<title>

Benjamin Yu Hang Bai

2020-09-17 22:21:11+01:00

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.

LD decay just takes a really really long time, but there are evo forces at work too that maintain LD

Genetic linkage is 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.

can i use published figures?

1.2 Genetic association studies for complex traits

1.2.1 Principles of genetic association

- Variation in human phenotypic traits arises from an interplay between genetics, environment and pure chance. 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 certain types of back-door path effects such as reverse causality, which 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

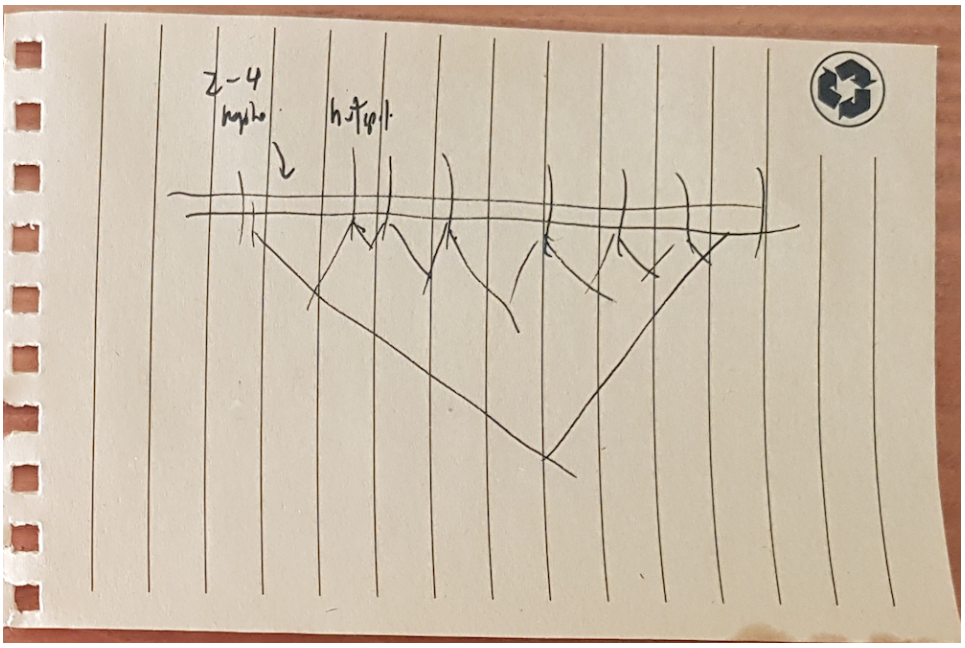

Figure 1.1: The genomic mosaic: block-like LD structure of the genome

frequencies, effect sizes, and interactions of trait-associated variants [8]. The number of genes defines a spectrum of traits from monogenic (where inheritance follows 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 the past 15 years

- 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 comparatively 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].
  - These new associations have ever smaller effect sizes https://www.pnas.org/content/early/2020/07/30/2005634117#F1. It is now appreciated that most heritable organism-level phenotypes are complex, and have remarkable polygenicity, with many hundreds or thousands of associated loci.
  - In general, the more organism level a phenotype, the more polygenic, but even molecular traits are very polygenic

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

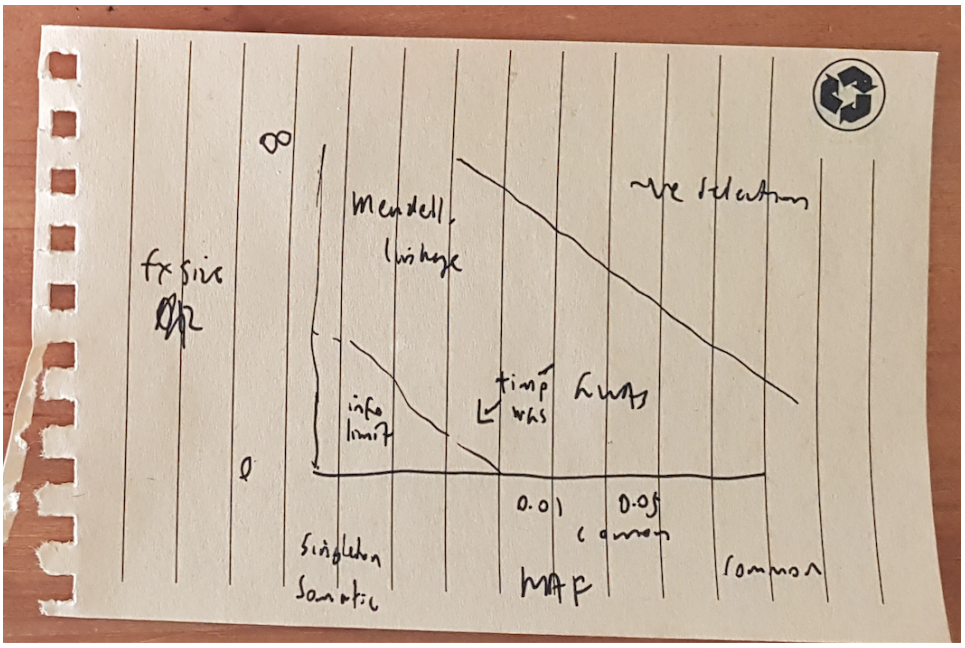

Figure 1.2: The reach of GWAS. OR vs MAF ala tam2019BenefitsLimitationsGenomewide, extended by imputation, sample size, WGS based genotypes, but may be indistinguishable from noise at the limits

that are poorly-tagged [14]. Modern imputation panels enable cost-effective GWAS including tens of millions of variants down to frequencies of ∼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 $0.05/10^6 = 5 \times 10^{-8}$ for genome-wide significance in European populations [20], akin* to controlling the family-wise type I error rate at using the Bonferroni correction.

*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].

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 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.

1.3 Gene expression as an intermediate phenotype

1.3.1 From causal variant to target gene

- For coding variants, there is a reasonable prior as to the target gene.
- 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 gene expression [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 https://www.nature.com/articles/s41586-020-2559-3.
  - For complex diseases, genomic enrichment of GWAS loci within regulatory elements are observed in disease-relevant tissues [14].
  - These enrichments put forth expression as an important intermediate linking non-coding GWAS variants to their associated traits (Fig. 1.3).

1.3.2 Expression is a complex trait

- Studies of the genetic architecture of expression have further reinforced this hypothesis.
  - Expression in itself, is a molecular phenotype that is heritable and complex [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 the most readily detectable eQTLs 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 variants 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 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.

add uses other vars

1.3.3 Context 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]

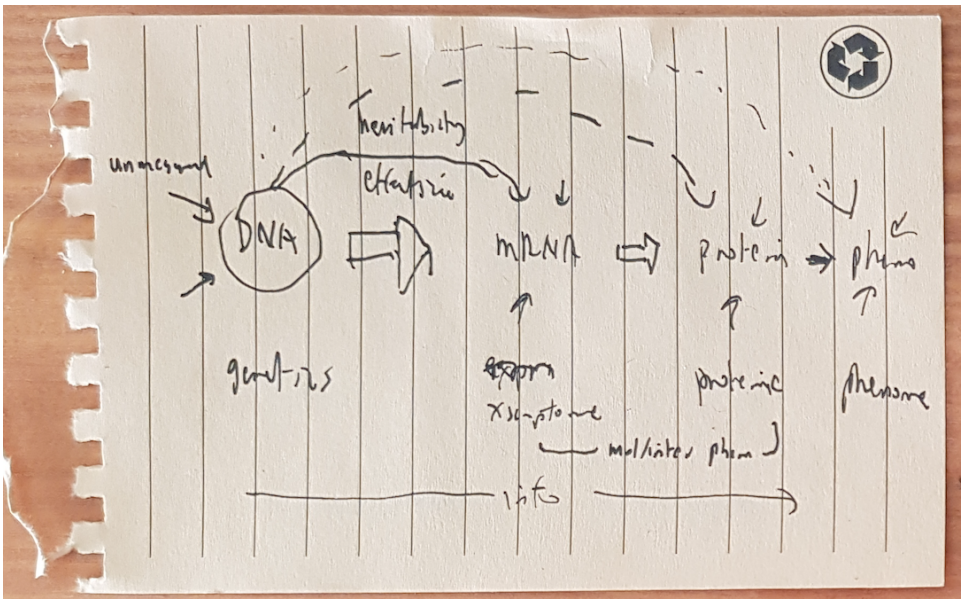

Figure 1.3: Mediation of genetic effect to phenotype, through the biological system

    - 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 1.3.4).

- 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.
- 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

1.3.4 Response expression quantitative trait loci in the immune system

- 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.
  - *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 more types and stims from [47] until [48]

- A complementary approach is *in vivo* reQTL mapping
  - There are 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

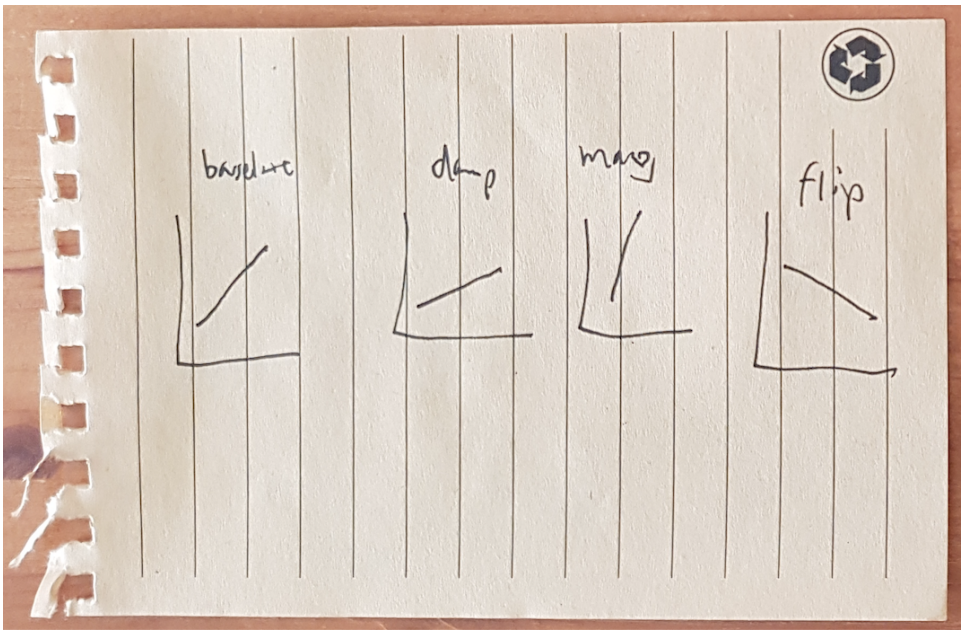

Figure 1.4: eqtl mech models: magnify, dampen, flip

  - 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 concluded they were 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.

  - Does not automatically reveal the mechanisms behind those differences, but provides a starting point for forming mechanistic hypotheses e.g. context-specific expression
  - Nevertheless useful for interpretation of GWAS signals, providing info on likely contexts that mediate the genetic effect
  - 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 Immune phenotypes are complex traits

1.4.1 Genetic effects on the healthy immune system

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

not sure if right order. Since most reQTL studies are immune, I went context-specific -> reQTL -> immune rather than context-specific -> immune -> reQTL

  - Systems studies of interindividual variation in the healthy immune system shows many aspects of the immune system are heritable and complex.
    - <Systems immunology: just getting started https://www.nature.com/articles/ni.3768>
  - 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? respecify scale of stability

  - 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].

- given genetic control of healthy system, perhaps not surprising that immune response to perturbation traits are also complex

  - also, as discussed in the context section above, context-specific genetic effects may not be apparent in the baseline healthy state, stimulation is required
  - since a central goal of systems immunology is to establish causal relationships between the many components of the immune system
    - Natural genetic variation can be leveraged, representing small scale perturbations that are causally anchored [60, 61]
  - In this context, immune in vivo reQTL studies can be considered as controllable perturbation studies of the activated immune system
    - Studies of natural infection are complicated by e.g. determining exposure, ethics, dose
  - Simultaneously provides insight in to the biology behind those specific responses
  - Two immune perturbations considered in this thesis are vaccines and biologic drugs.

1.4.2 Antibody response to vaccination is a complex trait

- Vaccination has enormous impact on global health [62]

  - <quick vaccine bio, specific flu vaccine goes in ch2>
    - 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>
  - 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 (sometimes dubbed as vaccinogenomics studies) are nowhere near as mature compared to the trait to gene pipeline described in above e.g. applied to complex disease

find best GWAS ref, probably mooney2013SystemsImmunogenetic then prune and reassign these citations

1.4.3 Response to biologic anti-TNF therapy is a complex trait

not sure about scope of this subsection, currently some overlap with PANTS chapter intro. tried to separate out only the non-IBD stuff here (mainly intro + RA context)

- <quick anti-tnf summary, specific ADA/IFX biology goes in ch4>
  - biologics are drugs synthesised using a living organism, typically proteins
  - cause immune response due to having immune targets, or immunogenecity because or their large and complex structure vs chem synth small molecule drugs
  - one of the largest classes are anti-TNFs
  - anti-TNFs (or TNF inhibitors), are drugs that suppress the activity of the TNF signalling pathway of the immune system
  - they are used to treat immune-mediated inflammatory diseases e.g. rheumatoid arthritis, Crohn's disease, psoriasis and ankylosing spondylitis.
    - indicated for many IMIDs e.g. rheumatoid arthritis, Psoriasis, ankylosing spondylitis. [69–71]
  - an enormous amount of money is spent on them: anti-TNF biologics are some of the largest market share pharmaceuticals
  - some proportion of patients fail. given the expenditures, it would be good to predict this

- <expression signatures of response to anti-TNFs>
  - have been detected e.g. for RA <"Validation study of existing gene expression signatures for anti-TNF treatment in patients with rheumatoid arthritis" https://pubmed.ncbi.nlm.nih.gov/22457743/>
  - most detected in small cohorts, many require validation
- <genetics of anti-TNF response>
  - pharmacogenomics is the study of the role of genetics in beneficial and adverse effects of drugs and theraputics https://doi.org/10.1016/S0140-6736(19)31276-0
  - some implementation in clinic already e.g. screening for certain allele-drug combos https://www.nature.com/articles/nature15817 https://academic.oup.com/bmb/article/124/1/65/4430783
  - GWAS in the pharmacogenomics field https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3003940/ https://www.futuremedicine.com/doi/full/10.2217/pgs-2018-0204
  - GWAS studies of anti-TNF response in RA https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6614444/
    - a few validation studies attempted e.g. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5937760/
- a lot of studies also done for CD and IBD (described in ch4)

1.5 Thesis overview

- My thesis focuses on two specific instances of in vivo immune response: antibody response to pandemic influenza vaccine in healthy individuals, and clinical response to biologic anti-TNF therapy for CD patients.
- <By chapter context-content-conclusion overview.>
  - <ch 2: systems vaccinology study of Pandemrix>

    - context: existing Sobolev study of expression differences between pandemic flu vaccine R/NR had small sample size and binary phenotype
    - content: meta-analysis of existing array with new RNAseq data and continuous phenotype
    - conclusion: distinct innate and adaptive expression response at d1 and d7; heterogeneity between array and RNAseq. significant expression differences between R/NR in meta-analysis at the gene set level

- <ch 3: in vivo reQTL study of Pandemrix>
    - context: relatively few studies have assessed the impact of genetic variation on expression response to flu vaccine
    - content: reQTL analysis for flu vaccine at d0, d1, d7. many reQTLs including sign flips. no particular gene set enrichments. evidence of cell type interactions at top hits.
    - conclusion: difficult to separate out modifying effect of cell composition. this may be a fundamental flaw in the study design

- <ch 4: systems immunology and reQTL study of response to anti-TNF treatment in CD>
    - context: studies on expression signatures of anti-TNF PNR have been small
    - content: R/NR comparison with larger n, at baseline, w14, and over time. reQTL analysis over 4 timepoints.
    - conclusion: a few hits for PNR at baseline. much stronger expression differences stronger at w14, then maintained until w54. Weak evidence for reQTLs, probably due to smaller magnitude of cell proportion changes over time vs the previous chapter.

- <discussion: limitations, future outlook>
    - main themes and parallels tying together the thesis
    - shared set of limitations permeating all chapters
    - recommendations for future analyses and study design

- future outlook for the fields of vaccinogenomics and pharmacogenomics

Chapter 2

multiPANTS: response to biologic anti-TNF therapy for CD

2.1 Introduction

2.1.1 Crohn's disease

- CD is a chronic inflammatory disease of the gastrointestinal tract.
  - CD is one of the two forms of IBD, characterised by patchy inflammation, where lesions are interspersed with regions of normal mucosa. The lesions can be distributed anywhere in the gastrointestinal tract, and tend to be transmural, affecting all layers of the gut wall.
  - The second form, UC is characterised by continuous inflammation, with lesions that are superficial rather than transmural, and restricted to the colon. [72]
  - Although these are two distinct forms of IBD, similarities in symptoms, therapies, genetic architecture mean they have historically been studied together.
    * The similarity is such that there is a subset of IBD-U patients with features of both CD and UC, and thus is difficult to classify as one or the other.

  - CD and UC are considered IMIDs, a group of related conditions involving immune dysregulation of common inflammatory pathways.
    * Other diseases include type 1 diabetes (T1D), systemic lupus erythematosus (SLE), rheumatoid arthritis (RA), multiple sclerosis (MS) and psoriasis. [73, 74]

- Pathogenesis of CD is not completely understood, but involves interaction of the immune system, environmental factors (e.g. smoking, stress, diet [72, 75]), and gut microbial factors in a genetically-susceptible individual [76].
  - Since the seminal discovery of association of *NOD2* with CD in 2001 https://www.nature.com/articles/35079223 (the first gene to be genetically-associated with a common disease), and the first IBD GWAS in 2006 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4410764/, a lot of progress has been made in establishing genetic risk factors.
  - The most recent GWAS studies define over 240 risk loci for IBD [77]
  - Genetic correlation between UC and CD is high [73, 74]
  - Most GWAS hits are shared, but there is some heterogeneity of effects between CD/UC, mostly notably at NOD2, which is much stronger for CD [78, 79]

add NOD2 OR

  - Concordance rates amongst monozygotic twins are higher for CD (50%) than for UC (∼15%) suggesting a greater heritable component [72]

- IBD has historically been considered a disease of the Western world.
  - The disease burden is now rising globally [80, 81].
  - The highest prevalence and incidence of new cases of CD are in North America and western Europe. [72]
  - CD is becoming increasingly common in newly industrialized countries in Asia, Africa and South America.

