Chapter 4

Transcriptomic associations with anti-TNF drug response in Crohn's disease patients

The work presented in this chapter is a collaboration between the Wellcome Sanger Institute, the Royal Devon and Exeter Hospital National Health Service (NHS) Foundation Trust, the University of Exeter, and AbbVie. I would like to thank Nicholas Kennedy, Tariq Ahmad, and the AbbVie team for kindly extending the opportunity to collaborate on the PANTS cohort; Mark Reppell, Samantha Lent, and others at AbbVie, for performing the RNA-seq library preparation and sequencing, initial quality control, alignment and quantification, and estimation of cell proportions from methylation data; Simeng Lin, for advice on the sample structure of PANTS; Aleksejs Sazonovs, for performing the genotype quality control; and other individuals in the Sanger-AbbVie-Exeter PANTS working group, for their feedback during our video conferences.

4.1 Introduction

4.1.1 Crohn's disease and inflammatory bowel disease

Crohn's disease (CD) is a chronic inflammatory disease of the gastrointestinal tract. Along with ulcerative colitis (UC), it is one of the two main forms of inflammatory bowel disease (IBD). CD is 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. In contrast, UC is characterised by continuous inflammation, with lesions that are superficial rather than transmural, and restricted to the colon [335]. Whilst the two are distinct forms of IBD, similarities in clinical presentation, available therapies, and genetic architecture mean they have often been studied together. Both are immune-mediated inflammatory diseases (IMIDs), a group of related diseases involving immune dysregulation of common inflammatory pathways. Other IMIDs include type 1 diabetes (T1D), systemic lupus erythematosus (SLE), rheumatoid arthritis (RA), multiple sclerosis (MS), and psoriasis [336, 337].

Pathogenesis of CD is not completely understood, but involves interaction of the immune

system, environmental factors (e.g. smoking, stress, diet [335, 338]), and gut microbial factors in a genetically-susceptible individual [339]. Since the seminal discovery by linkage analysis in 2001 that genetic variation in *NOD2* is linked to CD risk [340], much progress has been made in establishing the disease's genetic architecture. The most recent genome-wide association study (GWAS) studies catalogue over 240 risk loci for IBD [180]. Most associations are shared between CD and UC, but there is strong heterogeneity of effects at some loci, such as *NOD2* being only associated with CD risk [341, 342].

CD has historically been considered a disease of the Western world. The highest prevalence and incidence of new CD cases are in North America and Western Europe [335], although disease burden is now rising in newly industrialised countries in Asia, Africa, and South America [343, 344]. The modal age of onset is typically between late adolescence and early adulthood. The disease is progressive: within 10 years of diagnosis, approximately 50 % of CD patients develop further complications (strictures or penetrating lesions); within 20 years, approximately 15 % will require surgical intervention [335]. Given the rising prevalence and large impact on quality of life, there is active research into developing treatment regimens with the goal of inducing complete mucosal healing [335, 345].

4.1.2 Anti-TNF therapies for Crohn's disease

Tumour necrosis factor (TNF), also known by the archaic name TNF-α, is a proinflammatory cytokine produced mainly by immune cells such as monocytes, macrophages, natural killer (NK) cells, T cells, and B cells. It is synthesised in transmembrane form, then enzymatically cleaved into its soluble form. TNF binds to receptors TNFR1 and TNFR2; most cells in the body express one receptor or the other. Binding triggers a signalling cascade that in different contexts regulates inflammation, apoptosis, cell proliferation, and cell survival [346–348]. In the context of IBD pathogenesis, current models suggest high TNF levels promote apoptosis of monocytes, macrophages, and gut epithelial cells via TNFR1, while inhibiting apoptosis of mucosal $CD4^+$ T cells via TNFR2 [345, 348, 349], overall encouraging maintained gut inflammation.

The development of anti-TNF biologic therapies has revolutionised patient care for CD and a number of other IMIDs in the last two decades. Infliximab and adalimumab are the two major anti-TNF drugs in use. Both are IgG1 monoclonal antibodies that bind both soluble and transmembrane TNF, inhibiting their interactions with TNF receptors [349, 350]. Two main mechanisms for their action have been proposed: induction of $CD4^+$ T cell apoptosis in the gut mucosa by inhibiting the TNF-TNFR2 interaction; and binding of the antibody tail (Fc) region of the drug to Fc receptors on monocytes, inducing their differentiation into wound-healing M2 macrophages [345].

Adalimumab is a human antibody, typically administered subcutaneously via auto-injector pen, with two initial doses aimed to induce remission, then a dose every two weeks to maintain remission. Infliximab is a chimeric mouse-human antibody, administered via intravenous infusion, with a three-dose induction, then doses every eight weeks for maintenance [349]. Anti-TNF biologics consistently rank among the drugs generating the highest global revenues. In 2017, spending on adalimumab (Humira) in developed markets was estimated at 20.7 billion USD—almost double the spending on second-ranked insulin glargine (Lantus, 10.5 billion USD) [351].

4.1.3 Anti-TNF treatment failure

Unfortunately, anti-TNF therapy is not always effective at treating CD. Various types of treatment failure can occur: primary non-response (PNR) within the induction period (the first 12–14 weeks for adalimumab and infliximab), developing secondary loss of response (LOR) during maintenance after an initial response, failure to achieve remission after the treatment course, and adverse events that lead to treatment stoppage [352]. For IBD patients, the incidence of PNR is 10–40 %, and the incidence of secondary LOR is 24–46 % in the first year of treatment [353–355]. Another factor affecting treatment outcome is immunogenicity, the generation of antibodies against the drug, thought to increase the probability of treatment failure and LOR by increasing drug clearance rate [350, 355]. As a chimeric antibody, infliximab is more immunogenic than adalimumab [355, 356]. Although remission with complete mucosal healing remains the gold standard for treatment success or failure [335], PNR and LOR phenotypes can be defined much earlier in the treatment course, and help guide changes in treatment regimens, such as dose intensification or switching to a drug class with a different mechanism of action [350, 353].

Anti-TNF biologics are near the top of the therapeutic pyramid for CD in the UK, among the treatment options with the highest toxicity and costs [357]. The traditional approach to disease management in the UK is "step-up", beginning at the bottom of the pyramid with steroids [349, 357]. This may undertreat patients that require more aggressive therapies, allowing the disease time to progress. An inverted approach begins at biologic therapies, then steps down the pyramid if possible. This risks exposing patients to aggressive therapies they may not have needed [354]. The best approach would be to predict whether a particular treatment will be required and effective for a patient, especially given the costliness and patient risks associated with therapies near the top of the pyramid. Reliable baseline prediction would be especially valuable for stratifying patients to specific therapies from treatment initiation.

4.1.4 Predicting patient response to anti-TNFs

Clinical variables reported to have associations with anti-TNF response include age, disease duration, body mass index (BMI), smoking, C-reactive protein (CRP) levels, faecal calprotectin levels, serum drug concentrations, and anti-drug antibody concentrations. These associations have mostly been found in small retrospective cohorts, and have rarely been independently validated [348, 354, 358–361]. In the Personalised Anti-TNF Therapy in Crohn's Disease (PANTS) study, the largest study of infliximab and adalimumab response in CD patients to date (enrollment $n = 1610$), baseline obesity, smoking, and greater disease activity were associated with low serum drug concentration after induction. Low drug concentration was in turn associated with PNR and non-remission, suggesting immunogenicity may be mediating treatment failure [355].

Multiple studies have attempted to define transcriptomic predictors for anti-TNF response in gut biopsies and blood [348, 361]. In gut biopsies, expression of sets of "signature" genes were found to be predictive of mucosal healing after infliximab treatment in cohorts of UC (*TNFRSF11B*, *STC1*, *PTGS2*, *IL13RA2*, *IL11*; $n = 46$ [362]) and CD patients (*TNFAIP6*, *S100A8*, *IL11*, *G0S2*, *S100A9*; $n = 19$ [363]). Expression of *OSM* was associated with anti-TNF response defined by improved Mayo score, a multiparameter clinical score of UC activity ($n = 227$)

[364]. Most recently, single-cell RNA sequencing (RNA-seq) identified a module of IgG plasma cells, inflammatory mononuclear phagocytes, activated T cells, and stromal cells associated with clinical remission after anti-TNF therapy in two separate CD cohorts (total $n = 340$) [365].

As obtaining blood samples is non-invasive, there has been great interest in finding transcriptomic predictors of response in blood. While blood is not the main disease-relevant tissue for CD, many genes in gut biopsy signatures have high expression in infiltrating immune cells, and blood gene expression may capture the precursors of those cells [366]. Blood *TREM1* expression has been identified as a marker of anti-TNF response in two studies with inconsistent directions of effect. Gaujoux *et al.* [366] defined response based on "clinical and/or endoscopic improvement". *TREM1* expression was lower in infliximab responders in gut biopsies (total $n = 72$), but higher in responders in a separate cohort measuring baseline whole blood expression ($n = 22$). Verstockt *et al.* [367] defined response based on endoscopic remission, reporting *TREM1* to be a marker of response with lower expression in responders to infliximab and adalimumab in both baseline gut biopsies ($n = 44$) and baseline whole blood ($n = 54$). Proposed reasons for the discrepancy include false positives due to small sample sizes, differences in patient ethnicity, and different definitions of response [348, 368].

Attempts have also been made to find genetic markers for response. Anti-TNF response does not necessarily share the same genetic architecture as disease risk. Variants in TNF-regulated genes that are also associated with IBD risk (*NOD2*, *TNFR1*, *TNFR2*) are not associated with response to infliximab [348, 361]. A number of candidate gene studies found single nucleotide polymorphism (SNP) associations with response in genes such as apoptosis-related Fas ligand and caspase-9 that have yet to be validated [354, 369]. Recently, larger cohorts have enabled GWASs of anti-TNF response in IBD. In PANTS, although no associations to PNR were genome-wide significant, HLA-DQA1*05 carriage was found to be associated with higher anti-drug antibody levels, which was in turn associated with LOR [370], but larger samples may be needed to find direct associations between HLA-DQA1*05 carriage and LOR.

Overall, small sample sizes and variation among studies in analysis methods, anti-TNF drug, response definition, tissues sampled, and disease make a consensus hard to establish. Few markers of any type—clinical, transcriptomic (gut/blood), or genetic—have been validated in independent studies. No algorithms using such markers for predicting IBD patient response to anti-TNF therapy have yet been translated to clinical practice, although several are currently undergoing validation [361].

4.1.5 Chapter summary

This chapter focuses on identifying novel transcriptomic associations with anti-TNF primary response in a subset of the PANTS cohort with longitudinal RNA-seq data from the first year of follow-up. I model differential gene expression (DGE) between primary responders and non-responders at the gene and module-level at baseline (week 0), post-induction (week 14), and during maintenance (week 30 and week 54). As this is one of the largest datasets currently available for assessing transcriptomic associations with anti-TNF response in IBD, I attempt to validate and resolve conflicts in the literature for previously identified transcriptomic markers such as *TREM1*. Finally, I integrate existing genotype data to map response expression quantitative

trait loci (reQTLs) between timepoints, with the aim of identifying common genetic variants controlling expression response to anti-TNF drugs.

4.2 Methods

4.2.1 The Personalised Anti-TNF Therapy in Crohn's Disease (PANTS) cohort

PANTS is a UK-wide, prospective, observational cohort study of response to anti-TNF therapy in CD patients, described in detail by Kennedy *et al.* [355]. The study was registered with ClinicalTrials.gov identifier NCT03088449, and the protocol is available at `https://www.ibdresearch.co.uk/pants/`. 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 (ADA) and infliximab (IFX). The study also evaluated infliximab biosimilars; data from patients who received a biosimilar are not included in this chapter. All major visits were scheduled immediately prior to a drug dose. Adalimumab and infliximab have 2-week and 8-week dosing intervals respectively, so the timing of major visits was chosen such that the same visit structure could be used for patients on either drug. Additional visits were scheduled in case of secondary LOR or premature study exit due to drug withdrawal, usually replacing the next scheduled major visit.

The overall rate of primary non-response by week 14 was 21.9 % for infliximab and 26.8 % for adalimumab. The rate of secondary LOR by week 54 among primary responders was 36.9 % for infliximab and 34.1 % for adalimumab. Rate of remission by week 54 was 39.1 % for infliximab and 33.1 % for adalimumab.

4.2.2 Definition of timepoints

The RNA-seq data for this chapter comes from a subset of the cohort sampled around four timepoints: week 0, week 14, week 30, and week 54. These are the target timings for four major visits in the first year of follow-up. The week 0 major visit is the visit immediately prior to the first dose of drug. Week 0 to week 14 is the induction period. After week 14, patients continued to take their drug according to the prescribed schedule (a dose every 8 weeks for infliximab, and every 2 weeks for adalimumab). Whole blood samples at major visits were taken prior to the scheduled drug doses that aligned with those visits, labelled with the visit's name, and preserved for RNA-seq in Tempus Blood RNA Tubes. As sampling was always done prior to a scheduled dose, the measured transcriptome reflects the state at trough drug levels.

I mapped samples from major and additional visits to four discrete timepoints centered around the four major visits. As it could not be guaranteed that visits occurred on the exact day specified in the protocol, I considered the visit windows defined by Kennedy *et al.* [355]: week 0 (week −4–0)*, week 14 (week 10–20), week 30 (week 22–38), and week 54 (week 42–66). Samples

*Samples at negative weeks or study days indicate the patient was first sampled before the day they took the first drug dose (week or day 0).

were mapped to timepoints based on their sample label and study day according to the following mapping criteria:

- Labelled major visit samples were mapped to the corresponding timepoint, regardless of whether they fell within the corresponding window i.e. a sample *labelled* week 54 was always mapped to the week 54 timepoint.
- Samples taken at additional (LOR 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. This avoided any patient having multiple samples for a single timepoint.

Only a small minority of major visit samples fell outside their corresponding windows, mostly for later timepoints where there was more variation around the target day. Inclusion of samples from additional visits was important, as they often replaced major visits for patients with PNR or LOR. For example, a patient who developed PNR by week 14 and decided to exit the study would not have a labelled week 14 major visit, but may still have a sample taken at that time labelled as an additional exit visit. The mapping of samples to timepoints is shown in Fig. 4.1. Samples included under both of the above mapping criteria should be representative of trough drug levels, as major visits and LOR visits were always scheduled prior to a drug dose, and exit visits were scheduled for when the next drug dose would have been.

4.2.3 Definition of primary response and primary non-response

The definition of primary response and non-response was based on the clinical decision tree from Kennedy *et al.* [355]. Primary response and non-response could be assessed from week 12, with a final classification made by the scheduled week 14 visit. The criteria for primary non-response was *either* of the following:

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

The criteria used to define primary response was *all* of the following:

- not classified as a primary non-responder;
- compared to week 0, a decrease in CRP by $\geq 50\,\%$ or to $\leq 3\,\mathrm{mg\,l^{-1}}$, *and* a decrease in HBI by ≥ 3 points or to ≤ 4.

Grey zone patients that only partially met the criteria for either primary non-response or response were excluded in this chapter. There were also additional selection criteria used to choose the subcohort of PANTS patients put forward for RNA-seq. Patients were required to be at least 16 years old, and to have an available baseline serum sample. Within the patients on infliximab, there was propensity score matching between primary non-responders and other patients based on baseline immunomodulator use, baseline steroid use, age, sex, and BMI. As PANTS was

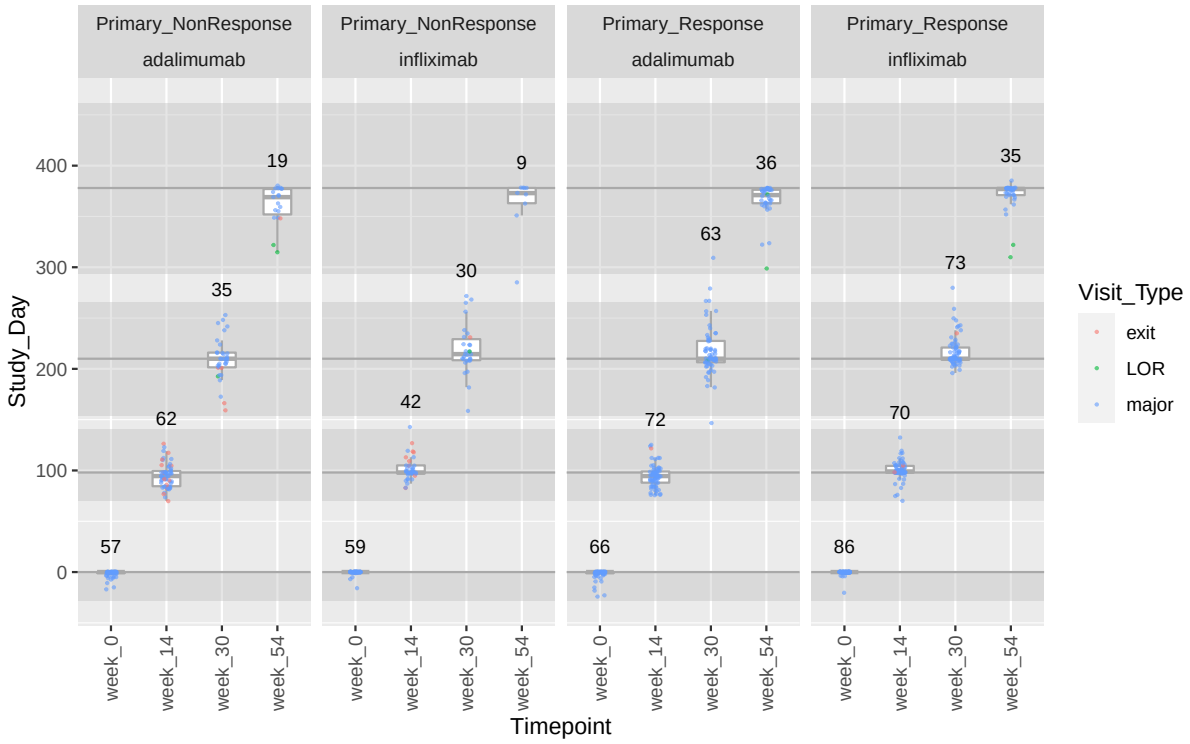

Figure 4.1: Sample size and study day distribution for PANTS study RNA-seq samples, stratified by timepoint and study group. Windows from Kennedy *et al.* [355] for the four major PANTS visits are colored in grey. Samples mostly come from major visits, but a small number of LOR and exit visit samples were included according to the criteria in Section 4.2.2.

an observational study that continued until drug withdrawal, a patient's clinician may have decided to continue anti-TNF therapy even if a patient demonstrated primary non-response, so it was possible for primary non-responders to remain in the study past week 14. Primary non-responders were selected excluding patients known to be in remission at week 54. Primary responders were selected from patients known to be in remission by week 30 or week 54. The primary non-responders and responders in the RNA-seq subcohort thus represent phenotypic extremes of response.

