

<title>

Benjamin Yu Hang Bai

March 31, 2020

<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	4
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	5
1.3.2.1 Context-specific immune response eQTLs in vitro	6
1.3.2.2 <i>in vivo</i> response QTL mapping	6
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	8
1.5.1 Genetic factors affecting biologic responses	8
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 Response to influenza vaccines	12
2.1.3 The Human Immune Response Dynamics (HIRD) study	12
2.1.4 Chapter summary	12
2.2 Methods	13
2.2.1 Existing HIRD study data and additional data	13
2.2.2 Genotype data generation	14
2.2.3 Genotype data preprocessing	14
2.2.4 Computing genotype PCs as covariates for ancestry .	15
2.2.5 Genotype phasing and imputation	15
2.2.6 RNA-seq data generation	18
2.2.7 RNA-seq quantification and filtering	18
2.2.8 Array data preprocessing	20
2.2.9 Computing baseline-adjusted measures of antibody response	22
2.2.10 Differential gene expression	22
2.2.11 differential gene expression meta-analysis	26
2.2.11.1 Cross-platform meta-analysis methods	26
2.2.11.2 Prior for between-studies heterogeneity	27
2.2.11.3 Prior for DGE effect size	29
2.2.11.4 Meta-analysis using bayesmeta	30
2.2.12 Gene set enrichment analysis	30
2.3 Results	30
2.3.1 Innate and adaptive immune response to Pandemrix .	30
2.3.1.1 Comparison to Sobolev et al.	30
2.3.2 Expression associated with antibody response	30
2.3.2.1 Comparison to Sobolev et al.	30
2.3.3 TODO Identifying molecular signatures for predicting antibody response	30
2.4 Discussion	34
2.4.1 Comparison to Sobolev R vs. NR	34
2.4.2 Inflammatory signatures of non-response	34

3 Genetic factors affecting Pandemrix vaccine response	35
3.1 Introduction	35
3.1.1 Genetic factors affecting influenza vaccine response . .	36
3.1.2 Context-specific immune response QTLs for influenza vaccine response	36
3.1.3 Chapter summary	36
3.2 Methods	36
3.2.1 Genotype imputation	36
3.2.2 Estimation of cell type abundances	36
3.2.3 Mapping cis-eQTLs with LMM	37
3.2.3.1 Expression normalisation	38
3.2.3.2 Finding hidden confounders with PEER . . .	38
3.2.4 eQTL mapping with mixed models	38
3.2.5 eQTL meta-analysis	38
3.2.5.1 Joint mapping with mashr	39
3.2.6 Defining shared and response eQTLs	40
3.3 Results	40
3.3.1 Overview of eQTLs at each timepoint	40
3.3.1.1 Estimation of eQTL sharing	40
3.3.1.2 TODO Replication of shared eQTLs in whole blood	40
3.3.2 Characterising re-eQTLs at each timepoint	40
3.3.3 The mechanism of reQTLs	40
3.3.4 TODO Colocalisation of re-eQTLs with known context-specific immune QTLs	40
3.3.5 TODO Disruption of binding site motifs as a model for re-eQTLs	40
3.4 Discussion	41
3.4.1 limitations: The mechanism of reQTLs	41
3.4.2 Conditional eQTL effects	41
4 Response to live attenuated rotavirus vaccine (Rotarix) in Vietnamese infants	43
4.1 Introduction	43
4.1.1 The genetics of vaccine response in early life	44
4.1.2 Rotavirus and rotarix in Vietnam	44

4.1.3 Known factors that affect rotavirus vaccine efficacy	44
4.2 Methods	44
4.2.1 RNA-seq data generation	44
4.2.2 Genotyping	44
4.3 Results	44
4.4 Discussion	44
5 multiPANTS	45
5.1 Introduction	45
5.2 Methods	45
5.2.1 Covariates to use	45
5.3 Results	45
5.4 Discussion	45
6 Discussion	47
A Supplementary Materials	49
A.1 Chapter 2	49
A.2 Chapter 3	49
A.3 Chapter 4	50
Bibliography	51
List of Abbreviations	55

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	Sample filters for missingness vs heterozygosity rate.	16
2.3	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.	17
2.4	FastQC sequence quality versus read position for Human Immune Response Dynamics (HIRD) RNA-seq samples. . . .	18
2.5	FastQC sequence duplication levels for HIRD RNA-seq samples.	19
2.6	FastQC GC profile for HIRD RNA-seq samples.	19
2.7	Distribution of short ncRNA and globin counts as a proportion of total counts in RNA-seq samples.	21
2.8	Proportion of samples in which genes are detected with non-zero counts. Vertical line shows 5% threshold.	21
2.9	Distribution of gene expressions for sample before and after filtering low expression genes. Vertical line shown at log(0.5) counts per million (CPM).	22
2.10	How TRI corrects fold changes for baseline titre.	23

2.11 TRI correlates with the standard responder definition (colored, 4-fold increase in either assay). An individual's TRI is the mean of their Z-transformed residuals from regressions of day 63 vs. day 0 fold-change against day 0 titre, over the two assays.	24
2.12 TRI vs platform	24
2.13 First 2 expression PCs	25
2.14 Fold-change comparison between array and RNAseq for day 1 vs day 0.	31
2.15 Priors for day 1 vs day 0 DGE meta-analysis.	31
2.16 Normalised gene expression for differentially expressed genes (adj. p < 0.05, $ \log_2 \text{FC} > 1.5$) across 208 RNA-seq samples from days 0, 1, and 7, clustered by gene.	32
2.17 Transcriptomic modules enriched in genes with expression positively and negatively associated with TRI. Blank cells n.s.	33

List of Tables

2.1 Transcriptomic modules enriched in highly up/downregulated genes in each expression cluster, based on ranking of $\log_2 FC$ vs. day 0. Blank cells n.s.	33
--	----

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
 - imputation
- 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
 - 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

CHAPTER 1. INTRODUCTION. IMMUNITY IS A COMPLEX TRAIT

- 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, $\approx 50\%$ variation in immune system driven by non-genetic factors, $\approx 30\text{--}40\%$ 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 cita-
tions

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,
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

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

1.5. IMMUNE RESPONSE TO BIOLOGICAL THERAPIES

[<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/PMC5350707/>

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

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

CHAPTER 2. TRANSCRIPTOMIC RESPONSE TO INFLUENZA A

2.1. INTRODUCTION (H1N1)PDM09 VACCINE (PANDEMRIX)

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

2.1.2 Response to influenza vaccines

- classical immunological response to influenza vaccines
 - differ by type of vaccine
- correlates of protection
 - mention HAI/MN assays
- Review influenza vaccine specific sysvacc papers (e.g. Nakaya's papers)
 - mention main timepoints
 - inclu. prevaccination signatures paper, and

2.1.3 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 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.

