

<title>

Benjamin Yu Hang Bai

2020-04-09 04:25:35+01:00

<dedication>

Abstract

<thesis abstract>

Acknowledgements

pipelines

oucru team

research assistants/research managers

family friends cuams churchill MCR, various badminton

stackexchange publication quality dialogue, model for future peer review?

Contents

List of Figures	xi
List of Tables	xiii
1 Introduction	1
1.1 A brief overview of genetic association for complex traits	1
1.1.1 Early days	1
1.1.2 The advent of GWAS	2
1.1.3 Narrowing the signal	2
1.1.4 Interpretation of genetic associations with molecular studies	3
1.1.5 So what? Translational directions [can cut this whole section]	4
1.2 The effects of genetic variation on expression: context is key	5
1.3 Immunity is a complex trait	5
1.3.1 Genetic factors affecting the healthy immune system	5
1.3.2 Genetic factors affecting immune response to challenge	6
1.3.2.1 Context-specific immune response eQTLs in vitro	6
1.3.2.2 <i>in vivo</i> response QTL mapping	7
1.4 Immune response to vaccination	7
1.4.1 Systems vaccinology: from empirical to rational vaccinology	7
1.4.2 Genetic factors affecting vaccine response	8
1.5 Immune response to biologic therapies	9
1.5.1 Genetic factors affecting biologic responses	9
1.6 Thesis overview	9

2 Transcriptomic response to influenza A (H1N1)pdm09 vaccine (Pandemrix)	11
2.1 Introduction	11
2.1.1 Influenza A (H1N1)pdm09 and Pandemrix	11
2.1.2 Immune response to influenza vaccines	12
2.1.3 Systems studies, and the Human Immune Response Dynamics (HIRD) study	12
2.1.4 Chapter summary	13
2.2 Methods	13
2.2.1 Existing HIRD study data and additional data	13
2.2.2 Computing baseline-adjusted measures of vaccine-induced antibody response	14
2.2.3 Genotype data generation	15
2.2.4 Genotype data preprocessing	15
2.2.5 Computing genotype PCs as covariates for ancestry	19
2.2.6 Genotype phasing and imputation	19
2.2.7 RNA-seq data generation	19
2.2.8 RNA-seq quantification and filtering	21
2.2.9 Array data preprocessing	23
2.2.10 Differential gene expression	25
2.2.10.1 Per-platform differential gene expression model	27
2.2.10.2 Choice of differential gene expression meta-analysis method	27
2.2.10.3 Prior for between-studies heterogeneity	28
2.2.10.4 Prior for effect size	29
2.2.10.5 Evaluation of priors	29
2.2.10.6 FDR	29
2.2.11 Gene set enrichment analysis	29
2.3 Results	31
2.3.1 Extensive global changes in expression after vaccination	31
2.3.2 Innate immune response at day 1 post-vaccination . .	31
2.3.3 Adaptive immune response at day 7 post-vaccination .	33
2.3.4 Expression signatures associated with antibody response	33
2.3.5 Identifying expression signatures for predicting antibody response	35
2.4 Discussion	35

3 Genetic factors affecting Pandemrix vaccine response	43
3.1 Introduction	43
3.1.1 Genetic factors affecting influenza vaccine response . .	44
3.1.2 Context-specific immune response QTLs for influenza vaccine response	44
3.1.3 Chapter summary	44
3.2 Methods	44
3.2.1 Genotype imputation	44
3.2.2 Estimation of cell type abundances	44
3.2.3 Mapping cis-eQTLs with LMM	45
3.2.3.1 Expression normalisation	46
3.2.3.2 Finding hidden confounders with PEER . . .	46
3.2.4 eQTL mapping with mixed models	46
3.2.5 eQTL meta-analysis	46
3.2.5.1 Joint mapping with mashr	47
3.2.6 Defining shared and response eQTLs	48
3.3 Results	48
3.3.1 Overview of eQTLs at each timepoint	48
3.3.1.1 Estimation of eQTL sharing	48
3.3.1.2 TODO Replication of shared eQTLs in whole blood	48
3.3.2 Characterising re-eQTLs at each timepoint	48
3.3.3 The mechanism of reQTLs	48
3.3.4 TODO Colocalisation of re-eQTLs with known context-specific immune QTLs	48
3.3.5 TODO Disruption of binding site motifs as a model for re-eQTLs	48
3.4 Discussion	49
3.4.1 DGE vs eqtl	49
3.4.2 limitations: The mechanism of reQTLs	49
3.4.3 Conditional eQTL effects	49
4 Response to live attenuated rotavirus vaccine (Rotarix) in Vietnamese infants	51
4.1 Introduction	51
4.1.1 The genetics of vaccine response in early life	52

4.1.2	Rotavirus and rotarix in Vietnam	52
4.1.3	Known factors that affect rotavirus vaccine efficacy	52
4.2	Methods	52
4.2.1	RNA-seq data generation	52
4.2.2	Genotyping	52
4.3	Results	52
4.4	Discussion	52
5	multiPANTS	53
5.1	Introduction	53
5.2	Methods	53
5.2.1	Covariates to use	53
5.3	Results	54
5.4	Discussion	54
6	Discussion	55
A	Supplementary Materials	57
A.1	Chapter 2	57
A.2	Chapter 3	57
A.3	Chapter 4	58
	Bibliography	59
	List of Abbreviations	67

List of Figures

2.1	Data types, timepoints, and sample sizes. Individuals were vaccinated after day 0 sampling. Vaccine-induced antibodies measured by haemagglutination inhibition (HAI) and microneutralisation (MN) assays. Array and RNA-sequencing (RNA-seq) gene expression measured in the peripheral blood mononuclear cell (PBMC) compartment.	14
2.2	How TRI corrects fold changes for baseline titre.	16
2.3	TRI vs platform	17
2.4	Sample filters for missingness vs heterozygosity rate.	18
2.5	HIRD samples (cyan) projected onto principal component (PC)1 and PC2 axes defined by PCA of HapMap samples. The first two PCs separate European (CEU, upper-right) from Asian (CHB and JPT, lower-right) and African (YRI, lower-left) populations.	20
2.6	FastQC sequence quality versus read position for Human Immune Response Dynamics (HIRD) RNA-seq samples. . .	21
2.7	FastQC sequence duplication levels for HIRD RNA-seq samples.	22
2.8	FastQC GC profile for HIRD RNA-seq samples.	22
2.9	Distribution of short ncRNA and globin counts as a proportion of total counts in RNA-seq samples.	24
2.10	Proportion of samples in which genes are detected with non-zero counts. Vertical line shows 5% threshold.	24
2.11	Distribution of gene expressions for sample before and after filtering low expression genes. Vertical line shown at log(0.5) counts per million (CPM).	25

2.12 First 2 expression PCs, expressions first standardized within platform to increase comparability, if not, even more variation would be driven by platform	26
2.13 Fold-change comparison between array and RNAseq for day 1 vs day 0.	30
2.14 Priors for tau for day 1 vs day 0 DGE meta-analysis.	30
2.15 Priors for mu for day 1 vs day 0 DGE meta-analysis.	31
2.16 Normalised gene expression for genes differentially expressed between any pair of timepoints ($\text{lfsr} < 0.05$, $ \text{FC} > 1.5$) across HIRD, clustered by gene. Distance metric: Manhattan.	32
2.17 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).	34
2.18 DGE effect sizes estimated in array vs RNA-seq. Significance colored by frequentist random effects meta-analysis $\text{FDR} < 0.05$	36
2.19 DGE effect sizes estimated in array vs RNA-seq. Significance colored by Bayesian random effects meta-analysis $\text{lfsr} < 0.05$	36
2.20 Transcriptomic modules enriched in genes with expression associated with antibody response (titre response index (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).	37

List of Tables

2.1	Sample descriptive statistics.	40
2.2	HIRD batch balance	41

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

- PheWAS¹
 - 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
- 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

- in the dreaded GxE interaction
 - "In genetics, context matters"
 - for both gwas, and molQTLs, context is key
- Architecture varies e.g. across cell type and tissues
 - tissue
 - cell type
 - interaction between cells *in vivo*
 - stimulation conditions
- 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 was², 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+ monocytes³) and stimulations (IFN γ and LPS⁴).

Take home messages: - reQTLs develop trans-effects on stimulation³
Overall, as the number of experimental systems and stimulations increases,

CHAPTER 1. INTRODUCTION IMMUNE RESPONSE TO VACCINATION

large number of eQTLs are only detected.

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 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 vulnerable 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 efficacy differences e.g. rotavirus] Particularly challenging populations for vaccination include the infants and elderly, pregnant, immuno compromised 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

1.4. IMMUNE RESPONSE TO VACCINES

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.2014.0146]. 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

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.].

Search for "variation in vaccine response genetics GA Poland" in google scholar

Genetics of adverse events e.g. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4845468/>

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)

Review paper on GWAS for vaccines mooney2013SystemsImmunogeneticsVaccines

CHAPTER 1. INTRODUCTION TO BIOLOGIC THERAPIES

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 (Pandemrix)

2.1 Introduction

2.1.1 Influenza A (H1N1)pdm09 and Pandemrix

- Basic seasonal flu biology
- Basic H1N1 biology
 - structure and life cycle.
 - relationship to other (seasonal) influenza viruses.
- The 2009 outbreak.
 - origins; timeline
- Vaccine development process in response to the outbreaks
 - Pandemrix was one of several vaccines licensed.
 - Efficacy, dosing: "...a single dose of monovalent 2009 H1N1 vaccine was recommended in adults, but young children were recommended to receive 2 doses (reviewed by [3••]). It is likely that a single dose was sufficient to induce immunity in adults because prior

CHAPTER 2. TRANSCRIPTOMIC RESPONSE TO INFLUENZA A

2.1. INTRODUCTION (H1N1)PDM09 VACCINE (PANDEMRIX)

exposure to seasonal H1N1 viruses had immunologically primed the population.”

- Inclusion of H1N1 strains into seasonal vaccines
 - * Later cohorts may have recall response to H1N1 from seasonal vaccination

2.1.2 Immune response to influenza vaccines

- classical immunological response to influenza vaccines
 - seasonal flu vaccines are a constant battle against flu virus antigenic drift
 - differ by type of vaccine
- correlates of protection
 - mention HAI/MN assays, do they measure IgM/IgG?
- Review influenza vaccine specific sysvacc papers (e.g. Nakaya’s papers)
 - mention main timepoints
 - inclu. prevaccination signatures paper, and

2.1.3 Systems studies, and the Human Immune Response Dynamics (HIRD) study

- Systems vaccinology of Pandemrix vaccine: Sobolev et al. 2016
 - Sobolev et al 2016 evaluated transcriptomic, cellular, antibody and adverse events after AS03-adjuvanted Pandemrix vaccination.
 - * Myeloid response similar to other unadjuvanted flu vaccines
 - * Early lymphoid response unlike other unadjuvanted vaccines
 - Knowns about the immune response to AS03
 - * Non responders had “reduced expression of genes associated with plasma cell development and antibody production at day 7”
 - * No consensus NR signatures at earlier timepoints day 0 or day 1 “many routes to failure”. One reason is variable baseline titres leading to variable trajectories of NR.