4.2.4 Library preparation and RNA-seq

Total RNA was extracted following the Qiagen QIAsymphony instrument protocol (RNA Isolation PAX RNA CR22332 ID 2915). RNA was quantified with the ThermoFisher QuBit BR RNA (Q10211), and RNA integrity assessed with the Agilent RNA ScreenTape assay (5067-5579, 5067-5577, 5067-5576) on the Agilent 4200 TapeStation.

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

4.2.5 RNA-seq quantification and preprocessing

A total of 1141 samples from 396 patients were sequenced to a median depth of ∼20 million read pairs. Sequencing data was demultiplexed with Picard*. Sequence quality, overrepresented sequences, adapter content, and sequence duplication rates were checked using FastQC [371]. Reads were mapped to GRCh38 using STAR (v2.6.1d) [372] and deduplicated to unique reads using UMI-tools [373]. Gene expression was quantified against the Ensembl 96 gene annotation with `featureCounts` (v1.6.4) [374].

Samples were filtered to remove outliers (>2 standard deviations from the mean) according to percentage of aligned reads in coding regions reported by Picard, percentage of unique reads, and number of unique reads. Samples that could not be mapped to a timepoint according to Section 4.2.2 were removed. Samples with sex mismatch were removed. Samples from patients with grey zone primary response were removed. Samples for which there was missingness in the data matrix for variables considered in the variable selection process (Section 4.2.6.1) were removed. A total of 814 samples remained after filtering. The number of samples mapping to each timepoint as defined in Section 4.2.2 is shown in Fig. 4.1. The number of samples per patient ranged from one to four, with a median of three (Fig. 4.2).

The Ensembl 96 gene annotation contains 58 884 genes, many of which are not expressed in whole blood. Effective library sizes were computed using the trimmed mean of M-values (TMM) method in `edgeR` [203], then between-sample normalisation for library size was performed using `edgeR::cpm`, converting counts to counts per million (CPM). Genes with low expression were filtered, requiring >1.25 CPM in >10 % of samples (1.25 CPM being approximately 10 counts at the median library size of 8 million unique mapped read pairs) and non-zero expression in >90 % of samples. Globin genes and short non-coding RNAs (ncRNAs) were removed. A total of 15 511 genes remained after filtering. Finally, CPMs were converted to the $\log_2$ scale, and precision weights to account for the expression mean-variance relationship were computed for each gene and sample using `variancePartition::voomWithDreamWeights` [375].

4.2.6 Differential gene expression

4.2.6.1 Variable selection by variance components analysis

For each gene, the DGE model was a regression expressing the response variable (gene expression), as a linear function of predictor variables of interest (primary response status, drug, timepoint), and other selected predictor variables. In estimating the association of predictor X to response Y by regression, adjustment for a third variable Z can increase, decrease, or even reverse the effect estimate for X (the regression coefficient). I aimed to select third variables for inclusion into the DGE model that were covariates, defined here as a Z that is associated with Y, but not with X. Such variables are also known as neutral controls [376], precision variables, or prognostic variables. At the cost of one degree of freedom (df), Z explains variation in Y that would otherwise be considered residual, so conditioning on Z increases the efficiency of estimating the effect of X on Y, but does not change the effect estimate.

*https://broadinstitute.github.io/picard/

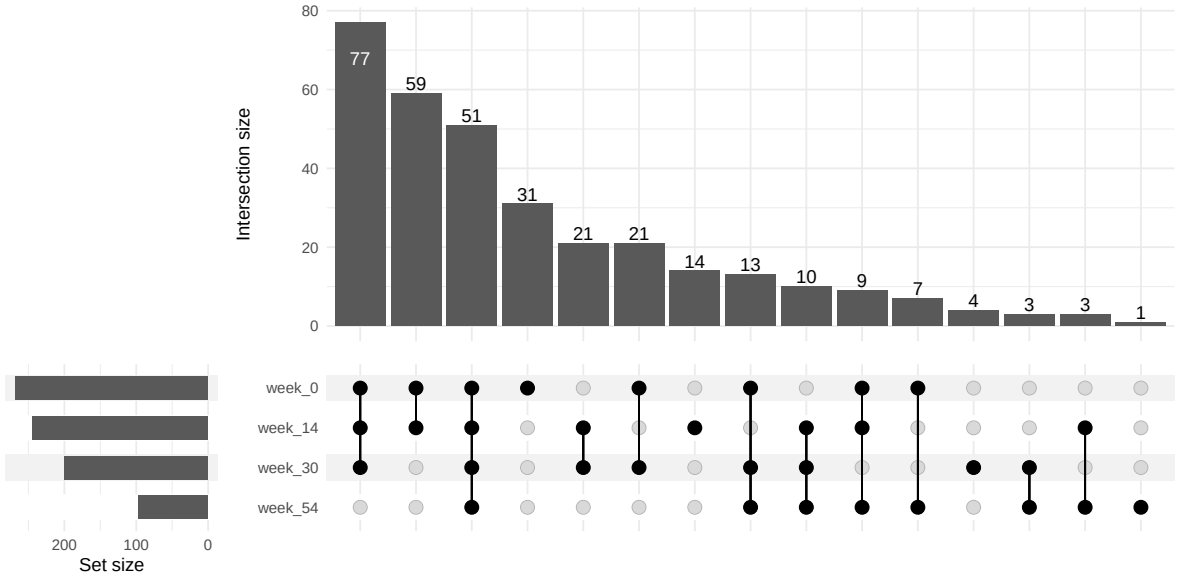

Figure 4.2: Distribution of RNA-seq samples from each patient among timepoints.

Many variables were available for selection; Fig. 4.3 shows their correlation matrix. These included three variables associated with primary response in Kennedy *et al.* [355]: baseline immunomodulator use, smoking, and BMI. Also available were proportions of six common cell types in whole blood ($CD4^+$ T cells, $CD8^+$ T cells, B cells, NK cells, monocytes, granulocytes), estimated using the Houseman method (`minfi::estimateCellCounts` [377]) from whole blood Illumina MethylationEPIC methylation array data collected from the same patients and timepoints. The Houseman method uses differentially methylated regions between immune cell types as cell type markers [378].

A variance components analysis was performed to quantify the proportion of expression variance explained by each variable for each gene using `variancePartition` [375]. Variables that do not explain much variation in the response are unlikely to improve efficiency if conditioned on. The model was a mixed effects regression model with variables in Fig. 4.3 included as predictors. Additional categorical variables were included for patient and RNA-seq library preparation plate. An additional continuous variable consisting of random numbers drawn from the standard normal distribution was included as a null. Granulocyte proportion estimates were dropped to relieve perfect multicollinearity. Categorical variables were coded as random intercepts, and continuous variables as fixed effects. Surprisingly, simulations from Hoffman *et al.* [375] showed variance proportion estimates were unbiased even when coding categorical variables with as few as two categories as random effects, as long as model parameters were estimated using maximum likelihood (ML) rather than restricted maximum likelihood (REML). It was also shown this approach avoids overestimates of variance proportions that occur if categorical variables with many levels are treated as fixed.

As downstream DGE methods require the same set of predictors for all genes, I aimed to select variables that explained a lot of variance for many genes. Variables that explained the most variance on average were patient, cell proportions, and RNA-seq plate (Fig. 4.4). Some variables that did not explain more variance on average than the null nevertheless had high maximum

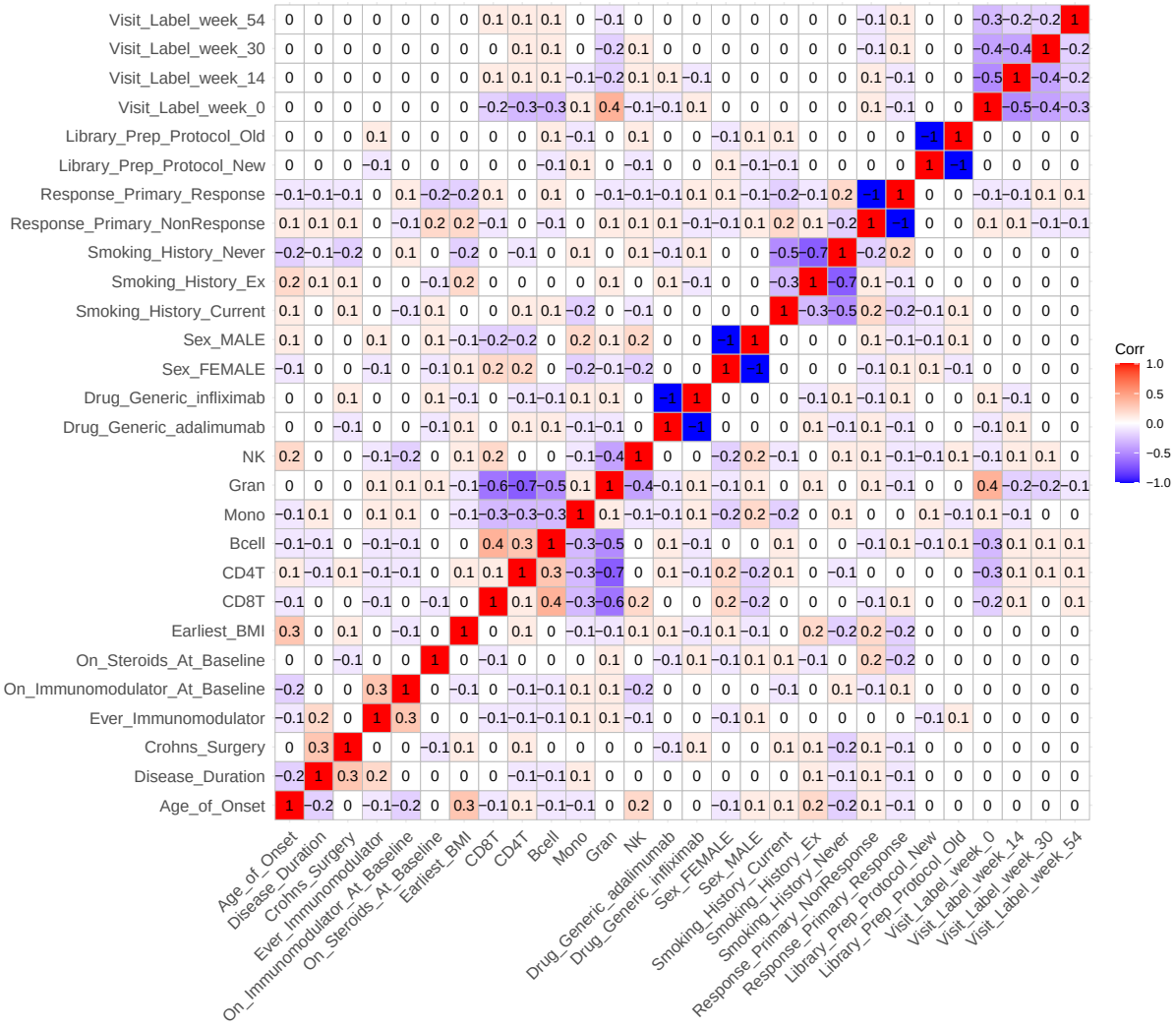

Figure 4.3: Correlation matrix of variables measured in PANTS that were considered as potential predictor variables. NK = NK cell, Gran = granulocyte, Mono = monocyte, Bcell = B cell, CD4T = $CD4^+$ T cell, CD8 = $CD8^+$ T cell.

values, indicating their importance for a relatively small number of genes. These included sex, library preparation protocol version, and smoking status. However primary response status—a variable of interest—also fell into this group, so it was difficult to justify excluding all variables with lower median variance explained than the null. Consequently, all non-null variables in Fig. 4.4 were selected as predictors in downstream models apart from "Ever_Immunomodulator" (whether the patient had ever had immunomodulator treatment), as that variable had both low median variance explained and was correlated with baseline immunomodulator use. This is a crude approach, but the sample size is large compared to number of df lost by including predictors that may not be relevant for some genes.

How might interpretations of effect sizes of interest be affected by including this suite of other variables, all of which can be considered as third variables? If a third variable Z is not a precision variable, but is also associated with X, conditioning on Z changes the effect estimate of X on Y. The regression model is mathematically agnostic to causal relationships between variables, but distinct types of third variable can be distinguished conceptually by assuming the direction of causal relationships [379]. Conditioning on a confounder ($X \leftarrow Z \rightarrow Y$) reduces bias of the effect estimate, conditioning on a collider ($X \rightarrow Z \leftarrow Y$) induces bias, and conditioning on a mediator in the causal pathway ($X \rightarrow Z \rightarrow Y$) changes the effect estimated by removing the indirect effect mediated by Z, usually biasing the effect estimate towards zero*.

From the variance components analysis shown in Fig. 4.4, cell proportions were among the biological factors that explained the most variance on average; they are one of the largest sources of variation in bulk blood expression data, and are a major driver of transcriptional response to immune perturbations [381]. Thus I decided to I fit two sets of separate DGE models including and excluding cell proportions as predictors, but otherwise identical. Assuming that cell proportions act as a mediator of the drug's effect on gene expression, these models have complementary interpretations. In models without cell proportions included, differential expression after drug perturbation could represent up or downregulation on a per-cell basis, but could also come from differences in cell proportions induced by the drug. The estimates from models adjusted for cell proportions are more likely to reflect up or downregulation on a per-cell basis. When comparing expression between responders and non-responders, one might also assume cell proportions can mediate the effect of a patient's response status on expression†. Analogously, estimates of expression differences between responders and non-responders from the two sets of models also have complementary interpretations: total difference, and per-cell differences not due to differences in cell proportions. Throughout this chapter, I interpret the estimates from both sets of models accordingly.

*It is not easy to determine the direction of bias (positive or negative) for any of these cases in general [380].

†The assumption that response status is a stable property of a patient that can be treated as a predictor/independent variable will be discussed in Section 5.2.

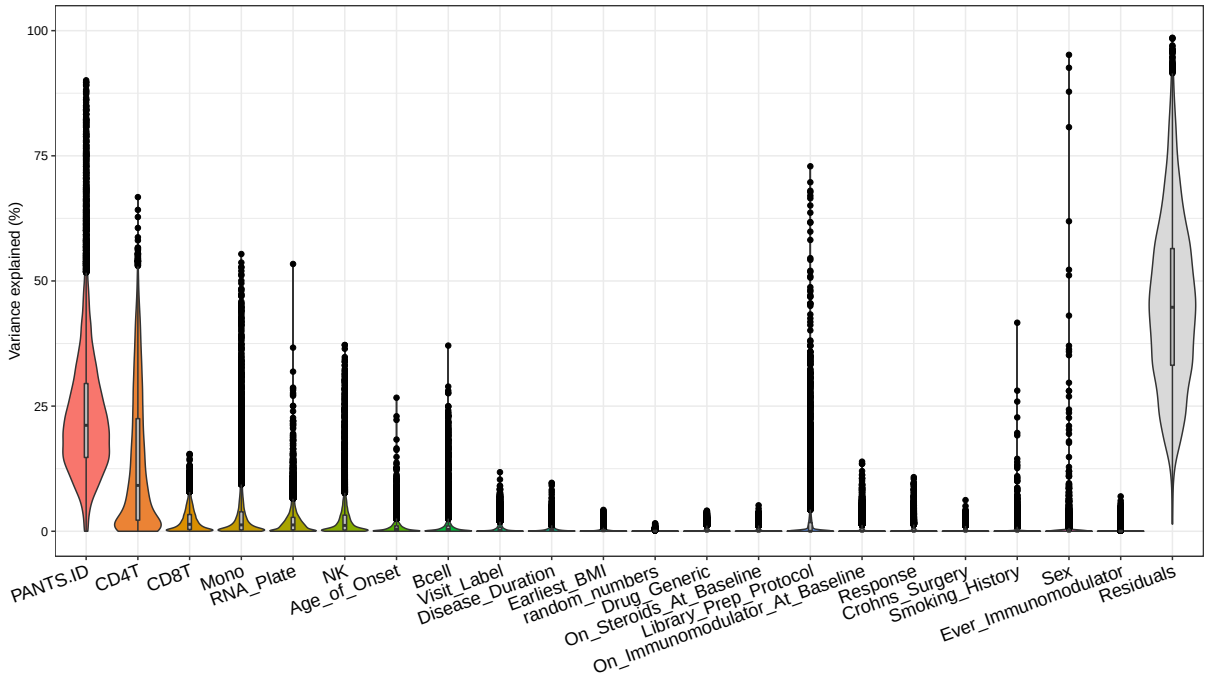

Figure 4.4: Variance components analysis showing the distribution of per-gene percentage of variance in expression explained by each variable. Variables are ordered by the median of per-gene variance explained estimates. random_numbers is a null drawn from the standard normal distribution. PANTS.ID = patient ID, NK = NK cell, Gran = granulocyte, Mono = monocyte, Bcell = B cell, CD4T = $CD4^+$ T cell, CD8 = $CD8^+$ T cell.

4.2.6.2 Contrasts for pairwise group comparisons

Per-gene DGE models were fit in `dream` [382]. Like the variance components analysis models, these DGE models were linear mixed models:

$$y = 0 + \beta_{trd} G_{trd} + \sum^{9} \beta_Z Z + (\sum^{5} \beta_C C) + u + v + \epsilon \tag{4.1}$$

where:

- The response variable is gene expression y.
- 0 indicates there is no intercept term.
- G_{trd} is a fixed effect for experimental group defined by combinations of the predictors of interest: timepoint (week 0, 14, 30, 54), response (responder, non-responder), and drug (infliximab, adalimumab). This is equivalent to having an intercept term and a three-way interaction between visit, response, and drug, including all lower order terms, but is more convenient for testing pairwise expression differences between groups, as the coefficient for each term is the estimate of mean expression for that group.
- $\sum^9 \beta_Z Z$ are the non-cell proportion fixed effects chosen in Section 4.2.6.1: sex (Sex), age of disease onset (Age_of_Onset), disease duration (Disease_Duration), smoking history (Smoking_History: ex, current or never), whether the patient has had surgery for CD (Crohns_Surgery), whether the patient was on immunomodulator at baseline

(On_Immunomodulator_At_Baseline), whether the patient was on steroids at baseline (On_Steroids_At_Baseline), BMI at baseline (Earliest_BMI), and library preparation protocol version (Library_Prep_Protocol).

