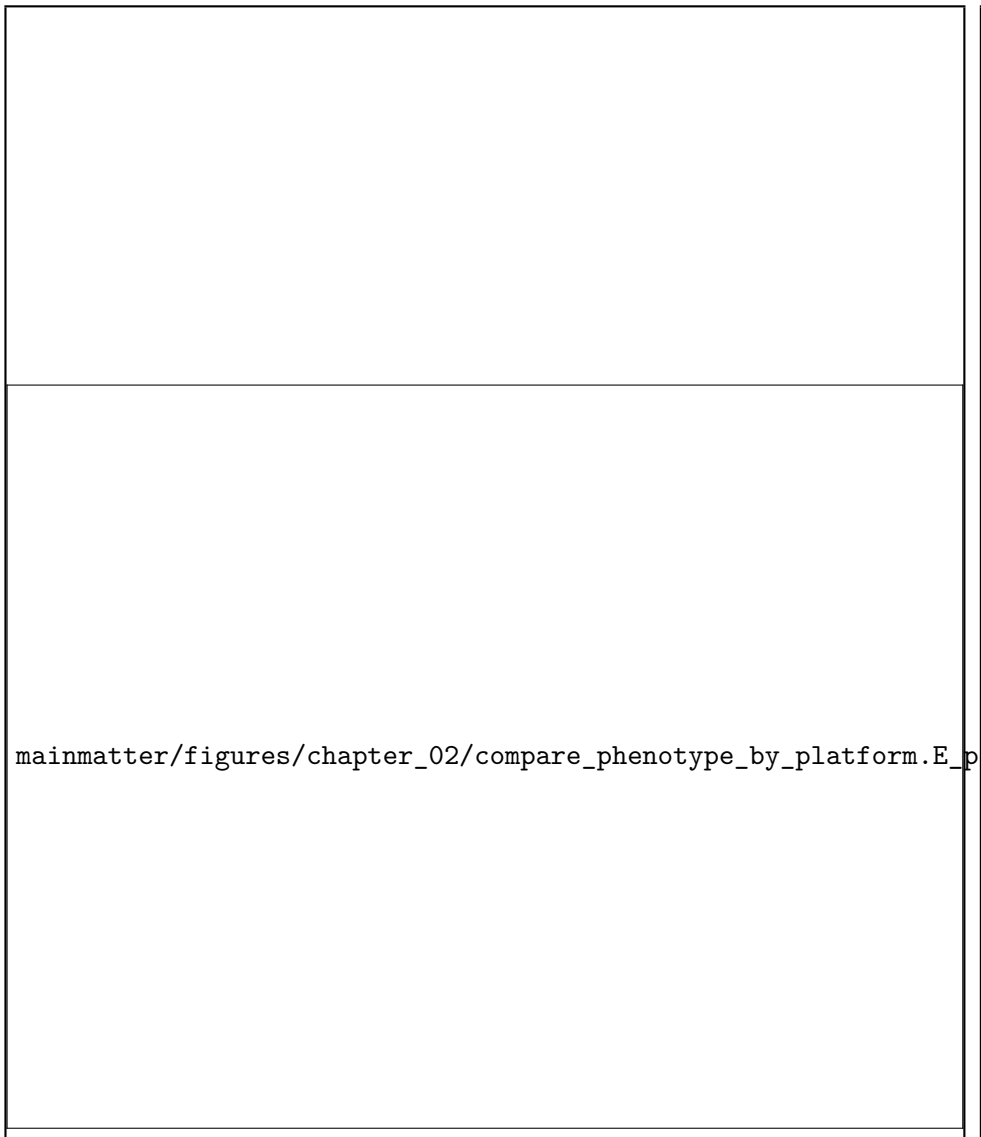


Figure 2.14: First four PCs in the HIRD expression data, colored by platform and batch (left), and timepoint (right).

using the trimmed mean of M-values (TMM) method[50] from edgeR[51]; then variance-stabilisation was performed using voom[52], resulting in expression values with units of $\log_2$ CPM.

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

Linear models were fit using limma[53], which is computationally fast, and performs well for sufficiently large ($n \geq 3$ per group) sample sizes[54]. 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. 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.

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

2.2.9.2 Choice of differential gene expression meta-analysis method

add section labels

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[55,56]. 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.

add label

Unfortunately, there is no optimal solution for directly estimating τ in random-effects meta-analyses with small k[58], in the case of $k = 2$ especially[59]. Many estimators are available[60], but lack of information with small k causes

make all the notation in this section consistent with, and add the equation 2.1. The normal-normal hierarchical model,[57]

estimation to be imprecise, and often results in boundary values of $\tau = 0$ that are incompatible with the assumed positive heterogeneity[61,62]. In such circumstances, the most sensible choice may be to incorporate prior information about model hyperparameters in a Bayesian random-effects framework[60–63]. For this study, I use the implementation in bayesmeta[57], which requires priors for both effect size and between-studies heterogeneity.

2.2.9.3 Prior for between-studies heterogeneity

The choice of prior for between-studies heterogeneity is influential when k is small[63]. Gelman [64] 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[61], 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 constrast to priors used in other studies from the log-normal (e.g.[65,66]) or inverse-gamma (e.g.[67]) 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.[63,64]), 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[60]) 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[62,64]. `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

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

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[68]. 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.

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

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 Fig. 2.15 (for τ) and Fig. 2.16 (for μ). For τ, the final prior used was Gamma(shape $= 1.5693$, scale $= 0.0641$). This is comparable to[61]'s default recommendation of a Gamma(shape $= 2$, scale $= \lambda$) 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.

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

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 than the confidence of a non-zero effect[69].

add comment on symmetry

2.2.10 Gene set enrichment analysis using blood transcription modules

Gene set enrichment analyses were conducted using `tmod::tmodCERNOtest`[70], 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

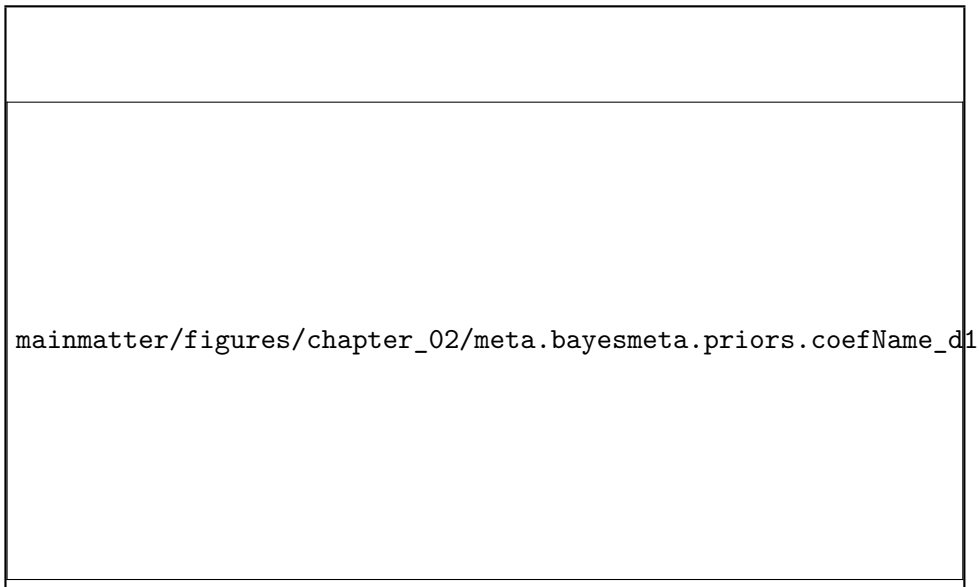


Figure 2.15: Gamma prior for τ used for `bayesmeta` (blue), compared to the empirical distribution of per-gene frequentist `metafor::rma` estimates for τ, for the day 1 vs. baseline effect (small estimates of $\tau < 0.01$ excluded). Empirical log-normal fit also shown (red).

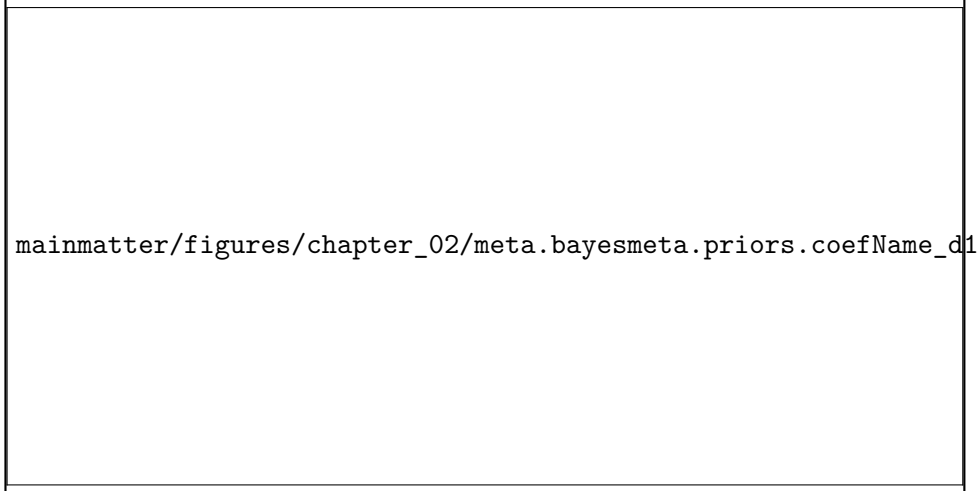


Figure 2.16: Normal prior for μ used for `bayesmeta` (blue), compared to the empirical distribution of per-gene frequentist `metafor::rma` estimates for τ, for the day 1 vs. baseline effect. The non-scaled normal fit is shown (black), as well as a Cauchy fit (red).

I used effect sizes from `bayesmeta`. The gene sets used were blood transcription modules (BTMs) from[71], 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

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 lfsr < 0.05 and absolute FC > 1.5 cutoff, 857/13593 genes were differentially expressed between any pair of timepoints, with their expression clustering into three main clusters (Fig. 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 (Fig. 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* ($\log_2$ FC $= 4.489$), an interferon-induced gene in monocytes and dendritic cells (DCs) involved in antiviral innate immune pathways[72]. Other key genes in the interferon signalling pathway[73] such as *STAT1* ($\log_2$ FC $= 2.1693060$), *STAT2* ($\log_2$ FC $= 0.9489341$), and *IRF9* ($\log_2$ 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 (Fig. 2.18), confirming that day 1 responses are dominated by signatures of innate immunity. 64 genes were downregulated at day 1, enriched

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

mainmatter/figures/chapter_02/plot_dge_eqtl.heatmap_dge.pdf

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

in modules associated with T cells and natural killer (NK) cells, with the largest absolute fold change observed for *FGFBP2* ($\log_2$ FC = −0.9141547). For both up and downregulated genes, there was a tendency to return to baseline expression levels by day 7.

