<title>

Benjamin Yu Hang Bai

2020-06-07 01:02:24+01:00

A little learning is a dangerous thing;
Drink deep, or taste not the Pierian spring:
There shallow draughts intoxicate the brain,
And drinking largely sobers us again.

Alexander Pope, *An Essay on Criticism*

Abstract

<thesis abstract>

Acknowledgements

<acknowledgements>

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Structure and diversity of the human genome

- The human genome is almost three billion base pairs (bps) in length, containing 20000-25000 protein-coding genes [1, 2] that span 1-3% of its length, with the remainder being non-coding. Each diploid human cell contains two copies of the genome; 46 chromosomes comprised of 23 maternal-parental pairs: 22 pairs of homologous autosomes and one pair of sex chromosomes.

- Variation in the genome between individuals in a population exists in the form of single nucleotide polymorphisms (SNPs), short indels, and structural variants—the vast majority of common variants (MAF $> 1-5\%$) are SNPs and short indels ($> 99.9\%$) [2]. On average, a pair of human genomes differs by one SNP per 1000-2000 bp [3]. Each version of a variant is called an allele; an individual has a maternal and parental allele at each variant.

consider moving awkward defs to margin notes, in the style of nature reviews

- The many variants in a population are inherited in a smaller number of haplotypes: contiguous stretches of the genome passed through generations via meiotic segregation. The fundamental sources of genetic diversity are mutation and meiotic recombination, generating new alleles and breaking apart haplotypes into shorter ones over time. Variants at locations on a chromosome (loci) that are physically close are less likely to flank a recombination event, hence more likely to cosegregate on the same haplotype, referred to as genetic linkage. Genetic linkage is

LD decay just takes a really really long time, but there are evo forces at work too that maintain LD one source of linkage disequilibrium (LD): the non-random association of alleles at two loci, differing from expectation based on their frequencies and the law of independent assortment [4]. LD is often quantified within a population by r^2, the squared correlation coefficient between alleles [4].

Recombination events are not distributed uniformly throughout the genome. The genome is a mosaic of blocks delimited by recombination hotspots, characterised by strong LD within blocks, and little LD between blocks [5, 6] (Fig. 1.1). The structure of correlated haplotypes reflects a population's unique evolutionary history, and can be used to trace the demography of human populations back through time [7].

Heard it's good for the reader's attention span to have figures in intro. Unless it's ok to use figures from papers, I only want to spend the time making the min that are necessary though.

1.2 Genetic association studies for complex traits

1.2.1 Principles of genetic association

- Variation in human traits arises from an interplay between genetics and environment. Traits for which genetic variation explains a non-zero fraction of phenotypic variation are heritable. Many measurable human traits are heritable and twin studies provide upper bounds on this heritability `https://www.nature.com/articles/ng.3285`. Discovering the specific genetic variants that contribute to heritability, through association of variants and phenotypes measured from the same individual, is a mainstay of the field of human genetics. Barring somatic mutation, an individual's genome is fixed at conception, providing a causally upstream anchor. Genetic association studies have intrinsic resistance to many back-door path effects that permeate observational studies of the causes of human phenotypes.

add something sweeping about utility here or elsewhere: e.g. insights into trait biology and clinical translational potential for disease traits, genetically support drug target identification

- Under the central dogma, information flows from gene to RNA to protein to phenotype via transcription and translation, thus it is assumed that genetic variants at loci in the genome affect phenotype by impacting on the function or regulation of target genes. How genetic variation contributes to any heritable trait defines it's genetic architecture: the number of genes affecting that trait; along with the allele frequencies, effect sizes, and interactions of trait-associated variants [8]. The mainmatter/figures/chapter_01/fig_mockup_haplotypeBlocks_Screenshot 2020-05-21 at 17.08

Figure 1.1: The genomic mosaic: block-like LD structure of the genome

number of genes defines a spectrum of traits from monogenic (where inheritance follow simple Mendelian patterns) to polygenic (where inheritance is complex). Many architectures have been proposed for complex traits; all have in common that the number of genes that affect a complex trait is large (ranging from dozens to many thousands), thus the average effect of each trait-associated loci is small [9, 10] https://www.pnas.org/content/106/23/9362.

1.2.2 Lessons from 15 years of genome-wide association

- For decades, linkage analysis had been successfully applied to map loci affecting Mendelian traits by tracing their cosegregation with the trait through pedigrees [11]. Small-scale genetic association studies were also performed, focusing on variants in or near candidate genes selected on the basis of prior biological knowledge [12]. These approaches were not successful for complex traits, as small effect sizes lead to low penetrance in pedigrees [11] and poor power at the sample sizes typically used in early candidate gene studies [13].
- Genome-wide association studies (GWAS) systematically test common

variants selected in a hypothesis-free manner across the genome for association with a trait (Fig. 1.2). Using large sample sizes to overcome small effects and large multiple testing burden, thousands of associations have been discovered for complex traits and disease, many robustly replicated across populations [11, 14]. Most genetic variance is explained by additive effects, the contribution of epistatic interactions is small [8], and pleiotropy is widespread [11]. Sample sizes in the millions are increasingly commonplace, and discovery of new associations with increasing sample size shows no sign of plateauing [15]. It is now appreciated that most heritable phenotypes are complex, and have remarkable polygenicity.

1.2.3 From complex trait to locus

GWAS rely on the tendency of common variants on the same haplotype to be in strong LD. As the number of haplotypes is comparatively few, it is possible to select a subset of tag variants such that all other known common variants are within a certain LD threshold of that subset. In practice, there is enough redundancy that the number of variants measured on a modern genotyping array (in the order of 10^5 to 10^6) is sufficient to tag almost all common variants [16, 17]. Associations with unmeasured variants are indirectly detected through their strong correlation with a tag variant. Furthermore, as unrelated individuals still share short ancestral haplotypes, study samples can be assigned haplotypes from a panel of haplotypes derived from reference samples by matching on the directly genotyped variants. This process of genotype imputation allows ascertainment of many more variants not directly genotyped [18], but helps to recover rarer variants that are poorly-tagged [14]. Modern imputation panels enable cost-effective GWAS including tens of millions of variants down to frequencies of $\sim$0.01% `https://www.biorxiv.org/content/10.1101/563866v1`.

seems like there is some connection to be made between the tagability of common variation and the feasibility of imputation both being enabled by the relatively small number of common haplotypes compared to variants

Testing such large numbers of variants incurs a massive multiple testing burden, but acknowledging the correlation between variants due to LD, there are only the equivalent of $\sim 10^6$ independent tests in the European genome, regardless of the number of tests actually performed [19]. The field has thus converged on a fixed discovery threshold of mainmatter/figures/chapter_01/fig_mockup_architecture_Screenshot 2020-05-21 at 17.08.41

Figure 1.2: The reach of GWAS. OR vs MAF ala tam2019BenefitsLimitationsGenomewide, extended by imputation, sample size, WGS based genotypes, but may be indistinguishable from noise at the limits

$0.05/10^6 = 5 \times 10^{-8}$ for genome-wide significance in European populations [20], akin* to controlling the family-wise type I error rate at using the Bonferroni correction.

1.2.4 From locus to causal variant

- By design, a significantly-associated variant from a GWAS needs not be a variant that causally affects the trait, and may only tag a causal variant.
  - Fine-mapping is the process of determining which of the many correlated variants at a GWAS locus are causal.
  - State-of-the-art methods (e.g. PAINTOR, CAVIARBF, FINEMAP `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6050137/`, SuSiE) provide Bayesian posterior probabilities that associated

*The Bonferroni procedure makes no assumptions about the dependence structure of the p-values, and is conservative (i.e. controls the family-wise error rate (FWER) at a stricter level than the chosen α) even for independent tests. In fact it is always conservative unless the p-values have strong negative correlations [21].

variants are causal, and some methods can consider the presence of multiple causal variants at the same locus [22].

  – Even if a single causal variant cannot be assigned, a credible set can

  – Power: to separate causal and tag variants depends on LD and sample size [14]. `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6050137/`

  – Resolution: Naturally, these methods assign probabilities assuming the causal variant is in the set of variants observed.

  – The causal variant must either be genotyped or confidently imputed. Denser genotyping e.g. by WGS, and larger imputation panels will help.

1.2.5 From causal variant to target gene via expression

- Unlike for Mendelian traits where most causal variants are coding `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4573249/`, over 90% of GWAS loci fall in non-coding regions of the genome [23], and often too far from the nearest gene to be in LD `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5291268/`. Thus even if the causal variant at a locus is fine-mapped, it may not be obvious how to find the target genes through which that variant affects the trait.

- Rather than directly impacting the coding sequence of a gene, many non-coding GWAS loci are thought to affect traits by affecting the regulation of target genes [23]. GWAS loci are enriched in regulatory elements annotated by functional genomics studies, such as regions of open chromatin, DNase I hypersensitive sites, splice sites, UTRs, histone binding sites, transcription factor (TF) binding motifs, and enhancers [23, 24] `https://genome.cshlp.org/content/22/9/1748.full`. For complex diseases, enrichment is observed in disease-relevant tissues [14]. As these regulatory elements not only have *cis*, but also *trans* regulatory effects on gene expression, these enrichments put forth expression as an important molecular phenotype linking non-coding GWAS variants to their associated traits (Fig. 1.3).

- Studies of the genetic architecture of quantitative molecular phenotypes have further reinforced this hypothesis.
  - Molecular phenotypes like expression are heritable complex traits [25]
  - Expression can be assayed by e.g. array or RNAseq
  - The variants associated with expression are called expression quantitative trait loci (eQTLs).
  - eQTLs can also be cis- or trans- to their target gene [26].
  - Their effect size declines with distance to the TSS, so most eQTLs detected are *cis*, and within 1Mb [27]
- GWAS variants are enriched for eQTLs `https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000888`
  - So GWAS loci that are also eQTL naturally prioritise target genes.
  - Is it a narrow view to assume that the effect of GWAS loci on complex traits not only act through a target gene, but are specifically mediated by eQTL effects?
  - Over many complex traits, a median of 11% heritability could be explained by mediation of GWAS loci by common (MAF > 0.01) cis-eQTL, and this proportion does not include *trans* or post-transcriptional effects.
- With increasing sample size, most genes (60-80%) have a detectable eQTL [27]. Assuming that a locus on the genome is associated with both a complex trait and an eQTL, how can we separate the scenario where one variant affects both trait and expression (pleiotropy), from coincidental overlap between distinct causal variants that may possibly in LD? Bayesian probabilistic colocalisation methods (e.g. eCAVIAR, Sherlock, coloc [28]) address this by estimating the posterior probability that the same causal variant is associated with both phenotypes. distinguishing pleiotropy from linkage, but not vertical pleiotropy (mediation) from horizontal pleiotropy (independent effects on trait and expression) [29]. As colocalisation of a GWAS loci with eQTLs is is necessary but not sufficient for mediation, it should be supported by

add uses other vars

complementary lines of evidence from other methods that integrate intermediate phenotypes (TWAS, MR, mediation analysis etc.) [29] to help untangle the multiplex of possible causal pathways from variant to trait.

1.3 Genetic effects on expression: environment is key

- The effects of eQTLs (and molecular quantitative trait loci (QTLs) in general) are incredibly context-dependent [26, 27].
  - This represents genotype-environment interactions at those eQTL.
  - A non-exhaustive list of environments that eQTLs have been found to interact with:
    * sex, age `https://academic.oup.com/hmg/article/23/7/1947/655184`
    * ancestry [30–32]
    * tissue [33, 34]
    * cell type composition in bulk samples [35–38]
    * individual cell type [30, 38–41]
    * disease status [40],
    * and experimental stimulation (see subsection 1.3.1).
- Given the effect of an eQTL can be starkly different between environments, it is difficult to determine the appropriate eQTL dataset to use for target gene prioritisation at GWAS loci.
  - It has already been shown that use of cell-type specific eQTLs increases coloc rates with GWAS hits [38] `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4498151/` `https://www.biorxiv.org/content/10.1101/2020.01.15.907436v1`
  - Successful colocalisation of GWAS loci with coloc may prioritise not only the target gene, but the specific environments most relevant to a trait.

mainmatter/figures/chapter_01/fig_mockup_systemsBio_Screenshot 2020-05-21 at 17.08.47.p

Figure 1.3: Mediation of genetic effect to phenotype, through the biological system

- What molecular mechanisms might facilitate genotype-environment interactions at eQTLs?
  - [42]: defines static, conditional, dynamic eQTLs
  - Fu *et al.* [43]: proposes TF-based mechanisms for cis-eQTL (here, define mag, damp, flip) (Fig. 1.4)
  - Gaffney [25] and Rotival [44]: suggests info on more regulatory layers will help break down transcriptional and post-transcriptional

1.3.1 Immune response expression quantitative trait loci

- A important subclass of context-dependent eQTL are response expression quantitative trait locus (reQTL), where the interacting environment is experimental stimulation [27, 45]. Most reQTL studies to date have been conducted on immune cells *in vitro*, not only because the immune system is specialised for responding to environmental exposures, but due to the abundance of immune cells easily accessible in peripheral blood, and amenable to separation (e.g. FACS) and stimulation.

mainmatter/figures/chapter_01/fig_mockup_reQTLs_Screenshot 2020-05-21 at 17.08.

Figure 1.4: eqtl mech models: magnify, dampen, flip

- *In vitro*, potential interacting variables such as cell type, and the nature, length, and intensity of stimulation can be precisely controlled.

- A seminal early study was conducted by [46], 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.
  - Since then, *in vitro* immune reQTL studies have been conducted for a variety of cell types (e.g. primary CD14+ monocytes [47]) and stimulations (IFNγ and LPS [47]).

list a few from [47] until [48]

- A complementary approach is *in vivo* reQTL mapping
  - There are numerous pros to *in vivo* stimulation.
    - the innumerable interactions in the immune system that are absent *in vitro*
    - ability to get whole organism phenotypes

    * ability to get repeated measures: can reason about change in expression over time
  - Major disadvantages: the choice of stimulation must be ethical *in vivo*, and many environmental factors (e.g. diet, lifestyle, immune exposures) cannot be controlled, leading to greater experimental noise (?), and more complex interpretations.
  - There are few published *in vivo* reQTL studies.
    * [49]: seasonal trivalent inactivated influenza vaccine (TIV), whole blood, antigen processing and intracellular trafficking genes, attempted mediation for Ab titres, but underpowered
    * [50]: fold-change expression after inactivated vaccinia vaccine, focus was on pairwise epistatic interactions, apoptosis pathways
    * [51]: whole blood, IFN status and anti-IL6 drug exposure, reQTL driven by ISRE and IRF4 motifs

- <why care about immune reQTLs>
  - Exposes differences in regulatory architecture between conditions, but does not automatically reveal the mechanisms behind those differences
  - Immune *in vitro* reQTL have been shown to be enriched more so than non-reQTL among GWAS loci for immune-related phenotypes such as susceptibility to infectious [46, 52] and immune-mediated diseases [52, 53].
  - Not yet clear whether *in vivo* reQTL have any utility on top of *in vitro* reQTL for interpreting GWAS loci: not that many studies, and complex interpretations.
  - Nevertheless, as the number of cell types systems and stimulations both *in vitro* and *in vivo* increases, the number of known reQTLs continues to grow.

1.4 Immunity is a complex trait

- Heritability of immune phenotypes is not only restricted to the expression phenotypes discussed above.

a little out of order, but since most reQTL studies are immune, I went context-specific -> reQTL -> immune rather than context-specific -> immune -> reQTL

- Studies of interindividual variation in the healthy immune system shows many aspects of the immune system are heritable and complex.
- Immune parameters are influenced by age, sex, seasonality, and chronic infection [54–58] `https://www.nature.com/articles/ncomms8000`, but most individuals have a healthy baseline immune state that is individual-specific, and relatively stable over time [55, 56, 59].

stable, yet varies by age?

- Overall estimates of the heritability of many immune parameters, such as cell composition and serum protein levels, lies between 20-40% [55–58]
- Genetic regulation is more important for the innate immune system than the adaptive immune system [57].

• A central goal of systems immunology is to establish causal relationships between the many parts and levels of the immune system

- Natural genetic variation represents small scale perturbation that is causally anchored [60, 61]
- But as discussed in the context section above, specific effects may not be apparent in the baseline state, stimulation is required
- Studies of natural infection are complicated by e.g. determining exposure.
- As in the immune in vivo reQTL studies, vaccines and drugs used as controlled immune perturbations to study the activated immune system
- reQTL may have utility for interpretation of these immune-related complex traits too, not just IMIDs/infectious disease discussed in reQTL section above

1.4.1 Immune response to vaccination is a complex trait

• Vaccination has enormous impact on global health [62]

- <quick vaccine bio>

        * 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.
        * <...>
    - A vaccine that is highly efficacious in one human population may have significantly lower efficacy in other populations. Particularly challenging populations for vaccination include the infants and elderly, pregnant, immuno comprimised patients, ethnically-diverse populations, and developing countries.
        * <1 example statistic on vaccine efficiacy differences e.g. rotavirus>
        * e.g. `https://www.sciencedirect.com/science/article/pii/S1473309918304900`
    - Traditional vaccine dev is empirical (classical "isolate, inactivate, inject" paradigm), often successful vaccine dev does not offer insights into the mechanisms of efficacy
    - The immunological mechanisms that underpin a specific vaccine's success or failure in a given individual are often poorly understood.
- A sub-discipline of systems immunology is systems vaccinology.
    - Systems vaccinology is the application of -omics technologies to provide a systems-level characterisation of the human immune system after vaccine-perturbation.
    - 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.

      define what a signature is
        * <add more to list of what vaccines have been studied, pull out of sysvacc_review_docx>

find best GWAS ref, probably mooney2013SystemsImmunogeneticsVaccines, then prune and reassign these citations

WIP, need to do more reading up on PANTS background.

Not yet sure if I need to go as wide as X=biologics for IMIDs, or only as narrow as anti-TNFs for CD

- Sysvacc informs more mechanism-based and cost-effective design (rational paradigm), and the move towards personalised vaccinology.
- Sysvacc has revealed many influences on vaccine response (age, sex, dose, adjuvants, expression signatures, microbiome, strain etc.)
- Studies of impact of host genetics is underrepresented [63]

- Like for other complex traits, from twin studies it's known that vaccine Ab responses are heritable.
  - Moving out of the candidate gene era (e.g. `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3570049/`) into GWAS.
  - [64] has heritability estimatates
  - Many loci have been implicated by GWAS e.g. HLA [63–68]
  - Overall, systems vacc studies that include genetics are nowhere near as mature compared to the trait to gene pipeline described in above e.g. for immune-mediated disease

1.4.2 Immune response to (biologic therapies for immune-mediated diseases) is a complex trait

- There is variation in response success to X for immune-mediated disease
- Many subphenotypes may be complex: immunogenicity, primary response, loss of response, remission rate, adverse fx
- Existing work on expression response to X
- Existing work on genetic factors for X

1.5 Thesis overview

- <By chapter context-content-conclusion overview.>
  - <ch 2: systems vaccinology study of Pandemrix>

        * meta analysis
    - <ch 3: in vivo reQTL study of Pandemrix>
        * Gap: relatively few studies have assessed the impact of human genetic variation on responses.
    - <ch 4: systems immunology and reQTL study of response to anti-TNF treatment in CD>
    - <discussion: limitations, future outlook>

Chapter 2

Transcriptomic response to influenza A (H1N1)pdm09 vaccine

2.1 Introduction

2.1.1 Seasonal and pandemic influenza

Influenza is an infectious disease, generally seasonal, caused by the influenza A and influenza B viruses in humans. Influenza A viruses circulate not only in humans, but also in a variety of other birds and mammal hosts. They are classified into antigenically-distinct subtypes by the combination of two surface proteins: haemagglutinin (HA) and neuraminidase (NA)[69].

