<title>

Benjamin Yu Hang Bai

January 28, 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

Why study human genetics?

Human variation [...] Nature vs nurture [...] Causal anchors. [...] Leveraging natural G variation.

1.1 A brief history of complex trait genetics

Early days, Prior to GWAS

Twin studies and heritability estimates. of complex traits

Mendelian genetics, family and linkage studies

Candidate gene studies (Border et al., 2019)

1.1.1 The advent of GWAS

Common disease, common variant

X years of GWAS

Missing heritability

comparison: array - WES - WGS

WES (about 40Mbp) go rarer variation vs WGS due to better coverage in just the exons

but lower n, so lower power than array genotyping to do single variant associations

WGS

structural variants

rare variants uncovered by arrays generally higher effect size often exonic

burden tests (e.g. SAIGE) aggregate based on variant consequence scores e.g. vep scores to get gene

future outlook

expanding into global populations, global biobanks Gains from Africa H3Africa and Asia

polygenic scores

use in the clinic e.g. polygenic background can modify penetrance

pathway prs challenge is variant to gene assignment/mapping e.g. restrictions to fine mapped eQTLs

Pathway analysis the great hairball gambit

1.1.2 Post-GWAS: 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

1.1.3 Post-GWAS: 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].

Why care? Understand mech. of causal genes

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

coloc methods (that photo on all the coloc methods that all attempt to solve the problem)

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

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

1.2 Context is key

for both gwas, and molQTLs, context is key

contexts tissue cell type stimulation conditions

QTLs can interact with sex and age

interaction between cells in vivo

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

types of conditinoal QTL ackerman conditional vs dynamic

Review of stimulation condition QTL mapping, invitro and invivo what models used? did they use change scores for longitudinal?

Mechanisms:

1.2.1 Context-specific immune response QTLs in vitro

A type of context is cell type Confounds actual context of stimulation

Review of cell type specific methods here

1.2.2 *in vivo* response QTL mapping

less popular

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

Review of in vivo mapping. Franco Lareau smallpox apoptosis Caliskan Rhinovirus Davenport

1.3 Immunity is a complex trait

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.

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.

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.].

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

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.

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 Genetics factors affecting vaccine response

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

1.6 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 outbreak
  - 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 Systems vaccinology of influenza vaccines

- Review influenza vaccine specific sysvacc papers (e.g. Nakaya's papers)
  - inclu. prevaccination signatures paper

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 ... TODO
      · Compare the d7 split to Sobolev TODO

2.2 Methods

2.2.1 Pre-existing HIRD study data and additional sampling

Sample demographics: age, sex, self-reported ethnicity

2.2.2 Genotype data generation

DNA extraction; genotyping array

2.2.3 Genotype data preprocessing

Sample and marker QC; phasing and imputation; post-imputation filters; PC projection; PC imputation

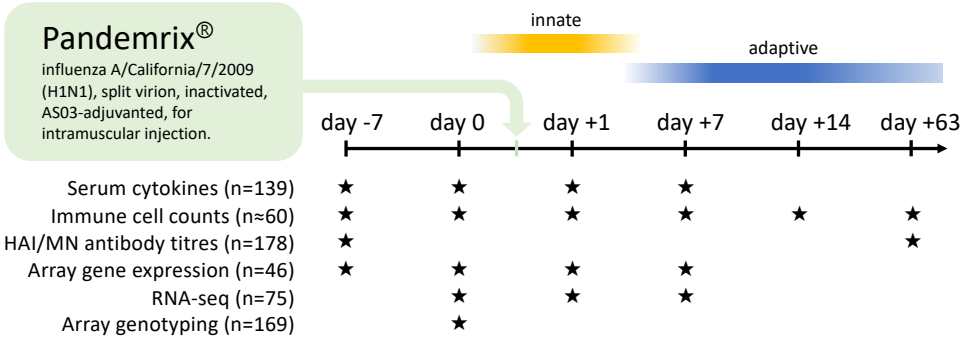

Figure 2.1: Data types, timepoints, and sample sizes. Individuals were vaccinated after day 0 sampling. Vaccine-induced antibodies measured by HAI and MN assays. Array and RNA-seq gene expression measured in the PBMC compartment.