    - Migrants from low- to high- prevelance regions are at higher CD risk, suggesting there is an influence of Western lifestyle on disease risk. [72]
  - The modal age of onset of CD is typically between late adolescence and early adulthood.
  - CD is progressive: Within 20 years of diagnosis, 50% of patients with CD develop gastrointestinal tract complications and approximately 15% require surgical intervention [72]
  - Given the rising prevelance and large impact on quality of life, there is active research into effective therapies that lead toward the end goal of complete mucosal healing [72, 82]

2.1.2 Anti-TNF biologic therapies for CD

as suggested, the subsection on anti-TNFs in the intro chapter will likely be merged here

- tumour necrosis factor (TNF) (sometimes referred to by it's archaic name TNF-α), is a proinflammatory cytokine
  - It is synthesised mainly by immune cells: macrophages, NK, T and B cells in transmembrane form, then cleaved into the soluble form.
  - Binding of TNF to either of its two downstream receptors (most cells in the body express one or the other) on immune and gut mucosal cells triggers a signalling cascade that in different contexts, can regulate inflammation, apoptosis, cell proliferation and survival. [70, 83, 84]
  - In the IBD pathogenesis context, TNF mediates gut inflammation. One current model is TNF promoting gut epithelial apoptosis; and inhibiting T cell apoptosis, which maintains chronic inflammation. [82, 84, 85]
- The use of anti-glsTNF therapy has revolutionised CD and IBD patient care in the last two decades.
  - The two major agents used are adalimumab and infliximab. Both are IgG1 monoclonal antibodies that binding to both soluble and transmembrane TNF, inhibiting interaction with it's receptors [69, 85]

  - two major mechanisms of anti-TNFs: have been proposed induction of T cell apoptosis in the gut mucosa, and Fc dependent mechanism inducing wound-healing M2 macrophages [82]
    * Fc regions fixing complement can also lyse cells expressing membrane TNF [85]
  - adalimumab is a human monoclonal antibody
    * typically administered subcutaneously (via auto-injector pen) every two weeks, after 2 dose induction [85]
  - infliximab is a chimeric (mouse/human)
    * taken via intravenous infusion, typically on a maintenance schedule of every eight weeks after an initial three-dose induction [85]
    * more immunogenic (more anti-drug antibodies that interfere with drug activity), thought to lead treatment failure via lowering drug concentrations [86] https://journals.sagepub.com/doi/10.1177/1756283X17750355
  - ADA and IFX together are very costly: 29B USD in 2017 "the-global-use-of-medicine-in-2019-and-outlook-to-2023.pdf"

- What determines which drug is given?
  - e.g. guidelines:
    * 2010 ECCO: "all currently available anti-TNF therapies appear to have similar efficacy and adverse-event profiles, so the choice depends on availability, route of delivery, patient preference, cost and national guidance." https://www.nature.com/articles/nrgastro.2015.135, also see "Figure 2: Biologic agents in IBD: a proposed algorithm for clinical practice."
    * [87]
    * 2018 <Infliximab and adalimumab for the treatment of Crohn's disease> https://www.nice.org.uk/guidance/ta187>
    * 2019 <British Society of Gastroenterology consensus guidelines on the management of inflammatory bowel disease in adults https://gut.bmj.com/content/68/Suppl_3/s1.full>

  - It is currently not able to predict efficacy. Treatment failure is not uncommon.

2.1.3 Treatment failure

- There many types of failure: primary non-response (PNR) within the induction period (12-14 weeks for ADA/IFX), secondary non-response, non-remission, adverse events https://journals.lww.com/ctg/Fulltext/2016/01000/Loss_of_Response_to_Anti_TNFs__Definition,.2.aspx
  - e.g. failure rate estimates
    * "The incidence of PNR varies between clinical trial and clinical practice from 10–20% to 13–30%.2, 3, 5" 2013 https://pubmed.ncbi.nlm.nih.gov/23792214/
    * "nonresponse for TNF antagonist therapy varies between clinical trial and clinical practice from 10 to 30%,2–4 and the annual risk of secondary nonresponse from 13% per patient year for infliximab (IFX)5 to 20.3% for adalimumab.6" 2018 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5784543/
    * "Unfortunately, anti-TNF treatment failure is common: 10–40% of patients do not respond to induction therapy (primary non-response),6–8 24–46% of patients have secondary loss of response in the first year of treatment,9 and approximately 10% have an adverse drug reaction that curtails treatment.10" [86]
  - also, immunogenicity leads to non-response via anti-drug Abs; serum drug levels correlated with efficacy, anti-drug Abs are inversely correlated [69]
  - although mucosal healing is the preferred gold standard [72], could be argued that PNR is important, as it's correlated with remission, and measurable early for stratification

- when they fail: anti-TNFs biologics just one part of the therapy pyramid for CD https://www.nature.com/articles/nrgastro.2013.158

  - a pyramid with higher toxicity/patient impact and high cost at the top
  - Two approaches to treat patients, neither are ideal
    * Step-up approach: may undertreats patients that require more aggressive therapy, allowing disease time to progress
    * Top-down approach: exposes patients to risks from more aggressive therapies they may not need from overtreating
    * recent moves towards a hybrid: start at biologics and go top-down if possible https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5784543/
  - after failure of biologics, other options include dose intensification, and switching to an agent with diff mech [69]
  - But better still would be to target the right therapy to those that require and respond to it
  - baseline predictors would be especially valuable for patient stratification

2.1.4 Predicting response to anti-TNFs for CD

- previously reported clinical predictors include age, disease duration, BMI, smoking, CRP, FCP, anti-TNF drug and anti-drug antibody concentrations, but these have mostly been found in small retrospective cohorts, and rarely independently validated [84, 87–91]
  - in PANTS: a recent prospective study of ADA and IFX for CD to date: large n=1610 [86]
  - PNR has observed at a 23.8% rate, evaluated at w14 via clinical decision algorithm.
  - low serum drug concentration in peripheral blood at w14 (ELISA) was associated with PNR, and also non-remission and immunogenicity for both drugs
  - no associations at baseline
- Studies have also attempted to define transcriptomic predictors for anti-TNF response in IBD [84, 91]

  - gut:
  - [92]: TNFRSF11B, STC1, PTGS2, IL13RA2 and IL11. endoscopic and histological healing. infliximab. UC. colon mucosal biopsy. total n 46.
  - [93]: (TNFAIP6, S100A8, IL11, G0S2, and S100A9). complete endoscopic and histologic healing. infliximab. Crohn's colitis CDc. gut biopsy. 19
  - [94]: OSM and OSM-hier cluster module. multiparameter measures clinial Mayo score. UC. gut mucosal biopsy. 227.
  - GIMATS [95]: expression of IgG plasma cells, inflammatory mononuclear phagocytes, activated T cells, and stromal cells, which we named the GIMATS module. various anti-TNF therapy. ileal CD. achieving durable (6 months) corticosteroid-free clinical remission (pediatric Crohn's disease index (PCDAI) < 10) at months 18 and 24 post-diagnosis, as responders. inflamed ileal biopsies. n=340 two cohorts with baseline.

- blood: prognostic tests performed on blood samples are non-invasive and hence of high value.
  - there is conflict in existing studies on TREM1
  - [96] : endoscopic score and/or clinial decision algorithm. infliximab. .
    * gut biopsy: proportion of plasma cells. following adjustment to plasma cells and activated monocytes from expression deconv, TREM1 is upstream reg of resulting DEGs: TREM-1 is expressed on myeloid lineage cells including monocytes and macrophages, has well-documented proinflammatory functions. 81 patients.
    * [96]: TREM1 down in NR. blood. endoscopic. CD. 22 patients.
  - [97]: opposite. Baseline whole blood TREM1 was downregulated in future anti-TNF responders. endoscopic remission. ifx and ada. IBD.
  - difference may be due to sample size, ethnicity, definition of response https://pubmed.ncbi.nlm.nih.gov/30007919/ [84]

  - [98] SMAD7, TLR2 and DEFA5. response based on decrease in Pediatric Crohn's Disease Activity Index (PCDAI) and Pediatric Ulcerative Colitis Activity Index (PUCAI). pIBD, IFX/ADA. n=31

- Finally, genetic markers for non-response [90]
  - baseline by definition
  - anti-TNF response does not necessarily share the same genetic architecture as disease risk loci e.g. polymorphisms in IBD risk lock NOD2, TNFR1, TNFR2 are not associated with NR to infliximab [84, 91]
  - IL-13 receptor (IL13Rα2), IL-23 receptor (IL23R), TNF-receptor I (TNFRI), IgG Fc receptor IIIa (FcYRIIIa), neonatal Fc receptor (VNTR2/VNTR3), apoptosis-related genes (Fas ligands, caspase 9), and MAP kinases, FAS-L, caspase 9 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6128143/ [90], and also multi-SNP risk scores https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6128143/
  - but like transcriptomic predictors, mostly small candidate gene studies
  - although no GWAS hits for PNR in PANTS, a related phenotype, immunogenicity is associated with HLA-DQA1*05 [99], so there may be some promise in the area

- in summary, varying in technology used to measure gene expression, analysis method, anti-TNF drug, response definition (both criteria and time of assessment), tissue, sample size, disease among all these studies make a consensus hard to establish.
  - none of these clinical, transcriptomic, or genetic markers have been sufficiently discriminative and validated for use predicting NR in the clinic yet, although several are undergoing validation [91]

2.1.5 Chapter summary

this subsection

- <context>: summary of above
  - conflicting results in the literature and need for validation for transcriptomic signatures for R/NR to anti-TNF
- <content: our approach>
  - use PANTS cohort data, the largest RNAseq dataset to date on CD patients with anti-TNF therapy, to define expression differences between PR/PNR
  - also able to evaluate evidence for genetic control of blood expression in CD patients
- <conclusion: our results>
  - there are some baseline differences between R and NR, weak effects that require further validation
  - there are strong transcriptomic differences after the induction period (12 weeks) between R and NR
  - these differences are maintained over time up to w54, suggesting NR phenotype is stable over time and many doses
  - change from baseline to w14 for expression is magnified in R vs NR, suggesting there may be a continuum of response
  - TREM1 baseline signature not replicated
  - weak evidence of interaction of anti-TNF therapy with genetic control of expression (reQTLs)

2.2 Methods

2.2.1 The PANTS cohort

Personalised Anti-TNF Therapy in Crohn's Disease (PANTS) is a prospective, observational, UK-wide cohort study of response to anti-glsTNF therapy in Crohn's disease (CD) patients, described in detail by Kennedy *et al.* [86]. The study was registered with ClinicalTrials.gov identifier NCT03088449, and the protocol is available at `https://www.ibdresearch.co.uk/pants/`. In brief,

total enrollment was 1610 patients, who were at least 6 years old, had active luminal CD, and were naive to anti-glsTNF therapy. Patients were invited to attend up to ten major study visits over a maximum follow-up period of three years, or until drug withdrawal.

The anti-glsTNF drugs evaluated were adalimumab and infliximab*. All major visits were scheduled immediately prior to a drug dose. Although adalimumab and infliximab have different dosing schedules, the timing of major visits was chosen such that the same visit structure could be used for patients on both drugs. Additional visits could also be scheduled in case of secondary loss of response, or exit due to drug withdrawal.

2.2.2 Definition of timepoints

The expression data for this chapter comes from PANTS samples centered around four timepoints from the first year of follow-up: week 0, week 14, week 30, and week 54. These are the specified timings for four major visits in the first year. Whole blood samples were taken prior to drug administration, and preserved for RNA-sequencing (RNA-seq) in Tempus Blood RNA Tubes. The study day that sampling occurred relative to the first drug dose was recorded.

For the purpose of measuring the transcriptome at trough drug levels, I mapped samples from major and additional visits to four timepoints centered around the four major visits. As it could not be guaranteed that visits occurred on the exact day specified in the protocol, I considered the visit windows defined by Kennedy *et al.* [86]: 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) Samples were mapped according to the following criteria:

* Major visit samples were mapped to the corresponding timepoint, regardless of whether they fell within the corresponding window i.e. an available week 0 sample is always mapped to

*The study also evaluated infliximab biosimilars. Data from patients who received a biosimilar is not included in this chapter

the week 0 timepoint.

* Samples taken at additional loss of response or exit visits falling within one of the windows were mapped to that timepoint, unless the patient also had a major visit sample inside that window.

Only a small minority of major visit samples fell outside their corresponding windows. Inclusion of samples from additional visits was important as they often replaced major visits for patients with primary non-response or loss of response. Samples included under both criteria should be representative of trough drug levels, as major visits and loss of response visits were always scheduled prior to a drug dose, and exit visits were scheduled for when the next drug dose would have been.

compute maximum deviation

Still discussing with Sim on the exact def of LOR and exit visits to decide whether this is sensible.

2.2.3 Definition of primary response and non-response

I used a definition of primary non-response based on the decision tree from Kennedy *et al.* [86]. Primary non-response was assessed at week 12, prior to the week 14 drug dose. The criteria for primary non-response was *either* of the following:

* exit for treatment failure before week 14 (e.g. as decided by physician global assessment), *or* corticosteroid use at week 14 (a continuing or new prescription).
* compared to week 0, a decrease in C-reactive protein (CRP) by less than 50% or to $>3\,\mathrm{mg\,l^{-1}}$, *and* a decrease in Harvey Bradshaw index (HBI) by less than 3 points or to >4.

In addition, the primary non-responders in the RNA-seq dataset were selected to exclude patients in remission at week 54. As this is an observational study that continues until drug withdrawal, so a patient's clinician may decide to continue anti-glsTNF therapy even if a patient has primary non-response.

The definition of primary response was to not have primary non-response, and also have CRP $\leq 3\,\mathrm{mg\,l^{-1}}$ and HBI < 4 by week 14. Grey zone patients that were intermediate between primary response and non-response were not analysed.

2.2.4 Library preparation and RNA-seq

Total RNA was extracted from whole blood samples preserved in Tempus Blood RNA Tubes, following the Qiagen QIAsymphony instrument protocol (RNA Isolation PAX RNA CR22332 ID 2915. RNA was then quantified with the ThermoFisher QuBit BR RNA (Q10211), RNA integrity was assessed with the Agilent RNA ScreenTape assay (5067-5579, 5067-5577, 5067-5576) on the Agilent 4200 TapeStation.

Library preparation was done using the Kapa mRNA HyperPrep Kit, including enrichment for mRNA using magnetic oligo-dT beads, depletion of rRNA and globin mRNA using the QIAseq FastSelect RNA Removal Kit, and adapter ligation with IDT xGEN Dual Index UMI adapters. Libraries were sequenced on the Illumina HiSeq 4000 using 75 bp paired-end reads.

2.2.5 RNAseq quantification and quality control

A total of 1141 samples were initially sequenced. Sequencing data was demultiplexed with Picard. Reads were mapped to GRCh38 using STAR (2.6.1d) and deduplicated to unique reads using UMI-tools. Gene expression was quantified against the Ensembl 96 gene annotation using featureCounts (1.6.4).

Total number of read pairs, sequence quality, overrepresented sequences, adapter content and sequence duplication rates were checked using FastQC. Samples were filtered to remove outliers (>2SD from the mean) according to percentage of aligned reads in coding regions reported by Picard, percentage of unique reads, and number of unique reads. Samples that could not be mapped to a timepoint according to subsection 2.2.2 were removed. Samples that came from patients with sex mismatch based on Y chromosome gene expression, grey zone primary response, or missing data for variables considered in the variable selection process (subsubsection 2.2.6.1), were removed. A total of 814 samples remained after filtering.

The Ensembl 96 gene annotation contains 58 900 genes, many of which are not expressed in whole blood. Normalised library sizes

were computed using edgeR::calcNormFactors(method='TMM') for use in CPM calculations. Genes were filtered to require >1.25 CPM in >10% of samples, where 1.25 CPM is approximately 10 counts at the median library size; and non-zero expression in >90% of samples. Globin genes and short ncRNAs were removed. A total of 15511 genes remained after filtering. Finally, counts were converted to the log2 CPM scale and precision weights to account for the mean-variance relationship were computed for each gene and sample using voomWithDreamWeights.

2.2.6 differential gene expression

2.2.6.1 Variable selection by variance components analysis

For each gene, the form of the differential gene expression (DGE) regression model has gene expression as the response variable, predictor variables of interest (such as primary non-response, drug, and timepoint), and other selected predictor variables. Many variables are available Variables in Fig. 2.1 were included as predictors.

- In estimating the effect X→Y, of predictor X on response Y by regression, adjustment for a third variable Z can increase, decrease, or even reverse the effect estimate. The regression model is agnostic to what causal role Z may play, but different types of third variable can be distinguished conceptually if some relationships are actually causal. In this section, I focus on identifying third variables that are covariates for inclusion into the DGE model: where Z is associated with Y and explains some variation in Y, and conditioning on Z increases the efficiency of estimating X →Y. conditioning on a covariate that is not associated with the predictor ($X \nleftrightarrow Z \leftrightarrow Y$, aka. a precision variable, prognostic variable in clinical trials lit) reduces variance of the estimate,

- If covariates are also associated with the predictor X, issues can arise depending on their causal role. Conditioning on a confounder (X ←Z →Y) reduces bias, conditioning on a collider (X →Z ←Y) induces bias,

and conditioning on a mediator in the causal path (X →Z →Y) changes the effect estimated by removing the indirect effect mediated by Z.