There are three classes of influenza vaccine against seasonal strains in use: inactivated vaccines, live attenuated influenza vaccines (LAIVs), and recombinant HA vaccines. These vaccines confer a degree of strain-specific protection, primarily by raising serum antibodies against the HA and/or NA proteins. Antigenic drift, the accumulation of mutations in these surface proteins over time, necessitates the annual reformulation of seasonal influenza vaccines to reflect circulating strains[70, 71]. On occasion, a novel subtype against which the majority of the population is immunogically naive can arise suddenly (antigenic shift), often from zoonotic origins. A recent example occurred in 2009, when an outbreak of a novel swine-origin strain, eventually termed influenza A (H1N1)pdm09, resulted in a global pandemic,

why? for diff groups of people

the fourth to occur in the last 100 years[69].

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

2.1.2 Quantifying immune response to influenza vaccines

The 2009 pandemic motivated the rapid development, trialing, and licensing of several novel vaccines[72]. Immune response to influenza vaccines in clinical trials is evaluated by assays that measure levels of antibodies specific to the vaccine strain(s). The haemagglutination inhibition (HAI) assay measures the levels of serum antibodies specific to the HA surface protein. The related microneutralisation (MN) assay measures levels of antibodies (which may or may not be anti-HA) that neutralise the infectivity of the virus in cell culture [73]. Values from these assays can be compared against thresholds for known correlates of protection: markers that associate with whether an individual is protected from the disease. For example, HAI titres are regarded as the primary correlate of protection for inactivated influenza vaccines. Targets that regulatory agencies expect a licensed vaccine to meet are based on thresholds such as the proportion of trial individuals achieving HAI titres ≥ 40 and seroconversion ($\geq$ 4-fold increase in titres)[74, 75].

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

Here, add few points about the immunological response to adjuvanted TIVs i.e. what happens after Pandemrix admin? Involve the innate -> B/CD4T response. Goto plotkins

2.1.3 Systems vaccinology of influenza vaccines

Although HAI titres are accepted as established correlates for inactivated seasonal influenza vaccines, they fail to account for alternate mechanisms such as T cell-mediated protection, and correlates for LAIV and pandemic influenza vaccines are less reliable[70]. For novel and emerging diseases, there may be no prior knowledge of robust correlates to use in the vaccine development process. In response, the last decade has seen the rise of systems vaccinology studies: the analysis of high-dimensional data measured using multiple technologies in vaccinated individuals, in order to characterise response to vaccination at multiple levels of the biological system[76]. Such information helps elucidate a vaccine's mode of action, discover "molecular signatures" predictive of vaccine safety and efficacy, and has become an increasingly important part of the modern vaccine development chain[77, 78].

is there a more recent review?

define 'signature'

Various systems vaccinology studies of seasonal influenza vaccines have been conducted, taking longitudinal measurements pre-vaccination, and commonly at some subset of days 1, 3, 7, and 28 post-vaccination. These measurements can be correlated to changes in antibody titres after vaccination to define signatures of antibody response with potential utility as correlates of protection. One of the earliest such studies by Zhu et al.[79] found that expression of type 1 interferon-modulated genes was a signature of response to LAIV. An expression signature including *STAT1*, CD74, and E2F2 correlated with serum antibody titres after vaccination with trivalent inactivated influenza vaccine[80]; kinase CaMKIV expression is also a strong predictor [81], as are genes related to B cell proliferation [82].

For these studies of seasonal influenza vaccines in adults, responses tend to be biased by recall from past vaccination or infection[80, 83]. There have also been few studies of adjuvanted influenza vaccines, despite their superior efficacy in comparison to non-adjuvanted counterparts[84, 85].

2.1.4 The Human Immune Response Dynamics (HIRD) study

The Human Immune Response Dynamics (HIRD) study conducted by Sobolev *et al.* [86] was conceived with the above limitations in mind. The vaccine studied was Pandemrix, an AS03-adjuvanted, split-virion, inactivated vaccine against the influenza A (H1N1)pdm09 strain, for which the majority of the cohort at the time would be unlikely to have immunological memory. A total of 178 individuals were vaccinated with a single dose of Pandemrix, and longitudinal transcriptomic, cellular, antibody titre, and adverse event phenotypes were collected. Gene expression was profiled using a microarray, and differential gene expression (DGE) analyses detected genes associated with both myeloid and lymphoid effector functions upregulated at day 1, most prominently for genes associated with interferon responses. These early myeloid responses were consistent with studies of unadjuvanted seasonal influenza vaccines, but the interferon gamma-associated lymphoid response was unique to this adjuvanted vaccine.

Genes related to plasma cell development and antibody production were more highly expressed in 23 vaccine responders compared to 18 non-responders at day 7 post-vaccination. However, due to high variability among the vaccine non-responders in variables such as baseline antibody titres, a consensus predictive model that segregated the two groups could not be built, even considering other measures such as frequencies of immune cell subsets and serum cytokine levels, suggesting there was no single contributing factor that led to vaccine failure. This is in contrast to several studies of seasonal

high variability, recheck this was the reason, or quote them make sure gap and how it is filled is emphed enough

influenza vaccines, where certain expression signatures are able to predict vaccine response even pre-vaccination[87–90].

2.1.5 Chapter summary

Transcriptomic measurements in the original HIRD study were restricted to a relatively small number (46/178) of individuals, potentially limiting power to detect a expression signatures associated with antibody response. In addition, the responder vs. non-responder phenotype definition used does not account for variation in pre-existing baseline titres, and the binary definition can result in loss of statistical power[91–93].

In this chapter, I integrate the original microarray data from HIRD with RNA-sequencing (RNA-seq) data on a larger subset (75) of newly sequenced individuals from the same cohort using Bayesian random-effects meta-analysis. The overall pattern of expression over time from my meta-analysis agrees with the patterns from the original study [86], with transient innate immune response at day 1 post-vaccination, progressing to adaptive immune response by day 7.

needs 1 more punchline sentence here

From existing HAI and MN data, I compute a baseline-adjusted, continuous measure of antibody response to vaccination, the titre response index (TRI)[80]. Effect sizes of genes with expression that correlated with TRI were very dependent on measurement platform (array or RNA-seq), and no robust hits were detected in the meta-analysis. Leveraging the greater power that rank-based gene set enrichment analyses affords, I find modules of coexpressed genes that correlate with antibody response, with the strongest effects observed for adaptive immune modules at day 7, but also in inflammatory modules at baseline.

2.2 Methods

2.2.1 Existing HIRD study data and additional data

The design of the HIRD study is described in [86]. In brief, the study enrolled 178 healthy adult volunteers in the UK. The vaccine dose was administered after blood sampling on day 0; five other longitudinal blood samples were taken on days -7, 0, 1, 7, 14 and 63. Serological responses were measured on days -7 and 63 using the HAI and MN assays, and various subsets of the

why blood? ready easy supply of immune cells, despite delivery being muscle?

cohort were also profiled for serum cytokine levels (Luminex panel, days -7, 0, 1 and 7), immune cell subset counts (fluorescence-activated cell sorting (FACS) panels, all days), and peripheral blood mononuclear cell (PBMC) gene expression (microarray, days -7, 0, 1 and 7). The gene expression microarrays were performed in two batches.

In addition to the existing data, array genotypes were generated for 169 individuals; and RNA-seq data for 75 individuals at days 0, 1, and 7. The sets of individuals with gene expression assayed by microarray and RNA-seq is disjoint, as no biological material for RNA extraction remained for the microarray individuals. An overview of datasets is shown in Fig. 2.1.

2.2.2 Computing baseline-adjusted measures of antibody response

In [86], Pandemrix responders were defined as individuals with ≥4-fold titre increases in either the HAI or MN assays. This is a threshold for seroconversion set out by the U.S. Food and Drug Administration[94], and is used in many studies of seasonal influenza vaccines[77]. 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, [86] 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 [91, 93].

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

heterogeneity: well of course there was

cite appropriate subfigures here

change score is usually negatively correlated to baseline [95]. hence TRI, whilst combining, is still not ideal

append change score bit, the only thing we are concerned wtih here is clifton2019CorrelationBaselineScore

mainmatter/figures/chapter_02/graphics_ashg19/hird_design-crop.pdf

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

To address these concerns, I computed the TRI as defined in Bucasas *et al.* [80]. For each assay, a linear regression was fit with the $\log_2$ day 63/day -7 titre fold change as the response, and the $\log_2$ day -7 baseline titre as the predictor. The residuals from the two regressions were each standardized to zero mean and unit variance, then averaged. The TRI expresses a continuous measure of change in antibody titres across both assays post-vaccination, compared to individuals with a similar baseline titre, and remains comparable to the binary 4-fold change definition (Fig. 2.2).

cite appropriate subfigures here, after adding proper subfigure labels

Descriptive statistics for the 114 individuals with both gene expression and antibody titre data are presented in Table 2.1. Although the proportion of responders between array (32/44) and RNA-seq (59/70) individuals is similar ($p = 0.1551$, Fisher's exact test), the variance of TRI in array individuals is higher ($p = 0.0002098$, Levene's test), suggesting more extreme antibody response phenotypes are present (Fig. 2.3). The cause of this is unknown, there is a possibility that individuals with more extreme phenotypes were prioritised for array transcriptomics in the original HIRD study*.

2.2.3 Genotype data generation

Add to collab note that extractions were done at KCL

DNA was extracted from frozen blood using the Blood and Tissue DNeasy kit (Qiagen), and genotyping was performed using on the Infinium CoreExome-24 BeadChip (Illumina). In total, 192 samples from 176 individuals in the HIRD cohort were genotyped at 550601 markers, including replicate samples submitted for individuals where extracted DNA concentrations were low.

2.2.4 Genotype data preprocessing

Using PLINK (v1.90b3w), genotype data underwent the following quality control procedures to remove poorly genotyped samples and markers: max marker missingness across samples $< 5\%$, max sample missingness across markers $< 1\%$, max marker heterozygosity rate within 3 standard deviations of the mean (threshold selected visually to exlude outliers, Fig. 2.4), removal of markers that deviate from Hardy–Weinberg equilibrium (`--hwe` option, $p < 0.00001$).

To exclude highly-related individuals and deduplicate replicate samples, pairwise kinship coefficients were computed on minor allele frequency (MAF)

*Personal communication with authors.

mainmatter/figures/chapter_02/phenotype_data_setup.tri_comparison.pdf

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

mainmatter/figures/chapter_02/compare_phenotype_by_platform.pheno_boxplots.pdf

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

mainmatter/figures/chapter_02/coreex_eQTLflu_20171204.gencall.smajor.impute_sex.qc2.pdf

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

< 0.05 pruned genotypes using KING (v1.4). For each pair of samples with pairwise kinship coefficient > 0.177 (first-degree relatives or closer), the sample with lower marker missingness was selected.

After filtering, 169 samples and 549414 markers remained.

2.2.5 Computing genotype principal components as covariates for ancestry

As shown in Table 2.1, the HIRD cohort is multi-ethnic, hence there is potential for confounding by population structure (sample structure due to genetic background) and genetic association studies [96, 97]. Large-scale population structure explains variation in gene expression [98], so including population structure as covariates can increase power. Treating HapMap 3 samples [99] as a reference population where the major axes of variation in genotypes are likely to be ancestry, principal component analysis (PCA) was performed using smartpca (v8000) on linkage disequilibrium (LD)-pruned genotypes (`PLINK --indep-pairwise 50 5 0.2`). HIRD sample principal components (PCs) were computed by projection onto the HapMap 3 PCA eigenvectors. For non-genotyped individuals, PC values were imputed as the mean value for all genotyped individuals with the same self-reported ancestry. The top PCs separate samples of European, African and Asian ancestry (Fig. 2.5), hence these PCs can be used as continuous covariates for ancestry downstream.

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

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

2.2.6 RNA-seq data generation

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

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

RNA-seq quality metrics were assessed using FASTQC* and Qualimap[100], then visualised with MultiQC[101]. Sequence quality was high (Fig. 2.6), and

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

mainmatter/figures/chapter_02/coreex_eQTLflu_20171204.gencall.smajor.impute_sex.qc3.pru

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

duplication levels were low (Fig. 2.7). The unimodal GC-content distribution suggested negligible levels of non-human contamination (Fig. 2.8).

2.2.7 RNA-seq quantification and filtering

add software versions

Reads were quantified against the Ensembl reference transcriptome (GRCh38) using Salmon[102] in quasi-mapping-based mode, which internally accounts for transcript length and GC composition. To combine technical replicates, as the sum of Poisson distributions remains Poisson-distributed, counts for technical replicates were summed for each sample. The mean number of mapped read pairs per sample after summing was 27.09 million read pairs (range 20.24-39.14 million), representing a mean mapping rate of 80.73% (range 75.57-90.10%), comfortably within sequencing depth recommendations for DGE experiments[103]. Relative transcript abundances were summarised to Ensembl gene-level count estimates using tximport (scaledTPM method) to improve statistical robustness and interpretability[104].

Genes with short noncoding RNA biotypes* were removed, as they are generally not polyadenylated, and expression estimates can be biased by misassignment of counts from overlapping protein-coding or lncRNA genes[105]. Globin genes, which are highly expressed in erythrocytes and reticulocytes, cell types expected to be depleted in PBMC [106], were also removed. Given the proportion of removed counts at this stage was low for most samples

*miRNA, miRNA_pseudogene, miscRNA, miscRNA pseudogene, Mt rRNA, Mt tRNA, rRNA, scRNA, snlRNA, snoRNA, snRNA, tRNA, tRNA_pseudogene. List from `https://www.ensembl.org/Help/Faq?id=468`

`mainmatter/figures/chapter_02/graphics_firstYearReport/fastqc/mqc_fastqc_per_ba`

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

mainmatter/figures/chapter_02/graphics_firstYearReport/fastqc/mqc_fastqc_sequence_dupli

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

mainmatter/figures/chapter_02/graphics_firstYearReport/fastqc/mqc_fastqc_per_sequence_g

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

(Fig. 2.9), poly(A) selection and PBMC isolation procedures were deemed to have been efficient.

Many of the genes in the reference transcriptome are not expressed in PBMC (Fig. 2.10), and many genes are expressed at counts too low for statistical analysis of DGE Genes were further filtered to require detection (non-zero expression) in at least 95% of samples, and a minimum of 0.5 counts per million (CPM) in at least 20% of samples. The 0.5 CPM threshold was chosen to correspond to approximately 10 counts in the smallest library, where 10-15 counts is a rule of thumb for considering a gene to be robustly expressed[107]. 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).

2.2.8 Array data preprocessing

Single-channel Agilent 4x44K microarray (G4112F) data for 173 samples from [86] were downloaded from ArrayExpress*. These arrays were originally processed in two batches, the effect of which is seen in the raw foreground intensities (Fig. 2.12).

VSN[108] was used to perform background correction, between-array normalisation, and variance-stabilisation of intensity values, resulting in expression values on a $\log_2$ scale.

Most genes are targetted by multiple array probes; 31208 probes were collapsed into 18216 Ensembl genes using by selecting the probe with the highest mean intensity for each gene (`WGCNA::collapseRows(method=MaxMean)`, recommended for probe to gene collapsing[109]). While it would be optimal to select a collapsing method to maximise the concordance between array and RNA-seq expression values, there were no samples assayed by both platforms in the HIRD dataset. The final normalised $\log_2$ intensity values for these 18216 genes over 173 samples is shown in Fig. 2.13.

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

mainmatter/figures/chapter_02/rnaseq_data_setup.per_sample.short_ncRNA_globin_levels_hi

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

mainmatter/figures/chapter_02/rnaseq_data_setup.gene_zero_prop.pdf

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

mainmatter/figures/chapter_02/rnaseq_data_setup.sample_cpm_density_filtered.pdf

Figure 2.11: Distribution of gene expressions for RNA-seq samples before and after filtering no expression and low expression genes. Vertical line shown at CPM = 0.5 threshold.

mainmatter/figures/chapter_02/array_data_setup.array_intensity_boxplots.pdf

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

mainmatter/figures/chapter_02/array_data_setup.array_intensity_boxplots.MaxMean.pdf

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

2.2.9 Differential gene expression

PCA of the expression data reveals although samples separate by experimental timepoint along PC3 (Fig. 2.14d), measurement platform is by far the largest source of variation. Normalisation was also not able to completely remove the batch effect within the array data (Fig. 2.14a). The large platform effect likely stems from systematic technological differences in how each platform measures expression. For example, arrays suffer from ratio compression due to cross-hybridisation[110]. RNA-seq has a higher dynamic range, resulting less bias at low expression levels, but estimates are more sensitive to changes in depth than array estimates are to changes in intensity [111]. There are also differences in the statistical models behind expression quantification and normalisation, as described above.

cite relevant preprocessing sections

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

Regarding the batch effect within the array data, a popular adjustment method is ComBat[112], which estimates centering and scaling parameters by pooling information across all genes using empirical Bayes. ComBat is the method used in [86]. In comparisons of microarray batch effect adjustment methods, ComBat performs favourably (vs. five other adjustment packages)[113] or comparably (vs. batch as a fixed or random effect in the linear model)[114]. However, where batches are unbalanced in terms of sample size[115] or distribution of study groups that have an impact on expression[116], ComBat can overcorrect batch differences or bias estimates of group differences respectively. In our data, sample size and timepoint groups are fairly balanced between the two array batches, but the proportion of responders is notTable 2.2, hence I elect not to use ComBat to pre-adjust the array expression data, and model the batches as fixed effects. In practice, results from the DGE analysis were not substantially affected by the choice of whether to use a ComBat pre-adjustment or a fixed effect.

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

this is not a very precise justification. actually, if I were to color R/NR in the PCA plot, R/NR doesn't really explain a lot of var in global gene expression. that's probably why the results don't change much.

mainmatter/figures/chapter_02/compare_phenotype_by_platform.E_pca.pdf

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

2.2.9.1 Per-platform differential gene expression model

For the array data, as [86] demonstrated no significant global differences in expression between day -7 and day 0, I likewise merge these two timepoints into a single "day 0" baseline timepoint in the following DGE models.

For the RNA-seq data, between-sample normalisation was performed using the trimmed mean of M-values (TMM) method[117] from edgeR[118]; then variance-stabilisation was performed using voom[119], resulting in expression values with units of $\log_2$ CPM.

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

Linear models were fit using limma[120], which is computationally fast, and performs well for sufficiently large ($n \geq 3$ per group) sample sizes[121]. For each gene, I fit a model (model 1) with expression as the response variable; with timepoint (baseline, day 1, day 7), TRI, batch, sex, age, and the first 4 genotype PCs as fixed-effect predictors; and individual as a random-effect predictor. Within-individual correlations for the random effect were estimated using limma::duplicateCorrelation. A second model (model 2) was also fit, including 3 additional terms for the interactions between each timepoint and TRI. Contrasts were defined, testing if linear combinations of estimated coefficients are different from zero. From model 1, I defined contrasts for day 1 vs. baseline, day 7 vs. baseline, day 7 vs. day 1, TRI, sex, and age. From model 2, I defined contrasts for the TRI specifically at each of the three timepoints. Corresponding coefficients and standard errors for the contrasts were extracted from the linear models, which represent effect size in units of $\log_2$ expression fold change per unit change in predictor value.