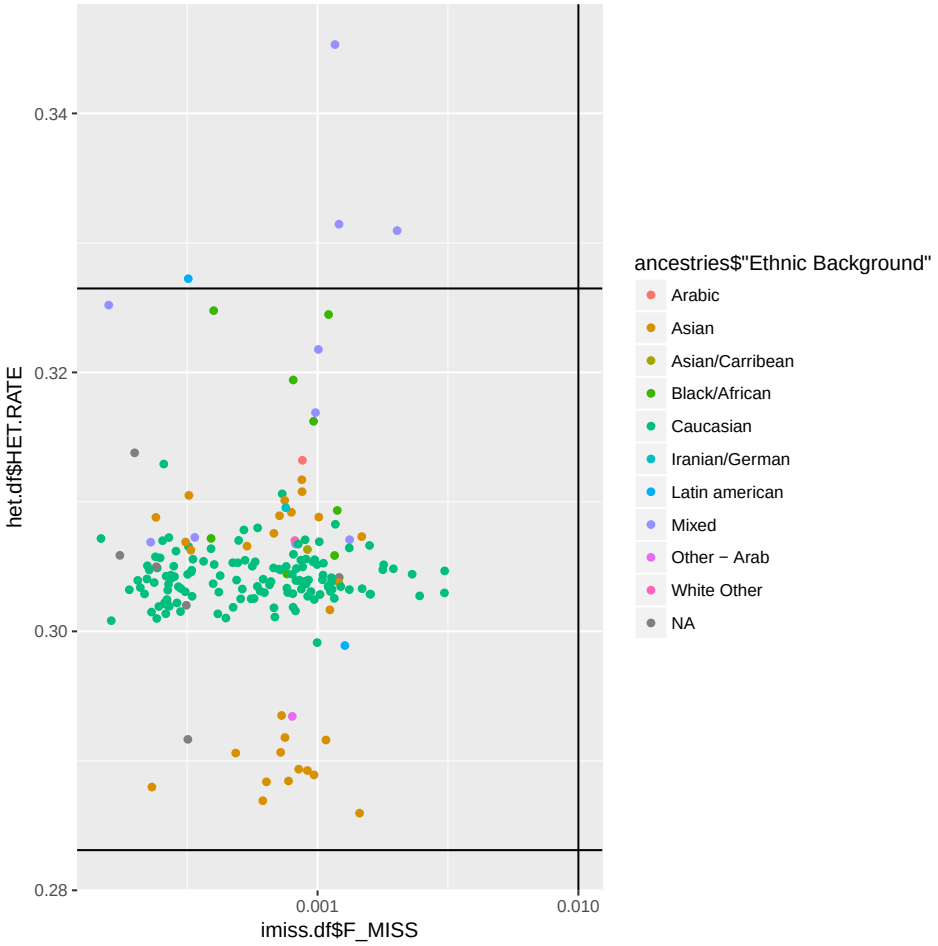

Figure 2.2: Sample filters for missingness vs heterozygosity rate.

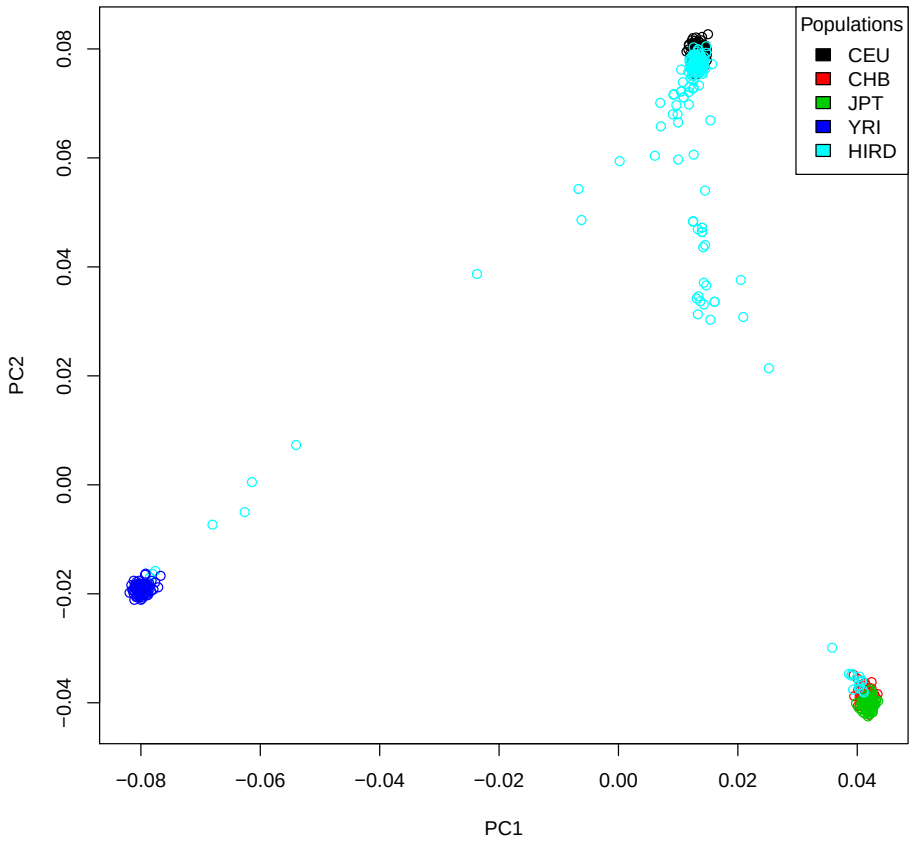

Figure 2.3: Samples projected onto HapMap PC axes.

2.2.4 RNA-seq data generation

Do we have enough reads for RNAseq analysis? https://www.ncbi.nlm.nih.gov/pubmed/24434847 and doi:10.1093/bioinformatics/btt688

Summing tech reps: sum of poison is poisson, average is not: https://www.biostars.org/p/30455

2.2.5 RNA-seq data preprocessing

QC

- fastqc (sequence quality, GC content, length, duplication, overrepresented sequences incl adapters)
- qualimap
- salmon qc

Quantification

Filtering

2.2.6 Array data preprocessing

Batch effect correction (see batch effects Zotero tag) Combat is best here. LM, LMM, Combat were comparable. 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 conficence intervals (1e).