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

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* ($\log_2$ FC = 1.7538617) and *MZB1* ($\log_2$ FC = 1.7369668). Plasma cell-specific genes including *SDC1* (encodes CD138 `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5437827/`) ($\log_2$ FC = 1.3673081) and *ELL2* (`https://www.nature.com/articles/ni.1786`) ($\log_2$ 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 (Fig. 2.18). Both the $CD4^+$ T cell and plasma cell response are indications of an adaptive immune response at day 7.

finish citing

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 FDR < 0.05, 6 genes had expression associated with TRI at baseline, 55 at day 7, and 11 pooling samples across timepoints (Fig. 2.19).[22] 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 FDR < 0.05, of which 58/62 fall into the 13593 genes considered in my meta-analysis (circled, Fig. 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 lfsr < 0.05 at any timepoint, or when pooling samples across all timepoints (Fig. 2.20).

add label

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

Figure 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).

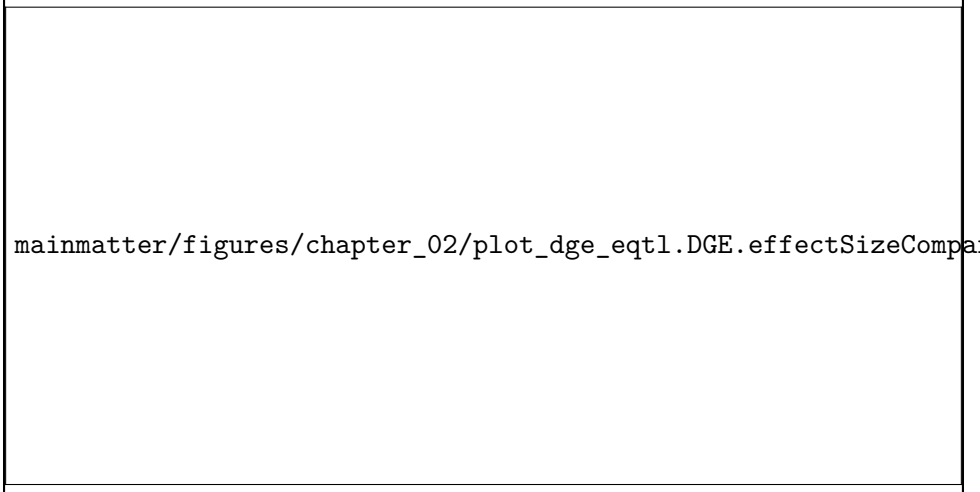


Figure 2.19: DGE effect sizes estimated in array vs. RNA-seq. Significance colored by frequentist random effects meta-analysis FDR < 0.05. Genes with day 7 expression associated with responder/non-responder status in[22] are circled for that contrast.

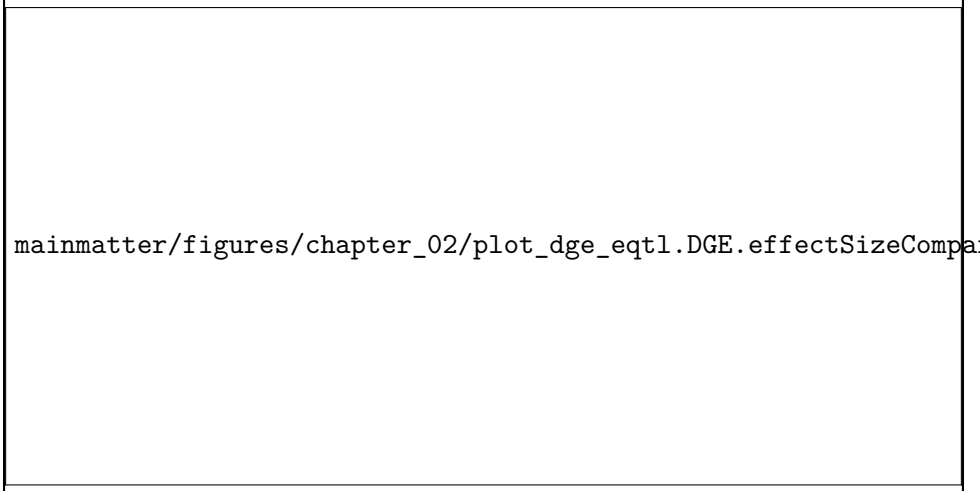


Figure 2.20: DGE effect sizes estimated in array vs RNA-seq. Significance colored by Bayesian random effects meta-analysis lfsr < 0.05. Genes with day 7 expression associated with responder/non-responder status in[22] are circled for that contrast.

inflammatory response in myeloid cells are also associated with high TRI (Fig. 2.21).

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

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

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 differention and proliferation of antibody-secreting plasmablasts and plasma cells[75]. 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[22], 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[22].

Not sure if there is a biological interpretion 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[74] though

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

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[22], 62 genes were reported as differentially expressed between vaccine responders and non-responders. Although[22] 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 (Fig. 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.

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

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

move numbers to results?

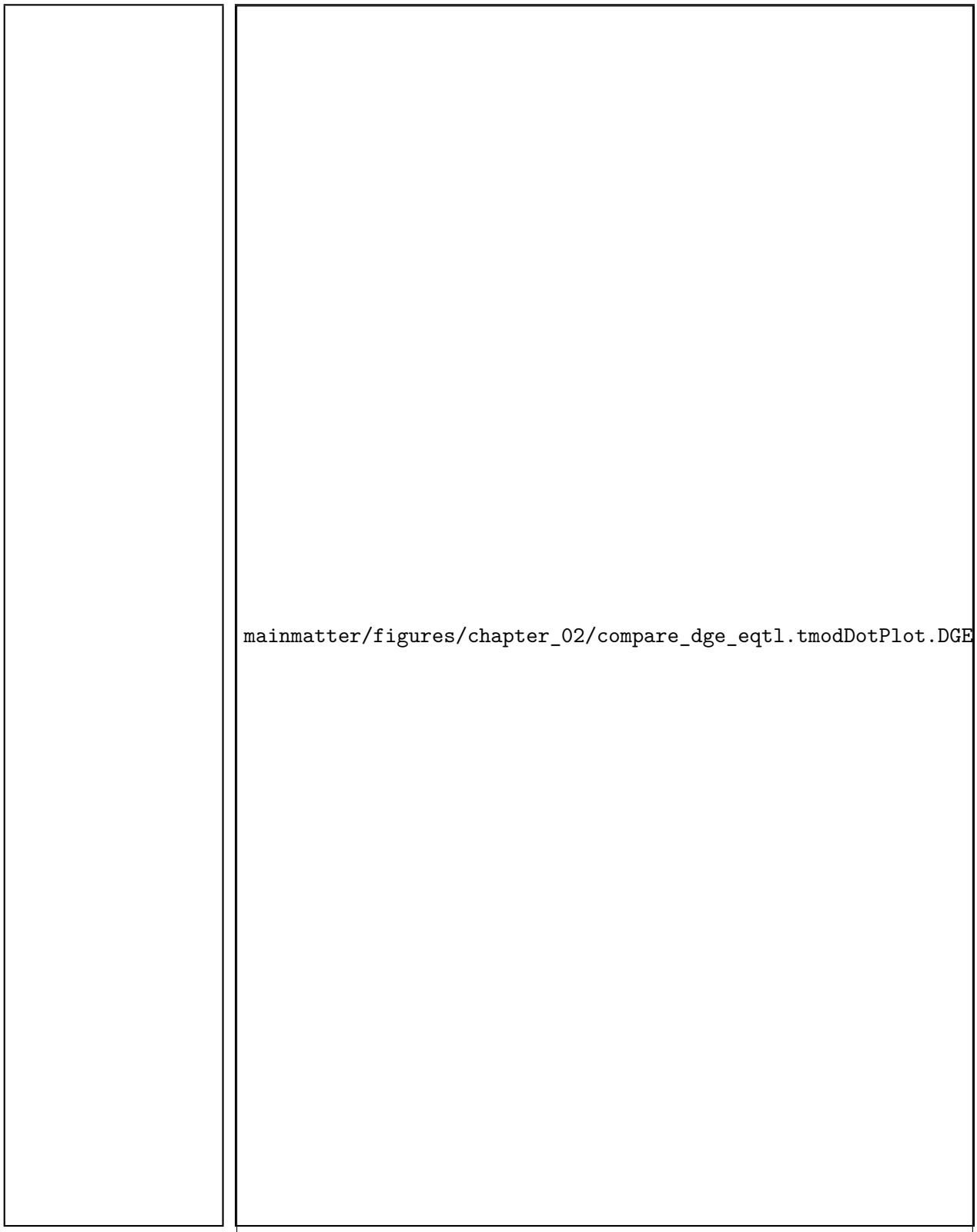


Figure 2.21: Transcriptomic modules enriched in genes with expression associated with antibody response (TRI) at each day. Size of circle indicates effect size. Color of circle indicates significance and direction of effect (red = expression positively correlated with TRI, blue = negative).

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 (Fig. 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, 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[23]; Fcγ receptor-mediated phagocytosis, TREM1 signaling[24]; enriched in B cells, T cell activation[25]; B cell receptor signalling, inflammatory response, platelet activation[26]; 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[20,22]. 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.

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

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.

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

Table 2.1: Sample descriptive statistics.

		platform	
	Total 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)

Table 2.2: HIRD batch balance

	Total	1	2	batch DN500165J	DN500166K	DN500167L
	n = 374	n = 87	n = 79	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)

Chapter 3

Genetic factors affecting Pandemrix vaccine response

3.1 Introduction

3.1.1 Genetic factors affecting influenza vaccine response

- Vaccine-induced antibody response is a complex trait. List GWASes

Specifically for response to seasonal influenza vaccines...

For seasonal

3.1.2 Response expression quantitative trait loci for seasonal influenza vaccination

A potential mechanism through which genetic variation can affect vaccine response is through altering the expression of genes.

eQTLs have condition specificity e.g. cell types or tissues

A reQTL is:[76] - an eQTL becomes more or less important after perturbation: Tells you something about the mechanism of perturbation. - Either expression regulatory activation/repression (signalling cascade -> TFs, chromatin remodelling etc.)

Little work done on reQTL for vaccine stimulation Summarise Franco et al In the case of inactivated trivalent influenza vaccine, genetic variation in membrane trafficking and antigen processing genes was associated with both transcriptomic and antibody responses in patients after vaccination [Franco].

3.1.3 Chapter summary

- In chapter 2, we observed massive changes in gene expression longitudinally after Pandemrix vaccination, as well as expression signatures correlated to degree of antibody responses. [variation observed in response to Pandemrix, e.g R vs. NR trajectories] - How does host genetics affect response to Pandemrix in the HIRD cohort?

