Chapter 3

Genetic architecture of transcriptomic response to Pandemrix vaccine

The work presented in this chapter is a collaboration between the Wellcome Sanger Institute, King's College London, the Francis Crick Institute, and the Biomedical Research Centre at Guy's and St Thomas' Hospital and King's College London. I would like to reiterate my thanks to the people and organisations mentioned at the beginning of Chapter 2.

3.1 Introduction

3.1.1 Host genetic factors affecting influenza vaccine response

Many human traits are heritable and complex—response to vaccination is no exception. Twin studies have demonstrated approximately 30–90 % heritability of antibody responses to many vaccines, including smallpox, hepatitis A and B, anthrax, pneumococcal, *Haemophilus influenzae* type b (Hib), diphtheria-tetanus-pertussis (DTP), and bacillus Calmette–Guérin (BCG) [252–255]. Candidate gene studies and genome-wide association studies (GWASs) have identified multiple genetic associations with antibody response [252, 253, 256, 257], including replicated associations for hepatitis B vaccine in a haplotype block in the human leukocyte antigen (HLA) region encompassing *HLA-DR* and *BTNL2*, and for measles vaccine in an intron of a receptor known to interact with measles virus, *CD46*.

In contrast, Brodin *et al.* [255] found anti-haemagglutinin (HA) antibody responses to seasonal influenza vaccine in 105 adult twin pairs (median age 44 yr) had no detectable heritability, alongside a general decrease in heritability of most immune parameters with age. They posited that the genetic contribution to response was overshadowed by environmental factors such as previous influenza vaccination or infection in adults, whereas the estimated heritability of the aforementioned vaccines was substantial because they are vaccines against non-circulating pathogens, or are childhood vaccines for which heritability was assessed in young children with shorter immune histories.

Nevertheless, a small number of candidate gene studies have identified genetic variants

associated with antibody response to influenza vaccines [257]. Gelder *et al.* [258] ($n = 73$) identified associations between HLA alleles in *HLA-DRB1* and *HLA-DQB1* with haemagglutination inhibition (HAI) seroconversion after trivalent inactivated influenza vaccine (TIV); Moss *et al.* [259] ($n = 185$) also found associations between HLA class II alleles (HLA-DRB1*04:01 and HLA-DPB1*04:01) and HAI seroconversion after seasonal influenza vaccination. Poland *et al.* [260] ($n = 184$) tested HLA alleles, and single nucleotide polymorphisms (SNPs) in coding and regulatory regions of cytokine or cytokine receptor genes, for association with post-TIV HAI titres specific to H1 and H3 subtypes (two of the components of the trivalent vaccine). They reported nominally significant associations for two *HLA-A* alleles with H1-specific titres, six SNP associations with H1-specific titres and ten SNP associations with H3-specific titres. Egli *et al.* [261] ($n = 196$) identified a SNP upstream of *IFNL3* (rs8099917) to be associated with seroconversion post-TIV, and also found the SNP to be an expression quantitative trait locus (eQTL) for *IFNL3* expression in H1N1-stimulated peripheral blood mononuclear cells (PBMCs) in a second cohort ($n = 49$). Lastly, Avnir *et al.* [262] focused on a coding variant (rs55891010) in the part of *IGHV1-69* that encodes the complementarity-determining region (CDR) of broadly neutralising antibodies that bind influenza HA. One month after H5N1 avian influenza vaccination ($n = 85$), associations were detected with usage of *IGHV1-69* in the antibody repertoire, and serum antibody binding efficiency to H5N1 HA. The associations listed above have all been found in small cohorts and have not been validated by subsequent studies, so it remains unknown whether robust genetic associations with antibody response to influenza vaccines exist.

3.1.2 reQTLs induced by influenza vaccination

Host genetic variation could play a causal role in influenza vaccine response by altering the expression of genes as eQTLs. As described in Section 1.2.3 and Section 1.2.4, the effect sizes of eQTLs can be highly context-dependent, and many eQTLs in the immune system are response expression quantitative trait loci (reQTLs) only detectable after stimulation, not at baseline. One can map reQTLs considering vaccination as an *in vivo* immune stimulation. This usually involves measuring the transcriptome of immune cells before and after vaccination in genotyped individuals, then testing for genotype-dependent changes in expression. As expression is a key molecular intermediate between genotype and phenotype, a genotype-dependent change in expression after vaccination may be a mechanism mediating genotype-dependent antibody responses.

As reviewed in Section 1.2.4, few *in vivo* reQTL studies have been conducted, and even fewer studies have been conducted where the *in vivo* stimulation is vaccination, despite the potential for learning about genetic regulation of vaccine-induced expression responses. To my knowledge, there is only one such study: by Franco *et al.* [94] on response to seasonal inactivated TIV. Franco *et al.* [94] enrolled healthy Europeans adults into discovery ($n = 119$ males) and validation ($n = 128$ females) cohorts in two consecutive influenza seasons*. In each cohort, peripheral blood gene expression was measured by expression array on day 0 (baseline); and on days 1, 3, and 14 post-vaccination. Serum HAI and microneutralisation (MN) titres were measured against each of the three vaccine components at days 0, 14, and 28. The titre response index (TRI) [153] was computed from these titres as a single measure of antibody response adjusted for baseline titres.

*Sex-dependence of effects was not addressed.

Individuals were genotyped by genotyping array.

Cis-eQTL were mapped using a linear mixed model jointly over all four days, with day, genotype, day-genotype interaction, and a random intercept for individual as predictors; and gene expression the response variable. At 467 (non-independent) eQTL for 78 genes replicated in both cohorts, there was both a significant day effect (indicating the gene was differentially expressed post-vaccination) and a significant genotype effect (indicating the eQTL effect). To call reQTLs, eQTLs were also mapped separately for each day with a linear model including only genotype as a predictor, from which the model R^2 was computed as a rough measure of the variance in expression explained by the eQTL at each day. Franco *et al.* [94] then computed delta-R^2: the maximum absolute deviation of the three post-vaccination R^2s from the day 0 R^2. Out of the eQTLs that replicated in both cohorts, 146 eQTLs for 34 genes ranking above the 99th percentile of the delta-R^2 distribution were defined as reQTLs. The union of the 78 and 34 genes from the above analyses (98 genes with differential gene expression (DGE) and an eQTL; or a reQTL) was enriched for pathways and gene sets related to antigen processing and presentation, $CD8^+$ T cell-mediated apoptosis, dendritic cell (DC) maturation and function, and membrane trafficking. Lastly, integrating antibody titre data, they filtered down to 20 genes with expression correlated to TRI at any day, with an eQTL, and with either post-vaccination differential expression *or* a reQTL effect. Seven genes out of these 20 were involved in antigen transport, processing, or presentation in antigen-presenting cells (APCs): *NAPSA*, *C1orf85*, *GM2A*, *SNX29*, *FGD2*, *TAP2*, and *DYNLT1*.