Figure 2.4: The mean quality value across each base position in the read.

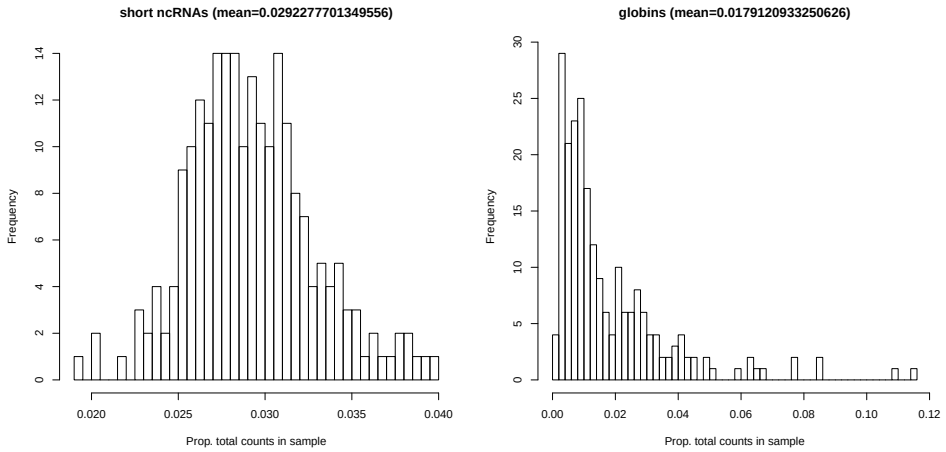

Figure 2.5: ncRNA and globin levels.

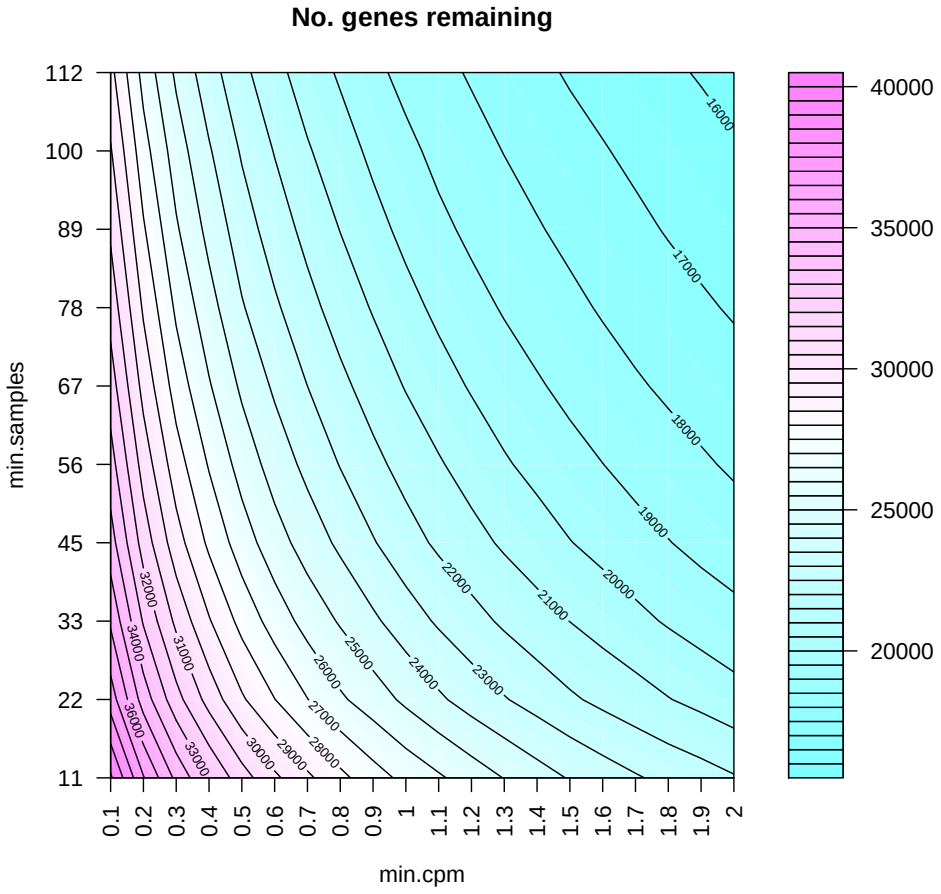

Figure 2.6: Number of features retained at different thresholds.