Sobolev pros Small effect expected? More variation will usually be explained by history of exposure rather than genetics, so may be harder to detect. but not here

also Knowns Sobolev: R vs NR, inconsistent variation in why people are NR

In this study, we model the influence of host genetics on longitudinal transcriptomic and antibody responses to Pandemrix, in vivo. overall strat: map per timepoint, joint analysis call reqtls characterise

20 genes from franco

12/17 DGE repliacted 14/17 eGenes no reQTLs

subtle enough st methodology wrecks it

3.2 Methods

3.2.1 Genotype phasing and imputation

better to just caveat, and leave numbers in

Prior to imputation, 213277 monomorphic markers that provide no information for imputation were removed. Imputation for the autosomes and X chromosome was conducted using the Sanger Imputation Service*, which involves pre-phasing with EAGLE2 (v2.4), then imputation with PBWT (v3.1) using the Haplotype Reference Consortium (r1.1) panel. Markers were lifted-over from GRCh37 to GRCh38 coordinates using CrossMap. Poorly-imputed markers with INFO < 0.4 or post-imputation missingness $> 5\%$ were removed, resulting in 40290981 markers.

3.2.2 Overall strategy for detecting reQTLs

Since one of the aims of this study is to identify genetic variation that affects expression response to vaccination, it may seem most direct to model the

*https://imputation.sanger.ac.uk/

change in each individual's expression after vaccination as the response variable. This approach has been used to identify condition-specific expression quantitative trait locus (eQTL), typically with the response taking units of log fold change between conditions (e.g.[77,78],`https://doi.org/10.1016/j.ygeno.2014.02.005`). Although a potentially powerful if eQTL effects are small and opposite between conditions[78], 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 expression between conditions is positively correlated[78–80] `https://www.researchgate.net/publication/221689734_Dichotomania_an_obsessive_compulsive_disorder_that_is_badly_affecting_the_quality_of_analysis_of_pharmaceutical_trials`. Instead, I map eQTLs within each of three conditions (pre-vaccination, day 1, and day 7), and find response expression quantitative trait loci (reQTLs) by looking for eQTLs that have different effects between conditions.

Within each timepoint, recall the the Human Immune Response Dynamics (HIRD) dataset includes expression measured by both array and RNA-sequencing (RNA-seq). As discussed in , it is difficult to directly estimate the between-studies heterogeneity when the number of studies is small, a Bayesian meta-analysis approach was preferred. That method does not scale to eQTL analysis, where the number of tests is very large, in the order of thousands of tests per gene versus a few differential gene expression (DGE) contrasts per gene. Instead, I perform a mega-analysis within each timepoint, first merging array and RNA-seq expression estimates into a single matrix with ComBat. For comparison purposes, analyses were also run using 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 due to incomplete power in each condition. Condition-by-condition analysis also makes no attempt to 'borrow information' across conditions to improve power to detect shared associations[81–83]. Counterintuitively, a joint multivariate analysis may be preferable even when associations are not shared across all conditions[84].

A variety of models have been proposed and developed to tackle the is-

no defense against why not just use interactions, apart from scalability genome wide, and additional complexity when also adding cell type and platform interactions, and assumption of homoscedascity between all groups

label to prev ch

no principled reason 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 in that context...

sue of joint eQTL mapping, including classical multivariate methods such as multivariate analysis of variance (MANOVA) (`https://www.nature.com/articles/ncomms6236`), frequentist meta-analyses (e.g. Meta-Tissue, Metasoft), and Bayesian models (e.g. eQtlBma, 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, mash, HT-eQTL). In this chapter, I apply `mash` for the estimation of eQTL effects across my three timepoint conditions. `mash` is a multivariate method that learns patterns of correlation among conditions empirically from condition-by-condition summary statistics, then applies shrinkage to provide improved posterior effect size estimates, along with measures of significance per condition.

3.2.3 Controlling for population structure with linear mixed models

As discussed in chapter 2, the HIRD cohort is multi-ethnic, and population structure can affect gene expression[32]. I addressed this by treating the top principal components (PCs) of the genotype matrix as covariates for large-scale population structure (ancestry). In the context of eQTL mapping, where the aim is to assess the marginal effect of a single genetic variant on expression, it is even more important that the confounding effect of population structure is accounted for.

An appealing approach is the linear mixed model (LMM), which includes a random effect that directly models genetic correlation between individuals as the covariance of that random effect[31,85,86] 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`https://projecteuclid.org/euclid.ss/1271770342`) from finer-scale effects such as family structure[86].

add some indication of how much inflation is reduced by LMMs

3.2.3.1 Estimation of kinship matrices

- The LMM requires a kinship matrix to scale(?) the covariance matrix of the random effect - When testing a variant for association, to avoid loss of power

from 'proximal contamination', kinship matrix used should not include that variant`https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3597090/`. - A simple way to avoid this is to compute leave-one-chromosome-out (LOCO) kinship matrices from all variants on chromosomes other than that variant's chromosome[87].

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)[88]. 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`). - The LOCO kinship matrix excluding chromosome 1 is shown [...]

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

3.2.4 Additional eQTL-specific expression preprocessing

There are a number of transformations that are often applied to expression data before eQTL mapping, such as the rank-based inverse normal transformation (INT) (e.g. used by GTEx v7 `https://storage.googleapis.com/gtex-public-data/Portal_Analysis_Methods_v7_09052017.pdf`), which conforms 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[89]. In multi-condition datasets, these transformations are also typically applied within conditions (e.g. within each tissue individually in GTEx). Another common

transform is standardising (centering and scaling to zero mean and unit variance) `https://github.com/molgenis/systemsgenetics/wiki/eQTL-mapping-analysis-cookbook-for-RNA-seq-data`, so that effects across all genes can be comparably intepreted in units of standard deviation expression change `https://www.nature.com/articles/s41467-018-04558-1`. The impact of these transformations on reQTL detection has not been explored.

I performed simulations to evaluate the effect of the transformation on reQTL calls 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 (Fig. 3.1). Scenario 0 has no eQTL, scenario 1 is a shared eQTL (beta = 1), scenario 2 is a reQTL where beta increases from 0 to 1, scenario 3 is a reQTL where beta increases from 0 to 2, scenario 4 is a reQTL where beta increases from 1 to 2, and scenario 6 is a reQTL where beta increases from 1 to 4. The simulated scenarios were subjected to rank-based INT (Blom method[89]), standardisation, scaling-only, and centering-only transformations. Transformations were applied both within each condition and without separating conditions.

The boxed facets in (Fig. 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 the 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 for my purposes of detecting reQTLs between conditions. Only the centering-only transformation avoids 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 , but automatic outlier exclusion based

need a note here on assumptions: preprocessing xforms inevitably scale, but philosophically I only start thinking about 'preserving' after all that

link to DGE low count filter max zeros section

on standard deviation (SD) thresholds at the eQTL mapping step could be considered.

3.2.5 Estimation of cell type abundance via expression deconvolution

not technically deconv

As 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, estimates for genes that have cell type specific expression depend on the relative proportions of each cell type in each sample, which shift after Pandemrix vaccination[22]. eQTL effects can also be cell type specific. The effect of genotype on expression can be compared between multiple timepoints to call reQTLs as genotype can be assumed to stay constant, but changes in cell type abundance confound this by modifying both expression and the effect of genotype on expression. Immune cell abundance also varies naturally between healthy individuals (`https://www.sciencedirect.com/science/article/pii/S0092867414015906?via%3Dihub`, `https://www.nature.com/articles/nri.2016.125`), so it is important to model these effects even at baseline.

determine appropriate citations from existing refs in ch1

Cell type abundance directly measured via fluorescence-activated cell sorting (FACS) are only available for a small subset of HIRD individuals, so I used expression deconvolution as an alternative to derive cell type abundances from bulk expression data for use in eQTL modelling[90,91]. I selected the `xCell` method, which previously been shown to outperform other deconvolution methods for cell type specific eQTL mapping in blood[91] `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[92]. `xCell` was originally developed for tumor samples, so many of the built-in cell types are not relevant to this study. Reviewing the literature to find which broad classes of peripheral blood cell types might be commonly-expected the PBMC compartment[90], I selected 7/64 of the built-in cell types: 'CD4+ T-cells', 'CD8+ T-cells', 'B-cells', 'Plasma cells', NK cells, Monocytes and DCs. RNA-seq and array expression data from sections were processed separately; the large batch effect present in the array expression was first removed using ComBat. Finally, enrichment scores were standardised, so that score of zero

deconv returns an aggregate measure, so should not confound results for any one gene

link in preproc sections ch2

mainmatter/figures/chapter_03/simulate_expression_transforms.pdf

Figure 3.1: expression xforms

estimates the average abundance of that cell type across all timepoints (**??**).

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 covariates 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`doi10.22237/jmasm/1162353960`, then selected only one cell type with high contribution to each of those components. In both array and RNA-seq datasets, the selected cell types were monocytes, natural killer (NK) cells, and plasma cells (Fig. 3.4). The choice to use actual enrichment scores over principal components directly as covariates is a sacrifice of orthogonality for interpretability.

note here that other cell types are correlated are in the model, but cannot be split

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 ([93] 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 » 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[22], overall correlation between xCell and FACS was weak (Fig. 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 continue the analysis using the xCell scores.

add exact defs for facs

just add all pops

3.2.6 Finding hidden confounders using factor analysis

Apart from cell type abundance, a myriad of other unmeasured variables also contribute to expression variation. Hidden determinants of expression variation were learnt using `PEER`[94]. As recommended by[94], between-sample normalisation and variance stabilisation RNA-seq data was performed using `DESeq2::vsn`, then ComBat was applied to first merge array and RNA-seq data into a single log scale expression matrix per timepoint, treating the two array batches and three RNA-seq library prep pools as known batch ef-