- $\sum^5 \beta_C C$ are the cell proportion fixed effects chosen in Section 4.2.6.1, for NK cells, monocytes, B cells, $CD4^+$ T cells, and $CD8^+$ T cells.
- u is a random intercept for RNA-seq plate (RNA_Plate).
- v is a random intercept for patient (PANTS.ID), nested inside RNA-seq plate.

As the interest is in estimating a single coefficient for each predictor's effect size on expression (rather than estimating variance components), most predictors above are modelled as fixed effects. Since RNA-seq plate and patient are nuisance variables with a large number of levels, they are modelled as random intercepts. A total of four sets of per-gene models were fit, with and without the cell proportion terms $\sum^5 \beta_C C$, and replacing $\beta_{trd} G_{trd}$ (separate drug models) with $\beta_{tr} G_{tr} + \beta_d d$ (pooled drug models) or not. Unlike with variance components analysis, to avoid small-sample bias in estimates of fixed effect standard errors, REML was used for estimation [383].

Specific hypotheses were tested using sum-to-zero contrasts, which are linear combinations of model coefficients with weights summing to zero. For example, to test for DGE between responders and non-responders to infliximab at baseline in the non-pooled model, I used a contrast where the weight for the week 0/responder/infliximab group coefficient was 1, the weight for the week 0/non-responder/infliximab group coefficient was -1, and all other coefficient weights were 0. To get p-values, the contrast divided by its standard error was compared to the t-distribution using the Satterthwaite approximation for df. False discovery rate (FDR) was controlled with the Benjamini-Hochberg (BH) method, with threshold set at 0.05, computed separately for each contrast*.

4.2.6.3 Spline model of expression over time

The aim was to use expression data from all four timepoints to find genes associated with response, while avoiding a large number of pairwise comparisons. I fit a natural cubic spline (`splines::ns`, R 3.6.2) to the study day to allow for non-linear trajectories of expression over time. A cubic spline is a continuous function defined piecewise in each successive interval between a set of k knots in the range of the input variable. The $k-1$ pieces between knots are polynomials of degree 3. For a natural spline, the function is constrained to be linear outside of the boundary (first and last) knots to avoid unpredictable behaviour at the boundaries [384]. I set two inner knots at week 14 and week 30, as expression is expected to change after each drug dose. To include all data within the boundaries, the two boundary knots were set at the minimum and maximum values of study day rather than week 0 and week 54. A basis matrix [384] was computed with `ns(Study_Day, knots=7*c(14, 30))`, which is a matrix with 3 columns, each column being a transformation of the input, study day. The columns are fit in the regression model in place

*FDR could also have been computed globally over all contrasts if it were necessary to have the same t-statistic threshold for statistical significance in all contrasts.

of study day to allow for non-linear effects of study day on expression. The model form used was as in Eq. (4.1), except with $\beta_{trd}G_{trd}$ replaced by $\beta_r r + \sum^3 \beta_b b + \sum^3 \beta_{rb} rb + \beta_d d$, where r is response status, d is drug, $\sum^3 \beta_b b$ are the three columns of the basis matrix, and $\sum^3 \beta_{rb} rb$ are the second-order interaction terms between response status and the basis matrix columns. Separate sets of per-gene models were again fit with and without cell proportions $\sum^5 \beta_C C$.

When testing for response-associated differences in the spline parameters, the predictors of interest are the interaction terms $\sum^3 \beta_{rb} rb$. The three terms were tested jointly with an F-test, and FDR correction was performed with the BH method, with the threshold set at 0.05. A significant result indicates a significant difference in the trajectory of expression over study day between responders and non-responders.

4.2.6.4 Clustering expression over all timepoints

I clustered genes by their expression trajectories to define sets of genes with similar trajectories over time. This was done to aid the interpretation of significant genes from the cell proportion-adjusted spline model using gene set enrichment analysis. Expression data was converted to the CPM scale using TMM normalisation factors, then regressed against cell proportions. Residuals were centered and scaled per gene. A distance matrix was computed using $1 - r$ as the distance metric, where r is the Pearson correlation. Hierarchical clustering was performed with complete agglomeration for inter-cluster distance (`fastcluster::hclust(method = "complete")`, [385]). The optimal number of clusters was assessed by the gap statistic (`factoextra::fviz_nbclust(method = "gap_stat", nboot = 500)`*), which determines when the change in within-cluster dispersions are no longer significantly improved by increasing the number of clusters [386]. The default `firstSEmax` criteria was used to choose the optimal number of clusters k, which finds the first local maximum at m clusters where $\text{Gap}(m) \geq \text{Gap}(m+1)$, then finds the smallest $k : 1 \leq k \leq m$ such that $\text{Gap}(k)$ is not less than $\text{Gap}(m)$ minus the bootstrapped standard error of $\text{Gap}(m)$. The hierarchical clustering tree was then cut into k clusters.

4.2.6.5 Gene set enrichment analyses

Rank-based gene set enrichment analyses were conducted using `tmod::tmodCERNOtest` [241] and blood transcription modules (BTMs), as described in Section 2.2.10. For each contrast, as the t-statistics are not comparable between genes due to the use of approximate df, I ranked genes by the signed z-score reported by `dream`, which is a monotonic transformation of the p-value. Similarly, moderated F-statistics from the spline model are not comparable between genes, so I used the signed F-statistic reported by `dream` from the transformation of the p-value.

Gene set overrepresentation analyses with the hypergeometric test were conducted with `tmod::tmodHGtest` as detailed in Section 3.2.11.

4.2.7 Genotyping and genotype data preprocessing

Genotype data were subsetted from the post-quality control PANTS cohort genotypes generated by Sazonovs *et al.* [370], where the genotyping and preprocessing pipeline is fully described. In brief,

*https://rpkgs.datanovia.com/factoextra/index.html

whole blood samples were collected into EDTA tubes at week 0 and genotyped on the Illumina CoreExome genotyping array. Pre-imputation quality control was performed in accordance with de Lange *et al.* [180]. Imputation was performed using the Sanger Imputation Service with the Haplotype Reference Consortium panel. Post-imputation, samples that were non-European, related (proportion identity-by-descent > 0.1875), or were outliers in genotype missingness or heterozygosity rate were removed; SNPs that were poorly imputed (INFO score < 0.4), deviated from Hardy-Weinberg equilibrium (HWE) ($p < 1 \times 10^{-10}$), had high missingness (>5 %), or low minor allele frequency (MAF) (<1 % before subsetting) were removed. 7 503 762 SNPs remained after filtering. Genotypes were converted to dosages of the non-reference allele.

4.2.8 reQTL mapping

The overall strategy and methods used were largely identical those to those used in Chapter 3, laid out in Section 3.2. Differences are described below.

4.2.8.1 Computing genotype principal components

Samples were projected onto principal components (PCs) defined by 1000 Genomes Project samples using SNP weights from akt*, confirming that samples were of European ancestry (Fig. 4.5). Here I chose the first five PCs for use as covariates in expression quantitative trait locus (eQTL) mapping downstream, one more than was chosen in Section 2.2.5 for Human Immune Response Dynamics (HIRD) by the Tracy-Widom test. This should be sufficient to adjust for large-scale population structure, as the PANTS cohort is less ethnically diverse than the HIRD cohort. PCs were centered and scaled before downstream use to improve model convergence.

4.2.8.2 Finding hidden confounders in expression data

Between-sample normalisation and variance stabilisation was applied to the counts matrix (`DESeq2::vst` [237]), resulting in $\log_2$ scale expression estimates. Akin to Section 3.2, given known factors (response, drug, five scaled genotype PCs, five cell proportions), PEER [182] was used to infer additional hidden factors that explain variance in the expression matrix for a large fraction of genes. This is similar in principle to the variance components analysis carried out in Section 4.2.6.1, except hidden factors can be unmeasured. To maximise efficiency for *cis*-eQTL mapping, the number of PEER factors retained for each timepoint was selected to maximise the number of genes with at least one significant eQTL (eGenes) detected on chromosome 1 (Fig. 4.6). The selected numbers were 25, 20, 15, and 5 factors; for weeks 0, 14, 30, and 54 respectively.

4.2.8.3 Computing kinship matrices

Akin to Section 3.2.3, leave-one-chromosome-out (LOCO) kinship matrices were computed on typed SNPs for each chromosome using LDAK [274].

*https://github.com/Illumina/akt

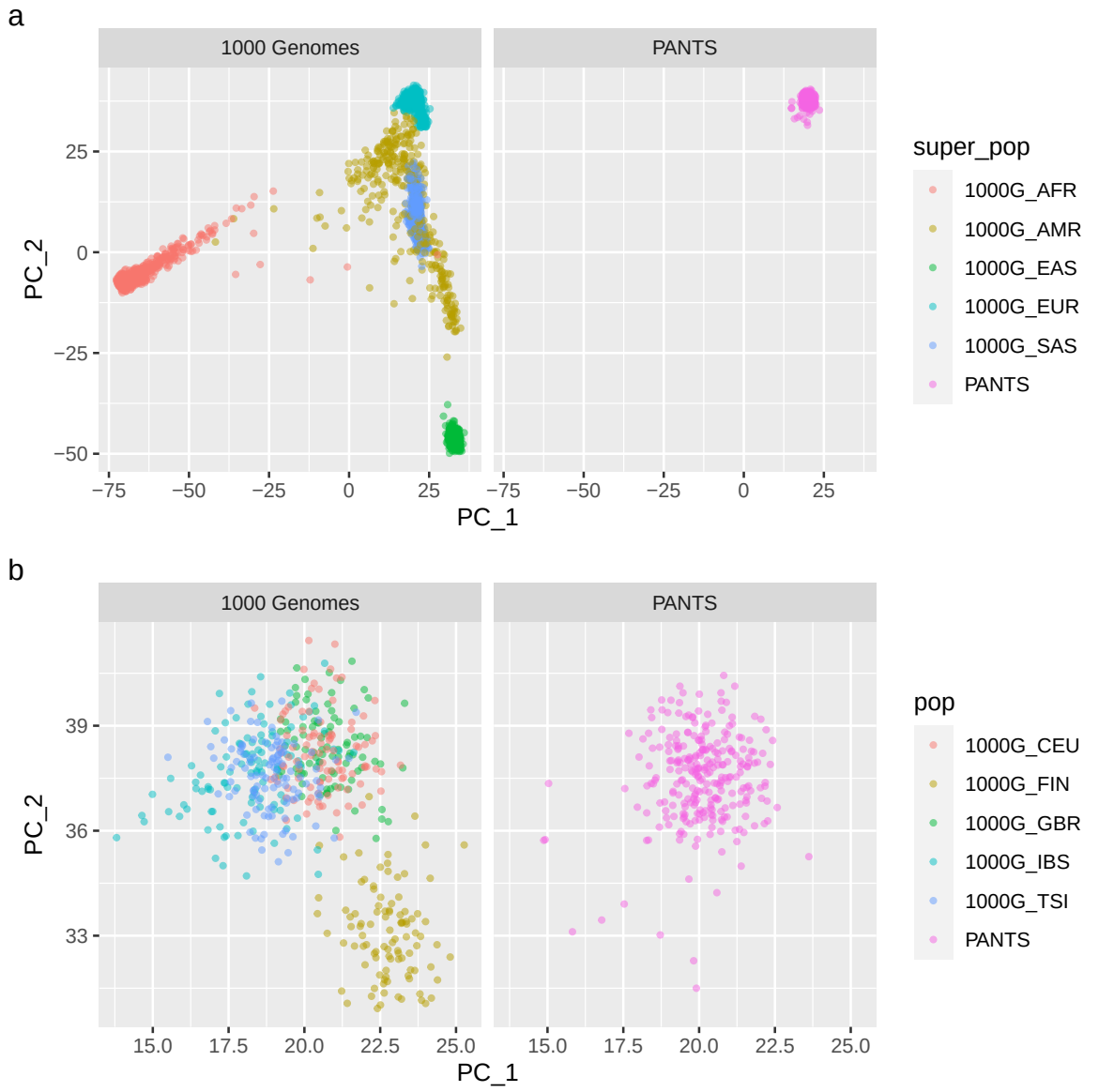

Figure 4.5: 1000 Genomes Project (1000G) samples and PANTS samples projected onto 1000G genotype PC1 and PC2 axes, colored by (a) superpopulation and (b) population. 1000G superpopulations: AFR = African, AMR = Ad Mixed American, EAS = East Asian, EUR = European, SAS = South Asian. 1000G European populations: CEU = Utah Residents (CEPH) with Northern and Western European Ancestry, FIN = Finnish in Finland, GBR = British in England and Scotland, IBS = Iberian Population in Spain, TSI = Toscani in Italia.

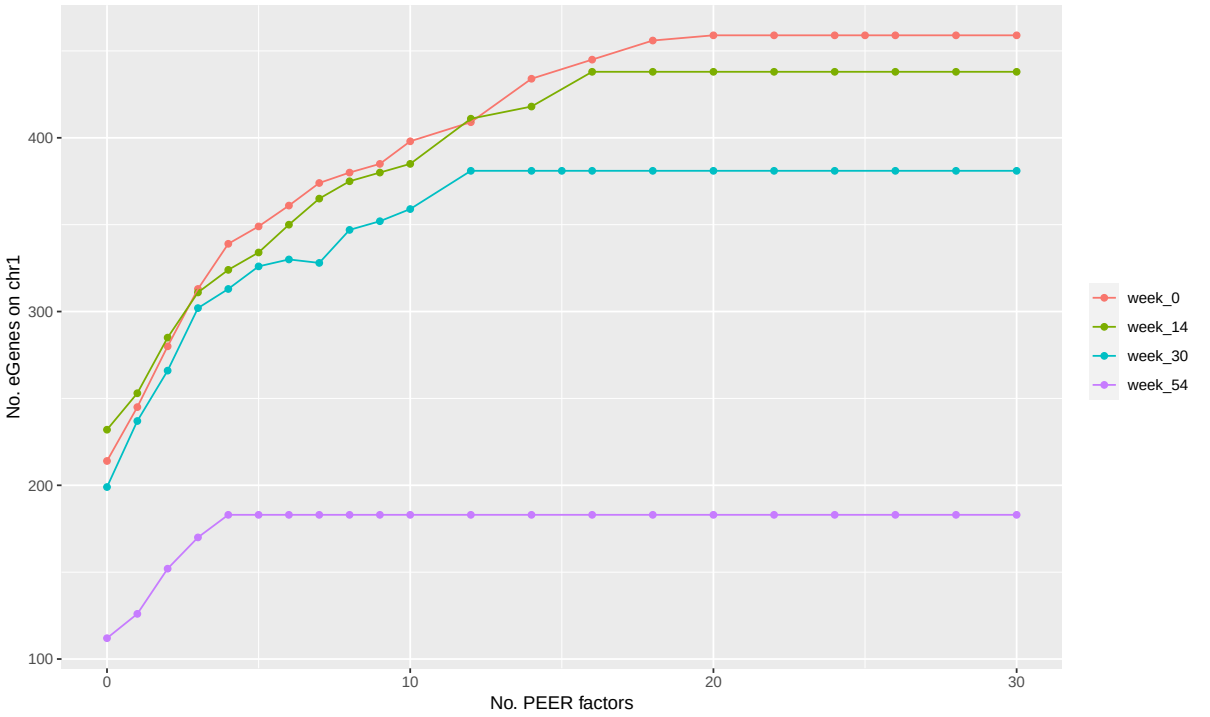

Figure 4.6: Number of eGenes on chromosome 1 vs. number of PEER factors included in eQTL mapping as covariates. FDR computed with hierarchical Bonferroni-BH [305] with significance threshold set at 0.05.

4.2.8.4 Mapping eQTLs per timepoint

As in Section 3.2, eQTLs were mapped in each timepoint using a linear mixed model in LIMIX [273]. The sample sizes with both genotype and expression data available for eQTL mapping at weeks 0, 14, 30 and 54 were 223, 205, 167, and 84 respectively.

For each autosomal gene, *cis*-SNPs within 1 Mbp of the Ensembl gene start (or gene end on the minus strand), were filtered to keep SNPs where the number of samples homozygous for the minor allele was at least five. Small group numbers lead to data points with high leverage that may be unduly influential on the genotype beta. Assuming HWE, this is equivalent to a MAF filter of $\sqrt{5/(223 \times 2)} = 0.11$ in the timepoint with the largest sample size (week 0), and $\sqrt{5/(84 \times 2)} = 0.17$ in the timepoint with the smallest sample size (week 54).

The LIMIX model for each SNP-gene pair had $\log_2$ expression as the response variable and genotype dosage as the predictor of interest. Other fixed effect predictors were the intercept, known factors (response, drug, five scaled genotype PCs, cell proportions), and PEER hidden factors (timepoint-specific number of PEER factors selected in Section 4.2.8.2). A random intercept term was also included with mean zero and covariance matrix proportional to the LOCO kinship matrix for the SNP's chromosome.

4.2.8.5 Joint reQTL mapping over all timepoints

Analogously to Section 3.2, summary statistics from per-timepoint mapping were input to `mashr` [281] to map eQTLs jointly over all timepoints. A total of 25 908 527 SNPs were testable at all four timepoints. The null correlation structure of the timepoints was estimated using null

tests within a random subset of 200 000 tests (`mashr::estimate_null_correlation_simple`). Data-driven covariance matrices representing patterns of effects across timepoints were estimated using a strong subset of tests. As the strong subset should contain eQTLs that are likely to have an effect in at least one timepoint, at each timepoint, for each gene with at least one nominally significant eQTL ($p < 0.05$), I selected the eQTL with the smallest p-value, resulting in a strong subset of 129 002 tests. The `mashr` model was then fit on the full random subset in exchangeable Z (EZ) mode, accounting for the computed null correlation and covariance matrices. Finally, posterior betas, standard errors, and local false sign rates (LFSRs) were computed for all tests using the fitted model parameters.