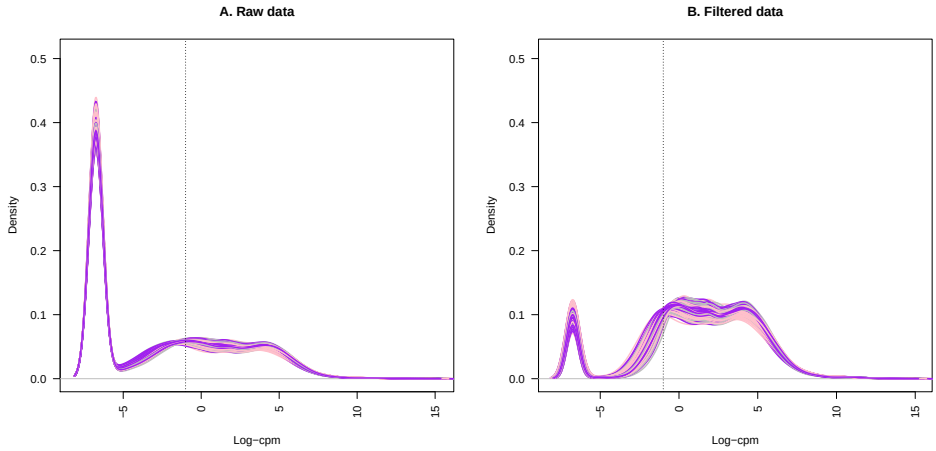

Figure 2.7: Choice of cpm filtering threshold.

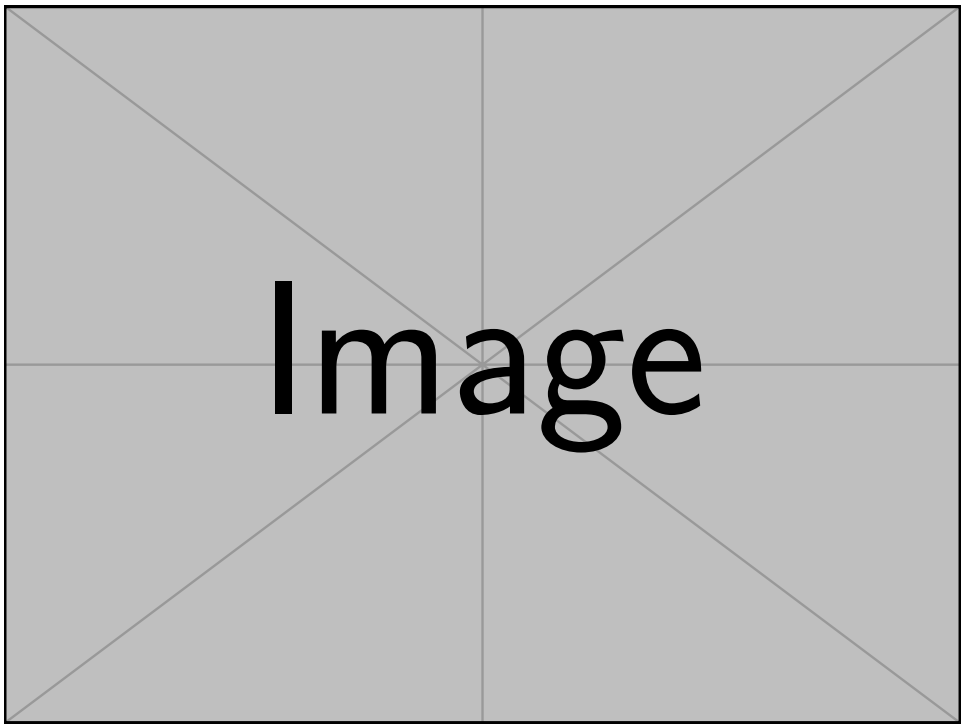

Figure 2.8: TABLE: Balance of timepoints and R/NR in the two array batches.

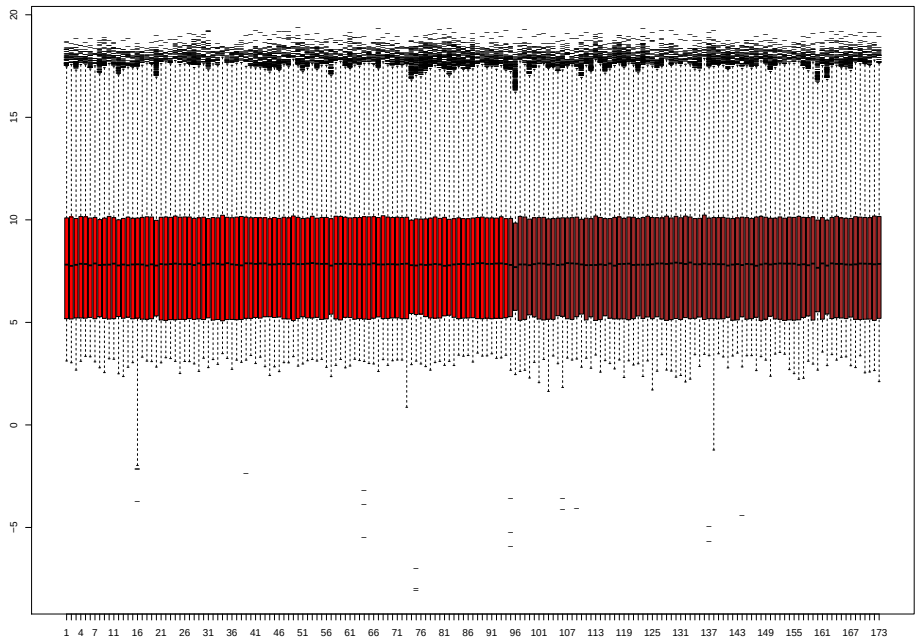

Figure 2.9: Array intensity estimates after normalisation and batch effect correction.