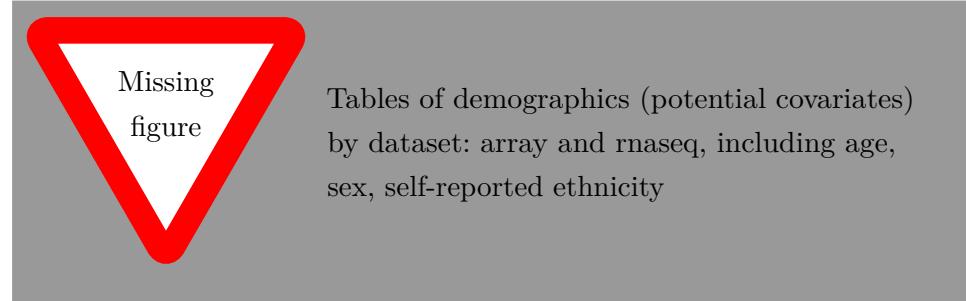
2.2 Methods

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).

ref demographics
table

In addition to the existing data, we generated array genotype data for 169 individuals; and RNA-seq data at days 0, 1, and 7, for 75 individuals not included in the microarray subset. An overview of all datasets is shown in Fig. 2.1.



Add to collab note

2.2.2 Genotype data generation

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.

2.2.3 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

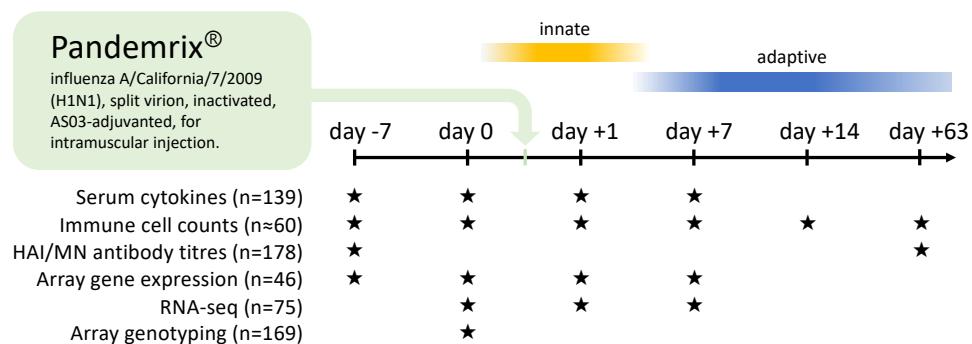


Figure 2.1: Data types, timepoints, and sample sizes. Individuals were vaccinated after day 0 sampling. Vaccine-induced antibodies measured by HAI and MN assays. Array and RNA-seq gene expression measured in the PBMC compartment.

markers < 1%, max marker heterozygosity rate within 3 standard deviations of the mean (threshold selected visually to exclude outliers, Fig. 2.2), 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.

After filtering, 169 samples and 549414 markers remained.

2.2.4 Computing genotype PCs as covariates for ancestry

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^{6–8}. 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. 2.3), hence these PCs can be used as covariates for ancestry downstream.

ref demographics table

Add Tracy-Widom statistics for PCs

2.2.5 Genotype phasing and imputation

Prior to imputation, 213277 markers with MAF < 0.001 (no variation in HIRD) 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.

¹<https://imputation.sanger.ac.uk/>

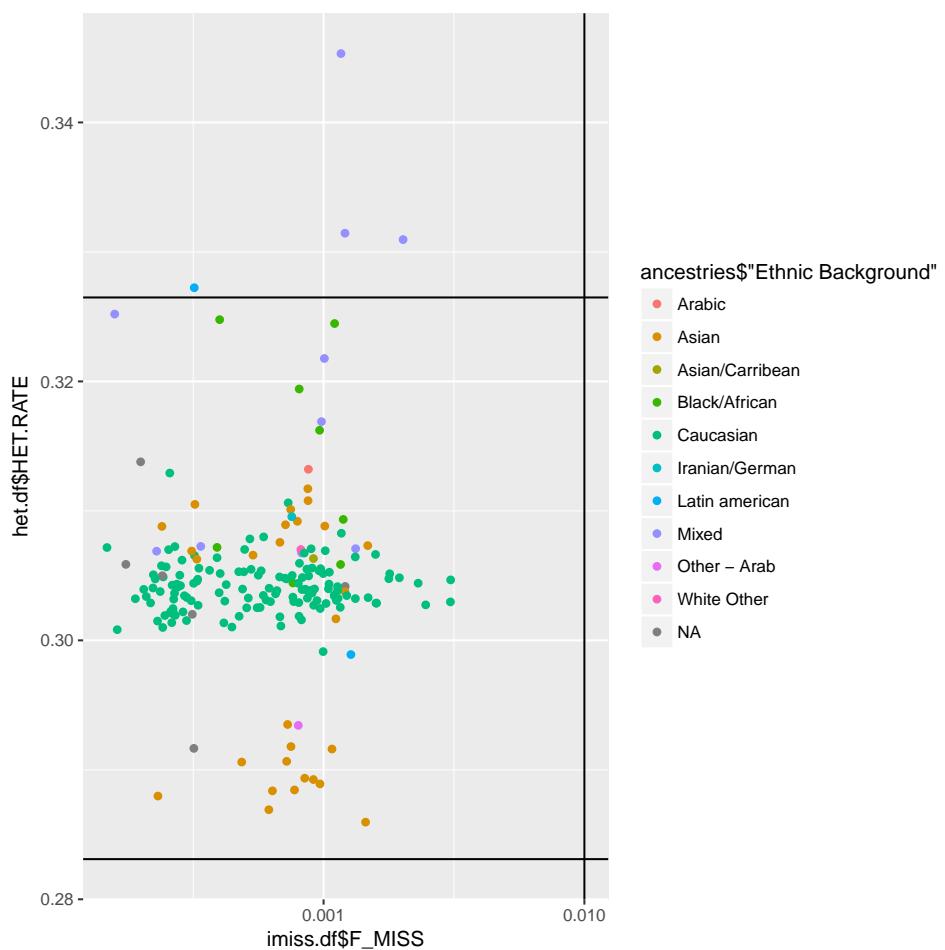


Figure 2.2: Sample filters for missingness vs heterozygosity rate.

