<title>

Benjamin Yu Hang Bai

April 6, 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

List of Tables

Chapter 1

Introduction

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

1.1 A brief overview of genetic association for complex traits

1.1.1 Early days

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

1.1.2 The advent of GWAS

- 10 years of GWAS
- "The case of the missing heritability"
- genotyping arrays
  - 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

- PheWAS[1]
- Fine-mapping
  - as sample sizes get larger, and provided that sequencing or imputation can more exhaustively identify all of the candidate SNPs on the haplotype, rare recombination events will pile up, helping to make the causal SNP stand out above the passenger SNPs that usually travel on its haplotype [Huang 2017].

  - tag snps: causal snps may not be directly typed, may need to be imputed

1.1.4 Interpretation of genetic associations with molecular studies

- Locus to gene mapping problem
  - nc snps
    - Genome-wide association studies have successfully identified genetic variants associated with immune-mediated disease, the majority of which are non-coding[10 Years of GWAS Discovery].
- using intermediate/endophenotypes
  - endophenotypes paper
  - expression as an important intermediate
  - 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

- inthe dreaded GxE interaction
  - "In genetics, context matters"
  - for both gwas, and molQTLs, context is key
- Architecture varies e.g. across cell type and tissues
  - 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 was[2], where eQTLs were mapped separately in monocyte-derived dendritic cells before and after 18h infection with *Mycobacterium tuberculosis*. reQTLs were detected for 198 genes, 102 specific to the uninfected state, and 96 specific to the infected state. These reQTLs were enriched for GWAS SNPs associated with host susceptibility to tuberculosis; this was not observed for eQTLs that were not reQTLs.

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

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

1.3.2.2 *in vivo* response QTL mapping

less popular A complementary approach.

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

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

Large cohorts:

1.4 Immune response to vaccination

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

Vaccines stimulate the immune system with pathogen-derived antigens to induce effector responses (primarily antigen-specific antibodies) and immunological memory against the pathogen itself. These effector responses are then be rapidly reactivated in cases of future exposure to the pathogen, mediating long-term protection.

1.4.1 Systems vaccinology: from empirical to rational vaccinology

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

- Vaccination coverage in vunerable populations is below optimal

However, a vaccine that is highly efficacious in one human population may have significantly lower efficacy in other populations. [1 statistic on vaccine efficiacy differences e.g. rotavirus] Particularly challenging populations for vaccination include the infants and elderly, pregnant, immuno comprimised patients, ethnically-diverse populations, and developing countries. For the majority of licensed vaccines, there is a lack of understanding regarding the molecular mechanisms that underpin this variation in host immune response. Immunological mechanisms that underpin a specific vaccine's success or failure in a given individual are often poorly understood[Immunological mechanisms of vaccination].

rational vacc, where the key is sys vacc

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

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

[http://dx.doi.org/10.1098/rstb.2014.0146]. 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/pubmed/18454680

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."

  - 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
  - seasinal flu vaccines are a constnat battle against flu virus antigenic dft
  - 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 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 in[5]. 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 medasured on days -7 and 63 using the HAI and MN assays, and various subsets of the cohort were

also profiled for serum cytokine levels (Luminex panel, days -7, 0, 1 and 7), immune cell 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

In[5], 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 titre response index (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

cite appropriate subfigures here

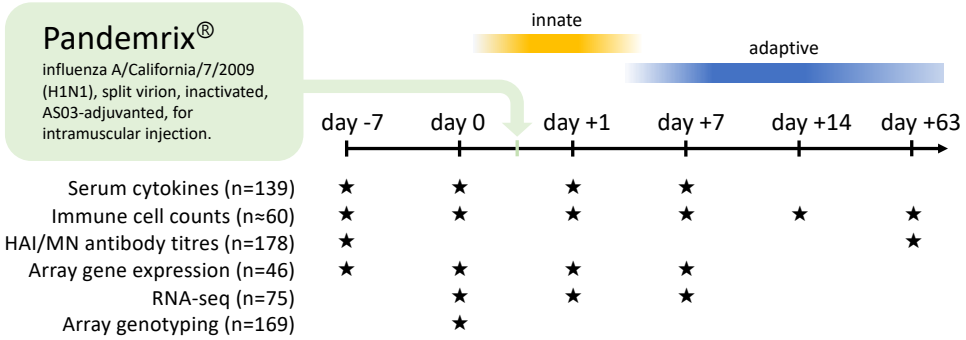

Figure 2.1: Data types, timepoints, and sample sizes. Individuals were vaccinated after day 0 sampling. Vaccine-induced antibodies measured by HAI and MN assays. Array and RNA-seq gene expression measured in the PBMC compartment.

$\log_2$ day -7 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).

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 on 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 exlude outliers, Fig. 2.4), removal of markers that deviate from Hardy–Weinberg equilibrium (`-hwe` option, $p < 0.00001$).

To exclude highly-related individuals and deduplicate replicate samples, pairwise kinship coefficients were computed on minor allele frequency (MAF) < 0.05 pruned genotypes using KING (v1.4). For each pair of samples with 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.