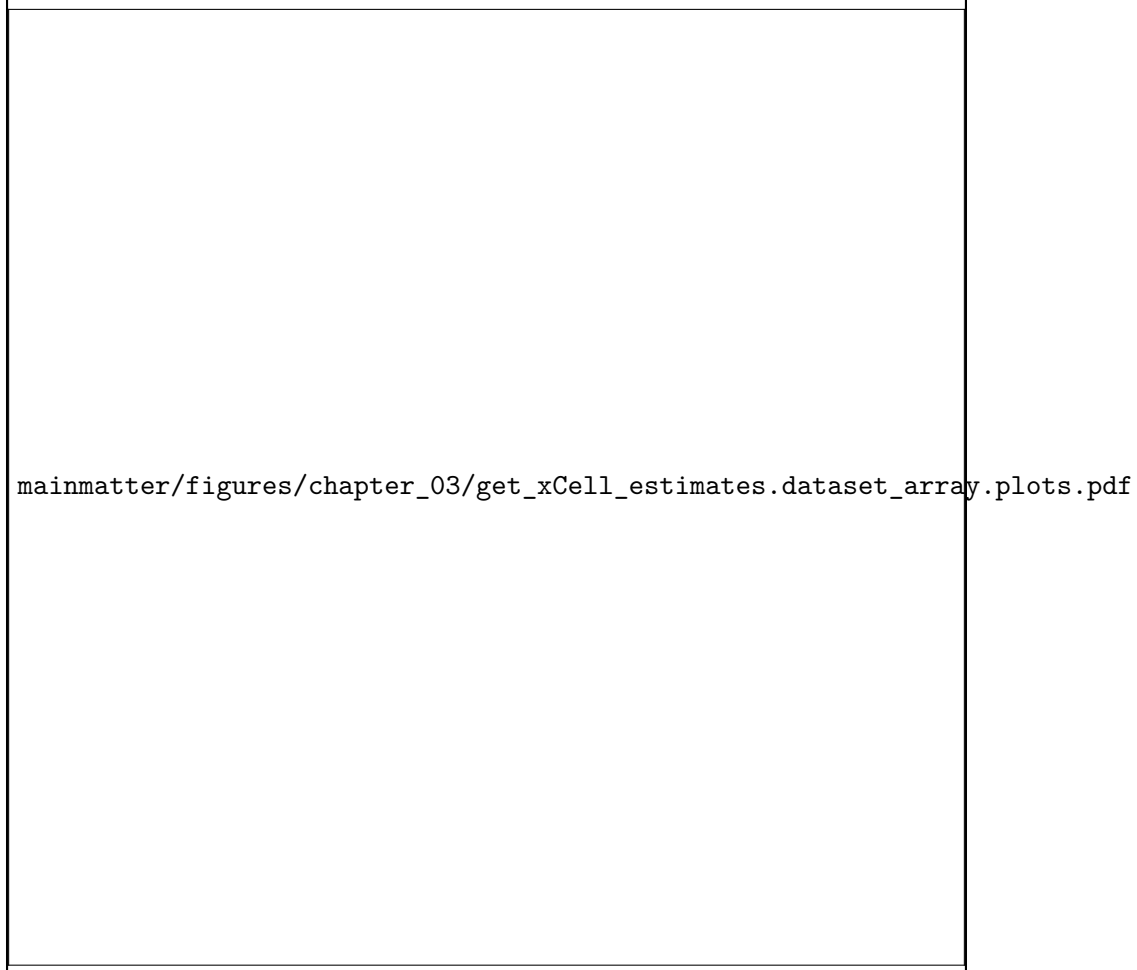


Figure 3.2: xCell enrichment scores in array data

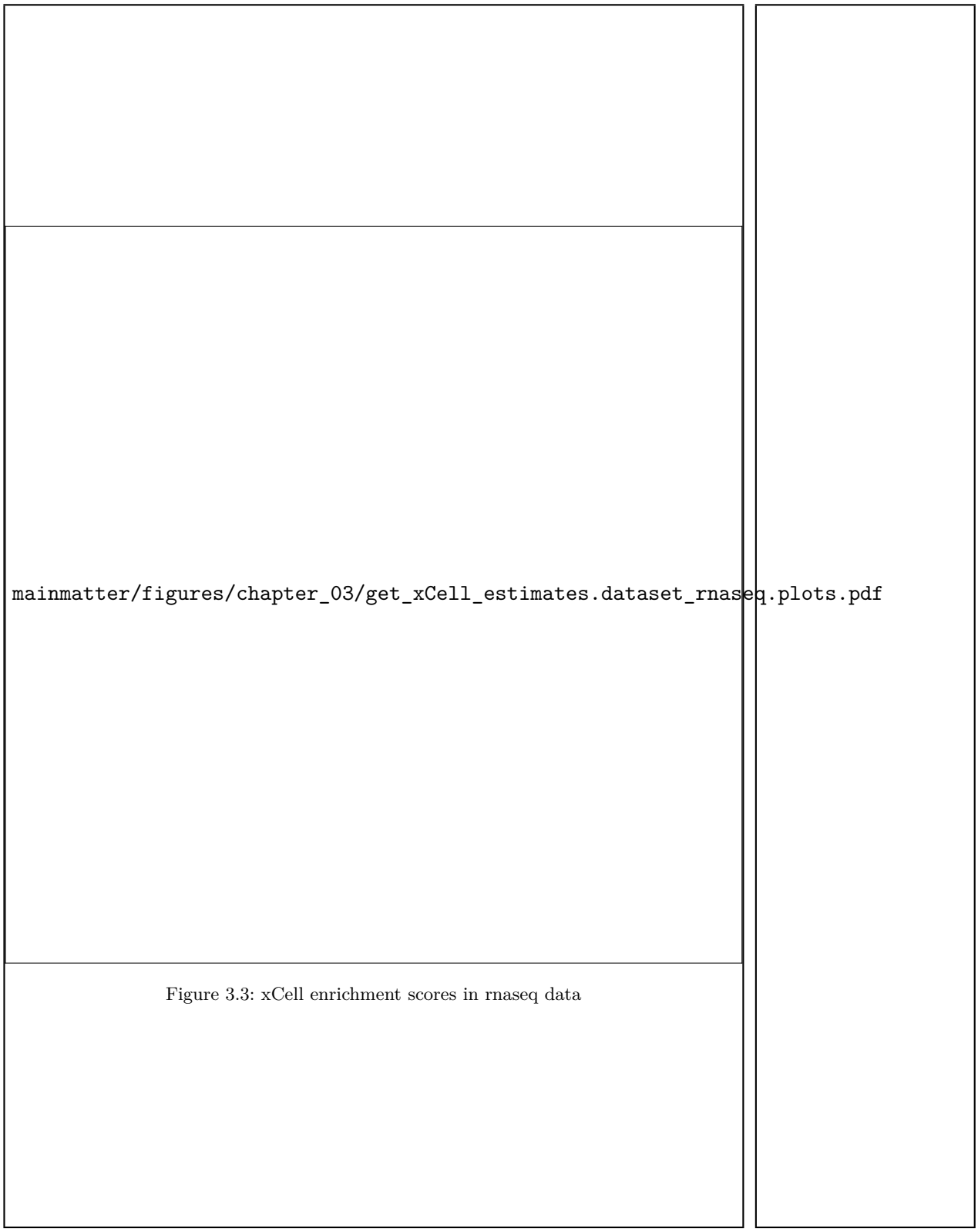
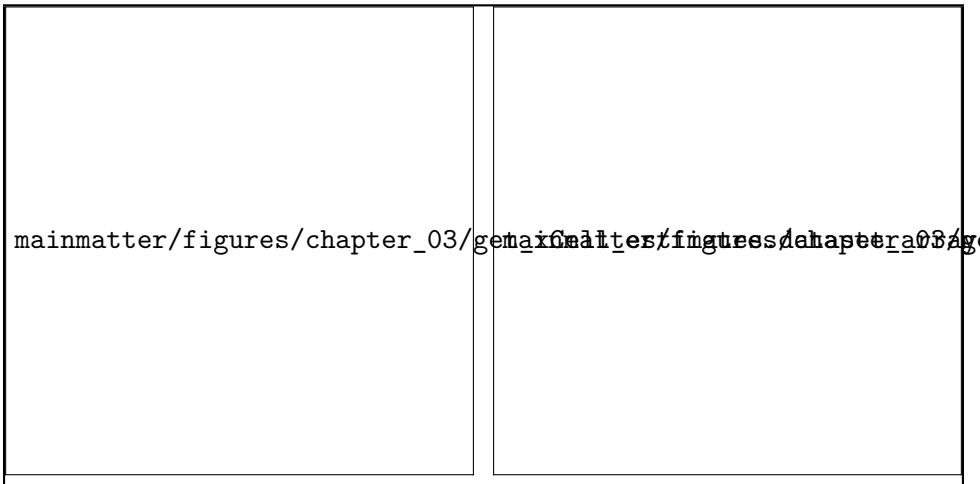


Figure 3.3: xCell enrichment scores in rnaseq data

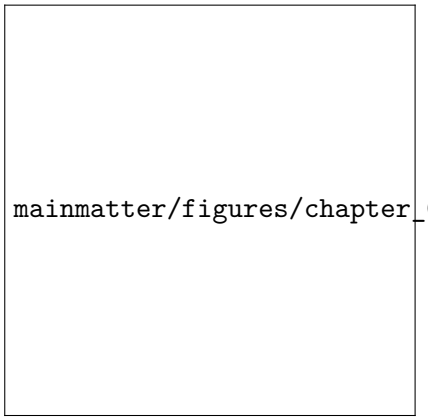


(a) array (b) rnaseq

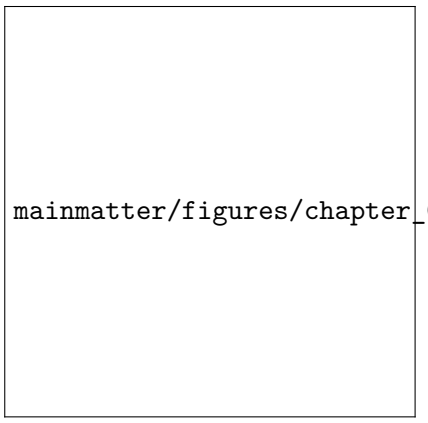
Figure 3.4: xCell cos2 contributions

fects. Given known covariates (intercept, sex, four genotype PCs from section representing ancestry, and the three xCell scores estimated above), PEER estimates additional hidden factors that explain variation in expression matrix. Factors are assumed to be unmeasured confounders that have global effects on a large fraction of genes, whereas a cis-eQTL will typically only have local effects, so factors should not interfere with the genotype term for the purposes of cis-eQTL mapping, but should soak up some of residual variation, hence 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.

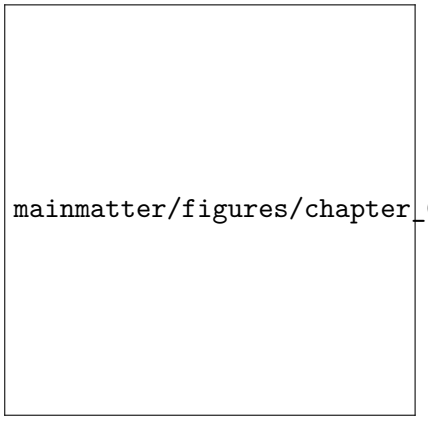
Correlating the estimated factors to a larger set of known covariates reveals many correlations with xCell estimates, indicating that cell type abundance does indeed have substantial global effects on the expression matrix. There is little correlation with known array or RNA-seq batch effects, indicating ComBat did an adequate job of removing batch- and platform-dependent global effects on expression (Fig. 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, which given the severity of the batch effect in this dataset (e Fig. 2.14), may be insufficient[48].