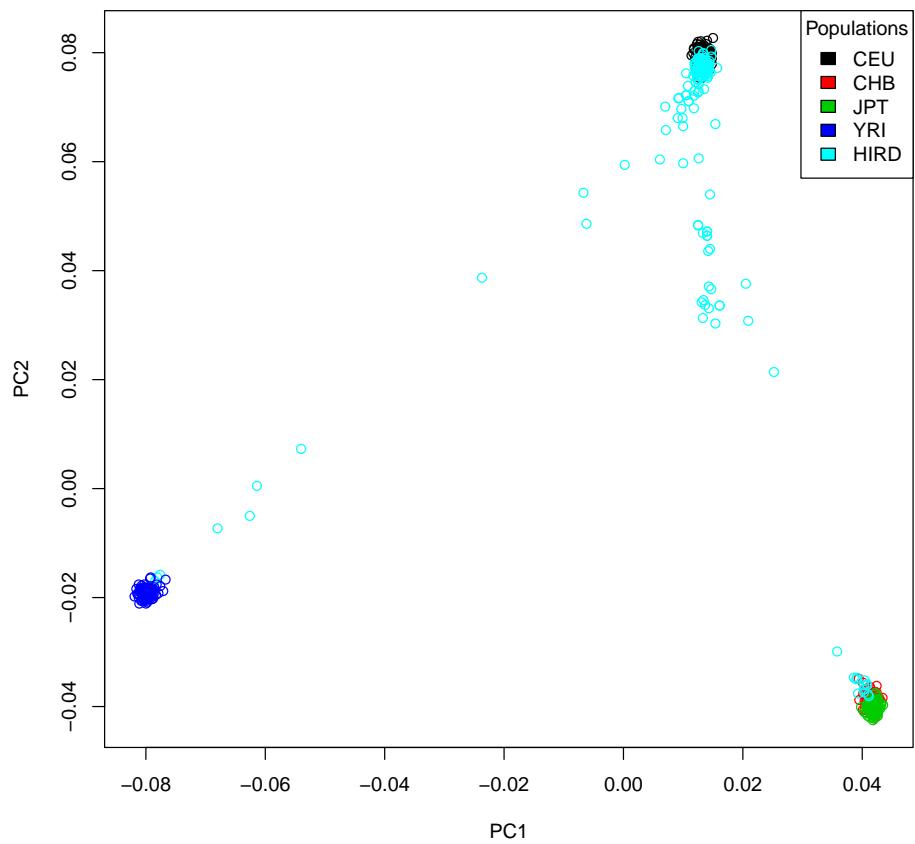


Figure 2.3: HIRD samples (cyan) projected onto PC1 and PC2 axes defined by PCA of HapMap samples. The first two PCs separate European (CEU, upper-right) from Asian (CHB and JPT, lower-right) and African (YRI, lower-left) populations.

2.2.6 RNA-seq data generation

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

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

RNA-seq quality metrics were assessed using FASTQC² and Qualimap⁹, then visualised with MultiQC¹⁰. Sequence quality was high (Fig. 2.4), and duplication levels were low (Fig. 2.5). The unimodal GC-content distribution suggested negligible levels of non-human contamination (Fig. 2.6).

add software versions

2.2.7 RNA-seq quantification and filtering

Reads were quantified against the Ensembl reference transcriptome (GRCh38) using Salmon¹¹ 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%

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

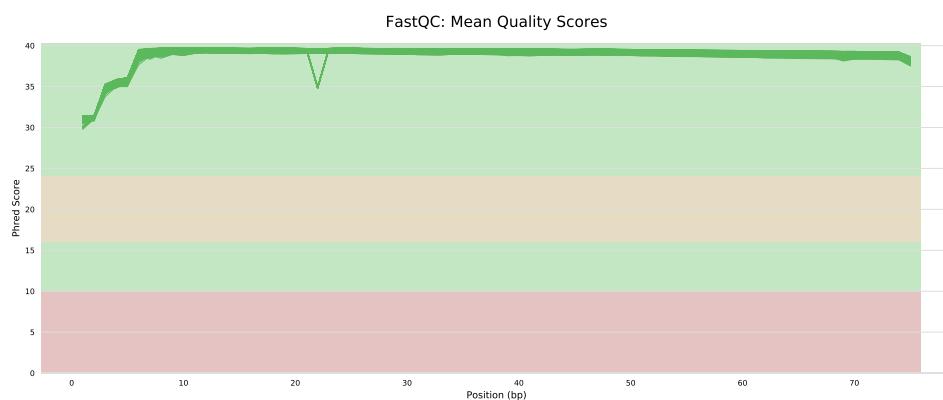


Figure 2.4: FastQC sequence quality versus read position for HIRD RNA-seq samples.

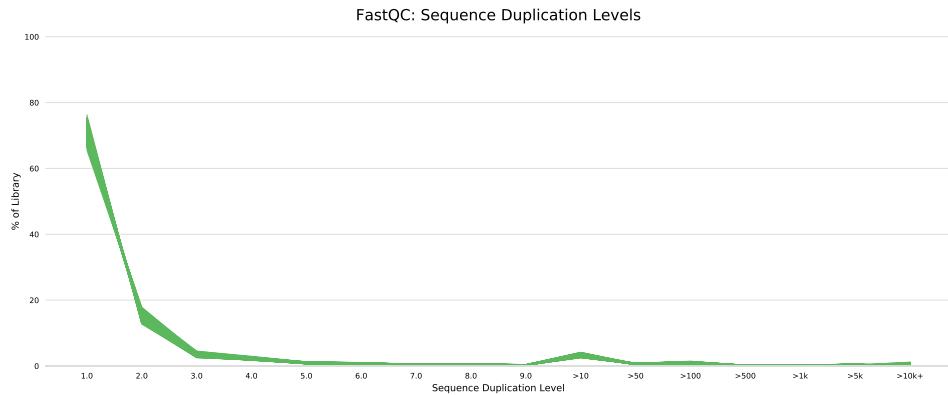


Figure 2.5: FastQC sequence duplication levels for HIRD RNA-seq samples.

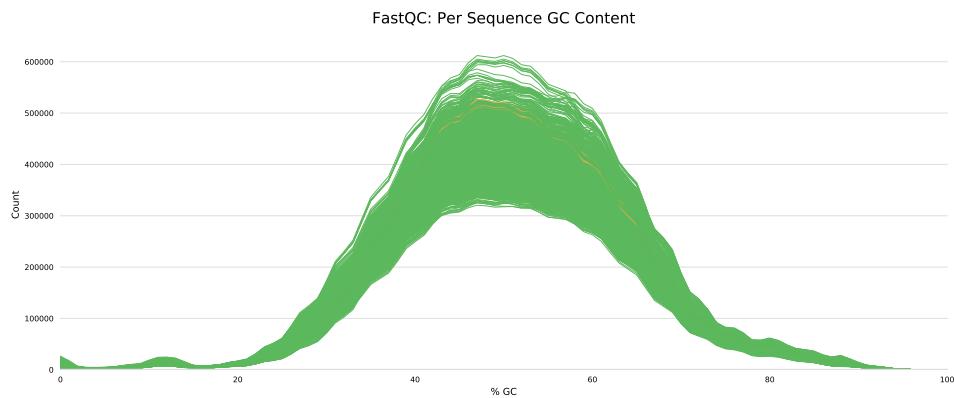


Figure 2.6: FastQC GC profile for HIRD RNA-seq samples.

(range 75.57-90.10%), comfortably within sequencing depth recommendations for **differential gene expression (DGE)** experiments¹². Relative transcript abundances were summarised to Ensembl gene-level count estimates using tximport to improve statistical robustness and interpretability¹³.

Genes with short noncoding RNA biotypes³ were removed, as they are generally not polyadenylated, and expression estimates can be biased by misassignment of counts from overlapping protein-coding or lncRNA genes¹⁴. Globin genes, which are highly expressed in erythrocytes and reticulocytes, cell types expected to be depleted in **PBMC**¹⁵, were also removed. Given the proportion of removed counts at this stage was low most samples (**Fig. 2.7**), 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. 2.8**), 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 expressed¹⁶. 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. 2.9**).

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

2.2.8 Array data preprocessing

Single-channel Agilent 4x44K microarray (G4112F) data for 173 samples from⁵ were downloaded from <https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-2313/>. VSN¹⁷ was used to perform background correction for non-specific hybridisation, between-array normalisation, and variance-stabilisation of intensity values.