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

add equation from ch3. especially justify having TRI in as predictor, by noting equiv of traditional lm to contrasts

2.2.9.2 Choice of differential gene expression meta-analysis method

In the section , I concluded that a two-stage meta-analysis approach would be appropriate. This meta-analysis is restricted to 13593 genes assayed by both the array and RNA-seq platforms.

add section labels

Two popular frameworks for effect size meta-analysis are fixed-effect and random-effects[122, 123]. Given k studies, the fixed-effect model assumes a common population effect size shared across all studies, with observed variation explained only by sampling error. The random-effects model assumes the k study-specific effect sizes are drawn from some distribution with variance τ^2 (standard deviation (SD) τ), representing an additional source of variation termed the between-studies heterogeneity, reducing to the fixed-effect model when $\tau = 0$. In the HIRD data, there are $k = 2$ 'studies' (array and RNA-seq), where the platform differences described in section contribute to considerable between-studies heterogeneity. The assumption of $\tau = 0$ is unrealistic, hence a random-effects model is more appropriate.

add label

make all the notation in this section consistent with, and add the equation 2.1. The normal-normal hierarchical model, [124]

Unfortunately, there is no optimal solution for directly estimating τ in random-effects meta-analyses with small k[125], in the case of $k = 2$ especially[126]. Many estimators are available[127], but lack of information with small k causes estimation to be imprecise, and often results in boundary values of $\tau = 0$ that are incompatible with the assumed positive heterogeneity[128, 129]. In such circumstances, the most sensible choice may be to incorporate prior information about model hyperparameters in a Bayesian random-effects framework[127–130]. For this study, I use the implementation in bayesmeta [124], which requires priors for both effect size and between-studies heterogeneity.

2.2.9.3 Prior for between-studies heterogeneity

The choice of prior for between-studies heterogeneity is influential when k is small[130]. Gelman [131] considers the case of $k = 3$, showing that a flat prior places too much weight on implausibly large estimates of τ, and recommends a weakly informative prior that acts to regularise the posterior distribution. Since I assumed zero estimates for τ are unrealistic, I use a weakly-informative gamma prior recommended by [128], which has zero density at $\tau = 0$, increasing gently as τ increases. This constrains τ to be positive, but still permits estimates close to zero if the data support it. This is in constrast to priors used in other studies from the log-normal (e.g. [132, 133]) or inverse-gamma (e.g. [134]) families that have derivatives or zero close to zero, thus ruling out small values of τ no matter what the data suggest; and in contrast to half-t family priors (e.g. [130, 131]), which have their mode at zero, and do not rule out $\tau = 0$.

To estimate the appropriate shape and scale parameters for the gamma empirically, a frequentist random-effects model using the restricted maximum likelihood (REML) estimator for τ (recommended for continuous effects[127]) was first for each gene using `metafor::rma`. Genes with small estimates of $\tau < 0.01$ were excluded, and a gamma distribution was fit to the remaining estimates using `fitdistrplus`.

2.2.9.4 Prior for effect size

While the choice of prior on τ is influential when k is small, there is usually enough data to estimate the effect size μ such that any reasonable non-informative prior can be used [129, 131]. `bayesmeta` implements both flat and normal priors for μ. Assuming that most genes are not differentially expressed with effect sizes distributed randomly around zero, I selected a normal prior with $N(\mu = 0, \sigma^2)$, over a flat prior. As in the section above, to determine an appropriate scale, a normal distribution with mean $\mu = 0$ was fit to the distribution of effect sizes from the gene-wise frequentist models to empirically estimate σ.

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

Heavy-tailed Cauchy priors have been proposed for effect size distributions in DGE experiments to avoid over-shrinkage of true large effects in the tails[135]. Since `bayesmeta` does not implement a Cauchy prior, to avoid over-shrinkage, I flatten the normal prior considerably by scaling up the variance to $N(0, 100\sigma^2)$. This is equivalent to assuming placing a 95% prior probability that effects are less extreme than approximately 20σ.

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

2.2.9.5 Evaluation of priors

An example of the empirically estimated hyperparameters for the priors for the day 1 vs. baseline contrast are shown in Fig. 2.15 (for τ) and Fig. 2.16 (for μ). For τ, the final prior used was $\text{Gamma}(\text{shape} = 1.5693, \text{scale} = 0.0641)$. This is comparable to [128]'s default recommendation of a $\text{Gamma}(\text{shape} = 2, \text{scale} = \lambda)$ prior where λ is small. For μ, the final prior used was $N(0, (0.3240 * 10)^2)$. The tails of the non-scaled normal fit (black) are light compared to the Cauchy fit (red), which may lead to over-shrinkage, especially since there are many genes with high positive fold changes for the day 1 vs. baseline effect.

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

2.2.9.6 Multiple testing correction

For the frequentist random-effects meta-analysis, nominal gene-wise p-values are converted to false discovery rate (FDR) estimates using the Benjamini-Hochberg (BH) procedure (`p.adjust` in R). For the Bayesian random-effects meta-analysis, posterior effect sizes and standard errors are supplied to `ashr`, which estimates the local false sign rates (lfsrs), which are analogous to FDR, but quantifies the probability of calling the wrong sign for an effect rather

add note on ositive regression dependency [21] mainmatter/figures/chapter_02/meta.bayesmeta.priors.coefName_d1.vs.d0.pdf

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

mainmatter/figures/chapter_02/meta.bayesmeta.priors.coefName_d1.vs.d0.pdf

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

add comment on symmetry

than than the confidence of a non-zero effect[136].

2.2.10 Gene set enrichment analysis using blood transcription modules

Gene set enrichment analyses were conducted using `tmod::tmodCERNOtest`[137], which assesses the enrichment of small ranks within specific sets of genes compared to all genes, when the genes are ranked by some metric—here I used effect sizes from `bayesmeta`. The gene sets used were blood transcription modules (BTMs) from[138], which are annotated sets of coexpressed genes mined from publicly available human blood transcriptomic data, and provide sets tailored for enrichment analyses in blood cells.

2.3 Results

more text

2.3.1 Extensive global changes in expression after vaccination

To gain an overview of how the transcriptome changes after vaccination, linear models were fit to identify genes differentially expressed at day 1 or day 7 compared to baseline (day -7 and day 0) in the HIRD array and RNA-seq expression data, accounting for covariates such as batch effects, sex, age, TRI, and ancestry. At 13593 genes with expression measured by both platforms, models were fit within each platform, then effect sizes were combined using Bayesian random-effects meta-analysis.

At a lfsr < 0.05 and absolute FC > 1.5 cutoff, 857/13593 genes were differentially expressed between any pair of timepoints, with their expression clustering into three main clusters (Fig. 2.17).

2.3.2 Innate immune response at day 1 post-vaccination

Consistent with global expression at day 1 being markedly different from expression at other timepoints (Fig. 2.14), the highest numbers of differentially expressed genes are observed at day 1, with 644 genes differentially expressed vs. baseline. The majority of these (580/644) were upregulated. The gene with the highest FC increase at day 1 compared to baseline was *ANKRD22* ($\log_2$ FC $= 4.49$), an interferon-induced gene in monomainmatter/figures/chapter_02/plot_dge_eqtl.heatmap_dge.pdf

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

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

not sure of interpretation at FGFBP2, it is indeed highly expressed in NKs through `https://dice-database.org/genes/FGFBP2`

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

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

finish citing

add label

cytes and dendritic cells (DCs) involved in antiviral innate immune pathways[139]. Other key genes in the interferon signalling pathway[140] such as *STAT1* ($\log_2$ FC = 2.1693060), *STAT2* ($\log_2$ FC = 0.9489341), and *IRF9* ($\log_2$ FC = 0.8153674) are also upregulated at day 1. Gene set enrichment analysis using `tmod` revealed that genes with the high FC increases at day 1 were enriched in modules associated with activated DCs, monocytes, toll-like receptor and inflammatory signalling (Fig. 2.18), confirming that day 1 responses are dominated by signatures of innate immunity. 64 genes were downregulated at day 1, enriched in modules associated with T cells and natural killer (NK) cells, with the largest absolute fold change observed for *FGFBP2* ($\log_2$ FC = −0.9141547). For both up and downregulated genes, there was a tendency to return to baseline expression levels by day 7.

2.3.3 Adaptive immune response at day 7 post-vaccination

59 genes were differentially expressed at day 7 vs. baseline, with expression fold changes more modest than those at day 1. The genes with the highest upregulation were the B cell-associated genes *TNFRSF17* ($\log_2$ FC = 1.7538617) and *MZB1* ($\log_2$ FC = 1.7369668). Plasma cell-specific genes including *SDC1* (encodes CD138 `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5437827/`) ($\log_2$ FC = 1.3673081) and *ELL2* (`https://www.nature.com/articles/ni.1786`) ($\log_2$ FC = 0.8679659) were also prominently upregulated. Strongly enriched modules at day 7 were related to mitosis and cell proliferation, particularly in CD4$^+$ T cells (Fig. 2.18). Both the CD4$^+$ T cell and plasma cell response are indications of an adaptive immune response at day 7.

2.3.4 Expression signatures associated with antibody response

I also looked for genes which have expression associated with baseline-adjusted antibody response, as quantified by TRI. At the initial frequentist meta-analysis stage, with a significance threshold of FDR < 0.05, 6 genes had expression associated with TRI at baseline, 55 at day 7, and 11 pooling samples across timepoints (Fig. 2.19). [86] also identified genes with day 7 expression associated with antibody response, where response was defined as a binary phenotype based on 4-fold change (described in section). They reported 62 significant associations at FDR < 0.05, of which 58/62 fall into the mainmatter/figures/chapter_02/compare_dge_eqtl.tmodDotPlot.DGE.timepoint.pdf

Figure 2.18: Transcriptomic modules significantly up or downregulated post-vaccination. Size of circle indicates effect size. Color of circle indicates significance and direction of effect (red = upregulation, blue = downregulation).

13593 genes considered in my meta-analysis (circled, Fig. 2.19), and 15/58 replicated, all with the same positive direction of effect (high expression with high TRI). In the Bayesian meta-analysis, no single gene was detected as significantly associated with TRI at lfsr < 0.05 at any timepoint, or when pooling samples across all timepoints (Fig. 2.20).

Significant enrichments were detected at the gene set level; the strongest effects are seen at day 7, where expression of cell cycle, $CD4^+$ T cells, and plasma cells are associated with high TRI. At day 0, modules related with inflammatory response in myeloid cells are also associated with high TRI (Fig. 2.21).

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

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

2.4 Discussion

There is extensive transcriptomic response to Pandemrix vaccination in the HIRD cohort. Upregulation of genes and modules related to the interferon signalling pathway, monocytes, inflammatory response, and other aspects of innate immunity were detected at day 1. This response is transient, with most such genes returning to baseline expression by day 7. Upregulation of cell cycle/proliferation, activated $CD4^+$ T cell, and B (plasma) cell genes and modules were detected at day 7. This is likely a signature indicating the shift to an adaptive immune response, involving $CD4^+$ T cell-supported differention and proliferation of antibody-secreting plasmablasts and plasma cells[142]. These patterns of expression change between timepoints in the RNA-seq data are consistent with the patterns in the array data in the original study[86], and with expansions of monocyte and plasma cell populations seen in the FACS data at days 1 and 7 respectively in the original HIRD study[86].

Not sure if there is a biological interpretion of downreg of T cells and NK cells gene sets at day 1, since it could be due to increase in other cell types in the sample. similar findings in [141] though

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

In contrast, I was not able to fully replicate the originally reported single gene-level associations between day 7 expression and antibody response in the RNA-seq data and subsequent and meta-analyses. In [86], 62 genes were reported as differentially expressed between vaccine responders and non-responders. Although [86] encodes responder status as a binary phenotype, whereas my analysis uses TRI, this is not the primary difference, as 51/62 mainmatter/figures/chapter_02/plot_dge_eqtl.DGE.effectSizeComparison.pdf

Figure 2.19: DGE effect sizes estimated in array vs. RNA-seq. Significance colored by frequentist random effects meta-analysis FDR < 0.05. Genes with day 7 expression associated with responder/non-responder status in [86] are circled for that contrast.

mainmatter/figures/chapter_02/plot_dge_eqtl.DGE.effectSizeComparison.pdf

Figure 2.20: DGE effect sizes estimated in array vs RNA-seq. Significance colored by Bayesian random effects meta-analysis lfsr < 0.05. Genes with day 7 expression associated with responder/non-responder status in [86] are circled for that contrast.

mainmatter/figures/chapter_02/compare_dge_eqtl.tmodDotPlot.DGE.TRI.pdf

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

genes replicated (FDR < 0.05) using TRI when considering just the array data. The same analysis using only the RNA-seq data replicated 0/62 genes. The majority of the effects for these genes were simply much stronger in the array dataset than in the RNAseq dataset (Fig. 2.19). Given that the range of TRI is higher in the array individuals (Table 2.1), this does not seem unusual that stronger TRI-associated effects are observed there.

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

move numbers to results?

58/62 reported hits were measured by both platforms and assessed in the meta-analysis. Only 15/58 signals replicated using frequentist random-effects meta-analysis to combine per-platform estimates. I do not consider these hits as robust, as the REML estimate of between-platform heterogeneity was zero for 8563/13593 for the day 7 TRI contrast overall, and zero for all 15 of these signals. None of these signals replicated in the Bayesian random-effects meta-analysis. The Bayesian meta-analysis is in general more conservative, calling fewer differentially expressed genes compared to the frequentist analysis for all contrasts (Fig. 2.20). Prior information about τ is incorporated, discouraging unrealistic estimates of zero heterogeneity. Given the between-platform heterogeneity coming from both platform-specific technical differences and TRI phenotype differences, relative to the modest effect size distributions compared to between-timepoint DGE comparisons, the data are not well-positioned to identify significant single-gene associations with antibody response.

Expression signatures of antibody response were, however, observed at the gene set level, for modules of coexpressed genes that are associated with TRI as a whole. The strongest effects were observed at day 7, where expression of adaptive immune response modules (cell cycle, stimulated CD4$^+$ cell, plasma cell modules) were positively associated with TRI. These are the same modules observed to be upregulated at day 7 compared to baseline; it seems that those individuals with the greatest antibody response to vaccination are most able to upregulate these gene sets by day 7 post-vaccination.

Module associations were also observed pre-vaccination (cell adhesion, enriched in B cells, proinflammatory cytokines, platelet activation), suggesting baseline immune state has some influence on long-term antibody response to Pandemrix. Over the years, a diverse range of gene sets have been found to be baseline predictors of serological response to influenza vaccination: apoptosis[87]; Fcγ receptor-mediated phagocytosis, TREM1 signaling[88]; enriched in B cells, T cell activation[89]; B cell receptor signalling, could comment on phenotype differences too, i.e. HIRD measure antibodies at d63, much later than is popular in the field: d28 usually

should probably emph sobolev didn't find pre-vacc signatures, and we did. But it's not exactly fair, as sobolev didn't use gene set enrichment as far as i can tell

There is also something to be said about 'prediction is not inference'. For use as correlates of protection, as promised by proponents of systems studies, prediction is what is important.

At no point in this chapter are we estimating causal effects

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

inflammatory response, platelet activation [90]; several of which I also observe. It should be noted that comparisons with these signatures from existing influenza systems vaccinology studies should caveated, as most existing studies are for non-adjuvanted influenza vaccines. Adjuvanted influenza vaccines are considerably more immunogenic, and post-vaccination expression patterns differ to those of non-adjuvanted vaccines [84, 86]. Hence, it is particularly important that the robustness of these observed baseline expression signatures be validated in an independent cohort for a comparable AS03-adjuvanted influenza vaccine.

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

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	batch 1	2	DN500165J	DN500166K	DN500167L
	n = 374	n = 87	n = 79	n = 70	n = 69	n = 69
visit						
v1	40 (10.7%)	20 (23%)	20 (25.3%)	0 (0%)	0 (0%)	0 (0%)
v2	114 (30.5%)	24 (27.6%)	20 (25.3%)	24 (34.3%)	23 (33.3%)	23 (33.3%)
v3	109 (29.1%)	21 (24.1%)	20 (25.3%)	22 (31.4%)	23 (33.3%)	23 (33.3%)
v4	111 (29.7%)	22 (25.3%)	19 (24.1%)	24 (34.3%)	23 (33.3%)	23 (33.3%)
responder						
FALSE	80 (21.4%)	12 (13.8%)	36 (45.6%)	11 (15.7%)	9 (13%)	12 (17.4%)
TRUE	294 (78.6%)	75 (86.2%)	43 (54.4%)	59 (84.3%)	60 (87%)	57 (82.6%)
TRI						
	-0.1 (1.0)	-0.1 (1.0)	-0.4 (1.4)	0.1 (0.6)	-0.0 (0.8)	0.2 (0.6)

Chapter 3

Genetic factors affecting Pandemrix vaccine response

3.1 Introduction

3.1.1 Genetic factors affecting influenza vaccine response

Vaccination is the most effective way by which seasonal influenza is controlled[70], and the mechanism by which influenza vaccines are efficacious is by raising strain-specific antibodies protective against future infection[143]. Humoral responses are influenced by vaccine-associated factors (e.g. type, dose, adjuvants), but are also a complex trait influenced by host genetics[63, 66]. Genetic variants associated with antibody response have been detected for vaccines such as hepatitis B, influenza, measles, rubella, and smallpox[63, 68]. For antibody response to seasonal influenza vaccines, studies have implicated genetic variation within cytokine genes, cytokine receptors[144]; antigen processing and intracellular trafficking genes[49]; immunoglobulin heavy-chain variable region loci[145]; and specific human leukocyte antigen (HLA) alleles [144, 146].

A potential mechanism through which genetic variation can play a causal role in influenza vaccine response is through altering the expression of genes as expression quantitative trait loci (eQTLs). eQTL can have condition-specificity: an interaction between their effect on expression and different environmental contexts such as tissue or cell type[26, 27]. The mechanisms by which eQTLs interact with environment are of great interest; for example, cell type specificity can inform us about how expression is regulated in pull in citations from intro

a cell type specific manner[38]. In a vaccination context, an important subset of environment-interacting eQTLs are response expression quantitative trait loci (reQTLs), defined as an eQTL whose effect interacts with external stimulation or perturbation. reQTL have been observed in many human cell types *in vitro*, or in the whole organism *in vivo*. As the pre- and post-stimulation environments are separated in time, a possible mechanism that leads to the observation of reQTL is a genotype-dependent change in gene expression between timepoints, which may underly genotype-dependent differences in antibody phenotypes.