Critically, Franco *et al.* [94] recognised that just assessing overlap of multiple filtering criteria cannot infer the direction of causal relationships between genetic variation, expression and TRI. They attempted a model comparison with the CIT [263] method to resolve the directionality of association between expression and TRI, finding suggestive evidence of causal effect on TRI mediated by expression at several eQTL. Unfortunately, they also evaluated that the power of the CIT was only ∼60 % at their total sample size of $n = 247$. Nevertheless, the study is proof of concept that integration of genotype, expression, and antibody response data in an *in vivo* reQTL framework can identify genes under genetic regulation likely to be involved in vaccine response.

3.1.3 Chapter summary

The Human Immune Response Dynamics (HIRD) cohort represents a unique opportunity for detecting host genetic contributions to influenza vaccine response. Similar to Franco *et al.* [94], expression, antibody response, and genotypes are all available for the same individuals. As Pandemrix is against a pandemic strain that had not been in seasonal circulation for decades at the time of cohort recruitment, responses will be less driven by individual immune history, so power to detect genetic associations is expected to be greater. In Chapter 2, I characterised differential expression induced by Pandemrix, as well as expression associations with antibody titres. In this chapter—given that HIRD is too small for a direct GWAS of antibody response—I focus on the genetic contribution to expression response. I apply the *in vivo* reQTL framework, aiming to characterise the association of common genetic variants with expression across multiple timepoints, and pinpoint genes important to Pandemrix response.

3.2 Methods

3.2.1 Overall strategy for reQTL mapping

A plethora of approaches to mapping eQTLs with linear models exist; each approach has its own advantages, disadvantages, and assumptions. When the task is also to define reQTL between multiple conditions, the diversity of possible approaches further multiplies. Here I will discuss aspects of the data and available methodologies that led to the final modelling strategy adopted in this chapter.

3.2.1.1 Adjusting for population structure using linear mixed models

Population structure occurs when the samples in a study are not independent, but structured due to genetic relatedness. Genetic association studies assume that the individuals in a sample are unrelated (or at least sufficiently distantly related) [264–266]. Relatedness, and thus population structure, occurs at different scales. Population stratification refers to systematic differences in allele frequencies and genetic background between human populations due to demographic history. This represents large-scale structure where individuals are related due to shared ancestry [188, 265]. At a smaller scale, sample individuals can be related due to being in the same family. The presence of more relatedness in a sample than is assumed is the problem of cryptic relatedness [264–266]; this can be at any scale, but more often the term refers to recent relatedness.

In the context of eQTL mapping (and genetic association studies in general), where the aim is to assess the effect of a single genetic variant on expression, there is potential for confounding. The issue (well-reviewed in [266, 267]) is that we fit a marginal model to estimate the effect of a single variant x_k on the phenotype y:

$$y = \mu + \beta_k x_k + \epsilon \tag{3.1}$$

where μ is the intercept, and $\epsilon \sim N(0, \sigma_e^2 I)$ is the error term that represents environmental and stochastic sources of variation. The variance-covariance matrix for error term is a scalar matrix, encoding the classic regression assumptions of homoscedasticity and uncorrelated errors. A more appropriate data generating model is:

$$y = \mu + \beta_k x_k + G + \epsilon \tag{3.2}$$

where $G = \sum_{i \neq k} \beta_i x_i$ represents the effect of the genetic background at all other variants in the genome. As many variants can be expected to affect a complex polygenic trait, G has some causal effect on y. Population structure means there can be a shared cause of G and x_k such as ancestry. This opens a backdoor path $x_k \leftarrow$ ancestry $\rightarrow G \rightarrow y$, confounding the relationship between x_k and y. In Eq. (3.1), when one estimates the coefficients, the effects of the omitted variable G will be attributed to x_k, resulting in spurious associations and genomic inflation of test statistics [188]. Here, G represents exactly the confounding due to genetic background, but there are other possible confounders, such as shared environmental factors that differ systematically between populations [268]. A popular approach to avoid confounding is to include genotype principal components (PCs) as fixed effects in the regression [185, 269], thus blocking the backdoor path

from x_k to y. Genotype PCs represent population stratification effects like ancestry, but also act to block confounding from genetic background and environmental effects by proxy [268].

Unfortunately, genotype PCs alone cannot account for smaller-scale population structure [188]. An approach that can explicitly model such population structure is the linear mixed model (LMM) [188, 267, 269], which expresses the idea that more genetically correlated individuals are expected to be more phenotypically correlated [268]. A typical model form is:

$$y = \mu + \beta_k x_k + u + \epsilon \tag{3.3}$$

where random effect $u \sim N(0, \sigma_g^2 K)$ has a variance-covariance matrix proportional to the genetic correlation between individuals: the kinship matrix K. This improves on Eq. (3.2) by recognising that the variants in G are correlated. σ_g^2 is often called the (genetic) variance component; the larger it is, the more phenotypic variance is explained by genetic background [267]. Although LMMs were originally developed in the context of animal breeding, where K is computed from a known pedigree, it can also be computed from genome-wide SNP data [266, 269]. Unlike pedigree-based kinships that range from zero (unrelated) to one (self or identical twin), SNP-based relatedness values represent average correlations of alleles between individuals [270], hence may be negative or greater than one [271]. This does not affect their usage in LMMs.