Most genes are covered by multiple array probes; 31208 probes were collapsed into 18216 Ensembl genes using by selecting the probe with the

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>

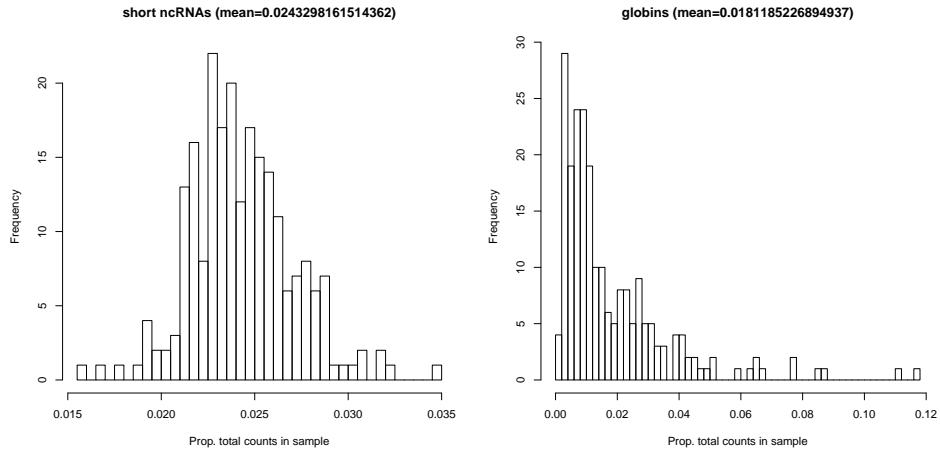


Figure 2.7: Distribution of short ncRNA and globin counts as a proportion of total counts in RNA-seq samples.

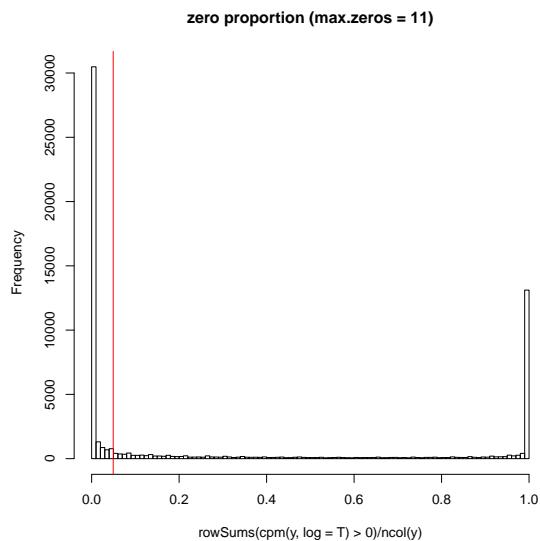


Figure 2.8: Proportion of samples in which genes are detected with non-zero counts. Vertical line shows 5% threshold.

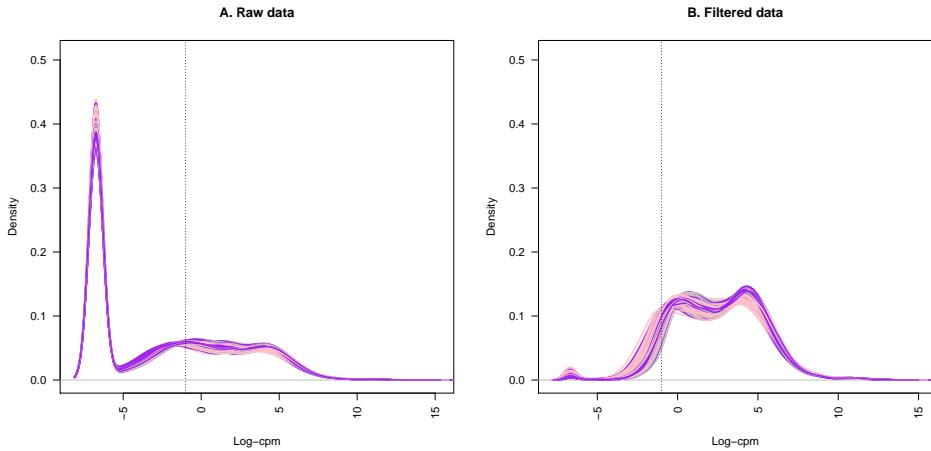


Figure 2.9: Distribution of gene expressions for sample before and after filtering low expression genes. Vertical line shown at $\log(0.5)$ CPM.

highest mean intensity for each gene (WGCNA::collapseRows, using the recommended MaxMean method)¹⁸. While it would be optimal to select a collapsing method to maximise the concordance between array and RNA-seq expression values, in this case there were no samples assayed by both platforms.

2.2.9 Computing baseline-adjusted measures of antibody response

We computed TRI for the HIRD dataset (Fig. 2.10) Relationship between TRI and the clinical responder definition used by Sobolev et al. (Fig. 2.11).

2.2.10 Differential gene expression

Batch effect correction (see batch effects Zotero tag) Major source of variation is batch: (Fig 2.13) Combat is best here. LM, LMM, Combat were comparable. LMM and LM correcting for batch provide a slightly safer option than Combat by identifying stronger relationships between big effect size and gene expression and better true/false positive rates for small effect size. In some cases, Combat overcorrects. Main issue is unbalanced design, which affects even 2-way anova. Rather than 2-step, Safest is to use a covariate, which seems to at least create appropriate confidence intervals (1e). Unbalanced batches: Batch 3 and batch 4 have smaller sample size than the other batches,

