<title>

Benjamin Yu Hang Bai

2020-08-20 01:38:31+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

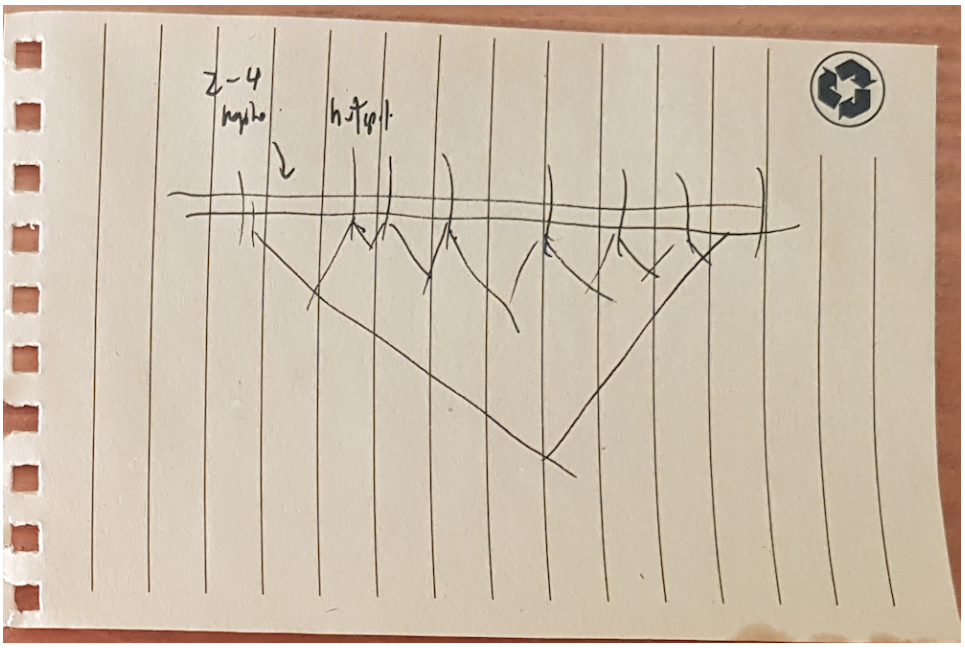


Figure 1.1: The genomic mosaic: block-like LD structure of the genome

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

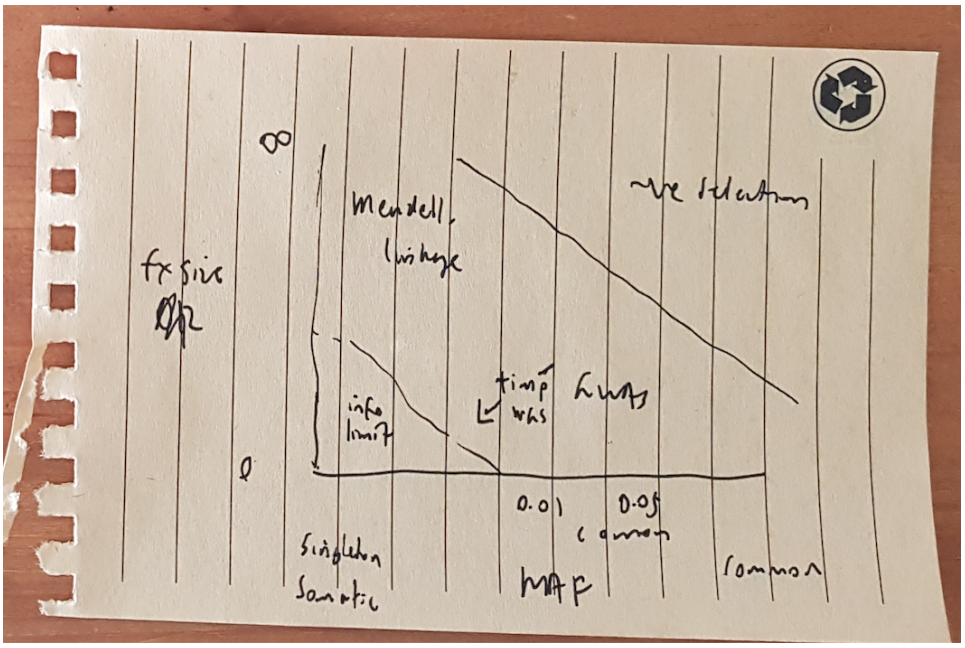


Figure 1.2: The reach of GWAS. OR vs MAF ala tam2019BenefitsLimitationsGenomewide, extended by imputation, sample size, WGS based genotypes, but may be indistinguishable from noise at the limits

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]

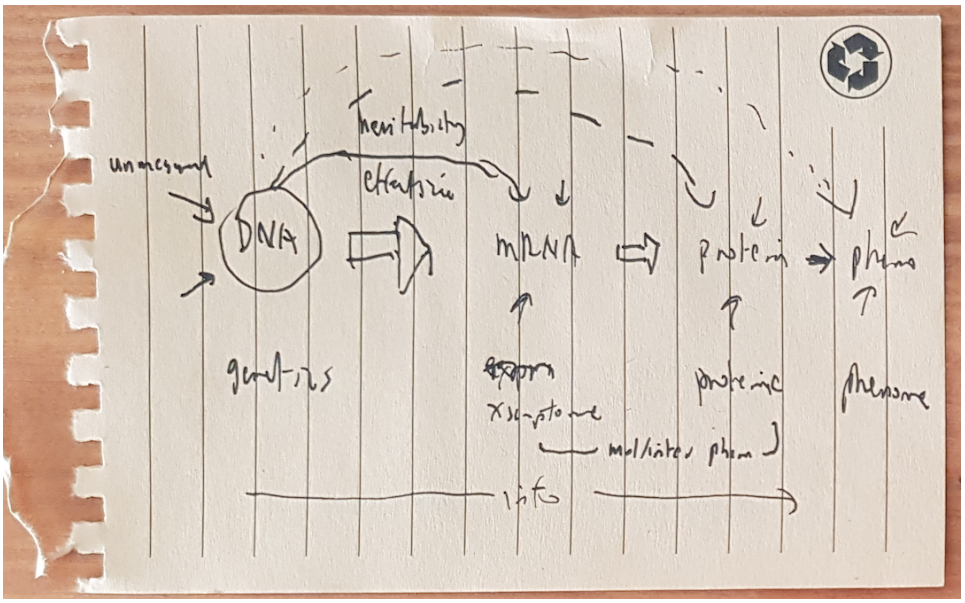


Figure 1.3: Mediation of genetic effect to phenotype, through the biological system

    - 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

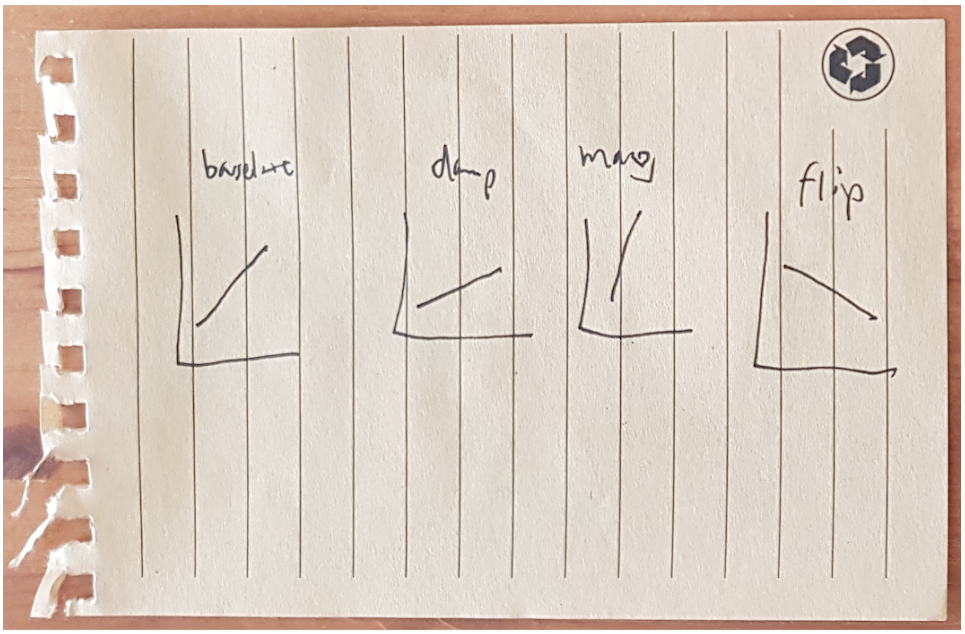


Figure 1.4: eqtl mech models: magnify, dampen, flip

  - 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.
  - 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. [69]
  - 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. [70, 71]

- Pathogenesis of CD is not completely understood, but involves interaction of the immune system, environmental factors (e.g. smoking, stress, diet [69, 72]), and gut microbial factors in a genetically-susceptible individual [73].
  - 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 [74]
  - Genetic correlation between UC and CD is high [70, 71]
  - 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 [75, 76]

add NOD2 OR

  - Concordance rates amongst monozygotic twins are higher for CD (50%) than for UC (∼15%) suggesting a greater heritable component [69]