2.1.4 Chapter summary

- Rationale for our study:
 - Sobolev uses array transcriptomic data for a subset of individuals; we use RNAseq data for a larger number of individuals, which allows us to look at a larger number of genomic features, and conduct a meta-analysis.
 - Instead of the binary definition for responder/NR used by Sobolev, we use a continuous response measure, for increased power. This also lets us normalise for baseline titre and combine HAI and microneutralization assay values.
 - * can we find consensus, and importantly prevaccination signatures of response?
 - main aims are de between timepoint and r/nr
- Main conclusions
 - The overall pattern of innate response at d1, adaptive response at d7, agrees with Sobolev.
 - Based on our continuous Ab phenotype, we find consensus response signatures
 - * plasma cells and inflammatory response overall
 - * at each timepoint, d0, d1, d7 ...
 - Compare the d7 split to Sobolev
- Finally, turn our focus to prediction, i.e. going from R status as a predictor to a response variable.

2.2 Methods

2.2.1 Existing HIRD study data and additional data

The design of the **HIRD** study is described in⁵. 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 samples were taken on days -7, 0, 1, 7, 14 and 63. Serological responses were measured on days -7 and 63 using the **HAI** and **MN** assays, and various subsets of the cohort were

also profiled for serum cytokine levels (Luminex panel, days -7, 0, 1 and 7), immune cell counts (fluorescence-activated cell sorting, all days), and **PBMC** gene expression (microarray, days -7, 0, 1 and 7).

In addition to the existing data, array genotype data was 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*, due to lack of biological material for the microarray individuals from which RNA could be extracted for **RNA-seq**. An overview of datasets is shown in Fig. 2.1.

2.2.2 Computing baseline-adjusted measures of vaccine-induced antibody response

In⁵, Pandemrix responders were defined, according to clinical convention (e.g.⁶), as individuals with ≥ 4 -fold titre increases in either the **HAI** or **MN** assays. 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,⁵ 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^{7,8}. To address these concerns, I also computed the **TRI** as defined in Bucasas *et al.* [9]. For each assay, a linear regression was fit with the \log_2 fold change \log_2 day 63/day -7 as the response, and the \log_2 baseline titre \log_2 day -7

cite appropriate subfigures here

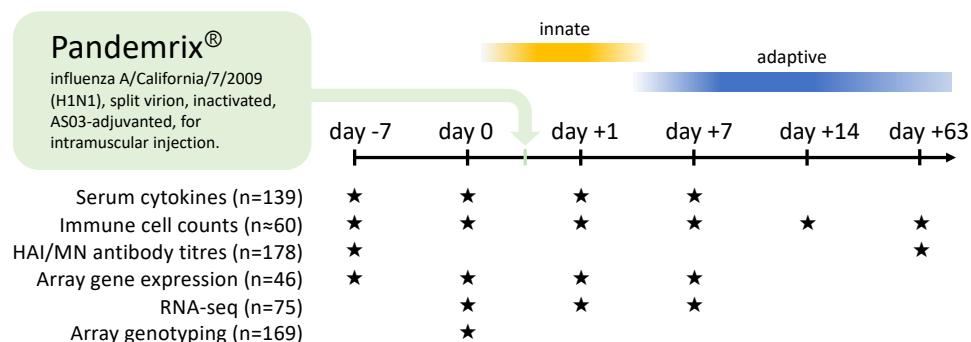


Figure 2.1: Data types, timepoints, and sample sizes. Individuals were vaccinated after day 0 sampling. Vaccine-induced antibodies measured by **HAI** and **MN** assays. Array and **RNA-seq** gene expression measured in the **PBMC** compartment.

as predictor. The residuals from the two regressions were each standardized to mean 0 and **standard deviation (SD)** 1, 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

Descriptive statistics for 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 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 in⁵.

2.2.3 Genotype data generation

[Add to collab note](#)

DNA was extracted from frozen blood using the Blood and Tissue DNeasy kit (Qiagen), and genotyping was performed using the Infinium CoreExome-24 BeadChip (Illumina). In total, 192 samples from 176 patients in the HIRD cohort were genotyped at 550601 markers, including replicate samples submitted for patients 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 exclude 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 pairwise kinship coefficient > 0.177 (first-degree relatives or closer), the sample with lower marker missingness was selected.

CHAPTER 2. TRANSCRIPTOMIC RESPONSE TO INFLUENZA A
2.2. METHODS (H1N1)PDM09 VACCINE (PANDEMRIX)

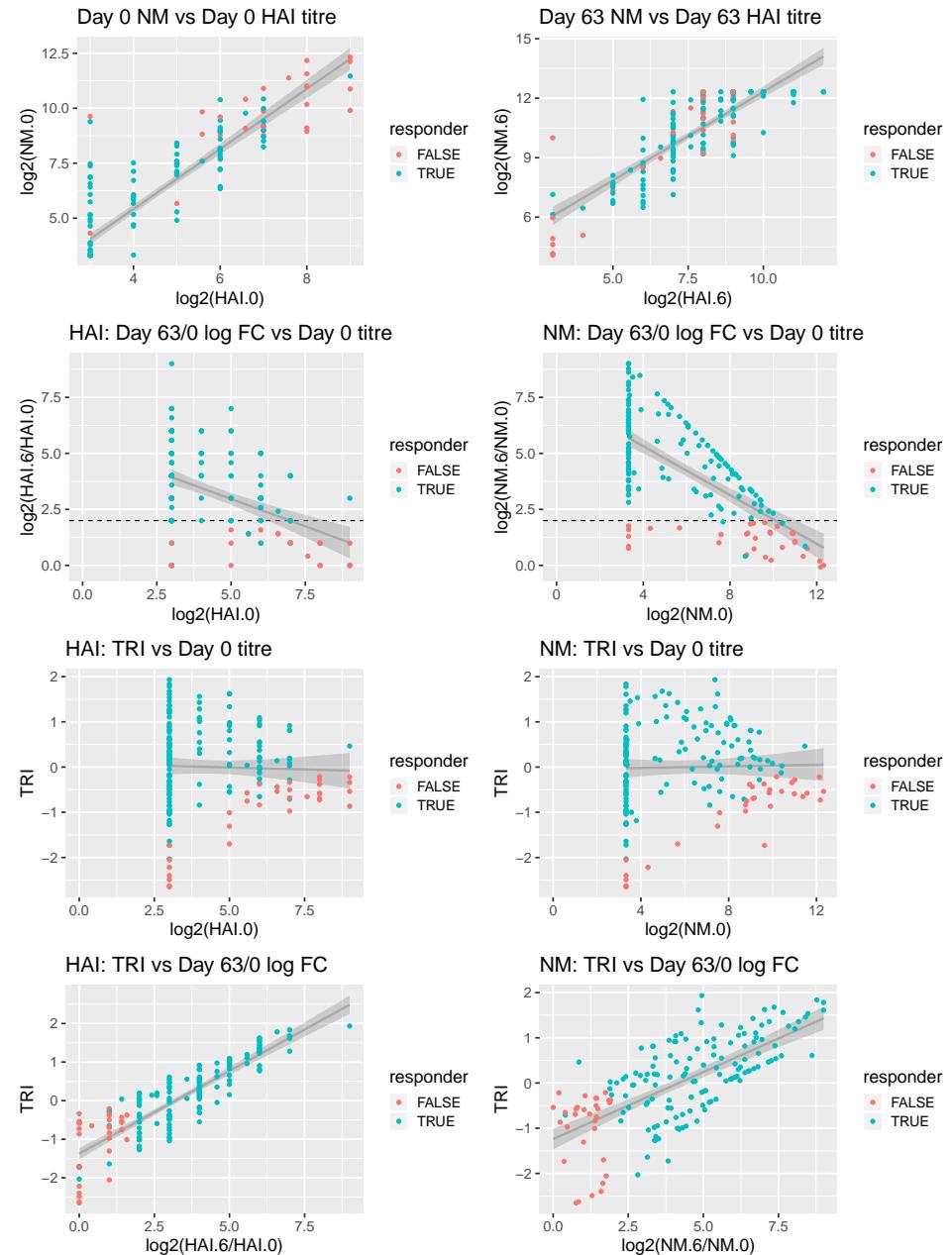


Figure 2.2: How TRI corrects fold changes for baseline titre.