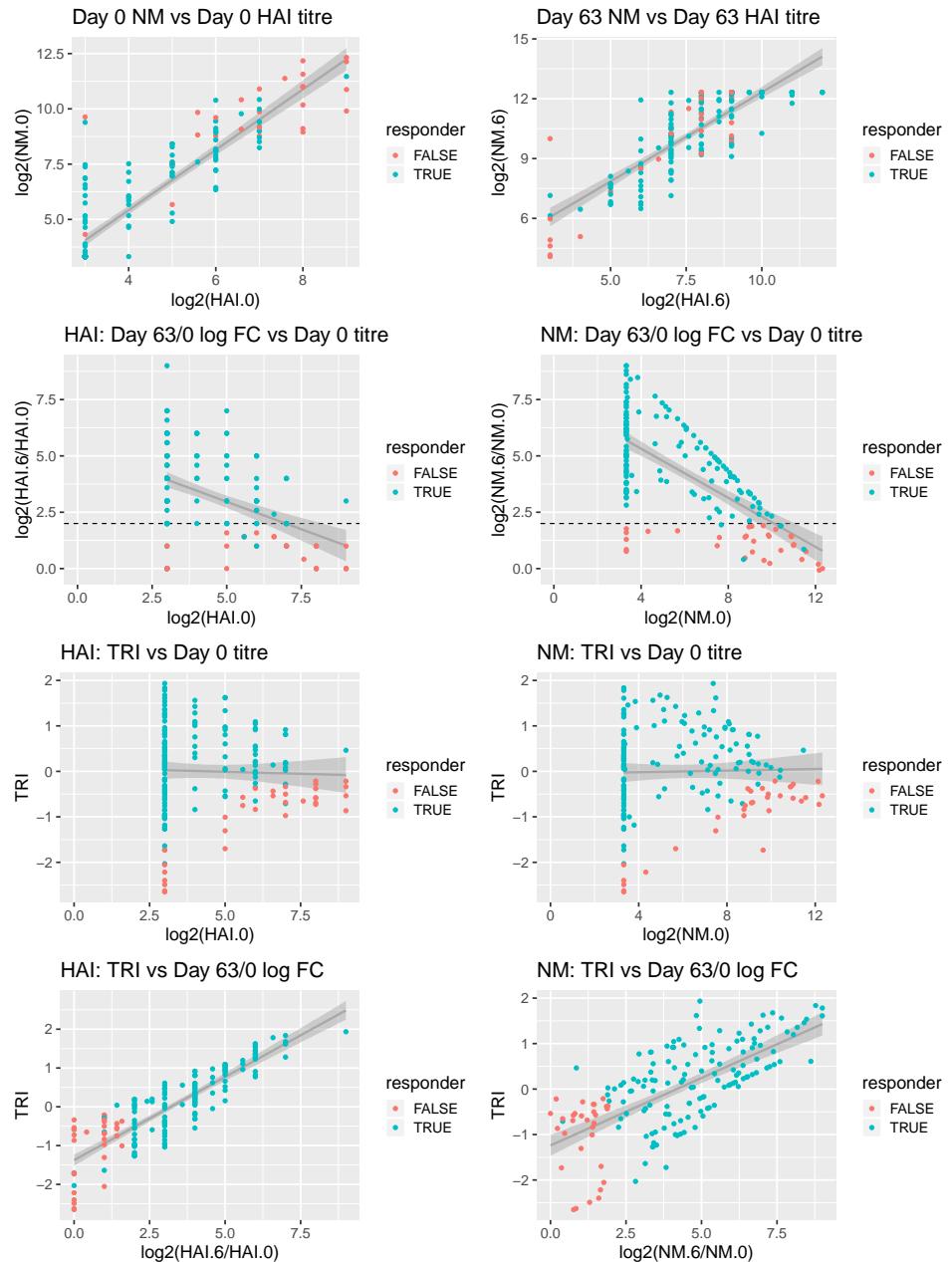


Figure 2.10: How TRI corrects fold changes for baseline titre.

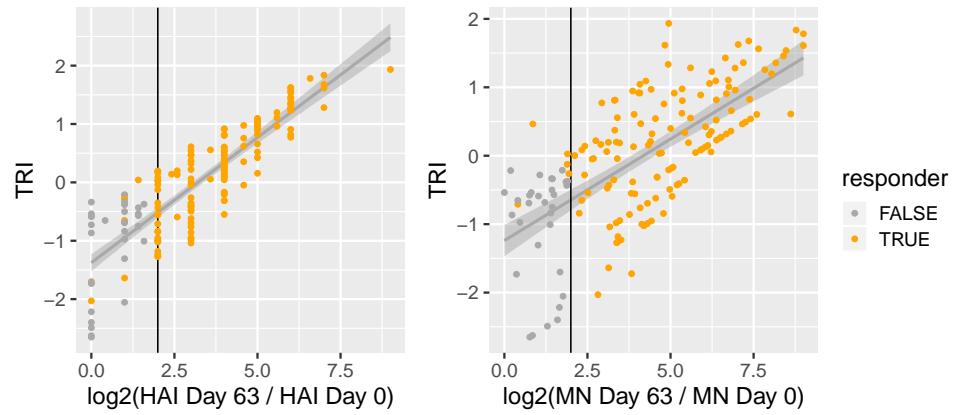


Figure 2.11: TRI correlates with the standard responder definition (colored, 4-fold increase in either assay). An individual's TRI is the mean of their Z-transformed residuals from regressions of day 63 vs. day 0 fold-change against day 0 titre, over the two assays.