3.1.2 Response expression quantitative trait loci for seasonal influenza vaccination

reQTL can be mapped considering a vaccine as an *in vivo* immune stimulation, looking for genotype-dependent changes in gene expression in immune cells. Little work has been done on vaccine-stimulated reQTLs, except one study conducted for the seasonal trivalent inactivated influenza vaccine (TIV). [49] collected longitudinal data in 247 European adults: peripheral whole blood gene expression measured at four timepoints (day 0, 1, 3, 14), and antibody titres measured at three timepoints (day 0, 14, 28). They identified 20 genes with a cis-eQTL effect, expression correlation with antibody response, and either post-vaccination differential expression *or* a reQTL effect at that cis-eQTL. Genes involved in intracellular antigen transport and processing were enriched among those 20 genes.

3.1.3 Chapter summary

distinction between expression/ab response is blurry here

The HIRD cohort represents a unique opportunity for detecting genetic contributions to influenza vaccine response. In chapter 2, we observed global changes in gene expression after Pandemrix vaccination, as well as expression signatures correlated to degree of antibody response. For seasonal influenza vaccines, the contribution is small: antibody responses in adults are largely driven by non-genetic influences such as previous influenza vaccination or infection[54]. As the Pandemrix vaccine is against the pdm09 pandemic strain that was not in seasonal circulation at the time the Human Immune Response Dynamics (HIRD) cohort was recruited (2010-11), with individuals mounting an expression response that was not recall-dominated

[86], the relative contribution of genetic factors to Pandemrix response may be greater.

straighten out tenses

In this chapter, I model the influence of common host genetic variation on longitudinal *in vivo* expression response to Pandemrix. I map cis-eQTL within each timepoint, accounting for ancestry, cell type abundance and unmeasured covariates, then call shared and reQTL effects from a joint model, looking for genes where the lead eQTL has a different effect size pre- and post-vaccination. Many of the strongest reQTL effects involve opposite signed effects on expression for the same variant at different timepoints. I detect a strong day 1 specific reQTL effect at *ADCY3*. Through modelling interaction of reQTL with cell type abundance estimates and statistical colocalisation with cell type specific QTL datasets, the reQTL signal was determined to be a monocyte-specific effect likely driven by increase monocyte abundance at day 1.

1 more sentence to round off context

3.2 Methods

3.2.1 Genotype phasing and imputation

Prior to imputation, 213277 monomorphic variants that provide no information for imputation were removed. Imputation for the autosomes and X chromosome was conducted using the Sanger Imputation Service, which involves pre-phasing with EAGLE2 (v2.4) and imputation with PBWT (v3.1) against the Haplotype Reference Consortium (r1.1) panel `https://www.ncbi.nlm.nih.gov/pubmed/27548312`. Variants were lifted-over from GRCh37 to GRCh38 coordinates using CrossMap. Poorly-imputed variants with INFO < 0.4 or post-imputation missingness $> 5\%$ were removed, leaving 40290981 variants.

3.2.2 Overall strategy for detecting reQTLs

Since the aim of this chapter is to identify genetic variation that affects expression response to vaccination, it may seem most direct to model the change in each individual's expression after vaccination as the response variable. This approach has been applied for identification of condition-specific eQTL, typically with the response taking units of log fold change between conditions (e.g. [42, 147, 148]). Although potentially powerful if eQTL effects are small and opposite between conditions[42], it is analogous to the "change score" approach, which can suffer from regression to the mean, and increased uncertainty from the variance sum law if effects between conditions have positive covariance[95, 149].

upend change score bit

Instead, I map eQTLs within each of three timepoint conditions (day 0 pre-vaccination, day 1, and day 7), and find reQTLs by looking for eQTLs that have different effects between conditions. Unlike a test for difference implemented using a genotype-condition interaction term in a joint regression model, homoscedasticity of errors is not assumed for all conditions[150]

the variable is not used as an inclusion/exclusion criterion for the study, otherwise regression to the mean will be strong

Can this really demonstrate genotype-dependent change in gene expression between timepoints? i.e. need understand how the change score/ANCOVA approaches differ from repeated measures ANOVA differ from the interaction/stratified approach I take?

Within each timepoint, recall the the HIRD dataset includes expression measured by both array and RNA-sequencing (RNA-seq). As discussed in subsubsection 2.2.9.2, it is difficult to directly estimate the between-studies heterogeneity when the number of studies is small, and Bayesian meta-analysis was preferred for combining array and RNA-seq differential gene expression (DGE) estimates. That method does not scale to eQTL analysis, where the number of tests is large, in the order of thousands of tests per gene, versus the handful DGE contrasts per gene performed in chapter 2. Instead, I perform a mega-analysis within each timepoint, first merging array and RNA-seq expression estimates into a single matrix with ComBat[112]. For comparison purposes, analyses were also run using in the array and RNA-seq samples separately.

why I didn't just do a mega-analysis in chapter 2 then, given I haven't any evidence if it's better or worse than Bayesian meta-analysis in that context.

add -7 note as with ch2

Defining whether an eQTL is shared between conditions can be a tricky business. Naively, one can map eQTLs separately in each condition, then assess the overlap of significant associations between conditions. This underestimates sharing due to the difficulty of distinguishing true lack of sharing from missed discoveries from incomplete power within each condition [40, 151]. Condition-by-condition analysis also makes no attempt to borrow information across conditions for mapping shared associations[151–153]. Counterintuitively, a joint multivariate analysis may be more powerful even when associations are not shared across all conditions[154].

A variety of models have been employed for joint eQTL mapping, including the use of classical multivariate methods such as multivariate analysis of variance (MANOVA)[155], frequentist meta-analyses (e.g. `Meta-Tissue`[156], METASOFT), and Bayesian models (e.g. eQtlBma[151], MT-HESS, MT-eQTL). Joint mapping has been repeatedly been demonstrated to be more powerful than condition-by-condition analysis, and recent methods are now computationally efficient when scaling to large numbers of conditions and variants tested (e.g. RECOV[157], mashr[152], HT-eQTL[153]). In this chapter, I apply `mashr`[152] for the estimation of eQTL effects across my three timepoints. `mashr` learns patterns of correlation among multiple conditions empirically from condition-by-condition summary statistics, then applies shrinkage to provide improved posterior effect size estimates, and compute measures of significance per condition.

3.2.3 Controlling for population structure with linear mixed models

There is population structure due to ancestry in the HIRD cohort, which was incorporated in DGE analyses by treating the top principal components (PCs) of the genotype matrix as continuous covariates for large-scale population structure (subsection 2.2.5). In the context of eQTL mapping (and genetic association studies in general), where the aim is to assess the marginal effect of a single genetic variant on expression, population structure can be correlated with both expression (e.g. through polygenic effects) and the tested variant (e.g. through ancestry-dependent frequency differences). This leads to omitted-variable bias (OVB) from confounding, which if not controlled for, leads to genome-wide inflation of test statistics [158]. An efficient approach is the linear mixed model (LMM) with a random effect that incorporates genetic correlation between individuals, usually in the form of a kinship matrix, into the covariance of that random effect[97, 158, 159] The LMM approach has the advantage of not only modelling large-scale population structure, but also cryptic relatedness (the presence of closely related individuals in a sample assumed to consist of unrelated individuals[160]) due to finer-scale effects such as family structure[159].

add some indication of how much inflation can be reduced by LMMs

3.2.3.1 Estimation of kinship matrices

When testing a variant for association using LMMs, to avoid loss of power from "proximal contamination", the kinship matrix used should not include that variant[161]. A simple way to avoid this is to compute a leave-one-chromosome-out (LOCO) kinship matrix using all variants on chromosomes other than the tested variant's chromosome[162].

I estimated kinship in the HIRD data from common autosomal variants, add chr1 loco kinship matrix as example, note the estimates for self-relatedness on the diagonals are not constrained to be 1

helps with coloc

emph here that the sims match what my def of reqtl is for rest of chapter

log scale: as interactions depend on the scale at which departure from additivity is detected

using `LDAK` (5.0), which computes kinship matrices adjusted for bias caused by linkage disequilibrium (LD)[163]. Filtered, pre-imputation sample genotypes from subsection 3.2.1 were pruned to MAF > 0.05. A kinship matrix was computed for each autosome, then combined into a single genome-wide matrix using `LDAK --join-kins`. To obtain a LOCO kinship matrix for each autosome, each autosome's kinship matrix was then subtracted from this genome-wide matrix (`LDAK --sub-grm`).

3.2.4 Additional eQTL-specific expression preprocessing

There are a number of transformations often applied to expression data before eQTL mapping, such as the rank-based inverse normal transformation (INT) (e.g. GTEx v8[164]), which conforms often non-normal expression data to an approximately normal distribution, and reduces the impact of expression outliers. In the context of genetic association studies, the practice of applying rank-based INT to phenotypes has been criticised for only guaranteeing approximate normality of residuals when effect sizes are small, and potential inflation of type I error, especially in linear models that include interactions[165]. In multi-condition datasets, these transformations are also typically applied within conditions (e.g. within each tissue individually in GTEx v8[164]). Another common transform is standardising (centering and scaling to zero mean and unit variance) (e.g. eQTLGen Consortium[166]), often done so that effects across genes and studies can be comparably interpreted in units of standard deviation expression[167].

I performed simulations to evaluate the effect of these transformations on reQTL detection between a hypothetical baseline and day 1 post-vaccination condition. Expression values on the log scale were simulated with the eQTL slope (beta) set to specific values corresponding to six scenarios for six gene-variant pairs (Fig. 3.1). The simulated scenarios were subjected to rank-based INT (Blom method[165]), standardisation (both centering and scaling), scaling-only, and centering-only transformations. Transformations were applied both within each condition and without separating conditions.

The boxed facets in Fig. 3.1 represent undesirable effects of transformations on reQTL calls. For example, rank-based INT induces false shared eQTL effects in scenarios 4 and 5. In general, transformations that scale within condition are not appropriate, as different variance between conditions can be what drives a reQTL effect. Scaling without separating conditions can also be problematic, since the total variance also contributes to the reQTL effect size. For example, scenarios 2 and 4 have the same 1 unit increase in slope pre-transformation (the same fold-change between conditions), but after scaling-only the beta increases are $0.75 - 0 = 0.75$ and $0.8 - 0.4 = 0.4$ respectively—eQTL 4 now looks like a weaker effect.

In light of these simulations, I decided that neither rank-based INT nor standardisation were appropriate given my intent of detecting reQTLs between conditions. Only the centering-only transformation avoided both false shared effects and preserves relative reQTL effect sizes between genes. The simple inclusion of an intercept term in the eQTL model already achieves this. Not performing any rank-based transform does lose the advantage of reining in outliers. The expression data have already been preprocessed to remove low-expression outliers in subsection 2.2.7, but automatic outlier exclusion based on standard deviation (SD) thresholds at the eQTL mapping step could be considered in future implementations[166]. Note that many preprocessing steps done prior to this stage in the pipeline (e.g. variance-stabilisation, ComBat batch effect correction) are also expression transformations, but I only consider the preservation of reQTL effects defined from expression values post-adjustment for those technical effects to be important.

add sample sizes and model for expression sim

3.2.5 Estimation of cell type abundance from expression

Peripheral blood mononuclear cell (PBMC) samples are a mixture of immune cells, and a fixed input of RNA extracted from that mixture is used to estimate expression, so estimates for genes that have cell type specific expression depend on the relative proportions of each cell type in each sample, These proportions shift after Pandemrix vaccination[86], and eQTL effects can also be cell type specific. As genotype can be assumed to stay constant, it is valid to compare the effect of genotype on expression between multiple timepoints to call reQTLs, but changes in cell type abundance influence this by both expression and the effect of genotype on expression. Immune cell abundance also varies naturally between healthy individuals[54, 56], so it is important to model these effects even at baseline.

determine appropriate citations from existing refs in intro

??

Cell type abundance directly measured via fluorescence-activated cell sorting (FACS) are only available for a small subset of HIRD individuals (subsection 2.2.1), so I derived cell type abundance estimates from the exmainmatter/figures/chapter_03/simulate_expression_transforms.pdf

Figure 3.1: Simulated log scale expression in two conditions for six genes (columns) representing six different scenarios: Scenario 0 has no eQTL, scenario 1 is a shared eQTL (beta = 1), scenario 2 is a reQTL where beta increases from 0 to 1, scenario 3 is a reQTL where beta increases from 0 to 2, scenario 4 is a reQTL where beta increases from 1 to 2, and scenario 6 is a reQTL where beta increases from 1 to 4. Rows represent the effect of different expression transformations across samples, conducted both within condition, and including both conditions.

pression data as an alternative. Such estimates have previously been used in eQTL analyses from bulk samples where cell type specific effects are expected[35, 38, 51]. As the estimates are based on the expression of multiple genes, is not entirely circular to use them as covariates in this way for genewise eQTL models. I selected `xCell`[168], which previously been shown to outperform other deconvolution methods for cell type specific eQTL mapping in blood[38]. `xCell` computes enrichment scores based on the expression ranks of approximately 10000 signature genes derived from purified cell types, works for both array and RNA-seq expression data, and implements "spillover compensation" to reduce dependency of estimates between related cell types[168]. `xCell` was originally developed for tumor samples, so many of the built-in cell types are not expected in PBMC. Reviewing the literature to find which broad classes of peripheral blood cell types are commonly-expected in the PBMC compartment[51, 169, 170], I selected 7/64 of the built-in cell types: $CD4^+$ T cells, $CD8^+$ T cells, B cells, plasma cells, natural killer (NK) cells, monocytes, and dendritic cells (DCs). Array and RNA-seq data from subsection 2.2.8 and subsection 2.2.7 were processed through `xCell` separately. The large batch effect present in the array expression was first removed using ComBat. Finally, enrichment scores were standardised, so that a score of zero estimates the average abundance of that cell type across all timepoints (Fig. 3.2 and Fig. 3.3).

add comment on existence of chosen cell types in samples, and clustering by visit

As with actual cell type abundances, the enrichment scores are correlated. Multicollinearity will be a problem for interpreting effect size estimates when these scores are used as indepedent variables in regression downstream. To prune the number of scores, I performed a principal component analysis (PCA) of the cell type scores across samples, determined the number of principal components that exceed the eigenvalues-greater-than-one rule of thumb[171], then selected only the one cell type with the highest contribution for each of those components. In both array and RNA-seq datasets, the number of components retained was three, and the selected cell types were monocytes, NK cells, and plasma cells (Fig. 3.4). The choice to use the actual cell type scores over principal components directly as covariates is a sacrifice of orthogonality for interpretability.

does not bias, but unstable and more var of estimates

no need for both size and color, use one for contribution percent

Scores were validated against FACS measurements in the subset of individuals that had them. Depending on each panel's gating strategy for each cell subset, the FACS data were in units of either absolute counts, or per-

add info on the markers used for the chosen FACS counterparts mainmatter/figures/chapter_03/get_xCell_estimates.dataset_array.plots.pdf

Figure 3.2: Standardised xCell enrichment scores for seven PBMC cell types in array samples.

mainmatter/figures/chapter_03/get_xCell_estimates.dataset_rnaseq.plots.pdf

Figure 3.3: Standardised xCell enrichment scores for seven PBMC cell types in RNA-seq samples.

mainmatter/figures/chapter_03/get_xCell_estimates.dataset_array.plots.pdf

(a) Array estimates.

mainmatter/figures/chapter_03/get_xCell_estimates.dataset_rnaseq.plots.pdf centage of the previously gated population. A rank-based INT was applied within each panel and cell subset, so that the transformed measure could be compared between individuals for each subset ([172] takes a similar approach for cell abundance data using a quantile-based INT). Missing values were imputed with `missForest`, a random forest imputation method suitable for high-dimensional data where $p \gg n$. Although the increase in xCell score for monocytes at day 1 and plasma cells at day 7 reflect the increases in these cell types observed by[86], overall correlation between xCell and FACS was weak (Fig. 3.5). Weighing the downside of having imperfect estimates of cell type abundance against the downsides of not accounting for abundance, or excluding samples without FACS measures, I chose to continue the analysis using the xCell scores.

get subset size

3.2.6 Finding hidden covariates using factor analysis

Apart from cell type abundance, a myriad of other unmeasured variables contribute to expression variation. Hidden determinants of expression variation were learnt using `PEER`[173]. As suggested by [173], between-sample normalisation and variance stabilisation RNA-seq data was performed using `DESeq2::vsn`. ComBat was applied to first merge array and RNA-seq data into a single log scale expression matrix per timepoint, treating the largest global effects on expression—the two array batches and three RNA-seq library prep pools (Fig. 2.14)—as known batch effects. Given known covariates (intercept, sex, four genotype PCs from subsection 2.2.5 representing ancestry, and the three xCell scores estimated above), PEER was used to estimate additional hidden factors that explain variation in expression matrix. Factors are assumed to be unmeasured covariates that have global effects on a large fraction of genes, whereas a cis-eQTL will typically only have local effects, so including factors as covariates should not introduce dependence with the genotype term, but should soak up some of residual variation, improving power to detect cis-eQTLs. The analysis was run per timepoint, otherwise global changes in expression between timepoints induced by the vaccine would be recapitulated as factors.

Correlating the estimated factors to a larger set of known covariates reveals many correlations with xCell estimates, indicating that cell type abundance does indeed have substantial global effects on the expression matrix. There is little correlation with known array or RNA-seq batch effects, mainmatter/figures/chapter_03/validate_xCell_estimates.cell_t

(a) Monocytes.

mainmatter/figures/chapter_03/validate_xCell_estimates.cell_t

(b) NK cells.

mainmatter/figures/chapter_03/validate_xCell_estimates.cell_t

(c) Plasma cells.

Figure 3.5: Correlation between standarised xCell scores and normalised FACS measurements for a similar immune subset, in the subset of individuals with FACS data.

indicating ComBat did an adequate job of removing batch- and platform-dependent global effects on expression (Fig. 3.6). Note that I did not leave this adjustment for PEER to perform, as ComBat estimates centering and scaling factors per gene to adjust for batch effects, whereas the use of PEER factors represent a mean-only adjustment Given the severity of the batch effect in this dataset, especially between platforms, mean-only adjustment may be insufficient[115].

remake this with only top k factors, and prune the possible covariates

3.2.7 eQTL mapping per timepoint

The performance of various software implementations of LMMs specialised for genetic association studies are highly comparable; the specific choice of implementation can usually be made on the basis of computational efficiency[97]. I map eQTLs within each timepoint using `LIMIX`[174], which implements univariate and multivariate LMMs with one or more random effects.

Imputed genotype probabilities were converted to alternate allele dosages using bcftools (1.7-1-ge07034a). Variants with sample AC < 15 within each timepoint were excluded.

add approximate MAFs, then cite hierarch paper

At each of 13570 genes, at all cis-variants within within $\pm 1Mb$ of the transcription start site (TSS), I fit the following model to map eQTL:

add note on treating x chrom variants with caution

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

where the eQTL effect size of interest is the slope of the genotype fixed effect β, the average additive effect of the alternate allele [8]; and $\mathbf{u}$ is a random effect with zero mean and covariance matrix proportional to the LOCO kinship matrix*.

lift proper vector notation from limix, then redo this with a timepoint subscript

add formulation of the 0-mean random effect to show exactly how the kinship matrix is used [175]

PEER factors are automatically weighted such that the variance of factors tends to zero as more factors are estimated, hence continuing to add more and more factors as covariates will not continue to improve eQTL detection power, and eventually the model degrees of freedom will be depleted. To optimise k, the number of factors to include as covariates†, Per-timepoint

note stacking of kinship for day -7 repeated measures

*For chromosome X variants, no LOCO matrix is available from LDAK, so the matrix for chromosome 1 is used.