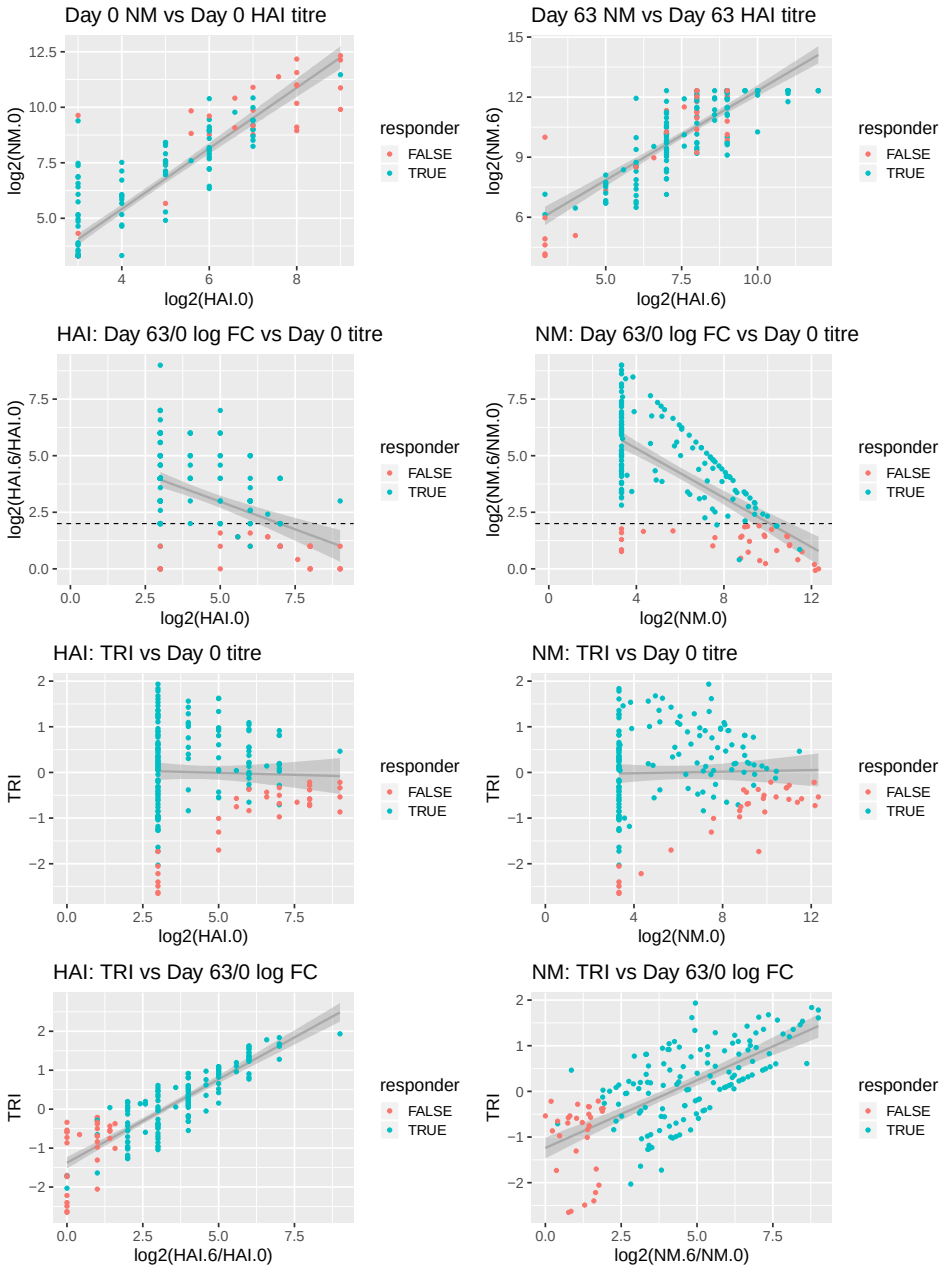

Figure 2.2: How TRI corrects fold changes for baseline titre.