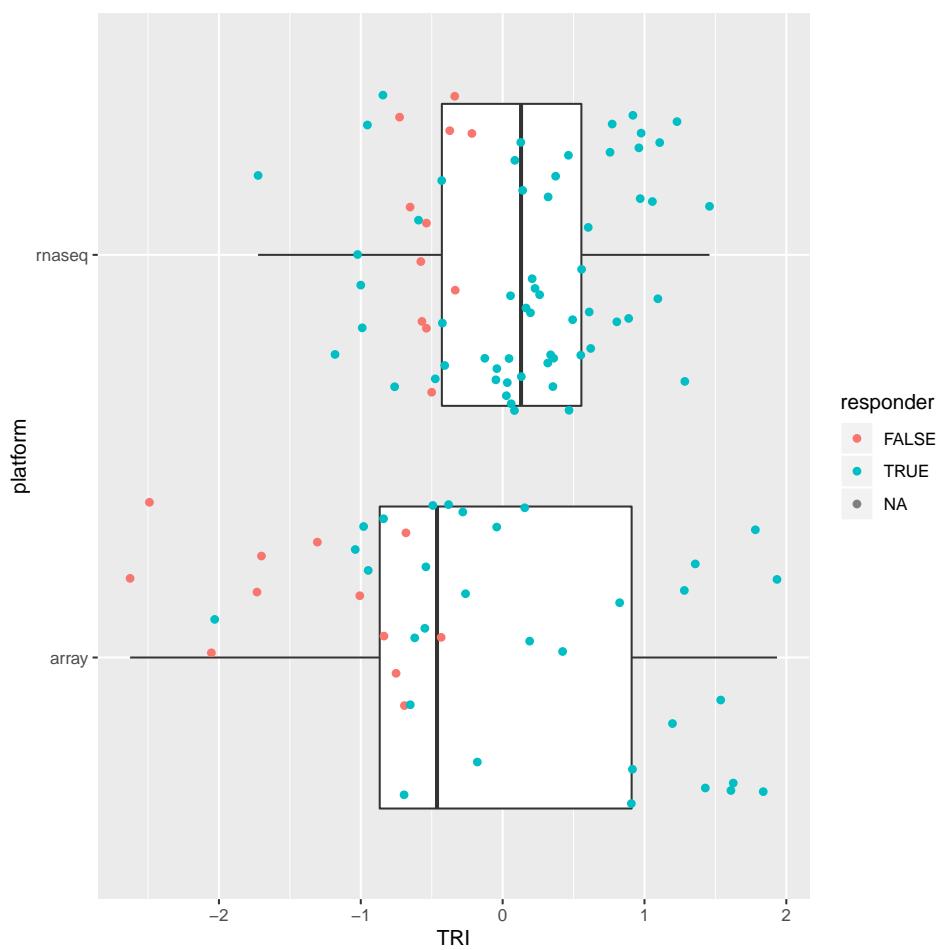


Figure 2.3: TRI vs platform

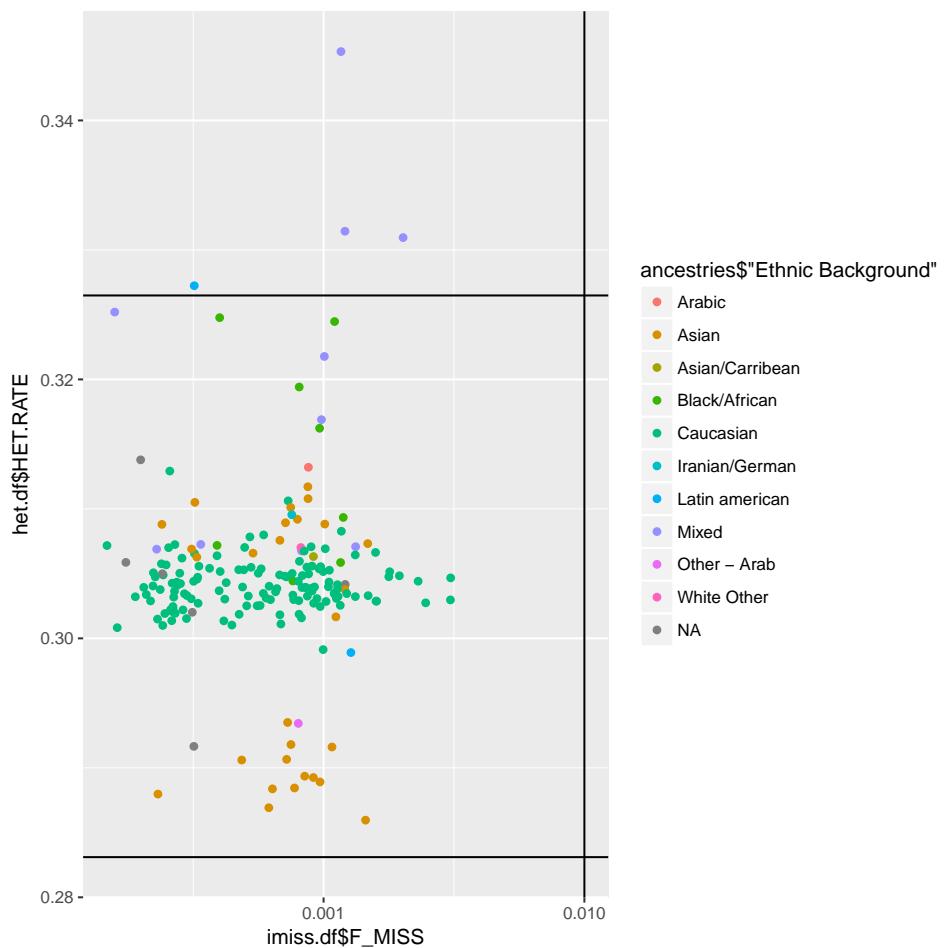


Figure 2.4: Sample filters for missingness vs heterozygosity rate.

After filtering, 169 samples and 549414 markers remained.

2.2.5 Computing genotype PCs 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^{10–12}. 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 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

2.2.6 Genotype phasing and imputation

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 Service¹, 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 missingness > 5% were removed, resulting in 40290981 markers.

2.2.7 RNA-seq data generation

Total RNA was extracted from PBMCs using the Qiagen RNeasy Mini kit, with on-column DNase treatment. RNA integrity was checked on the Agilent Bioanalyzer and mRNA libraries were prepared with the KAPA Stranded mRNA-Seq Kit (KK8421), which uses poly(A) selection. To avoid confounding of timepoint and batch effects from pooling, samples were pooled by library prep plate, ensuring libraries from all timepoints of an individual

¹<https://imputation.sanger.ac.uk/>

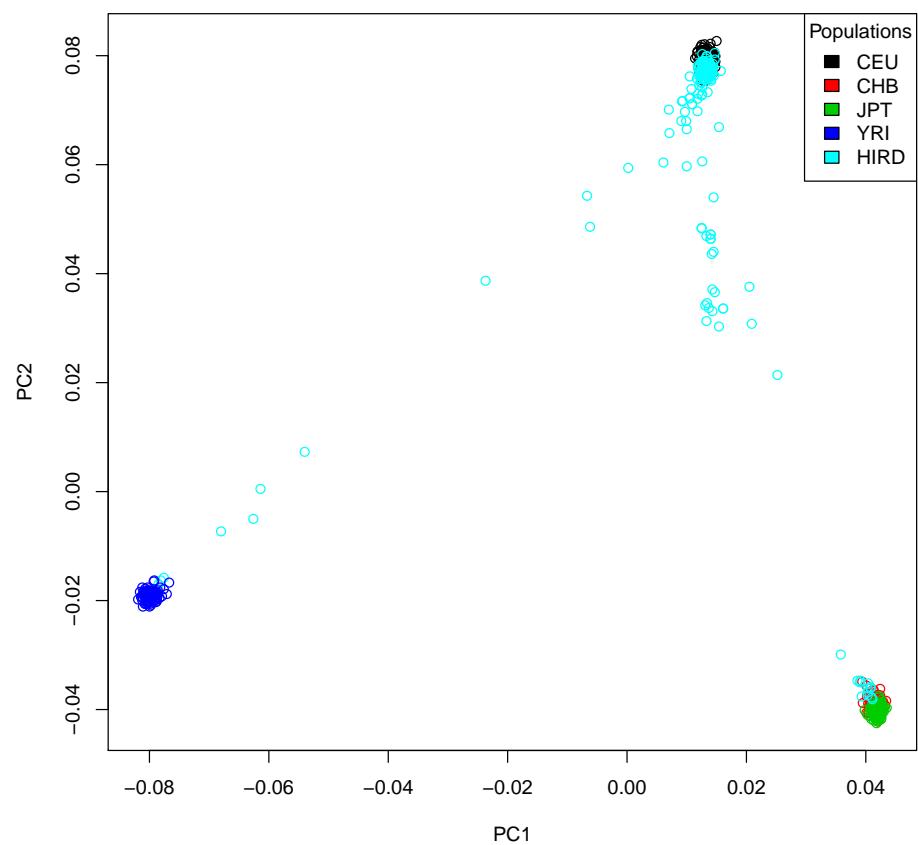


Figure 2.5: HIRD samples (cyan) projected onto PC1 and PC2 axes defined by PCA of HapMap samples. The first two PCs separate European (CEU, upper-right) from Asian (CHB and JPT, lower-right) and African (YRI, lower-left) populations.

were in the same pool, and then sequenced across multiple lanes as technical replicates (HiSeq 4000, 75bp paired-end).

RNA-seq quality metrics were assessed using FASTQC² and Qualimap¹³, then visualised with MultiQC¹⁴. 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).

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

2.2.8 RNA-seq quantification and filtering

[add software versions](#)

Reads were quantified against the Ensembl reference transcriptome (GRCh38) using Salmon¹⁵ 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 differential gene expression (DGE) experiments¹⁶. Relative transcript abundances were summarised to Ensembl gene-level count estimates using tximport (scaledTPM method) to improve statistical robustness and interpretability¹⁷.

Genes with short noncoding RNA biotypes³ were removed, as they are

²<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

³miRNA, miRNA_pseudogene, miscRNA, miscRNA pseudogene, Mt rRNA, Mt tRNA, rRNA, scRNA, snlRNA, snoRNA, snRNA, tRNA, tRNA_pseudogene. <https://www.ensembl.org/Help/Faq?id=468>



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

CHAPTER 2. TRANSCRIPTOMIC RESPONSE TO INFLUENZA A
2.2. METHODS (H1N1)PDM09 VACCINE (PANDEMRIX)

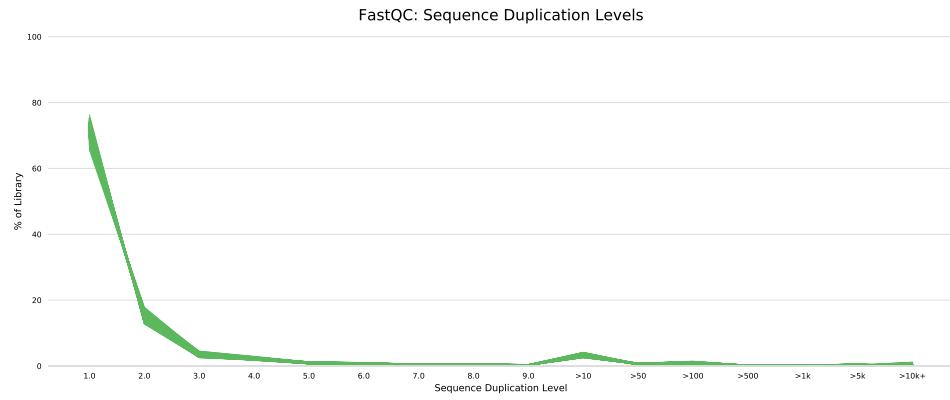


Figure 2.7: FastQC sequence duplication levels for HIRD RNA-seq samples.

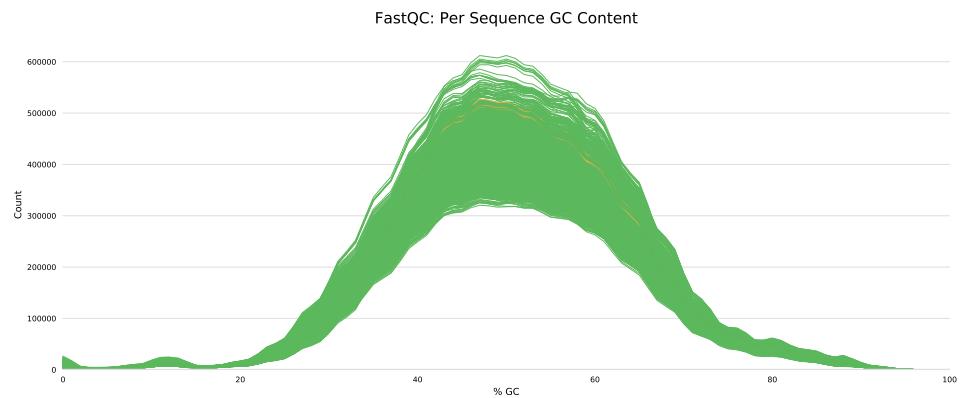


Figure 2.8: FastQC GC profile for HIRD RNA-seq samples.

generally not polyadenylated, and expression estimates can be biased by misassignment of counts from overlapping protein-coding or lncRNA genes¹⁸. Globin genes, which are highly expressed in erythrocytes and reticulocytes, cell types expected to be depleted in **PBMC**¹⁹, were also removed. Given the proportion of removed counts at this stage was low most samples (Fig. 2.9), poly(A) selection and **PBMC** isolation procedures were deemed to be 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 **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 expressed²⁰. 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).

2.2.9 Array data preprocessing

Single-channel Agilent 4x44K microarray (G4112F) data for 173 samples from⁵ were downloaded from <https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-2313/>. VSN²¹ was used to perform background correction for non-specific hybridisation, between-array normalisation, and variance-stabilisation of intensity values, resulting in expression values on a log₂ scale.