†I avoid the commonly-performed two-stage approach of treating PEER residuals as expression phenotypes, as the degrees of freedom seen downstream will be incorrect, which

mainmatter/figures/chapter_03/peer_mega/peer.factor_cor_matrix.v2.pdf

Figure 3.6: Correlation of PEER factors to known factors and other possible covariates. Note that PEER factors are not constrained to be orthogonal, so correlations to known factors are expected.

eQTL mapping was performed in chromosome 1, iteratively increasing the number of factors until the number of eQTLs detected plateaus. I settled on a final choice of $k = 10$ factors for pre-vaccination, 5 factors for day 1, and 5 factors for day 7 (Fig. 3.7).

i leave the pcs in to guard against unusually differentiated markers, where random effect alone may not be enough [158]

3.2.8 Joint eQTL analysis across timepoints

Joint analysis was conducted with `mashr`[152], at 40197618 gene-variant pairs (mean of 2962 tests per gene) for which summary statistics from within timepoint mapping were available in all three timepoint conditions. The mashr model incorporates multiple canonical (the identity matrix etc.) and data-driven covariance matrices to represent patterns of effects across conditions (in this case, $3x3$ matrices). Data-driven covariance matrices are derived by dimension reduction of a strong subset of tests likely to have an effect in at least one condition. I took the most significant variant per gene per condition, which ensures strong condition-specific effects are included (Fig. 3.8), then further filtered to only nominally significant tests, resulting in a strong subset of 45962 tests.

recheck if did I do a SNPs only filter

The mash model was trained on a random subset of 200000 tests, using the Exchangeable Z-scores model[152]. The correlation of null tests between conditions, critical to account for due to the repeated measures structure of the data, was estimated using `mashr::estimate_null_correlation`. The fitted model was used as a prior to compute posterior effects and standard errors for all tests through shrinkage. A condition-specific Bayesian measure of significance local false sign rate (lfsr) is returned, which can be interpreted as the the probability given the data, that the declared sign of the effect is incorrect.

note this is critical, since we know a priori not independent due to eqtl sharing

move lfsr explanation prior to ashr in dge chapter

3.2.9 Defining shared and response eQTLs

Many of the tested variants for each gene will be in high LD. To unambiguously select a lead eQTL variant per gene, I selected the variant with the lowest lfsr in any condition, breaking ties by highest imputation INFO, highest MAF, shortest distance to the TSS, and genomic coordinate. Sharing was then evaluated for that gene-variant pair across all three conditions.

can have a substantial effect on estimates at this modest sample size.

`mainmatter/figures/chapter_03/count_eGenes.signif_eGenes_vs_PEER_n.dataset_mega`

Figure 3.7: Number of significant eGenes detected on chromosome 1 (hierarchical Bonferroni-Benjamini-Hochberg (BH) FDR < 0.05) as a function of the number of PEER factors included as covariates k.

Thresholding on the lfsr is not appropriate for determining sharing, as the difference between significant and non-significant effect estimates in two conditions is not necessarily significant[176, 177]. [152] provides a heuristic that two effects are shared by magnitude if they have the same sign, and are also within a factor of 2 of one another, but this does not consider the posterior standard error of the estimates. Between a pair of effects in two conditions, I compute a z-score for the difference in effects[150, 176]:

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

This strategy has been applied to call reQTLs by [53], assuming posterior pairwise covariance of effects is zero $\sigma^2(x, y)$. A Wald test p-value for the difference can be computed, as under the null hypothesis of zero difference, asymptotically $z \sim \mathcal{N}(0, 1)$. I use nominal p-value < 0.05 as a heuristic threshold (like the mashr recommended 2-fold threshold) to define reQTL effects that are strong, rather than a formal measure of significance. Effects are only compared if at least one of the two effects has lfsr < 0.05, to avoid sharing being driven by null effects.

not sure whether this is conservative or anti-conservative

mashr does not provide by default

mainmatter/figures/chapter_03/mash_mega/mashr.strong_subset_zval_heatmap.cisDist_1e6.sa

Figure 3.8: Clustering of within-timepoint Z scores in the strong mashr subset (random sample of 10000/45962 tests), confirming the presence of strong condition-specific effects.

3.2.10 Replication of eQTLs in a reference dataset

To validate the eQTL mapping approach, I estimate the replication of significant eQTLs in a large independent reference. Due to the lack of large sample size eQTL maps specific to PBMC, I use the GTEx v8 whole blood dataset as my reference dataset (n=670, 51.2% eGene rate). For lead variants called as significant in the HIRD dataset at a given lfsr threshold, I lookup the nominal p-value for that variant in GTEx (where the variant exists in both datasets). I applied `qvalue::qvalue_truncp` to estimate the proportion of those GTEx nominal p-values that are null (π_0), the compute a measure of replication $\pi_1 = 1 - \pi_0$.

The mega-analysis has comparable replication rate to RNA-seq-only analysis for shared eQTLs at moderately stringent lfsr thresholds up to 10^{-5} (Fig. 3.9). Past this, as the π_1 procedure assumes a well-behaved p-valuedistribution in $[0, 1]$, reliability declines due to the number of p-values being too small*, or the maximum p-value being too far from 1. The numbers of reQTLs were too low to assess replication using this method, and one might not expect them to replicate in a baseline dataset such as GTEx whole blood, especially for those reQTLs significant only at post-vaccination timepoints. As the mega-analysis has a higher eGene rate (50.8 % vs. 29.9 %) compared to the RNA-seq-only analysis, with similar replication, I assume this represents a power advantage from having larger a sample size, rather than technical effects from merging the expression data.

RNAseq does test about 7000 more genes though...

3.2.11 Genotype interactions with cell type abundance

be more specific: "moderator", not 'modify'

If the abundance of a particular cell type does truly modify the eQTL effect, then an interaction term between genotype and cell type abundance is required, otherwise the regression slope of the eQTL will represent an average across the abundance range for that cell type; one can not correct for this modification just by including the main effect for cell type abundance. Given the modest sample size, I use the two-step approach used by others[35, 40, 51, 53], where tests for interaction are only performed at a subset of tests, often the lead eQTL variant for each gene. The key to the two-stage approach is that if the estimates for the interaction effect are sufficiently independent from the estimates of the main effect from main-effect only models, the

point is, doesn't make sense to assume the genotype effect is the same at all levels of cell type abundance

*https://github.com/StoreyLab/qvalue/pull/6#commitcomment-26277751 mainmatter/figures/chapter_03/compute_pi1.pi1_by_thresholds.pdf

Figure 3.9: Effect of HIRD lfsr threshold on GTEx whole blood replication rate (π_1), number of p-values used to compute π_1, and maximum p-value among those p-values; for shared and reQTL called from the array-only, RNA-seq-only and mega-analysis pipelines. Shaded region for π_1 represents the 5th-95th percentile range of 1000 bootstraps.

type I error can be controlled based on the number of interactions that are actually tested, rather the number of interactions that could have been tested for[40, 178]. It is unclear whether this assumption holds, as the size of the main effect may contribute to power for detecting interaction effects. As the main purpose of the interaction analyses is scanning for cell type effects at detected reQTLs, I chose to test for interactions only at the lead eQTL variant for each gene with a significant main eQTL, then apply the BH false discovery rate (FDR), as used by others[40, 53].

Models in interactions between genotype and other predictors were fit using `lme4qtl`. The model specification identical to Equation 3.1, with the addition of three interaction terms between genotype and each xCell score. Significance is assessed using the likelihood-ratio test versus the nested model with no interaction terms.

add note here that although peer is correlated with xcell, interactions are only formed with xcell, so the interaction term can be interpreted per unit of genotype increase when xcell=0

3.2.12 TODO Statistical colocalisation

- if adding coloc analysis, add `https://github.com/jrs95/hyprcoloc` methods here - cannot deal with multiple causal within each group

this analysis is incomplete, and is one of the things I would suggest to round off this chapter

3.3 Results

3.3.1 Mapping reQTLs to Pandemix vaccination

Within each timepoint condition (day 0 pre-vaccination, day 1, and day 7), cis-eQTLs (±1Mb of the TSS) were mapped using `LIMIX`, then joint analysis of effects was done using `mashr` to obtain posterior effect size and standard errors. At lfsr < 0.05, 6887/13570 genes (50.8 %) were eGenes (genes with a significant eQTL) in at least one timepoint. To sidestep the issue of multiple tested variants per gene being in LD, the most significant eQTL variant across all timepoints was selected as the lead variant for each eGene, then reQTLs were defined by comparing the effect size of this lead eQTL between each pair of timepoints. Most eQTLs were shared across timepoints; 1154/6887 (16.8 %) eQTLs were classified as reQTLs between any pair of timepoints (nominal p difference < 0.05).

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

Fig. 3.10 illustrates the difference between calling sharing using a significance threshold versus difference in betas approach. For instance, day 0 was the timepoint with the largest number of eGenes, reflecting the larger sample size compared to other timepoints. Although there are 1427 eGenes significant at only day 0, there are only 646/1427 reQTL among them, as the effect size at day 0 does not differ significantly when compared to day 1 or day 7 for the remainder. The strongest eQTLs with the highest proportion of variance explained (PVE)* are shared between timepoints, highlighting the power advantage for mapping shared effects granted by joint analysis.

actually, i've found that my PVE approximation is basically rescaled abs(Z), so pve is a bit pointless if we already have z, and doesn't really help with comparability between genes with diff var/MAF

3.3.2 Characterising reQTLs post-vaccination

As detection power is greatest at day 0, I focus on eQTLs that are reQTLs between day 0 and either day 1 or day 7 post-vaccination, and are significant at the corresponding timepoint: 819 reQTL between day 1 and day 0, and 1002 reQTL between day 7 and day 0 (Fig. 3.11). Gene set enrichment analysis on the eGenes targets for these sets of reQTLs did not detect any significant enrichments (`gprofiler2`, g:SCS adjusted $p < 0.05$). Many of the reQTL that satisfy this criteria have opposite effects pre- and post-vaccination—as lfsr quantifies uncertainty in the sign of the effect, I do not compare the sign unless the reQTL is also significant at day 0. Shared eQTLs are enriched close to the TSS, whereas reQTLs are distributed across the cis- window.

requiring signif post-vaccination may not be correct, as it excludes many dampening effects

the lack of any positional enrichment makes me concerned for false positives? check with ASE?

The strongest reQTL at day 1 was for *ADCY3* (p difference $= 8.68 \times 10^{-6}$, BH FDR $= 0.118$), where the reQTL variant explained approximately 1.9 % of expression variation at day 0, increasing to 14.1 % at day 1 (Fig. 3.12). At day 7 the strongest reQTL was at *SH2D4A* (p difference $= 1.37 \times 10^{-6}$, BH FDR $= 0.0175$). Here, the reQTL variant explained similar amounts of expression variation at day 0 (8.2 %) and day 7 (9.0 %), with opposite directions of effect (Fig. 3.13). Both *ADCY3* and *SH2D4A* have moderately high percentile expression at all timepoints, and are not differentially expressed post-vaccination. Overall, compared to genes without reQTL, reQTLs were less likely be differentially expressed post-vaccination at day 1 (26.5 % for reQTL vs. 42.3 %, Fisher's test $p < 2.20 \times 10^{-16}$), and no significant difference was observed at day 7 (2.2 % for reQTL vs. 1.4 %, Fisher's test $p = 0.0509$). Only 5/68 (13.2 %) genes with reQTLs that explain more variation at day 1 were upregulated at day 1 vs. day 0; 5/226 (2.2 %) for day 7 vs. day 0.

expand this to plot 1, list top 5 damp, flip, amp at each timepoint

note anything in lit about any of the 30

reword not significant

double check denoms

convert to subfigures

*TODO: add to methods `https://journals.plos.org/plosone/article/file?type=supplementary&id=info:doi/10.1371/journal.pone.0120758.s001`

mainmatter/figures/chapter_03/compare_dge_eqtl.upset.pdf

Figure 3.10: Summary of eQTL mapping results at 13570 genes-lead eQTL pairs, with intersections based on significance (lfsr < 0.05). Counts of shared eQTLs and reQTLs; and distribution of INFO score, min MAF across timepoints, and max PVE across timepoints for those lead variants are shown above each intersection.

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

Figure 3.11: Z score for difference in effect vs. day 0, of lead eQTLs for all eGenes significant at either day 1 or day 7; versus distance of the lead SNP to the TSS. Direction of effect is aligned so that the beta at day 0 is positive. Points with positive z score are magnified effects post-vaccination, points with negative z scores are dampening and opposite sign effects.

mainmatter/figures/chapter_03/plot_dge_eqtl_genotypes.ENSG00000138031,rs916485_

Figure 3.12: *ADCY3*, strongest reQTL at day 1.

mainmatter/figures/chapter_03/plot_dge_eqtl_genotypes.ENSG00000104611,rs7841346

Figure 3.13: *SH2D4A*, strongest reQTL at day 7. Top: Array and RNA-seq expression before merging with ComBat for mega-analysis. Bottom: eQTL effects at each timepoint condition in the mega-analysis.

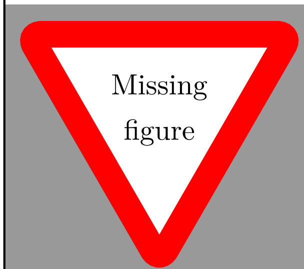
3.3.3 Genotype by cell type interaction effects

Given that many reQTLs are not explained by differential expression post-vaccination, the presence of cell type-specific eQTL effects was considered as an alternate explanation. As described in subsection 3.2.5, `xCell` enrichment scores were used to approximate abundance of seven PBMC cell types from the expression data. After pruning highly correlated cell types to avoid multicollinearity, standardised scores for monocytes, NK cells and plasma cells were tested for genotype interactions. Within-timepoint full eQTL models including the genotype main effect, the three cell type abundance main effects, and three cell type-genotype interaction terms, were fit using `lme4qtl`, then compared to a nested model excluding the three interaction terms.

siunitx permits uncertainties

Significant cell type interactions were detected at 16/1154 reQTLs (BH FDR < 0.05) in any timepoint, including *ADCY3* at day 1 ($\chi^2(3) = 26.3$, likelihood ratio test (LRT) BH FDR $= 9.54 \times 10^{-5}$). Although the genotype effect size was 0.256 (SE $= 0.0334$) in the nested model, the estimate in the full model was $-0.007\,22$ (0.0666); with the three cell type-genotype interaction term estimates being: monocyte=0.213 (0.0490), NK cells=$-0.009\,20$ (0.0447), and plasma cells=0.0162 (0.0663). The small magnitude of the genotype main effect in the full model vs. the nested model indicates the eQTL effect is driven largely by the monocyte score (or a cell type that is highly correlated with monocyte score, see Fig. 3.4). In the case where the monocyte score is zero (representing an average abundance across all samples, as scores are standardised), the effect of increasing genotype dosage on *ADCY3* expression is minimal.

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



expression vs monocyte xCell score, colored by genotype, to visually prove the point

3.3.4 TODO Genotype by platform interaction effects

- Perhaps using platform specific effects as a filter for reQTLs.

Need to consider Nikos' comment that there are too many (1069/13570 significant BH FDR) genotype-platform interactions to use mega-analysis. Consider filtering.

3.3.5 TODO Colocalisation of reQTLs with known *in vitro* condition-specific immune eQTLs

this analysis is incomplete, and is one of the things I would suggest to round off this chapter

- 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 reQTL by coloc with known immune QTL
- In a 500 Mb window around the lead *ADCY3* variant rs916485, `HyPrColoc` to colocalise with existing datasets and fine map.
- Day 1 HIRD colocs with BLUEPRINT and Fairfax monocytes (both stim and non stim), but not with Quach or Schmiedel monocytes (Fig. 3.14)
  - Biases from ethnicity-derived differences in LD?
  - Also, priors need tuning?
- `HyPrColoc` fine maps the signal to rs13407913 (credible set size=1, PP = 1), an intronic variant 45064 bp downstream of the TSS.

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

add obesity GWAS

3.4 Discussion

In the HIRD cohort, eQTL were detected for 50.8 % of genes in at least one timepoint, day 0, day 1, or day 7. Even in a joint mapping framework, defining reQTL by set significance thresholds, or change in the amount of expression variation explained, will miss classifying equal but opposite effect sizes. I account for the direction and magnitude of effect sizes, defining reQTL strength as the difference in effect size between timepoints. Most eQTL are shared between conditions and replicate well in GTEx whole blood; 16.8 % of lead eQTL for each eGene were reQTL that differed in effect size between timepoints.

compare sharing with mashr and ongen2017EstimatingCausalTissues

Multiple independent eQTLs are present for a large fraction of eGenes[179]. As the lead variant for reQTL assessment for each eGene was chosen based on significance across all conditions, I can not detect reQTL that are masked mainmatter/figures/chapter_03/perform_coloc.locusPlot.gene_ENSG00000138031.pdf

Figure 3.14: Multi-trait colocalisation of HIRD reQTL signal at ADCY3 (500 Kb window), with QTL studies from IHEC, BLUEPRINT, eQTL Catalogue, and GWAS Catalogue. Plots are colored by colocalised cluster. Black indicates non-colocalised datasets.

by a stronger shared eQTL at that gene. This is not expected to be uncommon, as the effective sample size for shared eQTLs is usually large due to borrowing of information across conditions. Secondary eQTL signals tend to be weaker, more distal to the TSS, more likely to be enriched in enhancers rather than promoters, and importantly, more context-specific[27, 180]. The proportion of genes with reQTL I detect based on only the lead signal likely represents a lower bound.

Given the larger global changes in expression vs. baseline at day 1 compared to day 7 described in chapter 2, the larger number of reQTLs detected at day 7 was unexpected (819 vs. 1002). Opposite sign effects among reQTL post-vaccination were common. Prevalance of opposite sign effects between pairs of conditions has been previously described in multi-tissue studies. In [181], the proportion of opposite sign effects as a percentage of all eGenes was 7.4 % (48 tissues); in HIRD, I find 39/6887 (0.6 %) at day 1, and 211/6887 (3.1 %) at day 7. In [43], the proportion of opposite sign effects as percentage of all reQTLs was 4.4 % (5 tissues); in HIRD, I find 39/819 (4.8 %) at day 1, and 211/1002 (21.1 %) at day 7. The enrichment of opposite sign effects in HIRD is also apparent at day 7. The strongest reQTL at day 7 is one such opposite sign effect; *SH2D4A* has constitutive expression in T cells, B cells, macrophages, and DCs, encoding a adapter protein involved in intracellular signal transduction*. An approach for validating these opposite sign reQTL using the existing HIRD RNA-seq data is allele-specific expression (ASE) (e.g. [182]), where one would expect true opposite sign reQTL effects would also be recapitulated as opposite directions of expression imbalance.

I'm not exactly sure why at the moment. Enrichment analyses so far have not turned up much. Up regulation of cell cycle TFs is a possibility.