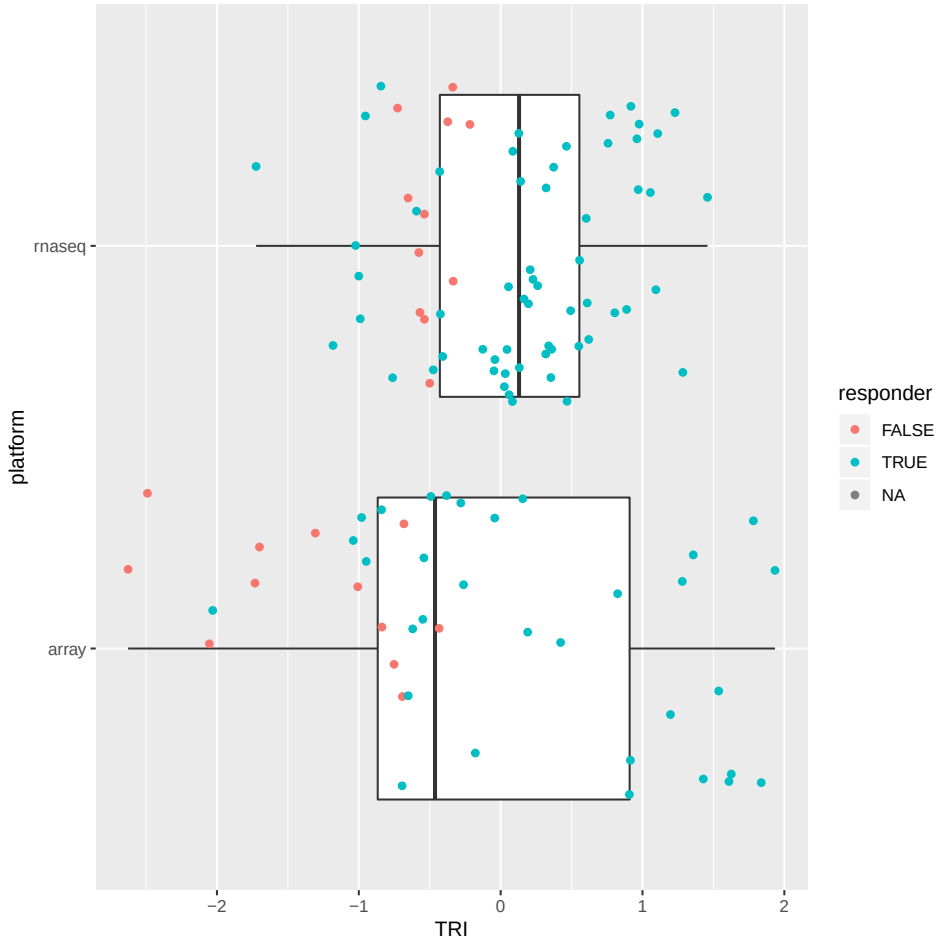

Figure 2.3: TRI vs platform

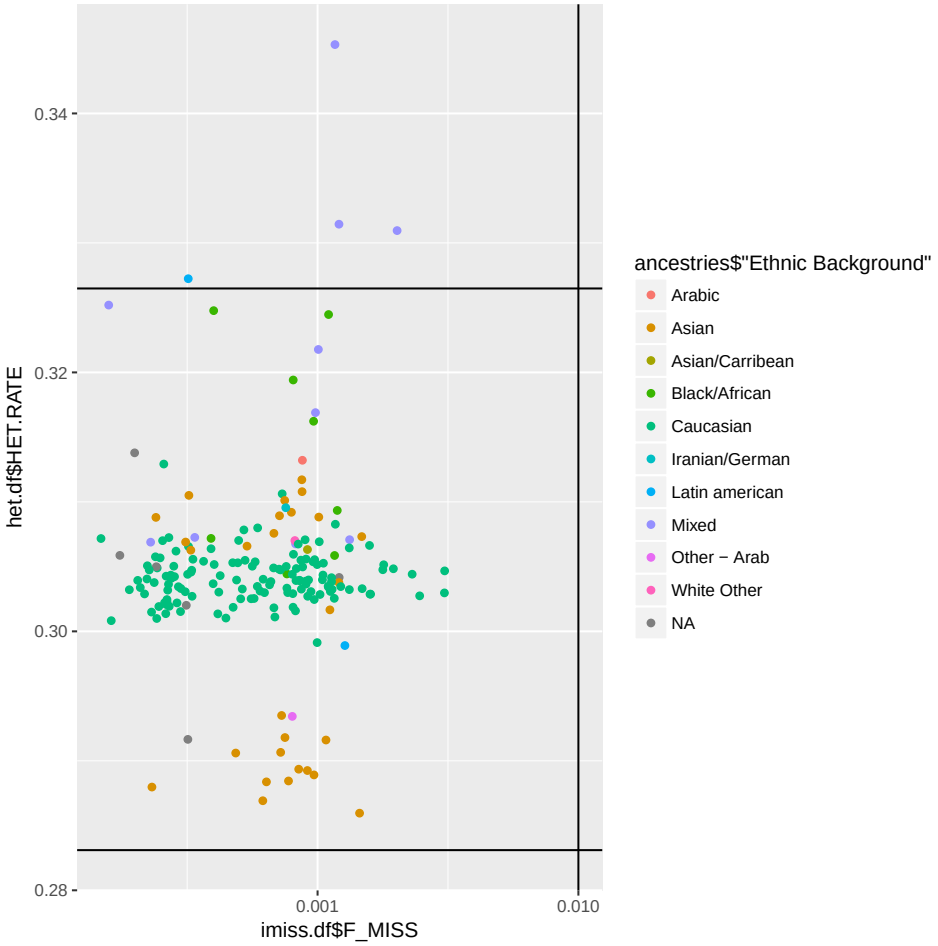

Figure 2.4: Sample filters for missingness vs heterozygosity rate.

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 markers with MAF < 0.001 (no variation in HIRD) were removed. Imputation for the autosomes and X chromosome was conducted using the Sanger Imputation Service[1], 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 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

[1] https://imputation.sanger.ac.uk/

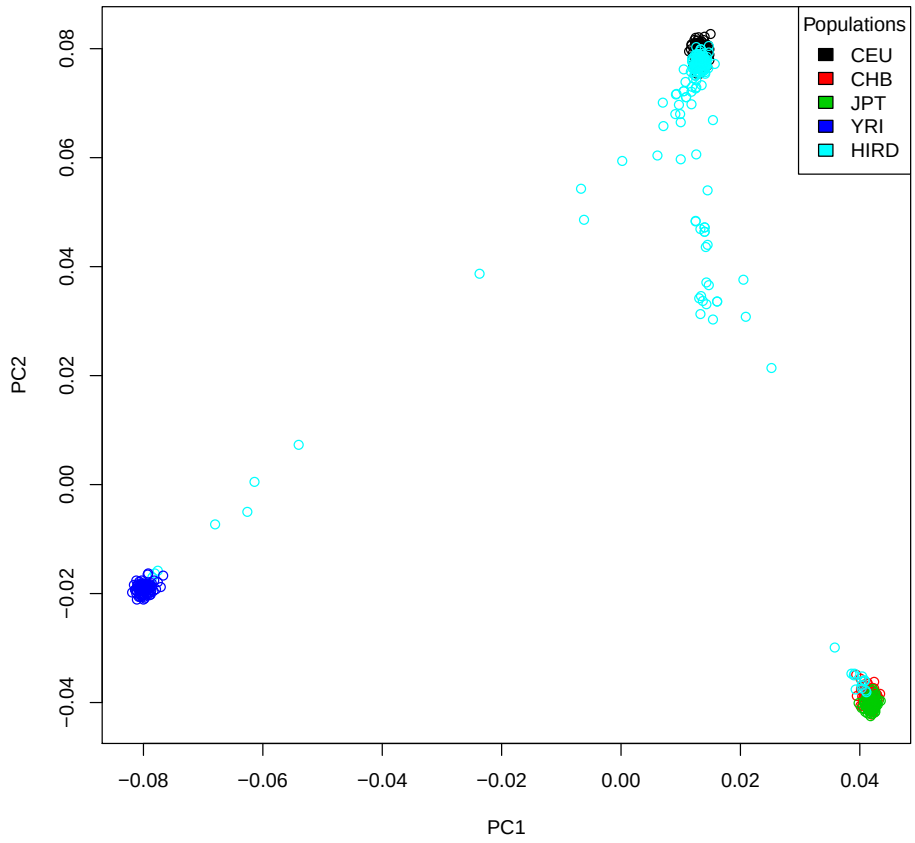

Figure 2.5: HIRD samples (cyan) projected onto PC1 and PC2 axes defined by PCA of HapMap samples. The first two PCs separate European (CEU, upper-right) from Asian (CHB and JPT, lower-right) and African (YRI, lower-left) populations.

RNA-seq quality metrics were assessed using FASTQC[2] and Qualimap[13], then visualised with MultiQC[14]. 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).

2.2.8 RNA-seq quantification and filtering

add software versions

Reads were quantified against the Ensembl reference transcriptome (GRCh38) using Salmon[15] 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) experiments[16]. Relative transcript abundances were summarised to Ensembl gene-level count estimates using tximport to improve statistical robustness and interpretability[17].