- 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; **??** shows their correlations with each other, and the predictor variables. 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.
- visualised main factors that influence global gene expression by PCA (Fig. 2.2)
  - main separation along PC1 is w0 anti-TNF naive samples from all other post-drug start samples
  - TODO: color other PCs by other variables: sex, response status, library prep protocol
- A variance components analysis was conducted to formally quantify the fractions of variation in expression explained by known variables using `variancePartition`[100], which fits a mixed regression model. Variables in Fig. 2.1 were included as predictors.
  - Includes prognostic factors from [86]
- Additional categorical variables were included for patient, 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 perfect multicollinearity. Categorical variables were coded as random effects, and continuous variables as fixed effects. Surprisingly, Hoffman *et al.* [100] 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

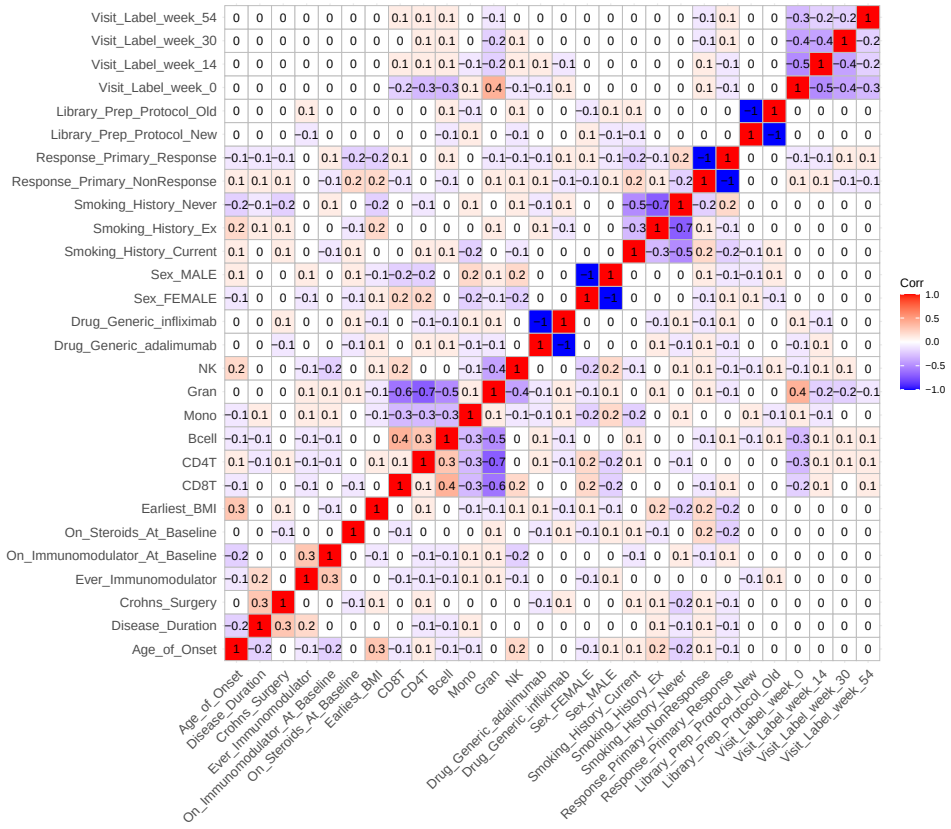

Figure 2.1: Correlation matrix of phenotypes considered as independent variables in DGE and eQTL models.

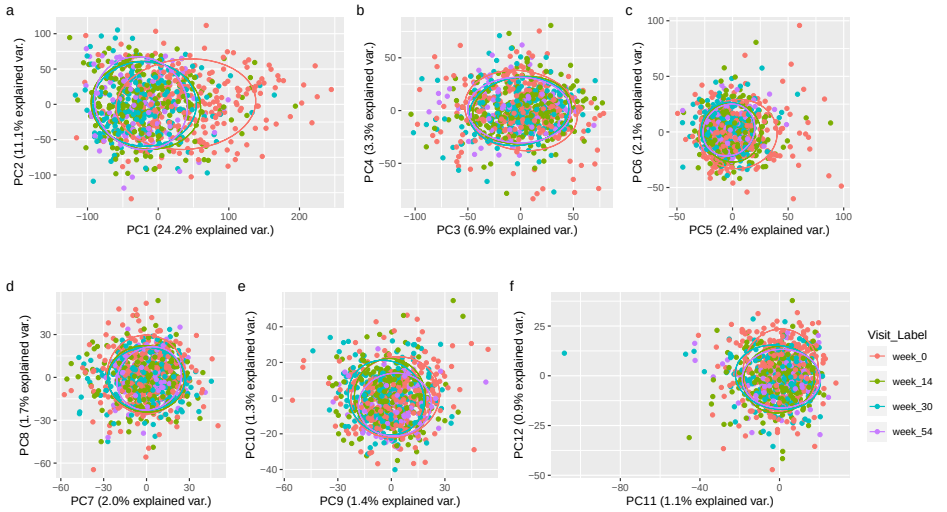

Figure 2.2: top 12 expression PCs of filtered expression data

restricted maximum likelihood (REML)*. It was also shown that this approach also avoids over-estimates of variance proportions that occur if categorical variables with many levels are treated as fixed.

- Variables were ranked by median variance proportion across all genes (Fig. 2.3). The variables that explained the most variance included patient, cell proportions and RNA-seq plate.
  - most influential on interpretation are cell counts: there are pros and cons to using them
  - Cell proportions explain a lot of variance: this is expected `https://genome.cshlp.org/content/early/2020/06/24/gr.256735.119.abstract?papetoc` and even more so as they change a lot over time (Fig. 2.4)

anova for each cell prop over time for p values

  - in the case of mediation i.e. PNR -> CC -> R
  - Should rarely find cell prop to be a collider, as in most genes, E -> cell prop is unlikely vs cell prop -> E
  - so keep them in as covariates

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

    - it's already popular for diff meth, and in DGE, can increase robustness https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1878-x
  - note that we end up with the adjusted effect: upregulation in this context is increase in transcripts because making more per cell, not more in the bulk sample
    - this may not be ideal in prediction context, as each var input needs to be measured, and may even attenuate ability to predict [96]

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?

- 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.
  - would be best to customise per gene, but less comparable interpr between genes
  - Included, as we need a consensus set of covariates, and the penalty is only 1 df.
  - final list of covariates is [TODO: basically select all, except Gran, since we have proportions; and ever immunomodulator, which is low in median and max var explained]

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

2.2.6.2 Contrasts for pairwise group comparisons

- model form is E [...] don't forget random fx!
- can we pool drugs with a mean term for this comparison? i.e. move from 8 to 4 means model
  - test interaction between drug and response at w0, and at w14 i.e. is there a diff in the diff between R vs NR between drugs?
  - no significant single gene hits at either timepoint
  - but unclear if preregistered power calculations would extend to this test https://statmodeling.stat.columbia.edu/2018/03/15/need-16-times-sample-size-estimate-interaction-estimate-main-effect/
  - there are baseline diffs are evident in clinical data

because this is non-randomised, baseline differences do matter??

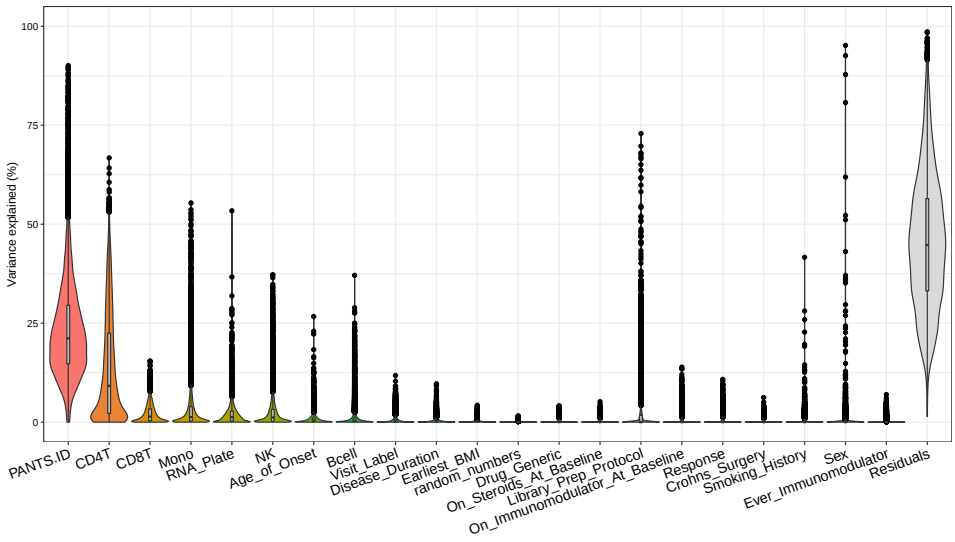

Figure 2.3: variance partition analysis, distribution of genewise % variance in expression explained by each variable

  - "Several baseline characteristics were significantly different between the infliximab-treated and adalimumab-treated patients, including age, smoking, body-mass index, disease duration, disease location, and disease behaviour. Patients treated with infliximab had more active disease at baseline than did patients treated with adalimumab, as evidenced by higher serum CRP and faecal calprotectin concentrations (table 1)." [86]
  - I include many but not all of these covariates in the DGE model

- used `dream` hoffman2018DreamPowerfulDifferential
- Group-means parametrisation with 8 means
  - equiv to to 3 way interactions between drug/response/visit
  - no intercept, so each group coef is a mean estimate
- specific hypotheses tested using sum to zero contrasts
- Model also fit that used Group-means parametrisation with 4 means: pooling the two drugs
- TODO: model equations

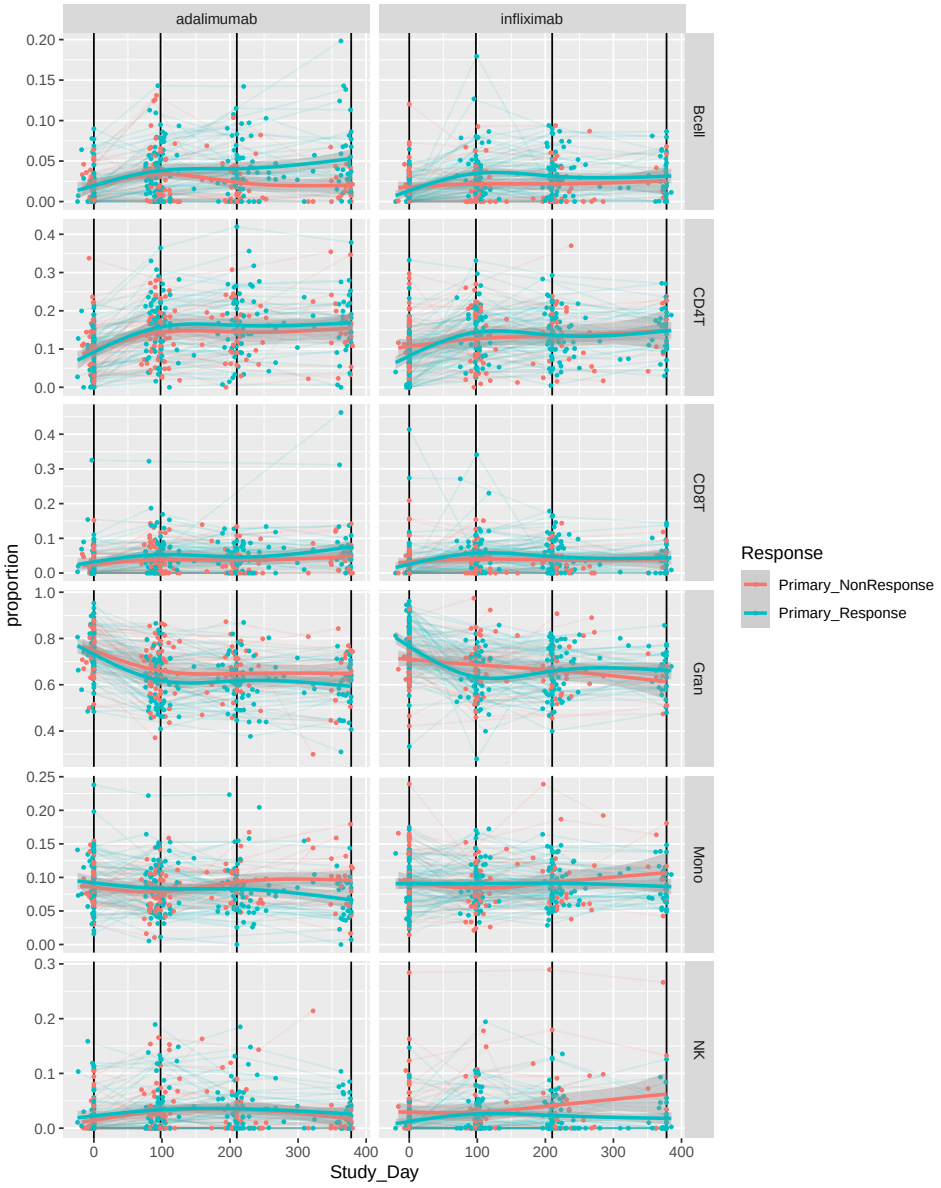

Figure 2.4: changes in cell proportions of 6 immune cell types over time

  - for dream analysis, unlike variancePartition, use REML over ML, so use fixed effects for small numbers of levels
  - also, need fixed effects for tested predictors
  - to get p values for papers, Dream uses lmerTest approximation Satterthwaite df https://link.springer.com/article/10.3758/s13428-016-0809-y with REML
  - this combo controls type 1 error for n>144 in lmerTest simulations
  - FDR BH separately per comparison: "The default method="separate" and adjust.method="BH" settings are appropriate for most analyses. method="global" is useful when it is important that the same t-statistic cutoff should correspond to statistical significance for all the contrasts." https://rdrr.io/bioc/limma/man/decideTests.html

2.2.6.3 Spline model for difference over all timepoints

- aim is to uses info in samples from other timepoints, avoiding a large number of pairwise comparisons
  - a simple study day x responder interaction over time assumes linear change
- could aosl treat time as categorical visits (like baseline/w14 analysis above), then f test all interactions
  - but there is variation in study day in the visit windows
- spline model tests over all timepoints, are there diff trajectories for R and NR?
  - Internal Knots set at w14 and w30, since drug dose just after the visit, so slope should be allowed to change until next dose
  - cubic between knots, linear outside external knots
  - f test over 3 interaction terms between spline basis and study day
    * can include all data
    * TODO: read this for maths behind basis functions https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-019-0666-3

- TODO: clustering spline hits
  - note more accurate to use partial expression, but complicated for DREAM, so used unadjusted expression
  - Centroids defined by simple mean in each visit

    less rudimentary

    a more direct connection to spline model fit e.g. clustering basis function results e.g. predicted values of expression

    how to account for levels of noise in expression data?

2.2.6.4 Gene set analysis ranked

- TODO: grab tmod paragraph from ch2
  - Genes are ranked by their test statistics, meaning we are ranking by significance
    - practice ranks are comparable between t and z.std, even though dream says otherwise, very high spearman cor
  - approx 8k genes in the gene set annotation for tmod

    The gene sets used were blood transcription modules (BTMs) from[101] "DC" (from Chaussabel et al. 2008)

2.2.7 Genotyping and genotype data preprocessing

- Whole blood samples were collected into EDTA tubes for genotypeing at w0
- TODO: scan Alex's thesis for genotyping and QC details
  - autosomal only
  - TODO describe strange limix behaviour that lead to deduping visits by patient in sample filtering

2.2.8 Computing genotype PCs

- used weights from akt for 1000G to do projection Fig. 2.5
- chose top 5 PCs for use in eQTL model

  - there is less pop structure here than HIRD. in HIRD, i did the PCA myself, and found 4 significant PCs with tracy-widom
  - from EIGENSTRAT paper, results not sensitive to number of PCs anyway, as long as it is sufficient https://www.nature.com/articles/ng1847

2.2.9 reQTL analysis

- same overall strat as ch3

2.2.9.1 Finding hidden confounders in expression data

- PEER (same as ch3)
  - Used DESeq2 vst for between sample norm e.g. sequencing depth
  - Chose n PEER to maximise cis-eQTLs on chr1 Fig. 2.6

2.2.9.2 Computing GRMs

- LDAK, same as ch3

2.2.9.3 limix model

- same as ch3, limix
  - n at each timepoint was [...]
  - AC thresh 15
  - extra filter to avoid small numbers of minor hom, as without sufficient numbers, leads to data points with high leverage that may be unduely influential on the beta
  - used a >5 min hom filter
- find lead eQTL for each gene in any condition by lowest lfsr
  - breaking ties by highest imputation INFO, highest minor allele frequency (MAF), shortest dist to transcription start site (TSS), and genomic coordinate.