2.2.7 Computing baseline-adjusted measure of antibody response: TRI

- Pre-process phenotypes
  - Compute responder status (>= 4-fold in HAI or MN)
  - Compute TRI (based on Bucasas 2009)
    - "We related the change in titer between pre- and postvaccination measurements (response variable) to the prevaccination titer (explanatory variable) using a simple linear model"
    - "We next determined the residuals from the above linear regressions and used them as the input values for the individual response scores."
    - "we standardized the residuals by dividing by the residual standard deviation for each component"
      - Based on their axis ranges, it appears they are plotting log2(post)-log2(pre)), equivalently log2(post/pre), a.k.a. log2 fold-change; against log2(pre)

- Note that log2(post-pre) does not make sense mathematically, as post-pre may well be negative
- The negative relationship indicates lower initial titres are more amenable to high fold-change increases, which is exactly what TRI is designed to correct for

- 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).

2.2.8 Differential gene expression (DGE)

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.

2.2.9 DGE meta-analysis

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."

2.2.9.1 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:

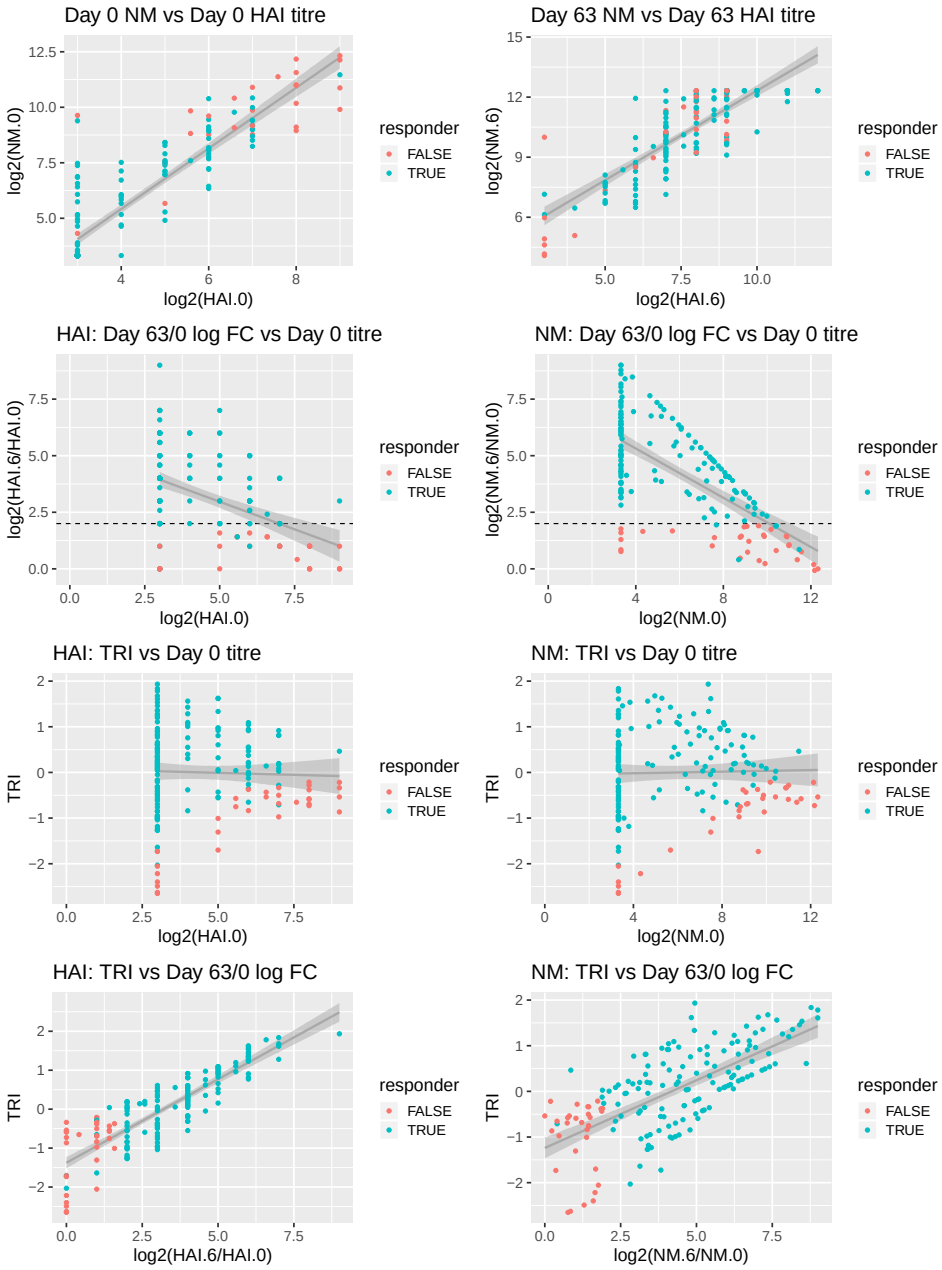

Figure 2.10: How TRI corrects fold changes for baseline titre.