Genes with short noncoding RNA biotypes [3] were removed, as they are generally not polyadenylated, and expression estimates can be biased by misassignment of counts from overlapping protein-coding or lncRNA genes[18]. Globin genes, which are highly expressed in erythrocytes and reticulocytes,

[2] https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

[3] miRNA, 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.

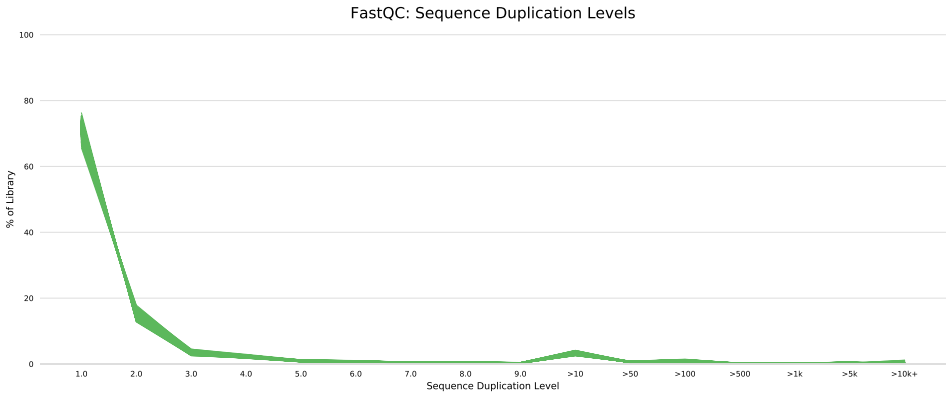

Figure 2.7: FastQC sequence duplication levels for HIRD RNA-seq samples.

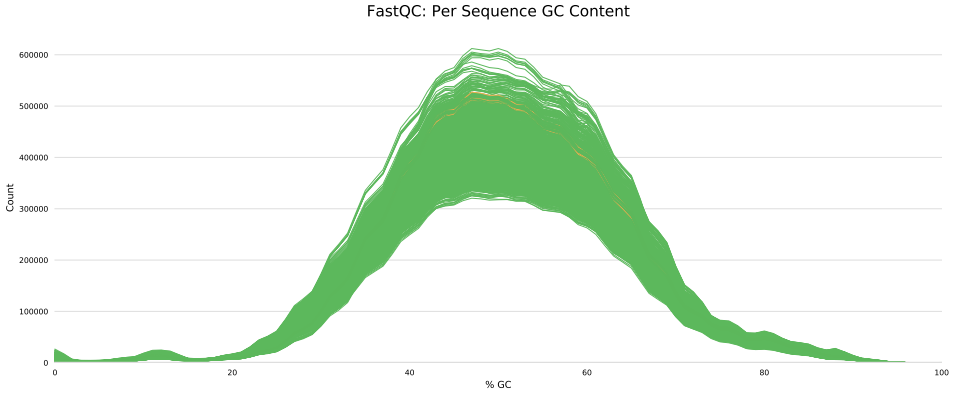

Figure 2.8: FastQC GC profile for HIRD RNA-seq samples.

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 expressed[20]. 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 from[5] were downloaded from `https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-2313/`. VSN[21] 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 $\log_2$ scale.

Most genes are covered 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.

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

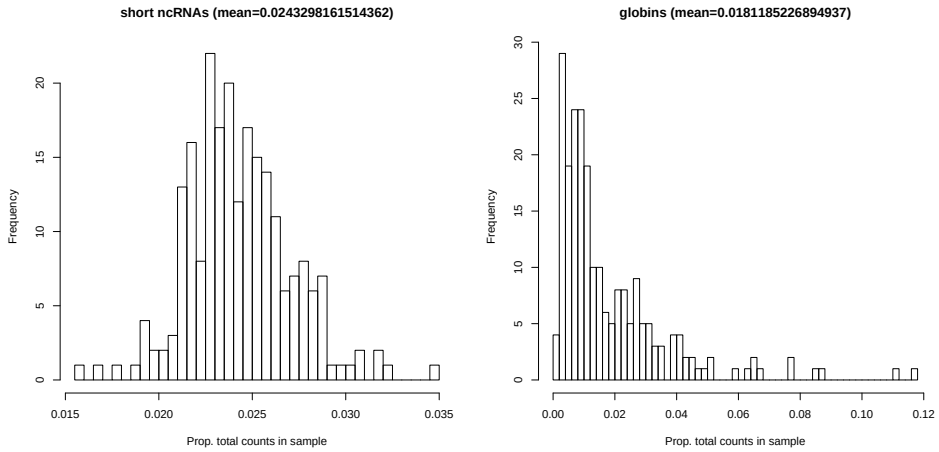

Figure 2.9: Distribution of short ncRNA and globin counts as a proportion of total counts in RNA-seq samples.

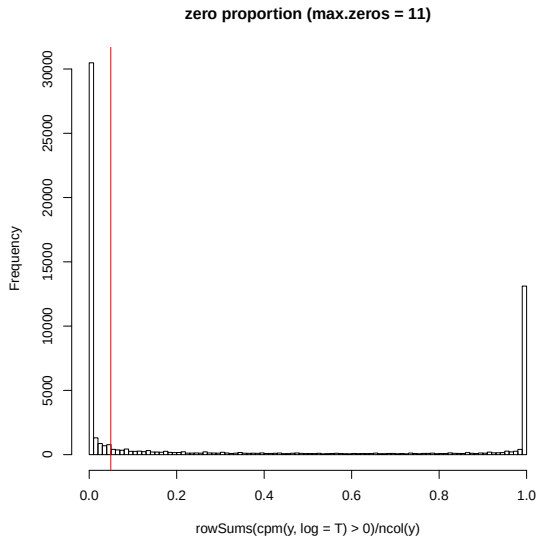

Figure 2.10: Proportion of samples in which genes are detected with non-zero counts. Vertical line shows 5% threshold.