- IBD has historically been considered a disease of the Western world.
  - The disease burden is now rising globally [77, 78].
  - The highest prevalence and incidence of new cases of CD are in North America and western Europe. [69]
  - 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. [69]
  - 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 [69]
  - 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 mucosal healing [69]

2.1.2 Anti-TNF biologic therapies for CD

- The use of anti-tumour necrosis factor (anti-TNF) therapy has revolutionised CD and IBD patient care in the last two decades.
  - "The pivotal role of TNF in IBD" [79]
  - The two major agents used for CD are adalimumab and infliximab IgG1 monoclonal antibodies that target TNF alpha (a proinflammatory cytokine) [80]
    - also indicated for many other IMIDs e.g. rheumatoid arthritis, Psoriasis, ankylosing spondylitis. [81, 82]
  - They work by binding to both soluble and transmembrane forms [81] [...] [83]
  - Inhibition of TNF by ADA/IFX leads to e.g. drop in markers like CRP [...] [83], also see [80, 84]
  - adalimumab is a human monoclonal antibody, infliximab is a mouse-human chimeric that is more immunogenic (more anti-drug antibodies that interfere with drug activity)
    - These two plus their biosimilars account for [...]
  - What determines which drug is given in the UK?
    - <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>

specific mechanisms of action in what cell type and tissue e.g. blood? gut?

introduce parts of the response algorithm here

how much market value?

    * <Infliximab and adalimumab for the treatment of Crohn's disease> https://www.nice.org.uk/guidance/ta187>
    * 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."
  - It is currently not able to predict efficacy. Treatment failure is not uncommon.

- 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

summarise rates

  - 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" [85]
  - also, immunogenicity leads to non-response via anti-drug Abs; serum drug levels correlated with efficacy, anti-drug Abs are inversely correlated [81]

  - 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, 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/
  - other options include dose intensification, and switching to an agent with diff mech [81]
  - But better still would be to target the right therapy to those that require and respond to it
    * although mucosal healing is the preferred gold standard [69], could be argued that PNR is important, as it's correlated with remission, and measurable early for stratification
    * baseline predictors would be especially valuable for patient stratification

2.1.3 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 [79, 86–89]
  - in PANTS: a recent prospective study of ADA and IFX for CD to date: large n=1610 [85]
  - PNR has observed at a 23.8% rate, evaluated at w14 via clinical 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 [79, 89]
  - OSM [90]
  - GIMATS [91]
  - there is conflict in existing studies on TREM1
  - [92]
    - being in blood: prognostic tests performed on blood samples are non-invasive and hence of high value.
  - [93]
  - difference may be due to sample size, ethnicity, definition of response https://pubmed.ncbi.nlm.nih.gov/30007919/ [79]

- Finally, genetic markers for non-response [88]
  - 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/ [88], and also multi-SNP risk scores https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6128143/
  - but like transcriptomic predictors, mostly small candidate gene studies
  - 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 [79, 89]
  - although no GWAS hits for PNR in PANTS, a related phenotype, immunogenicity is associated with HLA-DQA1*05 [94], so there may be some promise in the area

- In summary many studies with a variety of predictor types, tissues, cohort sizes, endpoints, analysis methods

  - 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 [89]

2.1.4 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-TNF therapy in Crohn's disease (CD) patients, described in detail by Kennedy *et al.* [85]. 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-TNF 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-TNF drugs evaluated were adalimumab and infliximab*. 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 for RNA-seq

The expression data for this chapter comes from PANTS whole blood 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. 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.

With the aim 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

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

be guaranteed that visits occurred on the exact day specified in the protocol, I considered the visit windows defined by Kennedy *et al.* [85]: 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 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 is important as they often replace 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.

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

- PNR was defined before w14 visit, according to the clinical algorithm in [85]
  - note PNR and remission at w54 are exclusive

2.2.4 RNAseq data generation

- <abbvie protocol>
  - total RNA extraction from tempus tubes
  - library prep with Kapa mRNA HyperPrep Kit, depleted of rRNA and globin mRNA using the QIAseq FastSelect RNA Removal Kit

  - sequenced on an Illumina HiSeq 4000 sequencer using 2x75bp read length.

2.2.5 RNAseq sample QC

- sample filtering
  - Starting from the full resequenced RNAseq dataset from AbbVie (Mar 2020), n=1141, remove:
  - Failed Mark's RNAseq QC (200 removed) e.g. number of reads TODO etc.
  - grey zone response (10 removed)
  - missing values for phenotypes used in section below.
  - 840 samples left
  - 814 mapped to timepoints

2.2.6 RNAseq quantification

- quantification with STAR+featurecounts
- gene filtering
- 58 900 genes in ENS annotation
- as in ch 2, filters:
  - >1.25 CPM in >10% (84) samples (15848 left)
  - >0 in >90% of samples (15645 left)
  - globins and short ncRNAs (15592 left)
- voom to get precision weights

2.2.7 DGE model selection

- In estimating the effect X→Y, of predictor X on response Y by regression, conditioning on 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. In this section, I focus on identifying third variables that are covariates for inclusion into the differential gene expression (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.

- If covariates are also associated with the predictor X, issues can arise depending on their causal role. In general, conditioning on a confounder (X ←Z →Y) reduces bias, conditioning on a collider (X →Z ←Y) induces bias, and conditioning on a mediator (X →Z →Y) changes the effect estimated by removing the indirect effect mediated by Z.

- 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`[95], which fits a mixed regression model. Variables in Fig. 2.1 were included as predictors.

  - Includes prognostic factors from [85]

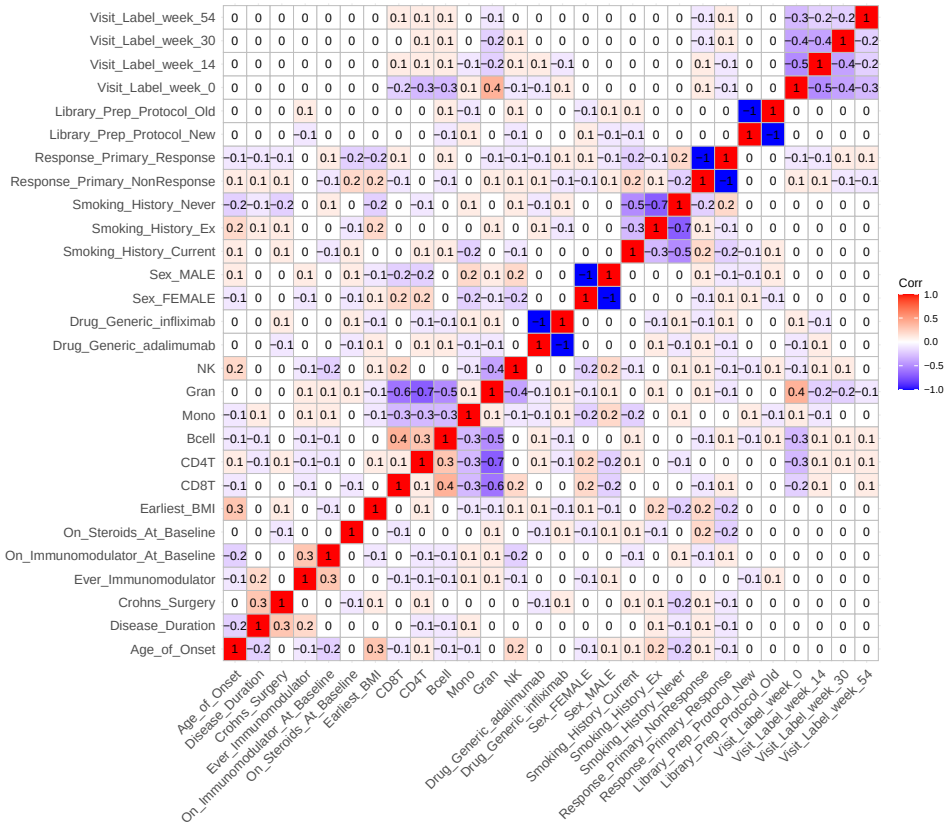


Figure 2.1: Correlation matrix of phenotypes considered as independent variables in DGE and eQTL models.