The lead eQTL for each gene was chosen as the eQTL with the lowest LFSR in any condition, breaking ties by highest INFO score, highest MAF, shortest distance to gene start (or end), and smallest genomic coordinate. Each lead eQTL was assessed for being a significant reQTL by a z-test for whether the difference in betas was zero, between the week 0 beta and each of the other three timepoints. Multiple testing for the number of genes was controlled using the BH FDR for each of the three comparisons separately.

4.3 Results

4.3.1 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 baseline before commencing anti-TNF therapy, and weeks 14, 30 and 54 after commencing anti-TNF therapy. After quality control, expression data was available for 15 584 genes and 814 samples. These samples come from 324 patients, whose characteristics are shown in Table 4.1. The proportion of primary non-responders is high (43.8 %) compared to the overall proportion in the PANTS cohort (23.8 % [355]). This is due to sample selection for RNA-seq to balance the sample size for each combination of drug and primary response status.

4.3.2 Baseline gene expression associated with primary response

Patient primary response to anti-TNF was defined at week 12–14 (after the induction period) according to the clinical decision algorithm from Kennedy *et al.* [355] described in Section 4.2.3, which integrates clinician assessment with changes in CRP level and HBI score. To identify differences in baseline gene expression associated with future primary response, I fit per-gene linear models at 15 511 genes, comparing week 0 gene expression in primary responders with week 0 gene expression in primary non-responders. Comparisons were performed both within infliximab-only and adalimumab-only subgroups, and with both drugs pooled. Models were run both adjusting for cell composition estimates of six immune cell types, and without adjustment. Throughout this section, the significance threshold was set at FDR < 0.05 for each comparison, and positive $\log_2$ FCs indicate increased expression in responders versus non-responders.

Without adjusting for cell composition, the largest effects were infliximab-only, with 859 genes differentially expressed. Only *KCNN3* ($\log_2$ FC $= -0.84$) was significant for the adalimumab-

Table 4.1: Patient characteristics for the PANTS RNA-seq subcohort. Values are count and percentage for categorical variables; mean and standard deviation for continuous variables; *p*-values are for the comparison between drugs.

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

only comparison, and only *SIGLEC10* ($\log_2$ FC = 0.35) was significant in the pooled analysis (Fig. 4.7). After adjustment for cell composition, there were no longer any significant genes in the infliximab-only analysis, with 856/859 genes that were significant before the comparison having a dampened effect size after correction (smaller absolute effect and same sign), suggesting many effects may be mediated by cell composition. *SIGLEC10* in the combined analysis was also non-significant after adjustment (adjusted $\log_2$ FC = 0.31, FDR = 0.05). Conversely, at the three genes downregulated in the adalimumab-only analysis that were the only significant genes post-adjustment, I observed increased significance: *PDIA5* (unadjusted $\log_2$ FC = −0.33, adjusted $\log_2$ FC = −0.35), *KCNN3* (−0.84, −0.88), and *IGKV1-9* (−1.15, −1.22).

To identify coordinately up and downregulated gene sets and increase sensitivity for detecting differences between responders and non-responders, I performed rank-based gene set enrichment analyses on the per-gene z-statistics using BTMs: annotated sets of coexpressed genes in peripheral whole blood from Li *et al.* [240] (prefixed "LI"). This module-level analysis was also run both unadjusted (Fig. 4.8) and adjusted for cell composition (Fig. 4.9).

Despite only *SAMD10* having a significantly different effect between drugs at the gene level (a significant interaction between drug and response at week 0), the large global differences observable in Fig. 4.7 were detected in the module-level analysis*. Without adjusting for cell composition, many of the most significantly upregulated modules in the pooled analysis—including upregulation of monocyte (LI.M11.0, LI.S4), neutrophil (LI.M37.1, LI.M11.2), and dendritic cell (LI.M165, LI.S11) modules—appear to be driven by an infliximab-specific effect. These modules had heavily reduced significance after adjusting for cell composition. The new modules that were most upregulated in the pooled analysis after adjustment had more consistent effects between drugs, such as MHC-TLR7-TLR8 cluster (LI.M146), antigen presentation (LI.M71, LI.M95.0), and myeloid cell enriched receptors and transporters (LI.M4.3).

For downregulated modules before adjustment, I observed infliximab-specific effects for NK cell (LI.M7.2) and T cell (LI.M7.0, LI.M7.1) modules. Adalimumab-specific effects were observed for plasma cell, B cell and immunoglobulin modules (LI.M156.0, LI.M156.0, LI.S3), and cell cycle and transcription modules (LI.M4.0, LI.M4.1). After adjustment, the significance of infliximab-specific modules was reduced, but the significance of adalimumab-specific modules and the corresponding interaction effects was increased. For both gene-level and module-level comparisons of baseline expression between responders and non-responders, there is a striking heterogeneity between patients on infliximab and adalimumab that is only partially reduced by cell proportion adjustment.

4.3.3 Assessing previously reported baseline predictors of primary response

In addition to significant genes from this study, Fig. 4.7 is annotated with genes whose expression in gut biopsies or blood has been previously evaluated for baseline prediction of primary response [362, 363, 367, 387]. Some genes expressed in gut mucosa (e.g. *IL13RA2*) were not appreciably expressed in this whole blood dataset, and most other genes that were expressed were

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

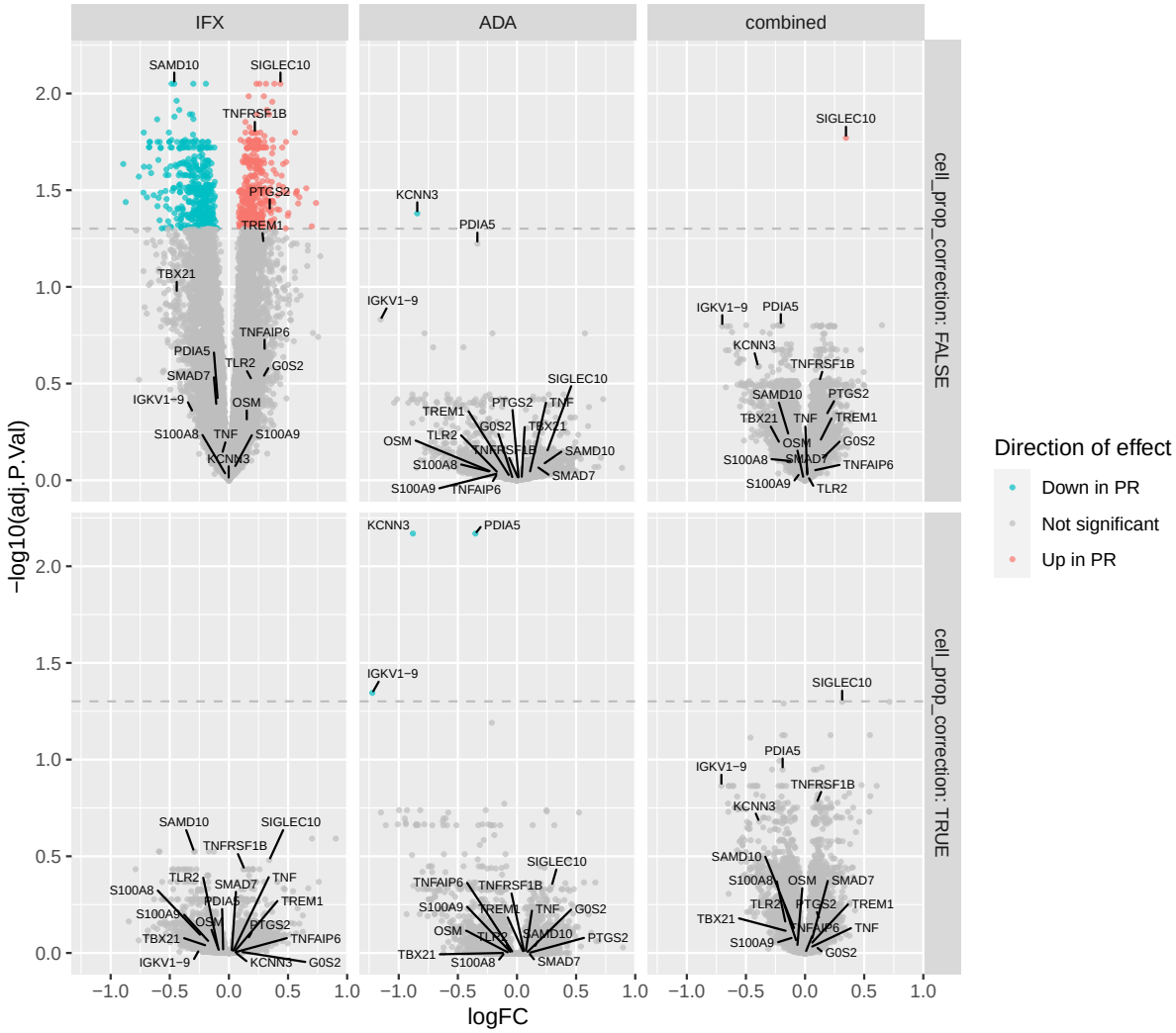

Figure 4.7: Volcano plots of DGE between primary responders (PR) and non-responders at week 0; unadjusted (top row) and adjusted (bottom row) for cell composition; for infliximab (IFX), adalimumab (ADA), or with both drugs pooled. Annotated genes include significant associations from this study and previously reported associations from Section 4.1.4. Dashed line shows significance threshold at FDR = 0.05.

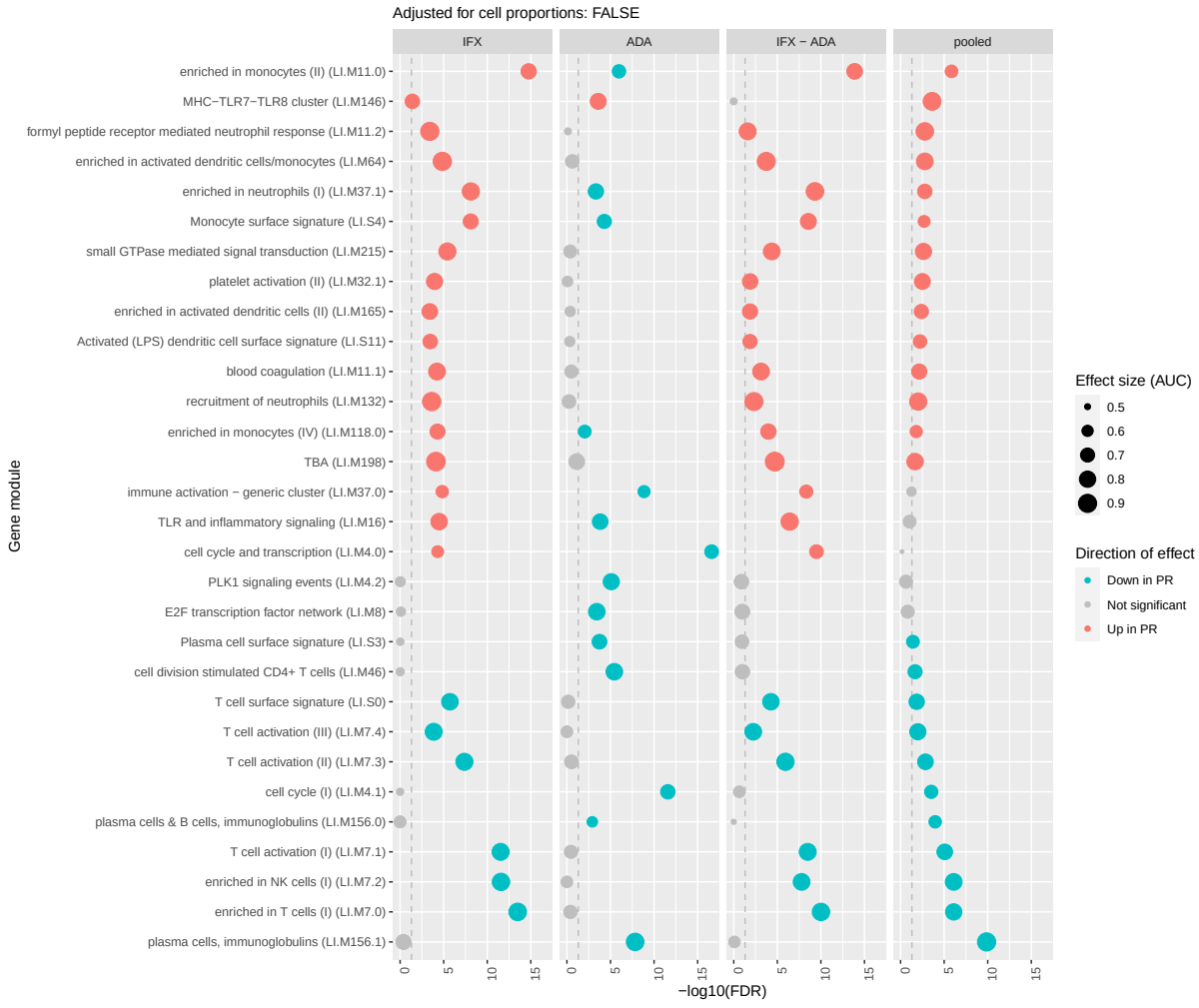

Figure 4.8: Top modules differentially expressed between primary responders (PR) and non-responders at week 0, unadjusted for cell composition. Columns correspond to results from infliximab (IFX), adalimumab (ADA), infliximab minus adalimumab difference, and pooled analyses. The top 30 modules ranked by minimum FDR in any column are shown. Vertical dashed line shows significance threshold at FDR = 0.05.

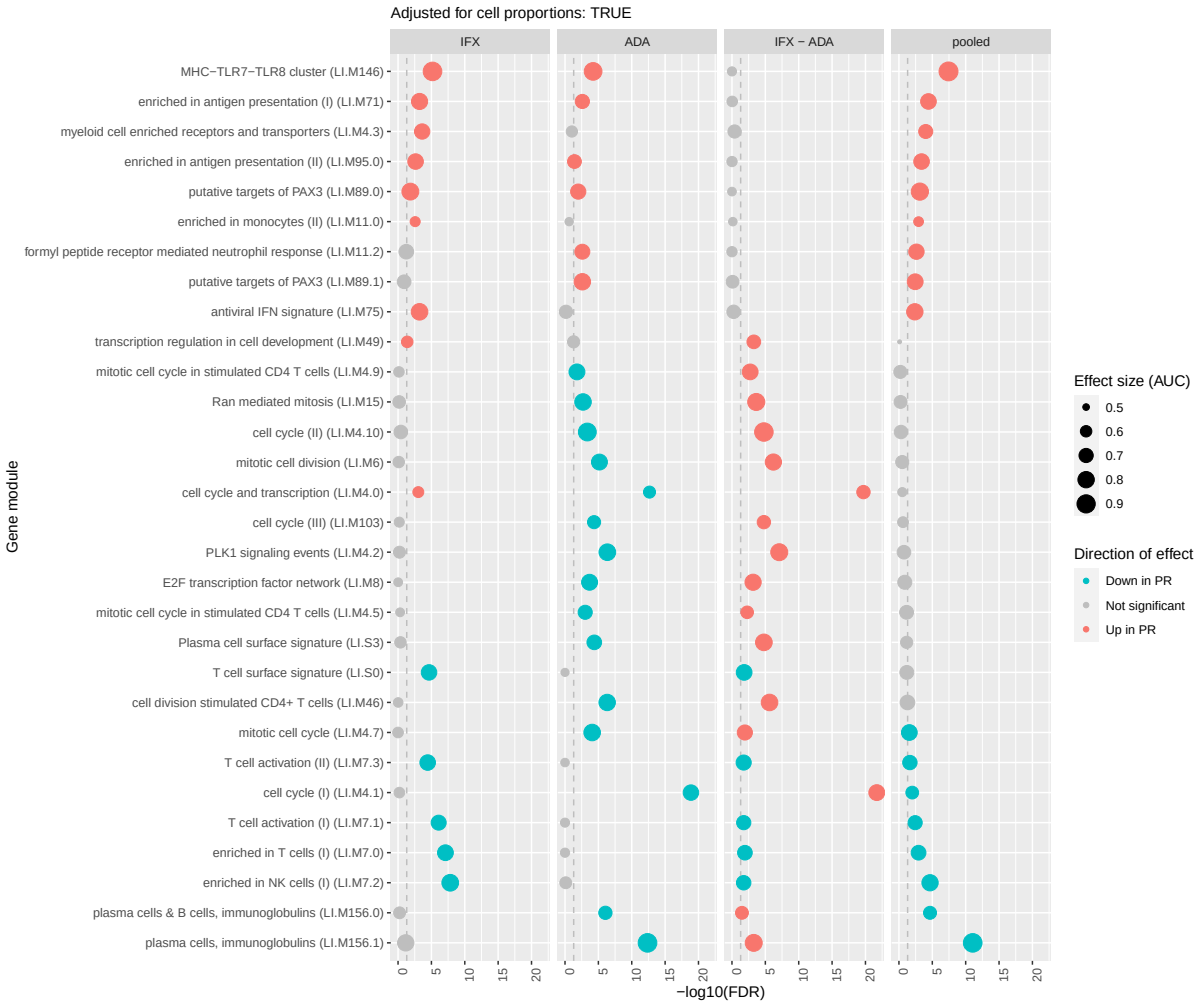

Figure 4.9: Top modules differentially expressed between primary responders (PR) and non-responders at week 0, adjusted for cell composition. Columns correspond to results from infliximab (IFX), adalimumab (ADA), infliximab minus adalimumab difference, and pooled analyses. The top 30 modules ranked by minimum FDR in any column are shown. Vertical dashed line shows significance threshold at FDR = 0.05.

not significantly differentially expressed. Only *TNFRSF1B* and *PTGS2* were associated with primary response, being upregulated at baseline in responders, specifically in the infliximab-only comparison, unadjusted for cell composition. *TNFRSF1B* was found by Verstockt *et al.* [367] to be downregulated in baseline inflamed mucosal biopsies of responders to anti-TNF therapy in IBD patients ($n = 44$, FC $= 0.72$, $p = 0.008$). *PTGS2* was found by Arijs *et al.* [362] to be downregulated in baseline mucosal biopsies of responders to infliximab for Crohn's colitis ($n = 46$). The directions of effect in both cases are opposite to this study, but comparisons are hard to draw between blood and mucosal biopsies.