The HIRD genotype data has already been filtered such that no pair of individuals are first-degree relatives or closer (Section 2.2.4), but cryptic relatedness may remain. The multi-ethnicity of the cohort means there is large-scale population structure from ancestry (Fig. 2.5). Genotype PCs were computed to represent axes of variation due to ancestry. These were included as covariates in DGE analyses in Chapter 2 to improve efficiency by explaining some variation in expression. For eQTL mapping in this chapter, I use both a random effect in an LMM and PC fixed effects to correct for population structure. This may not be strictly necessary if the random effect can correct for large-scale structure, but does not seem to impact power or type I error rate [272], and may have some benefit at SNPs with very different allele distributions between populations (unusually differentiated) [188].

The performance of various software implementations for kinship estimation from genome-wide SNP data and LMMs are highly comparable; the specific choice of implementation can usually be made on the basis of computational efficiency [269]. In this chapter, I use LMMs implemented in LIMIX [273] with kinship matrices estimated by LDAK [274].

3.2.1.2 Multi-condition models

Since the aim of this chapter is to identify genetic variation that affects expression response to vaccination, it may seem most direct to model the change in each individual's expression after vaccination as the response variable. This approach has been applied for identification of condition-specific eQTL, typically with the response taking units of log fold change between conditions (e.g. [275–277]). Although potentially powerful if eQTL effects are small and opposite between conditions [276], it is analogous to the "change score" approach, which can suffer from regression to the mean, and increased uncertainty from the variance sum law if effects between conditions have positive covariance [171, 278]. Instead, I map eQTLs within each of three

timepoint conditions (day -7/0 pre-vaccination baseline, day 1, and day 7), and find reQTLs by looking for eQTLs that have different effects between conditions. As in Section 2.2.9.2, day -7 and day 0 array measurements were treated as repeated measures of the baseline timepoint. Unlike a test for difference implemented using a genotype-condition interaction term in a joint regression model, homoscedasticity of errors is not assumed for all conditions [279].

Within each timepoint, the HIRD dataset includes expression measured by both array and RNA sequencing (RNA-seq). As discussed in Section 2.2.9.3, it is difficult to directly estimate the between-study heterogeneity when the number of studies is small, thus Bayesian meta-analysis was preferred for combining array and RNA-seq DGE estimates. That method does not scale to eQTL analysis, where the number of tests is large, in the order of thousands of tests per gene, versus the handful DGE contrasts per gene performed in Chapter 2. Instead, I perform a mega-analysis within each timepoint, first merging array and RNA-seq expression estimates into a single matrix with ComBat [211]. For comparison purposes, eQTL analyses were also run in the array and RNA-seq samples separately.

Defining whether an eQTL is shared between conditions can be a tricky business. Naively, after mapping eQTLs separately in each condition, one can assess the overlap of significant associations between conditions. This underestimates sharing due to the difficulty of distinguishing true lack of sharing from missed discoveries as a consequence of incomplete power within each condition [68, 280]. Condition-by-condition analysis also cannot borrow information across conditions for mapping shared associations [280–282]. Counterintuitively, a joint multivariate analysis may be more powerful even when associations are not shared across all conditions [283].

A variety of models have been employed for joint eQTL mapping, including the use of classical multivariate methods such as multivariate analysis of variance (MANOVA) [80], frequentist meta-analyses (e.g. Meta-Tissue [284], METASOFT [285]), and Bayesian models (e.g. eQtlBma [280], MT-HESS [286], MT-eQTL [287]). Joint mapping has repeatedly been demonstrated to be more powerful than condition-by-condition analysis, and recent joint methods are now computationally efficient when scaling to large numbers of conditions and variants tested (e.g. RECOV [288], `mashr` [281], HT-eQTL [282]). In this chapter, I apply `mashr` [281] for the joint estimation of eQTL effects across my three timepoints. The method learns patterns of correlation among multiple conditions empirically from condition-by-condition summary statistics, then applies shrinkage to provide improved posterior effect size estimates, and computes measures of significance per condition.

3.2.1.3 Additional expression preprocessing

There are a number of transformations often applied to expression data before eQTL mapping, such as the rank-based inverse normal transformation (INT) (e.g. GTEx v8 [54]). INTs work by matching sample quantiles to quantiles of the standard normal distribution, which conforms often non-normal expression data to an approximately normal distribution, improving computation speed in large samples [289] and reducing the impact of expression outliers. In the context of genetic association studies, the practice of applying rank-based INT to phenotypes has been criticised for only guaranteeing approximate normality of residuals when effect sizes are small, and potentially inflating the type I error in linear models that include interaction terms [290].

More recent simulations suggest that INTs have good power and type I error control across commonly-encountered distributions of non-normal residuals [289]. Another common transform is standardising (centering and scaling to zero mean and unit variance e.g. eQTLGen Consortium [58]), often done so that effects across genes and studies can be comparably interpreted in units of standard deviation expression [291]. In multi-condition datasets, data transformations are typically applied within conditions (e.g. within each tissue individually in GTEx v8 [54]).

Simulations were performed to evaluate the effect of the aforementioned transformations on reQTL detection between a hypothetical day 0 baseline and day 1 post-vaccination condition. The size of a reQTL effect depends on the scale of the expression data; here I define the size of a reQTL as the difference in eQTL slopes (betas) for the same variant-gene pair between conditions with expression measured on the $\log_2$ scale. The boxed facets in Fig. 3.1 represent undesirable effects of transformations on reQTL effect sizes. Rank-based INT induces false shared eQTL effects between conditions in scenarios 4 and 5 (e.g. row "Rank-based INT" in Fig. 3.1). In general, transformations that scale within condition are not appropriate, as different variances within conditions contribute to the reQTL effect (e.g. "Scale within day"). Scaling without separating conditions is also problematic, since the total variance also affects the reQTL effect size. For example, in "Scale", scenarios 2 and 4 have the same 1 unit increase in beta pre-transformation (the same fold-change between conditions), but after scaling-only the beta increases are $0.75 - 0 = 0.75$ and $0.8 - 0.4 = 0.4$ respectively—scenario 4 now looks like a reQTL of weaker effect.