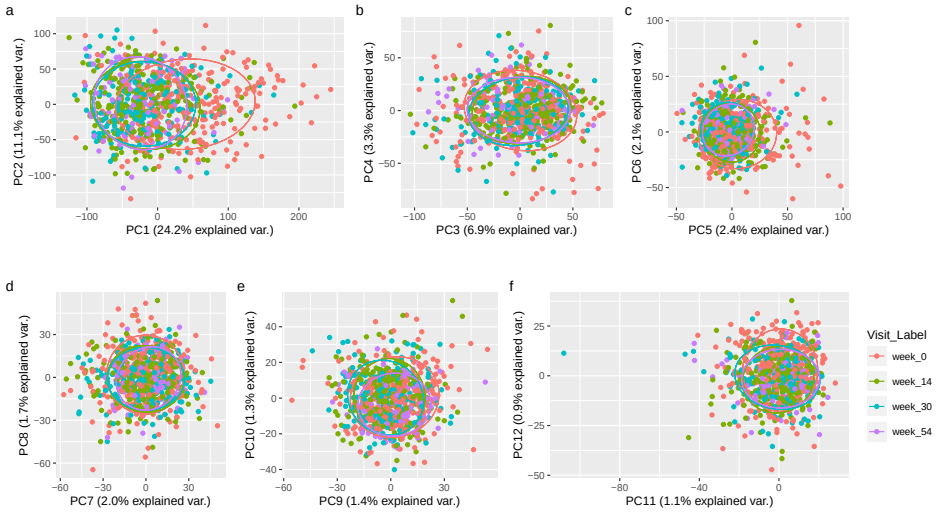


Figure 2.2: top 12 expression PCs of filtered expression data

- 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.* [95] showed that variance proportion estimates are unbiased even when coding categorical variables with as few as two categories as random, as long as all model parameters are estimated jointly using maximum likelihood (ML) rather than restricted maximum likelihood (REML)*. 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

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

  - 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)
  - 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
    * 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 [93]

anova for each cell prop over time for p values

- 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]

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?

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

2.2.8 Fitting Differential gene expression models

- model form is E [...]
- can we pool drugs with a mean term for this comparison? i.e. move from 8 to 4 means model

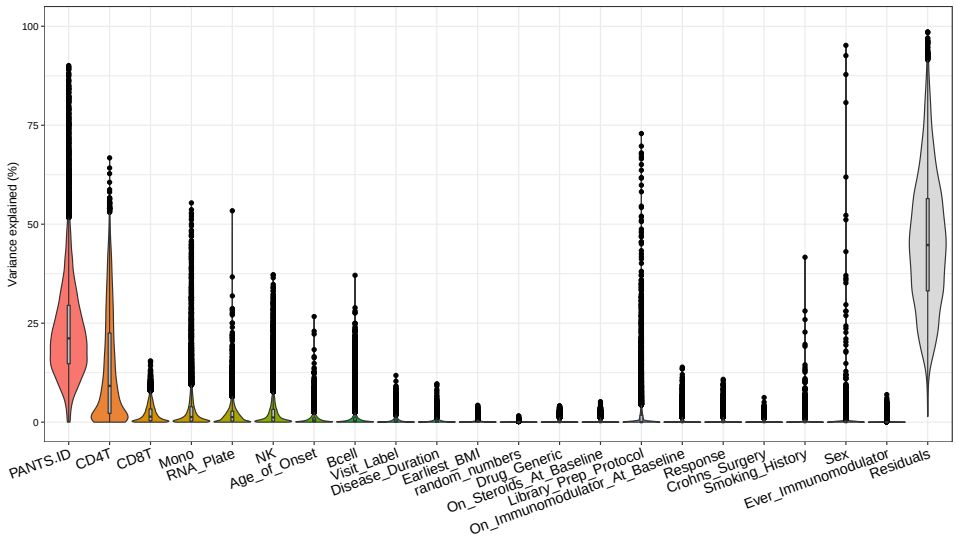


Figure 2.3: variance partition analysis, distribution of genewise % variance in expression explained by each variable

- 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??

- "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)." [85]
- I include many but not all of these covariates in the DGE model

2.2.8.1 Contrasts model for pairwise comparisons

- used `dream` hoffman2018DreamPowerfulDifferential

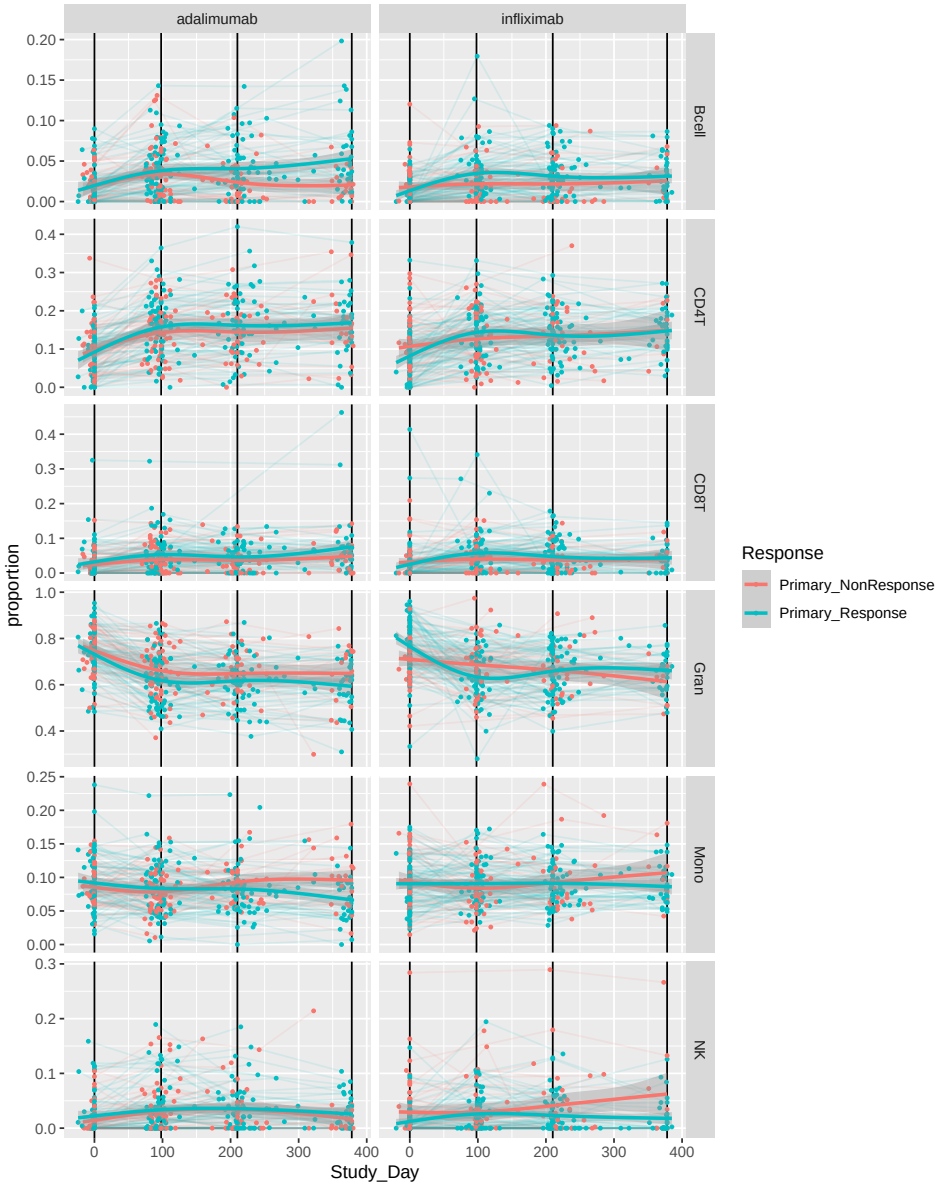


Figure 2.4: changes in cell proportions of 6 immune cell types over time

- 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
  - 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, 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