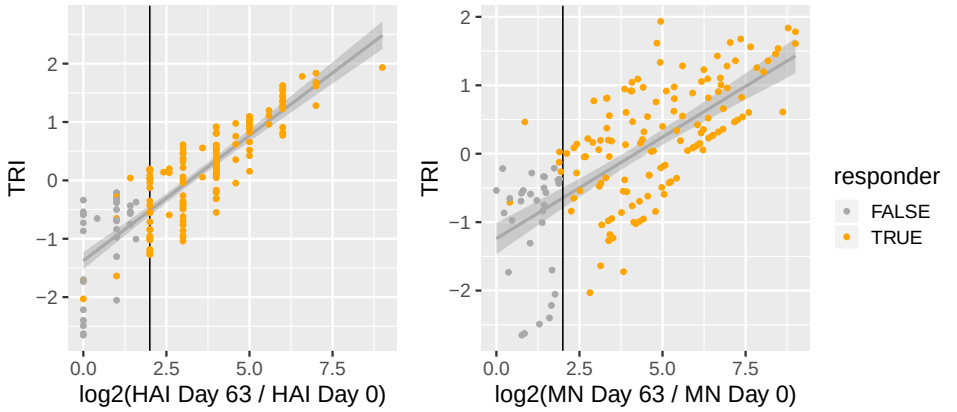

Figure 2.11: TRI correlates with the standard responder definition (colored, 4-fold increase in either assay). An individual's TRI is the mean of their Z-transformed residuals from regressions of day 63 vs. day 0 fold-change against day 0 titre, over the two assays.

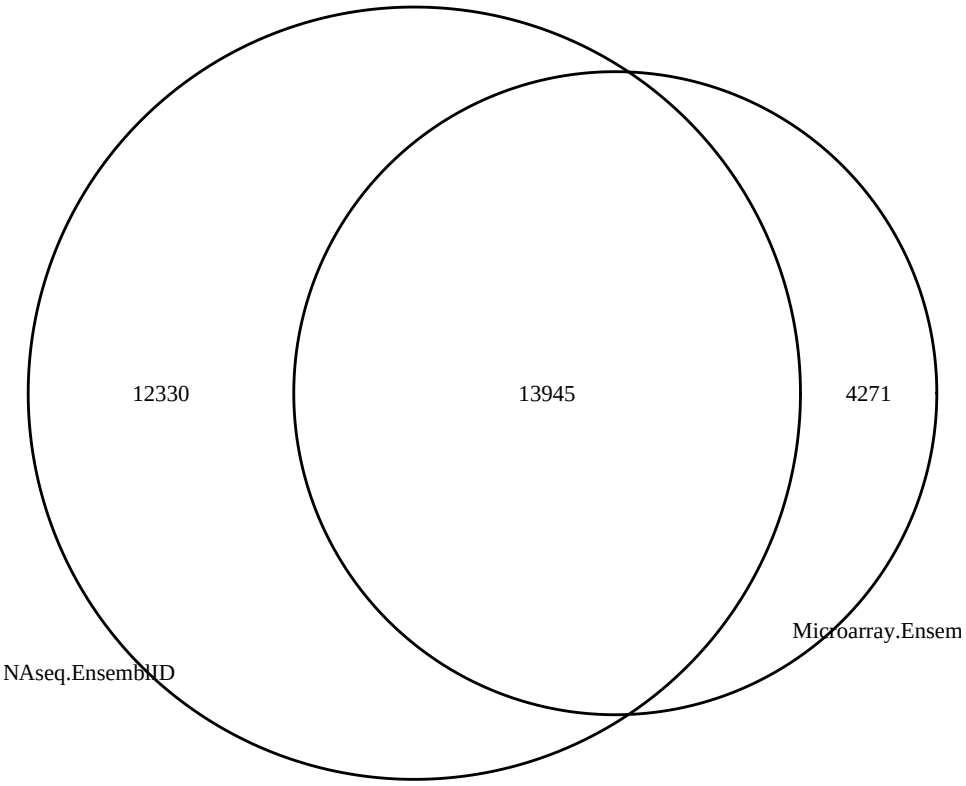

Figure 2.12: Feature overlap between array and RNAseq data post-filtering.

  - 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 combinding platform
    * cannot acually 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?

2.2.9.2 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.

2.2.9.3 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 ashr, 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 log2FC of 20.

2.2.9.4 Meta-analysis using bayesmeta

2.2.10 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. 2.15).