The strongest reQTL detected at day 1 was *ADCY3*, a membrane-bound enzyme that catalyses the conversion of ATP to the second messenger cAMP[183]. Genome-wide association studies (GWAS) have identified *ADCY3* as a candidate gene for diseases such as obesity[183] and IBD[184]. *ADCY3* has been identified as a target for reQTLs in multiple studies involving stimulated blood immune cells: in PBMC 24h post-infection with rhinovirus[185], in whole blood *in vivo* day 1 after vaccination with seasonal TIV[49], and in whole blood after stimulation with M. leprae antigen for 26-32 h[52]. Given the diversity of stimulations and tissue types, the effect is likely a consequence of general immune activation, rather than a Pandemrix-specific response.

replace mcgovern2015GeneticsInflammatoryBowel with more recent

*https://doi.org/10.1111/j.1600-065X.2009.00829.x

Statistical colocalisation suggests that the day 1 reQTL signal identified here is likely to be a monocyte-specific effect—and independent to the IBD signal, which colocalises with T cell and macrophage datasets. The proportion of monocytes in the PBMC increase at day 1, supported by both FACS[86] measurements, and an increase in monocyte xCell score. Expression of *ADCY3* is not monocyte-specific, as despite the increase in monocyte proportion, no upregulation is observed at day 1. Colocalisation is also not restricted to stimulated monocytes, hence the signal could be hypothesised to result simply from the increased proportion of the bulk sample taken up by monocytes, rather than a upregulation-driven increase in detection power, or a vaccine-induced activation of the locus at day 1.

add lfsr.dge

need to consider: if this kind of thing is what bulk in vivo reQTL can find, they what is the additional value over FACS?

Changes in relative abundances for many cell types occur in the bulk PBMC samples after vaccination. I accounted for the effect of abundance on mean expression including xCell scores and PEER factors as fixed effects in the model, and also considered the effect of abundance on the genotype effect using interaction terms between xCell scores and genotype. Due to the modest sample size, and computational requirements for `lme4qtl`, I focused only whether reQTLs that have a detectable main effect may be driven by cell type interactions, testing only for interactions at significant lead reQTL, Compared to FACS measurements in a cohort subset, the xCell scores used above were only weakly correlated. Some discrepancy is expected, as the cell types as defined in the xCell signatures do not directly correspond to the combinations of surface markers used for FACS. The FACS gating strategy also meant that for some cell populations, the only available FACS measure was a proportion of the previously gated population, whereas xCell attempts to estimate scores that represent proportions of the whole mixture. The accuracy of the built-in signatures is lower when applied to the expression matrix for a stimulated state, likely because the enrichment-based method can not distinguish differential expression of signature genes due to stimulation from actual changes in cell abundance. Nevertheless, as assuming a single genotype where cell-type specific slopes are likely is inappropriate, so xCell scores were used as a best approximation. At 16/1154 reQTLs, the genotype effect was detected to interact with abundance of one or more of the tested cell types (or a correlated cell type). At the day 1 *ADCY3* reQTL, the genetic effect can be mainly attributed to the monocyte score-genotype interaction term, further supporting the hypothesis that it is monocyte-specific.

dge is coupled to reqtl, if you do an enrichment of dge+reqtl overlap genes, enrichment is driven by DGE signal harmonise terminology for 'opposite'

check "rs2223286 is associated with profound directional effects in the expression of SELL dependent upon genotype, with the minor C allele associated with increased expression of SELL in B-cells and reduced expression of SELL in monocytes "

note coloc doesn't distinguish pleiotropy from mediation?

A pressing question remains: what molecular mechanisms underlie the *ADCY3* reQTL, and indeed the remainder of the reQTLs? Power differences due to condition-specific expression are unlikely to explain a large proportion of reQTLs. As in [51, 53], the overlap between differentially expressed genes and genes with reQTL was poor, and reQTL were not more likely to be differentially expressed compared to genes without reQTL. One mechanism by which cis-eQTL affect expression is through their impact on transcription factor (TF) binding affinity to motifs in promoters and enhancers[186]. Immune cells, including monocytes, are regulated by cell type specific TFs[187]. Cell type specific expression of different TFs have been proposed as a model for explaining magnifying, dampening and opposite reQTL effects; for example, opposite effects can result from TFs regulating the same gene, that are activating in one cell type and suppressive in another[43]. There is evidence that TF activity is important for *in vivo* immune reQTL: [185] found rhinovirus reQTLs in PBMCs were enriched in ENCODE ChIP-seq peaks for the TFs *STAT1* and *STAT2*, and [51] found interferon and anti-IL6 drug reQTLs likely disrupt *ISRE* and *IRF4* binding motifs. Rather than condition-specific expression of the eGene, what may be condition-specific is the expression of TFs whose activity is affected by the reQTL*.

Finally, I address the prospect that common genetic variation may explain some variation in antibody response to Pandemrix. I have indirectly demonstrated genotype-dependent effects on expression response by identifying reQTLs with differing effect size between timepoints, but have not yet to determined resulting genotype-dependent differences in antibody phenotypes. Some of the identified reQTLs will undoubtedly affect genes whose expression or post-vaccination expression change correlates with antibody response, but correlation is not transitive[188], and a formal tests such as the causal inference test (CIT)[189] are required to distinguish mediation of genotype-antibody associations through gene expression from competing models. [49] realised this, but concluded that they had insufficient power with a greater sample size and comparable study design to HIRD. The

*A cursory scan of TF motifs disrupted by the location of the fine-mapped *ADCY3* reQTL intronic variant rs13407913 on `https://ccg.epfl.ch/snp2tfbs/snpviewer.php`, does indeed show several motifs (for NR2C2, HNF4A, HNF4G, NR2F1) where the PWM score is higher for the ALT allele, consistent with the direction of effect for the day 1 reQTL.

HIRD cohort is also too small for a direct GWAS of Pandemrix antibody response. A suitable approach for prioritising reQTL that contribute to the antibody response to Pandemrix will be to leverage external genetic associations to similar phenotypes, for example, colocalisation with existing GWAS summary statistics for antibody response to a similar type of adjuvanted, inactivated vaccine.

add 1 concluding line

Overall I feel like the chapter is too descriptive, and falls short of making biological insights into Pandemrix response. Any additional analyses would hope to address that.

Chapter 4

multiPANTS

4.1 Introduction

4.1.1 IBD

- IBD is a complex IMID of the GI tract.
  - Prevelance of IBD in the Western world is at least 0.5%, and rising `https://www.nature.com/articles/nrgastro.2015.150`.
  - Although often seen as a disease of the Western world, the disease is increasingly common in non-Western countries as they industrialise.
  - Pathogenesis defined by interaction of the host genetic, environmental (e.g. diet) and gut microbial factors `https://www.nature.com/articles/nrgastro.2015.186`
- It has two major forms, UC and CD.
- UC is distinguished by ...
  - CD is distinguished by ... [190]
- IBD is one of the most-well studied diseases by GWAS
  - Over TODO hits at TODO loci (dig up latest paper out of [191–193])
  - Genetic correlation between UC and CD is high
  - Hits unique to CD are TODO

4.1.2 Anti-TNF therapies for IBD

- anti-tnfs in use for IBD
  - also used in related conditions e.g. RA [194]
  - "Biologic therapy with anti-TNF agents." for IBD https://www.nature.com/articles/nrgastro.2015.135
    * Two big players: adalimumab and infliximab [195]
    * https://www.nice.org.uk/guidance/ta329
    * promote mucosal healing
    * Their mechanisms of action on the target pathway [196]
- failure of anti-tnfs is common (TODO%) https://journals.lww.com/ctg/Fulltext/2016/01000/Loss_of_Response_to_Anti_TNFs__Definition,.2.aspx
  - types of failure: primary non-response (PNR), non-remission, adverse events.
    * clinical predictors [197]
    * immunogenicity (not a failure, but mediates it) via anti-drug Abs
- although reported failure rates (single-measure) do not necessarily reflect that there is something inherent to an individual that causes it https://www.ncbi.nlm.nih.gov/pmc/articles/PMC524113/ senn2016MasteringVariationVariance
  - GWAS studies in RA https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6614444/
  - attempts in IBD [198]
  - furthermore we know immunogenicity has genetic determinant [199]
  - does not necessarily share the same genetic arch as disease risk
  - if heritable, and amenable to gwas, follow the same strat of gene prioritisation for failure phenotypes, to drug target prioritisation as outlined in ch1

make sure some statement of drug target prioritisation is in ch1

- Biologics are part of a whole treatment pyramid `https://www.nature.com/articles/nrgastro.2013.158`
    - biologics are the 2nd most intense therapies below surgical intervention
    - Step-up approach: undertreats patients
    - Step-down approach: exposes patients to risks from more aggressive therapies
    - Neither are ideal
- The promise of transcriptomic signatures for response prediction and stratifying patients to the right therapies
    - much like for sysvacc in ch2
    - starts with DGE R vs NR in biopsies
    - cause or effect?
    - conflict in existing studies e.g. "The difference in results of these two studies could not be more stark: one found that TREM1 was downregulated in anti-TNF responders, and the other found that TREM1 was downregulated in anti-TNF non-responders." `https://www.nature.com/articles/s41575-019-0228-5`
    - TREM1 signature [200]
    - prospective study of 54 active IBD patients (24CD, 30UC)

4.1.3 The PANTS cohort

- prospective, observational cohort study UK wide, with total enrolled n=1610 , 92.29pc EUR [197]
    - patients at least 6yo, with active luminal CD, and antitnf naive
        - active defined by CRP and faecal calprotectin
    - up to 12 mo of followup (until withdrawl due to remission or otherwise), possible 2y extension
    - 2 drugs: ada and inf
    - evaluates several aspects of antitnf response: PNR at week 14, non-remission at w54, adverse events

    * also immunogenicity (defined by antidrug ab levels)
  - PNR evaluated at w14 via algo, after PNR (24%), rarely helpful to keep dosing
  - immunogenicity (63pc adalimumab), (28.5pc infliximab)
    * use of immunomods had protective effect on time to immunogenicity
  - 8% have an adverse drug reaction that curtails treatment
  - clinical factors associated with response
    * low serum drug concentration in peripheral blood at w14 (ELISA) assoc with PNR and non-remission and immunogenicity, (in multivariate models, for both drugs)
    * optimal is conc above which there is no improvement
  - suggest Dose intensification
- in this cohort, immunogenicity has a genetic association [199]

4.1.4 chapter summary

- What is the hypothesis???
- Identifying signatures of PNR
  - replicate TREM1?
- Identifying reQTL
  - Why require a reQTL approach?
  - There are IBD specific reQTL (TODO is that relevant?) [201]
  - identify reQTL for use in e.g. mediation analysis of the genetic causes of non-response via expression

4.2 Methods

4.2.1 Overall strategy

4.2.2 Study design

Patients recruited to the Personalised Anti-TNF Therapy in Crohn's Disease (PANTS) study may attend up to 10 study visits, a detailed overview of which can be found in Kennedy *et al.* [197]. This chapter focuses on the four major study visits: week 0 (week −4 to 0), week 14 (week 10 to 20), week 30 (week 22 to 38), and week 54 (week 42 to 66). Time is measured relative to the day of the first drug dose. Ranges indicate the eligibility windows defined in Kennedy *et al.* [197]. Major visits were scheduled prior to drug doses (IFX infusion or ADA injection). At each major visit, peripheral whole blood samples were collected and preserved in Tempus™ Blood RNA Tubes For this chapter, samples from scheduled major visits that fall outside the windows were included. Samples from additional visits (e.g. scheduled due to loss of response (LOR) or early study exit) that fall into major visit windows were also included, as additional visits often replaced major visits for patients with PNR or LOR.

figure out how many doses happen at additional visits

4.2.3 Definition of primary non-response (PNR)

- PNR was defined as <...>
- "In the event of loss of response (LOR) an additional LOR visit will be scheduled to occur immediately prior to the next anti-TNF infusion / injection. Patients will remain in the study if they continue with the same anti-TNF drug, even if LOR or ADRs occur."

4.2.4 RNAseq data generation and quantification

4.2.5 RNAseq quality control

4.2.6 Differential gene expression

4.2.6.1 Covariate selection

In estimating the effect X→Y, of predictor X on response variable Y by regression, conditioning on a third variable Z can increase, decrease, or even reverse the effect estimate. The regression model does not statistically distinguish what causal role Z may play, but different types of third variable can be distinguished conceptually. In this section, I focus on identifying third variables that are covariates: where Z is mainmatter/figures/chapter_04/process_pheno.pheno_filtered_dge.Study_Day_vs_Vis

Figure 4.1

mainmatter/figures/chapter_04/process_pheno.pheno_filtered_dge.Visit_Label_upse

Figure 4.2

associated with Y and explains some variation in Y, and conditioning on Z increases the efficiency of estimating X $\rightarrow$Y. Here, the predictors in question are primary response status, drug, and study visit; and the response variable is gene expression.

Many phenotypes and technical variables are available as potential covariates in the PANTS cohort (Fig. 4.3). These include proportions of six common cell types in whole blood, estimated using the Houseman method (`minfi::estimateCellCounts` `https://academic.oup.com/bioinformatics/article/30/10/1363/267584`) from whole blood Illumina MethylationEPIC data also collected for the same patients and timepoints.

A variance components analysis was conducted to identify variables that explain large fractions of variation in expression using `variancePartition`[202], which fits a mixed regression model. Variables in Fig. 4.3 were included as predictors. Additional categorical variables were included for patient, RNA-sequencing (RNA-seq) plate, and library prep protocol version. An additional continuous variable consisting of random numbers drawn from the standard normal distribution was also included as a null. Granulocyte proportion estimates were dropped to relieve multicollinearity. Categorical variables were coded as random effects, and continuous variables as fixed effects. Surprisingly, Hoffman *et al.* [202] showed that variance proportion estimates are unbiased even when coding categorical variables with as few as two categories as random, as long as all model parameters are estimated jointly using maximum likelihood (ML) rather than restricted maximum likelihood (REML)*. This approach also avoids over-estimates of variance proportions that occur if categorical variables with many levels are treated as fixed.

the var explained by Gran will be redistributed among highly cor vars anyways

- Variables were ranked by median variance proportion across all genes (**??**). The variables that explained the most variance included patient, cell proportions and RNA-seq plate. Variables that did not explain more variation on average than the null could still have high maximum values, indicating their importance for specific genes only, such as genes with sex-specific expression.

*REML treats random effects as nuisance parameters and estimates fixed effects after first integrating out random effects).

mainmatter/figures/chapter_04/process_pheno.pheno_filtered_dge.ggcorrplot.pdf

Figure 4.3

  - choice: penalty is 1 df, so include some of these low median, high max variables as covariates.
  - so basically select all, except Gran, and ever immunomod

- If covariates are also associated with the predictor X, issues can arise depending on their causal role. In general, conditioning on a confounder (X ←Z →Y) reduces bias, conditioning on a collider (X →Z ←Y) induces bias, and conditioning on a mediator (X →Z →Y) changes the effect estimated by removing the indirect effect mediated by Z.
  - Do the selected covariates have potential roles as confounders, colliders or mediators?
  - cell counts are potential ... confounders/mediators ... depends ...

don't know if it matters if there are colliders among covariates, since we can't estimate any causal effects in DGE due to lack of a control group

because this is non-randomised, baseline differences do matter

4.2.6.2 Contrasts model

`dream` hoffman2018DreamPowerfulDifferential

8 groups equiv to 3 way interactions 12 contrasts

But for dream, REML is TRUE, so use fixed for small numbers of levels also, need fixedeffects for tested covariates

Dream uses lmerTest approximation Satterthwaite df `https://link.springer.com/article/10.3758/s13428-016-0809-y` and REML combo controls type 1 error for n>144 in lmerTest simulations

journals need p values

4.2.6.3 Spline model

simple time x responder interaction over time assumes linear change

treat time as categorical visits (like baseline/w14 analysis above), then f test all interactions not clean defintion of visits there are many intermediate visits that are not main 4

over all timepoints, are there diff trajectories for R and NR? 3 interaction terms Internal Knots set at w14 and w30, since drug admin here, so slope should be allowed to change until next admin TODO check What is a basis i.e. what is in the design matrix?

mainmatter/figures/chapter_04/dream.plotVarPart.pdf

Figure 4.4

cubic or linear between knots? Not sure if time is ok as continuous. Knots are approximate

4.2.7 GSE

camera is dev to use mod t, but in practice ranks are comparable between t and z.std, even though dream says otherwise

4.2.8 pred

Sparse partial least squares regression for simultaneous dimension reduction and variable selection

Sparse Partial Least Squares Classification for High Dimensional Data

4.2.9 Genotype data preprocessing

4.2.10 reQTL

ANCOVA vs repeated measures vs mixed model

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

4.3 Results

4.3.1 DGE

not much DGE R vs NR at baseline PDIA5, IGKV1-9, KCNN3 ADA only Top ADA genes are full of IG segments, not so for IFX Not sure which of the baseline diffs are relevant, not sure what informs the choice of ada vs ifx some DGE R vs NR at w14 downregulated in R: immune activation, TLR and inflammatory signalling e.g. CD177 lots of DGE baseline -> w14 more so for R than NR more so for IFX than ADA consistent with Abbvie Not much diff between drugs TODO: Add contrasts with drugs combined? i.e. if theres no interaction, drop it TODO: prettify GSE

306 signif Not sure how much intersect with w14 pairwise

4.4 Discussion

- can we expand the PANTS conclusions to IBD and other IMIDs?
- source of multiomics data 1000IBD cohort [203]

"Comparative analysis of differential gene expression tools for RNA sequencing time course data" Surprisingly, TC tools were outperformed by the classical pairwise comparison approach on short time series (<8 time points) in terms of overall performance and robustness to noise, mostly because of high number of false positives, with the exception of ImpulseDE2.

PNR definition its very complex kennedy2019PredictorsAntiTNFTreatment says once PNR, no point in continuing approx of remission, which has it's own def

Binary pheno Def not rubbish marks fc analysis And remission is rare in PNR in general

Utility of the other timepoints: mainly seems to be maintained

Try predict drug conc?

Chapter 5

Discussion

summary of chapters

Tie ch 2 to 3 using baseline predictors? A response eqtl is not always a response eqtl

Limitations, and the perfect study.

Prediction is not inference two chapters of mine have always treated the R as indep in DGE kinda strange: response is a organism pheno should be downstream of E for modelling convenicnece cant make causal claims anyway without a control, we cannot observe the counterfactual of what if an individual was a non-responder

correlates of protection

error in var models

sysvacc: need genetics to move beyond prediction

gene signatures, rise and fall expression as a biomarker

rise and fall paper: 15 years experience from cancer: list challenges to clinical implementation from Discussion

The design of more longitudinal cohorts in the future

avoid bulk: Even if cell type interaction/proxy gene approach Cannot distinguish between correlated cell types

more chance for in vivo but how to take advantage of it

systems immunology/vacc still needed: generate mol phenotypes, in the right context, but make sure to include genotypes

all the change score nonsense

gone are the days of GWAS marginals (just a screening approach) under time pressure, or convention throwing in covariates as complex disease genetics moves computation not limits correct stat must no longer be divorced from rigourous statistics statistical considerations from small scale studies

move away from dichotomania

predictive claims

rely of stability of 'responder' `https://discourse.datamethods.org/t/responder-analysis-loser-x-4/1262` senn2016MasteringVariationVariance senn2018StatisticalPitfallsPersonalized

and to Cost-effectiveness and clinical implementation if you can identify NRs, what are you going to do about it?

Extending the sample size

longitudinal studies are smallish e.g. IBD bioresource TWAS: predicted gene expression, then associate with phenotypes

will everything be gwas associated as n continues to increase sensitive to the smallest differences in case/control

equiavlence testing but what is really smallest effect size of interest?