Most genes are targeted 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, using the MaxMean method recommended for probe to gene collapsing)²². 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.

standardise form here.
final sentence at end of
each preprocessing should
be n genes x samples x
units

CHAPTER 2. TRANSCRIPTOMIC RESPONSE TO INFLUENZA A
2.2. METHODS **(H1N1)PDM09 VACCINE (PANDEMRIX)**

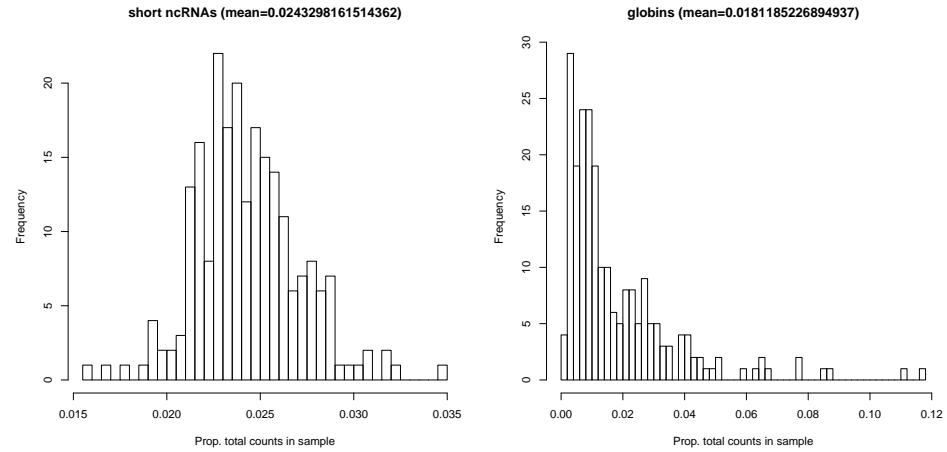


Figure 2.9: Distribution of short ncRNA and globin counts as a proportion of total counts in RNA-seq samples.

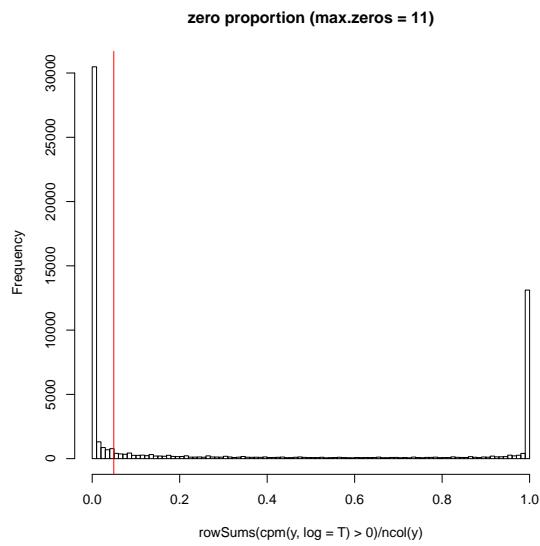


Figure 2.10: Proportion of samples in which genes are detected with non-zero counts. Vertical line shows 5% threshold.

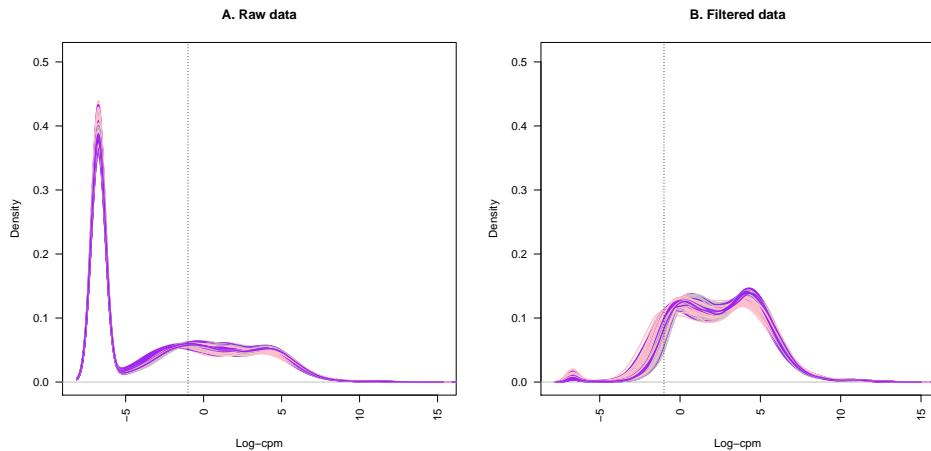


Figure 2.11: Distribution of gene expressions for sample before and after filtering low expression genes. Vertical line shown at $\log(0.5)$ CPM.

2.2.10 Differential gene expression

PCA of the expression data reveals although samples separate by experimental timepoint along PC3 (Fig. 2.12d), measurement platform is by far the largest source of variation, and a batch effect exists within the array data (Fig. 2.12a). The large platform effect likely stems from fundamental technological differences in how each platform measures expression. For example, arrays suffer from ratio compression due to cross-hybridisation²³. RNA-seq has a higher dynamic range, resulting less systematic bias at low expression levels, but estimates are more sensitive to changes in depth than array estimates are to changes in intensity²⁴. There are also differences in the statistical models behind expression quantification, as described above.

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 accounting for between-platform heterogeneity.

cite relevant preprocessing sections

Regarding the batch effect within the HIRD array data, a popular adjustment method is ComBat²⁵, which estimates centering and scaling parameters per-gene pooling information across all genes using empirical Bayes. ComBat is the method used in⁵. In comparisons of microarray batch effect adjustment

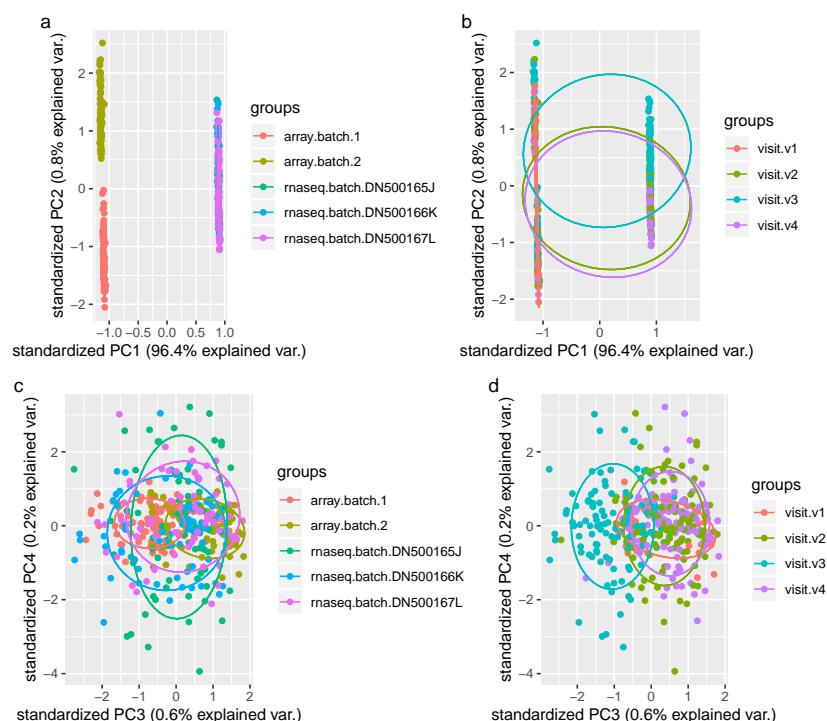


Figure 2.12: First 2 expression PCs, expressions first standardized within platform to increase comparability, if not, even more variation would be driven by platform

methods, ComBat performs favourably (vs. five other adjustment packages)²⁶ or comparably (vs. batch as a fixed or random effect in the linear model)²⁷. However, where batches are unbalanced in terms of sample size²⁸ or distribution of study groups that have an impact on expression²⁹, ComBat can overcorrect batch differences or bias estimates of group differences respectively. In our data, sample size and timepoint groups are fairly balanced between the two array batches, but the proportion of responders is not^{Table 2.2}, hence I elect not to use ComBat to pre-adjust the array expression data, and model the batches as fixed effects. In practice, results from the DGE analysis were not substantially affected by the choice of whether to use a ComBat pre-adjustment or a fixed effect.

2.2.10.1 Per-platform differential gene expression model

As⁵ demonstrated no significant global difference in array expression between day -7 and day 0, I likewise treat merge these two timepoints into a single baseline timepoint “day 0” in the following DGE models.

For RNA-seq samples, between-sample normalisation was performed using the trimmed mean of M-values (TMM) method³⁰ from edgeR³¹; then variance-stabilisation was performed using voom³², resulting in expression values with units of log₂ CPM.

Linear models were fit using limma³³, which is computationally fast, and performs well for sufficiently large ($n \geq 3$ per group) samples³⁴. For each gene, I fit a model (model 1) with expression as the response; and 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. Coefficients and their standard errors were extracted from the linear models. These coefficients represent effect size in units of log2 expression fold-change per unit change in predictor value.

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. then the justification can be simplicity. should probably do this at some point.

The point is, combat has pros such as pergene scaling, that fixed fx don't do

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

specify which particular contrasts are extracted from which model

2.2.10.2 Choice of differential gene expression meta-analysis method

In the section , I concluded that a two-stage meta-analysis approach would add section labels 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^{35,36}. 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 (**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 between-studies heterogeneity, the assumption of $\tau = 0$ is unrealistic, hence a random-effects model is more appropriate.

Unfortunately, there is no optimal solution for directly estimating τ in random-effects meta-analyses with small k ³⁸, especially when $k = 2$ ³⁹. Many estimators are available⁴⁰, but lack of information with small k causes estimation to be imprecise, and can often result in boundary values of $\tau = 0$ that are incompatible with the assumed positive heterogeneity^{41,42}. In such circumstances, the most sensible choice may be to incorporate prior information about model hyperparameters in a Bayesian random-effects framework^{40–43}. For this study, I use the implementation in `bayesmeta`³⁷, which requires priors for both effect size and between-studies heterogeneity.

add label

make all the notation in this section consistent with, and add the equation 2.1. The normal-normal hierarchical model,³⁷

2.2.10.3 Prior for between-studies heterogeneity

The choice of prior for between-studies heterogeneity is influential when k is small⁴³. Gelman [44] 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 non-positive values for τ are unrealistic, I use a weakly-informative gamma prior recommended by⁴¹, which has zero density at $\tau = 0$, increasing gently as τ increases. This constrains τ to be positive, but still permits estimates close to zero if the data support it. This is in contrast to log-normal (e.g.^{45,46}) or inverse-gamma priors (e.g.⁴⁷) which have derivatives close to zero, and rule out small values of τ no matter what the data suggest; and in contrast to half-t family priors (e.g.^{43,44}), 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 random-effects model using the **restricted maximum likelihood (REML)** estimator for τ (recommended for continuous effects⁴⁰) was first fit for each gene using **metafor**. Genes with small estimates of $\tau < 0.01$ were excluded, and a gamma distribution was fit to the remaining estimates using **fitdistrplus**.

2.2.10.4 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^{42,44} **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 choose a very weak normal prior with $\mu = 0$ over a flat prior.

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

As in the section above, to determine an appropriate scale, a normal distribution with mean $\mu = 0$ is fit to the distribution of effect sizes from the gene-wise metafor models to empirically estimate the **SD** σ . Heavy-tailed Cauchy priors have been proposed for effect size distributions in **DGE** experiments to avoid over-shrinkage of large effects⁴⁸. Since **bayesmeta** does not implement a Cauchy prior, the normal prior used for meta-analysis is flattened to $N(0, 10\sigma)$ to avoid over-shrinkage in the tails. This prior is equivalent to assuming a 95% probability that effects are less extreme than approximately 20σ .

the derivation here is

$$\text{qnorm}(0.975, \text{mean}=0, \text{sd}=1*10) = 1*19.59964$$

2.2.10.5 Evaluation of priors

⁴¹ recommends (shape=2, scale= λ), with small λ .

From Fig. 2.14 we are similar.