In light of these simulations, I decided that neither rank-based INT nor standardisation were appropriate. Only the centering-only transformations (e.g. "Center") avoided both false shared effects and preserved relative reQTL effect sizes. The simple inclusion of an intercept term in the eQTL model already achieves this, so no additional expression transformations were applied before eQTL mapping. Not performing a rank-based transform does lose the advantage of reining in outliers, but the expression data have already been preprocessed to remove low-expression outliers in Section 2.2.7. Many other preprocessing steps done prior to this stage in the pipeline (e.g. variance-stabilisation, ComBat batch effect correction) are also expression transformations, but I only consider the preservation of reQTL effects defined from expression values post-adjustment for technical effects to be important, so I did not consider those steps in my simulations.

3.2.2 Genotype phasing and imputation

Genotyping and pre-imputation processing are described in Section 2.2.3 and Section 2.2.4. Prior to imputation, 213 277 monomorphic variants that provide no information for imputation were removed. Variant alleles were aligned such that the reference allele matched the GRCh37 reference, and 358 indels were removed, leaving only SNPs. Imputation for the autosomes and X chromosome was conducted using the Sanger Imputation Service*, which involved pre-phasing (separate estimation of haplotypes before imputation to improve imputation speed) with EAGLE2 (v2.4) [292] and imputation with PBWT (v3.1) [293] against the Haplotype Reference Consortium (r1.1) panel [294]. Imputed SNPs were lifted-over from GRCh37 to GRCh38 coordinates using CrossMap [295]. Poorly-imputed SNPs with imputation information score INFO < 0.4 were

*https://www.sanger.ac.uk/tool/sanger-imputation-service/

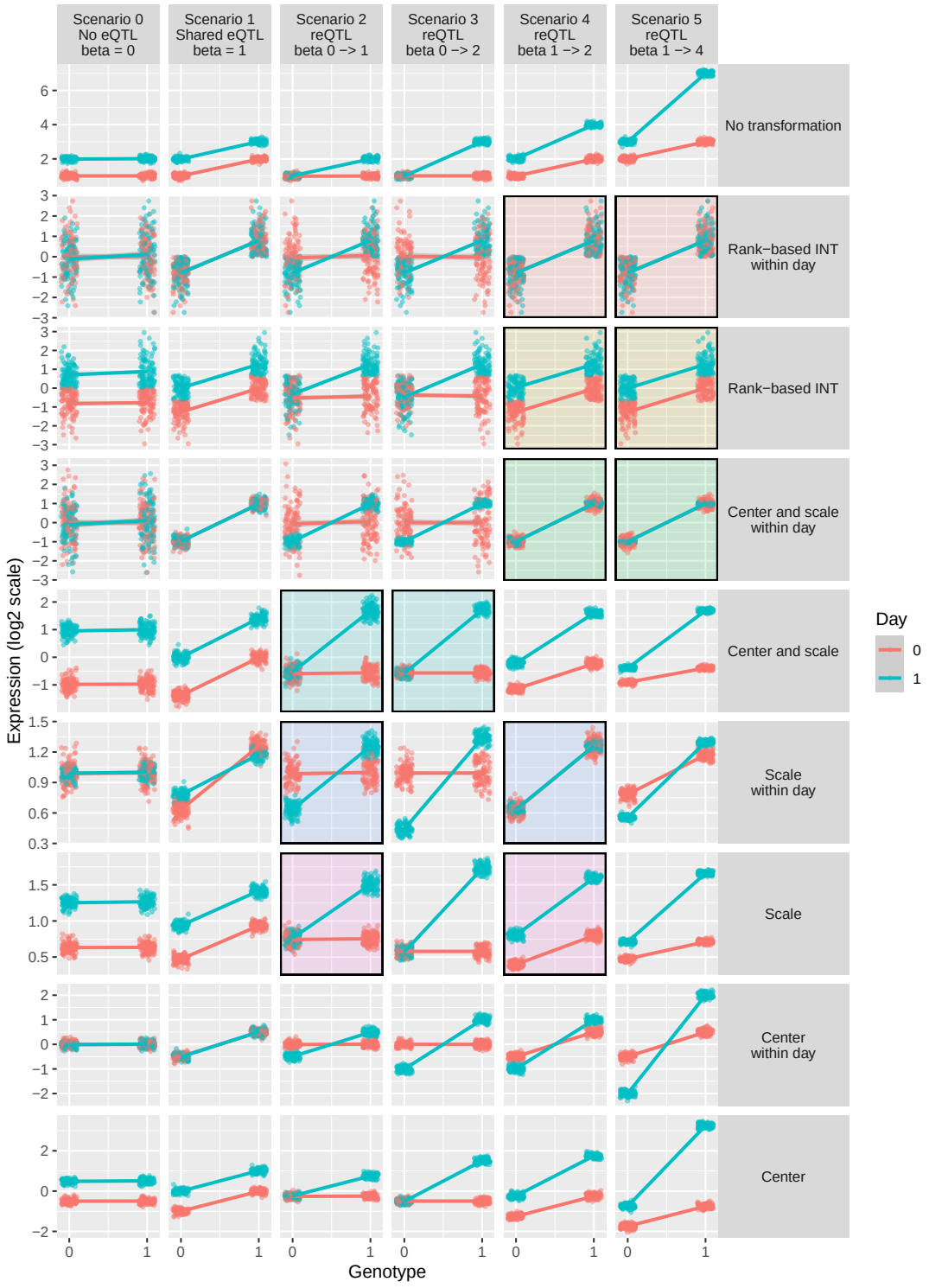


Figure 3.1: Simulating the effect of data transformation on reQTL effects. Expression values on the log scale were simulated for 200 individuals (100 with each genotype) at a day 0 baseline and day 1 post-vaccination timepoint. Gene expression is upregulated at day 1 by $\log_2 \text{FC} = 1$. Six scenarios were simulated with different gene-variant pairs (columns) corresponding to different eQTL and/or reQTL effects between day 1 and day 0; the size of the reQTL effect (difference in beta between day 1 and day 0) was set to 0, 0, 1, 2, 1, and 3 for the six scenarios. Gaussian noise with mean = 0 and standard deviation = 0.1 was added. The top row is the ground truth. In following rows, a different transformation was applied within each row: a rank-based INT (Blom offset for fractional ranks [289, 290]), standardising (centering and scaling), centering only, or scaling only. Highlighted pairs of scenario-transform combinations on each row represent false positives or negatives where the size of the relative reQTL effects are no longer correct.

removed, leaving 40 290 981 SNPs measured for the 169 genotyped individuals.