Finer and finer context in the intro: gwas to function pipeline now, the future

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

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

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

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

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

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

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

Deep phenotyping

disease specific biobanks e.g. ibd bioresource/predicct

vaccines: the cellular response, beyond ab titires cao2016SystemsImmunologyAntibody

PheWAS[204] PheWAS has the advantage of identifying genetic variants with pleiotropic properties.

Translational directions

- 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
    - 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
          unification immunology and vaccine dev: deep phenotyping, small cohorts achieved -> larger cohorts human genetics and gwas: large cohorts achieved -> deeper phenotyping
          MOFA multiomics Pqtls more accurate?
          already lots of expression data
          combining systems immunology studies of genetic arch of immune parameters e.g. dejager2015ImmVarProjectInsights, zalocusky201810000Immunomes giving gmore intermediate phenotypes layering of evidence (triangulation)

with GWAS of immune phenotypes

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. The ENCODE Project Consortium. An Integrated Encyclopedia of DNA Elements in the Human Genome. *Nature* **489,** 57–74. doi:`10.1038/nature11247` (2012).
2. 1000 Genomes Project Consortium *et al.* A Global Reference for Human Genetic Variation. *Nature* **526,** 68–74. doi:`10.1038/nature15393` (2015).
3. The International SNP Map Working Group. A Map of Human Genome Sequence Variation Containing 1.42 Million Single Nucleotide Polymorphisms. *Nature* **409,** 928–933. doi:`10.1038/35057149` (2001).
4. Slatkin, M. Linkage Disequilibrium — Understanding the Evolutionary Past and Mapping the Medical Future. *Nature Reviews Genetics* **9,** 477–485. doi:`10.1038/nrg2361` (2008).
5. Wall, J. D. & Pritchard, J. K. Haplotype Blocks and Linkage Disequilibrium in the Human Genome. *Nature Reviews Genetics* **4,** 587–597. doi:`10.1038/nrg1123` (2003).
6. The International HapMap Consortium. A Second Generation Human Haplotype Map of over 3.1 Million SNPs. *Nature* **449,** 851–861. doi:`10.1038/nature06258` (2007).
7. Karczewski, K. J. & Martin, A. R. Analytic and Translational Genetics. *Annual Review of Biomedical Data Science* **3.** doi:`10.1146/annurev-biodatasci-072018-021148` (2020).
8. Visscher, P. M. & Goddard, M. E. From R.A. Fisher's 1918 Paper to GWAS a Century Later. *Genetics* **211,** 1125–1130. doi:`10.1534/genetics.118.301594` (2019).
9. Gibson, G. Rare and Common Variants: Twenty Arguments. *Nature reviews. Genetics* **13,** 135–145. doi:`10.1038/nrg3118` (2011).

10. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169,** 1177–1186. doi:10.1016/j.cell.2017.05.038 (2017).

11. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five Years of GWAS Discovery. *The American Journal of Human Genetics* **90,** 7–24. doi:10.1016/j.ajhg.2011.11.029 (2012).

12. Hirschhorn, J. N., Lohmueller, K., Byrne, E. & Hirschhorn, K. A Comprehensive Review of Genetic Association Studies. *Genetics in Medicine* **4,** 45–61. doi:10.1097/00125817-200203000-00002 (2002).

13. Border, R. *et al.* No Support for Historical Candidate Gene or Candidate Gene-by-Interaction Hypotheses for Major Depression Across Multiple Large Samples. *American Journal of Psychiatry* **176,** 376–387. doi:10.1176/appi.ajp.2018.18070881 (2019).

14. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *The American Journal of Human Genetics* **101,** 5–22. doi:10.1016/j.ajhg.2017.06.005 (2017).

15. Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G. & Meyre, D. Benefits and Limitations of Genome-Wide Association Studies. *Nature Reviews Genetics.* doi:10.1038/s41576-019-0127-1 (2019).

16. The International HapMap Consortium. A Haplotype Map of the Human Genome. *Nature* **437,** 1299–1320. doi:10.1038/nature04226 (2005).

17. Barrett, J. C. & Cardon, L. R. Evaluating Coverage of Genome-Wide Association Studies. *Nature Genetics* **38,** 659–662. doi:10.1038/ng1801 (2006).

18. Das, S., Abecasis, G. R. & Browning, B. L. Genotype Imputation from Large Reference Panels. *Annual Review of Genomics and Human Genetics* **19,** 73–96. doi:10.1146/annurev-genom-083117-021602 (2018).

19. Pe'er, I., Yelensky, R., Altshuler, D. & Daly, M. J. Estimation of the Multiple Testing Burden for Genomewide Association Studies of Nearly All Common Variants. *Genetic Epidemiology* **32,** 381–385. doi:10.1002/gepi.20303 (2008).

20. Jannot, A.-S., Ehret, G. & Perneger, T. $P < 5 \times 10^{-8}$ Has Emerged as a Standard of Statistical Significance for Genome-Wide Association Studies. *Journal of Clinical Epidemiology* **68,** 460–465. doi:10.1016/j.jclinepi.2015.01.001 (2015).
21. Goeman, J. J. & Solari, A. Multiple Hypothesis Testing in Genomics. *Statistics in Medicine* **33,** 1946–1978. doi:10.1002/sim.6082 (2014).
22. Schaid, D. J., Chen, W. & Larson, N. B. From Genome-Wide Associations to Candidate Causal Variants by Statistical Fine-Mapping. *Nature Reviews Genetics* **19,** 491–504. doi:10.1038/s41576-018-0016-z (2018).
23. Gallagher, M. D. & Chen-Plotkin, A. S. The Post-GWAS Era: From Association to Function. *The American Journal of Human Genetics* **102,** 717–730. doi:10.1016/j.ajhg.2018.04.002 (2018).
24. Trynka, G. *et al.* Disentangling the Effects of Colocalizing Genomic Annotations to Functionally Prioritize Non-Coding Variants within Complex-Trait Loci. *The American Journal of Human Genetics* **97,** 139–152. doi:10.1016/j.ajhg.2015.05.016 (2015).
25. Gaffney, D. J. Global Properties and Functional Complexity of Human Gene Regulatory Variation. *PLoS Genetics* **9** (ed Abecasis, G. R.) e1003501. doi:10.1371/journal.pgen.1003501 (2013).
26. Albert, F. W. & Kruglyak, L. The Role of Regulatory Variation in Complex Traits and Disease. *Nature Reviews Genetics* **16,** 197–212. doi:10.1038/nrg3891 (2015).
27. Vandiedonck, C. Genetic Association of Molecular Traits: A Help to Identify Causative Variants in Complex Diseases. *Clinical Genetics.* doi:10.1111/cge.13187 (2017).
28. Wallace, C. Eliciting Priors and Relaxing the Single Causal Variant Assumption in Colocalisation Analyses. *PLOS Genetics* **16** (ed Epstein, M. P.) e1008720. doi:10.1371/journal.pgen.1008720 (2020).
29. Hemani, G., Bowden, J. & Davey Smith, G. Evaluating the Potential Role of Pleiotropy in Mendelian Randomization Studies. *Human Molecular Genetics* **27,** R195–R208. doi:10.1093/hmg/ddy163 (2018).

30. De Jager, P. L., Hacohen, N., Mathis, D., Regev, A., Stranger, B. E. & Benoist, C. ImmVar Project: Insights and Design Considerations for Future Studies of "Healthy" Immune Variation. *Seminars in Immunology* **27,** 51–57. doi:10.1016/j.smim.2015.03.003 (2015).

31. Nédélec, Y. *et al.* Genetic Ancestry and Natural Selection Drive Population Differences in Immune Responses to Pathogens. *Cell* **167,** 657–669.e21. doi:10.1016/j.cell.2016.09.025 (2016).

32. Quach, H. & Quintana-Murci, L. Living in an Adaptive World: Genomic Dissection of the Genus Homo and Its Immune Response. *Journal of Experimental Medicine* **214,** 877–894. doi:10.1084/jem.20161942 (2017).

33. Nica, A. C. *et al.* The Architecture of Gene Regulatory Variation across Multiple Human Tissues: The MuTHER Study. *PLoS Genetics* **7** (ed Barsh, G.) e1002003. doi:10.1371/journal.pgen.1002003 (2011).

34. Aguet, F. *et al.* Genetic Effects on Gene Expression across Human Tissues. *Nature* **550,** 204–213. doi:10.1038/nature24277 (2017).

35. Westra, H.-J. *et al.* Cell Specific eQTL Analysis without Sorting Cells. *PLOS Genetics* **11** (ed Pastinen, T.) e1005223. doi:10.1371/journal.pgen.1005223 (2015).

36. Zhernakova, D. V. *et al.* Identification of Context-Dependent Expression Quantitative Trait Loci in Whole Blood. *Nature Genetics* **49,** 139–145. doi:10.1038/ng.3737 (2017).

37. Glastonbury, C. A., Couto Alves, A., El-Sayed Moustafa, J. S. & Small, K. S. Cell-Type Heterogeneity in Adipose Tissue Is Associated with Complex Traits and Reveals Disease-Relevant Cell-Specific eQTLs. *The American Journal of Human Genetics* **104,** 1013–1024. doi:10.1016/j.ajhg.2019.03.025 (2019).

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

39. Dimas, A. S. *et al.* Common Regulatory Variation Impacts Gene Expression in a Cell Type-Dependent Manner. *Science* **325,** 1246–1250. doi:10.1126/science.1174148 (2009).

40. Peters, J. E. *et al.* Insight into Genotype-Phenotype Associations through eQTL Mapping in Multiple Cell Types in Health and Immune-Mediated Disease. *PLOS Genetics* **12** (ed Plagnol, V.) e1005908. doi:10.1371/journal.pgen.1005908 (2016).

41. Chen, L. *et al.* Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. *Cell* **167,** 1398–1414.e24. doi:10.1016/j.cell.2016.10.026 (2016).

42. Ackermann, M., Sikora-Wohlfeld, W. & Beyer, A. Impact of Natural Genetic Variation on Gene Expression Dynamics. *PLoS Genetics* **9** (ed Wells, C. A.) e1003514. doi:10.1371/journal.pgen.1003514 (2013).

43. Fu, J. *et al.* Unraveling the Regulatory Mechanisms Underlying Tissue-Dependent Genetic Variation of Gene Expression. *PLoS Genetics* **8** (ed Gibson, G.) e1002431. doi:10.1371/journal.pgen.1002431 (2012).

44. Rotival, M. Characterising the Genetic Basis of Immune Response Variation to Identify Causal Mechanisms Underlying Disease Susceptibility. *HLA* **94,** 275–284. doi:10.1111/tan.13598 (2019).

45. Huang, Q. *The Genetics of Gene Expression: From Simulations to the Early-Life Origins of Immune Diseases* (2019).

46. 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 (2012).

47. 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 (2014).

48. Alasoo, K., Rodrigues, J., Danesh, J., Freitag, D. F., Paul, D. S. & Gaffney, D. J. Genetic Effects on Promoter Usage Are Highly Context-Specific and Contribute to Complex Traits. *eLife* **8.** doi:10.7554/eLife.41673 (2019).

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

50. Lareau, C. A., White, B. C., Oberg, A. L., Kennedy, R. B., Poland, G. A. & McKinney, B. A. An Interaction Quantitative Trait Loci Tool Implicates Epistatic Functional Variants in an Apoptosis Pathway in Smallpox Vaccine eQTL Data. *Genes & Immunity* **17,** 244–250. doi:10.1038/gene.2016.15 (2016).

51. Davenport, E. E. *et al.* Discovering in Vivo Cytokine-eQTL Interactions from a Lupus Clinical Trial. *Genome Biology* **19.** doi:10.1186/s13059-018-1560-8 (2018).

52. Manry, J. *et al.* Deciphering the Genetic Control of Gene Expression Following Mycobacterium Leprae Antigen Stimulation. *PLOS Genetics* **13** (ed Sirugo, G.) e1006952. doi:10.1371/journal.pgen.1006952 (2017).

53. Kim-Hellmuth, S. *et al.* Genetic Regulatory Effects Modified by Immune Activation Contribute to Autoimmune Disease Associations. *Nature Communications* **8.** doi:10.1038/s41467-017-00366-1 (2017).

54. Brodin, P. *et al.* Variation in the Human Immune System Is Largely Driven by Non-Heritable Influences. *Cell* **160,** 37–47. doi:10.1016/j.cell.2014.12.020 (2015).

55. Liston, A., Carr, E. J. & Linterman, M. A. Shaping Variation in the Human Immune System. *Trends in Immunology* **37,** 637–646. doi:10.1016/j.it.2016.08.002 (2016).

56. Brodin, P. & Davis, M. M. Human Immune System Variation. *Nature Reviews Immunology* **17,** 21–29. doi:10.1038/nri.2016.125 (2017).

57. Patin, E. *et al.* Natural Variation in the Parameters of Innate Immune Cells Is Preferentially Driven by Genetic Factors. *Nature Immunology.* doi:10.1038/s41590-018-0049-7 (2018).

58. Liston, A. & Goris, A. The Origins of Diversity in Human Immunity. *Nature Immunology* **19,** 209–210. doi:10.1038/s41590-018-0047-9 (2018).

59. Lakshmikanth, T. *et al.* Human Immune System Variation during One Year. *bioRxiv.* doi:10.1101/2020.01.22.915025 (2020).

60. Tsang, J. S. Utilizing Population Variation, Vaccination, and Systems Biology to Study Human Immunology. *Trends in Immunology* **36,** 479–493. doi:10.1016/j.it.2015.06.005 (2015).

61. Villani, A.-C., Sarkizova, S. & Hacohen, N. Systems Immunology: Learning the Rules of the Immune System. *Annual Review of Immunology* **36,** 813–842. doi:10.1146/annurev-immunol-042617-053035 (2018).

62. Greenwood, B. The Contribution of Vaccination to Global Health: Past, Present and Future. *Philosophical Transactions of the Royal Society B: Biological Sciences* **369,** 20130433. doi:10.1098/rstb.2013.0433 (2014).

63. Linnik, J. E. & Egli, A. Impact of Host Genetic Polymorphisms on Vaccine Induced Antibody Response. *Human Vaccines & Immunotherapeutics* **12,** 907–915. doi:10.1080/21645515.2015.1119345 (2016).

64. O'Connor, D. & Pollard, A. J. Characterizing Vaccine Responses Using Host Genomic and Transcriptomic Analysis. *Clinical Infectious Diseases* **57,** 860–869. doi:10.1093/cid/cit373 (2013).

65. Mooney, M., McWeeney, S. & Sékaly, R.-P. Systems Immunogenetics of Vaccines. *Seminars in Immunology* **25,** 124–129. doi:10.1016/j.smim.2013.06.003 (2013).

66. Mentzer, A. J., O'Connor, D., Pollard, A. J. & Hill, A. V. S. Searching for the Human Genetic Factors Standing in the Way of Universally Effective Vaccines. *Philosophical Transactions of the Royal Society B: Biological Sciences* **370,** 20140341–20140341. doi:10.1098/rstb.2014.0341 (2015).

67. Scepanovic, P. *et al.* Human Genetic Variants and Age Are the Strongest Predictors of Humoral Immune Responses to Common Pathogens and Vaccines. *Genome Medicine* **10.** doi:10.1186/s13073-018-0568-8 (2018).

68. Dhakal, S. & Klein, S. L. Host Factors Impact Vaccine Efficacy: Implications for Seasonal and Universal Influenza Vaccine Programs. *Journal of Virology* **93** (ed Coyne, C. B.) doi:10.1128/JVI.00797-19 (2019).

69. Krammer, F. *et al.* Influenza. *Nature Reviews Disease Primers* **4.** doi:10.1038/s41572-018-0002-y (2018).

70. Houser, K. & Subbarao, K. Influenza Vaccines: Challenges and Solutions. *Cell Host & Microbe* **17,** 295–300. doi:10.1016/j.chom.2015.02.012 (2015).

71. Sautto, G. A., Kirchenbaum, G. A. & Ross, T. M. Towards a Universal Influenza Vaccine: Different Approaches for One Goal. *Virology Journal* **15.** doi:10.1186/s12985-017-0918-y (2018).

72. Broadbent, A. J. & Subbarao, K. Influenza Virus Vaccines: Lessons from the 2009 H1N1 Pandemic. *Current Opinion in Virology* **1,** 254–262. doi:10.1016/j.coviro.2011.08.002 (2011).

73. Klimov, A. *et al.* in *Influenza Virus* (eds Kawaoka, Y. & Neumann, G.) 25–51 (Humana Press, Totowa, NJ, 2012). doi:10.1007/978-1-61779-621-0_3.

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

75. Cox, R. Correlates of Protection to Influenza Virus, Where Do We Go from Here? *Human Vaccines & Immunotherapeutics* **9,** 405–408. doi:10.4161/hv.22908 (2013).

76. Pulendran, B. Systems Vaccinology: Probing Humanity's Diverse Immune Systems with Vaccines. *Proceedings of the National Academy of Sciences* **111,** 12300–12306. doi:10.1073/pnas.1400476111 (2014).

77. Hagan, T., Nakaya, H. I., Subramaniam, S. & Pulendran, B. Systems Vaccinology: Enabling Rational Vaccine Design with Systems Biological Approaches. *Vaccine* **33,** 5294–5301. doi:10.1016/j.vaccine.2015.03.072 (2015).

78. Raeven, R. H. M., van Riet, E., Meiring, H. D., Metz, B. & Kersten, G. F. A. Systems Vaccinology and Big Data in the Vaccine Development Chain. *Immunology* **156,** 33–46. doi:10.1111/imm.13012 (2019).

79. Zhu, W. *et al.* A Whole Genome Transcriptional Analysis of the Early Immune Response Induced by Live Attenuated and Inactivated Influenza Vaccines in Young Children. *Vaccine* **28,** 2865–2876. doi:10.1016/j.vaccine.2010.01.060 (2010).

80. 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 (2011).

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

82. Tan, Y., Tamayo, P., Nakaya, H., Pulendran, B., Mesirov, J. P. & Haining, W. N. Gene Signatures Related to B-Cell Proliferation Predict Influenza Vaccine-Induced Antibody Response. *European Journal of Immunology* **44,** 285–295. doi:10.1002/eji.201343657 (2014).

83. Nakaya, H. I., Li, S. & Pulendran, B. Systems Vaccinology: Learning to Compute the Behavior of Vaccine Induced Immunity. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* **4,** 193–205. doi:10.1002/wsbm.163 (2012).

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

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

86. 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 (2016).

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

88. Tsang, J. S. *et al.* Global Analyses of Human Immune Variation Reveal Baseline Predictors of Postvaccination Responses. *Cell* **157,** 499–513. doi:`10.1016/j.cell.2014.03.031` (2014).

89. Nakaya, H. I. *et al.* Systems Analysis of Immunity to Influenza Vaccination across Multiple Years and in Diverse Populations Reveals Shared Molecular Signatures. *Immunity* **43,** 1186–1198. doi:`10.1016/j.immuni.2015.11.012` (2015).

90. HIPC-CHI Signatures Project Team & HIPC-I Consortium. Multicohort Analysis Reveals Baseline Transcriptional Predictors of Influenza Vaccination Responses. *Science Immunology* **2,** eaal4656. doi:`10.1126/sciimmunol.aal4656` (2017).