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 TODO 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 (Table 2.2).
- TODO: split by day and look for signatures per day

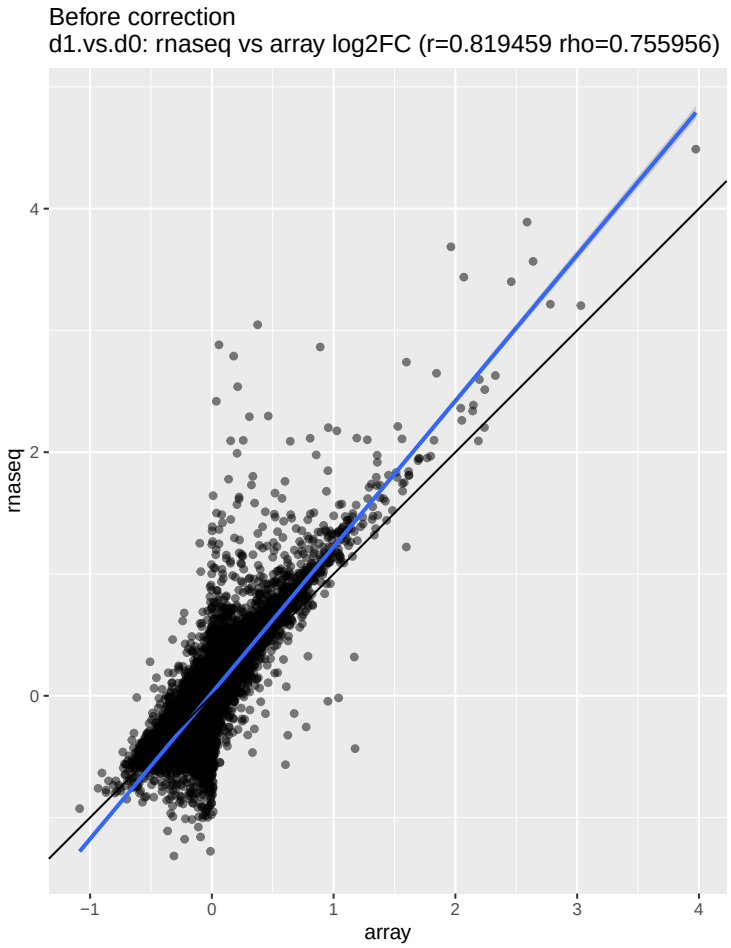

Figure 2.13: Fold-change comparison between array and RNAseq for day 1 vs day 0.

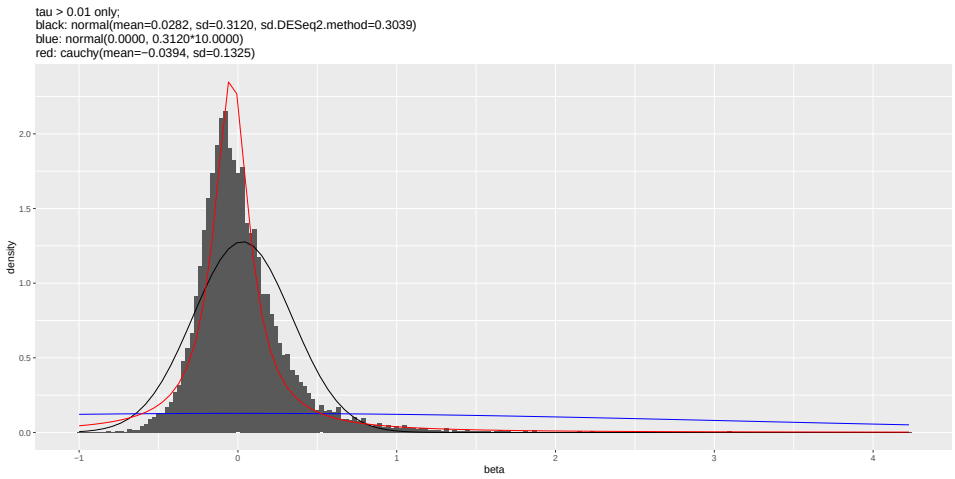

Figure 2.14: Priors for day 1 vs day 0 DGE meta-analysis.