2.2.8.2 Spline model for difference over time

- 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

2.2.9 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

2.2.10 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

2.2.11 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

2.2.12 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

2.2.13 Computing GRMs

- LDAK, same as ch3

2.2.14 reQTL analysis

- same overall strat as ch3

2.2.14.1 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

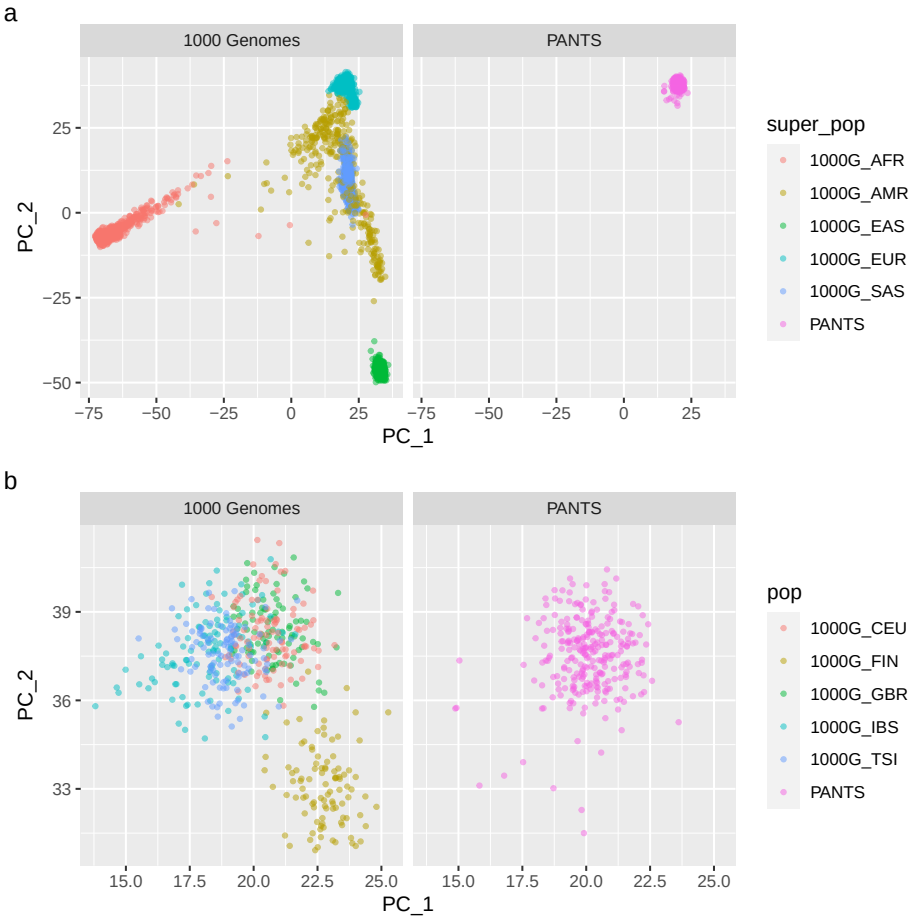


Figure 2.5: projection of PANTS samples onto 1000G genotype PC axes

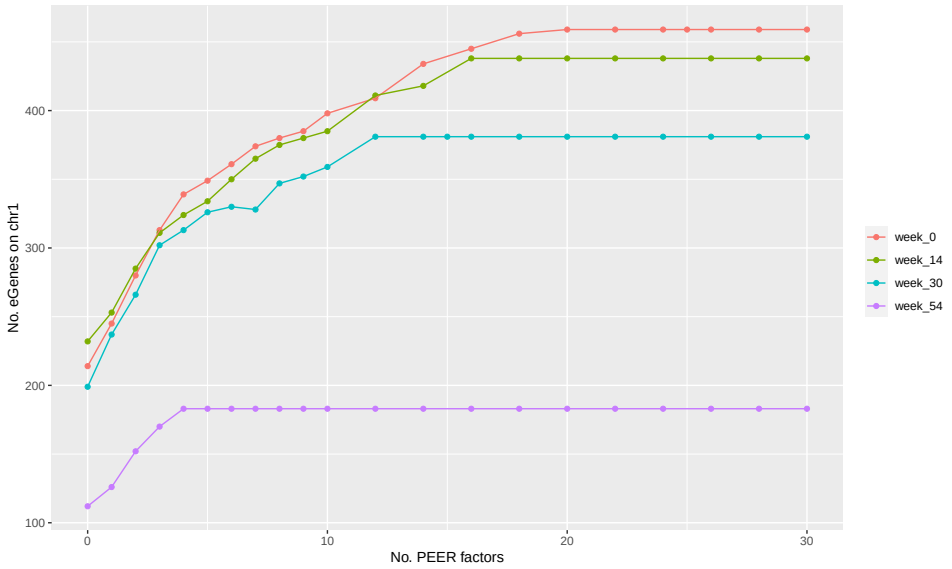


Figure 2.6: number of eGenes on chr1 used to choose number of PEER factors for each timepoint

  - breaking ties by highest imputation INFO, highest minor allele frequency (MAF), shortest dist to transcription start site (TSS), and genomic coordinate.

2.2.14.2 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

2.2.14.3 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

2.2.14.4 gprofiler

2.3 Results

2.3.1 Longitudinal RNA-seq data from the PANTS cohort

To define transcriptomic differences between primary responders and non-responders to anti-TNF therapy in the PANTS cohort, I analysed whole blood RNA-seq gene expression measured at up to four timepoints per patient: week 0 (anti-TNF naive baseline), 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 **??**.

autoref to table 1

2.3.2 Gene expression differences after anti-TNF induction

Primary response was assessed after the anti-TNF induction at week 14. "This is an observational study and the clinician may decide to continue with anti-TNF therapy even if the patient meets the study definition of primary non-response."

- week 14 R vs NR: single gene Fig. 2.9, module Fig. 2.11
  - much stronger effect sizes than at week 0 for single genes
    - SIGLEC10 and CROCC2 are also DE at week 14, same direction of effect as week 0
    - top downregulated hit KREMEN1 (Fig. 2.10)

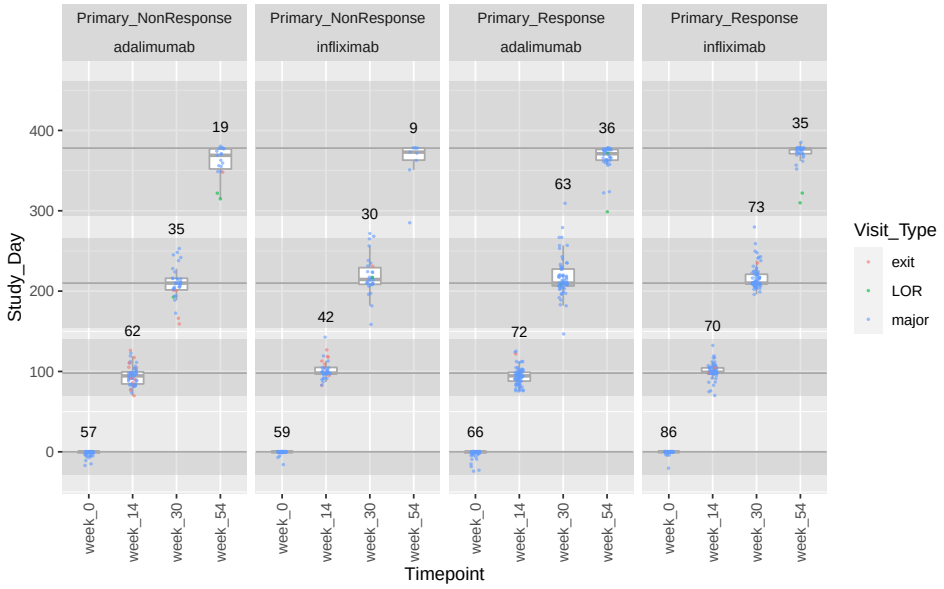


Figure 2.7: Distribution of samples among defined study visit windows. lor and exit are additional visits that fall into the windows.

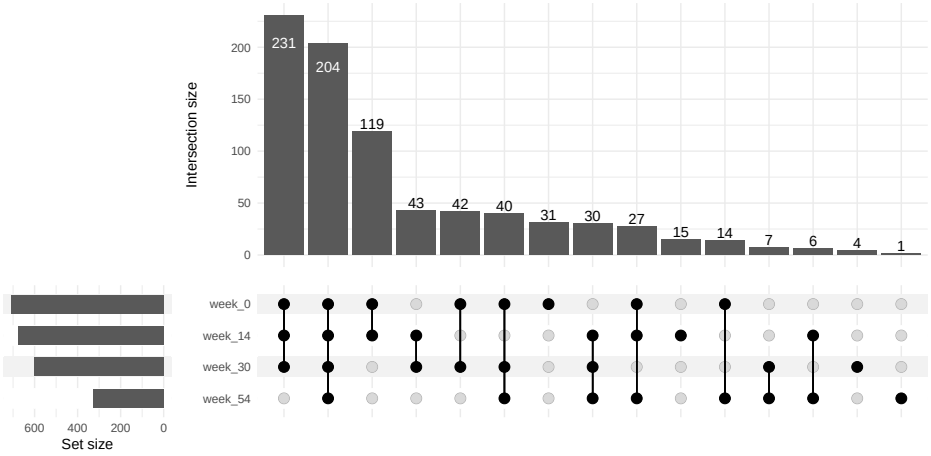


Figure 2.8

  - more consistent picture between drugs from the module analysis vs week 0
    * downregulation of innate TLR/inflam, monocyte, DC modules in responders
    * upregulation of B/T cell modules in responders

2.3.3 Gene expression differences baseline

I compared gene

- week 0 R vs NR: single gene Fig. 2.12, module Fig. 2.14
  - ADA only drug specific: IGKV1-9, KCNN3, PDIA5 downregulated in NR
    * IGKV1-9 is part of many B cell, plasma cell, immunoglobulin genes that are downreg in NR in the ADA specific module analysis
    * as seen in the IFX-ADA interaction, this is an ADA specific baseline phenomena, and seems to drive the signal in the pooled module analysis
  - pooled analysis: SIGLEC10 (sialic acid binding Ig like lectin 10) (Fig. 2.13) and CROCC2 (Ciliary Rootlet Coiled-Coil, Rootletin Family Member 2) upregulated in R
    * consistent direction of effect for both hits in separate drug analyses
    * pooled module analysis shows upregulation of myeloid cell receptor and antigen presentation modules