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

92. Senn, S. Dichotomania: An Obsessive Compulsive Disorder That Is Badly Affecting the Quality of Analysis of Pharmaceutical Trials, 14 (2005).

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

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

95. Clifton, L. & Clifton, D. A. The Correlation between Baseline Score and Post-Intervention Score, and Its Implications for Statistical Analysis. *Trials* **20.** doi:`10.1186/s13063-018-3108-3` (2019).

96. 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` (2006).

97. 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 (2014).

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

99. The International HapMap 3 Consortium. Integrating Common and Rare Genetic Variation in Diverse Human Populations. *Nature* **467,** 52–58. doi:10.1038/nature09298 (2010).

100. 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 (2015).

101. 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 (2016).

102. 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 (2017).

103. 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 (2014).

104. 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 (2016).

105. 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 (2018).

106. 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).

107. 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 (2016).

108. 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 2002).

109. 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).

110. 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 (2006).

111. 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 (2015).

112. 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 (2007).

113. 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 (2011).

114. 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 (2018).

115. 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` (2018).

116. 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 2015).

117. 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` (2018).

118. 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` (2010).

119. 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).

120. 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` (2015).

121. 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` (2013).

122. 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).

123. 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` (2010).

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

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

126. 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` (2015).

127. 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` (2016).

128. Chung, Y., Rabe-Hesketh, S., Dorie, V., Gelman, A. & Liu, J. A Nondegenerate Penalized Likelihood Estimator for Variance Parameters in Multilevel Models. *Psychometrika* **78,** 685–709. doi:`10.1007/s11336-013-9328-2` (2013).

129. 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` (2017).

130. 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` (2019).

131. 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` (2006).

132. 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` (2011).

133. 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. *Statistics in Medicine* **34,** 984–998. doi:`10.1002/sim.6381` (2015).

134. 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 (1996).

135. 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 (2019).

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

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

138. Li, S. *et al.* Molecular Signatures of Antibody Responses Derived from a Systems Biology Study of Five Human Vaccines. *Nature Immunology* **15,** 195–204. doi:10.1038/ni.2789 (2013).

139. Bin, L., Li, X., Feng, J., Richers, B. & Leung, D. Y. M. Ankyrin Repeat Domain 22 Mediates Host Defense Against Viral Infection Through STING Signaling Pathway. *The Journal of Immunology* **196,** 201.4 LP –201.4 (1 Supplement 2016).

140. Schneider, W. M., Chevillotte, M. D. & Rice, C. M. Interferon-Stimulated Genes: A Complex Web of Host Defenses. *Annual Review of Immunology* **32,** 513–545. doi:10.1146/annurev-immunol-032713-120231 (2014).

141. Nakaya, H. I. *et al.* Systems Biology of Immunity to MF59-Adjuvanted versus Nonadjuvanted Trivalent Seasonal Influenza Vaccines in Early Childhood. *Proceedings of the National Academy of Sciences* **113,** 1853–1858. doi:10.1073/pnas.1519690113 (2016).

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

143. Nauta, J. J., Beyer, W. E. & Osterhaus, A. D. On the Relationship between Mean Antibody Level, Seroprotection and Clinical Protection from Influenza. *Biologicals* **37,** 216–221. doi:`10.1016/j.biologicals.2009.02.002` (2009).

144. Poland, G. A., Ovsyannikova, I. G. & Jacobson, R. M. Immunogenetics of Seasonal Influenza Vaccine Response. *Vaccine* **26,** D35–D40. doi:`10.1016/j.vaccine.2008.07.065` (2008).

145. Avnir, Y. *et al.* IGHV1-69 Polymorphism Modulates Anti-Influenza Antibody Repertoires, Correlates with IGHV Utilization Shifts and Varies by Ethnicity. *Scientific Reports* **6,** 20842. doi:`10.1038/srep20842` (2016).

146. Moss, A. J. *et al.* Correlation between Human Leukocyte Antigen Class II Alleles and HAI Titers Detected Post-Influenza Vaccination. *PLoS ONE* **8** (ed Sambhara, S.) e71376. doi:`10.1371/journal.pone.0071376` (2013).

147. Maranville, J. C. *et al.* Interactions between Glucocorticoid Treatment and Cis-Regulatory Polymorphisms Contribute to Cellular Response Phenotypes. *PLoS Genetics* **7** (ed Gibson, G.) e1002162. doi:`10.1371/journal.pgen.1002162` (2011).

148. Shpak, M. *et al.* An eQTL Analysis of the Human Glioblastoma Multiforme Genome. *Genomics* **103,** 252–263. doi:`10.1016/j.ygeno.2014.02.005` (2014).

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

150. Clogg, C. C., Petkova, E. & Haritou, A. Statistical Methods for Comparing Regression Coefficients Between Models. *The American Journal of Sociology* **100,** 1261–1293 (1995).

151. Flutre, T., Wen, X., Pritchard, J. & Stephens, M. A Statistical Framework for Joint eQTL Analysis in Multiple Tissues. *PLOS Genet* **9,** e1003486. doi:`10.1371/journal.pgen.1003486` (2013).

152. Urbut, S. M., Wang, G., Carbonetto, P. & Stephens, M. Flexible Statistical Methods for Estimating and Testing Effects in Genomic Studies with Multiple Conditions. *Nature Genetics.* doi:`10.1038/s41588-018-0268-8` (2018).

153. Li, G., Jima, D., Wright, F. A. & Nobel, A. B. HT-eQTL: Integrative Expression Quantitative Trait Loci Analysis in a Large Number of Human Tissues. *BMC Bioinformatics* **19.** doi:`10.1186/s12859-018-2088-3` (2018).

154. Stephens, M. A Unified Framework for Association Analysis with Multiple Related Phenotypes. *PLoS ONE* **8** (ed Emmert-Streib, F.) e65245. doi:`10.1371/journal.pone.0065245` (2013).

155. Kim, S. *et al.* Characterizing the Genetic Basis of Innate Immune Response in TLR4-Activated Human Monocytes. *Nature Communications* **5.** doi:`10.1038/ncomms6236` (2014).

156. Sul, J. H., Han, B., Ye, C., Choi, T. & Eskin, E. Effectively Identifying eQTLs from Multiple Tissues by Combining Mixed Model and Meta-Analytic Approaches. *PLoS Genetics* **9** (ed Schork, N. J.) e1003491. doi:`10.1371/journal.pgen.1003491` (2013).

157. Duong, D. *et al.* Applying Meta-Analysis to Genotype-Tissue Expression Data from Multiple Tissues to Identify eQTLs and Increase the Number of eGenes. *Bioinformatics* **33,** i67–i74. doi:`10.1093/bioinformatics/btx227` (2017).

158. Price, A. L., Zaitlen, N. A., Reich, D. & Patterson, N. New Approaches to Population Stratification in Genome-Wide Association Studies. *Nature Reviews Genetics* **11,** 459–463. doi:`10.1038/nrg2813` (2010).

159. Golan, D., Rosset, S. & Lin, D.-Y. in Borgan, Ø., Breslow, N. E., Chatterjee, N., Gail, M. H., Scott, A. & Wild, C. J. *Handbook of Statistical Methods for Case-Control Studies* (eds Borgan, Ø., Breslow, N., Chatterjee, N., Gail, M. H., Scott, A. & Wild, C. J.) 1st ed., 495–514 (Chapman and Hall/CRC, 2018). doi:`10.1201/9781315154084-27`.

160. Astle, W. & Balding, D. J. Population Structure and Cryptic Relatedness in Genetic Association Studies. *Statistical Science* **24,** 451–471. doi:`10.1214/09-STS307` (2009).

161. Listgarten, J., Lippert, C., Kadie, C. M., Davidson, R. I., Eskin, E. & Heckerman, D. Improved Linear Mixed Models for Genome-Wide Association Studies. *Nature Methods* **9,** 525–526. doi:`10.1038/nmeth.2037` (2012).

162. Lippert, C., Listgarten, J., Liu, Y., Kadie, C. M., Davidson, R. I. & Heckerman, D. FaST Linear Mixed Models for Genome-Wide Association Studies. *Nature Methods* **8,** 833–835. doi:10.1038/nmeth.1681 (2011).

163. Speed, D., Hemani, G., Johnson, M. R. & Balding, D. J. Improved Heritability Estimation from Genome-Wide SNPs. *The American Journal of Human Genetics* **91,** 1011–1021. doi:10.1016/j.ajhg.2012.10.010 (2012).

164. Aguet, F. *et al.* The GTEx Consortium Atlas of Genetic Regulatory Effects across Human Tissues. *bioRxiv.* doi:10.1101/787903 (2019).

165. Beasley, T. M., Erickson, S. & Allison, D. B. Rank-Based Inverse Normal Transformations Are Increasingly Used, But Are They Merited? *Behavior Genetics* **39,** 580–595. doi:10.1007/s10519-009-9281-0 (2009).

166. Võsa, U. *et al.* Unraveling the Polygenic Architecture of Complex Traits Using Blood eQTL Meta-Analysis. *bioRxiv.* doi:10.1101/447367 (2018).

167. Qi, T. *et al.* Identifying Gene Targets for Brain-Related Traits Using Transcriptomic and Methylomic Data from Blood. *Nature Communications* **9.** doi:10.1038/s41467-018-04558-1 (2018).

168. Aran, D., Hu, Z. & Butte, A. J. xCell: Digitally Portraying the Tissue Cellular Heterogeneity Landscape. *Genome Biology* **18.** doi:10.1186/s13059-017-1349-1 (2017).

169. Kleiveland, C. R. in *The Impact of Food Bioactives on Health* (eds Verhoeckx, K. *et al.*) 161–167 (Springer International Publishing, Cham, 2015). doi:10.1007/978-3-319-16104-4_15.

170. Van der Wijst, M. G. P. *et al.* Single-Cell RNA Sequencing Identifies Celltype-Specific Cis-eQTLs and Co-Expression QTLs. *Nature Genetics* **50,** 493–497. doi:10.1038/s41588-018-0089-9 (2018).

171. Kanyongo, G. Y. The Influence of Reliability on Four Rules for Determining the Number of Components to Retain. *Journal of Modern Applied Statistical Methods* **5,** 332–343. doi:10.22237/jmasm/1162353960 (2005).

172. Astle, W. J. *et al.* The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell* **167,** 1415–1429.e19. doi:10.1016/j.cell.2016.10.042 (2016).

173. Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using Probabilistic Estimation of Expression Residuals (PEER) to Obtain Increased Power and Interpretability of Gene Expression Analyses. *Nature protocols* **7,** 500–507. doi:10.1038/nprot.2011.457 (2012).

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

175. Sul, J. H., Martin, L. S. & Eskin, E. Population Structure in Genetic Studies: Confounding Factors and Mixed Models. *PLOS Genetics* **14** (ed Barsh, G. S.) e1007309. doi:10.1371/journal.pgen.1007309 (2018).

176. Schenker, N. & Gentleman, J. F. On Judging the Significance of Differences by Examining the Overlap Between Confidence Intervals. *The American Statistician* **55,** 182–186 (2001).

177. Gelman, A. & Stern, H. The Difference Between "Significant" and "Not Significant" Is Not Itself Statistically Significant. *The American Statistician* **60,** 328–331. doi:10.1198/000313006X152649 (2006).

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

179. Zeng, B. *et al.* Comprehensive Multiple eQTL Detection and Its Application to GWAS Interpretation. *Genetics* **212,** 905–918. doi:10.1534/genetics.119.302091 (2019).

180. Dobbyn, A. *et al.* Landscape of Conditional eQTL in Dorsolateral Prefrontal Cortex and Co-Localization with Schizophrenia GWAS. *The American Journal of Human Genetics* **102,** 1169–1184. doi:10.1016/j.ajhg.2018.04.011 (2018).

181. Mizuno, A. & Okada, Y. Biological Characterization of Expression Quantitative Trait Loci (eQTLs) Showing Tissue-Specific Opposite Directional Effects. *European Journal of Human Genetics* **27,** 1745–1756. doi:10.1038/s41431-019-0468-4 (2019).

182. Kumasaka, N., Knights, A. J. & Gaffney, D. J. Fine-Mapping Cellular QTLs with RASQUAL and ATAC-Seq. *Nature Genetics* **48,** 206–213. doi:10.1038/ng.3467 (2016).

183. Wu, L., Shen, C., Seed Ahmed, M., Östenson, C.-G. & Gu, H. F. Adenylate Cyclase 3: A New Target for Anti-Obesity Drug Development: ADCY3 and Anti-Obesity. *Obesity Reviews* **17,** 907–914. doi:10.1111/obr.12430 (2016).

184. McGovern, D. P., Kugathasan, S. & Cho, J. H. Genetics of Inflammatory Bowel Diseases. *Gastroenterology* **149,** 1163–1176.e2. doi:10.1053/j.gastro.2015.08.001 (2015).

185. Çalışkan, M., Baker, S. W., Gilad, Y. & Ober, C. Host Genetic Variation Influences Gene Expression Response to Rhinovirus Infection. *PLOS Genetics* **11** (ed Gibson, G.) e1005111. doi:10.1371/journal.pgen.1005111 (2015).

186. Pai, A. A., Pritchard, J. K. & Gilad, Y. The Genetic and Mechanistic Basis for Variation in Gene Regulation. *PLoS Genetics* **11** (ed Lappalainen, T.) e1004857. doi:10.1371/journal.pgen.1004857 (2015).

187. Choudhury, M. & Ramsey, S. A. Identifying Cell Type-Specific Transcription Factors by Integrating ChIP-Seq and eQTL Data-Application to Monocyte Gene Regulation. *Gene Regulation and Systems Biology* **10,** GRSB.S40768. doi:10.4137/GRSB.S40768 (2016).

188. Langford, E., Schwertman, N. & Owens, M. Is the Property of Being Positively Correlated Transitive? *The American Statistician* **55,** 322–325 (2001).

189. Millstein, J., Zhang, B., Zhu, J. & Schadt, E. E. Disentangling Molecular Relationships with a Causal Inference Test. *BMC Genetics* **10.** doi:10.1186/1471-2156-10-23 (2009).

190. Roda, G. *et al.* Crohn's Disease. *Nature Reviews Disease Primers* **6.** doi:10.1038/s41572-020-0156-2 (2020).

191. De Lange, K. M. *et al.* Genome-Wide Association Study Implicates Immune Activation of Multiple Integrin Genes in Inflammatory Bowel Disease. *Nature Genetics* **49,** 256–261. doi:10.1038/ng.3760 (2017).

192. Huang, H. *et al.* Fine-Mapping Inflammatory Bowel Disease Loci to Single-Variant Resolution. *Nature* **547,** 173–178. doi:`10.1038/nature22969` (2017).

193. Luo, Y. *et al.* Exploring the Genetic Architecture of Inflammatory Bowel Disease by Whole-Genome Sequencing Identifies Association at ADCY7. *Nature Genetics* **49,** 186–192. doi:`10.1038/ng.3761` (2017).

194. Mulhearn, Barton & Viatte. Using the Immunophenotype to Predict Response to Biologic Drugs in Rheumatoid Arthritis. *Journal of Personalized Medicine* **9,** 46. doi:`10.3390/jpm9040046` (2019).

195. Adegbola, S. O., Sahnan, K., Warusavitarne, J., Hart, A. & Tozer, P. Anti-TNF Therapy in Crohn's Disease. *International Journal of Molecular Sciences* **19,** 2244. doi:`10.3390/ijms19082244` (2018).

196. Levin, A. D., Wildenberg, M. E. & van den Brink, G. R. Mechanism of Action of Anti-TNF Therapy in Inflammatory Bowel Disease. *Journal of Crohn's and Colitis* **10,** 989–997. doi:`10.1093/ecco-jcc/jjw053` (2016).

197. Kennedy, N. A. *et al.* Predictors of Anti-TNF Treatment Failure in Anti-TNF-Naive Patients with Active Luminal Crohn's Disease: A Prospective, Multicentre, Cohort Study. *The Lancet Gastroenterology & Hepatology* **4,** 341–353. doi:`10.1016/S2468-1253(19)30012-3` (2019).

198. Gaujoux, R. *et al.* Cell-Centred Meta-Analysis Reveals Baseline Predictors of Anti-TNFα Non-Response in Biopsy and Blood of Patients with IBD. *Gut* **68,** 604–614. doi:`10.1136/gutjnl-2017-315494` (2019).

199. Sazonovs, A. *et al.* HLA-DQA1*05 Carriage Associated With Development of Anti-Drug Antibodies to Infliximab and Adalimumab in Patients With Crohn's Disease. *Gastroenterology.* doi:`10.1053/j.gastro.2019.09.041` (2019).

200. Verstockt, B. *et al.* Low TREM1 Expression in Whole Blood Predicts Anti-TNF Response in Inflammatory Bowel Disease. *EBioMedicine* **40,** 733–742. doi:`10.1016/j.ebiom.2019.01.027` (2019).

201. Piasecka, B. *et al.* Distinctive Roles of Age, Sex, and Genetics in Shaping Transcriptional Variation of Human Immune Responses to Microbial Challenges. *Proceedings of the National Academy of Sciences* **115,** E488–E497. doi:10.1073/pnas.1714765115 (2018).

202. Hoffman, G. E. & Schadt, E. E. variancePartition: Interpreting Drivers of Variation in Complex Gene Expression Studies. *BMC Bioinformatics* **17.** doi:10.1186/s12859-016-1323-z (2016).

203. Imhann, F. *et al.* The 1000IBD Project: Multi-Omics Data of 1000 Inflammatory Bowel Disease Patients; Data Release 1. *BMC Gastroenterology* **19.** doi:10.1186/s12876-018-0917-5 (2019).

204. 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 (2017).

List of Abbreviations

AC allele count

ASE allele-specific expression

BH Benjamini-Hochberg

bp base pair

BTM blood transcription module

CIT causal inference test

CPM counts per million

DC dendritic cell

DGE differential gene expression

eQTL expression quantitative trait locus

FACS fluorescence-activated cell sorting

FC fold change

FDR false discovery rate

FWER family-wise error rate

GWAS genome-wide association study

HA haemagglutinin

HAI haemagglutination inhibition

HIRD Human Immune Response Dynamics

HLA human leukocyte antigen

INT inverse normal transformation

LAIV live attenuated influenza vaccine

LD linkage disequilibrium

lfsr local false sign rate

LMM linear mixed model

LOCO leave-one-chromosome-out

LOR loss of response

LRT likelihood ratio test

MAF minor allele frequency

MANOVA multivariate analysis of variance

ML maximum likelihood

MN microneutralisation

NA neuraminidase

NK natural killer

OVB omitted-variable bias

PANTS Personalised Anti-TNF Therapy in Crohn's Disease

PBMC peripheral blood mononuclear cell

PC principal component

PCA principal component analysis

PNR primary non-response

PVE proportion of variance explained

QTL quantitative trait locus

REML restricted maximum likelihood

reQTL response expression quantitative trait locus

RNA-seq RNA-sequencing

SD standard deviation

SNP single nucleotide polymorphism

TF transcription factor

TIV trivalent inactivated influenza vaccine

TMM trimmed mean of M-values

TRI titre response index

TSS transcription start site

spell-check

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

add automatic rounding to x decimal places using num and sisetup

collaboration note in italics at start of each chapter

fncychap

Todo list