A previously identified marker in blood, *TREM1* was found to have opposing effects in two studies by Gaujoux *et al.* [366] ($n = 22$) and Verstockt *et al.* [367] ($n = 54$). Here, *TREM1* showed the strongest differences between responders and non-responders at baseline in the infliximab subcohort, but did not reach significance before ($\log_2$ FC $= 0.29$, FDR $= 0.06$) nor after adjusting for cell composition ($\log_2$ FC $= 0.05$, FDR $= 0.99$). The sample size in this study for the infliximab subcohort at baseline is $n = 145$ (Fig. 4.1), so it is expected the power in this study is greater.

4.3.4 Post-induction gene expression associated with primary response

The same methodology applied at week 0 was applied at week 14 to identify differences in post-induction expression associated with primary response. A larger proportion of the transcriptome was differentially expressed between responders and non-responders at week 14: 1364 genes for the infliximab-only comparison, 1544 genes for the adalimumab-only comparison, and 4841 genes pooling both drugs (Fig. 4.10). No significant interactions between drug and response were detected at the per-gene level. Given that sample sizes at week 0 and week 14 are comparable (Fig. 4.1), the overall signal-to-noise ratio is much stronger than at baseline.

After adjusting for cell composition, 1320/1364, 1515/1544, and 4653/4841 genes had dampened effects; and the numbers of significant genes dropped to 379, 177, and 1302; for the infliximab, adalimumab, and pooled analyses respectively. This again suggests many effects are mediated by differences in immune cell composition between responders and non-responders.

Modules including generic immune activation, monocytes, TLR and inflammatory signalling, and neutrophils were downregulated in responders; whereas B cell and plasma cell modules were upregulated (Fig. 4.11). These modules remained differentially expressed with the same direction of effect after adjusting for cell composition (Fig. 4.12), suggesting there is per-cell up or downregulation on top of abundance changes of the cell types expressing these modules. Modules related to antigen presentation (LI.M71, LI.M97.0, LI.M5.0), interferon (LI.M75, LI.M127, LI.M111.1), and dendritic cells (LI.M64, LI.M165) also appeared among significantly downregulated modules after cell composition adjustment. Directions of effect for the most significant modules were largely consistent between drugs, and there were few significant drug by response interaction effects. This is in contrast to the baseline responder vs. non-responder comparison, where many of the strongest effects in the pooled analysis were driven by stronger effects in one drug.

SIGLEC10 from the baseline analysis retained its significant association with primary response post-induction, with the same direction of effect (adjusted $\log_2$ FC $= 0.37$). Some genes previously

proposed as baseline markers of response in gut mucosa—*G0S2*, *TNFAIP6*, *S100A8*, and *S100A9* by Arijs *et al.* [363]; and *OSM* by West *et al.* [364]—were differentially expressed in post-induction blood in this study. The direction of effect for both sets of markers, downregulation in primary responders, also matches this study.

4.3.5 Magnification of expression changes from baseline to post-induction in responders

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

Without adjusting for cell composition, 12 862 genes were differentially expressed in primary responders comparing week 14 vs. week 0, 8310 genes in primary non-responders, and 6320 genes had a significant interaction between responders and non-responders. After adjusting for cell composition, 5572 genes were differentially expressed in primary responders, 626 genes in primary non-responders, and 179 genes had a significant interaction. Of the genes differentially expressed between week 14 and week 0 in both primary responders and non-responders, and with a significant interaction between timepoint and response, nearly all (4885/4891 unadjusted for cell composition, 31/32 adjusted) were magnified by primary response, with the same genes having larger FCs in the same direction for primary responders (Fig. 4.13).

The most significant modules that changed from week 0 to week 14 in responders included upregulation of B cell (LI.M47.0), plasma cell (LI.M156.0), and T cell activation (LI.M7.1); and downregulation of immune activation (LI.M37.0), monocyte (LI.M11.0), neutrophil (LI.M37.1), and TLR and inflammatory signalling (LI.M16) modules (Fig. 4.14). Many of these are the same modules associated with response within the week 14 timepoint, with concordant directions of effect between the two analyses (Fig. 4.11), suggesting that a change in expression of these gene sets from week 0 to week 14 is what leads to the difference between responders and non-responders at week 14.

Adjusting for cell composition decreased the significance of a majority of modules (Fig. 4.15), especially for T cell modules in the adalimumab-only analysis. Magnification was also observed at the module level, with nearly all module effects aligned in the same direction in responders and non-responders, with significant interactions also in the same direction. In general, responders seem to experience greater changes in their gene expression from week 0 to week 14, presumably due to the drug.

4.3.6 Interferon modules with opposing expression changes in responders and non-responders

Fig. 4.13 also contains genes that were downregulated from week 0 to week 14 in responders, but upregulated in non-responders ("flipped"). At the module level, these flipped effects were apparent in the cell composition-adjusted analysis, for antiviral interferon signature (LI.M175),

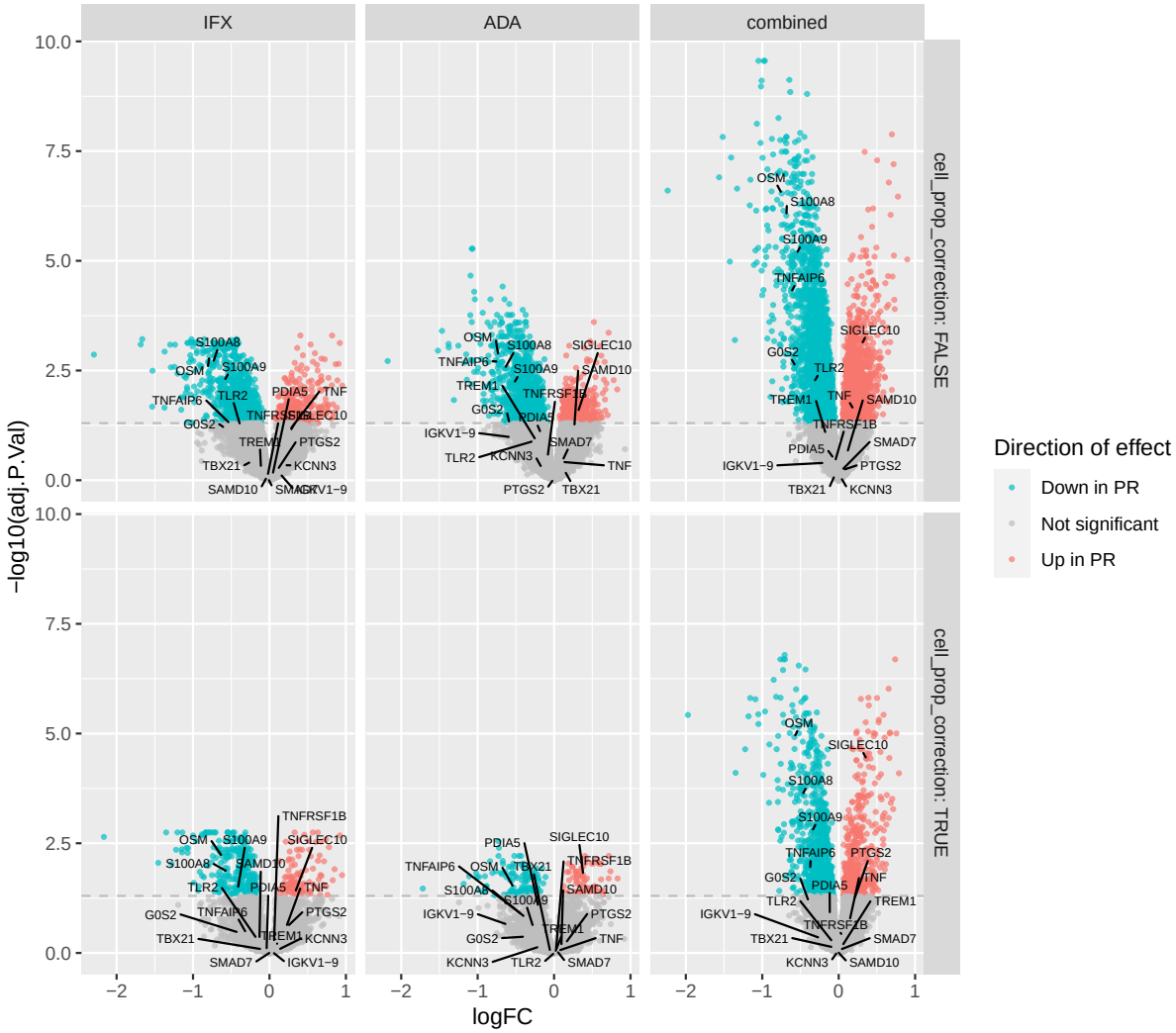

Figure 4.10: Volcano plots of DGE between primary responders (PR) and non-responders at week 14; unadjusted (top row) and adjusted (bottom row) for cell composition; for infliximab (IFX), adalimumab (ADA), or with both drugs pooled. Annotated genes include significant associations from this study and previously reported associations from Section 4.1.4. Dashed line shows significance threshold at FDR = 0.05.

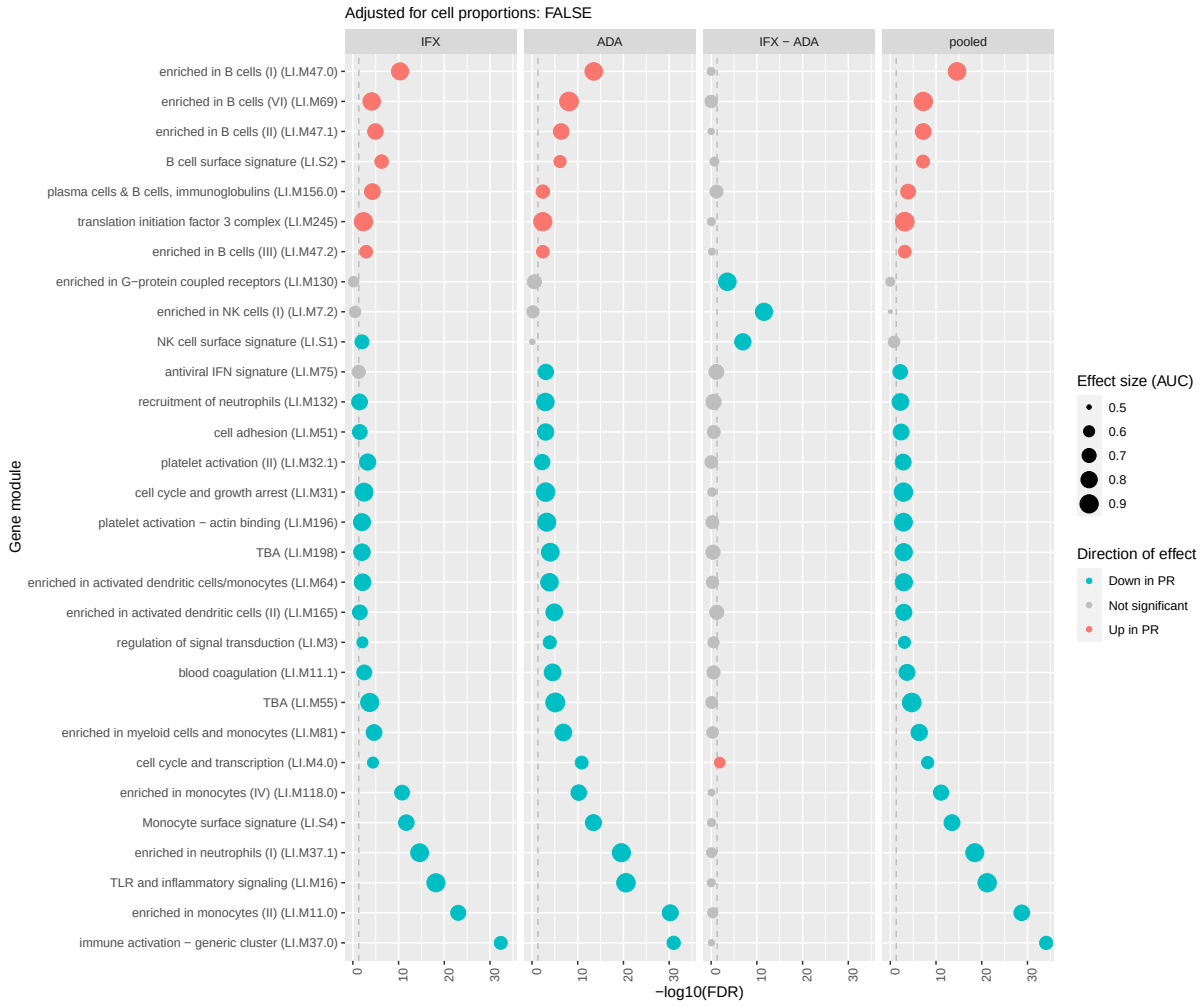

Figure 4.11: Top modules differentially expressed between primary responders (PR) and non-responders at week 14, unadjusted for cell composition. Columns correspond to results from infliximab (IFX), adalimumab (ADA), infliximab minus adalimumab difference, and pooled analyses. The top 30 modules ranked by minimum FDR in any column are shown. Vertical dashed line shows significance threshold at FDR = 0.05.

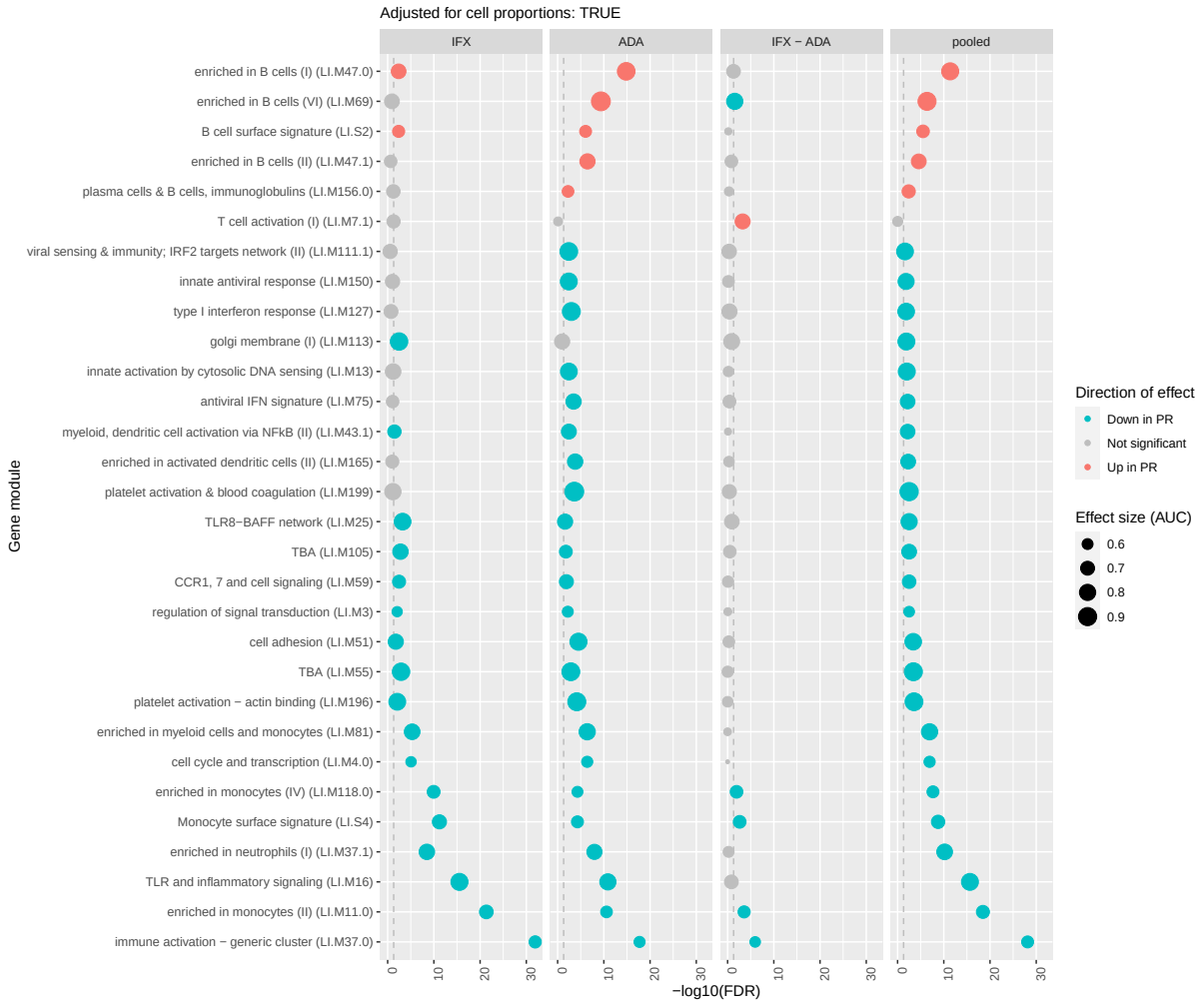

Figure 4.12: Top modules differentially expressed between primary responders (PR) and non-responders at week 14, adjusted for cell composition. Columns correspond to results from infliximab (IFX), adalimumab (ADA), infliximab minus adalimumab difference, and pooled analyses. The top 30 modules ranked by minimum FDR in any column are shown. Vertical dashed line shows significance threshold at FDR = 0.05.