2.3.4 Replication of known signatures baseline

- Previously TREM1 blood expression found to be a predictor, where it was downregulated in responders.
  - In our data, strongest TREM1 effect is in IFX only analysis, but it's not significant at FDR 0.05: log2FC = 0.05 (1.04-fold up in PR at week 0); adj.p = 0.96 (Fig. 2.15)
    * stronger effect without cell prop adjustment, but still n.s.: log2FC = 0.29 (1.22-fold up in PR at week 0); adj.p = 0.06

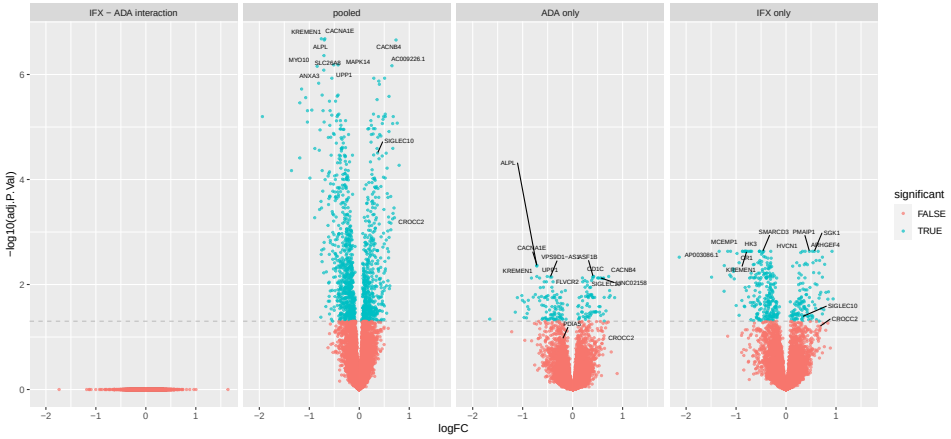


Figure 2.9: DGE volcano plot for PR vs PNR at week 14

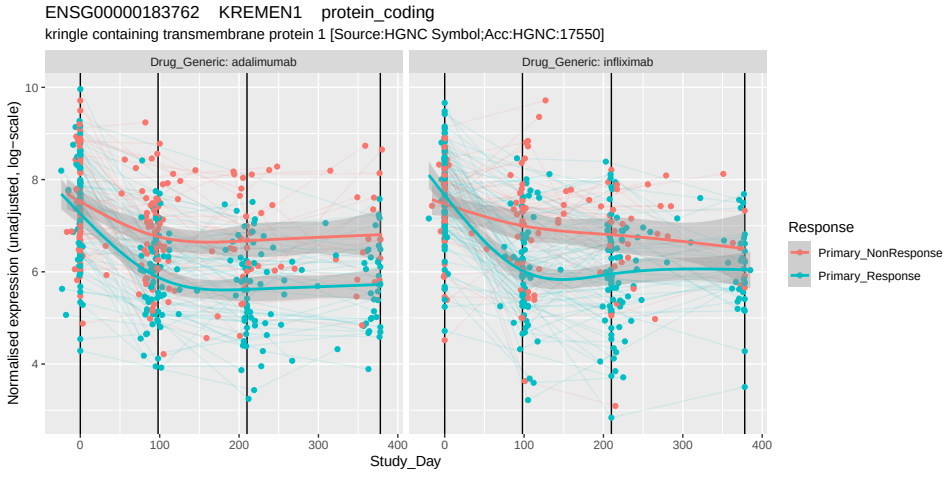


Figure 2.10: Unadjusted, normalised KREMEN1 expression over time

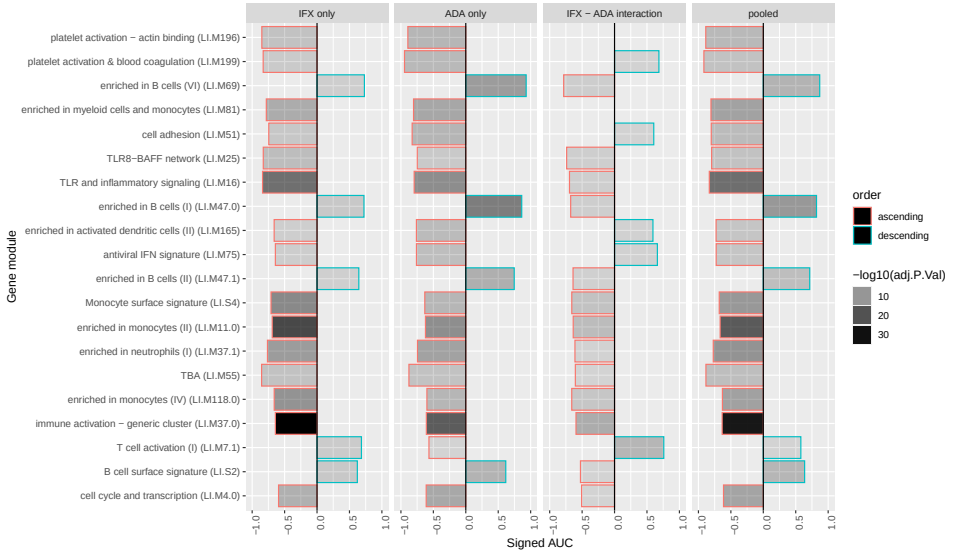


Figure 2.11: Panel plot of module enrichent analysis for PR vs PNR at week 14. Length of bar is effect size, shade is FDR. Blue is upreg in R, red is downreg in R.

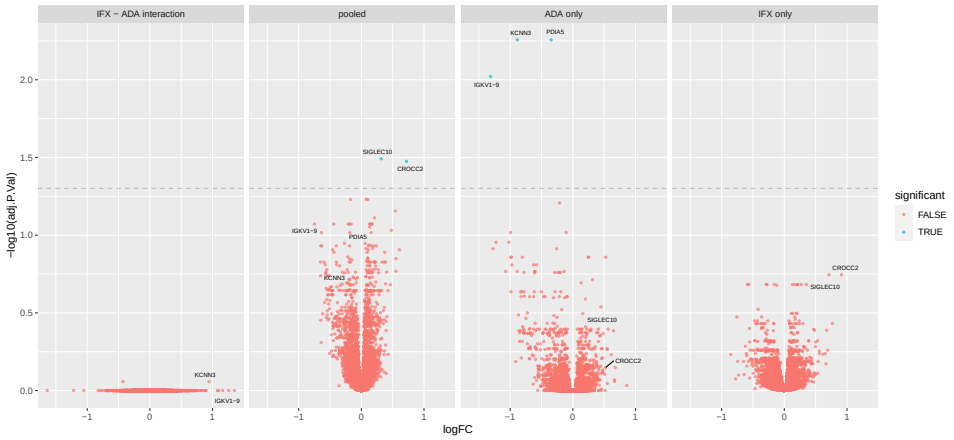


Figure 2.12: DGE volcano plot for PR vs PNR at week 0

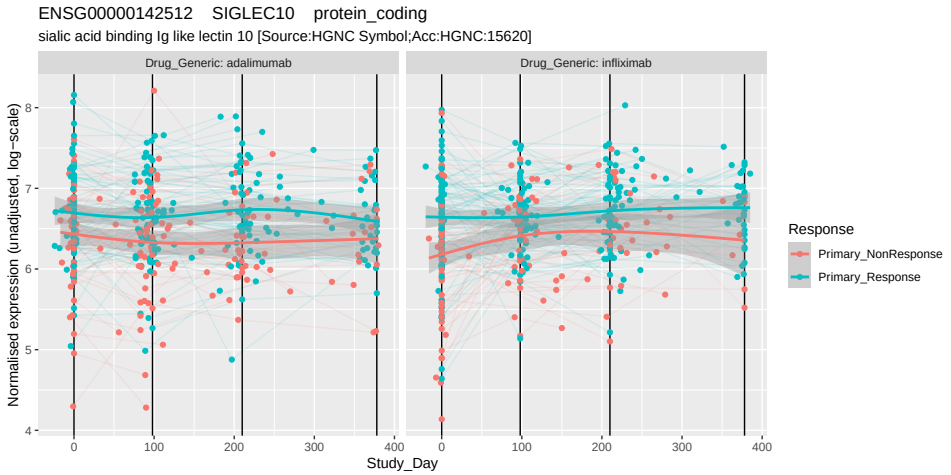


Figure 2.13: Unadjusted, normalised SIGLEC10 expression over time

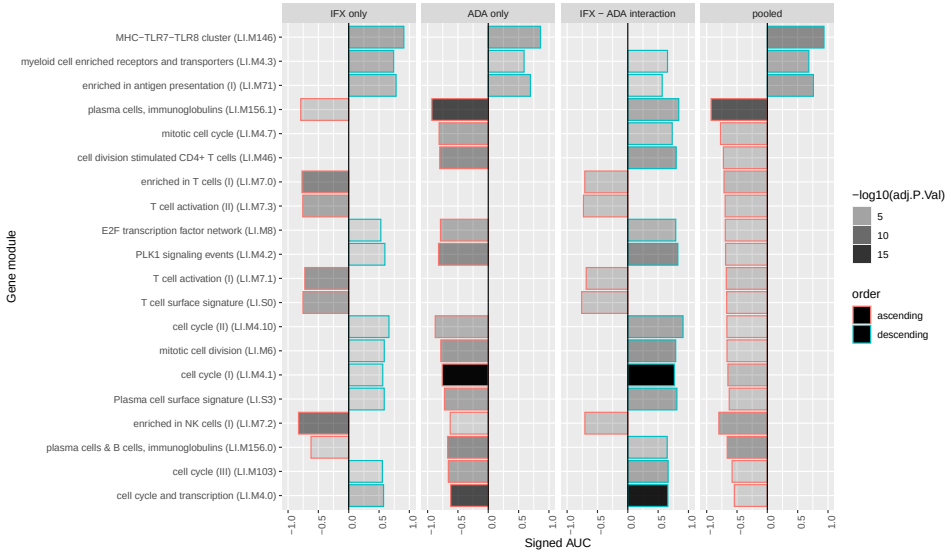


Figure 2.14: Panel plot of module enrichent analysis for PR vs PNR at week 0. Length of bar is effect size, shade is FDR. Blue is upreg in R, red is downreg in R.