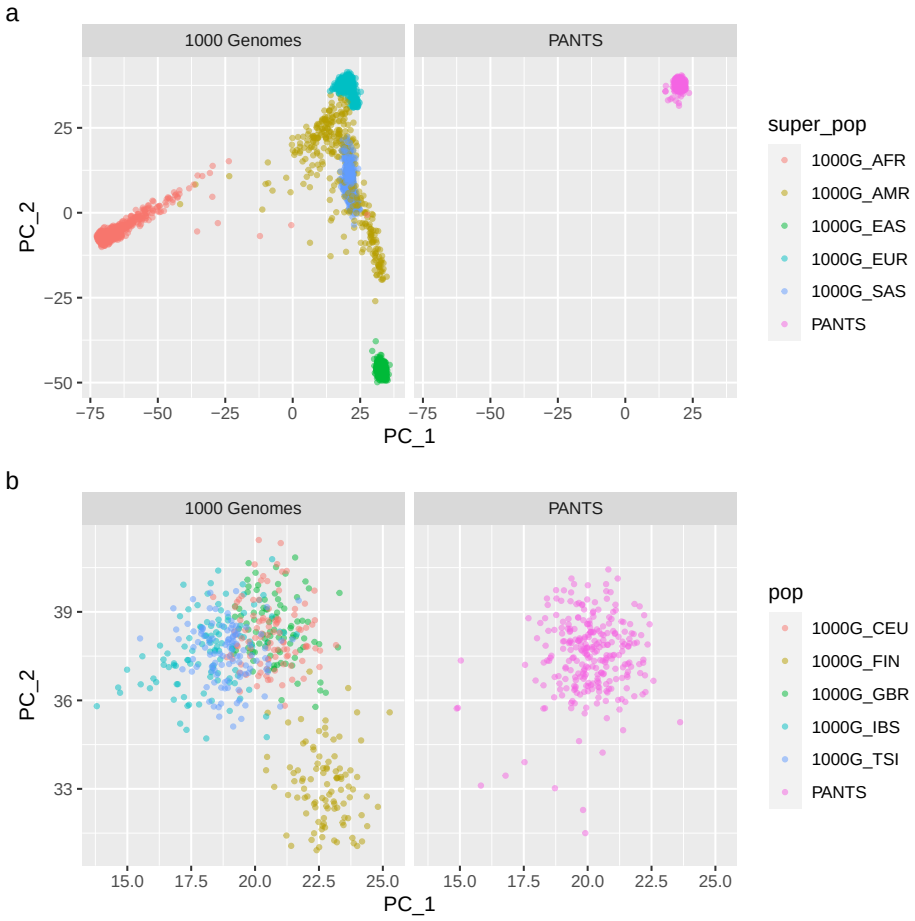

Figure 2.5: projection of PANTS samples onto 1000G genotype PC axes

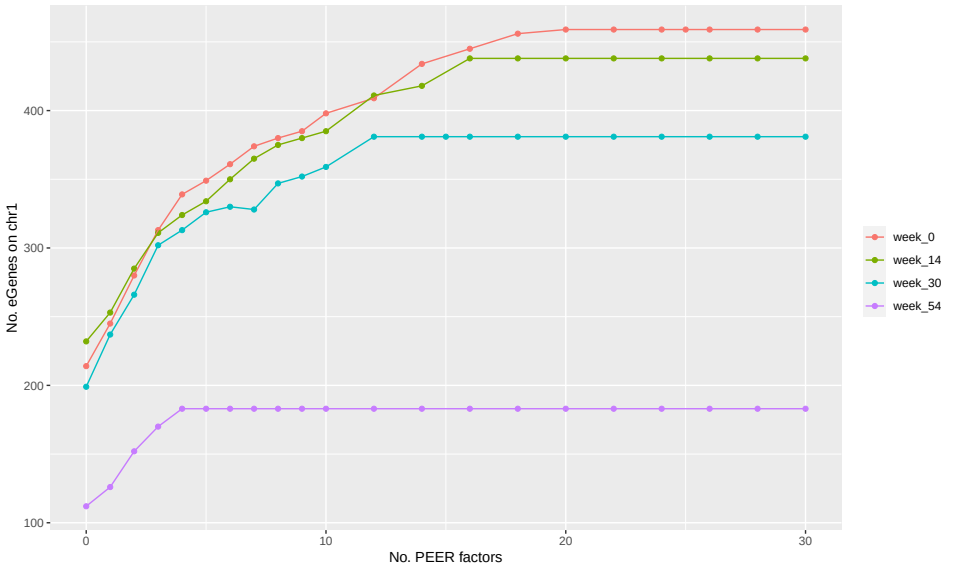

Figure 2.6: number of eGenes on chr1 used to choose number of PEER factors for each timepoint

2.2.9.4 mashr joint analysis

- same as ch3
  - TODO describe mashr bug for negative betas
  - reQTLs defined by difference in betas test between timepoints
  - BH FDR separate per comparison, not globally

2.2.9.5 Clustering reQTLs

- <pipeline>
  - align
  - Centering, no scaling
    - ensure comparability between gene
    - Amplifies noise? Migitate by prefiltering on nominal signif diff between two timepoints
  - dist_cor(method='pearson')
  - fastcluster::hclust(method='complete')
  - distance metric 1-cor(pearson)
  - Number of clusters: gap stat fviz_nbclust

2.2.9.6 gprofiler

2.3 Results

what to put in results vs discussion. going with the pattern of providing enough info for the reader to intepret the data in the results, then doing a summary and my own interpretation in the discussion

2.3.1 Longitudinal RNA-seq data from the PANTS cohort

To define transcriptomic differences between primary responders and non-responders to anti-glsTNF therapy in the PANTS cohort, I analysed whole blood RNA-seq gene expression measured at up to four timepoints per patient: week 0 baseline before commencing anti-TNF therapy, and weeks 14, 30 and 54 after commencing anti-TNF therapy. After quality control, expression data was available for 15584 genes and 814 samples (Fig. 2.7). These samples come from 324 patients, with a median of three samples per patient (Fig. 2.8).

Patient characteristics are shown in Table 2.1. The proportion of primary non-responders is high (43.8%) compared to the overall proportion in the PANTS cohort (23.8%, [86]). This is due to sample selection for RNA-seq to balance the sample size for each combination of drug and primary response status.

2.3.2 Gene expression associated with response at baseline

Patient primary response to anti-TNF was defined after the induction period, between week 12 and week 14 according to the clinical decision algorithm from Kennedy *et al.* [86] described in subsection 2.2.3, which integrates clinician assessment with change in CRP level and HBI score. To identify differences in baseline gene expression associated with future primary response, I fit gene-wise linear models at 15511 genes, comparing week 0 gene expression in primary responders with primary non-responders. Comparisons were performed both within infliximab-only and adalimumab-only subgroups, and with both drugs pooled. Models were run both adjusting for cell composition estimates of six immune cell types,

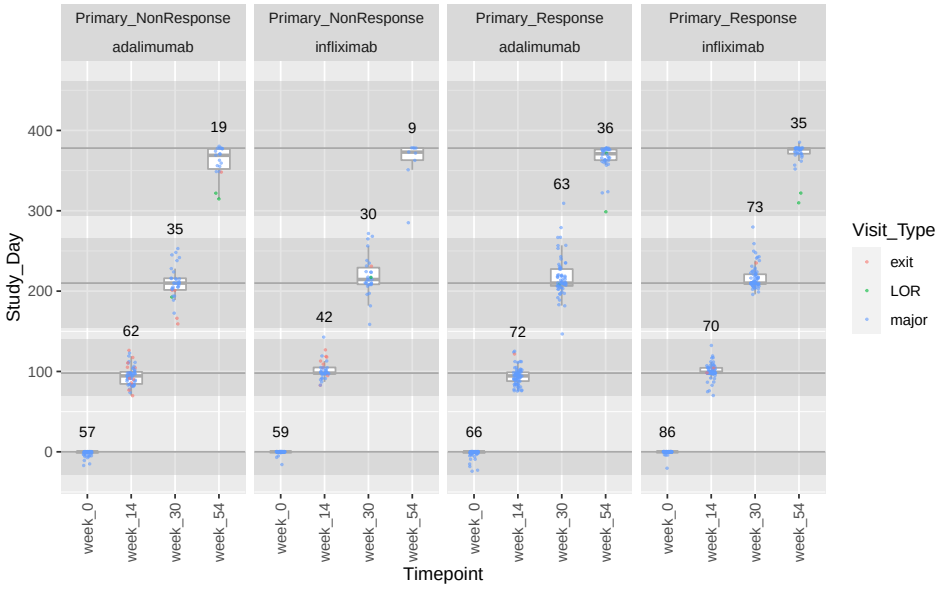

Figure 2.7: Distribution of samples among defined study visit windows. lor and exit are additional visits that fall into the windows.

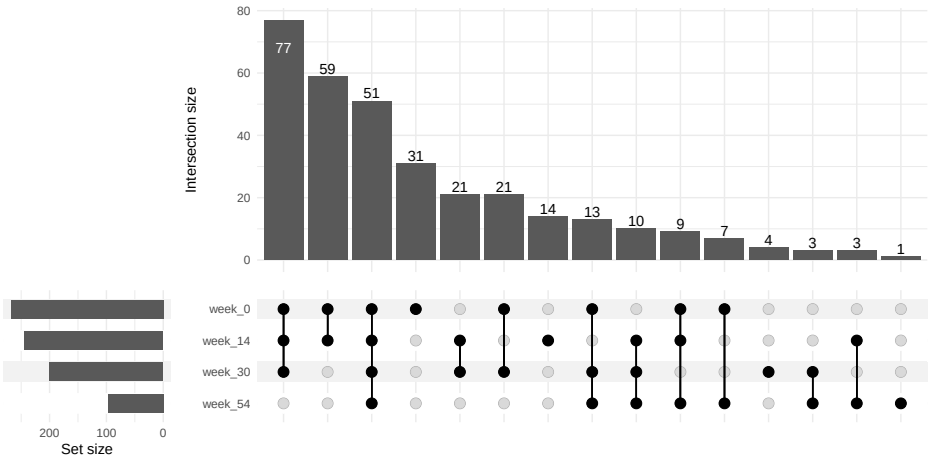

Figure 2.8

and without adjustment. Throughout this section, the significance threshold was set at FDR < 0.05 for each comparison, and up-regulation refers to increased expression in primary responders versus non-responders.

Without adjusting for cell composition, the largest effects were infliximab-only, with 859 genes differentially expressed. Only *KCNN3* (log2FC=−0.843) was significant for the adalimumab-only comparison, and only *SIGLEC10* (log2FC=0.346) was significant in the pooled analysis (Fig. 2.9).

After adjustment for cell composition, there were no longer any significant genes in the infliximab-only analysis, with 856/859 genes that were significant before the comparison having a dampened effect size after correction (smaller absolute effect and same sign), suggesting many effects may be mediated by cell composition. *SIGLEC10* in the combined analysis was also non-significant after adjustment (adjusted log2FC=0.314, FDR=0.0503). Conversely, at three genes downregulated in the adalimumab-only analysis that were the only significant genes post-adjustment, I observed increased significance: *PDIA5* (unadjusted log2FC=−0.335, adjusted log2FC=−0.351), *KCNN3* (−0.843, −0.880), and *IGKV1-9* (−1.15, −1.22).

To identify coordinately up and downregulated gene sets and increase sensitivity for detecting differences between primary responders and non-responders, I performed ranked gene set analysis in on the gene-wise z-statistics using blood transcriptomic modules: annotated sets of coexpressed genes in peripheral whole blood from Li *et al.* [101] (prefixed "LI"). This module-level analysis was also run unadjusted (Fig. 2.10) and adjusted for cell composition (Fig. 2.11).

Despite only *SAMD10* having a significantly different effect between drugs at the gene-level (an interaction between drug and response at week 0), the large global differences observable in Fig. 2.9 were detected in the module-level analysis*. Without

*It is likely the study is not powered to detect gene-level three-way interaction effects between timepoint, drug and response, although I am not aware of which subgroup analyses may have been prespecified during the study design and sample size calculations for the

adjusting for cell composition, many of the most significantly upregulated modules in the pooled analysis, including upregulation of monocyte (LI.M11.0, LI.S4), neutrophil (LI.M37.1, LI.M11.2), and dendritic cell (LI.M165, LI.S11), appear to be driven by infliximab. These modules have heavily reduced significance after adjusting for cell composition. The new modules that are most upregulated in the pooled analysis after adjustment have more consistent effects between drugs, such as MHC-TLR7-TLR8 cluster (LI.M146), antigen presentation (LI.M71, LI.M95.0), and myeloid cell enriched receptors and transporters (LI.M4.3).

For downregulated modules before adjustment, I observed infliximab-specific effects for for NK cell (LI.M7.2) and T cell (LI.M7.0, LI.M7.1) modules. Adalimumab-specific effects were observed for plasma cell, B cell and immunoglobulin modules (LI.M156.0, LI.M156.0, LI.S3); and cell cycle and transcription modules (LI.M4.0, LI.M4.1). After adjustment, the significance of infliximab-specific modules was reduced, but the significance of adalimumab-specific modules and the corresponding interaction effects was increased.

2.3.3 Assessing previously reported baseline predictors of response

In addition to hits from this study, Fig. 2.9 is annotated with genes whose expression in gut biopsies or blood have been tested as baseline predictors of primary response in the literature[92, 93, 97, 98]. Some genes expressed in gut mucosa (e.g. *IL13RA2*) were not appreciably expressed in this whole blood dataset, and most other genes that were expressed were not differentially expressed. Only *TNFRSF1B* and *PTGS2* were associated with primary response, specifically in the infliximab-only comparison unadjusted for cell composition.

A previously identified marker in blood *TREM1*, found to have opposing in two studies from [96, 97] was not significantly associated with response in this study neither before (logFC=0.293, FDR=0.0595) nor after adjusting for cell composition (logFC=0.0463,

PANTS RNA-seq cohort.

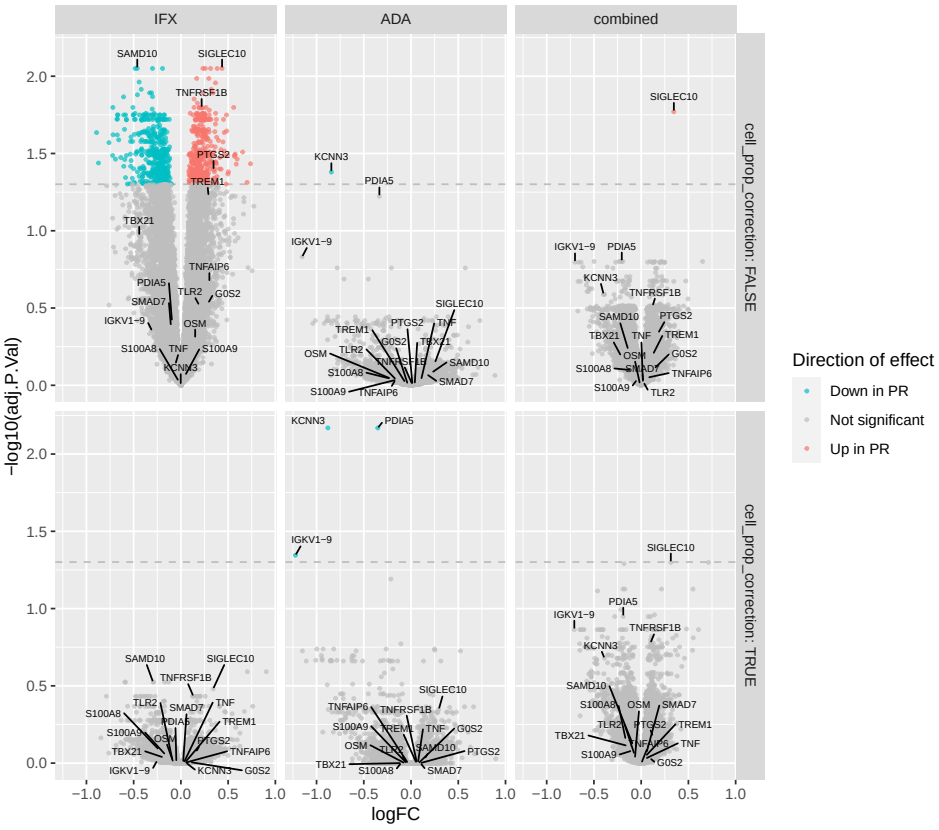

Figure 2.9: DGE volcano plot for PR vs PNR at week 0

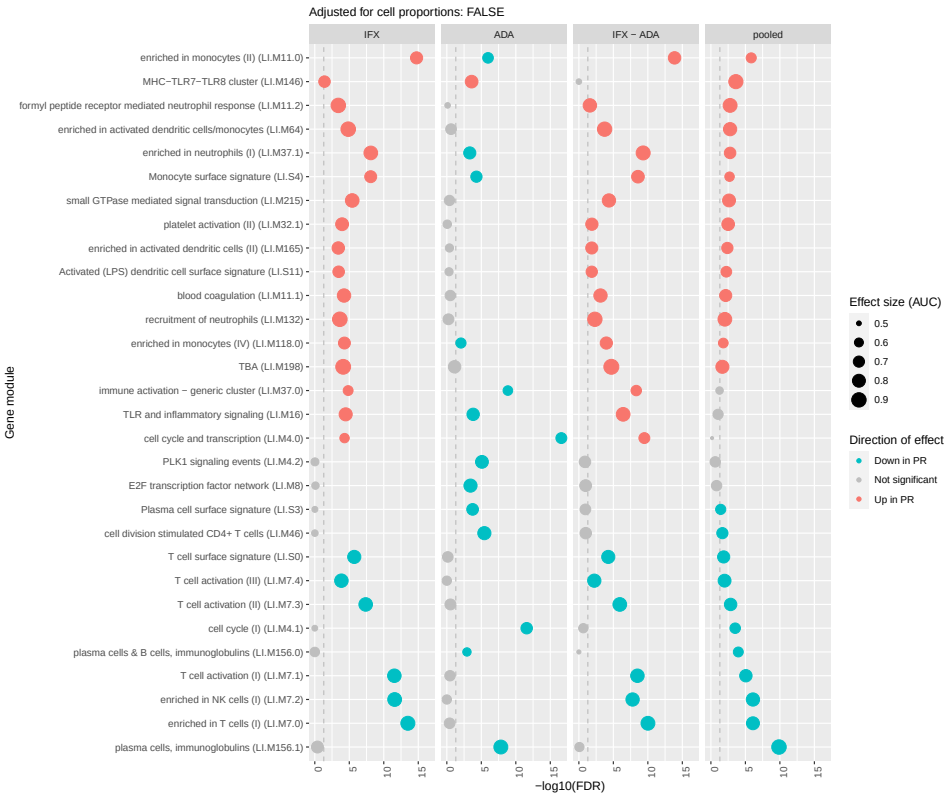

Figure 2.10: Panel plot of module enrichent analysis for PR vs PNR at week 0. Length of bar is effect size, shade is FDR. Blue is upreg in R, red is downreg in R.