2.2.10.6 FDR

ashr

2.2.11 Gene set enrichment analysis

tmod; gprofileR; CAM-ERA

Before correction
 $d1\text{.vs}\text{.d0: rnaseq vs array log2FC } (r=0.894289 \text{ rho}=0.851984)$

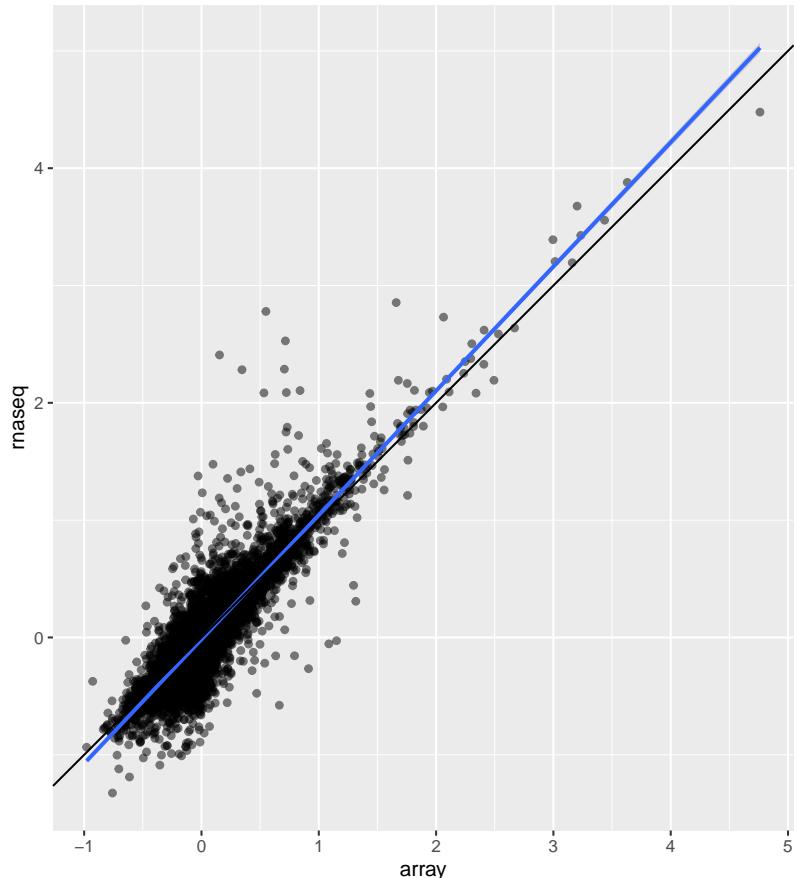


Figure 2.13: Fold-change comparison between array and RNAseq for day 1 vs day 0.

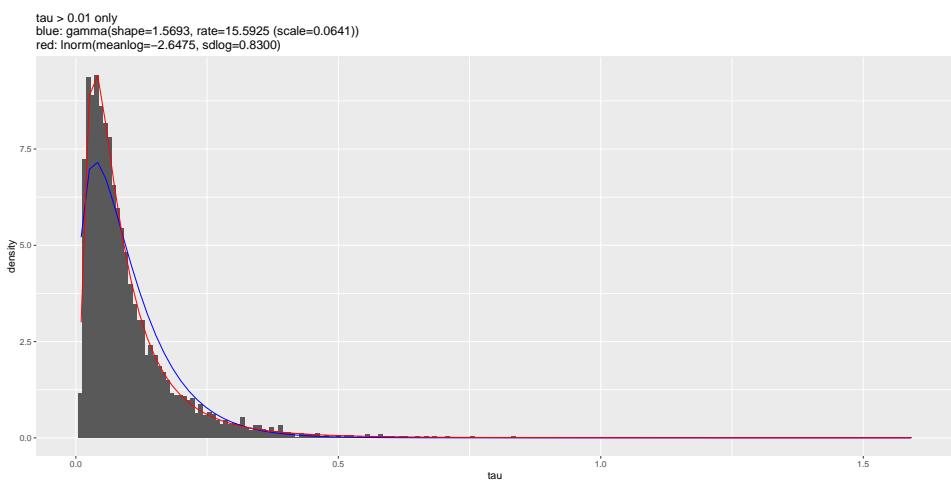


Figure 2.14: Priors for tau for day 1 vs day 0 DGE meta-analysis.

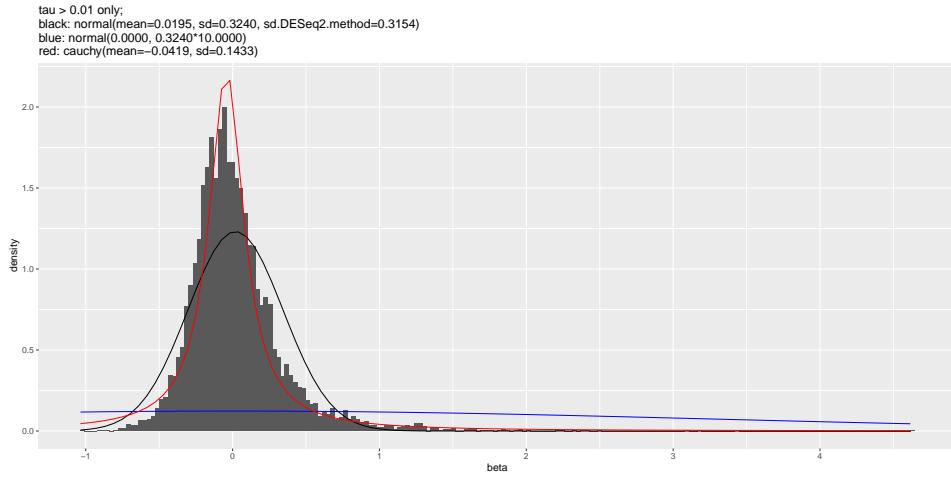


Figure 2.15: Priors for mu for day 1 vs day 0 DGE meta-analysis.

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 local false sign rate (lfsr) < 0.05 and absolute fold change > 1.5 cutoff, 857/13593 genes were differentially expressed between any pair of timepoints. Clustering the expression of these genes revealed three main clusters: 692 genes that tend to be upregulated at day 1, 94 gene upregulated at day 7, and 59 genes downregulated at day 1 (Fig. 2.16).

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. 2.12), the highest numbers of differentially expressed genes are observed at day 1, with 580 genes differentially expressed vs. baseline, the majority of which return to baseline by day 7.

CHAPTER 2. TRANSCRIPTOMIC RESPONSE TO INFLUENZA A
2.3. RESULTS (H1N1)PDM09 VACCINE (PANDEMRIX)

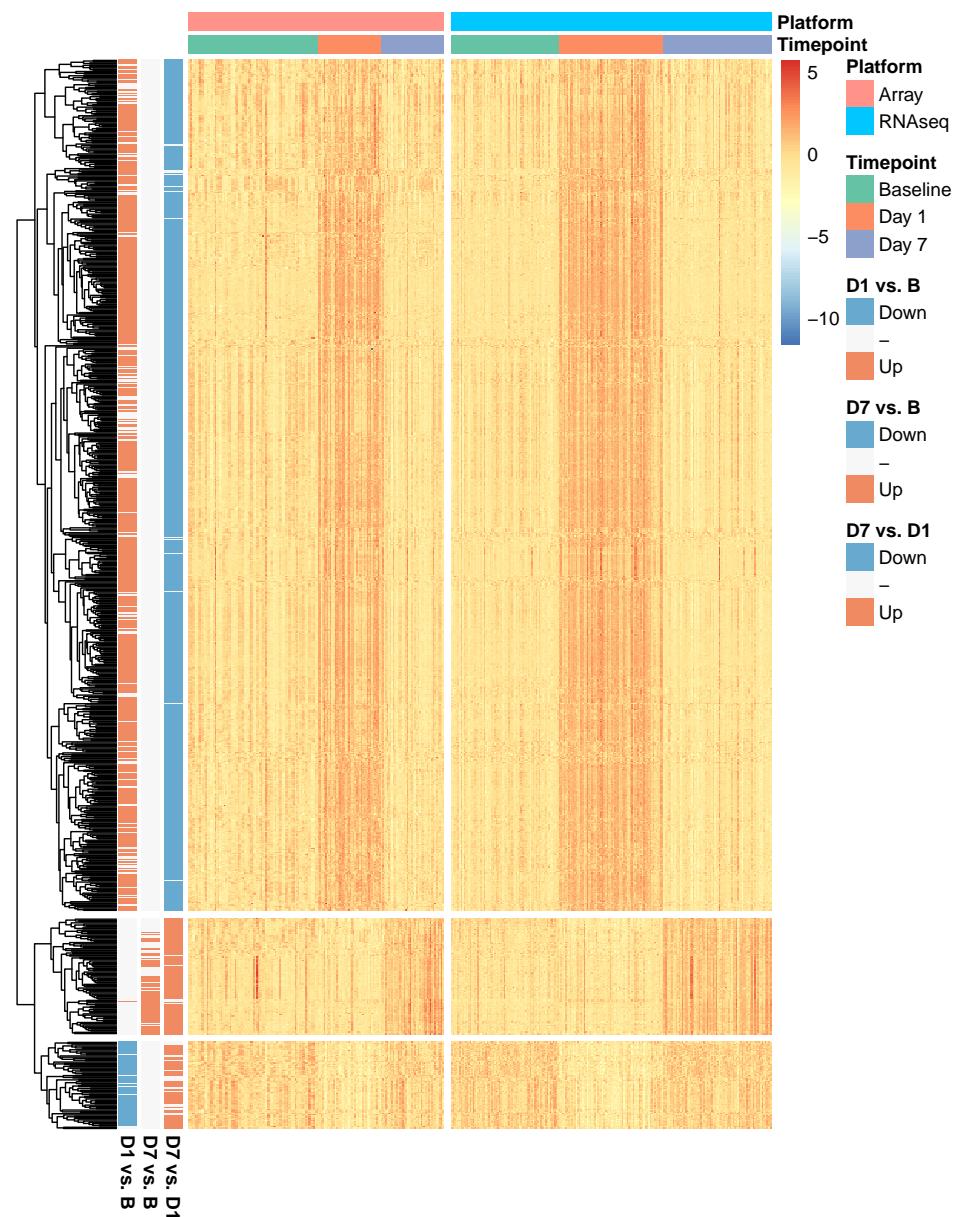


Figure 2.16: Normalised gene expression for genes differentially expressed between any pair of timepoints ($\text{lfsr} < 0.05$, $|\text{FC}| > 1.5$) across HIRD, clustered by gene. Distance metric: Manhattan.

The gene with the highest fold change (FC) increase at day 1 compared to baseline was *ANKRD22* ($\log_2 \text{FC} = 4.489150$), an interferon-induced gene in monocytes and dendritic cells (DCs) involved in antiviral innate immune pathways⁴⁹. Other key genes in the interferon signalling pathway⁵⁰ such as *STAT1* ($\log_2 \text{FC} = 2.1693060$), *STAT2* ($\log_2 \text{FC} = 0.9489341$), and *IRF9* ($\log_2 \text{FC} = 0.8153674$) are also upregulated at day 1. Gene set enrichment analysis using tmod revealed that genes with the high FC increases at day 1 were enriched in modules associated with activated DCs, monocytes, toll-like receptor and inflammatory signalling (Fig. 2.17), confirming that day 1 responses are dominated by signatures of innate immunity.