  - Direction of effect is upregulated in PR, opposite to [92], consistent with [93]
  - potential differences in clinical vs endoscopic endpoint between all 3 studies

other known signatures from intro

2.3.5 Change magnified

- given the much strong differences in R vs NR expression at w14 than w0, interested in whether the change in expression upon starting anti-TNF treatment from week 0 vs week 14 is different for R and NR
  - contrast for interaction between R/NR and w14/w0
  - only found single gene hits for the pooled analysis Fig. 2.16
  - module analysis Fig. 2.17
    * finds many of the same modules as in w14-only R/NR comparison, with the same direction of effect

not sure about extra platelet activation modules yet

- TODO: add plot of w0 vs w14 effect size
  - more DGE w0 vs w14 for R than NR
  - most effects are magnifiying, such that R have larger foldchanges in the same direction for the same gene than NR

2.3.6 Expression differences during maintenance

- despite pnr at w14, some patients continue
- Finally, spline model as a formal way to test general differences between PR and PNR over all timepoints
  - instead of doing every pairwise R vs NR comparison
  - 266 hits at FDR < 0.05
  - clustering the expression of spline hits finds two main patterns: upregulation and downregulation after starting anti-TNFs, magnified in responders (Fig. 2.18)
  - from this analysis, can additionally see that expression differences are maintained even out to week 54

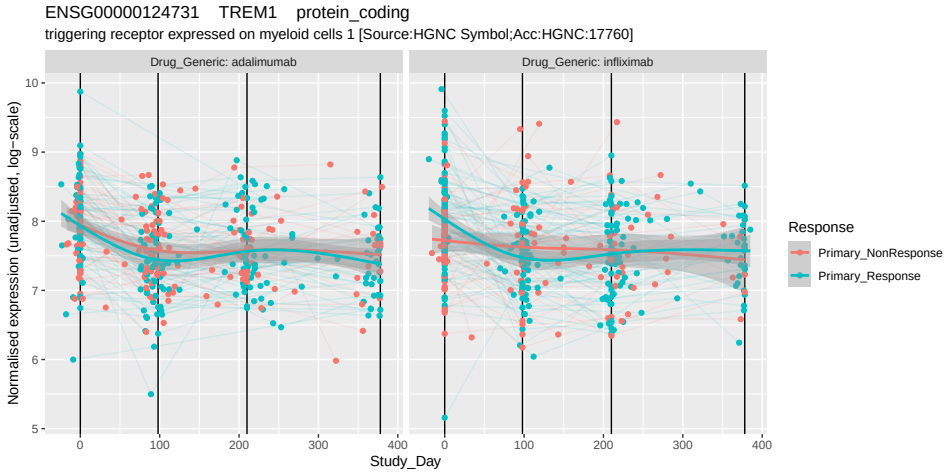


Figure 2.15: Unadjusted, normalised TREM1 expression over time

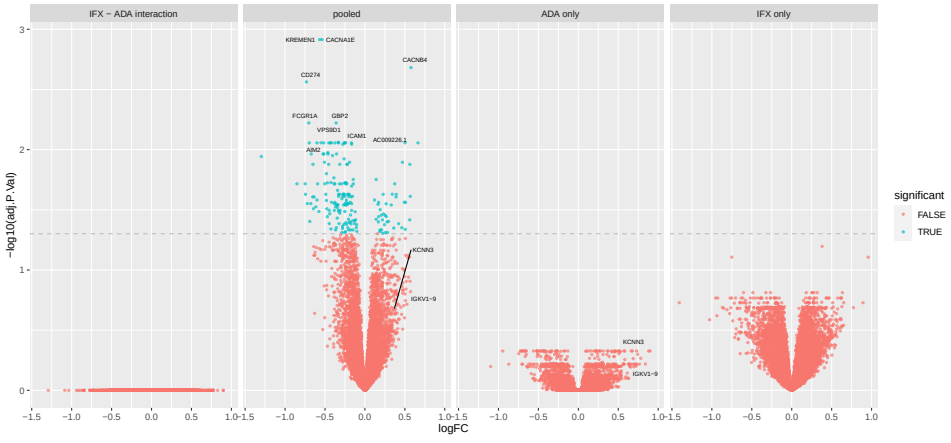


Figure 2.16: DGE volcano plot for PR vs PNR for week 14 - week 0 change

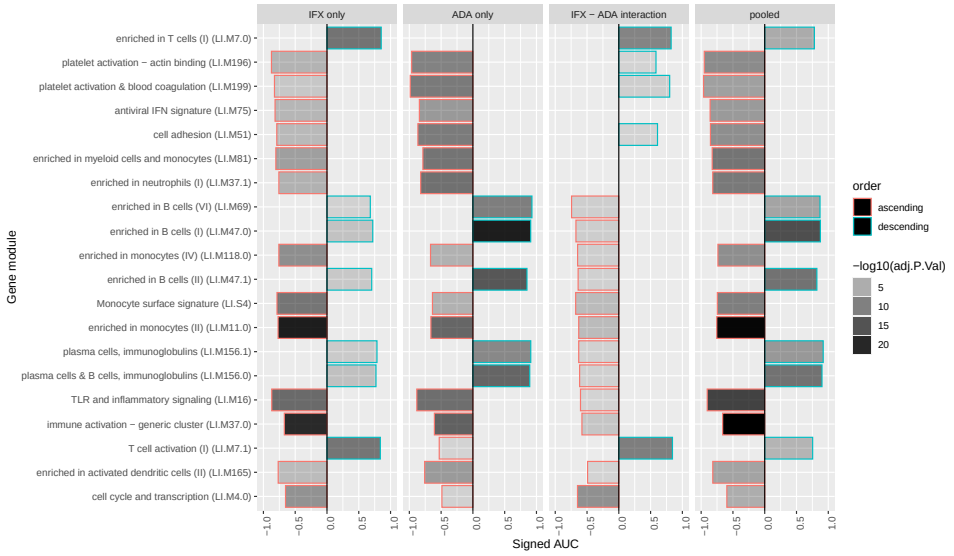


Figure 2.17: Panel plot of module enrichent analysis for PR vs PNR for week 14 - week 0 change. Length of bar is effect size, shade is FDR. Blue is upreg in R, red is downreg in R.

  - most hits in the spline have already been found in the separate w0, w14 or w14-w10 R vs NR comparisons (only 81 unique), but confirms the maintain

2.3.7 Genetics of gene expression over time

- given the substantial changes in expression after starting the drug, are there differences in genetic control of expression of the course of taking the drug?

have not put in an interaction between R and NR, not sure if powered

- mapped eQTLs per timepoint, then did joint analysis with mashr
  - 15040 genes tested after filtering
  - 11156/15040 (0.74175531914) of genes are eGenes: have at least 1 eQTL in at least 1 timepoint (mashr lfsr < 0.05)
  - TODO: assess pc that are DGE
- based only on lfsr threshold, most eQTLs are shared: 999 significant in 1 timepoint, 381 significant in 2 timepoints, 526 significant in 3 timepoints, 9250 significant in all 4 timepoints
  - formal test: compared 3 post-drug timpoints with baseline: did test for difference of betas between w0 and w14, w0 and w30, w0