(a) mono



(b) nk



(c) plasma

Figure 3.5: xCell vs facs int

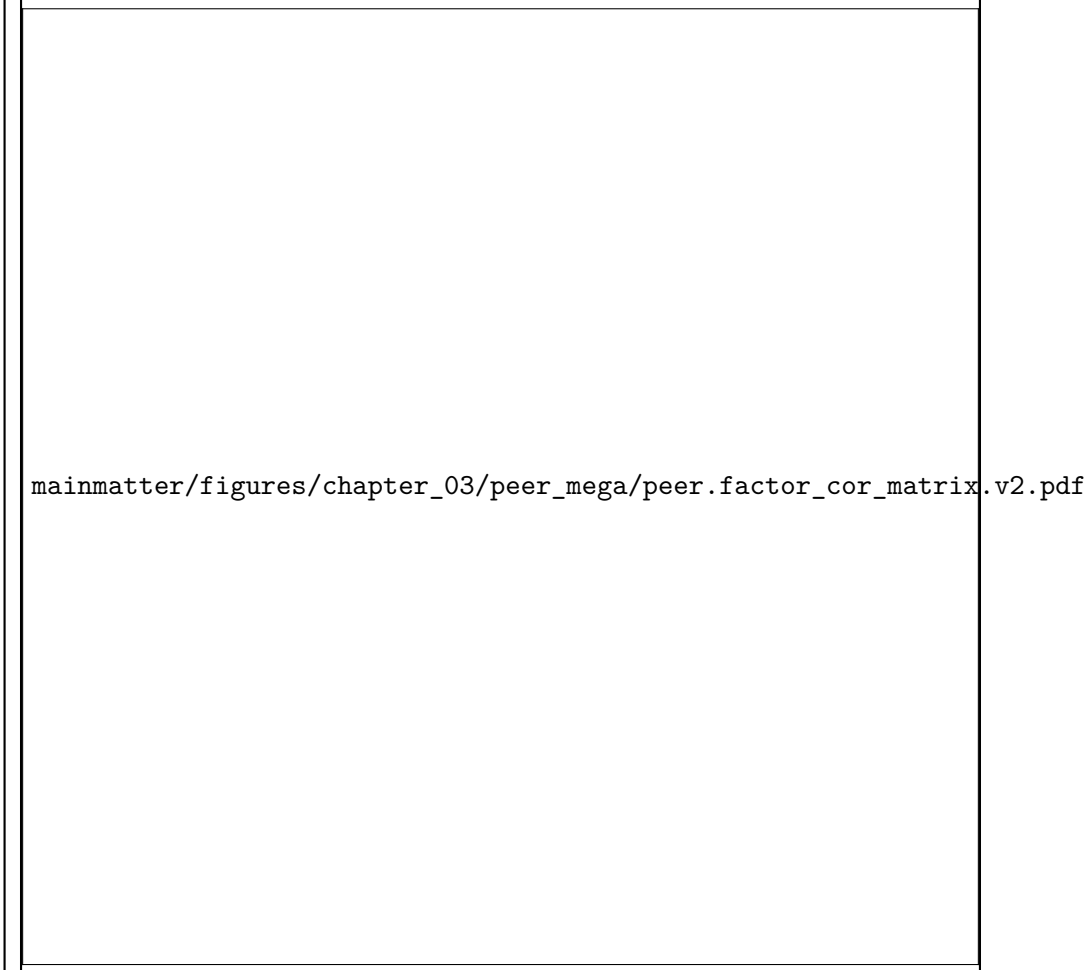


Figure 3.6: Note that PEER factors are not constrained to be orthogonal, so correlations to the provided known factors are expected.

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[31].
I map eQTLs within each timepoint using `LIMIX`[95], which implements efficient univariate and multivariate LMMs with one or more random effects.

Imputed genotype probabilities were converted to alternate allele dosages using bcftools (1.7-1-ge07034a). Variants with sample allele count (AC) <= 15 within each timepoint were excluded. X chromosome variants were excluded, as the number of copies differ between males and females, and X-inactivation makes it difficult to determine the active allele `https://www.nature.com/articles/ng.467`, so sex-specific methods are required `http://www.biomedcentral.com/1471-2105/15/392`.

xchrom

At each of 13126 autosomal genes, at all cis-variants within within +-1 Mb of the transcription start site (TSS), I fit the following model to map eQTL:

$$Y = 1 + sex + \sum_{i=1}^{4} PC_i + \sum^{3} xCell + \sum_{i=1}^{k} PC_i + G + \mathbf{u} + \epsilon \qquad (3.1)$$

lift proper vector notation from limix

where .

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 just 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 (Fig. 3.7).

3.2.8 Joint eQTL analysis across timepoints

snps only?

*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`https://onlinelibrary.wiley.com/doi/abs/10.1002/gepi.20607`.

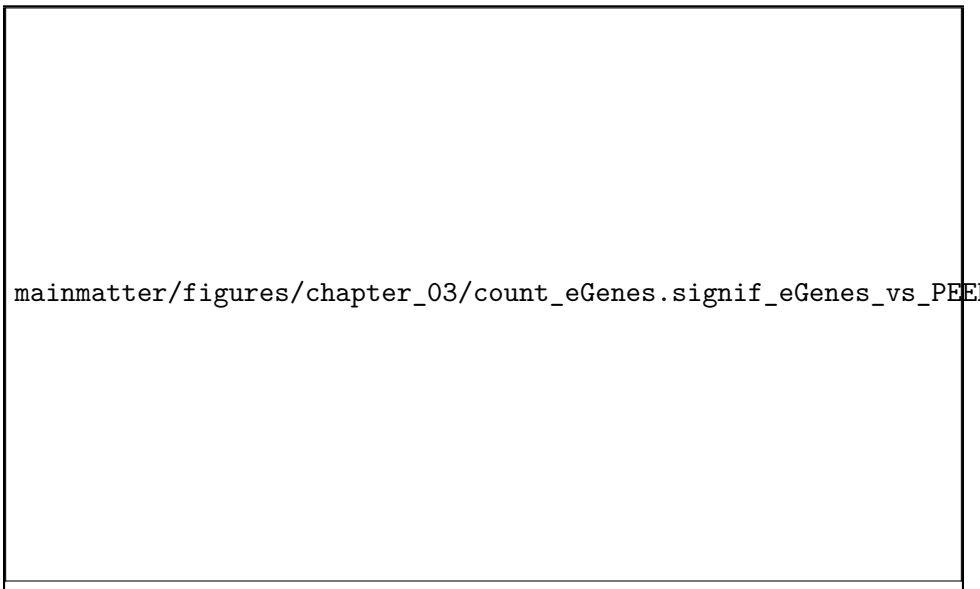


Figure 3.7: optimlise

Joint analysis was conducted with mashr, at 40197618 gene-variant pairs (tests) for which summary statistics from within timepoint mapping were available in all three timepoint conditions.

eQTL

The mashr model incorporates both canonical (e.g. the identity matrix) and data-driven covariance matrices to represent patterns of effects across conditions (in this case, 3 x 3 matrices). Data-driven covariance matrices are derived by dimension reduction 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 (Fig. 3.8, then further filtered to only nominally significant tests.

The mash model was trained on a random subset of 200000 tests, using the Exchangeable Z-scores model. The correlation of null tests between conditions, important 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.

move lfsr to dge chapter

Figure 3.8: optimlise, sample 10k

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 tested, I selected the variant with the lowest lfsr in any condition, breaking ties by highest imputation INFO, highest MAF, shortest distance to the TSS. 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 `http://www.stat.columbia.edu/~gelman/research/unpublished/signif3.pdf`.[82] 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. Effects are also only compared if at least one of the two effects have lfsr < 0.05, to avoid sharing being driven by null effects. I combine this approach with the beta-comparison approach `https://psycnet.apa.org/record/1995-27766-001` `https://doi.org/10.1111/j.1745-9125.1998.tb01268.x`[96] (applied to reQTLs by[97]), that also considers the standard error of both effects in the computation of a z score for the difference:

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

The posterior pairwise covariance of effects for test $\sigma^2(x,y)$ is difficult to estimate, so here I assume $\sigma^2(x,y) = 0$, a generally conservative assumption if effects with opposite signs between conditions are generally rare. Unlike a test for difference implemented using a genotype x condition interaction term in a joint regression model, homoscedasticity of errors is not assumed for all conditions `https://psycnet.apa.org/record/1995-27766-001`. The z score can be compared to a standard normal to obtain a nominal Z-test p value for the difference in betas between each pair of conditions, at each gene's lead variant. I use nominal p value < 0.05 as a heuristic threshold to define reQTL effects that are interesting (like the mashr recommended 2-fold threshold), rather than a formal measure of significant difference.

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 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 for shared eQTLs at moderately stringent lfsr thresholds up to 10^{-5} (Fig. 3.9). Past this, as the π_1 procedure assumes a well-behaved p value distribution in $[0, 1]$, reliability declines due to the 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 compared to the RNA-seq-only analysis, with similar replication, I assume this represents is due to the power advantage from having larger a sample size, rather than technical effects from merging the expression data.

get exact numbers, roughly 50 vs 30pc

3.2.11 Genotype interactions with non-timepoint predictors

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 slope of the eQTL will represent an average across the abundance range for that cell type; one can not 'correct' 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[90,97–99], 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

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

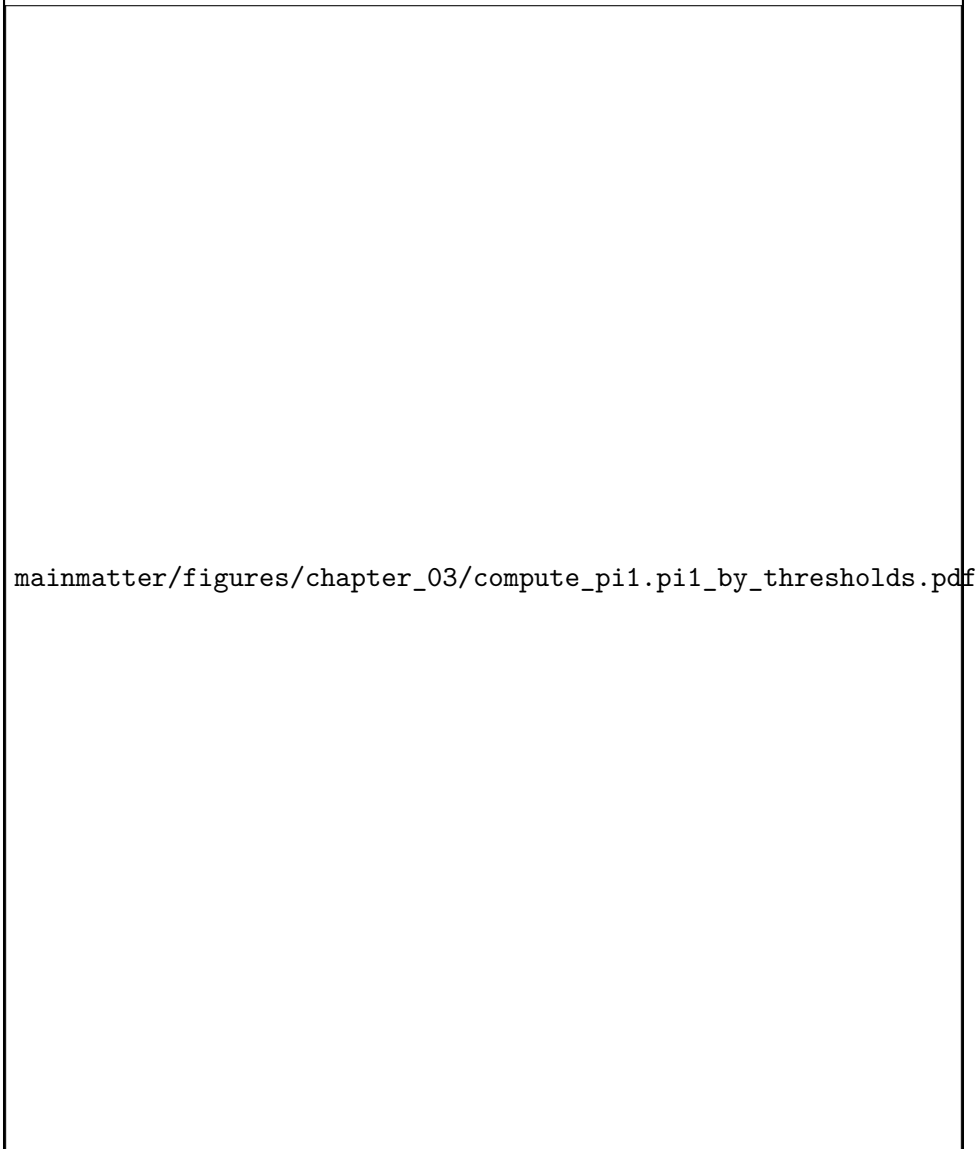

mainmatter/figures/chapter_03/compute_pi1.pi1_by_thresholds.pdf

Figure 3.9

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[99,100]. 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 choose to test for interactions only at the lead eQTL variant for each gene with a significant main eQTL, then apply the Benjamini-Hochberg (BH) false discovery rate (FDR) used by others[97,99].