3.2.3 Estimation of kinship matrices

When testing a variant for association using LMMs, the kinship matrix used should not include that variant to avoid loss of power from "proximal contamination" [296]. A simple way to avoid this is to compute a leave-one-chromosome-out (LOCO) kinship matrix using all variants except the ones on the tested variant's chromosome [297]. I estimated kinship in the HIRD data from common autosomal variants, using LDAK (v5.0) [274], which computes SNP-based kinship matrices weighting SNPs by linkage disequilibrium (LD) and accounting for genotype uncertainty. Filtered pre-imputation sample genotypes from Section 3.2.2 were pruned to MAF > 0.05. A kinship matrix was computed for each autosome, then combined into a single genome-wide matrix using LDAK `--join-kins`. To obtain a LOCO kinship matrix for each autosome, each autosome's kinship matrix was then subtracted from this genome-wide matrix (LDAK `--sub-grm`).

3.2.4 Estimation of cell type abundance from expression

PBMC samples are a mixture of immune cells, and a fixed input of RNA extracted from that mixture is used to estimate expression, so estimates for genes that have cell type-specific expression depend on the relative abundances of each cell type in each sample. Sobolev *et al.* [162] showed these abundances shift after Pandemrix vaccination. As genotype can be assumed to stay constant, it is valid to compare the effect size of genotype on expression between multiple timepoints to call reQTLs, but changes in cell type abundance complicate this by modifying both expression (i.e. cell type-specific expression) and the effect of genotype on expression (i.e. cell type-specific eQTL effects). Immune cell abundance also varies naturally between healthy individuals [109, 255], so it is important to model these effects not only post-vaccination, but also at baseline.

Cell type abundance directly measured via fluorescence-activated cell sorting (FACS) was only available for a small subset of HIRD individuals (Section 2.2.1), so I computed cell type abundance estimates from the expression data as an alternative. *In silico* estimates have previously been used as covariates for eQTL analyses in bulk samples where cell type-specific effects are expected [71, 72, 74, 96]. As the estimates are based on the expression of multiple genes, it is not entirely circular to use them as covariates in this way for per-gene eQTL models. I selected `xCell` [298], which has been shown to outperform other deconvolution methods for cell type-specific eQTL mapping in blood [74]. `xCell` computes enrichment scores based on the expression ranks of approximately 10 000 signature genes derived from purified cell types, works for both array and RNA-seq expression data, and implements "spillover compensation" that reduces dependency of estimates between related cell types [298]. `xCell` was originally developed for tumour samples, so many of the built-in cell types are not expected to be in PBMCs. Reviewing the literature to find which broad classes of peripheral blood cell types are commonly expected in the PBMC compartment [96, 299, 300], I selected 7/64 of the built-in cell types: $CD4^+$ T cells, $CD8^+$ T cells, B cells, plasma cells, natural killer (NK) cells, monocytes, and DCs. Array and RNA-seq data from Section 2.2.8 and Section 2.2.7 were processed separately, as different internal `xCell` parameters are used for each platform. The large batch effect present in the array expression was

first removed using ComBat [211]. Finally, enrichment scores were standardised across timepoints, so that a score of zero represents the average abundance of that cell type across all timepoints.

As with actual cell type abundances, `xCell` enrichment scores are correlated (Fig. 3.2). Imprecise coefficient estimates due to multicollinearity may be a problem when these scores are included as independent variables in eQTL models*. To select a subset of cell type scores, I performed a principal component analysis (PCA) of the cell type scores separately in array and RNA-seq datasets (to prevent axes reflecting platform rather than cell type), then determined the number of PCs that exceeded the eigenvalues-greater-than-one rule of thumb [302]. In both array and RNA-seq datasets, this number of PCs was three. The cumulative percentage of variance explained by the top three PCs was 81.02 % and 74.58 % in the array and RNA-seq datasets respectively. Since the PCs between the array and RNA-seq are not directly comparable, I selected three cell types with high contributions to the top three PCs in both datasets: monocytes, NK cells, and plasma cells (Fig. 3.3)—Sobolev *et al.* [162] also reported monocytes and plasma cells to be the cell types with the highest abundance increases at days 1 and 7 respectively. Additionally, using the actual cell type scores rather than PCs as covariates provides more interpretable regression coefficients for those terms.

Scores were validated against FACS measurements from Sobolev *et al.* [162] in the subset of ∼40 individuals that had both expression and FACS data. Depending on each FACS panel's gating strategy for each cell subset, the data were in units of either absolute counts or percentage of the previously gated population. Values were normalised by rank-based INT within each panel and cell subset (Astle *et al.* [303] took a similar approach for cell abundance data using a quantile-based INT). Missing values were then imputed with MissForest [304], a random forest imputation method suitable for high-dimensional mixed-type data where $p \gg n$. The method establishes an initial guess for missing values using mean- or mode-imputation, then a random forest is trained on the observed part of the data and used to predict and update the values of the missing part. The process repeats iteratively until convergence.

Although the increases in `xCell` score for monocytes at day 1 and plasma cells at day 7 do reflect the increases in those cell types observed by Sobolev *et al.* [162], overall correlation between `xCell` and FACS was poor (Fig. 3.4). Substantial discrepancy is expected, as the cell types as defined in the `xCell` signatures do not directly correspond to the combinations of surface markers used for FACS; the comparison is against the closest match. The FACS gating strategy also meant that for some cell populations, the only available FACS measure was a proportion of the previously gated population, whereas `xCell` attempts to estimate scores that represent enrichments in the whole mixture. The accuracy of the built-in signatures may also be lower when applied to the expression matrix for a stimulated state, where an enrichment-based method may not be able to distinguish per-cell differential expression of signature genes from changes in cell abundance. A custom signature matrix can be used for `xCell`, perhaps drawn from an independent study with similar stimulation conditions as HIRD, such as Franco *et al.* [94], but this would not solve the issue of coupled differential expression and cell abundance. Weighing

*High correlation between predictors is not necessary nor sufficient by itself to induce multicollinearity (predictors being linearly-related), but multiple correlation (how well predictors can be predicted as linear combinations of other predictors) does have an inverse relationship with the standard error of coefficient estimates [301].