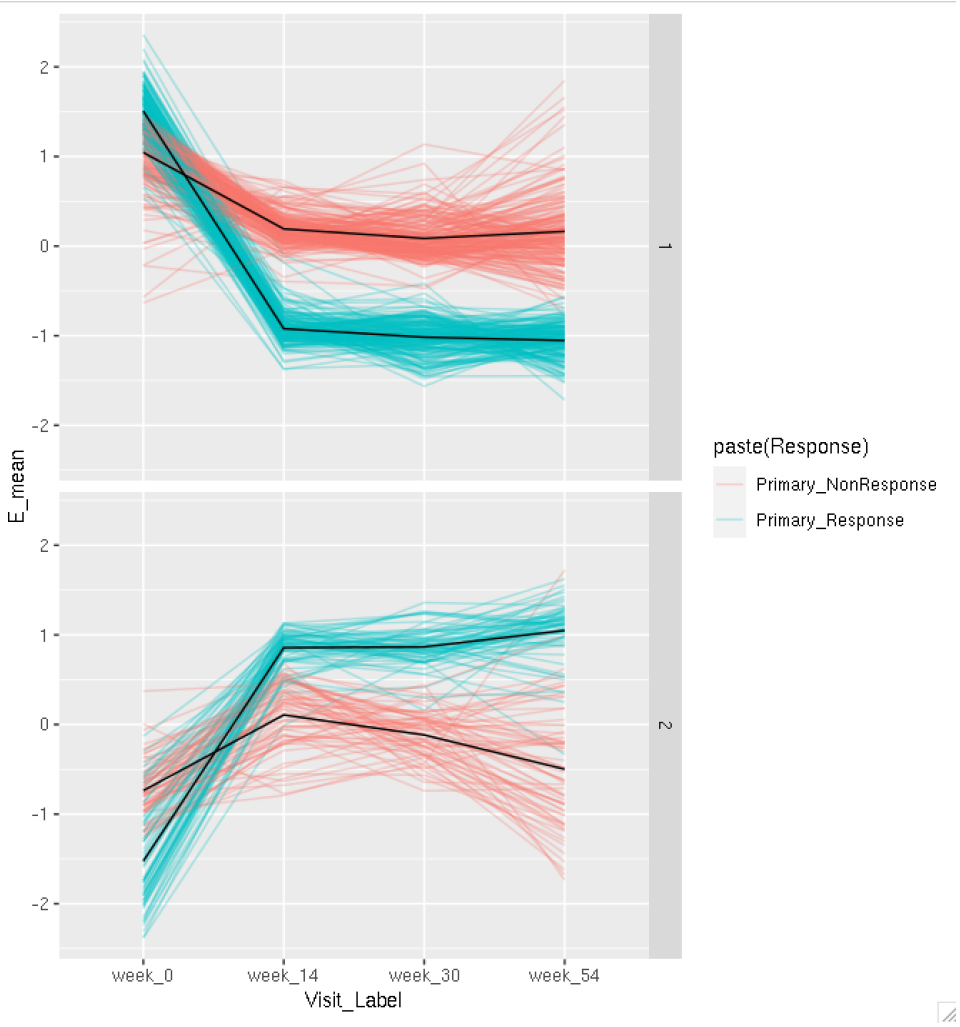


Figure 2.18: Clustered expression over time for DGE genes in spline analysis

and w54 to identify reQTLs with difference in effects while avoid thresholding effects

  - only 6 hits at BH FDR < 0.05. Strongest effects at w0 vs w54: five of the six are for w0 vs w54 (Fig. 2.19)

- clustering of eQTL effect sizes across 4 timepoints to identify general patterns of change in beta

similar to the idea of moving from single gene to gene set analyses for more sensitivity

  - start with prefilter for any reQTL significant at nominal p < 0.05. There were 344.
    * 327/344 significant in all timepoints, so align ok
  - 3 main clusters: high effect at w54, high effect at w0 and intermediate cluster (Fig. 2.20)

check if ADCY3 is in cluster 3

  - GSEA on cluster 1 revealed enrichment of genes with interferon regulatory motifs in cluster 1 (Fig. 2.21)
  - 2 INF genes

2.4 Discussion

- <study summary>
  - in PANTS, a largest to date cohort of anti-TNF naive CD patients who then got ADA/IFX
  - measured gene expression differences between PR and PNR over 4 timepoints from 0 to 54 weeks
  - reQTL mapping to identify changes in genetic control of expression over the timepoints

- at baseline, SIGLEC10 and CROCC2 were significantly upregulated in future PR to anti-TNF in the pooled analysis, with concordant direction of effect in both drugs
  - high levels of SIGLEC10 in nc-monocytes https://jbiol.biomedcentral.com/articles/10.1186/jbiol206 https://www.researchgate.net/profile/Siew-Cheng_Wong/publication/221723208_The_three_human_monocyte_subsets_Implications_for_health_and_disease/links/09e415101253562a3c000000.pdf

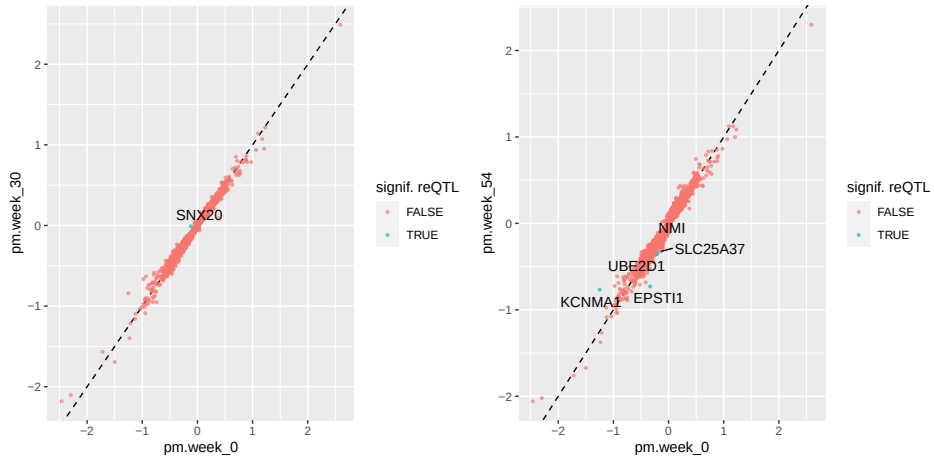


Figure 2.19: Week 30 and week 54 eQTL effect sizes vs baseline. Significant reQTLs in blue.

    * nc monocytes have a role in many chronic inflam diseases https://www.annualreviews.org/doi/full/10.1146/annurev-immunol-042617-053119#_i16

- SIGLEC10 is one of the innate immune cell-surface Ig superfamily that binds with CD24 and repress DAMP-mediated inflammation. https://www.nature.com/articles/mi201614?draft=collection
- in the context of IBD: Levels of DAMPS are increased in IBD.
    * chronic inflam -> tissue death -> release of small proinflam molecules called DAMPs [73]
    * "Interestingly, faecal calprotectin, the most frequently used and most sensitive marker of IBD clinical activity, is a complex of S100A8–S100A9, two prototypical DAMPs" https://www.nature.com/articles/mi201614?draft=collection [73]
- from [85] "In our study, higher baseline markers of inflammation predicted lower drug concentrations at week 14, suggesting that higher inflammatory load might contribute to faster drug elimination."
- proposed model: baseline high SIGLEC10 -> low DAMP levels -> low inflam -> (higher drug conc at w14) -> primary response
    * it is stressed that this is a hypothetical model, no way to