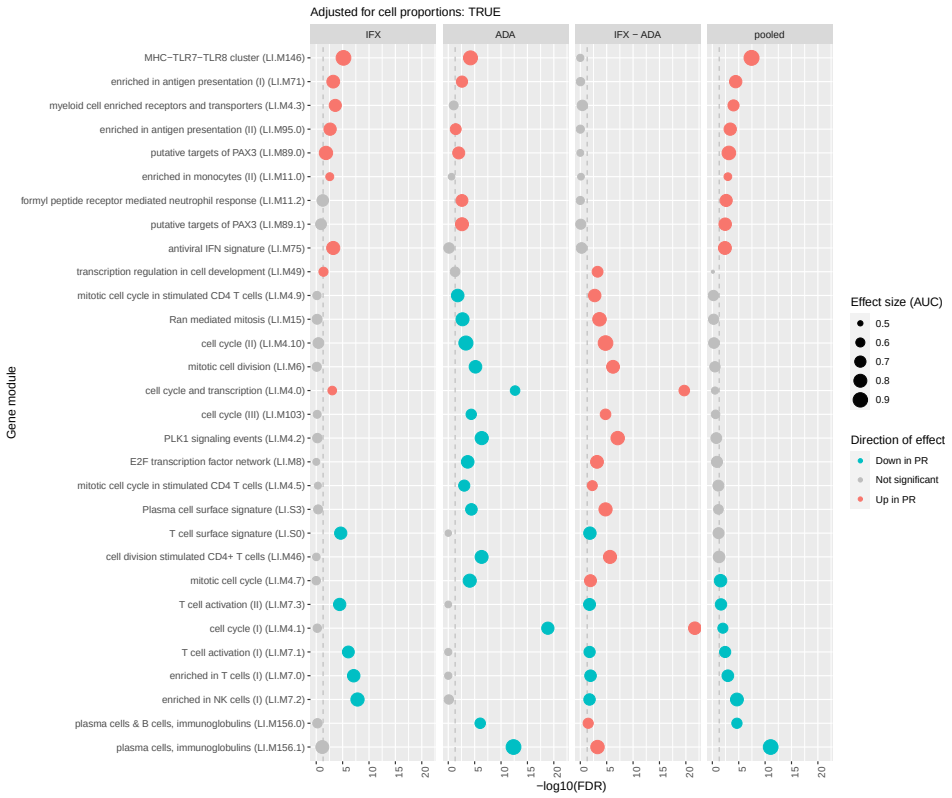

Figure 2.11: Panel plot of module enrichent analysis for PR vs PNR at week 0. Length of bar is effect size, shade is FDR. Blue is upreg in R, red is downreg in R.

FDR=0.986).

2.3.4 Gene expression associated with response post-induction

The same methodology applied at week 0 was applied at week 14 to identify differences in post-induction expression associated with primary response. A larger proportion of the transcriptome is differentially expressed at week 14: 1364 for the infliximab-only comparison, 1544 for the adalimumab-only comparison, and 4841 pooling both drugs (Fig. 2.12). No significant interactions between drug and primary response were detected at the gene-wise level. Given that sample sizes at week 0 and week 14 are comparable (Fig. 2.7), the overall signal is much stronger than at baseline.

Adjusting for cell proportions, 1320/1367, 1515/1544 and 4653/4841 have dampened effects; and the numbers of significant genes drop to 379, 177, and 1302 for infliximab, adalimumab, and pooled analyses respectively. This again suggests many effects are mediated by differences in immune cell composition between primary responders and non-responders.

Modules including generic immune activation, monocytes, TLR and inflammatory signalling, and neutrophils were downregulated in primary responders; whereas B cell and plasma cell modules were upregulated Fig. 2.13. These modules remained differentially expressed with the same direction of effect after adjusting for cell composition Fig. 2.13, suggesting there is per-cell up or downregulation on top of abundance changes of the cell types expressing these modules. Modules related to antigen presentation (LI.M71, LI.M97.0, LI.M5.0) and interferon (LI.M75, LI.M127, LI.M111.1) also appear among significantly upregulated modules after correction. Directions of effect for the most significant modules were largely consistent between drugs, with little evidence for interactions at the module level.

SIGLEC10 from my baseline analysis retains its significant association with primary response post-induction, with the same direction of effect (adjusted logFC=0.366). Some genes previously

highlight a few more individual genes specific to this analysis too? not sure how to pick them at the moment.

proposed as baseline markers of response in gut mucosa: *G0S2*, *TNFAIP6*, *S100A8* and *S100A9* by [93]; and *OSM* by [94], were differentially expressed in post-induction blood in this study. The direction of effect in both cases, downregulation of markers in primary responders, also matches this study.

2.3.5 Magnification of expression change from baseline to post-induction in responders

Given the stronger differences in expression between primary responders and non-responders at week 14 versus week 0, I estimated the change in expression from week 0 to week 14 within the two groups, and also estimated the timepoint by response interaction. I perform only the pooled comparison here both to simplify the analysis, and because similarly to the within week 14 comparison, change from week 0 to week 14 was relatively consistent between drugs, with exceptions noted.

Without adjusting for cell proportions, 12862 genes were differentially expressed in primary responders at week 14 vs week 0 in the pooled analysis, 8310 genes in primary non-responders, and 6320 genes had a significant interaction. After adjusting for cell proportions, 5572 genes were differentially expressed in primary responders, 626 genes in primary non-responders, and 179 genes had a significant interaction. Of the genes differentially expressed between week 14 and week 0 in both primary responders and non-responders, and with a significant interaction between timepoint and response, nearly all (4885/4891 unadjusted for cell composition, 31/32 adjusted) were magnified by primary response, such that the same genes have larger fold-changes in the same direction for primary responders (Fig. 2.15).

The most significant modules that change from week 0 to week 14 in responders included upregulation of B cell (LI.M47.0), plasma cell (LI.M156.0), and T cell activation (LI.M7.1); and downregulation of immune activation (LI.M37.0), monocyte (LI.M11.0), neutrophil (LI.M37.1) and TLR and inflammatory signalling (LI.M16) modules (Fig. 2.16). Many of these are the same modules asso-

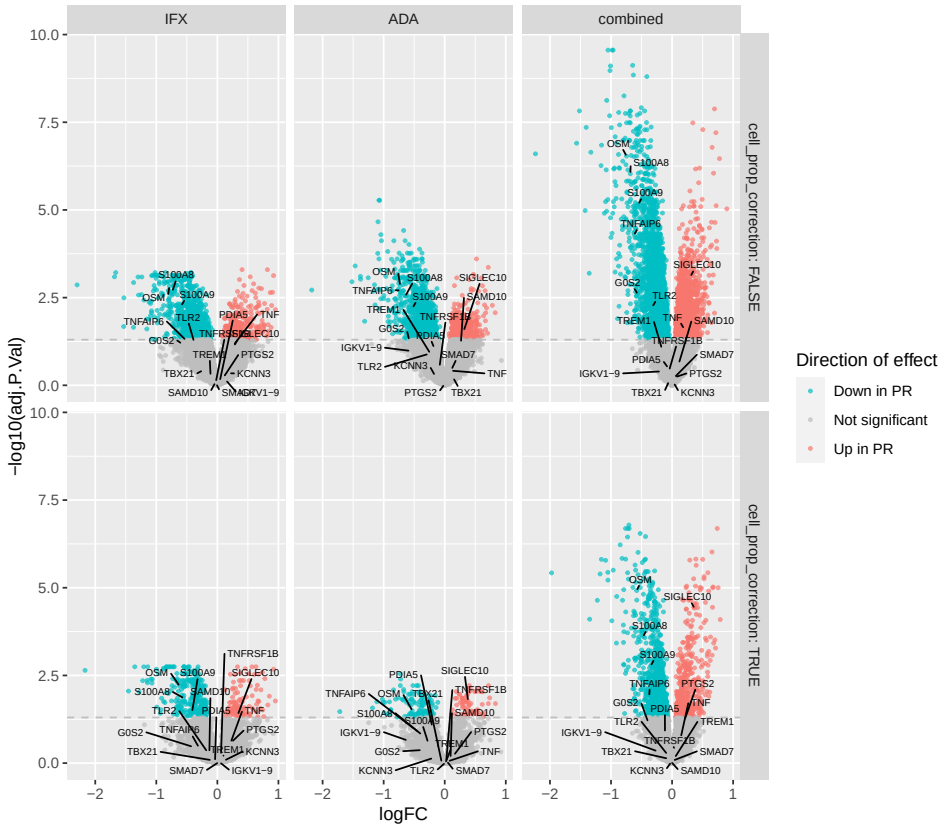

Figure 2.12: DGE volcano plot for PR vs PNR at week 14

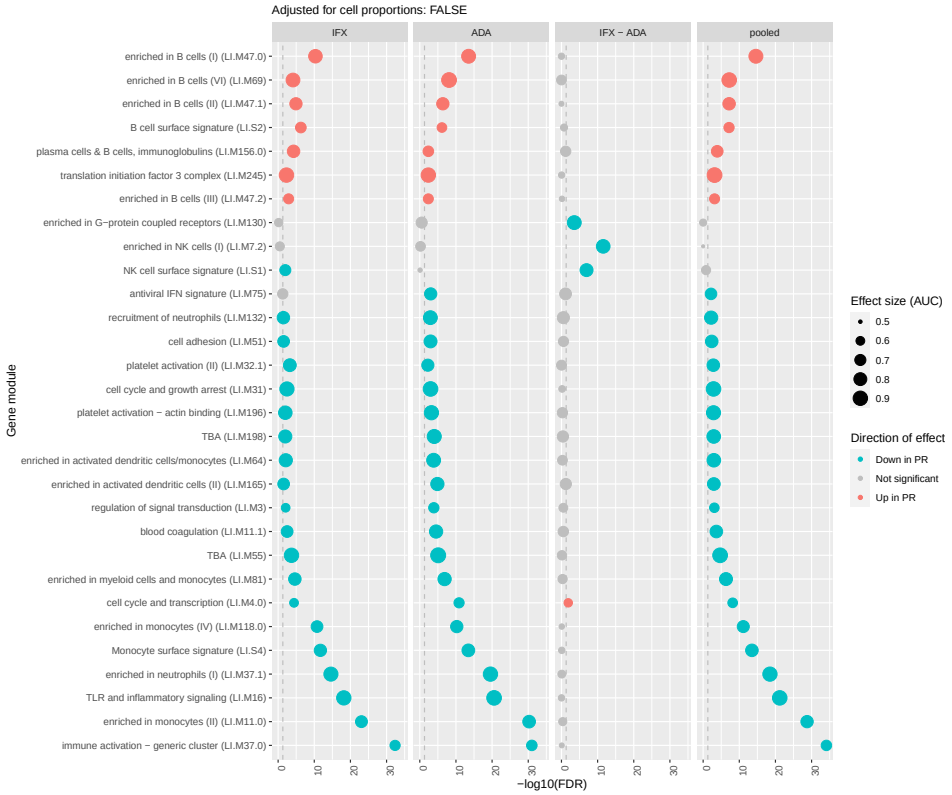

Figure 2.13: Panel plot of module enrichent analysis for PR vs PNR at week 14, adjusted for cell comp. Length of bar is effect size, shade is FDR. Blue is upreg in R, red is downreg in R. top 20 by max

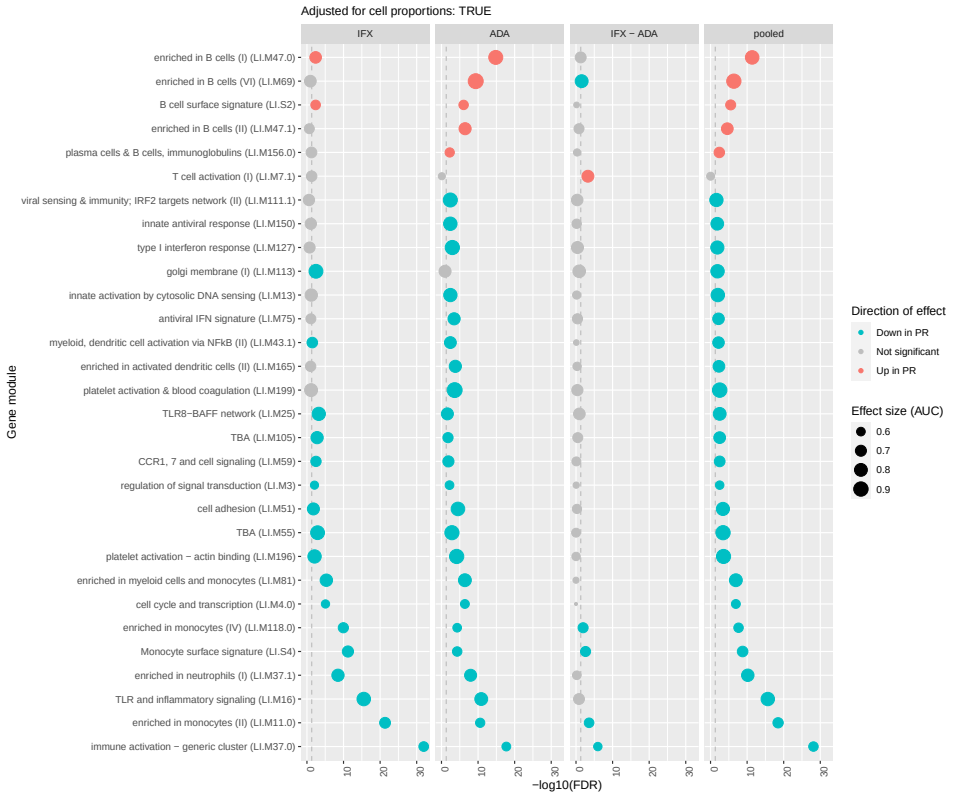

Figure 2.14: Panel plot of module enrichent analysis for PR vs PNR at week 14, adjusted for cell comp. Length of bar is effect size, shade is FDR. Blue is upreg in R, red is downreg in R. top 20 by max

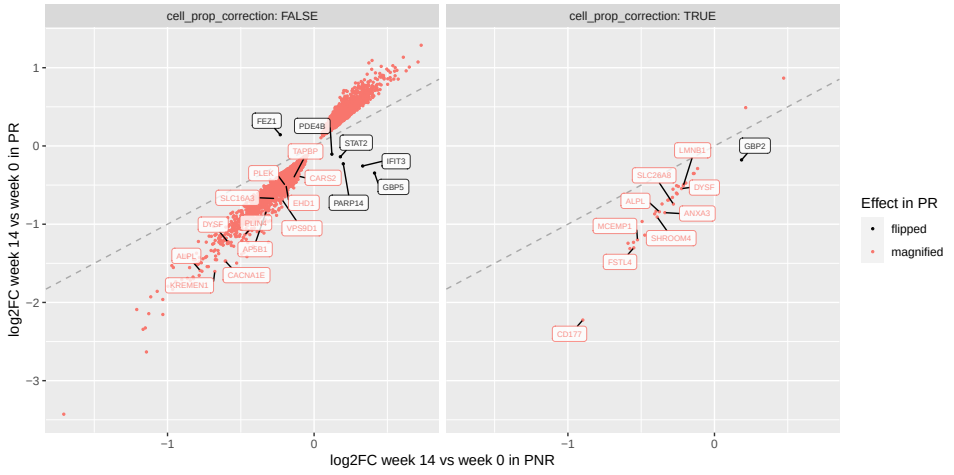

Figure 2.15

ciated with response within timepoints. The differences between primary responders and non-responders at week 14 are qualitatively similar to the differences between primary responders at week 14 and baseline.

Adjusting for cell composition decreases the significance of a majority of modules (Fig. 2.17), with T cell modules in the adalimumab-only analysis especially decreased. Magnification is also observed at the module level, with nearly all module effects aligned in the same direction in responders and non-responders, with significant interactions also in the same direction.

2.3.6 Interferon modules with opposing differential expression in responders and non-responders

Fig. 2.15) also contained genes that were downregulated from week 0 to week 14 in responders, but upregulated in non-responders. Considering modules from [102] in `tmod` (prefixed "DC"), *STAT2*, *GBP5*, and *PARP14* are annotated into an interferon module (DC.M3.4, tmodHGtest FDR=0.000 126). *IFIT3* and *GBP2* are also annotated into interferon modules (DC.M1.2, DC.M5.12). Adjusted for cell composition, these modules are significantly upregulated at week 14 in non-responders only: DC.M3.4 FDR=3.45×10^{-21}, DC.M1.2 FDR=9.49×10^{-16}, DC.M5.12 FDR=1.36×10^{-13} (Fig. 2.18). Similar opposing effects after cell composition adjustment were also observed in [101] modules for antiviral interferon signature (LI.M175), type I interferon response (LI.M127), and antigen presentation (LI.M95.0) (Fig. 2.17).