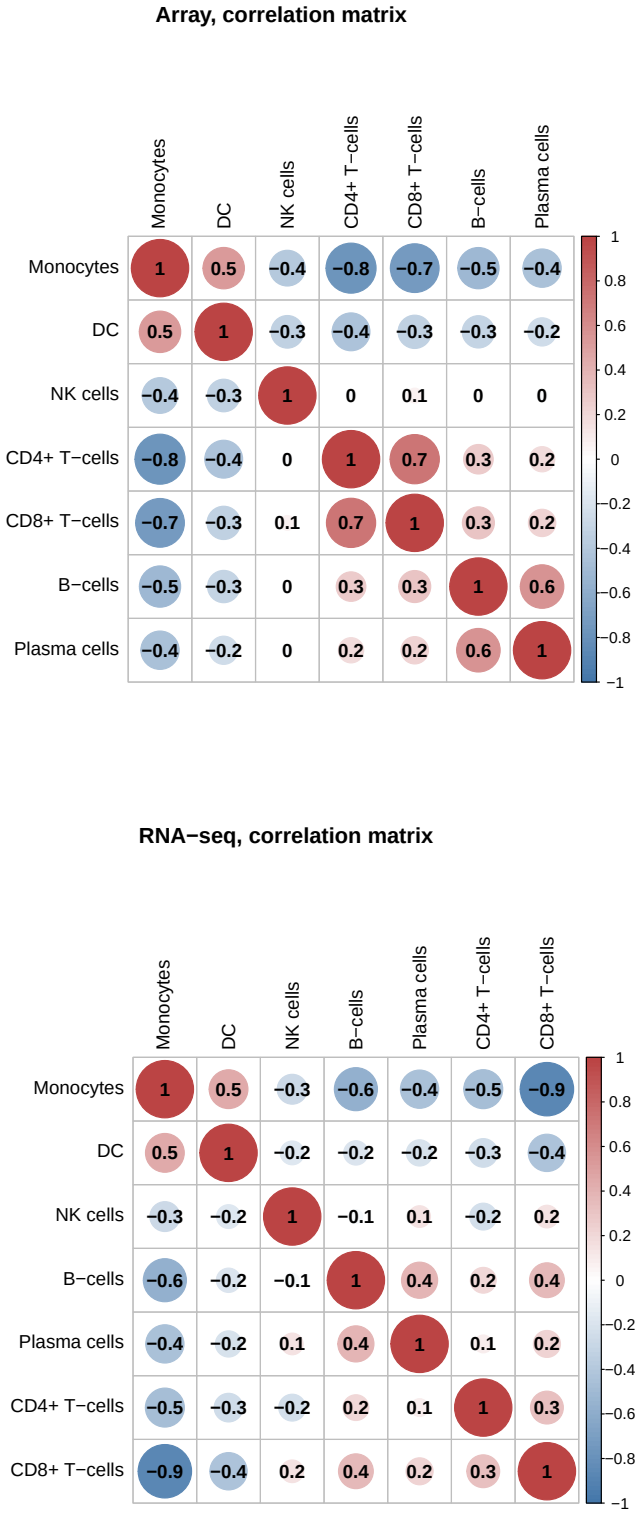


Figure 3.2: Correlation matrix of standardised `xCell` cell type enrichment scores in HIRD array and RNA-seq datasets. Rows and columns are hierarchically-clustered.

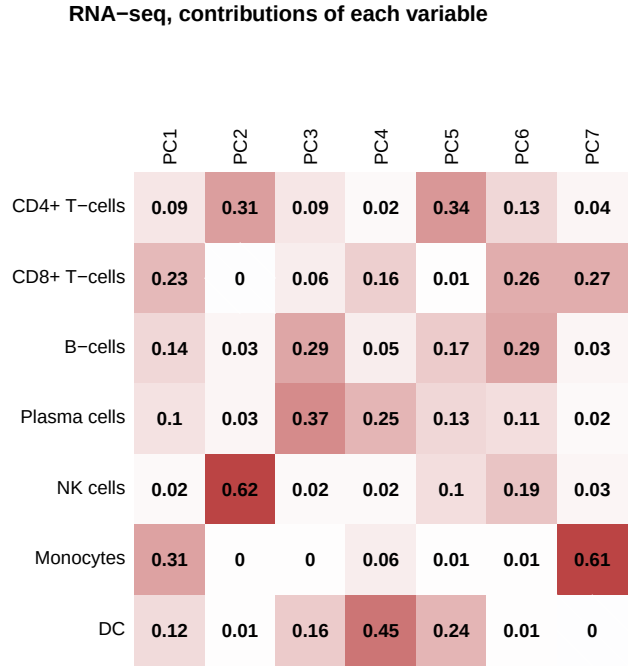


Figure 3.3: Contribution of each cell type score to each PC dimension after PCA of standardised `xCell` cell type enrichment scores. Contribution is calculated as the squared correlation between a variable and a PC (cos2), scaled to the sum of cos2 for all variables with that PC. High contributions indicate variables that are highly correlated with the PC.

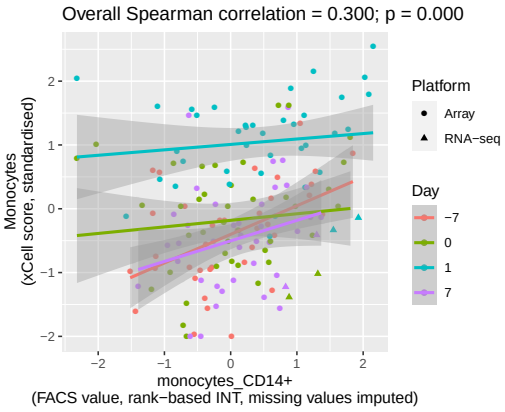
the downsides of having imperfect estimates of cell type abundance against the downsides of not accounting for abundance, or excluding samples without FACS measurements, I chose to continue the analysis using the `xCell` scores. These scores can distinguish large changes in cell abundances between days, but may not be reliable for distinguishing small differences in abundance between individuals within a timepoint.