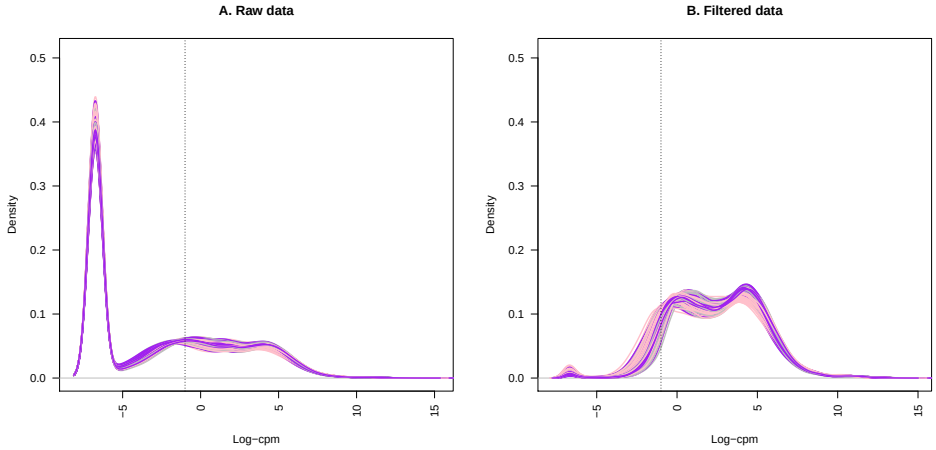

Figure 2.11: Distribution of gene expressions for sample before and after filtering low expression genes. Vertical line shown at log(0.5) CPM.

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-hybridisation[23]. 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 intensity[24]. There are also differences in the statistical models behind expression quantification, as described above.

cite relevant preprocessing sections

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

Regarding the batch effect within the HIRD array data, a popular adjustment method is ComBat[25], which estimates centering and scaling parameters per-gene pooling information across all genes using empirical Bayes. ComBat is the method used in[5]. In comparisons of microarray batch effect adjustment 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 size[28] or distribution of study groups that have an impact on expression[29],

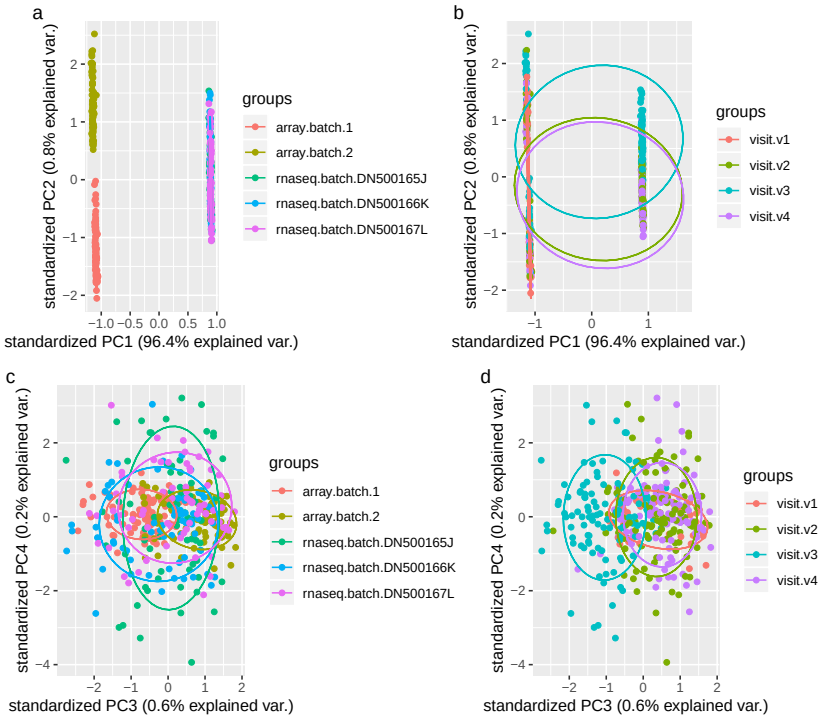

Figure 2.12: First 2 expression PCs, expressions first standardized within platform to increase comparability, if not, even more variation would be driven by platform

ComBat can overcorrect batch differences or bias estimates of group differences respectively. In our data, sample size and timepoint groups are fairly balanced between the two array batches, but the proportion of responders is notTable 2.2, hence I elect not to use ComBat to pre-adjust the array expression data, and model the batches as fixed effects.

2.2.10.1 Per-platform Differential gene expression model

As[5] 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 in the following DGE models.

For RNA-seq samples, between-sample normalisation was performed using the trimmed mean of M-values (TMM) method[30] from edgeR[31]; then variance-stabilisation was performed using voom[32], resulting in expression values with units of $\log_2$ CPM.

Linear models were fit using limma[33], which is computationally fast, and performs well for sufficiently large ($n \geq 3$ per group) samples[34] 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.

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

2.2.10.2 Choice of differential gene expression meta-analysis metho

In the section , I concluded that a two-stage meta-analysis approach would be appropriate. This meta-analysis is restricted to 13945 genes assayed by both the array and RNA-seq platforms.

add section labels

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, representing an additional source of variation termed the between-studies heterogeneity. In the HIRD data, I have $k = 2$ 'studies' (array and RNA-seq), where the platform differences described in section contribute to between-

add label

studies heterogeneity, hence a random-effects model is appropriate.

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

2.2.10.3 Prior for between-studies heterogeneity

The choice of prior for between-studies heterogeneity is influential when k is small[42].[44] considers the case of $k = 3$, showing that a flat prior places too much weight on large estimates of τ, and recommends a weakly informative prior that acts to constrain the posterior distribution. Since I assumed non-positive values for tau^2 are unrealistic, I use a weakly-informative gamma prior recommended by[40], which has zero density at $\tau = 0$, increasing gently as τ increases. This constrains τ to be positive, but still permits estimates close to zero if the data support it. This is in constrast to 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 tau no matter what the data suggest; and in contrast to half-t family priors (e.g.[42,44]), 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 effects[39]) was first 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[41,44] `bayesmeta` implements both flat and

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

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.

As in the section above, to get the 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 effects[48]. 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 qnorm(0.975, mean=0, sd=1*10) = 1*19.59964

2.2.10.5 FDR

ashr

2.2.11 Gene set enrichment analysis

tmod; gprofileR; CAMERA

2.3 Results

We conducted differential gene expression (DGE) analysis to detect genes significantly up or downregulated (log2 fold-change > 0.5 at FDR < 0.05) between the three timepoints: pre-vaccination on day 0, and days 1 and 7 post-vaccination; and between vaccine responders and non-responders. Few differentially expressed genes were detected between different covariate groups included in the model (library preparation pool, sex, and ethnic background), suggesting that vaccination was the dominant influence on transcriptomic response. Overall response clusters into two distinct patterns (Fig. 2.16).