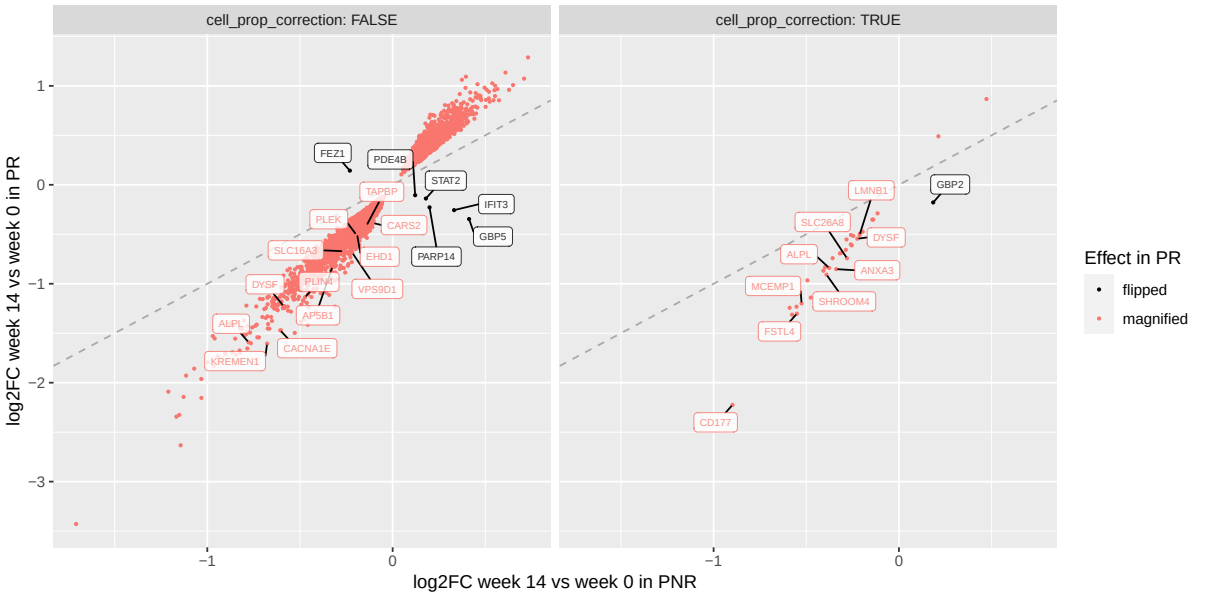

Figure 4.13: Expression $\log_2$ FC from week 0 to week 14 in primary responders (PR) versus non-responders (PNR), for genes that differentially expressed from week 0 to week 14 in both responders and non-responders, with a significantly different effect size between responders and non-responders. Results adjusted (right) and unadjusted (left) for cell proportions are shown. The identity line is shown by the dashed line. Most expression changes from week 0 to week 14 are magnified in primary responders, with a small proportion of changes in the opposite direction.

type I interferon response (LI.M127), and antigen presentation (LI.M95.0) modules (Fig. 4.15). I extended my gene set enrichment analyses to include modules from Chaussabel *et al.* [239] (prefixed "DC"); although these modules are on the whole poorly annotated compared to modules from Li *et al.* [240], interferon modules are well-annotated. *STAT2*, *GBP5*, and *PARP14* from Fig. 4.13 are annotated into an interferon module, DC.M3.4. *IFIT3* and *GBP2* are also annotated into separate interferon modules, DC.M1.2 and DC.M5.12. Adjusted for cell composition, these modules were all significantly upregulated at week 14 in non-responders only (DC.M3.4, FDR $= 3.45 \times 10^{-21}$; DC.M1.2, FDR $= 9.49 \times 10^{-16}$; DC.M5.12, FDR $= 1.36 \times 10^{-13}$; Fig. 4.16).

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

As PANTS is an observational study, it was able to include patients who continued with anti-TNF therapy even after meeting the definition of primary non-response at week 14. For both responders and non-responders, expression data was also available from blood samples around week 30 and week 54, and at additional visits scheduled in the event of secondary LOR. To test for general differences in expression over time between responders and non-responders, I fit a natural cubic spline to the expression of each gene as a function of study day. This analysis was performed only with drugs pooled due to lower sample sizes at later timepoints.

Without adjusting for cell composition, 4426 genes were differentially expressed between responders and non-responders; 210 genes were differentially expressed after adjustment. To identify distinct trajectories of expression over time, I hierarchically clustered those 210 genes by their mean expression in responders and non-responders at each timepoint, and determined the

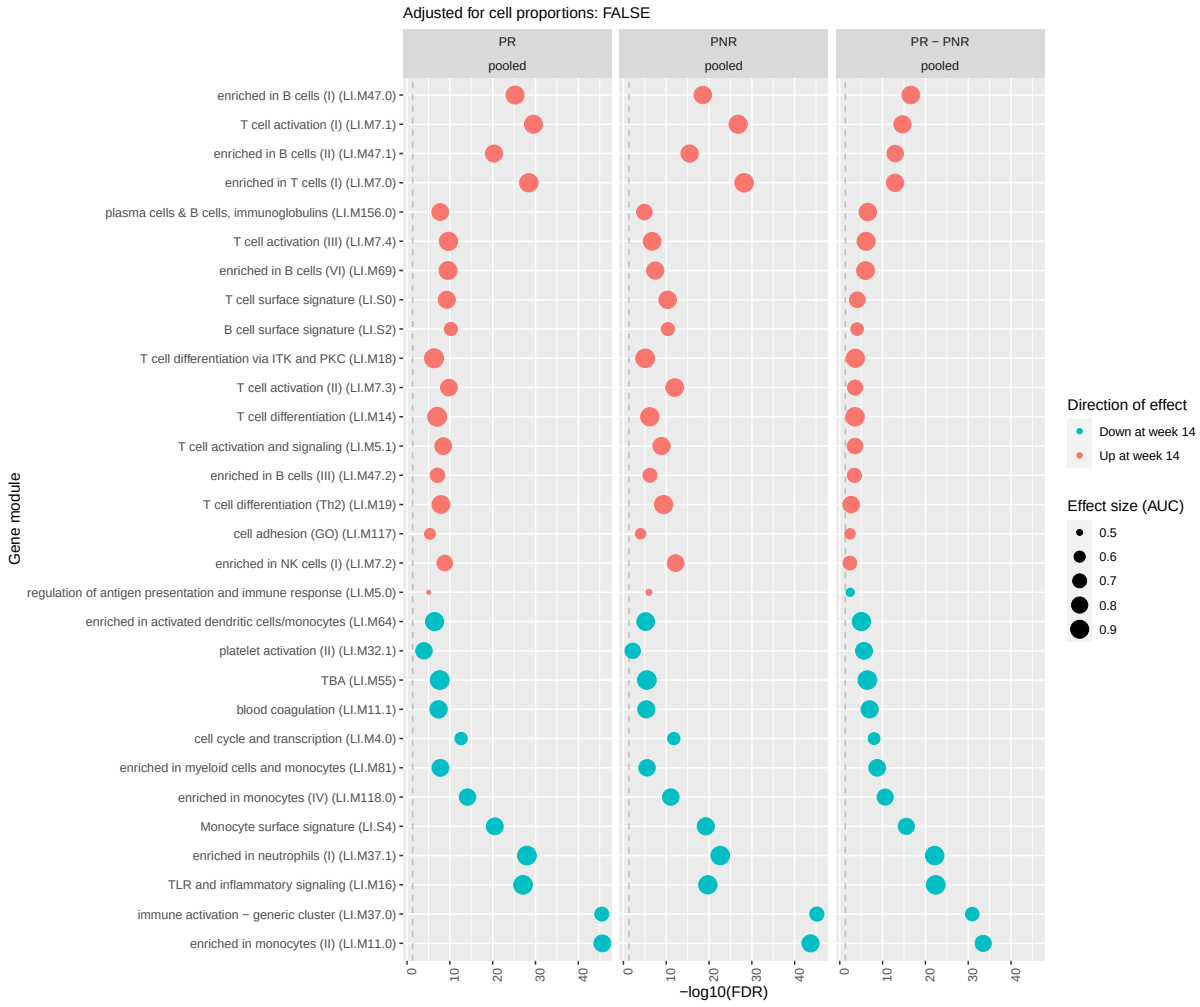

Figure 4.14: Top modules differentially expressed between week 14 and week 0, unadjusted for cell composition. Columns show effects in primary responders (PR), non-responders (PNR), and the primary responder minus non-responder difference. The top 30 modules ranked by minimum FDR in any column are shown. Vertical dashed line shows significance threshold at FDR = 0.05.

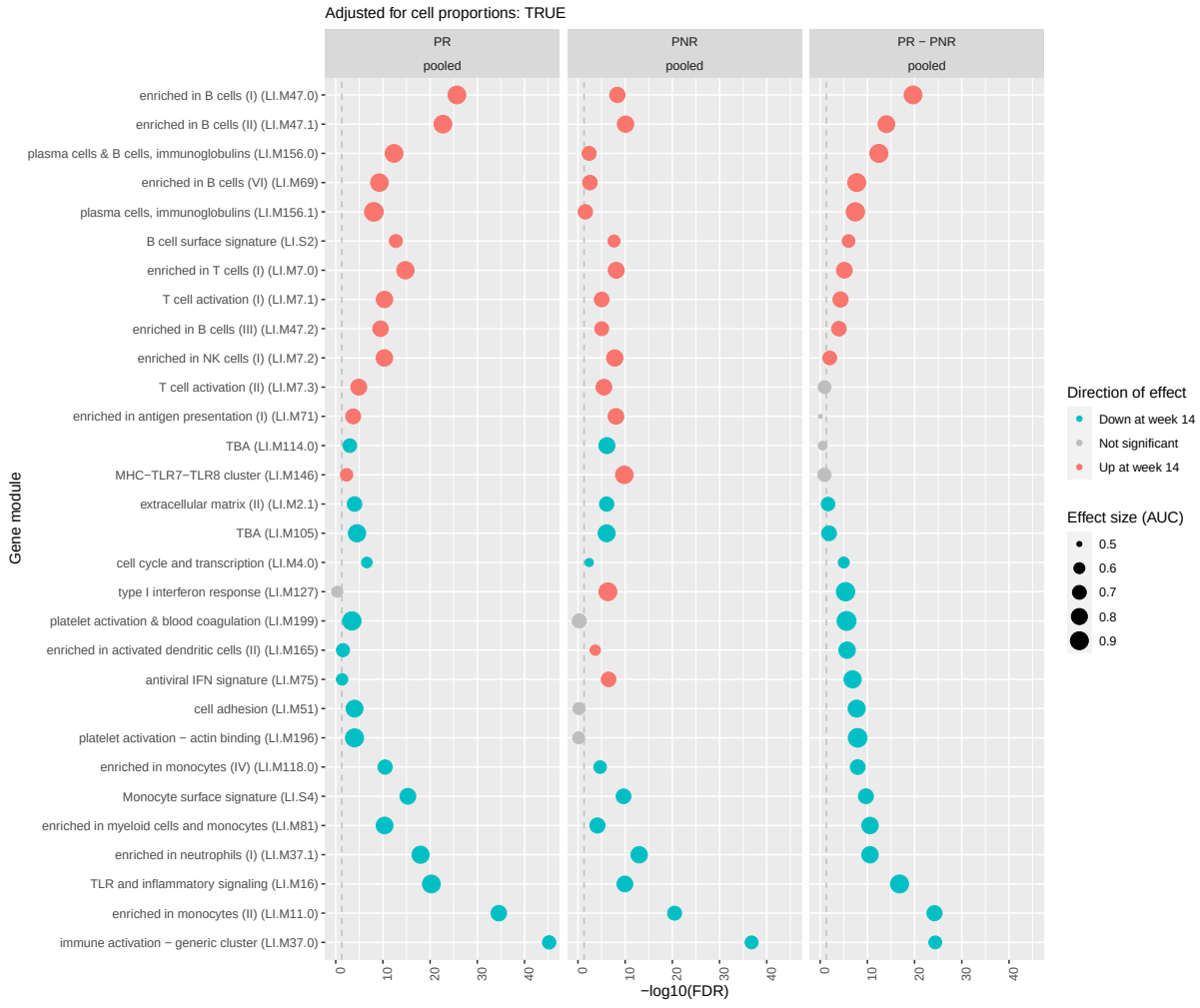

Figure 4.15: Top modules differentially expressed between week 14 and week 0, adjusted for cell composition. Columns show effects in primary responders (PR), non-responders (PNR), and the primary responder minus non-responder difference. The top 30 modules ranked by minimum FDR in any column are shown. Vertical dashed line shows significance threshold at FDR = 0.05.

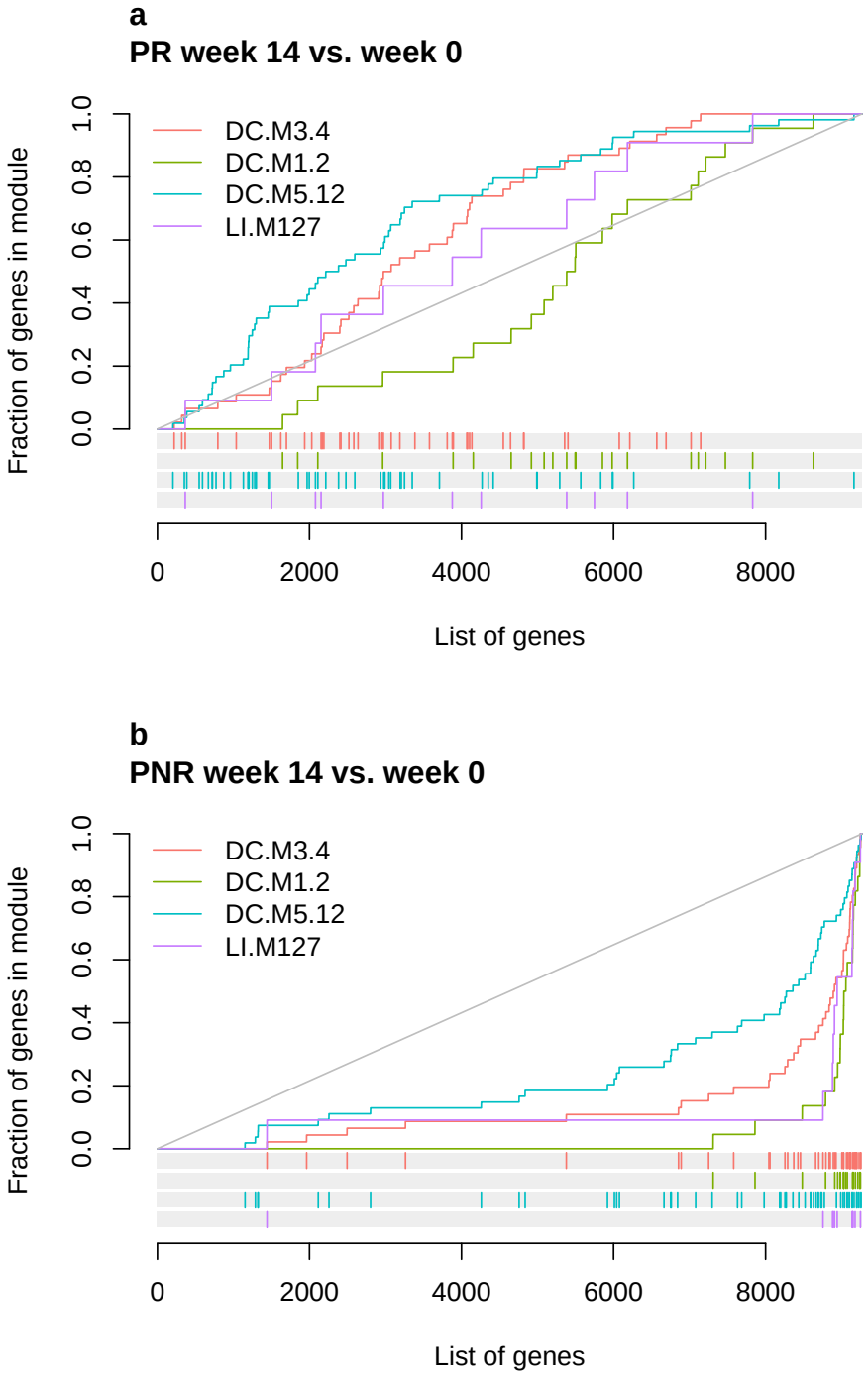

Figure 4.16: `tmod` evidence plots showing interferon-related modules specifically upregulated from week 0 to week 14 in primary non-responders (PNR), but not in primary responders (PR). Genes were ranked in ascending order by week 14 versus week 0 DGE z-statistic. The ranks of genes in interferon-related modules are indicated by colored rug plots. Colored curves show the cumulative fraction of genes in each module. For non-responders, these modules are enriched for large ranks (large, positive z-statistics). The area under the colored curves are the effect sizes (area under the curves (AUCs)). The null of randomly-distributed ranks is shown by the grey diagonal line.

optimal number of clusters by the gap statistic method (Fig. 4.17). Six distinct clusters were proposed (Fig. 4.18). Many of the 210 genes had previously been identified as having significant differences in expression between responders and non-responders either within week 14, or for change in expression from week 0 to week 14. Cluster 1 contained mainly previously identified genes (Fig. 4.19); and was enriched for modules such as myeloid cells and monocytes (LI.M81, hypergeometric test, FDR $= 2.11 \times 10^{-6}$), platelet activation (LI.M196, FDR $= 1.35 \times 10^{-5}$), immune activation (LI.M37.0, FDR $= 1.44 \times 10^{-4}$), and TLR and inflammatory signalling (LI.M16, FDR $= 2.36 \times 10^{-3}$). The spline analysis highlights that expression differences at week 14 are maintained at week 30 and week 54.

The highest proportions of genes uniquely identified as significant by the spline analysis were in cluster 2 (26/31) and cluster 4 (15/20). Cluster 2 was enriched in Li *et al.* [240] B cell modules (LI.M47.0, FDR $= 1.53 \times 10^{-6}$; LI.M47.1, FDR $= 4.53 \times 10^{-5}$) previously identified as having a greater increase from week 0 to week 14 in primary responders than in primary non-responders (Fig. 4.15), matching the observed cluster trajectory. Cluster 4 was not enriched in any modules from Li *et al.* [240], but was enriched for a B cell module (DC.M4.10, FDR $= 1.37 \times 10^{-3}$) from Chaussabel *et al.* [239]. Although no genes were significantly associated with response at week 0 (Fig. 4.7), the genes in cluster 4 were coordinately downregulated as a set in primary responders (CERNO test, $p = 6.18 \times 10^{-25}$).