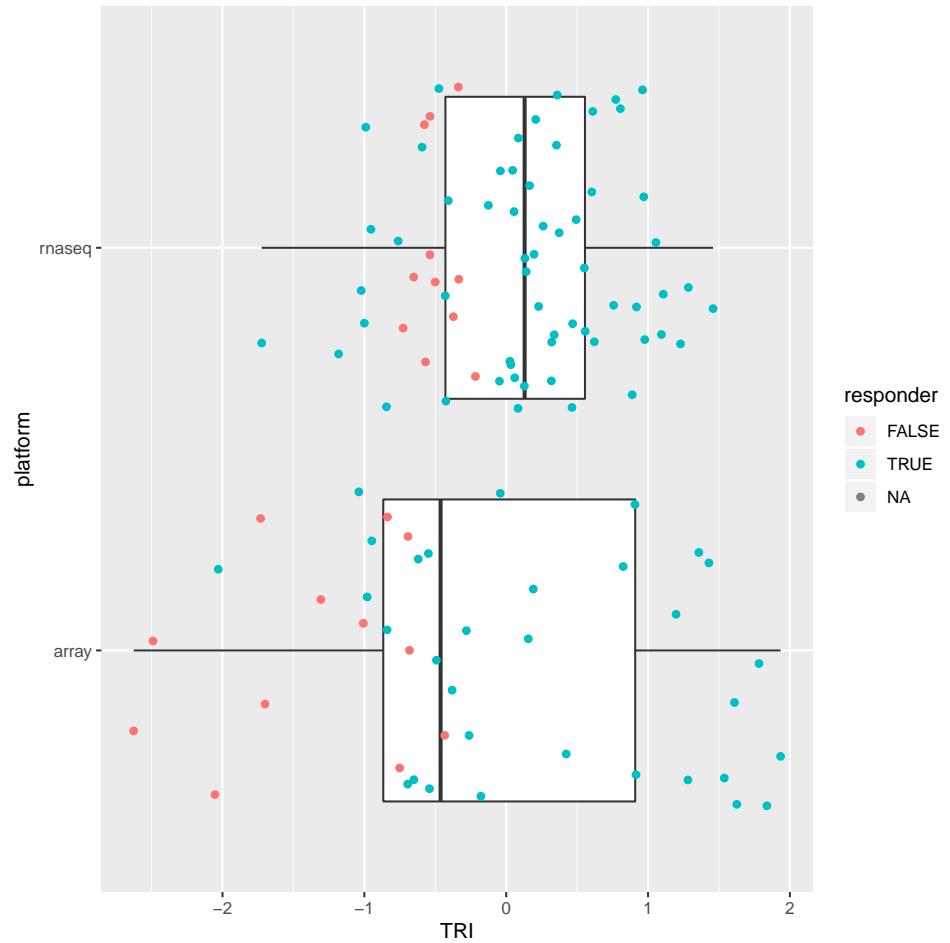


Figure 2.12: TRI vs platform

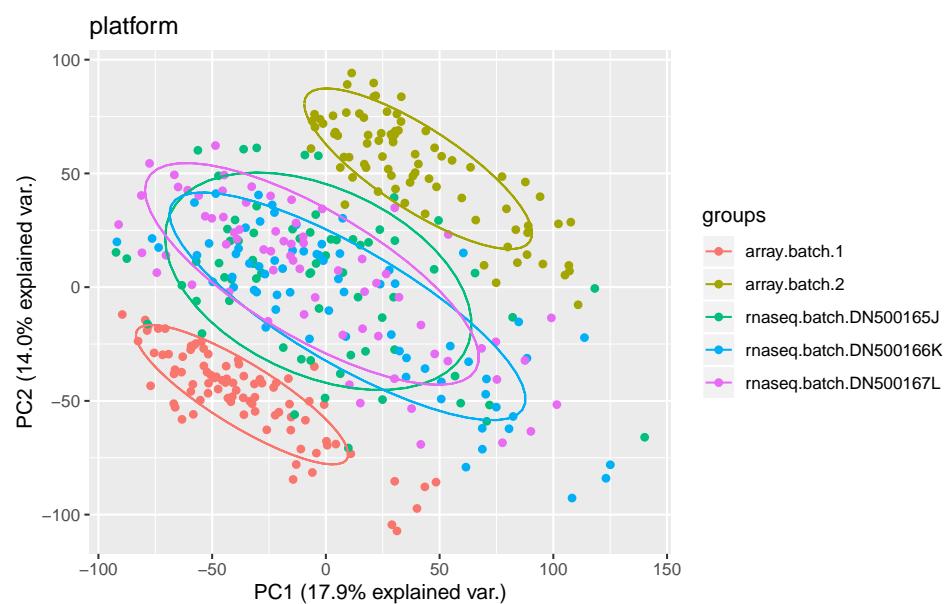


Figure 2.13: First 2 expression PCs

thus their variance estimates are impacted more by outlying samples.

Why limma over edgeR/DESeq2? Comparable at sufficient sample sizes, and faster.

Why combine -7 and 0? See Sobolev: (a) Observed values of multivariate statistic t (m.v.t.) quantifying global PBMC gene-expression dissimilarity in comparison of two study days (red dots) to values expected when days are randomly assigned between groups.

Equation for linear models used in limma.

2.2.11 differential gene expression meta-analysis

13945

Should we meta-analyse? "In conclusion, we found that underpowered studies play a very substantial role in meta-analyses reported by Cochrane reviews, since the majority of meta-analyses include no adequately powered studies. In meta-analyses including two or more adequately powered studies, the remaining underpowered studies often contributed little information to the combined results, and could be left out if a rapid review of the evidence is required."

2.2.11.1 Cross-platform meta-analysis methods

- Whilst there is a slew of literature on meta-analysis of rnaseq and array (e.g. metaMA), combining platforms is fraught with difficulties.
 - different tech -> diff statistical models
- Expected heterogeneity:
 - Platform effect (ratio compression, differences in preprocessing to genes).
 - Different sets of samples (more extreme in array)
- Examples of past meta:
 - sva: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3617154/>
 - MetaVolcano: vote counting, REM (note small k)
 - CorMotif first applies limma (Smyth, 2004) to each study separately.

- * CorMotif for microarray data since it was motivated by the microarray analysis in the SHH study. However, the idea behind CorMotif is general, and it should be straightforward to develop a similar framework for RNA-seq data.
- CBM (“Cross-platform Bayesian Model”), see CBM paper for discussion of difficulties of combining platform
 - * cannot actually use CBM, as it operates on expressions, with a binary case vs control, so no covariates
 - * same limitation for cormotif, although it takes any number of groups
- Rankprod (focus on case/control design)
- Mayday seasight
- Classic models: Two schools of thought for frequentist meta-analysis:
 - fixed-effect
 - or in the presence of het, random-effects.
 - * e.g. random effects model of approx 24 datasets
 - We have het, so def use random effects.
- How to estimate het?
 - Many methods and estimators.
 - The problem: we only have k=2, and MLE estimates of tau are not very good with k=2.
 - * Sweeny tests the effect of varying k.
 - * Highly imprecise, and often boundary estimate problems, and we know 0 het is inappropriate.
- Bayesian random-effects meta is attractive, but what priors should we use?

2.2.11.2 Prior for between-studies heterogeneity

Prior for tau.

- A general rec is: Use distribution in the half-t family e.g. Cauchy ($df=1$) when the number of groups is small and in other settings where a weakly-informative prior is desired.
 - In their 3-schools examples, choose a value of scale just higher than expected, this is to weakly constrain the posterior, and not to actually represent prior knowledge.
 - Warn against inverse-gamma(e, e), as it can influence the posterior mean.
- But weak priors are not recommended, as k is small, so there is little information in the data.
- We can get empirical distribution of many genes.
 - fit a default reml model, exclude 0 ests.
- Advantage of getting the correct parameter scale for our data.
- So use Empirical Bayes:
 - aside: empirical bayes is popular for high dim data e.g. edgeR, DESeq2, limma-voom, combat (method of moments)
- Papers that fit empirical datasets for tau2: Most of these are inverse-gamma/log-t family
 - Fit inverse gamma distribution on method of moments estimates from 18 gastroenterology trials with similar endpoints.
 - This paper has described the distribution of the between-study variance amongst Cochrane reviews published between 2008 and 2009, and investigating a binary outcome. A log-normal distribution incorporating the association between the between-study variance and the pooled effect size gave the best fit.
 - Predictive distributions are presented for nine different settings, defined by type of outcome and type of intervention comparison. For example, for a planned meta-analysis comparing a pharmacological intervention against placebo or control with a subjectively measured outcome, the predictive distribution for heterogeneity is a log-normal (2.13, 1.582) distribution, which has a median value of 0.12.

- Model selection based on the deviance information criterion (DIC) [8] led to the choice of the log-t model for t2. (5df)
- The priors are derived as log-normal distributions for the between-study variance, applicable to meta-analyses of binary outcomes on the log odds-ratio scale.
- We choose gamma: as Density at tau=0 is 0, but increases linearly from 0, so values close to 0 are still permitted if the data suggests it.
 - For lognormal/inverse gamma, they have a derivative of 0 at tau=0, so they rule out small tau no matter what the data suggest.
 - For The exponential and half-Cauchy families, for example, do not decline to zero at the boundary, so they do not rule out posterior mode estimates of zero.

2.2.11.3 Prior for DGE effect size

Prior for logFC

- Not as much discussion in the lit:
- There is Typically enough data to estimate this to use a non informative prior.
- Even Friede uses noninformative flat.
- Two choices in bayesmeta are uniform and normal.
 - We know Mean is 0: most genes are not DE, so flat prior makes no sense
- To avoid overshrinking, could consider heavy-tailed priors (e.g. cauchy) for mu rather than normal
 - Cauchy 2.5
 - DEseq/apeglm: prior on logfc, cauchy with scale adapted.
- But this is not possible in bayesmeta, bayesmeta is normal. So weaken further to place more prior on larger values. This means less shrinkage.
- Also: we will shrink again with ash, which can fit a more complicated (mixture?) distr

- So we use a very weak normal prior, scaled to each coef, as we still want some scaling based on parameter scales.
 - Equiv to saying 95pc chance that effect is within 20sds (of what?).

2.2.11.4 Meta-analysis using bayesmeta

PPI possible, but too slow.