2.3.7 Sustained expression differences between primary responders and non-responders during maintenance

As PANTS is an observational study, it was able to include some patients who continued with anti-TNF therapy even after meeting the definition of primary non-response at week 14. For both primary responders and non-responders, expression data could

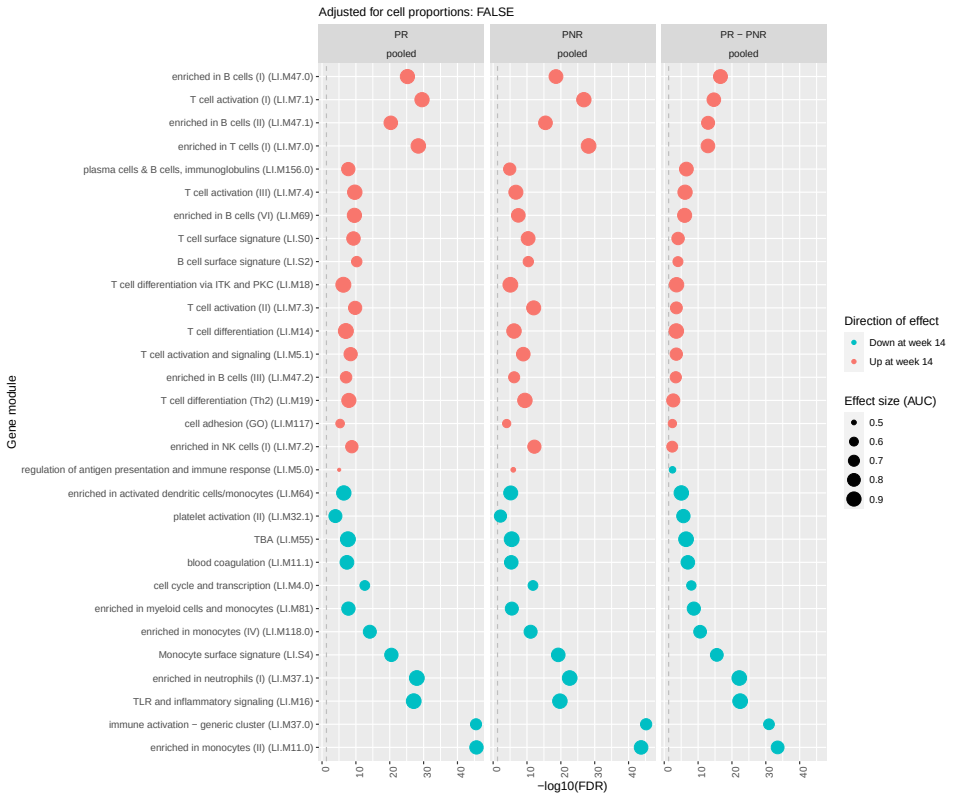

Figure 2.16: Panel plot of module enrichent analysis for PR vs PNR for week 14 - week 0 change. Length of bar is effect size, shade is FDR. Blue is upreg in R, red is downreg in R.

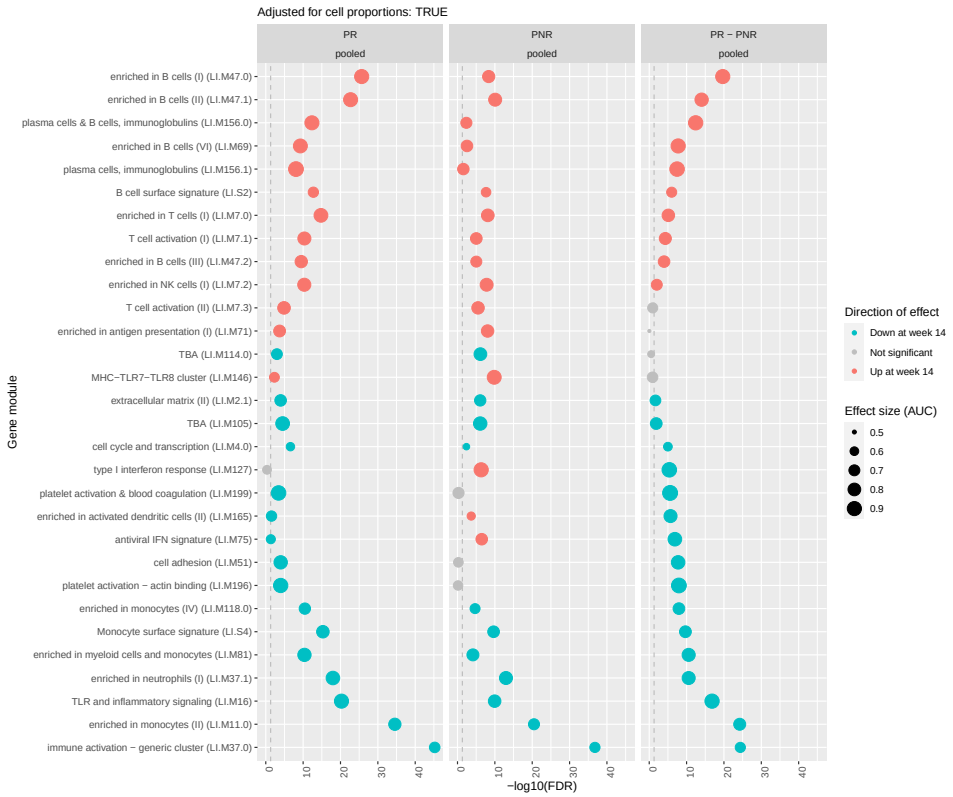

Figure 2.17: Panel plot of module enrichent analysis for PR vs PNR for week 14 - week 0 change. Length of bar is effect size, shade is FDR. Blue is upreg in R, red is downreg in R.

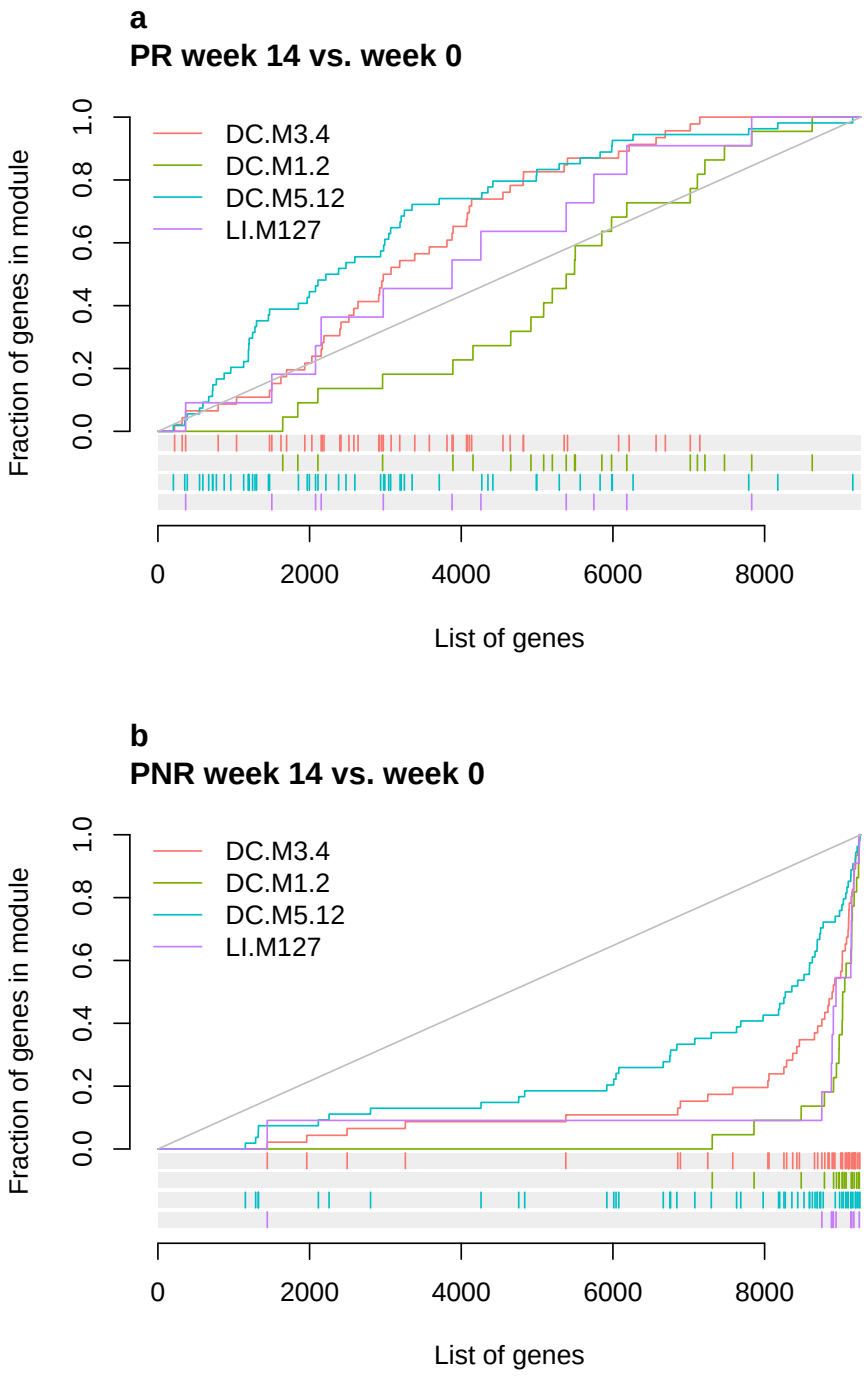
a
PR week 14 vs. week 0
Fraction of genes in module
0.0
0.2
0.4
0.6
0.8
1.0
DC.M3.4
DC.M1.2
DC.M5.12
LI.M127
0
2000
4000
6000
8000
List of genes
b
PNR week 14 vs. week 0
Fraction of genes in module
0.0
0.2
0.4
0.6
0.8
1.0
DC.M3.4
DC.M1.2
DC.M5.12
LI.M127
0
2000
4000
6000
8000
List of genes

Figure 2.18

also be available from blood samples around week 30 and week 54, and at additional visits scheduled in the event of secondary LOR.

I fit a natural cubic spline to the expression of each gene as a function of study day, and tested for general differences in expression over time between primary responders and non-responders. This analysis was done only with drugs pooled due to lower sample sizes for later timepoints. Without adjusting for cell composition, 4426 genes were differentially expressed between primary responders and non-responders; 210 genes were differentially expressed after adjustment. To identify distinct trajectories of expression over time, I hierarchically clustered those 210 genes by their mean expression in responders and non-responders at each timepoint, and determined the optimal number of clusters by the gap statistic method (**??**). Six distinct clusters were proposed (Fig. 2.20).

Many of these genes had previously been identified as having significant differences in expression between responders and non-responders either within week 14, or for change in expression from week 0 to week 14. Cluster 1 contained mainly previously identified genes (Fig. 2.21), and was enriched for modules including myeloid cells and monocytes (LI.M81, hypergeometric test, FDR=2.11×10^{-6}), platelet activation (LI.M196, FDR=1.35×10^{-5}), immune activation (LI.M37.0, FDR=1.44×10^{-4}), and TLR and inflammatory signalling (LI.M16, FDR=2.36×10^{-3}). The spline analysis highlighted that expression differences at week 14 are maintained at week 30 and week 54.

The highest proportion of genes uniquely identified as significant by the spline analysis were in cluster 2 (26/31) and cluster (15/20). Cluster 2 was enriched in [101] B cell modules (LI.M47.0, FDR=1.53×10^{-6}; LI.M47.1, FDR=4.53×10^{-5}) previously identified as having a greater increase from week 0 to week 14 in primary responders versus primary non-responders (Fig. 2.17), matching the observed cluster trajectory. Cluster 4 was not enriched in any modules from Li *et al.* [101], but is enriched for a B cell module (DC.M4.10, FDR=0.001 37) from Chaussabel *et al.* [102]. Although no genes were significantly associated with response at week 0 (Fig. 2.9), the genes in cluster 4 are coordi-

nately downregulated as a set in primary responders (CERNO test, p=6.18×10^{-25}).

Cluster 3 is also of interest, enriched for type I interferon response (LI.M127, FDR=0.005 68) and interferon (DC.M3.4, FDR=0.000 527) modules, as well as genes that contain putative transcription factor binding motifs for interferon regulatory factors *IRF7* (`g:Profiler` term ID TF:M00453_1, adj. p value=0.005 05) and *IRF8* (TF:M11684_1, adj. p value=0.007 77; TF:M11685_1, adj. p=0.0103). The cluster trajectory shows direction of expression change is opposing in responders and non-responders from week 0 to week 14, followed by sustained differences at week 30 and week 54. The trajectory and interferon-related gene set enrichments are consistent with those identified in subsection 2.3.6. Of the 9 genes in this cluster, 8 genes have significant interaction between week 0 to week 14 expression change and response status, whether or not correcting for cell composition. However, only *GBP5* was differentially expressed from week 0 to week 14 in both responders and non-responders, and only when unadjusted for cell composition (Fig. 2.15). This indicates that such small and opposite effects in responders and non-responders may only be detected at a single gene level in the interaction analysis where the difference is amplified, and the spline analysis, with the support of additional data from week 30 and week 54.

2.3.8 Limited evidence for change in genetic architecture of gene expression over time

Given the substantial changes in expression from baseline to post-induction,after starting the drug, and the differing trajectories observed in responders and non-responders, I performed expression quantitative trait locus (eQTL) mapping to detect common genetic variants associated with expression. Variants cis (within 1 Mb of the TSS) to 15040 genes were tested for association. Mapping was done within each timepoint (weeks 0, 14, 30, and 54), followed by joint analysis of per-timepoint eQTL summary statistics and control for multiple testing were done using `mashr`.

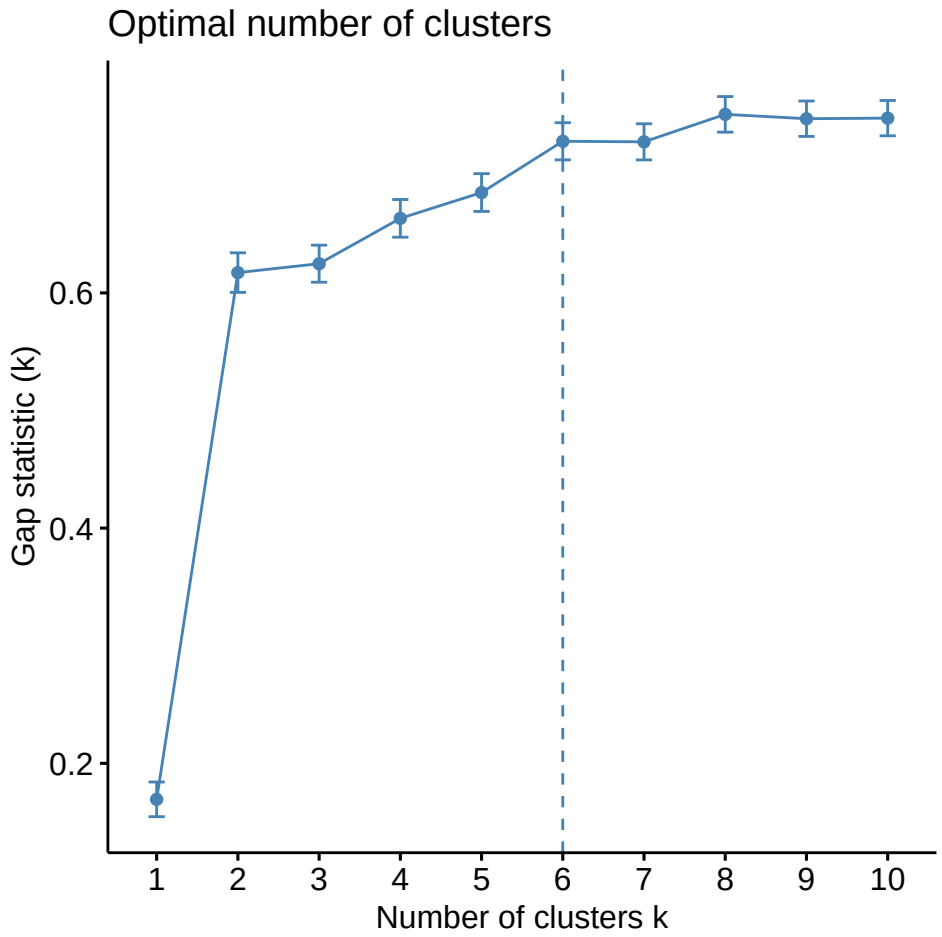
Optimal number of clusters
Gap statistic (k)
0.6
0.4
0.2
1
2
3
4
5
6
7
8
9
10
Number of clusters k

Figure 2.19

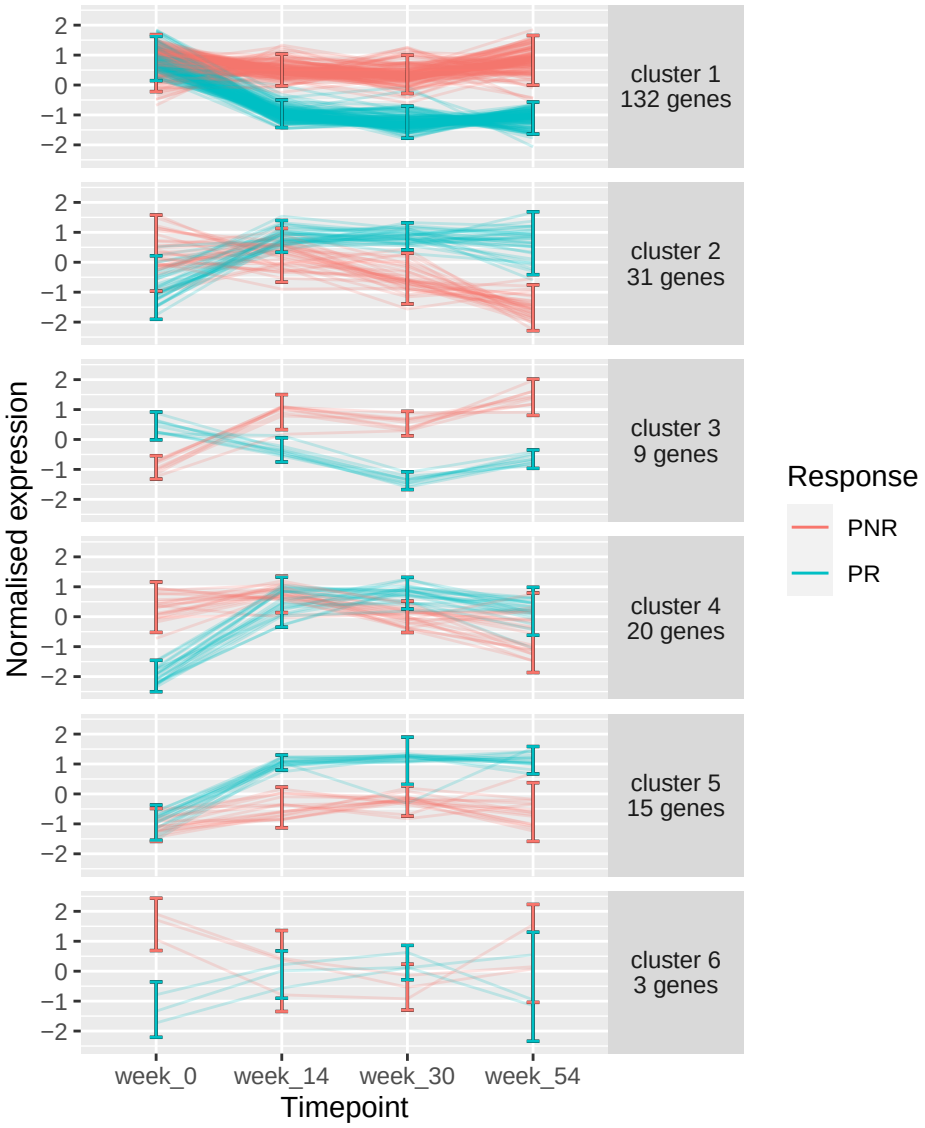
cluster 1
132 genes
cluster 2
31 genes
cluster 3
9 genes
cluster 4
20 genes
cluster 5
15 genes
cluster 6
3 genes
Normalised expression
2
1
0
−1
−2
week_0
week_14
week_30
week_54
Timepoint
Response
PNR
PR