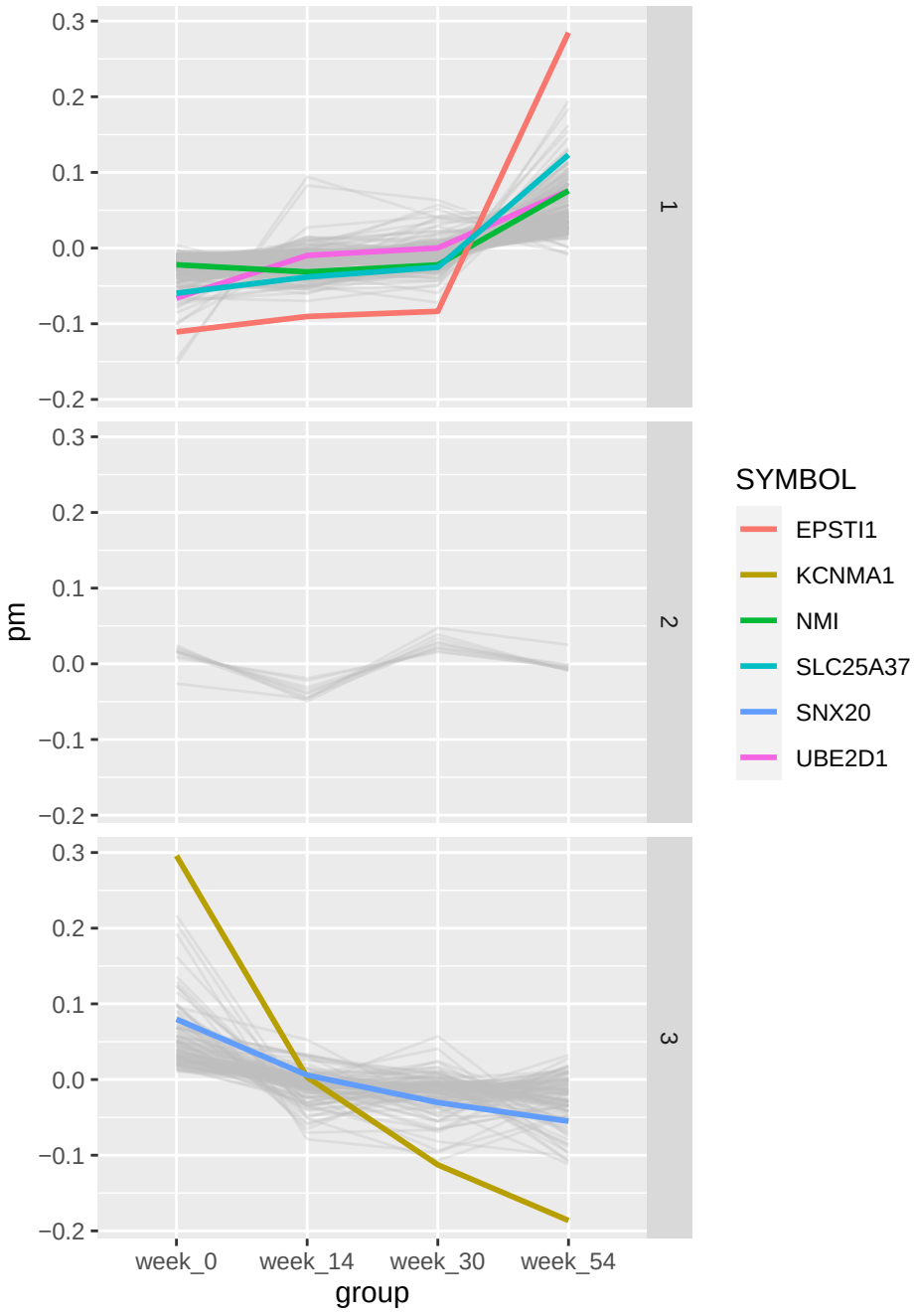
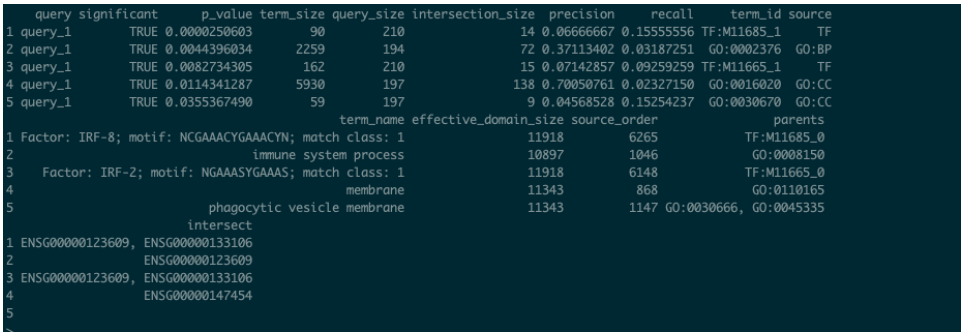


Figure 2.20: Clustering of eQTL betas over the 4 visits



	query	significant	p_value	term_size	query_size	intersection_size	precision	recall	term_id	source
1	query_1	TRUE	0.0000250603	90	210	14	0.06666667	0.15555556	TF:M11685_1	TF
2	query_1	TRUE	0.0044396034	2259	194	72	0.37113402	0.03187251	GO:0002376	GO:BP
3	query_1	TRUE	0.0082734305	162	210	15	0.07142857	0.09259259	TF:M11665_1	TF
4	query_1	TRUE	0.0114341287	5930	197	138	0.70050761	0.02327150	GO:0016020	GO:CC
5	query_1	TRUE	0.0355367490	59	197	9	0.04568528	0.15254237	GO:0030670	GO:CC

	term_name	effective_domain_size	source_order	parents
1	Factor: IRF-8; motif: NCGAAACYGAAACYN; match class: 1	11918	6265	TF:M11685_0
2	immune system process	10897	1046	GO:0008150
3	Factor: IRF-2; motif: NGAAASYGAAAS; match class: 1	11918	6148	TF:M11665_0
4	membrane	11343	868	GO:0110165
5	phagocytic vesicle membrane	11343	1147	GO:0030666, GO:0045335

	intersect
1	ENSG00000123609, ENSG00000133106
2	ENSG00000123609
3	ENSG00000123609, ENSG00000133106
4	ENSG00000147454
5	

Figure 2.21: gene set enrichment using gprofileR for cluster 1 genes

suggest causality due to the model used, and this being an uncontrolled cohort study

- not much known about CROCC2 (aka AC104809.3), but it's expression is nc-monocyte specific https://dice-database.org/genes/AC104809.3 https://www.proteinatlas.org/ENSG00000226321-CROCC2
  - potential source of multiomics data for potential validation is [96], contains drug response phenotypes
  - also, validate using protome/serological data in PANTS, although there is a known disconnect between proteome and transcriptome
- there ADA-specific downreg of gene modules in PR, especially plasma cell and immunoglobulins
  - includes one of the 3 hits: IGKV1-9
  - B cell genes down -> ??? -> response to anti-TNF
    - consistent, as TNF is involved in T cell-dependent B-cell responses, so inhibiting it should lower B cell response https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6235207/
  - Gaujoux *et al.* [93] find lower props of inflam macrophage and plasma cell in IFX responders in gut biopsy
  - also, could be due to cell props, as B cell prop used in model is not perfectly correlated with plasma cell freq (see ch3)
  - differences in baseline characteristics between drugs discussed later on in limitations

I have no idea why the effect is stronger in ADA

- replication of known baseline signatures
  - TREM1 signal from Verstockt *et al.* [92] is not replicated in this study, n.s. and opposite direction of effect
  - effect decreases when adding cell proportion covariates: perhaps TREM1 mainly a surrogate for (monocyte) cell props?
  - the replication is sensitive to covariates, end points, drug
    - e.g. Verstockt *et al.* [92] did not use any covariates (i.e. t test), so they report an unadjusted effect
  - an aside: just as I examined previously known biomarkers, I expect replication of SIGLEC10 and CROCC2 before putting much credence into them
- at week 14, difference in transcriptome between R/NR becomes 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?
  - spline analysis confirmed that the diff starting at w14 is maintained at w30 and w54
  - [85]: "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>
  - 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?
  - 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 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

- <more limitations: internal validity>
  - 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.
        * 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 [81]
    - was pooling drugs ok?
        * included these as covariates
        * residual confounding may be an issue since this is uncontrolled, unrandomised
    - cell prop correction
        * 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

- <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?
          richness of dataset although other mediators of NR could be modelled using genetic instruments e.g drug level
    - blood, not gut

- <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
    - 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

Variable	adalimumab	infliximab	Overall	
Sex				
(Col %)				F
FEMALE	78 (48.4%)	89 (54.6%)	167 (51.5%)	
MALE	83 (51.6%)	74 (45.4%)	157 (48.5%)	
Age of onset				
Mean (SD)	33.3 (15.4)	32.8 (15.3)	33.1 (15.3)	Wilcoxo
Disease duration				
Mean (SD)	6.1 (8.1)	5.9 (7.7)	6.0 (7.9)	Wilcoxo
Smoking status				
(Col %)				F
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 surgery				
(Col %)				F
FALSE	114 (70.8%)	110 (67.5%)	224 (69.1%)	
TRUE	47 (29.2%)	53 (32.5%)	100 (30.9%)	
Ever immunomodulator				
(Col %)				F
FALSE	23 (14.3%)	28 (17.2%)	51 (15.7%)	
TRUE	138 (85.7%)	135 (82.8%)	273 (84.3%)	
On immunomodulator at baseline				
(Col %)				F
FALSE	79 (49.1%)	81 (49.7%)	160 (49.4%)	
TRUE	82 (50.9%)	82 (50.3%)	164 (50.6%)	
On steroids at baseline				
(Col %)				F
FALSE	113 (70.2%)	92 (56.4%)	205 (63.3%)	
TRUE	48 (29.8%)	71 (43.6%)	119 (36.7%)	
Earliest BMI				
Mean (SD)	25.2 (6.2)	24.3 (5.5)	24.8 (5.9)	Wilcoxo

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

70. 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).

71. 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).

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

73. 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).

74. 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).

75. 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).

76. 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).

77. 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).

78. 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).

79. 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).

80. 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).

81. 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).

82. 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).

83. 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).

84. Danese, S., Vuitton, L. & Peyrin-Biroulet, L. Biologic Agents for IBD: Practical Insights. *Nature Reviews Gastroenterology & Hepatology* **12,** 537–545. doi:10.1038/nrgastro.2015.135 (2015).

85. 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).

86. 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).

87. 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).

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

89. 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).

90. 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).

91. 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).

92. 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).

93. 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).

94. 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).

95. 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).

96. 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

anti-TNF anti-tumour necrosis factor

bp base pair

CD Crohn's disease

DGE differential gene expression

eQTL expression quantitative trait locus

FWER family-wise error rate

GWAS genome-wide association study

LD linkage disequilibrium

MAF minor allele frequency

ML maximum likelihood

PANTS Personalised Anti-TNF Therapy in Crohn's Disease

PNR primary non-response

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

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