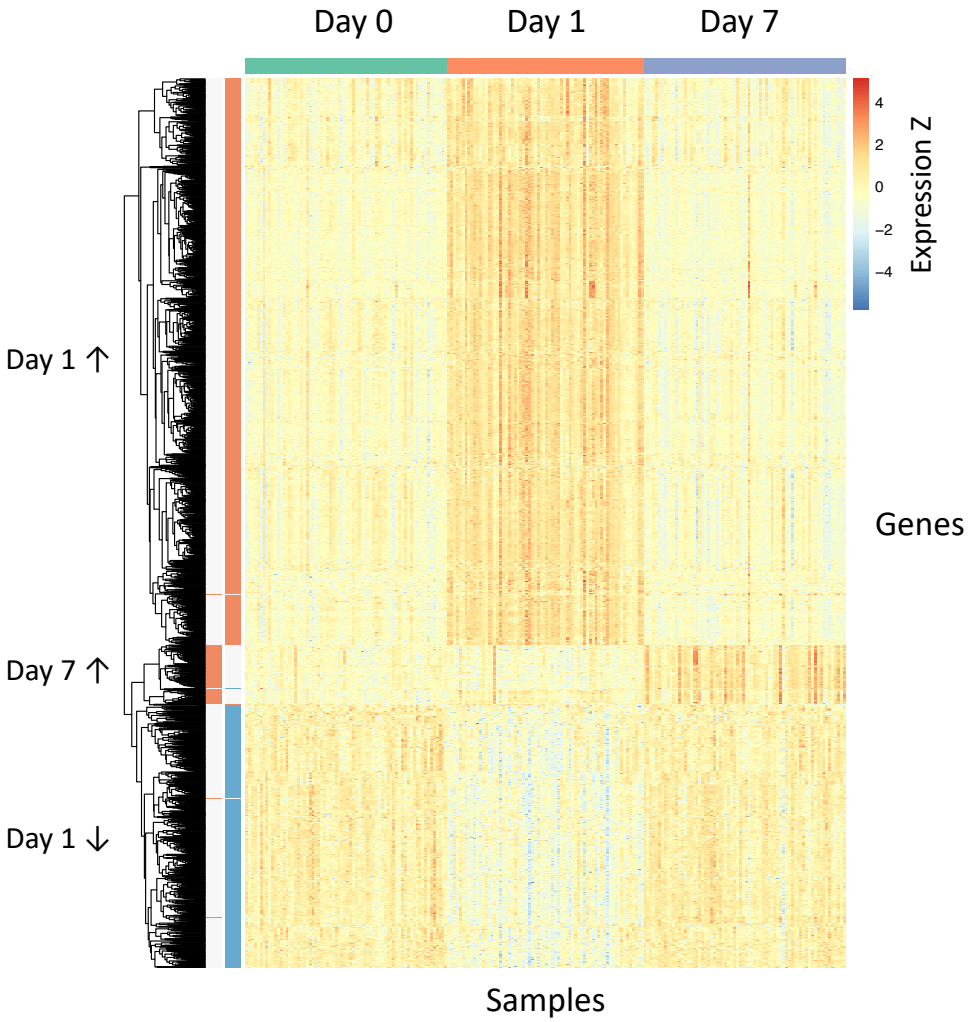

Figure 2.15: Normalised gene expression for differentially expressed genes (adj. $p < 0.05$, $|\log_2 FC| > 1.5$) across 208 RNA-seq samples from days 0, 1, and 7, clustered by gene.

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.

	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}	

Table 2.2: Transcriptomic modules enriched in genes with expression positively and negatively associated with TRI. Blank cells n.s.

	adj. p value	
Module	High TRI	Low TRI
plasma cells & B cells, immunoglobulins	8.0×10^{-4}	
innate activation by cytosolic DNA sensing		3.7×10^{-3}
proinflammatory cytokines and chemokines		3.7×10^{-3}
AP-1 transcription factor network		1.4×10^{-2}
enriched in neutrophils		2.3×10^{-2}

2.3.2.1 TODO Comparison to Sobolev et al.

2.3.3 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.

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).

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 Context-specific immune response QTLs for flu

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

3.2 Methods

3.2.1 expression norm

2018-03-15 in log

3.2.2 Genotyping data generation

3.2.3 Genotyping quality control

3.2.4 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.

3.2.5 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:

3.2.5.1 Estimation of cell type abundances

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

3.2.5.2 Kinship matrix computation

LDAK kinship matrix construction http://dougspeed.com/method-overview/

Note: can be negative https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6157025/ LDAK version 4.9 [3] and IBDLD version 3.33 [4] were used to derive 2 empirical kinship matrices based on the GAW20 genotype data. For LDAK, in principle, this kernel should correspond to a genetic relationship matrix; in practice, however, we observed that LDAK estimates of self-relatedness were widely spread around their expectation of 1 (Fig. 1a). For IBDLD the estimates of self-relatedness were closer to 1 (Fig. (Fig.1b).1b). The empirical kinship estimate matrices from LDAK and IBDLD were postprocessed to remove negative nonzero values and scaled to have a diagonal equal to 1.

3.2.5.3 Expression normalisation

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

3.2.5.4 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

3.2.5.5 Correction for cell type abundances

3.2.6 eQTL mapping with mixed models

3.2.7 eQTL meta-analysis

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

3.2.7.1 Joint mapping

3.2.7.2 mashr smoothing

review: condition/Cell-type specific methods refere to 2019-11-19 Cell-count specific eQTL mapping papers

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

normally eqtls use perms for FDR

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

3.2.7.3 Defining shared and response eQTLs

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

3.2.8 Colocalization

Due to the increasingly abundant

For example, ran

Coloc and assumptions

Hypercoloc and assumptions

large numbers of traits

Confounding by multiple causal
Fine mapping

3.3 Results

3.3.1 eQTLs at each timepoint

3.3.2 Estimation of eQTL sharing

3.3.3 replication of shared eQTLs in whole blood

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

3.3.5 (pathway) Polygenic score to predict antibody response

3.4 Discussion

Current limitations Confounded by changes in immune cell proportions in bulk PBMCs

No conditional eQTL analysis to disentangle conditional effects Unclear connection to vaccine biology e.g. what genesets/pathways/cell types are driving the observed transcriptomic and eQTL response? Future work to address limitations 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

Chapter 4

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

4.1 Introduction

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

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

5.2 Methods

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

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

Cost-effectiveness and clinical implementation

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).

List of Abbreviations

HAI haemagglutination inhibition

MN microneutralisation

PBMC peripheral blood mononuclear cell

RNA-seq RNA-sequencing

Todo list