Figure 2.20

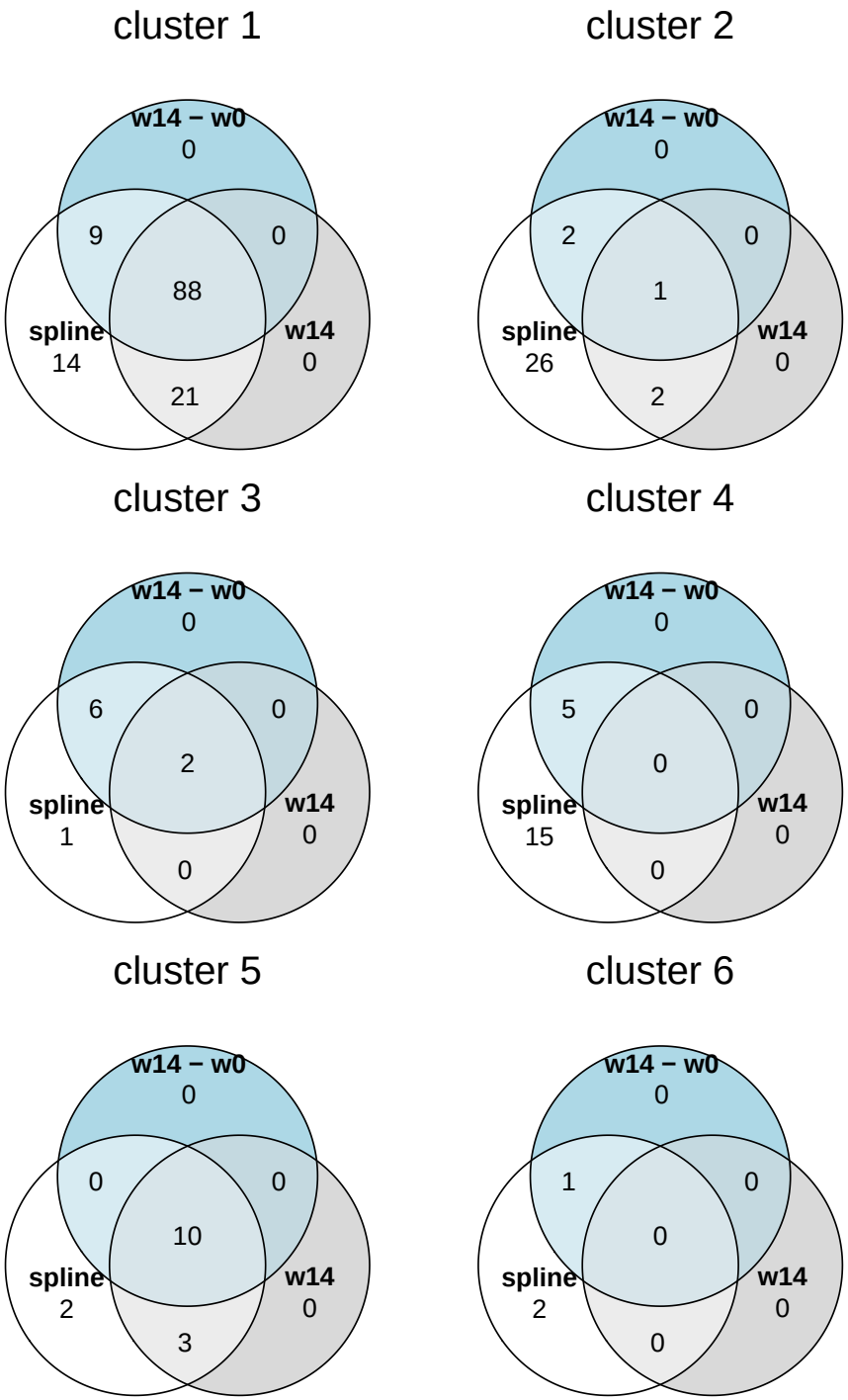
cluster 1
w14 − w0
0
9
0
88
spline
14
w14
0
21
cluster 2
w14 − w0
0
2
0
1
spline
26
w14
0
2
cluster 3
w14 − w0
0
6
0
2
spline
1
w14
0
0
cluster 4
w14 − w0
0
5
0
0
spline
15
w14
0
0
cluster 5
w14 − w0
0
0
0
10
spline
2
w14
0
3
cluster 6
w14 − w0
0
1
0
0
spline
2
w14
0
0

Figure 2.21

The majority 11156/15040 (74.2 %) of genes were eGenes (a gene with at least one significant cis-eQTL) in at least 1 timepoint (lfsr < 0.05). The variant with the lowest lfsr in any timepoint for each gene was chosen as the lead variant (eSNP) for that gene. Most eSNPs are significant in multiple timepoints: 999 significant in 1 timepoint, 381 significant in 2 timepoints, 526 significant in 3 timepoints, 9250 significant in all 4 timepoints. I compared eSNP effect sizes between week 0 and each of weeks 14, 30 and 54. to identify response expression quantitative trait loci (reQTLs) with significant difference in effect versus baseline, and may be associated with changes in expression from baseline. Most eSNPs were shared across timepoints; only six eSNPs-eGene pairs were significant at BH FDR < 0.05: with 1/6 for week 30 versus week 0 and 5/6 for week 54 versus week 0 (Fig. 2.22). Of the six eGenes, *NMI* and *EPSTI1* both had their respective eSNPs having magnified effects on expression at week 54 compared to week 0, and both are annotated to contain putative binding motifs for *IRF8* and *IRF2* (`g:Profiler` term IDs TF:M11685_1 and TF:M11665_1). However, there may be complications of confounding by cell composition for reQTLs in bulk expression data (discussed in **??**).

any need to add w14 to plot?

right word???

2.4 Discussion

In PANTS, a cohort of CD patients receiving infliximab or adalimumab anti-TNF therapy for the first time, there were substantial differences in whole blood gene expression between primary responders and non-responders. At baseline, the greatest differences in expression were observed between future responders and non-responders to infliximab, with increased expression of monocyte, neutrophil and dendritic cell gene modules in responders, and decreased expression of T cell and NK cell modules. These effects appear to be infliximab-specific, and are attenuated after adjusting for the proportions of six major immune cell types, suggesting expression differences may be driven by mediation via the proportions of these cell types.

So many modules associations here, maybe try to Google some of them...

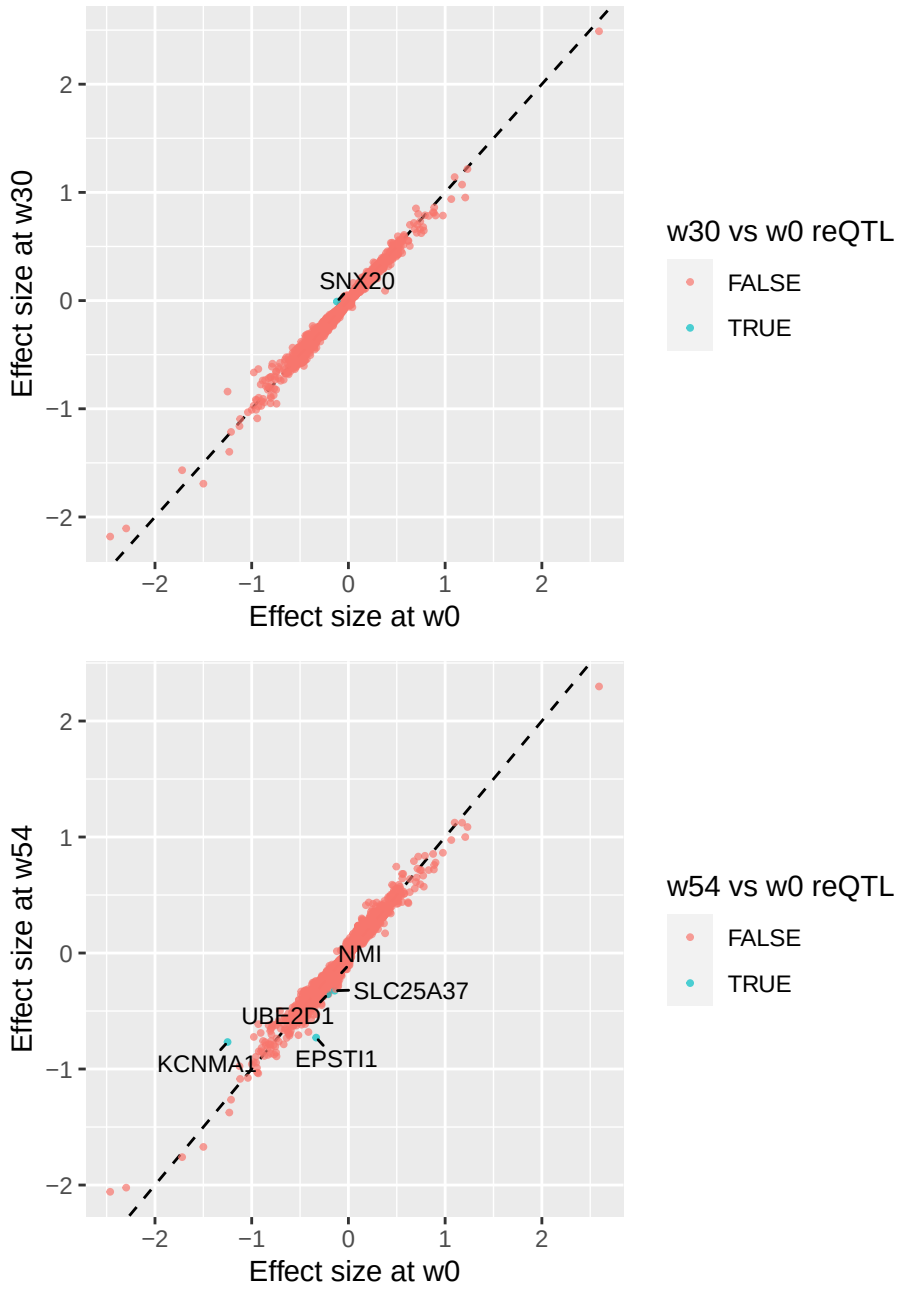

Figure 2.22: Week 30 and week 54 eQTL effect sizes vs baseline. Significant reQTLs in blue.

In contrast, future responders to adalimumab had lower baseline expression of plasma cell and cell division modules. Th module-level results line up with the three gene-wise hits for the adalimumab-only analysis: *IGKV1-9* encodes the immunoglobulin light chain variable region that forms part of antibodies produced by plasma cells, *KCNN3* is annotated to plasma cell surface signature module (LI.S3[101]), and the expression of both *KCNN3* and *PDIA5* are correlated with blood plasmablast frequencies[103]. It was reported by Gaujoux *et al.* [96] that baseline plasma cell abundances are lower for infliximab responders, hypothesising that plasma cell survival is supported by increased TNF levels in non-responders. Plasma cells also formed a part of a correlated module of cell populations identified by [95], where lower module expression was associated with better response to anti-TNF in a cohort with patients taking both infliximab and adalimumab. However, both these studies were done in gut biopsy samples, and there was no report of strong between-drug heterogenenity.

The adalimumab-specific associations I find were more significant after cell proportion adjustment, which may indicate per-cell downregulation rather than cell abundance being associated with response. However, cell composition differences mediated by rarer cell types that have abundances poorly captured by the six major types used in the model will be poorly adjusted for. For example, plasma cell proportions are only weakly correlated with other immune cell types in the healthy immune system[104], although the relationship may differ for CD patients. If this is the case, the role of cell composition estimates for adalimumab-specific may be more akin to precision variables.

The differences between drugs are puzzling, especially the greater effect of cell composition adjustment for infliximab. Baseline patient differences between drugs may offer a partial explanation. In the full PANTS cohort, baseline characteristics other than those in Table 2.1 differ between patients on different drugs[86]. In particular, lower albumin, higher CRP, and higher faecal calprotectin in infliximab patients suggest that they may have had greater disease severity. Differences may be driven by patient or physician

preference, for example, patients with more severe disease are often given infliximab rather than adalimumab*. I have not yet been able to access clinical variables such as CRP and faecal calprotectin levels to consider as variables to adjust for in my modelling. A richer phenotype dataset containing some of these variables has been requested from collaborators.

The strongest single-gene association in the pooled analysis was *SIGLEC10*, which had reduced significance post-adjustment with a comparable effect size, where baseline expression was approximately 25% higher in responders. Direction of effect was consistent between drugs, but most significant in infliximab without cell composition adjustment. In inflammatory bowel disease (IBD), small molecules called damage-associated molecular patterns (DAMPs) are released due to tissue damage and cell death, and further promote inflammation through pathogen sensing pattern recognition receptor (PRR) pathways that include Toll-like receptors (TLR) family receptors[76, 105]. For instance, faecal calprotectin, a marker for IBD activity, is a complex of two DAMPs, S100A8 and S100A9[76]. SIGLEC10 has been shown to repress DAMP-mediated inflammation through binding CD24[105]. *SIGLEC10* is expressed on B cells, monocytes and eosinophils `https://www.nature.com/articles/nri2056`, and of these cell types, module level results posit monocytes as the most likely candidate cell type to have increased module expression in responders. In monocytes, *SIGLEC10* expression is more specific to the CD16+ monocytes[106], and in particular the CD14+CD16++ non-classical monocytes rather than the classical CD14++CD16- or intermediate CD14++CD16+ subsets[107]. In PANTS, it was suggested by Kennedy *et al.* [86] that higher inflammatory load as indicated by low baseline albumin levels may result in low week 14 drug levels due to faster drug clearance, and low drug levels at week 14 were in turn associated with non-response. A hypothetical model might be imagined where high baseline *SIGLEC10* expression reflecting higher proportions of CD16+ monocytes (or lower proportions of CD16- monocytes), decreased DAMP-mediated

*Kennedy, N. A., personal communication, 4 June (2020).

inflammation, and increased chance of primary response, possibly by affecting drug clearance rate. This is an extremely tentative model: both the cell proportion estimates and module definitions used thus far only represent monocytes as a whole, lacking the resolution to properly explore shifts in the three monocyte subsets. It may be possible to use expression of monocyte subset marker genes such as those identified by Villani *et al.* [107] to improve the resolution of the cell proportion estimates.

Despite the strong heterogenenity in effects between drugs, one consistent effect that emerged after adjusting for cell composition was baseline upregulation of MHC-TLR7-TLR8, antigen presentation, and interferon modules in responders. As mentioned above, TLR receptors are involved in pathogen sensing, and TLR7 and TLR8 are endosomal proteins primarily expressed in monocytes, macrophages and dendritic cells (DCs), part of an antigen presentation pathway that senses bacterial DNA and activates downstream innate immune pathways including type I interferon response https://www.nature.com/articles/cmi201238. Type I interferons have pathogenic or protective roles in many immune-mediated inflammatory diseases (IMIDs)https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4084561/. It has been suggested that type I interferon responses induced via TLR7 and TLR8 can suppress colitis in mouse models, and play a role in maintaining gut homeostasis https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5797585/ https://www.frontiersin.org/articles/10.3389/fmed.2018.00032/full, so upregulation here may again represent a less severe baseline disease in future responders.

How does the MHC come in here?

I also assessed the previously reported association of *TREM1* expression in blood with anti-TNF response. Gaujoux *et al.* [96] found *TREM1* expression to be lower in infliximab responders in gut biopsies, but higher in responders in a separate whole blood cohort. Verstockt *et al.* [97] also reported *TREM1* to be a marker of response, with lower expression in responders to infliximab and adalimumab in both gut biopsies and blood. I did not find *TREM1* to be significantly differentially expressed in PANTS. although the direction of effect is increased expression in

responders— matching the Gaujoux *et al.* [96] direction of effect in blood— *TREM1* is expressed on myeloid lineage cells such as monocytes and macrophages; Villani *et al.* [107] reported that *TREM1* expression is most specific to classical monocytes and a newly identified subtype within the intermediate monocytes ("Mono3"). The *TREM1* effect is one of the infliximab-specific differentially expressed genes that is much stronger without cell proportion adjustment, so it may reflect association of baseline monocyte cell proportions with response.

Most previously reported baseline markers in blood and gut biopsies were non-significant in this study. For gut markers, this may not be unexpected. Although a subset of gut infiltrating immune cells and their precursors may also be circulating, genes specific to epithelium and some immune cell types as monocyte-derived macrophages that differentiate after they migrate into tissues, will be difficult to observe in PANTS.

a consensus for any marker seems far from established

Gaujoux *et al.* [96] note that changes in cell abundance from baseline to post-treatment occur to a greater extent in responders.

richness of dataset although other mediators of NR could be modelled using genetic instruments e.g drug level

- blood, not gut

  final: even if it turns out that the between-drug heterogenenity in this study is resolved by adjusting for one or even starker differences are likely to be found between studies Many studies consider anti-TNF drugs as a single class, aggregating cohorts for infliximab, adalimumab and other biologic anti-TNF drugs. due to the difficulty of large cohorts.