2.3.3 Adaptive immune response at day 7 post-vaccination

Expression fold changes were more modest at day 7; the genes with the highest upregulation being the B cell-associated genes *TNFRSF17* ($\log_2 \text{FC} = 1.7538617$) and *MZB1* ($\log_2 \text{FC} = 1.7369668$). Plasma cell-specific genes including *SDC1* (encodes CD138 <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5437827/>) ($\log_2 \text{FC} = 1.3673081$) and *ELL2* (<https://www.nature.com/articles/ni.1786>) ($\log_2 \text{FC} = 0.8679659$) were also prominently upregulated. The enriched gene sets at day 7 were related to mitosis and cell proliferation, particularly in CD4⁺ T cells (Fig. 2.17). Both the CD4⁺ T cell and plasma cell response are indications of an adaptive immune response at day 7.

Not sure if there is a biological interpretation of downreg of T cells and NK cells gene sets, since it could be due to increase in other cell types in the sample

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

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

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

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-BH < 0.05, 6 genes had expression associated with TRI at day 0, 55 at day 7, and 11 pooled across timepoints (red, Fig. 2.18).⁵ 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 false discovery rate (FDR) < 0.05, of which 58/62 fall into the 13593 genes considered in my meta-analysis (circled, Fig. 2.18), and 15/58 replicated, all with the same positive direction of effect (high expression correlates with high TRI). In the Bayesian meta-analysis, no single

add label

CHAPTER 2. TRANSCRIPTOMIC RESPONSE TO INFLUENZA A
2.3. RESULTS (H1N1)PDM09 VACCINE (PANDEMRIX)

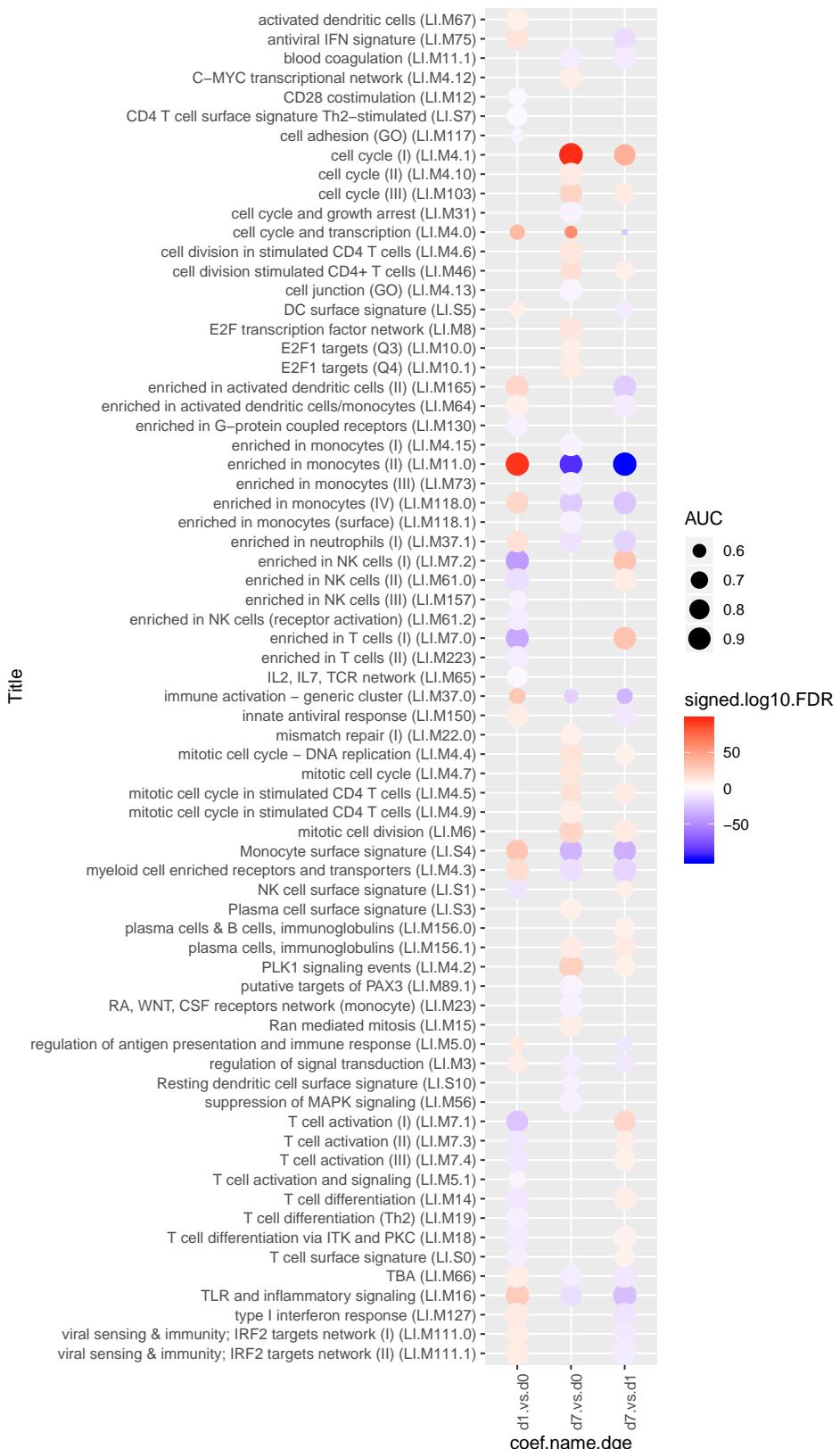


Figure 2.17: Transcriptomic modules significantly up or downregulated post-vaccination. Size of circle indicates effect size. Color of circle indicates significance and direction of effect (red = upregulation, blue = downregulation).

gene was detected as significantly associated with **TRI** at $\text{lfsr} < 0.05$ at any timepoint, or when pooling samples across all timepoints (circled, Fig. 2.19).

Significant enrichments can be detected at the gene set level; the strongest effects are seen at day 7, where expression of cell cycle, CD4⁺ T cells, and plasma cells are associated with high **TRI**. At day 0, modules related with inflammatory response in myeloid cells are also associated with high **TRI** (Fig. 2.20).

have you explained how tmod works?

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

predictive analysis not complete, and possibly not even required for this chapter

2.3.5 Identifying expression signatures for predicting antibody response

2.4 Discussion

There is extensive transcriptomic response to Pandemrix vaccination in the **HIRD** cohort, with characterised genes and gene sets differentially expressed between timepoints, and some heterogeneity in expression associated with baseline-adjusted antibody response. 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 for the adaptive immune response, involving CD4⁺ T cell-supported differentiation and proliferation of antibody-secreting plasmablasts and plasma cells⁵¹. These patterns of expression change between timepoints are consistent with the patterns of (array-only) expression, and with expansions of monocyte and plasma cell populations seen in the **fluorescence-activated cell sorting (FACS)** data at days 1 and 7 respectively in the original **HIRD** study⁵.

mention downreg at d1 section in results, then add here

In contrast, I was not able to fully replicate single gene-level associations between day 7 expression and antibody response reported by⁵. Of the 62 genes reported as differentially expressed between vaccine responders and non-responders, 58/62 were measured by both platforms and assessed in the meta-analysis.

Although⁵ encodes responder status as a binary phenotype, whereas my analysis uses **TRI**, which is continuous and adjusts for baseline titres, this is

Add numbers up and down reg to facet titles

CHAPTER 2. TRANSCRIPTOMIC RESPONSE TO INFLUENZA A

2.4. DISCUSSION (H1N1)PDM09 VACCINE (PANDEMRIX)

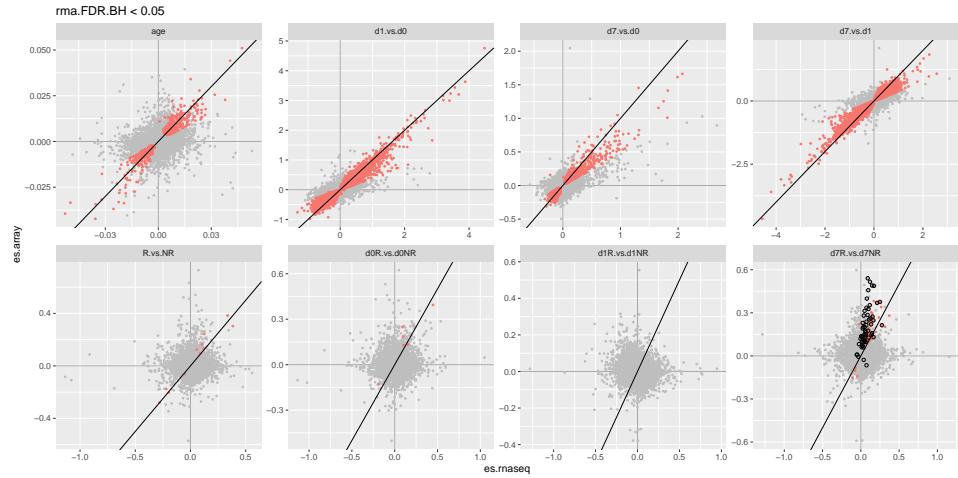


Figure 2.18: DGE effect sizes estimated in array vs RNA-seq. Significance colored by frequentist random effects meta-analysis FDR < 0.05.

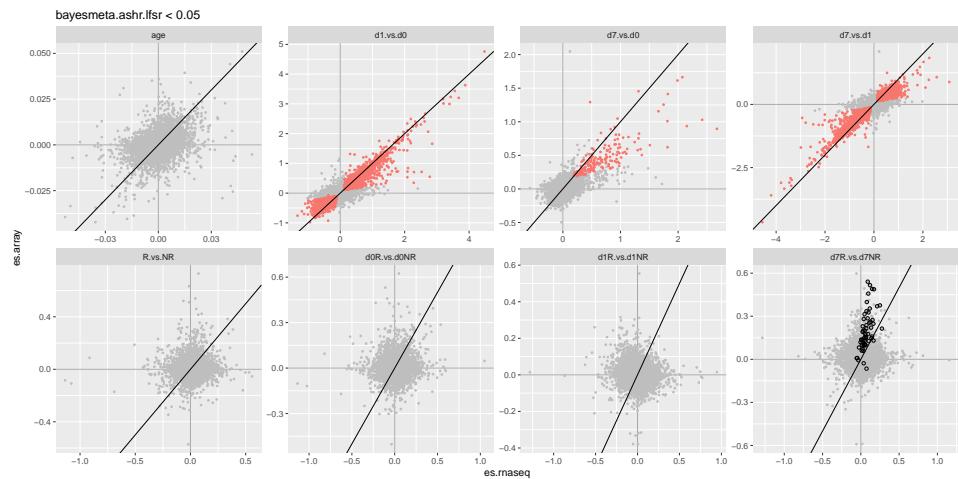


Figure 2.19: DGE effect sizes estimated in array vs RNA-seq. Significance colored by Bayesian random effects meta-analysis lfsr < 0.05.