Cluster 3 was enriched for type I interferon response (LI.M127, FDR $= 0.01$) and interferon (DC.M3.4, FDR $= 5.27 \times 10^{-4}$) modules, as well as genes that contain putative transcription factor (TF) binding motifs for interferon regulatory factors *IRF7* (g:Profiler [313] term ID TF:M00453_1, adj. $p = 0.01$) and *IRF8* (TF:M11684_1, adj. $p = 0.01$; TF:M11685_1, adj. $p = 0.01$). The cluster trajectory shows direction of expression change is opposing in responders and non-responders from week 0 to week 14, followed by sustained differences at week 30 and week 54. The trajectory and interferon-related gene set enrichments are consistent with those identified in Section 4.3.6. Of the nine genes in this cluster, eight genes (*STAT1*, *BATF2*, *GBP1*, *GBP5*, *IRF1*, *TAP1*, *APOL1*, *APOL2*) had significant interaction between week 0 to week 14 expression change and response status, whether or not correcting for cell composition. However, only *GBP5* was differentially expressed from week 0 to week 14 in both responders and non-responders, and only when unadjusted for cell composition (Fig. 4.13). This indicates that small and opposite effects in responders and non-responders at the gene level are best detected in the interaction analysis that explicitly tests the difference, and in the spline analysis with the support of additional data from week 30 and week 54.

4.3.8 Limited evidence for changes in genetic architecture of gene expression over time

Given the substantial changes in expression from baseline to post-induction after starting the drug, and the differing trajectories observed in responders and non-responders, I performed eQTL mapping to identity common genetic variants associated with expression that may contribute to these differences. Variants *cis* (within 1 Mb of the transcription start site (TSS)) to 15 040 genes were tested for association. Mapping was performed within each timepoint (weeks 0, 14, 30, and 54), followed by joint analysis of per-timepoint eQTL summary statistics and control for

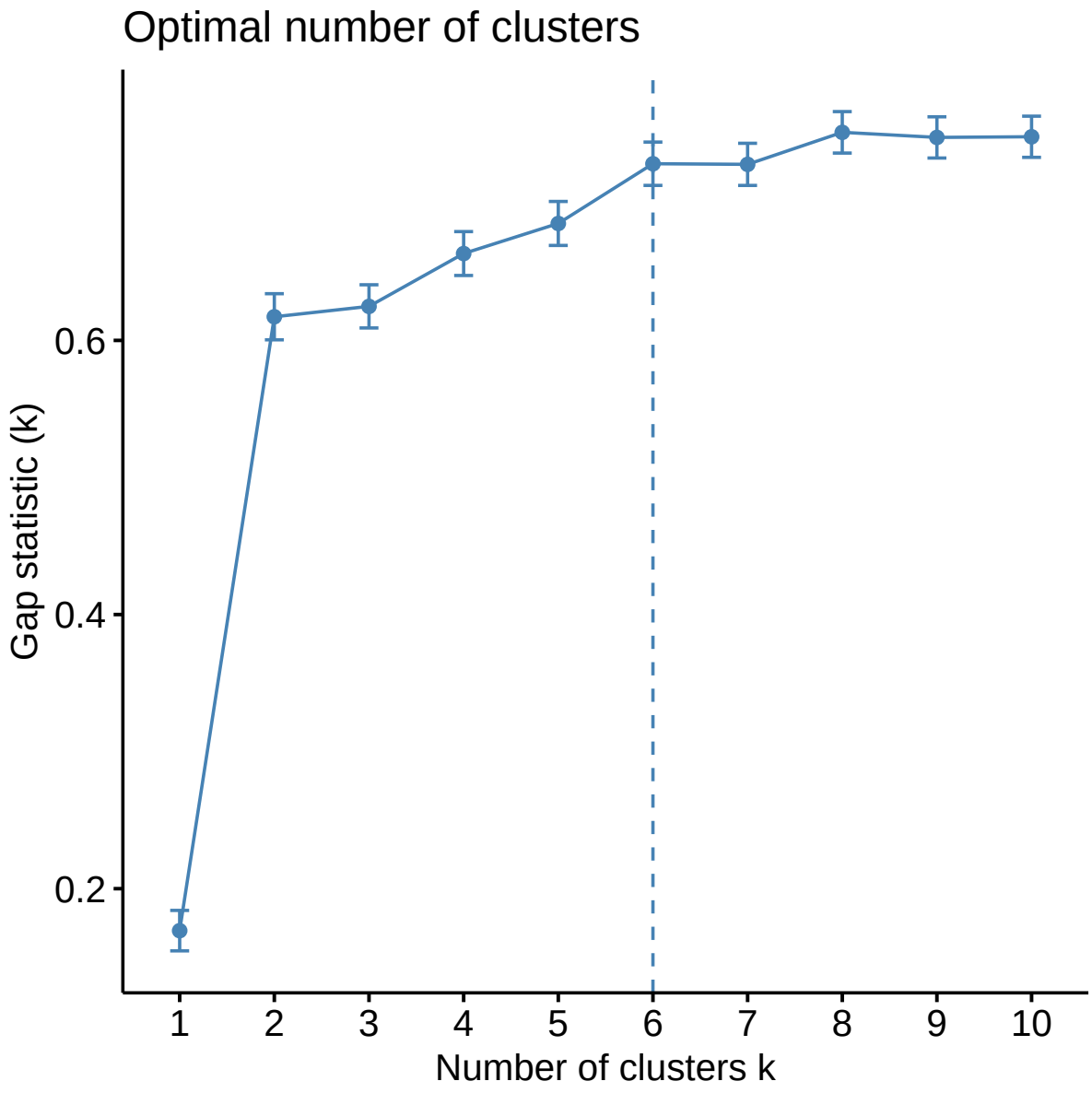

Figure 4.17: Gap statistic versus cluster number k**.** Error bars derived from 500 bootstraps. Here, the optimal number was $k = 6$ by the `factoextra::fviz_nbclust firstSEmax` criteria (https://rpkgs.datanovia.com/factoextra/index.html).

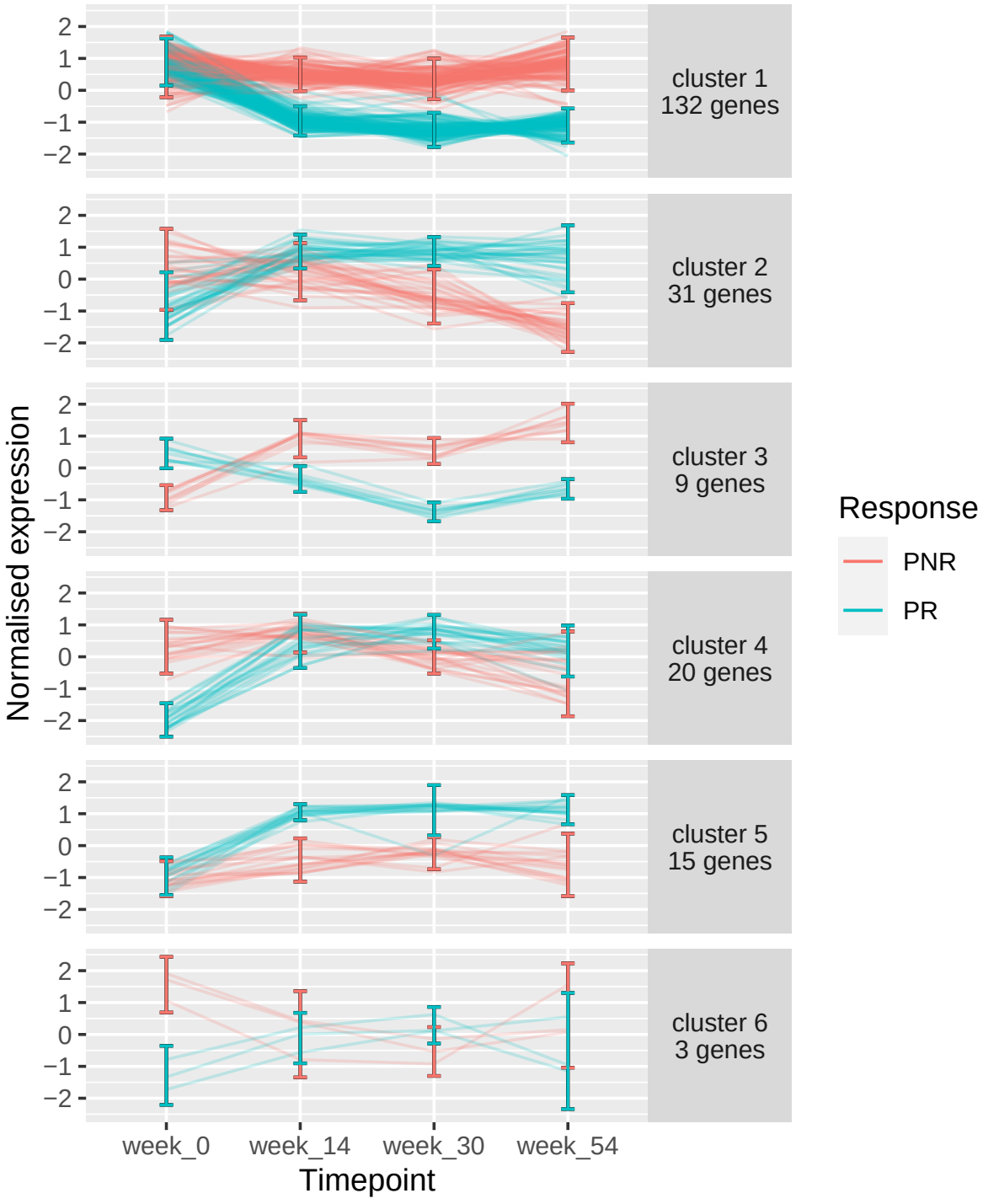

Figure 4.18: Normalised expression over the timepoints for genes in the six identified clusters. Log scale expression for each gene was standardised before clustering. Error bars show the mean ± standard deviation for the genes at each timepoint in primary responders (PR) and non-responders (PNR).

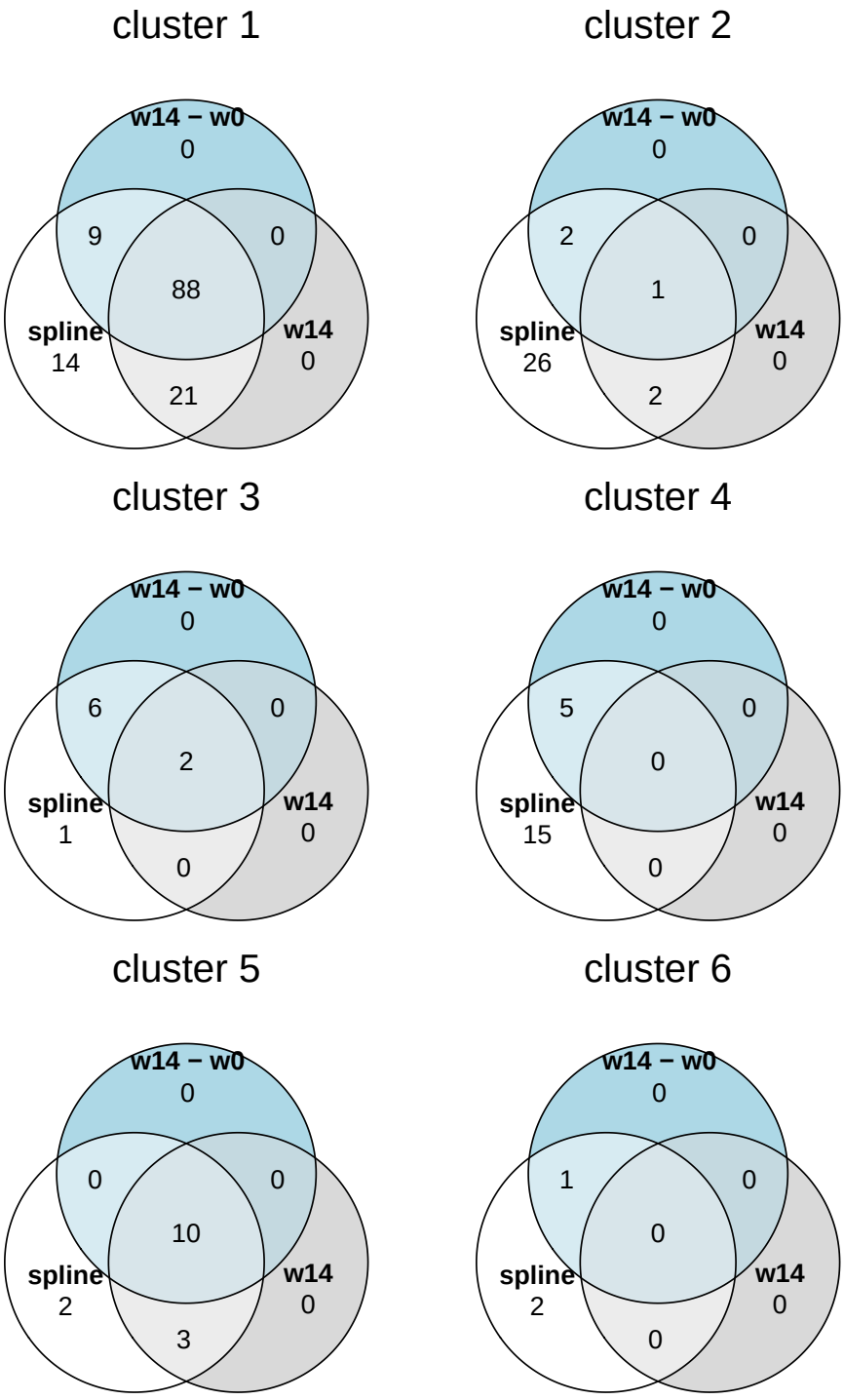

Figure 4.19: Overlap of genes differentially expressed between responders and non-responders from the spline model, the week 14 responder versus non-responder contrast (w14), and the interaction between week 0 to week 14 change and response status contrast (w14 − w0). Clusters 2 and 4 largely contained genes identified as significant only by the spline model.

multiple testing using `mashr` [281].

The majority 11 156/15 040 (74.2 %) of genes were eGenes (a gene with at least one significant *cis*-eQTL) in at least 1 timepoint (LFSR < 0.05). The variant with the lowest LFSR over all timepoints for each gene was chosen as the lead variant (eSNP) for that gene. Most eSNPs were significant at multiple timepoints: 999 at one timepoint, 381 at two timepoints, 526 at three timepoints, and 9250 at all four timepoints. I compared eSNP effect sizes between week 0 and each of weeks 14, 30, and 54 to identify reQTLs with a significant difference in effect versus baseline, as they may explain changes in expression from baseline. Most eSNPs were shared across timepoints; only six eSNP-eGene pairs were significant reQTLs (difference in betas BH FDR < 0.05): 1/6 between week 30 and week 0, and 5/6 between week 54 and week 0 (Fig. 4.20). Of the six eGenes with reQTLs between week 54 and week 0, *NMI* and *EPSTI1* both have magnified eQTL effect sizes at week 54 compared to week 0, and both are annotated to contain putative binding motifs for *IRF8* and *IRF2* (g:Profiler term IDs TF:M11685_1 and TF:M11665_1). However, direct interpretation of these reQTLs is complicated by changing cell composition in bulk expression data (discussed in Section 3.2.10 and Section 5.3).

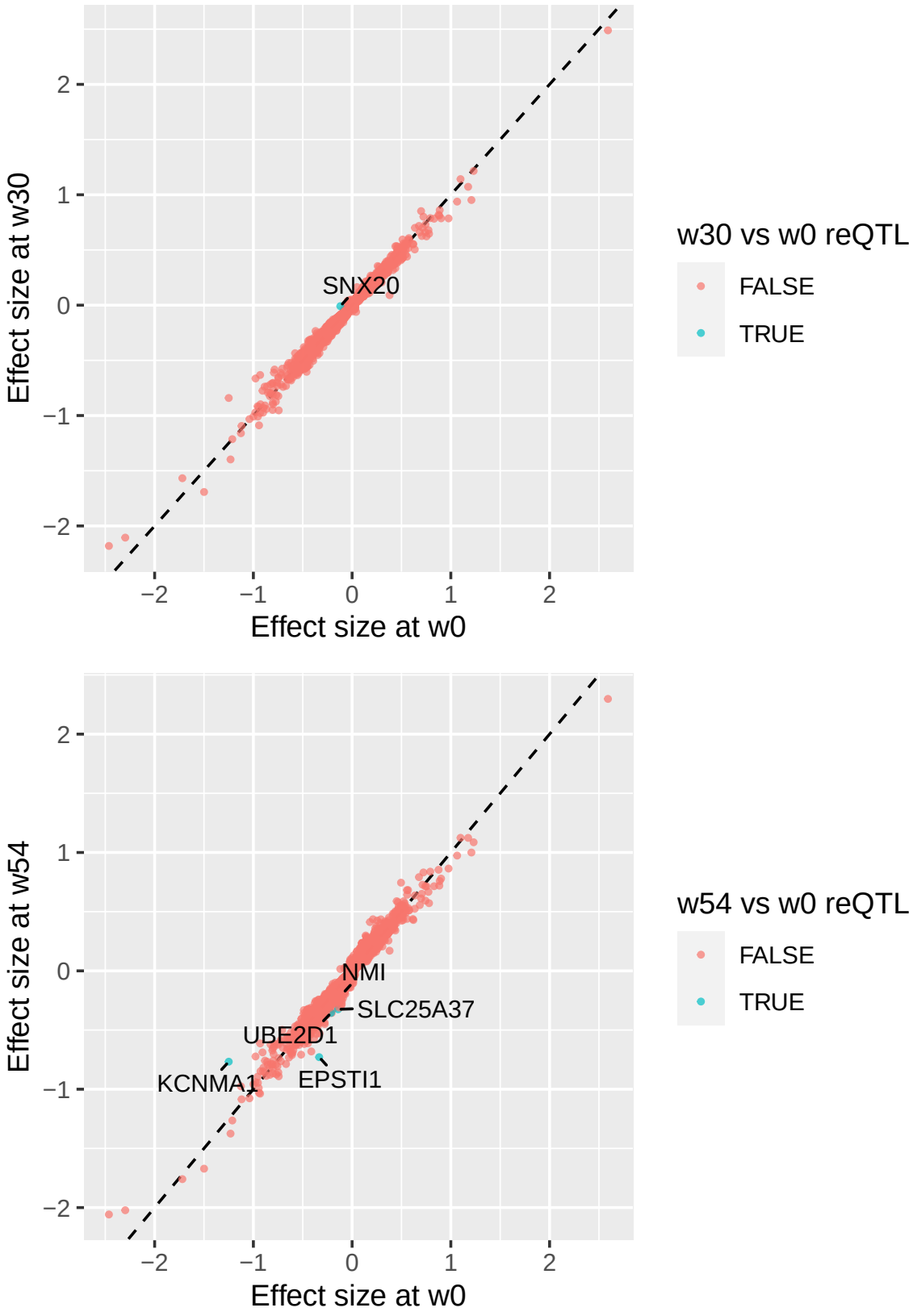

Figure 4.20: Post-treatment eQTL effect sizes at week 30 and week 54 compared to baseline effect sizes at week 0. Significant reQTLs at BH FDR < 0.05 are labelled.