2.3.0.1 Evaluation of priors

[40] recommends (shape=2, scale=λ), with small λ.

From Fig. 2.14 and Fig. 2.15, we are similar.

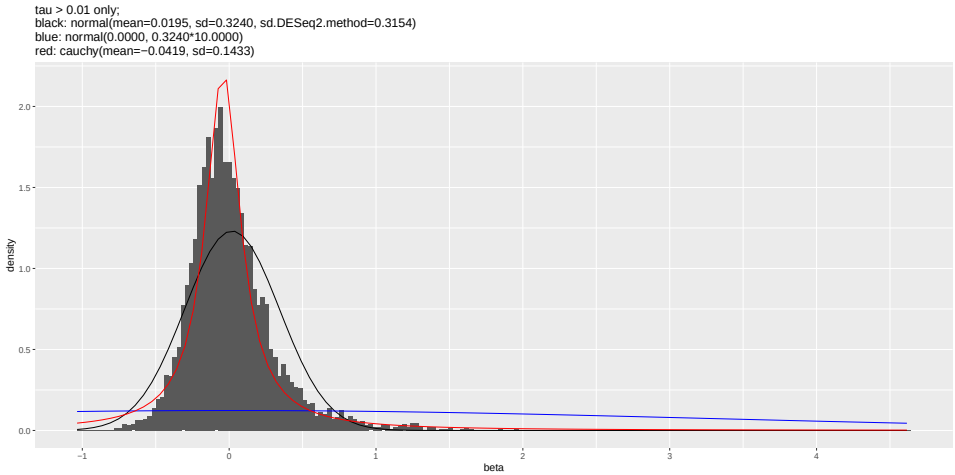

Figure 2.13: Fold-change comparison between array and RNAseq for day 1 vs day 0.

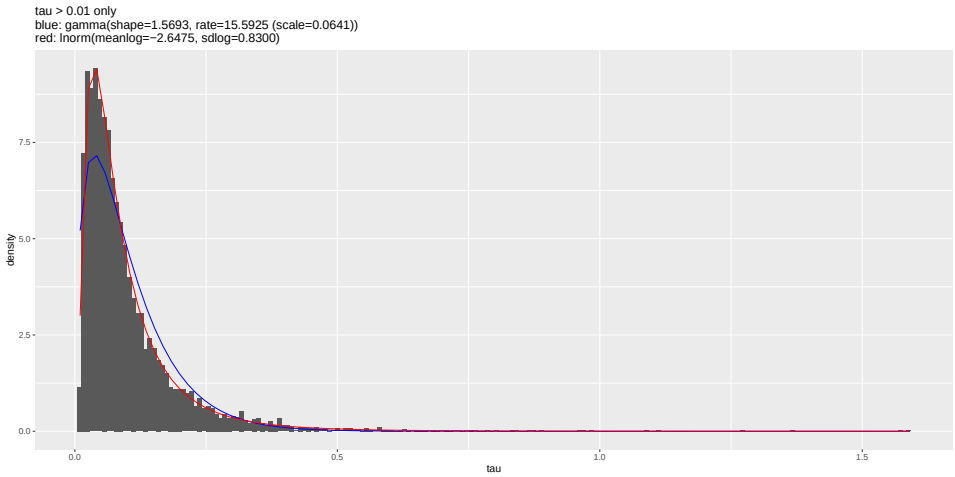

Figure 2.14: Priors for tau for day 1 vs day 0 DGE meta-analysis.

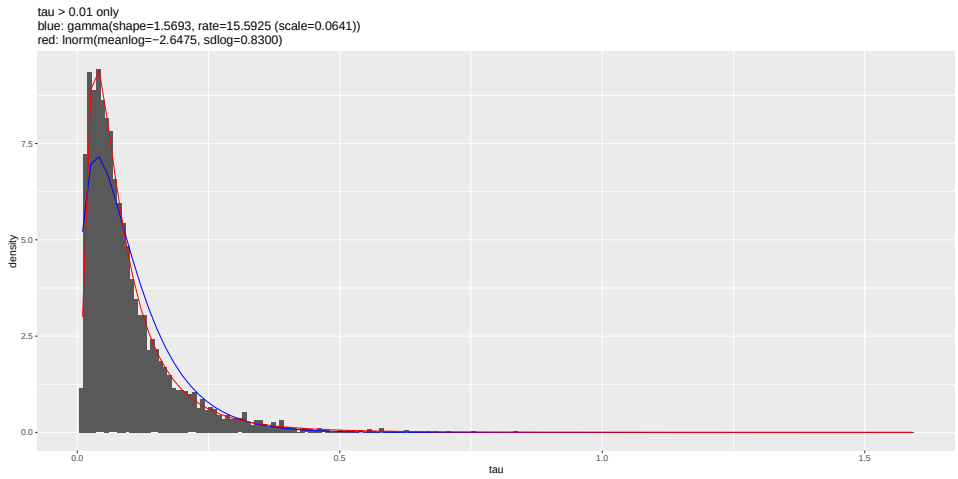

Figure 2.15: Priors for mu for day 1 vs day 0 DGE meta-analysis.

2.3.0.2 Transient upregulation of interferon, interferon-inducible, and other innate immunity genes observed at day 1 post-vaccination

The highest number of upregulated genes (799) was observed at day 1 compared to day 0, and the majority of these genes (757) were then downregulated from day 1 to day 7 (Fig. **??**). Genes with high fold change (FC) increases at day 1 included ANKRD22 (log2 FC = 4.545, viral defence[**Bin2016**]), the interferon gamma-induced cytokine CXCL10 (log2 FC = 3.933) and C1-inhibitor SERPING1 (log2 FC = 3.766, complement system[4]). Upregulation of interferons and interferon-inducible genes were particularly prominent at day 1 (Fig. **??**). We also detected upregulation of non-coding transcripts such as macrophage-associated[**Zhang2017b**] lincRNA RP11-10J5.1 (log2 FC = 3.747).