2.2.12 Gene set enrichment analysis

tmod; gprofileR; CAMERA

2.3 Results

2.3.1 Innate and adaptive immune response to Pandemrix

Overall response clusters into two distinct patterns (Fig. 2.16).

Day 1 response is characterised by innate response: monocyte genes, inflammatory response, type I interferon response. Note type I interferons are alpha/beta, not gamma. Day 7 response is characterised by adaptive B cell response: plasma cell genes, immunoglobulins, proliferation (Table 2.1).

2.3.1.1 Comparison to Sobolev et al.

2.3.2 Expression associated with antibody response

- Overall, B cell module positively associated, inflammatory modules negatively associated with TRI (Fig).
- Signatures split by day

2.3.2.1 Comparison to Sobolev et al.

Overlap of day 7 R vs NR DGE genes with original pipeline

2.3.3 TODO Identifying molecular signatures for predicting antibody response

- For inference, don't dichotomise due to statistical concerns

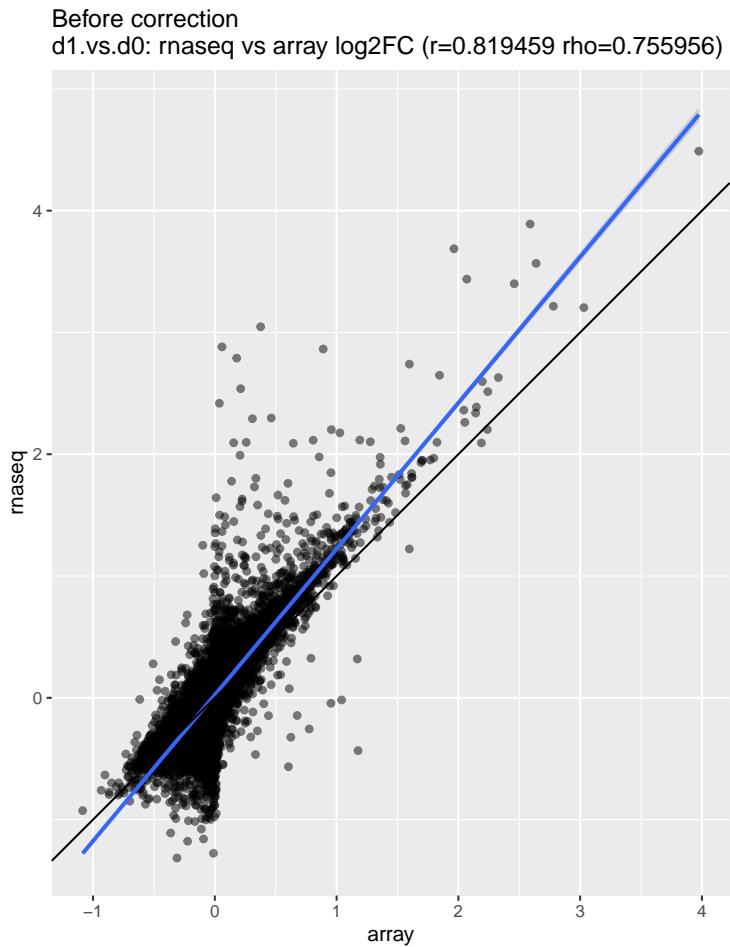


Figure 2.14: Fold-change comparison between array and RNAseq for day 1 vs day 0.

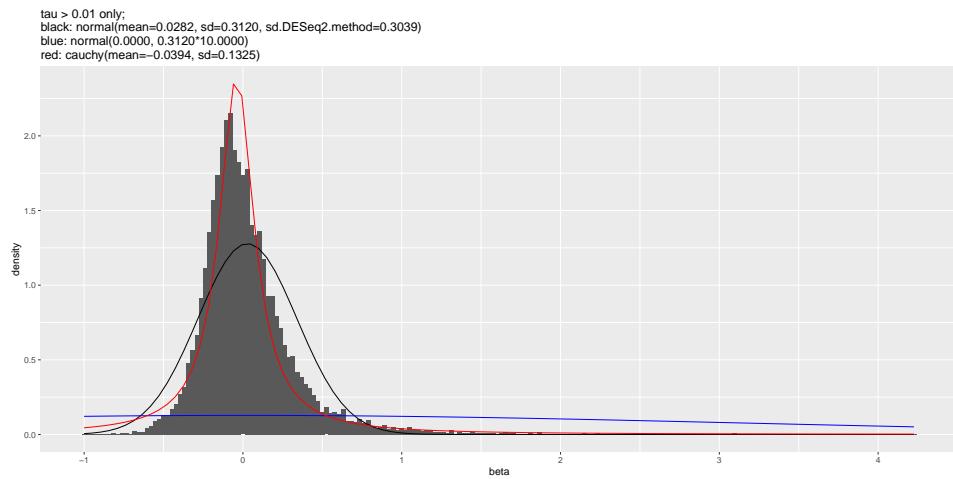


Figure 2.15: Priors for day 1 vs day 0 DGE meta-analysis.