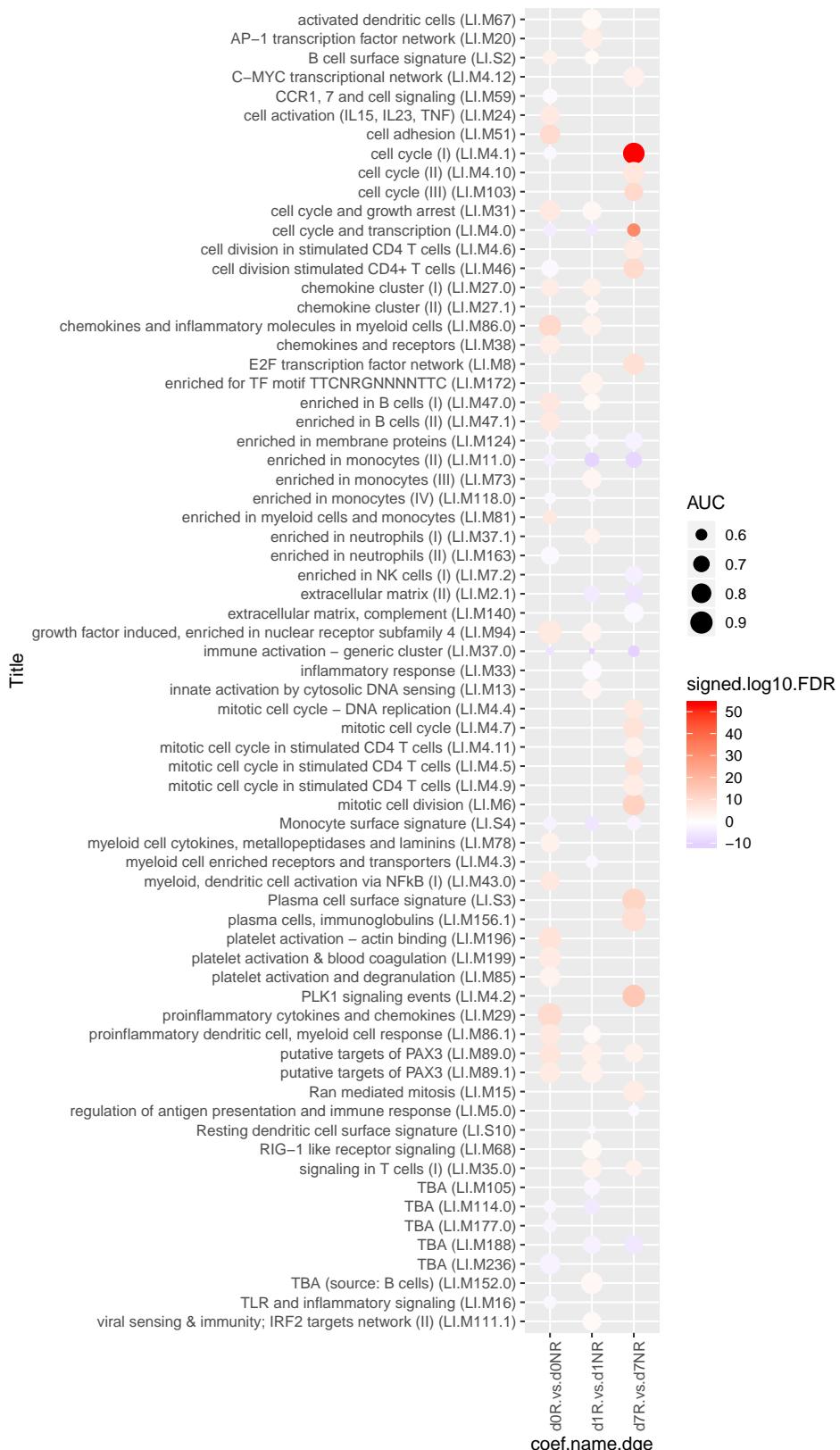


Figure 2.20: Transcriptomic modules enriched in genes with expression associated with antibody response (TRI) at each day. Size of circle indicates effect size. Color of circle indicates significance and direction of effect (red = expression positively correlated with TRI, blue = negative).

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

not the primary difference, as 51/62 genes replicated ($FDR < 0.05$) when considering the array data in isolation, despite the difference in phenotype. The majority of the effects for these genes were simply much stronger in the array dataset than in the RNAseq dataset (Fig. 2.18). 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.

As such, only 15/58 signals replicated using frequentist random-effects meta-analysis to combine per-platform estimates. I do not consider these hits as robust, as the **REML** estimate of between-platform heterogeneity was zero for 8563/13593 for the day 7 **TRI** contrast overall, and zero for all 15 of the replications. 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. 2.19). 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 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: apoptosis⁵²; Fc γ receptor-mediated phagocytosis, TREM1 signaling⁵³; enriched in B cells, T cell activation⁵⁴; B cell receptor signalling, inflammatory response, platelet activation⁵⁵; several of which I also observe. It should

be noted that comparisons with these signatures from existing influenza systems vaccinology studies should be caveated, as most existing studies are for non-adjuvanted influenza vaccines, whereas Pandemrix is AS03-adjuvanted. Adjuvanted influenza vaccines are considerably more immunogenic, and post-vaccination expression patterns differ to those of non-adjuvanted vaccines (<https://www.frontiersin.org/articles/10.3389/fimmu.2017.01760/full>). Hence, it is particularly important that the robustness of these observed baseline expression signatures be validated in an independent cohort for a comparable AS03-adjuvanted influenza vaccine.

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

There is also something to be said about 'prediction is not inference'

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

CHAPTER 2. TRANSCRIPTOMIC RESPONSE TO INFLUENZA A
2.4. DISCUSSION (H1N1)PDM09 VACCINE (PANDEMRIX)

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 n = 374	1 n = 87	2 n = 79	batch		
				DN500165J n = 70	DN500166K n = 69	DN500167L 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

[The influence of host genetics on vaccines response has also been explored] Vaccine-induced antibody response is a complex trait, with heritability estimates ranging from ... [e.g. seaonsal influenza 10.1016/j.vaccine.2008.07.065 Poland e.g. smallpox e.g. measeks 10.1080/21645515.2015.1119345.]

Narcolepsy controversy (evidence for genetics)

A potential mechanism through which genetic variation can affect vaccine response is through altering the expression of nearby genes (cis-eQTLs). 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]. [summary of Sobolev findings]

In this study, we model the influence of host genetics on longitudinal transcriptomic and antibody responses to Pandemrix, *in vivo*.

also, we have phenotype data, *in vivo*

[main aim: how much variation in response is genetic?] [other aims: assess differences to seasonal influenza vaccines] [summary of main results] Why Sobolev? More variation will be explained by history of exposure rather than genetics, so may be harder to detect.

Knowns Sobolev: R vs NR, inconsistent variation in why people are NR
Prevacc signatures of Tri Using larger transcriptomic dataset Are they

genetic

Good points of our study Repeated measures in vivo perturbation

Utility of genetics: allows coloc How does common genetic variation affect response to vaccine?

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.)

3.1.1 Genetic factors affecting influenza vaccine response

Impact of host genetic polymorphisms on vaccine induced antibody response

3.1.2 Context-specific immune response QTLs for influenza vaccine response

if change in expression vs d0 is under genetic control, we should see change in effect size of eqtl vs d0

Summarise Franco et al

3.1.3 Chapter summary

Given the large changes in expression in ch2, detect context-specific fx.

3.2 Methods

3.2.1 Genotype imputation

why exclude x chrom? As is standard for imputation, we excluded all X-linked SNPs for the following reasons: (i) the X chromosome has to be treated differently from the autosomes; (ii) it cannot be predicted which allele is active on the X chromosome, (iii) testing males separately from females results in different sample sizes and power. Imputation of SNPs in the HapMap CEU population was performed using either MACH46 or IMPUTE47. All SNPs with a MAF <0.01 were excluded from analysis. In total, up to 2.11 million genotyped or imputed SNPs were analyzed.

3.2.2 Estimation of cell type abundances

FACS data norm; imputation; scaling

deconv

decon eqtl decon2 has an interesting method: no genotype main effect requires full data i.e. it's an eqtl mapper

cell type interaction terms from proxy genes

Why impute for cell counts but not for eQTL? expression matricse are mostly complete, and we only exclude genes based on low expression in RNAseq we cannot drop whole panels so easily like we can drop genes

Note, the use of gene signatures for deconv in stimulated samples does not distinguish upreg from prolif either if expression goes up, the method will detect more of the signature i.e. it may correct away some signal of upregulation

3.2.3 Mapping cis-eQTLs with LMM

lmms: use a kinship matrix to scale the sample-sample genetic covariance
see: 2018-11-16 notes in log

this is good background

Choice of lmm method for various methods, see 2018-03-05 and 2018-07-25
in log

for discussion of how lmm implementation doesn't matter (Eu-ahsunthornwattana et al., 2014)

Can also refer to previous notes in "2017_Book_SystemsGenetics"

why including known covariates: why not a two stage approach?

Why not mapping on deltas? (if we are interested in the direct question of G on change) ackermann: change scores are prone to increased noise from franco: "We attempted analyses with an approach similar to that proposed by the reviewers in the course of our work, but found that the approach that was ultimately chosen to explore the day differences was the most powerful. Specifically, utilizing a pairwise comparison (difference) between time points as the substrate for the eQTL analysis would lead to an increase in the technical variance of the phenotype, as the sum of two independent (technical) errors has twice the variance of an individual measurement. "

NOTE: peer factors would need to be computed on the foldchange phenotype

The final model:

3.2.3.1 Expression normalisation

2018-03-15 in log

Rank-based int: heavily used in genetics, Although criticised: "Rank-Based Inverse Normal Transformations are Increasingly Used, But are They Merited?"

3.2.3.2 Finding hidden confounders with PEER

Why RANKINT before PEER? "Many statistical tests rely on the assumption that the residuals of a model are normally distributed [1]. In genetic analyses of complex traits, the normality of residuals is largely determined by the normality of the dependent variable (phenotype) due to the very small effect size of individual genetic variants [2]. However, many traits do not follow a normal distribution." "applying rank-based INT to the dependent variable residuals after regressing out covariates re-introduces a linear correlation between the dependent variable and covariates, increasing type-I errors and reducing power."

PEER: expression PCs: if too many, will explain away the signal Not a problem with cis-eQTLs, but trans might have more global effects

GWAS on PEER factors would pick up trans fx, cell count QTL effects

Unlike PCs, PEER factors are not constrained to be orthogonal: adding more and more factors will not explain more of the variance Also, they are weighted

why include genetic PCs see stegle 2012 PEER paper: if PCs are not included, they can be recapitulated in the factors

3.2.4 eQTL mapping with mixed models

Sample AC thresh note they are dosages. if they were not, use ac thresh to estimate number of hom minor expected

3.2.5 eQTL meta-analysis

Why not do a mega analysis? Using a fixed effect assumes mean diff between rnaseq and array and forces the slope to the average. Adding a Gxplatform interaction again leads to diff effect sizes problem.

meta-analysis? can't do a bayesian, which would be ideal. also, small n for array

Restricted to non-full bayesian methods. For small k, Sidik MVa or Ruhkin RBp recommended. Sidik-Jonkman estimator, also called the ‘model error variance estimator’, is implemented in metafor (SJ method).

Starts with an init estiamte of $ri=\sigma^2_i/\tau^2_i$ i.e. ratio of study-specific and between-studies het variance, then updates.

They recommend using Hedges [1], to init, but this is bad???

We use mode of gamma as an apriori estiamte of tau.

compuationally challenging Note we can't just meta the top eqtls from RNAseq as a shortcut , as there is no guarantee the top would have been the top from a meta analysis in the beginning

3.2.5.1 Joint mapping with mashr

In the same period that condition specific eQTL mapping was getting started (as discussed in section...), tools were being created to identify these locitoools were being created to identify these loci review: condition/Cell-type specific methods refere to 2019-11-19 Cell-count specific eQTL mapping papers PANAMA, LIMMI

How much sharing is expected? - overestimates of specificity? e.g. (fair-fax2014InnateImmuneActivity More than half of cis-eQTLs identified, involving hundreds of genes and associated pathways, are detected specifically in stimulated monocytes.)