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.

2.3.0.3 Upregulation of plasmablast and cell proliferation genes at day 7 post-vaccination

A total of 421 genes were upregulated at day 7 compared to day 0, day 1, or both (Fig. **??**). The transcriptomic profile compared to day 0 was dominated by increases in immunoglobulin genes: 68 of the top 100 most

[4] https://www.uniprot.org/uniprot/P05155

significantly DE genes from day 0 to day 7 were immunoglobulins (IGH, IGK or IGL), including IGHV1-69 (log2 FC = 2.365), a known modulator of anti-influenza antibodies[Avnir2016]. Other highly up-regulated genes included the B cell-specific genes MZB1 (log2 FC = 2.014) and TNFRSF17 (log2 FC = 1.837).

The enriched gene sets at day 7 were related to cellular proliferation; the top three MSigDB sets were targets of E2F, targets of MYC, and genes associated with the G2/M check point (Table **??**). Overrepresented Gene Ontology categories amongst upregulated genes included "mitotic cell cycle" (p = 8.46e-21), "cell division" (p = 1.59e-16), and "chromosome segregation" (p = 6.80e-15).

Day 7 response is characterised by adaptive B cell response: plasma cell genes, immunoglobulins, proliferation (Table 2.3).

2.3.0.4 Comparison to Sobolev et al.

2.3.1 Expression associated with antibody response

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

2.3.1.1 Comparison to Sobolev et al.

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

2.3.2 TODO Identifying molecular signatures for predicting antibody response

- For inference, don't dichotimise due to statistical concerns
  - "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?
- Interpretatability: DAMIP gives rulesets composed of small sets of genes, amenable to rapid qPCR assays.

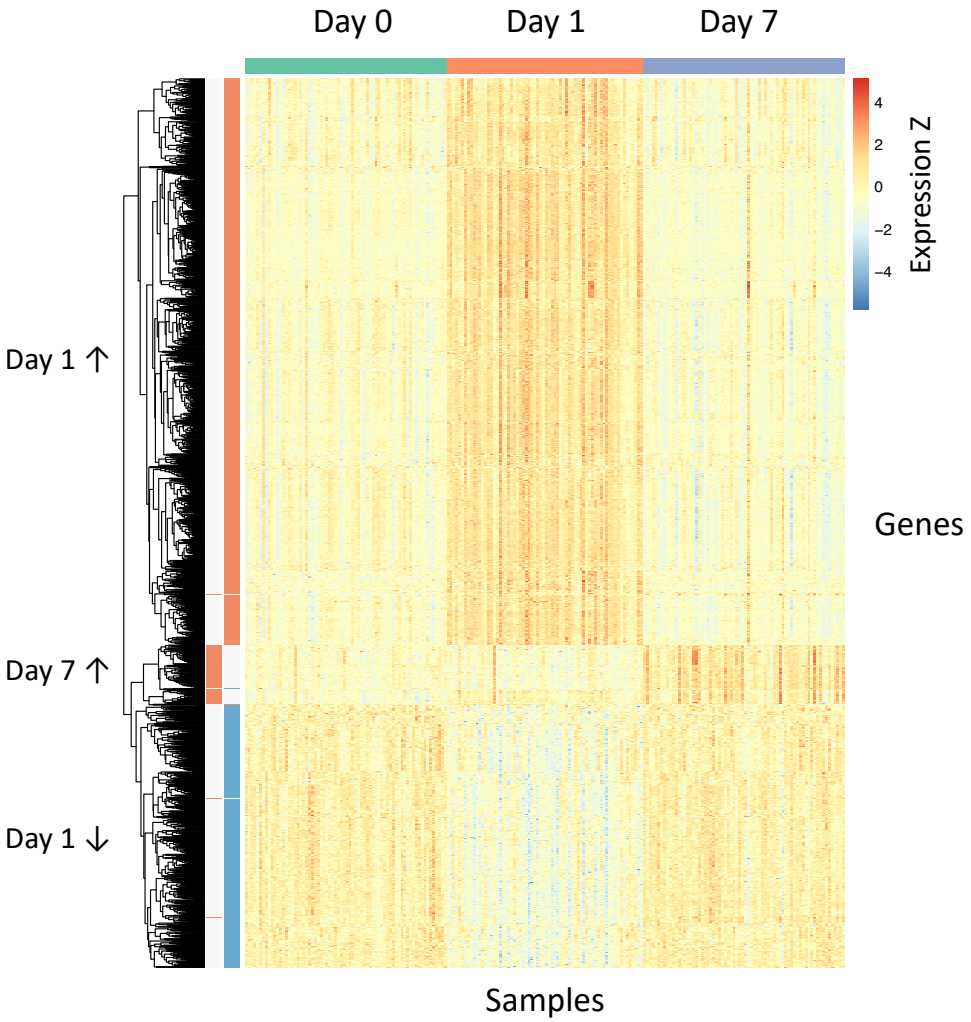

Figure 2.16: Normalised gene expression for differentially expressed genes (adj. $p < 0.05$, $|\log_2 FC| > 1.5$) across 208 RNA-seq samples from days 0, 1, and 7, clustered by gene.