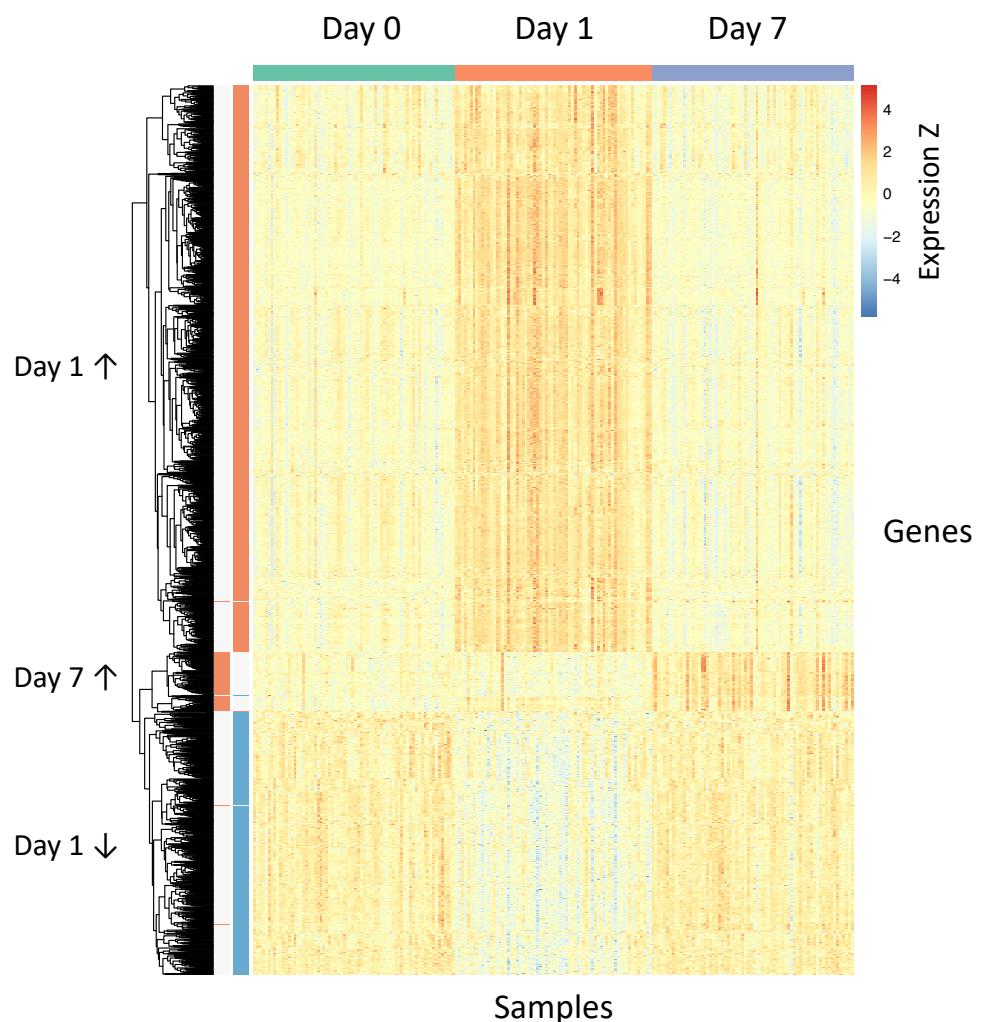


Figure 2.16: Normalised gene expression for differentially expressed genes (adj. $p < 0.05$, $|\log_2 \text{FC}| > 1.5$) across 208 RNA-seq samples from days 0, 1, and 7, clustered by gene.

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

2.3. RESULTS

Table 2.1: Transcriptomic modules enriched in highly up/downregulated genes in each expression cluster, based on ranking of \log_2 FC vs. day 0. Blank cells n.s.

Module	adj. p value		
	Day 1 ↑	Day 1 ↓	Day 7 ↑
cell cycle and transcription	2.3×10^{-36}		7.7×10^{-61}
immune activation - generic cluster	1.4×10^{-32}		
enriched in monocytes	2.9×10^{-90}		
TLR and inflammatory signaling	1.7×10^{-28}		
type I interferon response	8.9×10^{-13}		
cell division stimulated CD4+ T cells			3.5×10^{-18}
PLK1 signaling events			2.7×10^{-25}
plasma cells, immunoglobulins			5.8×10^{-12}
enriched in NK cells		4.9×10^{-50}	
enriched in T cells		3.8×10^{-46}	
T cell activation		8.7×10^{-29}	

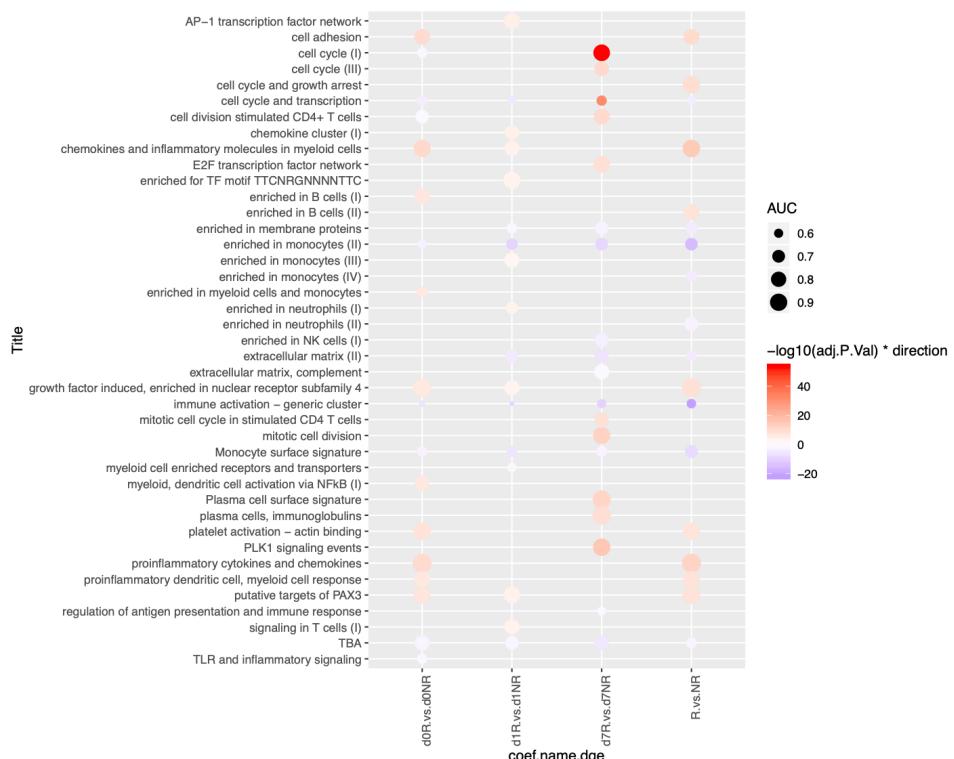


Figure 2.17: Transcriptomic modules enriched in genes with expression positively and negatively associated with TRI. Blank cells n.s.

- "In clinical studies seroprotection is normally defined as a specific antibody titer or antibody titer increase (seroconversion)."
- For prediction, what rules can be easily implemented in the clinic?
- Interpretability: DAMIP gives rulesets composed of small sets of genes, amenable to rapid qPCR assays.

2.4 Discussion

- Recap of results with limitations
- Cannot directly separate adjuvant effect

2.4.1 Comparison to Sobolev R vs. NR

- Differences between array-only and rnaseq-only DGE results for R/NR comparison (see 1st year report).
- Caveats of the meta analysis
- Is the R vs NR bayesmeta -> lfsr method too stringent due to auto reg by prior?

2.4.2 Inflammatory signatures of non-response

"The reduced efficacy of vaccination has also been linked to excessive inflammation for influenza,³¹ yellow fever,³² tuberculosis,³³ and hepatitis B³⁴ vaccines."

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.

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 limitations: The mechanism of reQTLs

3.4.2 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 Find out optimal spline degree.

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

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.coि.2014.07.001](https://doi.org/10.1016/j.coि.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. 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).
7. 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).

APPENDIX A. BIBLIOGRAPHY

8. 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).
9. 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).
10. 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).
11. 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).
12. 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).
13. 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).
14. 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).
15. 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).
16. 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).

APPENDIX A. BIBLIOGRAPHY

17. 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).
18. 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).

APPENDIX A. BIBLIOGRAPHY

List of Abbreviations

CPM counts per million

DGE differential gene expression

eQTL expression quantitative trait locus

HAI haemagglutination inhibition

HIRD Human Immune Response Dynamics

LD linkage disequilibrium

MAF minor allele frequency

MN microneutralisation

PBMC peripheral blood mononuclear cell

PC principal component

PCA principal component analysis

reQTL response expression quantitative trait locus

RNA-seq RNA-sequencing

Todo list

add more pros for in vitro reQTLs here, and find citations	6
ref demographics table	13
Figure: Tables of demographics (potential covariates) by dataset: array and rnaseq, including age, sex, self-reported ethnicity	14
Add to collab note	14
ref demographics table	15
Add Tracy-Widom statistics for PCs	15
Can add other fastqc plots e.g. kmers, overrepresented seqs, seq length	18
add software versions	18