- cell prop correction

  * overall, adjusting for cc drops number of DGE, and dampens effects, rather than increase, so mediation rather than precision?
  * if the asusmption of mediation is believed, then two complementary interpretations from adjusted and non

    - in general, more consistentcy after correction, but probs due to inflix more reliant on cell props
    - highlight that cell comp makes a big diff: likely mediation
    - 6 main cell types corrected, but doesn't mean that's enough for rare types e.g. see https://www.biorxiv.org/content/10.1101/2020.05.28.120600v1
    - and did not sep out effect cell count modification on eQTL effect (e.g. recruitment vs stimulation)
    - need further interaction models like in ch3
      Gaujoux *et al.* [96] noted that adjusting expression for cell composition resulted in gene signatures that were worse at discriminating responders from non-responders.
      even within study, far from resolved
      Even for the there are different
      Even within

- replication of known baseline signatures
  - potential differences in clinical vs endoscopic endpoint between all 3 studies
    similar to drug het, between study het in fx could be due to same reasons also subject to differences in the definitions of primary response also all cohorts are small.
  - the replication is sensitive to covariates, end points, drug
    - e.g. Verstockt *et al.* [97] did not use any covariates (i.e. t test), so they report an unadjusted effect
      Whether there are robust baseline markers The interpretation is muddied by differences between studies Additional to the PANTS samples differing in the characteristics described for differences between drugs, which also applies in a between-studies context, there are differences in the definition of primary response.
      baseline still hard for anti-tnf disrep between us and existing even within study between drugs
      further complicated gut vs blood
      It is hoped that module associations such as will be explored further in blood

then, in terms of prediction, eskew modules and measure cell counts of relevant populations??
changes in large propostions of the transcriptome main var explained by anti-TNF drug effect baseline diffs are diluted hence more consistent
although IFN up for R at baseline changes more for NR to w14

- at week 14, difference in transcriptome between R/NR very distinct
  - generally more consistent between drugs
    - e.g. ADA-specific B cell downreg in PR effect seems to vanish in the tmod results at week 14
  - top hit KREMEN1 is part of an inflammatory apoptotic pathway https://academic.oup.com/ibdjournal/article/14/suppl_1/S4/4653822, makes sense that it is downreg in responders
  - modules: innate (monocyte/TLR and inflam) down in R. makes sense
  - modules: T and B cells up in R — why???
- went from few differences at w0 to many differences at w14 between R and NR
  - looking at change from baseline to week 14 confirmed mostly magnifying effects in R
    - could this suggest that there is a continuum of response?
  - the ones that are not consistent: what are they and could they give insights?
  - spline analysis confirmed that the diff starting at w14 is maintained at w30 and w54
  - [86]: "Continuing standard dosing regimens after primary non-response was rarely helpful; only 14 (12 · 4% [95% CI 6 · 9–19 · 9]) of 113 patients entered remission by week 54."
    - may be reflected in the transcriptomics too
  - attrition or loss to followup bias for reqtl effect

- reQTL mapping to identify changes in genetic control of expression over the timepoints
- <more limitations: internal validity> There are several threats to the internal validity of the study.
  - a key question for interpreting both the DGE and reQTL results is the definition of visit and study day
    - arguably, time is only meaningful wrt to drug naive/on drug, and time since last dose and to next dose i.e. everyone is at trough drug levels
    - real drug levels over time will be peaky
    - I included LOR and EXIT visits based on a time window, but in reality, patients that have more LOR have
    - It should be noted that one option following patient loss of response is dose escalation, so samples from the same patient *after* recorded loss of response may actually represent a different trough drug level to the rest.
  - * why R change E, but NR don't? faster clearance? adjust for drug level to see this
    - but I did not use trough drug level measured on or near the same study day as a covariate
    - this would explain more variance, but more missingness, overall 319/840 missing corresponding drug levels, cuts n considerably
    - also, difficulty in pooled drug analyses
    - differ in their pharmokinetics peakiness, IFX has a shorter half life and dose less often [69]
    - may need to fit completely drug strat models, that forgo pooling on all params
    - or some clever non-parametric normalisation method
  - was pooling drugs ok?
    - included these as covariates, not all avail though
    - residual confounding may be an issue since this is uncontrolled, unrandomised

    - drug prescribe diffs
  - stat issues with missingess: MCAR assumption violated

- <even more limitations: external validity>
  - PNR definition
    - its a very complex binary, but certainly useful
    - kennedy2019PredictorsAntiTNFTreatment Univariable analysis showed, for both drugs, that the most significant determinant of non-remission at week 54 was clinical status at week 14 (table 4; appendix pp 21–22).
    - DGE analysis also agrees with kennedy2019PredictorsAntiTNFTreatment: once PNR, no point in continuing
    - continuum of PNR? perhaps model split pheno?

- potential source of multiomics data for potential validation is [108], contains drug response phenotypes
  - also, validate using protome/serological data in PANTS, although there is a known disconnect between proteome and transcriptome

- <reQTL>
  - only 6 reQTLs at per-comparison FDR 0.05
  - 2 main patterns of reQTLs over time
  - change from baseline expected, but no enrichments
  - but can only speculate on why INF-stimulated genes had change in E from baseline to w54 genetically controlled
  - biases
    - winner's curse caused by combo of low power and a signif threshold?
      since E is diff between R and NR, why not put in an interaction between R and NR, not sure if powered not sure if estmating the interaction effect would require more power than losing the averaging effect
  - encourage ASE for validation like for ch3, gutierrez-arcelus2020AllelespecificExpressionChanges, to check hits are not artifacts of my pipeline

  - in summary, little evidence for interaction of eQTL with anti-TNF useage
    * although need more conditional modelling etc.
    * although there may be some disease-specific eQTLs, would need to check het of effect vs healthy controls with suitable cell composition control though
    * future: add interaction for response status, since no change of E in NR may dilute signal for reQTL, but it is not clear if this would boost power, or decrease due to dicing
  - given issues about reQTL in bulk discussed in ch3, do not pursue this, but outline future solutions in ch5 also drugs pooled due to much more power for eqtl required than DGE, may be inappropriate

- <conclusion of how the field has moved forward, and future work>
  - DGE at SIGLEC10 and CROCC2 in baseline whole blood, consistent in both drugs
  - ADA-specific plasma cell signature
  - limited evidence for strong reQTL effects, so i have not attempted to evaluate further
  - finally, as with results from any single study, future validation needed to generalise outside this cohort, to other CD cohorts, and to IBD and relevance to other IMIDs if any
  - given evidence of DGE between R/NR, and the presence of eQTLs, can begin to conceive of potential causal mechanisms
  - and also, how to translate from inference into the language of prediction models (e.g. sensitivity/spec)

- how to move to causal inference + prediction further discussed in ch5, due to overlap with ch3
  - deliberate avoidance of 'signature'

Table 2.1: Table caption.

	ADA	IFX	pooled	p-value
Sex				0.317
(Col %)				Fisher exact
FEMALE	78 (48.4%)	89 (54.6%)	167 (51.5%)	
MALE	83 (51.6%)	74 (45.4%)	157 (48.5%)	
Age of onset (years)				0.774
Mean (SD)	33.3 (15.4)	32.8 (15.3)	33.1 (15.3)	Wilcoxon rank-sum
Missing	0	0	0	
Disease duration (years)				0.546
Mean (SD)	6.1 (8.1)	5.9 (7.7)	6.0 (7.9)	Wilcoxon rank-sum
Missing	0	0	0	
Smoking status				0.263
(Col %)				Fisher exact
Current	28 (17.4%)	36 (22.1%)	64 (19.8%)	
Ex	55 (34.2%)	43 (26.4%)	98 (30.2%)	
Never	78 (48.4%)	84 (51.5%)	162 (50.0%)	
Crohn's-related surgery				0.549
(Col %)				Fisher exact
FALSE	114 (70.8%)	110 (67.5%)	224 (69.1%)	
TRUE	47 (29.2%)	53 (32.5%)	100 (30.9%)	
On immunomodulator ever				0.543
(Col %)				Fisher exact
FALSE	23 (14.3%)	28 (17.2%)	51 (15.7%)	
TRUE	138 (85.7%)	135 (82.8%)	273 (84.3%)	
On immunomodulator at baseline				0.912
(Col %)				Fisher exact
FALSE	79 (49.1%)	81 (49.7%)	160 (49.4%)	
TRUE	82 (50.9%)	82 (50.3%)	164 (50.6%)	
On corticosteroids at baseline				0.011
(Col %)				Fisher exact
FALSE	113 (70.2%)	92 (56.4%)	205 (63.3%)	
TRUE	48 (29.8%)	71 (43.6%)	119 (36.7%)	
Baseline BMI				0.237
Mean (SD)	25.2 (6.2)	24.3 (5.5)	24.8 (5.9)	Wilcoxon rank-sum
Missing	0	0	0	
Primary response status				0.263
(Col %)				Fisher exact
Primary non-response	76 (47.2%)	66 (40.5%)	142 (43.8%)	
Primary response	85 (52.8%)	97 (59.5%)	182 (56.2%)	
CD8+ T cell (%)				0.380
Mean (SD)	2.8 (4.2)	2.8 (5.2)	2.8 (4.7)	Wilcoxon rank-sum
Missing	38	18	56	
CD4+ T cell (%s)				0.752
Mean (SD)	9.2 (6.3)	9.2 (6.8)	9.2 (6.5)	Wilcoxon rank-sum
Missing	38	18	56	
B cell (%s)				0.094
Mean (SD)	1.9 (2.0)	1.5 (1.9)	1.7 (1.9)	Wilcoxon rank-sum
Missing	38	18	56	
Monocyte (%s)				0.497
Mean (SD)	8.9 (3.5)	9.2 (3.7)	9.0 (3.6)	Wilcoxon rank-sum
Missing	38	18	56	
NK cell (%s)				0.683
Mean (SD)	1.9 (3.2)	1.9 (3.8)	1.9 (3.5)	Wilcoxon rank-sum
Missing	38	18	56	
Granulocyte (%s)				0.911
Mean (SD)	74.3 (9.7)	74.3 (10.8)	74.3 (10.3)	Wilcoxon rank-sum
Missing	38	18	56	

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 1 Year. *Cell Reports* **32,** 107923. doi:10.1016/j.celrep.2020.107923 (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. Lichtenstein, G. R. Comprehensive Review: Antitumor Necrosis Factor Agents in Inflammatory Bowel Disease and Factors Implicated in Treatment Response. *Therapeutic Advances in Gastroenterology* **6,** 269–293. doi:10.1177/1756283X13479826 (2013).

70. Kalliolias, G. D. & Ivashkiv, L. B. TNF Biology, Pathogenic Mechanisms and Emerging Therapeutic Strategies. *Nature Reviews Rheumatology* **12,** 49–62. doi:10.1038/nrrheum.2015.169 (2016).

71. 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).

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

73. Cotsapas, C. & Hafler, D. A. Immune-Mediated Disease Genetics: The Shared Basis of Pathogenesis. *Trends in Immunology* **34,** 22–26. doi:10.1016/j.it.2012.09.001 (2013).

74. David, T., Ling, S. F. & Barton, A. Genetics of Immune-Mediated Inflammatory Diseases. *Clinical & Experimental Immunology* **193,** 3–12. doi:10.1111/cei.13101 (2018).

75. Ananthakrishnan, A. N. Epidemiology and Risk Factors for IBD. *Nature Reviews Gastroenterology & Hepatology* **12,** 205–217. doi:10.1038/nrgastro.2015.34 (2015).

76. De Souza, H. S. P. & Fiocchi, C. Immunopathogenesis of IBD: Current State of the Art. *Nature Reviews Gastroenterology & Hepatology* **13,** 13–27. doi:10.1038/nrgastro.2015.186 (2016).

77. 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).

78. Jostins, L. *et al.* Host–Microbe Interactions Have Shaped the Genetic Architecture of Inflammatory Bowel Disease. *Nature* **491,** 119–24. doi:10.1038/nature11582 (2012).

79. Liu, J. Z. *et al.* Association Analyses Identify 38 Susceptibility Loci for Inflammatory Bowel Disease and Highlight Shared Genetic Risk across Populations. *Nature Genetics* **47,** 979–986. doi:10.1038/ng.3359 (2015).

80. Kaplan, G. G. The Global Burden of IBD: From 2015 to 2025. *Nature Reviews Gastroenterology & Hepatology* **12,** 720–727. doi:10.1038/nrgastro.2015.150 (2015).

81. Alatab, S. *et al.* The Global, Regional, and National Burden of Inflammatory Bowel Disease in 195 Countries and Territories, 1990–2017: A Systematic Analysis for the Global Burden of Disease Study 2017. *The Lancet Gastroenterology & Hepatology* **5,** 17–30. doi:10.1016/S2468-1253(19)30333-4 (2020).

82. 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).

83. Aggarwal, B. B. Signalling Pathways of the TNF Superfamily: A Double-Edged Sword. *Nature Reviews Immunology* **3,** 745–756. doi:10.1038/nri1184 (2003).

84. Digby-Bell, J. L., Atreya, R., Monteleone, G. & Powell, N. Interrogating Host Immunity to Predict Treatment Response in Inflammatory Bowel Disease. *Nature Reviews Gastroenterology & Hepatology.* doi:10.1038/s41575-019-0228-5 (2019).

85. 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).

86. 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).

87. D'Haens, G. R. *et al.* The London Position Statement of the World Congress of Gastroenterology on Biological Therapy for IBD With the European Crohn's and Colitis Organization: When to Start, When to Stop, Which Drug to Choose, and How to Predict Response?: *American Journal of Gastroenterology* **106,** 199–212. doi:10.1038/ajg.2010.392 (2011).

88. Ding, N. S., Hart, A. & De Cruz, P. Systematic Review: Predicting and Optimising Response to Anti-TNF Therapy in Crohn's Disease - Algorithm for Practical Management. *Alimentary Pharmacology & Therapeutics* **43,** 30–51. doi:10.1111/apt.13445 (2016).

89. Kopylov, U. & Seidman, E. Predicting Durable Response or Resistance to Antitumor Necrosis Factor Therapy in Inflammatory Bowel Disease. *Therapeutic Advances in Gastroenterology* **9,** 513–526. doi:10.1177/1756283X16638833 (2016).

90. Flamant, M. & Roblin, X. Inflammatory Bowel Disease: Towards a Personalized Medicine. *Therapeutic Advances in Gastroenterology* **11,** 1756283X1774502. doi:10.1177/1756283X17745029 (2018).

91. Noor, N. M., Verstockt, B., Parkes, M. & Lee, J. C. Personalised Medicine in Crohn's Disease. *The Lancet Gastroenterology & Hepatology* **5,** 80–92. doi:10.1016/S2468-1253(19)30340-1 (2020).

92. Arijs, I. *et al.* Mucosal Gene Signatures to Predict Response to Infliximab in Patients with Ulcerative Colitis. *Gut* **58,** 1612–1619. doi:10.1136/gut.2009.178665 (2009).

93. Arijs, I. *et al.* Predictive Value of Epithelial Gene Expression Profiles for Response to Infliximab in Crohn's Disease. *Inflammatory Bowel Diseases* **16,** 2090–2098. doi:10.1002/ibd.21301 (2010).

94. West, N. R. *et al.* Oncostatin M Drives Intestinal Inflammation and Predicts Response to Tumor Necrosis Factor–Neutralizing Therapy in Patients with Inflammatory Bowel Disease. *Nature Medicine* **23,** 579–589. doi:10.1038/nm.4307 (2017).

95. Martin, J. C. *et al.* Single-Cell Analysis of Crohn's Disease Lesions Identifies a Pathogenic Cellular Module Associated with Resistance to Anti-TNF Therapy. *Cell* **178,** 1493–1508.e20. doi:10.1016/j.cell.2019.08.008 (2019).

96. 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).

97. 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).

98. Salvador-Martín, S. *et al.* Gene Signatures of Early Response to Anti-TNF Drugs in Pediatric Inflammatory Bowel Disease. *International Journal of Molecular Sciences* **21,** 3364. doi:10.3390/ijms21093364 (2020).

99. 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).

100. 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).

101. 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).

102. Chaussabel, D. *et al.* A Modular Analysis Framework for Blood Genomics Studies: Application to Systemic Lupus Erythematosus. *Immunity* **29,** 150–164. doi:10.1016/j.immuni.2008.05.012 (2008).

103. 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).

104. Zalocusky, K. A. *et al.* The 10,000 Immunomes Project: Building a Resource for Human Immunology. *Cell Reports* **25,** 513–522.e3. doi:10.1016/j.celrep.2018.09.021 (2018).

105. Boyapati, R. K., Rossi, A. G., Satsangi, J. & Ho, G.-T. Gut Mucosal DAMPs in IBD: From Mechanisms to Therapeutic Implications. *Mucosal Immunology* **9,** 567–582. doi:10.1038/mi.2016.14 (2016).

106. Martinez, F. O. The Transcriptome of Human Monocyte Subsets Begins to Emerge. *Journal of Biology* **8,** 99. doi:10.1186/jbiol206 (2009).

107. Villani, A.-C. *et al.* Single-Cell RNA-Seq Reveals New Types of Human Blood Dendritic Cells, Monocytes, and Progenitors. *Science* **356,** eaah4573. doi:10.1126/science.aah4573 (2017).

108. 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).

List of Abbreviations

bp base pair

BTM blood transcription module

CD Crohn's disease

CRP C-reactive protein

DAMP damage-associated molecular pattern

DC dendritic cell

DGE differential gene expression

eQTL expression quantitative trait locus

FWER family-wise error rate

GWAS genome-wide association study

HBI Harvey Bradshaw index

IBD inflammatory bowel disease

IMID immune-mediated inflammatory disease

LD linkage disequilibrium

MAF minor allele frequency

ML maximum likelihood

PANTS Personalised Anti-TNF Therapy in Crohn's Disease

PNR primary non-response

PRR pattern recognition receptor

QTL quantitative trait locus

REML restricted maximum likelihood

reQTL response expression quantitative trait locus

RNA-seq RNA-sequencing

SNP single nucleotide polymorphism

TF transcription factor

TIV trivalent inactivated influenza vaccine

TLR Toll-like receptors

TNF tumour necrosis factor

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