3.2.5 Finding unmeasured covariates using factor analysis

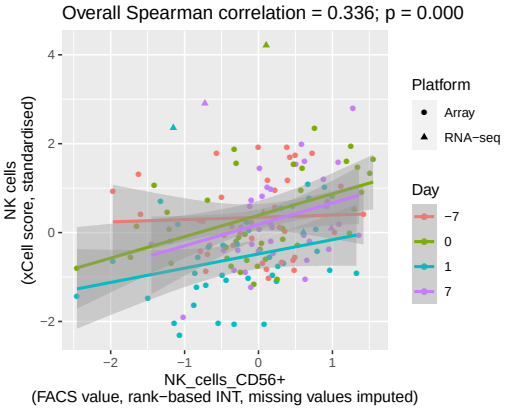
Apart from cell type abundance, a myriad of unmeasured variables contribute to expression variation. Hidden determinants of expression variation were learnt using PEER [182]. As suggested by Stegle *et al.* [182], I used `DESeq2::vst` [237] to perform between-sample normalisation and variance stabilisation on the RNA-seq count data*. ComBat [211] was then applied to merge array and RNA-seq data into a single log scale expression matrix per timepoint, treating the largest global effects on expression—the two array batches and three RNA-seq library preparation pools (Fig. 2.13)—as known batch effects. Given a set of known covariates (intercept, sex, four genotype PCs from Section 2.2.5 representing ancestry, and the three `xCell` scores estimated above in Section 3.2.4), PEER was used to estimate additional hidden factors that explain variation in the expression matrix. These can be technical (e.g. sample quality/concentration, library preparation plate/reagents, processing time, lane/flow cell) or biological (e.g. cell type composition, ancestry). Factors are assumed to be unmeasured covariates that have global effects on a large fraction of genes, whereas a *cis*-eQTL will typically only have local effects, so including factors as covariates should not introduce dependence with the genotype term, but should explain some of the residual variation, improving power to detect *cis*-eQTLs. The analysis was run per timepoint, otherwise global changes in expression between timepoints induced by the vaccine would be recapitulated as factors.

Correlating the estimated factors to a larger set of known covariates revealed many correlations with `xCell` estimates, indicating that cell type abundance does indeed have substantial global effects on the expression matrix (Fig. 3.5). These factors likely represent additional cell types with abundances that have a global effect on the expression matrix, and when used as covariates in combination with the three major cell type scores selected in Section 3.2.4, should improve overall adjustment for cell composition. There was little correlation with known array or RNA-seq batch effects, indicating ComBat did an adequate job of removing batch- and platform-dependent global effects on expression prior to PEER. Note that I did not leave this adjustment for PEER to perform, as ComBat estimates centering and scaling factors per gene and batch, whereas the use of PEER factors represents a mean-only per-gene adjustment. Given the severity of the batch effect in this dataset, especially between platforms, mean-only adjustment may be insufficient [214], particularly in the context of *cis*-eQTL mapping where associated variants will only explain a small fraction of expression variance.

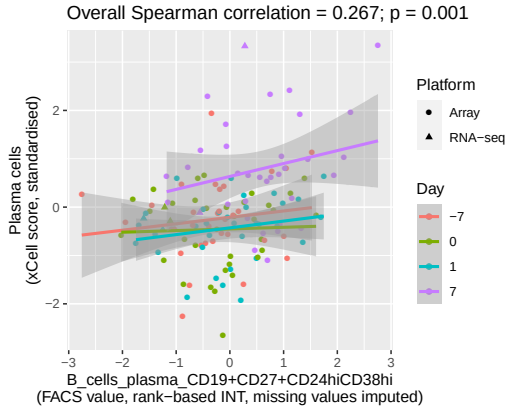
*The count data were taken from Section 2.2.7 before trimmed mean of M-values (TMM) normalisation and `limma::voom` transformation, as PEER cannot use the weights output by those methods for between-sample normalisation and variance stabilisation as `limma` can.



(a) Monocytes.



(b) NK cells.



(c) Plasma cells.

Figure 3.4: Comparison of standardised `xCell` scores with normalised HIRD FACS measurements, for monocytes, NK cells, and plasma cells. The comparisons are against the most comparable measurements available in the FACS data of Sobolev *et al.* [162]: $CD14^+$ monocyte count, $CD56^+$ NK cell count, and the proportion of $CD19^+$ B cells that were $CD19^+CD27^+CD24^{hi}CD38^{hi}$ plasma cells. Missing FACS values were first imputed with MissForest after rank-based INT transformation.