eQTL models interactions between genotype and other predictors were fit using `lme4qtl`. The model specification is as in , with the addition of . Significance is assessed using the likelihood-ratio test versus the nested model with no interaction terms.

above section

interactions with cel type-/platform

note here that although peer is correlated with xcell, interactions are only formed with 3, so the interaction term can be interpreted per unit of genotype increase at e.g. mono=0

3.3 Results

3.3.1 Mapping reQTLs to Pandemix vaccination

Within each timepoint condition (day 0 pre-vaccination, day 1, and day 7), cis-eQTLs (±1Mb 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 lfsr < 0.05, 6887/13570 genes (50.75 %) were eGenes (genes with a significant eQTL) in at least one timepoint. The most significant eQTL variant across all timepoints was selected as the lead variant for each eGene. reQTLs were defined by comparing the effect size of this lead eQTL between each pair of timepoints using beta-comparison approach. Most eQTLs were shared across timepoints; 1154/6887 (16.76 %) eQTLs were classified as significant reQTLs (nominal p difference < 0.05).

Fig. 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, at 646/1427 eGenes, the effect size at day 0 does not differ significantly when compared to day 1 or day 7. The most significant eQTLs with the highest abs(z) in any timepoint are shared between timepoints, highlighting the power advantage for mapping shared effects granted by joint analysis.

upset has changed

Figure 3.10

3.3.2 Characterising reQTLs post-vaccination

I focus on eQTLs that are significant post-vaccination, and explain more variation in expression post-vaccination, as the converse may be caused by greater power at day 0 rather than being a result of vaccine stimulation. Many of the reQTL that satisfy this criteria have opposite effects pre- and post-vaccination (Fig. 3.11)— as lfsr quantifies uncertainty in the sign of the effect, I do not compare the sign unless the lfsr < 0.05 at day 0 also. Shared eQTLs are enriched close to the TSS, and reQTLs are distributed across the cis- window.

Gene set enrichment analysis on eGenes for these sets of reQTLs at day 1 (68 eGenes) and day 7 (226 eGenes) did not detect any significant enrichments (`gprofiler2`, g:SCS adjusted $p < 0.05$).

The strongest reQTL at day 1 was for *ADCY3* (p difference $= 8.677 \times 10^{-6}$, BH FDR $= 0.1177$), where the reQTL variant explained approximately 1.86 % of expression variation at day 0, increasing to 14.08 % at day 1 (Fig. 3.12). At day 7 the strongest reQTL was at *SH2D4A* (p difference $= 1.370 \times 10^{-6}$, BH FDR $= 0.017\,48$). Here, the reQTL variant explained similar amounts of expression variation at day 0 (8.23 %) and day 7 (8.96 %), with opposite directions of effect (Fig. 3.13).

Both *ADCY3* and *SH2D4A* have moderately high percentile expression at all timepoints, and are not differentially expressed post-vaccination. Overall, comparing reQTLs to genes without reQTL, they were less likely be differentially expressed post-vaccination at day 1(26.50 % vs. 42.27 %, Fisher's test $p < 2.200 \times 10^{-16}$), and no significant difference was observed at day 7 (2.20 % vs. 1.37 %, Fisher's test $p = 0.050\,88$). Only 5/68 (13.24 %) genes with reQTLs that explain more variation at day 1 were upregulated at day 1 vs. day 0; 5/226 (2.21 %) 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. Standardised `xCell` enrichment scores were used to approximate abundance of 7 PBMC cell types from the expression data. After pruning highly correlated cell types, scores for monocytes, NK cells and plasma cells remained. Within-timepoint full eQTL models including the genotype main effect, the

put pve formula in methods, include point that pve norms to var(y), so can compare between timepoints and gene

lets hope they are not all false positives

could put in reqtl ranked [redacted]erno enrichments here

could put in reqtl gosts here

change all these numbers, remove the pve requirements, since equal and opp is important