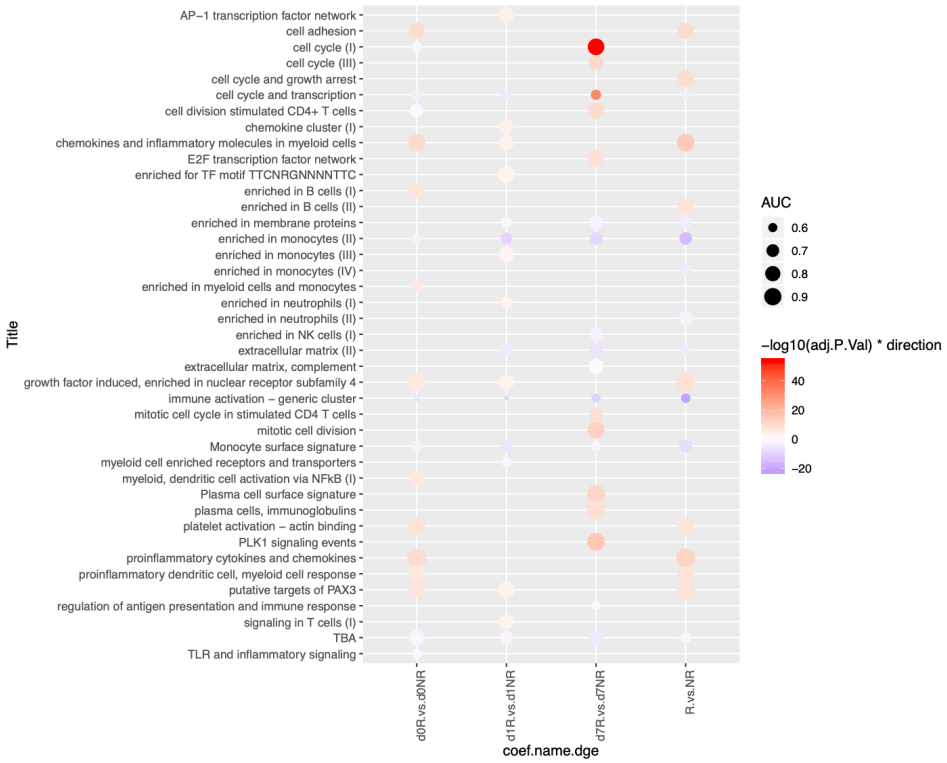

Figure 2.17: Transcriptomic modules enriched in genes with expression positively and negatively associated with TRI. Blank cells n.s.

2.4 Discussion

We have measured the transcriptomic response to influenza-H1N1 vaccination in the HIRD cohort, and characterised genes and gene sets differentially expressed between timepoints, and between vaccine responders/non-responders. Transient upregulation of many genes associated with interferons (a class of cytokines important for response to viral infection) and other aspects of innate immunity were detected at day 1, with most genes returning to baseline expression by day 7. Upregulation of immunoglobulin, B cell-related, and cell proliferation genes were detected at day 7—likely related to expansion of plasmablasts. Previous studies on seasonal influenza vaccines also reported induction of interferon-related genes 1-3 days post-vaccination, and enrichment of plasmablast and cell-cycle genes at day 7[Nakaya2011, Franco2013, Nakaya2015a].

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

the day 63 timepoint is too late? late compared to many other studies
Use hai and MN in split models No change scores
other measures of seroconversion https://www.who.int/biologicals/vaccines/Annex_2_WHO_TRS_963-3.pdf
and seroconversion may not correspond well to protection...

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."

Table 2.1: Sample descriptive statistics.

		platform	
	Total	array	rnaseq
	n = 114	n = 44	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)

Table 2.3: 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.

	adj. p value		
Module	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}	

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=sigma2i/tau2i 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 locitools 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. (fairfax2014InnateImmuneActivity 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 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/predicct

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

Appendix A

Supplementary Materials

A.1 Chapter 2

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

A.2 Chapter 3

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

luctus mauris.

A.3 Chapter 4

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

Bibliography

1. Verma, A. & Ritchie, M. D. Current Scope and Challenges in Phenome-Wide Association Studies. *Current Epidemiology Reports* **4,** 321–329. doi:10.1007/s40471-017-0127-7 (Dec. 2017).

2. Barreiro, L. B., Tailleux, L., Pai, A. A., Gicquel, B., Marioni, J. C. & Gilad, Y. Deciphering the Genetic Architecture of Variation in the Immune Response to Mycobacterium Tuberculosis Infection. *Proceedings of the National Academy of Sciences* **109,** 1204–1209. doi:10.1073/pnas.1115761109 (Jan. 24, 2012).

3. Fairfax, B. P. *et al.* Innate Immune Activity Conditions the Effect of Regulatory Variants upon Monocyte Gene Expression. *Science* **343,** 1246949–1246949. doi:10.1126/science.1246949 (Mar. 7, 2014).

4. Fairfax, B. P. & Knight, J. C. Genetics of Gene Expression in Immunity to Infection. *Current Opinion in Immunology* **30,** 63–71. doi:10.1016/j.coi.2014.07.001 (Oct. 2014).

5. Sobolev, O. *et al.* Adjuvanted Influenza-H1N1 Vaccination Reveals Lymphoid Signatures of Age-Dependent Early Responses and of Clinical Adverse Events. *Nature Immunology* **17,** 204–213. doi:10.1038/ni.3328 (Jan. 4, 2016).

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 (June 1983).

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

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 (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 (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 (July 17, 2014).

12. Brown, B. C., Bray, N. L. & Pachter, L. Expression Reflects Population Structure. doi: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 (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 (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 (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 (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 (Feb. 29, 2016).

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 (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 (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 (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 (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 (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 (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 (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 (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 (Feb. 28, 2011).

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 (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 (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 (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 (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 (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 (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 (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 (2003).

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 (Apr. 2010).

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

38. Gonnermann, A., Framke, T., Großhennig, A. & Koch, A. No Solution yet for Combining Two Independent Studies in the Presence of Heterogeneity. *Statistics in Medicine* **34,** 2476–2480. doi:10.1002/sim.6473 (July 20, 2015).

39. Veroniki, A. A. *et al.* Methods to Estimate the Between-Study Variance and Its Uncertainty in Meta-Analysis. *Research Synthesis Methods* **7,** 55–79. doi:10.1002/jrsm.1164 (Mar. 2016).

40. 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 (Oct. 2013).

41. Friede, T., Röver, C., Wandel, S. & Neuenschwander, B. Meta-Analysis of Few Small Studies in Orphan Diseases. *Research Synthesis Methods* **8,** 79–91. doi:10.1002/jrsm.1217 (Mar. 2017).

42. Seide, S. E., Röver, C. & Friede, T. Likelihood-Based Random-Effects Meta-Analysis with Few Studies: Empirical and Simulation Studies. *BMC Medical Research Methodology* **19.** doi:10.1186/s12874-018-0618-3 (Dec. 2019).

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

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 (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 (Nov. 20, 2011).

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 (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 (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 (June 1, 2019).

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

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

spell-check

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

Todo list