Simple, mixed models, joint models, multilocus models; Ending with why we chose mashr

normally eqtls use perms for FDR

used for smoothing, info sharing, fdr

mashr beats out stuff it compared to in the paper e.g. metasoft

Choice of strong effects If there is a particular condition with much greater power, choosing the lowest p value for each gene across all conditions could bias strong effects towards including just condition-specific effects for that particular condition. how to ensure condition specific effects are present? look at heatmap of strong subset

lfsr:

3.2.6 Defining shared and response eQTLs

beta-comparison approach from Sarah Kim-Hellmuth 2017 note they correct for FDR

3.3 Results

3.3.1 Overview of eQTLs at each timepoint

3.3.1.1 Estimation of eQTL sharing

Look at diff in beta, not multiples, other 0->1 will be inf

3.3.1.2 TODO Replication of shared eQTLs in whole blood

3.3.2 Characterising re-eQTLs at each timepoint

Ranking metrics: PVE: prefers large maf and high betas since it squares the beta. even if the beta does not change so much. ignores sign. beta: p: ignores sign Z score:

3.3.3 The mechanism of reQTLs

3.3.4 TODO Colocalisation of re-eQTLs with known context-specific immune QTLs

Colocalisation with known associations; Colocalisation is used to understand the molecular basis of GWAS associations (of a variety of human disease traits) (Giambartolome, 2014); Here the inverse: coloc is used to understand the biological relevance of observed expression variation

Choice of method; Coloc and assumptions; Hypercoloc and assumptions

3.3.5 TODO Disruption of binding site motifs as a model for re-eQTLs

See models from Fu et al, Unraveling the Regulatory Mechanisms Underlying Tissue-Dependent Genetic Variation of Gene Expression

3.4 Discussion

Current limitations; Confounded by changes in immune cell proportions in bulk PBMCs; Unclear connection to vaccine biology e.g. what genesets/pathways/cell types are driving the observed transcriptomic and eQTL response?;

3.4.1 DGE vs eqtl

Why dge eqtl overlap poor? Peer accounts for cell prop means.....

It must be said, overlap is not rigorous Formal Mediation analysis required

3.4.2 limitations: The mechanism of reQTLs

3.4.3 Conditional eQTL effects

Confounding by multiple causal variants?; No conditional eQTL analysis to disentangle conditional effects; Are re qtls more likely to be distal and secondary?

CHAPTER 3. GENETIC FACTORS AFFECTING PANDEMRIX

3.4. DISCUSSION

VACCINE RESPONSE

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

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

5.3 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

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/predict

unification immunology and vaccine dev: deep phenotyping, small cohorts

achieved -> larger cohorts human genetics and gwas: large cohorts achieved

-> deeper phenotyping

CHAPTER 6. DISCUSSION

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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/10.1016/j.coi.2014.07.001) (Oct. 2014).
5. 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](https://doi.org/10.1038/ni.3328) (Jan. 4, 2016).
6. Food and Drug Administration. *Guidance for Industry: Clinical Data Needed to Support the Licensure of Pandemic Influenza Vaccines* (Jan. 6, 2007), 20.
7. Cohen, J. The Cost of Dichotomization. *Applied Psychological Measurement* **7**, 249–253. doi:[10.1177/014662168300700301](https://doi.org/10.1177/014662168300700301) (June 1983).
8. Fedorov, V., Mannino, F. & Zhang, R. Consequences of Dichotomization. *Pharmaceutical Statistics* **8**, 50–61. doi:[10.1002/pst.331](https://doi.org/10.1002/pst.331) (Jan. 2009).

APPENDIX A. BIBLIOGRAPHY

9. 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](https://doi.org/10.1093/infdis/jiq156) (Apr. 2011).
10. 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](https://doi.org/10.1038/ng1847) (Aug. 2006).
11. 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](https://doi.org/10.1371/journal.pgen.1004445) (July 17, 2014).
12. Brown, B. C., Bray, N. L. & Pachter, L. Expression Reflects Population Structure. doi:[10.1101/364448](https://doi.org/10.1101/364448) (July 8, 2018).
13. 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](https://doi.org/10.1093/bioinformatics/btv566) (Oct. 1, 2015).
14. 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](https://doi.org/10.1093/bioinformatics/btw354) (Oct. 1, 2016).
15. 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](https://doi.org/10.1038/nmeth.4197) (Apr. 6, 2017).
16. 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](https://doi.org/10.1093/bioinformatics/btt688) (Feb. 1, 2014).
17. 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](https://doi.org/10.12688/f1000research.7563.2) (Feb. 29, 2016).

APPENDIX A. BIBLIOGRAPHY

18. 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](https://doi.org/10.1038/s41598-018-23226-4) (Dec. 2018).
19. 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](https://doi.org/10.1186/1471-2164-11-96) (2010).
20. 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](https://doi.org/10.12688/f1000research.8987.2) (Aug. 2, 2016).
21. 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](https://doi.org/10.1093/bioinformatics/18.suppl_1.S96) (Suppl 1 July 1, 2002).
22. 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](https://doi.org/10.1186/1471-2105-12-322) (2011).
23. 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](https://doi.org/10.1016/j.tig.2005.12.005) (Feb. 2006).
24. 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](https://doi.org/10.1093/nar/gkv636) (June 30, 2015).
25. 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](https://doi.org/10.1093/biostatistics/kxj037) (Jan. 1, 2007).
26. 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](https://doi.org/10.1371/journal.pone.0017238) (Feb. 28, 2011).

APPENDIX A. BIBLIOGRAPHY

27. 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](https://doi.org/10.1371/journal.pone.0202947) (Aug. 30, 2018).
28. 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](https://doi.org/10.1186/s12859-018-2263-6) (Dec. 2018).
29. 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](https://doi.org/10.1093/biostatistics/kxv027) (January Aug. 13, 2015).
30. 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](https://doi.org/10.1093/bib/bbx008) (Sept. 28, 2018).
31. 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](https://doi.org/10.1093/bioinformatics/btp616) (Jan. 1, 2010).
32. 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).
33. 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](https://doi.org/10.1093/nar/gkv007) (Apr. 20, 2015).
34. 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](https://doi.org/10.1186/1471-2105-14-91) (Dec. 2013).
35. 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](https://doi.org/10.1037/1082-989X.8.3.243) (2003).

APPENDIX A. BIBLIOGRAPHY

36. 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](https://doi.org/10.1002/jrsm.12) (Apr. 2010).
37. Röver, C. Bayesian Random-Effects Meta-Analysis Using the Bayesmeta R Package (Nov. 23, 2017).
38. Bender, R. *et al.* Methods for Evidence Synthesis in the Case of Very Few Studies. *Research Synthesis Methods*. doi:[10.1002/jrsm.1297](https://doi.org/10.1002/jrsm.1297) (Apr. 6, 2018).
39. 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](https://doi.org/10.1002/sim.6473) (July 20, 2015).
40. 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](https://doi.org/10.1002/jrsm.1164) (Mar. 2016).
41. Chung, Y., Rabe-Hesketh, S., Dorie, V., Gelman, A. & Liu, J. A Non-degenerate Penalized Likelihood Estimator for Variance Parameters in Multilevel Models. *Psychometrika* **78**, 685–709. doi:[10.1007/s11336-013-9328-2](https://doi.org/10.1007/s11336-013-9328-2) (Oct. 2013).
42. 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](https://doi.org/10.1002/jrsm.1217) (Mar. 2017).
43. 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](https://doi.org/10.1186/s12874-018-0618-3) (Dec. 2019).
44. 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](https://doi.org/10.1214/06-BA117A) (Sept. 2006).
45. 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](https://doi.org/10.1002/sim.4326) (Nov. 20, 2011).

APPENDIX A. BIBLIOGRAPHY

46. 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](https://doi.org/10.1002/sim.6381) (Mar. 15, 2015).
47. 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](https://doi.org/10.1002/(SICI)1097-0258(19961230)15:24<2733::AID-SIM562>3.0.CO;2-0) (Dec. 30, 1996).
48. 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](https://doi.org/10.1093/bioinformatics/bty895) (June 1, 2019).
49. 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).
50. 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](https://doi.org/10.1146/annurev-immunol-032713-120231) (Mar. 21, 2014).
51. Murphy, K. & Weaver, C. *Janeway's Immunobiology* 9th edition. 904 pp. (Garland Science/Taylor & Francis Group, LLC, New York, NY, 2016).
52. Furman, D. et al. Apoptosis and Other Immune Biomarkers Predict Influenza Vaccine Responsiveness. *Molecular Systems Biology* **9**, 659. doi:[10.1038/msb.2013.15](https://doi.org/10.1038/msb.2013.15) (Apr. 16, 2013).
53. 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](https://doi.org/10.1016/j.cell.2014.03.031) (Apr. 10, 2014).
54. 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.jimmuni.2015.11.012](https://doi.org/10.1016/j.jimmuni.2015.11.012) (Dec. 2015).

APPENDIX A. BIBLIOGRAPHY

55. 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](https://doi.org/10.1126/sciimmunol.aal4656) (Aug. 25, 2017).

APPENDIX A. BIBLIOGRAPHY

List of Abbreviations

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

HAI haemagglutination inhibition

HIRD Human Immune Response Dynamics

LD linkage disequilibrium

lfsr local false sign rate

MAF minor allele frequency

MN microneutralisation

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

TMM trimmed mean of M-values

TRI titre response index

List of Abbreviations

List of Abbreviations

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

add more pros for in vitro reQTLs here, and find citations	6
define what a signature is	8
cite appropriate subfigures here	14
cite appropriate subfigures here	15
Add to collab note	15
Add Tracy-Widom statistics for PCs	19
Can add other fastqc plots e.g. kmers, overrepresented seqs, seq length	21
add software versions	21
standardise form here. final sentence at end of each preprocessing should be n genes x samples x units	23
cite relevant preprocessing sections	25
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. then the justification can be simplicity. should probably do this at some point.	27
The point is, combat has pros such as pergene scaling, that fixed fx don't do	27
link to papers justifying sex, age, ancestry as significant effects on immune gene expression	27
specify which particular contrasts are extracted from which model . .	27
add section labels	27
add label	28
make all the notation in this section consistent with, and add the equation 2.1. The normal-normal hierarchical model, ³⁷	28
why is this? is it having well powered studies? gelman is vague . . .	29
the derivation here is qnorm(0.975, mean=0, sd=1*10) = 1*19.59964 .	29

From Fig. 2.14 we are similar.	29
ashr	29
tmod; gprofileR; CAMERA	29
Not sure if there is a biological interpretation of downreg of T cells and NK cells gene sets, since it could be due to increase in other cell types in the sample	33
any point in a table of e.g. top 20 DE genes, or is the gene set analysis already enough?	33
can also add MSigDB hallmark sets, which include interferon sets; and of course gene ontology sets	33
change x axis labels to baseline, specify top 10 procedure in figure caption	33
add label	33
have you explained how tmod works?	35
figure x labels here should be TRI, not R.vs.NR	35
predictive analysis not complete, and possibly not even required for this chapter	35
mention downreg at d1 section in results, then add here	35
Add numbers up and down reg to facet titles	35
might have to rerun everything using the original binary R/NR if this line of reasoning isn't strong enough	38
There is also something to be said about 'prediction is not inference'	39
found signatures, but so what? Feels like chapter lacks a punchline?	39
spell-check	69
make sure package versions are in, and package names are monospace	69
add automatic rounding to x decimal places using num and sisetup	69