4.4 Discussion

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

In contrast, future responders to adalimumab had lower baseline expression of plasma cell and cell division modules. The three hits from the gene-level adalimumab-only analysis implicate similar cell types; *IGKV1-9* encodes the immunoglobulin light chain variable region that forms part of antibodies produced by plasma cells, *KCNN3* is annotated to a plasma cell surface signature module (LI.S3 [240]), and the expression of both *KCNN3* and *PDIA5* are correlated with blood plasmablast frequencies [161]. Gaujoux *et al.* [366] observed lower baseline plasma cell abundances in infliximab responders than in non-responders, and hypothesised that plasma cell survival is supported by increased TNF levels in non-responders. Plasma cells also formed part of a correlated module of cell populations identified by Martin *et al.* [365], where lower module expression was associated with better response to anti-TNFs in a cohort with both infliximab and adalimumab patients. However, both these studies were conducted in gut biopsy samples, and there was no mention of strong between-drug heterogeneity.

The adalimumab-specific associations were more significant after cell proportion adjustment, which may indicate per-cell downregulation rather than cell abundance being associated with response. However, cell composition differences mediated by rarer cell types that have abundances poorly captured by the six major types used in the model will be poorly adjusted for. For example, plasma cell proportions are only weakly correlated with other cell types in the healthy immune system [388], although the relationship may differ in CD patients. It has also been shown that DGE analyses with correction for only common blood cell types can identify associations that are proxies for rare cell types [389]. If this is the case, the role of the cell composition estimates for adalimumab-specific effects may be more akin to precision variables, which would be consistent with increased significance after adjusting.

The between-drug heterogeneity for baseline associations is puzzling, especially the greater effect of cell composition adjustment for the infliximab-only model. Baseline patient differences between drugs could offer a partial explanation. There may be characteristics not listed in Table 4.1 that differ between patients on different drugs [355]. In the full PANTS cohort, lower albumin, higher CRP, and higher faecal calprotectin levels in infliximab patients suggest that they may have had greater disease severity. Differences may be driven by patient or physician preference, for example, patients with more severe disease are often given infliximab rather than adalimumab*. I have not yet been able to access clinical variables such as CRP and faecal calprotectin levels, variables that one could consider adjusting for in the DGE models. A richer

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

phenotype dataset containing clinical variables has been requested from collaborators.

The strongest single-gene association in the pooled analysis was *SIGLEC10*, which had reduced significance and a comparable effect size post-adjustment, where baseline expression was approximately 25 % higher in responders. Direction of effect was consistent between drugs, although the association was most significant in infliximab without cell composition adjustment. In IBD, small molecules called damage-associated molecular patterns (DAMPs) are released due to tissue damage and cell death, and further promote inflammation through pathogen-sensing pattern recognition receptor (PRR) pathways that include toll-like receptor (TLR) family receptors [339, 390]. For instance, faecal calprotectin, a marker for IBD activity, is a complex of two DAMPs, S100A8 and S100A9 [339]. SIGLEC10 has been shown to repress DAMP-mediated inflammation through binding CD24 [390]. SIGLEC10 is expressed on B cells, monocytes, and eosinophils [391]. Of these cell types, module level results posit monocytes as the most likely candidate cell type to be increased in anti-TNF responders. In monocytes, *SIGLEC10* gene expression is more specific to the $CD16^{+}$ monocytes [392], and in particular the $CD14^{+}CD16^{++}$ non-classical monocytes rather than the classical $CD14^{++}CD16^{-}$ or intermediate $CD14^{++}CD16^{+}$ subsets [393]. In PANTS, it was suggested by Kennedy *et al.* [355] that higher inflammatory load, as indicated by low baseline albumin levels, may result in low week 14 drug levels due to faster drug clearance. Low drug levels at week 14 were in turn associated with non-response. A hypothetical model might be high baseline *SIGLEC10* expression reflecting higher proportions of $CD16^{+}$ monocytes (or lower proportions of $CD16^{-}$ monocytes), decreased DAMP-mediated inflammation and increasing chance of primary response, possibly by affecting drug clearance rate. This is an extremely tentative model: both the cell proportion estimates and module definitions used thus far only represent monocytes as a whole, lacking the resolution to properly explore shifts in monocyte subsets. It may be possible to use expression of monocyte subset marker genes, such as those identified by Villani *et al.* [393], to improve the resolution of the cell proportion estimates.

Despite the strong heterogeneity in effects between drugs, consistent module-level effects that emerged after adjusting for cell composition included baseline upregulation of MHC-TLR7-TLR8, antigen presentation, and interferon modules in responders. As mentioned above, TLR receptors are involved in pathogen sensing, and TLR7 and TLR8 are endosomal proteins primarily expressed in monocytes, macrophages, and dendritic cells (DCs), part of an antigen presentation pathway that senses bacterial DNA and activates downstream innate immune pathways including type I interferon response [318]. Type I interferons have pathogenic or protective roles in many IMIDs [394]. It has been suggested that type I interferon responses induced via TLR7 and TLR8 can suppress colitis in mouse models, and play a role in maintaining gut homeostasis [395, 396], so upregulation here may again represent a less severe baseline disease in future responders.

Most previously reported baseline markers in blood and gut biopsies were non-significant in this study. For gut markers, this may not be unexpected. Although a subset of gut infiltrating immune cells and their precursors may also be circulating, genes specific to epithelium and immune cell types that differentiate after they migrate into tissues (e.g. monocyte-derived macrophages) are difficult to observe in blood. For blood markers, I sought to clarify the conflicting results in the literature about the association of *TREM1* expression in blood with anti-TNF response

[366, 367]. *TREM1* is expressed in myeloid lineage cells such as monocytes and macrophages; Villani *et al.* [393] reported that *TREM1* expression is most specific to classical monocytes and a newly identified subtype within the intermediate monocytes ("Mono3"). I did not find *TREM1* to be significantly differentially expressed in PANTS. The direction of effect corresponded to increased expression in responders, matching the Gaujoux *et al.* [366] direction of effect in blood. The strongest effect was observed in the infliximab-specific comparison without cell proportion adjustment, which may be another indication that baseline monocyte cell proportions are associated with response. There are many factors that could explain failures to replicate reported markers or identification of different markers from study to study. Many existing studies pool cohorts with different anti-TNF drugs due to the scarcity of large datasets, yet even within the PANTS cohort, there appears to be heterogeneity between drugs. There are between-study differences in the definition of primary response, such as endoscopic healing [366] versus scoring on clinical parameters [367]. Any two studies are unlikely to have adjusted for the same combinations of covariates in modelling, and some covariates like cell composition are very influential for bulk expression data. Finally, small sample sizes have considerable sampling error. Set-based associations that draw on changes in multiple genes, such as the module associations from this study, may be more reproducible compared to single-gene markers. A future aim will be to see if the identified baseline module associations also imply that response status can be predicted from baseline expression. Because modules associated with response appear to be mediated by cell proportions, much of the predictive ability may also lie in differences in cell proportions between responders and non-responders. Indeed, Gaujoux *et al.* [366] noted that adjusting expression for cell composition resulted in gut gene signatures that were worse at discriminating responders from non-responders. Testing the abundances of specific subpopulations for association with response (e.g. $CD16^+$ monocytes or plasma cells) can also be viewed as a type of set-based test that represents a set of cell-type specific genes, and thus may also be more reproducible than single-gene markers.

A much larger proportion of the transcriptome was associated with response after the induction period at week 14. Module associations showed downregulation of immune activation, TLR, inflammatory, monocyte, and neutrophil modules in responders; and upregulation of B and T cell modules. Similar module associations were also found when considering modules differentially expressed from week 0 to week 14. The differences between responders and non-responders at week 14 were qualitatively similar to the differences between week 14 and week 0 in responders, suggesting there may be relatively little change in the transcriptome of non-responders after anti-TNF induction. Associations were generally consistent between drugs for both the within week 14 and change from week 0 to week 14 analyses, perhaps because any baseline transcriptomic differences between patients taking different drugs were diluted by the large transcriptomic perturbation caused by taking an anti-TNF drug. Many of the same modules were also significant regardless of cell proportion correction. A general reduction in immune activation in responders at week 14 is presumably due to successful inhibition of TNF-mediated inflammation by the anti-TNF drug. Decreased inflammation correlates with reduced neutrophil activation and reduced monocyte recruitment [397], supported by the observed downregulation of neutrophil and monocyte modules. Apoptosis of monocytes induced by anti-TNF in CD patients

has also been previously described [398]. Certain B cell subsets are reduced in the blood of IBD patients compared to controls [399], so upregulation of B cell modules at week 14 may represent a shift towards health. Another potential explanation would be increased immunogenicity due to higher drug levels in responders [355], although lack of between-drug heterogeneity for the B cell signal is not consistent with the greater immunogenicity of infliximab. Overall, it is difficult to glean exact mechanisms from an observational study design, with bulk expression data, and using such broad module definitions.

Some previously identified baseline gut markers of response that were not differentially expressed in blood at week 0, were differentially expressed at week 14. *S100A8* and *S100A9*, identified as markers by Arijs *et al.* [363], which encode components of the inflammatory marker CRP, were downregulated in week 14 responders. The cytokine *OSM*, which promotes inflammation in gut stromal cells [364], was similarly downregulated. Although it is pointless to use a week 14 marker to predict a response that is defined at week 14, this does demonstrate that gut markers can coincide with blood markers if expressed in immune cells present in both tissues.

When considering the interaction between change from week 0 to week 14 and response, the general pattern is magnification in responders, where the same expression changes occur with greater magnitude than in non-responders. A potential hypothesis is a continuum of response from non-response to response. Gaujoux *et al.* [366] found changes in cell proportions in response to anti-TNF treatment were magnified in responders, also supporting response as continuous phenotype. This study demonstrates a similar trend at the transcriptional level. There were some rare exceptions to magnification for genes and modules in the type I interferon pathway. These showed upregulation in non-responders from week 0 to week 14, yet were either downregulated or not significantly different in responders. Single-gene examples include the interferon-induced guanylate-binding proteins *GBP2* and *GBP5* [400], and *STAT2*, a key transcription factor for interferon-stimulated genes [244]. Genes such as *IFIT3* and *STAT2* are more strongly induced by type I interferons compared to type II [401]. A study of RA, an IMID also treated with anti-TNF drugs, likewise found increases in type I interferon-regulated gene expression in blood after infliximab treatment to be associated with poor clinical response [402].

A spline model of expression over all four timepoints confirmed the above observations made in week 0 and week 14 samples. Two main clusters of genes (clusters 1 and 5) contained mostly genes significantly associated with response in the two pairwise comparisons: within week 14, and change from week 0 to week 14. An example is the most significant single-gene association from cluster 1 in spline model, *KREMEN1*, which is also one of most significant associations in the pairwise comparisons. *KREMEN1* is part of an inflammatory apoptotic pathway in gut epithelium [403], and is downregulated in responders post-induction. The trajectories of expression for genes in clusters 1 and 5 confirmed that changes in expression post-induction were generally greater for responders, and in addition demonstrates that post-induction expression differences between responders and non-responders are sustained in samples taken around week 30 and week 54 during the anti-TNF maintenance period. In PANTS, Kennedy *et al.* [355] found that "continuing standard dosing regimens after primary non-response was rarely helpful" for inducing remission by week 54. This phenomenon may have a transcriptomic basis, although non-responders in the PANTS RNA-seq data were selected to exclude patients in remission by week 54, so trajectories

for non-responders at week 14 that eventually achieved remission could not be observed.

Making use of data from later timepoints allowed more subtle effects to be detected in the spline analysis. Clusters 2 and 4 were enriched for B cell genes that were not significantly different at the gene-level in the within week 0 comparison, although some downregulation of B cell and plasma cell modules was detected. Cluster 3 reproduced the observation that interferon-induced genes have opposing trajectories of expression in responders and non-responders. Expression of these genes was higher in responders at week 0 and lower at all post-treatment timepoints. The cluster contains genes such as *STAT1*, *IRF1*, and *TAP1* that are induced by both type I and type II interferons [401]. I propose that blood expression of interferon-related genes is an attractive target for future studies of the biological basis of anti-TNF response, and for use in building predictive models of primary response status. Since the difference is maintained until week 54, by which time patients would have received many doses of drug, it is more likely that response is due to some biological property of an individual patient. Studies of anti-TNF response in RA patients have also found high baseline interferon activity in blood to be associated with good clinical response [404, 405]. It should be noted that the number of clusters is only the optimal number determined in this dataset, and does not imply that genes in different clusters represent biologically distinct pathways. Clusters 2 and 4 have similar trajectories and enrichments for B cell genes, and interferon pathway genes appear in both clusters 1 and 3.

Finally, I attempted to determine if there were changes in genetic architecture of expression over time, which could indicate that expression response to anti-TNF has a genetic component. Out of all significant lead eQTLs for 11 156 genes, only six reQTLs were detected between baseline and any one of the three post-treatment timepoints. Although no enrichment analyses were attempted due to the small number of associations, *NMI* and *EPSTI1* are both interferon-induced genes with significant reQTLs that had their strongest effect size on expression at week 54. Given the issues with doing a reQTL analysis in bulk expression data are similar to those encountered in Chapter 3, I did not place emphasis on interpreting these small numbers of associations. I would also like to verify that these significant reQTLs are not artifacts from shrinkage of effect sizes in the joint eQTL model, as their posterior effect sizes from `mashr` were very different from the input effect sizes from the per-timepoint models. If these hits are indeed reproducible by complementary methods such as allele-specific expression (ASE) [406], it may then be worth introducing genotype-response interaction terms into the eQTL models to identify eQTLs with differing effects in responders and non-responders. Given there is prior interest in the interferon pathway from DGE analyses, a more statistically powerful approach may be to generate a continuous interferon pathway score for each sample, which would then act as the interacting variable, similar to the approach of Davenport *et al.* [96].

Several threats to the validity of the study remain to be discussed. The most pressing may be the meaning of time in the study. For pairwise DGE comparisons, expression trajectory clustering, and reQTL mapping, samples were divided into four discrete timepoints that corresponded to the major visits in PANTS; whereas the DGE spline model was fit to study day directly. Study day has substantial variation around the target for later timepoints. The particular target timings for post-baseline visits (weeks 14, 30, 54) were chosen so that patients on infliximab (8 weeks between doses) and adalimumab (2 weeks between doses) could both be sampled with the same

visit structure. Drug levels peak sharply after each dose and decline exponentially over time. To capture trough drug levels, visits were scheduled to be as close as possible to the next scheduled drug dose (within a week) Neither modelling approach is perfect; matching patients by timepoint and study day are merely attempts to gather samples matched by trough drug level.

A further complication is the inclusion of LOR samples in analyses. One treatment option after LOR is dose escalation, which may raise trough drug levels for all subsequent visits for those patients. However, since the PANTS protocol allows for LOR visits that coincide with major visits to be labelled as a major visit, there is no guarantee that simply excluding samples labelled as LOR would resolve this. The best solution may to explicitly model measured serum drug levels as a covariate, where like cell proportions, it would likely act as a mediator of some associations with response. I did not do this as data missingness would reduce the sample size by about 40 % in this study. Finding a suitable normalisation of drug level for use in pooled drug analyses would also be challenging. Infliximab and adalimumab have differing pharmacokinetics; infliximab has higher peak concentrations, higher peak-trough ratios, and shorter half-life. The same serum concentrations of infliximab and adalimumab also have different biological effects due to differing therapeutic windows [350, 407, 408].

The effect of differential drop out in responders and non-responders has not been explored. There are three main mechanisms of missing data: missing completely at random (MCAR), where probability of data being missing is independent of both observed and missing data; missing at random (MAR), where probability of data being missing conditional on observed data is independent of missing data; and missing not at random (MNAR), where probability of data being missing depends on missing data [409]. Even conditional on response status, it is more likely that expression data from more extreme non-responders is missing for later timepoints, so the likely mechanism here is MNAR, thus the linear mixed models used in this study may be biased. If it is indeed the most extreme non-responders dropping out, the estimation of responder versus non-responder effects may be conservative in the spline analysis. Note there is no sidestepping a MNAR mechanism by analysing only the complete cases, since they will differ systematically from the sample as a whole [410].

In conclusion, it remains unclear whether there are any robust single-gene expression markers for anti-TNF response in the whole blood of CD patients at baseline. Baseline module associations were observed, but there was unexpected heterogeneity between infliximab and adalimumab patients, so it remains to be seen if such associations will be replicated in other cohorts. Large upcoming datasets with drug response phenotypes such as the 1000IBD project [411] will be invaluable for attempted replication of the associations found in PANTS. Expression differences between responders and non-responders were more distinct at timepoints after the induction period. I found type I interferon genes were more upregulated post-treatment in non-responders, going against the general trend of magnified transcriptomic change in responders, Given that type I interferon expression in blood has also been associated with anti-TNF response in RA patients, there may be an opportunity to consider the shared biology of anti-TNF response in IBD and RA. Much work has been done generating and validating signatures for anti-TNF response in RA [412]; but not much work on validating RA signatures in IBD cohorts and vice versa.

This chapter has been purely descriptive. Although there are expression differences at many genes between responders and non-responders, I do not know which cause non-response, and which are a consequence of disease reduction in responders. I have deliberately avoided the term "signature" in describing my own results, as I have not yet had a chance to assess the predictive capability of associated gene modules. I also did not find evidence for many strong and interpretable reQTL effects over time in whole blood, and therefore was unable to form hypotheses on the genetic mechanisms influencing anti-TNF response via expression. However, the presence of eQTLs for most genes and the presence of strong differences in expression post-induction may allow testing for causal mechanisms where genotype affects drug response via expression. Strategies for moving on to both prediction and causal inference will be discussed in Chapter 5.