can use alpha for reqtl status and color for dge status instead

align d0 is plus

cahnge numbers

move this up to model

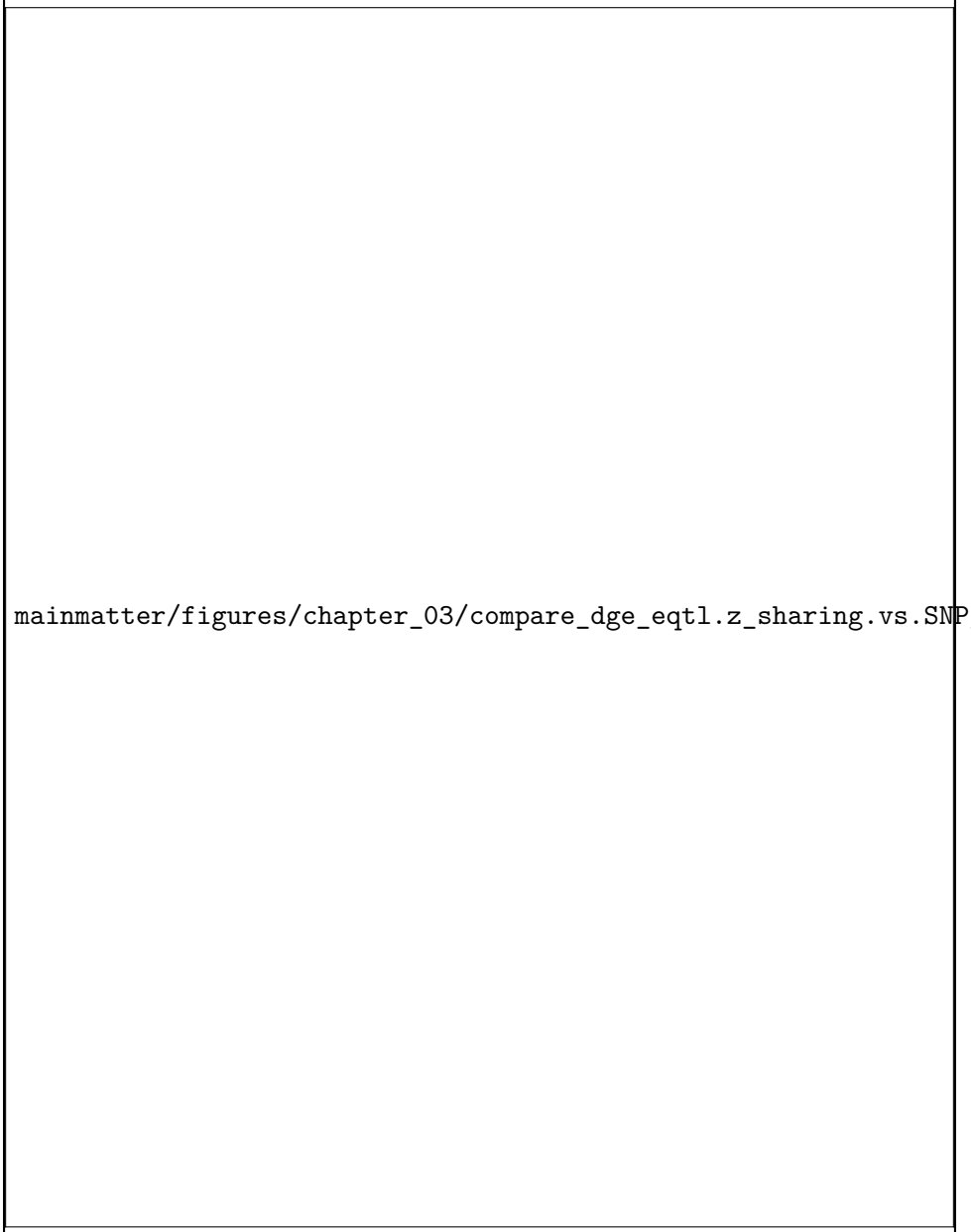

mainmatter/figures/chapter_03/compare_dge_eqtl.z_sharing.vs.SNP_gene_TSS_dist.p

Figure 3.11

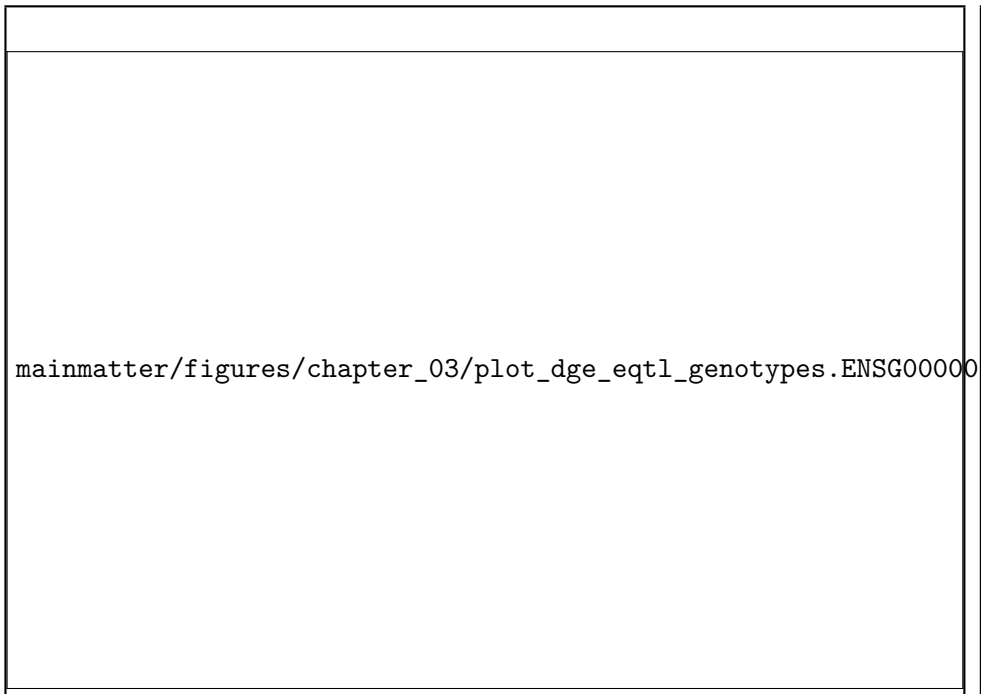


Figure 3.12

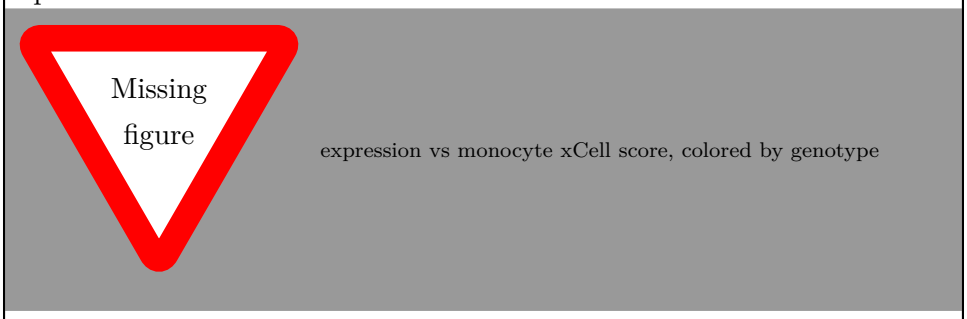
Figure 3.13

three cell type abundances, and three cell type-genotype interaction terms, were fit using `lme4qtl`, 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) For *ADCY3*, at day 1 post-vaccination, the full model had significantly better model fit ($chisq(3) = 26.290769$, likelihood ratio test (LRT) BH FDR $= 9.540 \times 10^{-5}$). Although the genotype effect size was 0.2560 (SE = 0.033 39) in the nested model, the estimate in the full model was −0.007 217 (0.066 56); with the three cell type-genotype interaction term estimates being: monocyte 0.2130 (0.048 98), NK cells −0.009 195 (0.044 70), and plasma cells 0.016 15 (0.066 33). This indicates the eQTL effect is actually driven largely by the monocyte abundance; in the case when monocyte abundance is zero (representing an average abundance across all samples, as scores are standardised), the effect of increasing genotype dosage on *ADCY3* expression is minimal.

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

this is probably what tables are for



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

- 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 (Fig. 3.14) ?!
  - Biases from ethnicity-derived differences in LD?

  - Also, priors need tuning?

- Fine mapped to rs13407914 (PP = 1), an intronic variant 45064 bp downstream of the TSS.

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

3.4 Discussion

add obesity

leadin

eQTL were detected for 50.75 % of genes in at least one condition. Each method for determining reQTL has it's own biases. 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. but these dampening effects are a mix of same sign and smaller magnitude, and opposite sign effect, which may represent distinct molecular mechanisms[101]. I chose a beta-comparison approach, defining reQTL strength as the difference in effect size between timepoints. Most eQTL are shared between conditions, with 16.76 % of lead eQTL being reQTL that differ in effect size between timepoints.

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

if rank by pve, put it here

Multiple independent eQTLs are present for a large fraction of eGenes. 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 especially problematic, as the effective sample size for shared eQTLs is usually large due to borrowing of information across conditions. In studies that performed step-wise conditioning, secondary signals are more distal, more likely to be enriched in enhancers rather than promoters, and more context-specific. 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, that these changes are mostly tied to innate immune activation , and that innate immunity is under stronger genetic control than adaptive immunity[102], the larger number of reQTLs detected at day 7 was unexpected. Opposite sign effects among reQTL post-vaccination were common: 39/88 at day 1, 211/269 at day 7. Prevalance of opposite sign effects between pairs of conditions has been previously described in multi-tissue studies. In[103], the proportion of opposite sign effects as a percentage of all eGenes was 7.40 % (48 tissues); in HIRD, I find 39/6887 (0.57 %) at day

sectionref

sectionref

but why are there so many at day 7?


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

Figure 3.14

1, and 211/6887 (3.06 %) at day 7. In[101], the proportion of opposite sign effects as percentage of all reQTLs was 4.40 % (5 tissues); in HIRD, I find 39/819 (4.76 %) at day 1, and 211/1002 (21.06 %) at day 7. The enrichment of opposite sign effects in HIRD is also most apparent at day 7. An approach for validating these opposite sign reQTL using the existing HIRD RNA-seq data is allele-specific expression (ASE) (e.g.[104]), where one would expect true opposite sign reQTL effects would also be recapitulated as opposite directions of expression imbalance.

but why are my reQTLs opposite? consult fu

The strongest reQTL detected at day 1 was *ADCY3*, a membrane-bound enzyme that catalyses the conversion of ATP to the second messenger cAMP `https://onlinelibrary.wiley.com/doi/full/10.1111/obr.12430`. genome-wide association study (GWAS) have identified *ADCY3* as a candidate gene for diseases such as obesity `https://onlinelibrary.wiley.com/doi/full/10.1111/obr.12430` and IBD `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4915781/`. *ADCY3* has been identified as a reQTL in multiple studies involving stimulated blood immune cells: in PBMC 24h post-infection with rhinovirus[105], in whole blood *in vivo* day 1 after vaccination with seasonal trivalent inactivated influenza vaccine (TIV)[106], and in whole blood after stimulation with M. leprae antigen for 26-32 h[107]. The effect is likely a consequence of general innate 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[22] 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 (which correlate with xCell scores) as fixed effects in the model, and also consider 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 `lme4qtl`, I used a two-step approach, testing only for interactions at significant lead reQTL. This means that the analysis addresses only whether reQTLs that can be detected based on only the main effect, may be driven by cell type interactions. At /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). The cell type interaction analysis at the day 1 *ADCY3* reQTL shows the genetic effect is mainly attributed to the monocyte score-genotype interaction term, which further supports the hypothesis that it is monocyte-specific.

Considering FACS measurements as a gold standard, the xCell scores used above were only moderately reliable. 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 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 also lower when applied to the expression matrix for a stimulated state, as the enrichment method can not distinguish differential expression of signature genes due to stimulation from actual changes in cell abundance. A custom signature matrix can be used for xCell, but this would need to be drawn from an independent study under the same stimulation conditions as HIRD, and would not solve the issue of coupled DE and cell abundance.

A pressing question still 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[90,97], 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 `https://doi.org/10.1371/journal.pgen.1004857`, and many immune cells including monocytes, have cell type specific TFs `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5156548/`. Cell type

specific expression of different TFs form a model that can explain magnifying, dampening and opposite reQTL effects; for example, opposite effects can result from TFs for the same gene, that are activating in one cell type and suppressive in another[101]. There is evidence that TF activity is important for *in vivo* immune reQTL:[105] found rhinovirus reQTLs were ENCODE ChIP-seq peaks for the TFs *STAT1* and *STAT2*, and[90] 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 raised in the previous chapter, that common genetic variation may explain some variation in antibody response to Pandemrix. I have indirectly demonstrated the effect of genetic variation on expression response by identifying reQTLs with differing effect size between timepoints, and some of these reQTLs will undoubtedly affect genes whose expression correlates with antibody response, but for proper causal inference, a mediation analysis is required that formally tests the effect of genotype on antibody response, through the intermediate phenotype of gene expression.[106] attempted this, but concluded that they had insufficient power with a greater sample size than HIRD, and a comparable study design assessing response to seasonal TIV. The HIRD cohort is also too small for a direct GWAS for antibody response. A suitable approach for prioritising reQTL that contribute to the antibody response to Pandemrix may be to colocalise with existing GWAS summary statistics from a separate cohort, ideally antibody response to another adjuvanted, inactivated influenza vaccine.

overall summary

It is difficult to make any conclusions regarding the effects on antibody response or vaccine efficacy.

*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

Response to live attenuated rotavirus vaccine (Rotarix) in Vietnamese infants

4.1 Introduction

Summary

Rotavirus vaccine efficacy is lower in LMICs than EU and NA. Protective response to many vaccines is linked with genetic variation. Hypothesis: difference in efficacy is due to differences in genetic variation.

Aim: identify genetic and transcriptomic markers associated with Rotarix protective response primary outcome will be Rotarix vaccine failure events secondary outcomes will be antibody responses and genotypic characterization of the infection virus in Rotarix failure events

4.1.1 The genetics of vaccine response in early life

4.1.2 Rotavirus and rotarix in Vietnam

4.1.3 Known factors that affect rotavirus vaccine efficacy

4.2 Methods

4.2.1 RNA-seq data generation

Stranded RNAseq AUTO with Globin Depletion (>47 samples) uses the NEB Ultra II directional RNA library kit for the poly(A) pulldown, fragmentation, 1st and 2nd strand synthesis and the flowing cDNA library prep (with some minor tweaks e.g. at during the PCR we use kapa HiFi not NEB's Q5 polymerase). Between the poly (A) pulldown and the fragmentation we use a kapa globin depletion kit (it's very similar to their riboerase kit but the rRNA probes are swapped for globin ones).

4.2.2 Genotyping

We will also use the SNP data to accurately impute ABO blood groups and secretor status.

4.3 Results

Transcriptomic response to rotavirus vaccination (pre- vs. post-, prime vs. boost dose, responders vs. non-responders)

Genetic contribution to transcriptomic response

4.4 Discussion

Chapter 5

multiPANTS

5.1 Introduction

Why do some people not respond?

Explore time series transcriptomic

Multilevel model where individual is a RE, Find out optimal spline degree. Then work out if genetics changes trajectories for any gene i.e. DGE models with a snp as predictor First need to eQTL scan in general with mashr and find the snps in the most reQTLish genes, since this modelling is probably expensive

Creating composite features to conduct genetic associations on.

Identifying signatures of response.

5.2 Methods

Responder analysis is a 4-fold loser. It fails to give the needed clinical interpretation, has poor statistical properties, increases the cost of studies because binary outcomes require significantly higher sample sizes, and raises ethical issues because more patients are required to be randomized than would be needed were the endpoint and analysis to be fully efficient.

immunomods

In the IFX+ADA cohort, DE PR vs PNR baseline PR vs PNR and w14

n patients with data for each number of visits

5.2.1 Covariates to use

Sex Age BMI Age of Onset Crohn's Surgery Ever Immunomodulator Current Smoker PCA Proportions of the 6 cell types: CD4+ T cells, CD8+ T cells, B cells, NK cells, monocytes, and granulocytes

5.2.2 reQTL

ANCOVA vs repeated measures vs mixed model

5.3 Results

5.4 Discussion

Chapter 6

Discussion

Tie ch 2 to 3 using baseline predictors?

Limitations, and the perfect study.

A response eqtl is not always a response eqtl

More timepoints Allele-specific expression changes dynamically during T cell activation in HLA and other autoimmune loci

More conditions e.g. 250 e.g. StructLMM Identifies eQTLs with GxE, where the number of environments in E is large (modelled as a random effect)

as datasets and conditions get larger, proportion of eGenes is going to be 100pc, then the question is what are the most relevant ones

Era of single cell. 1st Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs https://www.nature.com/articles/s41588-018-0089-9

"Single-cell eQTLGen Consortium: a personalized understanding of disease" https://arxiv.org/abs/1909.12550

Optimal design of single-cell RNA sequencing experiments for cell-type-specific eQTL analysis https://www.biorxiv.org/content/biorxiv/early/2019/09/12/766972.full.pdf

Single-cell genomic approaches for developing the next generation of immunotherapies Ido Yofe, Rony Dahan and Ido Amit

reQTL detection: bulk, sorted, sc current sc will only detect highly expressed genes

Cost-effectiveness and clinical implementation

if you can identify NRs, what are you going to do about it?

Deep phenotyping

disease specific biobanks e.g. ibd bioresource/predicct

unification immunology and vaccine dev: deep phenotyping, small cohorts achieved -> larger cohorts human genetics and gwas: large cohorts achieved -> deeper phenotyping

Appendix A

Supplementary Materials

A.1 Chapter 2

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

A.2 Chapter 3

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque

cursus luctus mauris.

A.3 Chapter 4

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Bibliography

1. Verma, A. & Ritchie, M. D. Current Scope and Challenges in Phenome-Wide Association Studies. *Current Epidemiology Reports* **4,** 321–329. doi:10.1007/s40471-017-0127-7 (Dec. 2017).
2. 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 (Jan. 24, 2012).
3. 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 (Mar. 7, 2014).
4. Fairfax, B. P. & Knight, J. C. Genetics of Gene Expression in Immunity to Infection. *Current Opinion in Immunology* **30,** 63–71. doi:10.1016/j.coi.2014.07.001 (Oct. 2014).
5. Krammer, F. *et al.* Influenza. *Nature Reviews Disease Primers* **4.** doi:10.1038/s41572-018-0002-y (Dec. 2018).
6. Houser, K. & Subbarao, K. Influenza Vaccines: Challenges and Solutions. *Cell Host & Microbe* **17,** 295–300. doi:10.1016/j.chom.2015.02.012 (Mar. 2015).
7. 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 (Dec. 2018).
8. 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 (Oct. 2011).

9. 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.

10. Plotkin, S. A. Correlates of Protection Induced by Vaccination. *Clinical and Vaccine Immunology* **17,** 1055–1065. doi:10.1128/CVI.00131-10 (July 2010).

11. 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 (Feb. 2013).

12. 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).

13. 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).

14. 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 (Jan. 2019).

15. 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 (Apr. 2010).

16. 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 (Apr. 2011).

17. Nakaya, H. I. *et al.* Systems Biology of Vaccination for Seasonal Influenza in Humans. *Nature Immunology* **12,** 786–795. doi:10.1038/ni.2067 (July 10, 2011).

18. 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 (Jan. 2014).

19. 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 (Mar. 2012).

20. Wilkins, A. L. *et al.* AS03- and MF59-Adjuvanted Influenza Vaccines in Children. *Frontiers in Immunology* **8.** doi:10.3389/fimmu.2017.01760 (Dec. 13, 2017).

21. Tregoning, J. S., Russell, R. F. & Kinnear, E. Adjuvanted Influenza Vaccines. *Human Vaccines & Immunotherapeutics* **14,** 550–564. doi:10.1080/21645515.2017.1415684 (Mar. 4, 2018).

22. 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 (Jan. 4, 2016).

23. Furman, D. *et al.* Apoptosis and Other Immune Biomarkers Predict Influenza Vaccine Responsiveness. *Molecular Systems Biology* **9,** 659. doi:10.1038/msb.2013.15 (Apr. 16, 2013).

24. 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 (Apr. 10, 2014).

25. 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 (Dec. 2015).

26. 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 (Aug. 25, 2017).

27. Cohen, J. The Cost of Dichotomization. *Applied Psychological Measurement* **7,** 249–253. doi:10.1177/014662168300700301 (June 1983).

28. Fedorov, V., Mannino, F. & Zhang, R. Consequences of Dichotomization. *Pharmaceutical Statistics* **8,** 50–61. doi:10.1002/pst.331 (Jan. 2009).

29. Food and Drug Administration. *Guidance for Industry: Clinical Data Needed to Support the Licensure of Pandemic Influenza Vaccines* (Jan. 6, 2007), 20.

30. 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 (Aug. 2006).

31. 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 (July 17, 2014).

32. Brown, B. C., Bray, N. L. & Pachter, L. Expression Reflects Population Structure. doi:10.1101/364448 (July 8, 2018).

33. 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 (Oct. 1, 2015).

34. 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 (Oct. 1, 2016).

35. 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 (Apr. 6, 2017).

36. 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 (Feb. 1, 2014).

37. 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 (Feb. 29, 2016).

38. 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 (Dec. 2018).

39. 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).

40. 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 (Aug. 2, 2016).

41. 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 July 1, 2002).

42. 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).

43. 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 (Feb. 2006).

44. 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 (June 30, 2015).

45. 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 (Jan. 1, 2007).

46. 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 (Feb. 28, 2011).

47. Espín-Pérez, A., Portier, C., Chadeau-Hyam, M., van Veldhoven, K., Kleinjans, 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 (Aug. 30, 2018).

48. 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 (Dec. 2018).

49. 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 Aug. 13, 2015).

50. Evans, C., Hardin, J. & Stoebel, 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 (Sept. 28, 2018).

51. 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 (Jan. 1, 2010).

52. 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).

53. 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 (Apr. 20, 2015).

54. 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 (Dec. 2013).

55. 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).

56. 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 (Apr. 2010).

57. Röver, C. Bayesian Random-Effects Meta-Analysis Using the Bayesmeta R Package (Nov. 23, 2017).

58. Bender, R. *et al.* Methods for Evidence Synthesis in the Case of Very Few Studies. *Research Synthesis Methods.* doi:10.1002/jrsm.1297 (Apr. 6, 2018).

59. 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 (July 20, 2015).

60. 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 (Mar. 2016).

61. 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 (Oct. 2013).

62. 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 (Mar. 2017).

63. 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 (Dec. 2019).

64. 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 (Sept. 2006).

65. 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 (Nov. 20, 2011).

66. 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: R. M. TURNER *ET AL. Statistics in Medicine* **34,** 984–998. doi:10.1002/sim.6381 (Mar. 15, 2015).

67. 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 (Dec. 30, 1996).

68. 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 (June 1, 2019).

69. Stephens, M. False Discovery Rates: A New Deal. *Biostatistics,* kxw041. doi:10.1093/biostatistics/kxw041 (Oct. 17, 2016).

70. Weiner 3rd, J. & Domaszewska, T. Tmod: An R Package for General and Multivariate Enrichment Analysis. doi:10.7287/peerj.preprints.2420v1 (2016).

71. 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 (Dec. 15, 2013).

72. 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 May 1, 2016).

73. 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 (Mar. 21, 2014).

74. 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 (Feb. 16, 2016).

75. Murphy, K. & Weaver, C. *Janeway's Immunobiology* 9th edition. 904 pp. (Garland Science/Taylor & Francis Group, LLC, New York, NY, 2016).

76. Vandiedonck, C. Genetic Association of Molecular Traits: A Help to Identify Causative Variants in Complex Diseases. *Clinical Genetics.* doi:10.1111/cge.13187 (Dec. 1, 2017).

77. 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 (July 7, 2011).

78. 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 (June 6, 2013).

79. Allison, P. D. Change Scores as Dependent Variables in Regression Analysis. *Sociological Methodology* **20,** 93. doi:10.2307/271083 (1990).

80. 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 (Dec. 2019).

81. 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 (May 9, 2013).

82. 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 (Nov. 26, 2018).

83. 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 (Dec. 2018).

84. 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 (July 5, 2013).

85. 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 (July 2010).

86. 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, June 27, 2018). doi:10.1201/9781315154084-27.

87. 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 (Oct. 2011).

88. 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 (Dec. 2012).

89. 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 (Sept. 2009).

90. 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 (Dec. 2018).

91. Kim-Hellmuth, S. *et al.* Cell Type Specific Genetic Regulation of Gene Expression across Human Tissues. *bioRxiv.* doi:10.1101/806117 (Oct. 17, 2019).

92. 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 (Dec. 2017).

93. 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 (Nov. 2016).

94. 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 (Feb. 16, 2012).

95. Lippert, C., Casale, F. P., Rakitsch, B. & Stegle, O. LIMIX: Genetic Analysis of Multiple Traits. doi:10.1101/003905 (May 22, 2014).

96. 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).

97. 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 (Dec. 2017).

98. 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 (May 8, 2015).

99. 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 (Mar. 25, 2016).

100. Kooperberg, C. & LeBlanc, M. Increasing the Power of Identifying Gene × Gene Interactions in Genome-Wide Association Studies. *Genetic Epidemiology* **32,** 255–263. doi:10.1002/gepi.20300 (Apr. 2008).

101. 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 (Jan. 19, 2012).

102. 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 (Feb. 23, 2018).

103. 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 (Nov. 2019).

104. 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 (Feb. 2016).

105. 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 (Apr. 13, 2015).

106. Franco, L. M. *et al.* Integrative Genomic Analysis of the Human Immune Response to Influenza Vaccination. *eLife* **2,** e00299. doi:10.7554/eLife.00299 (July 16, 2013).

107. 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 (Aug. 9, 2017).

List of Abbreviations

AC allele count

ASE allele-specific expression

BH Benjamini-Hochberg

BTM blood transcription module

CPM counts per million

DC dendritic cell

DGE differential gene expression

eQTL expression quantitative trait locus

FACS fluorescence-activated cell sorting

FC fold change

FDR false discovery rate

GWAS genome-wide association study

HA haemagglutinin

HAI haemagglutination inhibition

HIRD Human Immune Response Dynamics

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

LRT likelihood ratio test

MAF minor allele frequency

MANOVA multivariate analysis of variance

MN microneutralisation

NA neuraminidase

NK natural killer

PBMC peripheral blood mononuclear cell

PC principal component

PCA principal component analysis

REML restricted maximum likelihood

reQTL response expression quantitative trait locus

RNA-seq RNA-sequencing

SD standard deviation

TF transcription factor

TIV trivalent inactivated influenza vaccine

TMM trimmed mean of M-values

TRI titre response index

TSS transcription start site

spell-check

make sure package versions are in, and package names are monospace

add automatic rounding to x decimal places using num and sisetup

Todo list