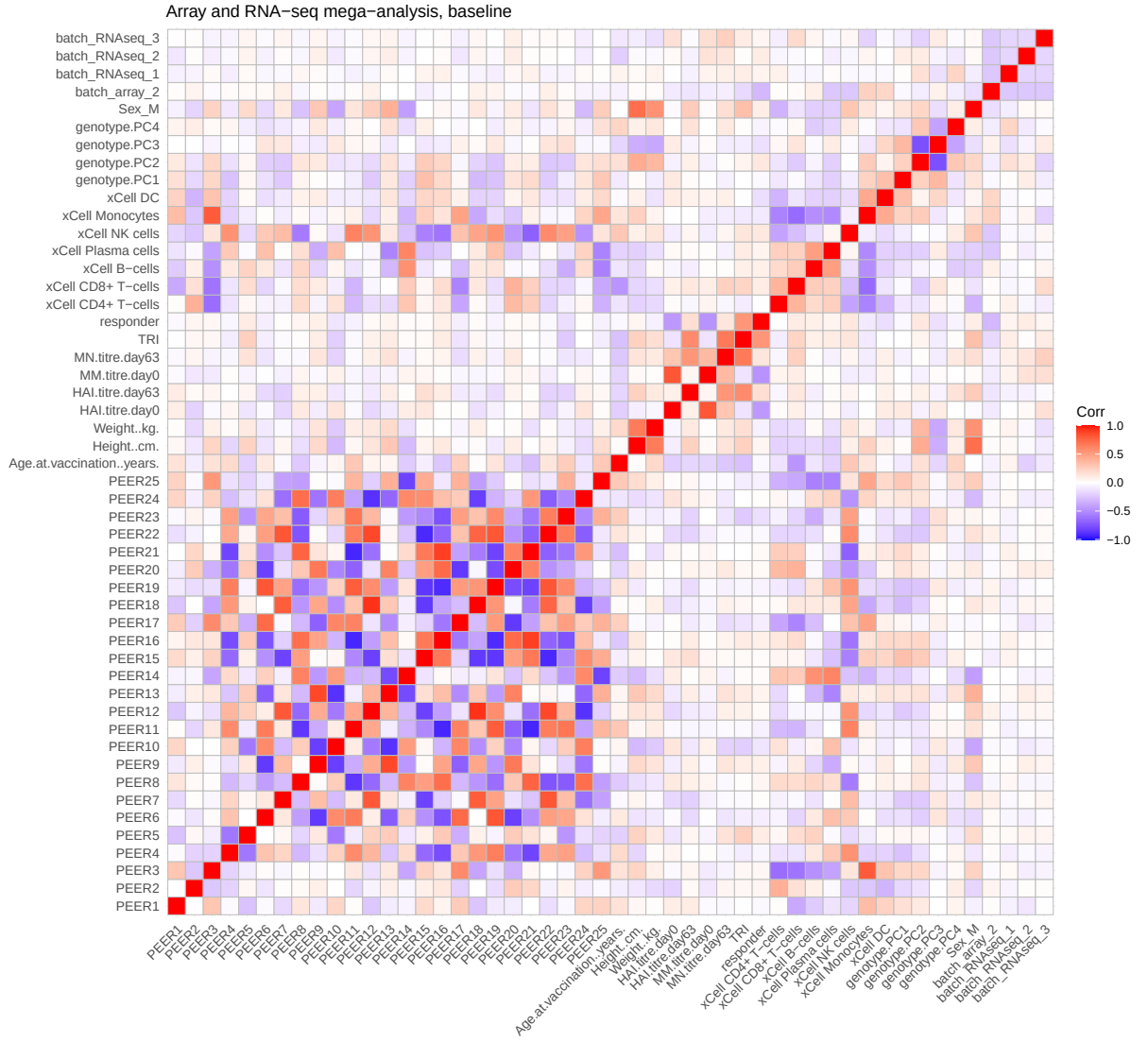


Figure 3.5: Correlation of known variables to the first 25 PEER factors estimated from the array and RNA-seq mega-analysis expression data at baseline. The known factors provided to PEER were sex, four genotype PCs, and monocyte, NK cell and plasma cell `xCell` scores. PEER factors are not constrained to be orthogonal like PCs, so correlations to known factors and other PEER factors are expected. The estimated factors have zero mean, and PEER implements automatic relevance determination [182], which decreases the variance of successive estimated factors to zero if they no longer explain additional expression variance. Although there are extensive correlations between higher numbered factors, these factors have near-zero variance.

3.2.6 eQTL mapping per timepoint

I mapped eQTLs within each timepoint using LIMIX [273], which implements univariate and multivariate LMMs with one or more random effects. Imputed genotype probabilities were converted to continuous alternate allele dosages using bcftools (1.7-1-ge07034a)*. Variants with sample AC < 15 of the minor allele within each timepoint were excluded, corresponding to a 5–7 % MAF depending on sample size (145 at baseline, 105 at day 1, and 107 at day 7). At these sample sizes, false discovery rate (FDR) cannot be controlled by standard hierarchical FDR methods without a MAF filter of approximately 5–10 % [305].

At each of 13 570 genes, at all *cis*-variants within within ±1 Mbp of the gene transcription start site (TSS), I fit the following model to map eQTLs:

$$y = 1 + \text{sex} + \sum_{i=1}^{4} \text{PC}_i + \sum^{3} \text{xCell} + \sum_{i=1}^{k} \text{PEER}_i + \beta x + u + \epsilon \tag{3.4}$$

where the eQTL effect size is the slope of the genotype fixed effect β, the average additive effect of the alternate allele [13]; and $u \sim N(0, \sigma_g^2 K)$ is a random effect with zero mean and covariance matrix proportional to the LOCO kinship matrix for variant x. For chromosome X variants, no LOCO matrix was available from LDAK, so the matrix for chromosome 1 was used. Known covariates and PEER factors from Section 3.2.5 were included. PEER factors are automatically weighted such that the variance of factors tends to zero as more factors are estimated, hence continuing to add more and more factors as covariates will not continue to improve eQTL detection power, and eventually the model degrees of freedom will be depleted. To optimise the number of factors k to include†, per-timepoint eQTL mapping was performed in chromosome 1, iteratively increasing the number of factors until including additional factors provides no further benefit, and the number of eQTLs detected stabilises. I settled on a final choice of $k = 10$ factors for baseline, 5 factors for day 1, and 5 factors for day 7 (Fig. 3.6).

3.2.7 Joint eQTL analysis across timepoints

Joint analysis was conducted with `mashr` [281] at 40 197 618 gene-variant pairs (mean of 2962 tests per gene) for which summary statistics from within timepoint mapping were available in all three timepoint conditions. For n conditions, the `mashr` model incorporates multiple $n \times n$ canonical and data-driven covariance matrices to represent patterns of effects across conditions. Canonical matrices include the identity matrix (representing independent effects between conditions), singleton matrices (effects only in one condition), a matrix of ones (equal effects in all conditions), and other patterns of correlations. Data-driven covariance matrices represent patterns of effect observed empirically, derived from dimension reduction of a strong subset of tests likely to have an effect in at least one condition. I took the most significant variants per gene per condition, which ensures strong condition-specific effects are included, then further filtered to only nominally significant tests, resulting in a strong subset of 45962 tests used

*https://samtools.github.io/bcftools/

†I avoid the commonly performed two-stage approach of using PEER residuals as expression phenotypes, as the degrees of freedom for the eQTL model will be incorrect, which can have a substantial effect on estimates at this modest sample size.

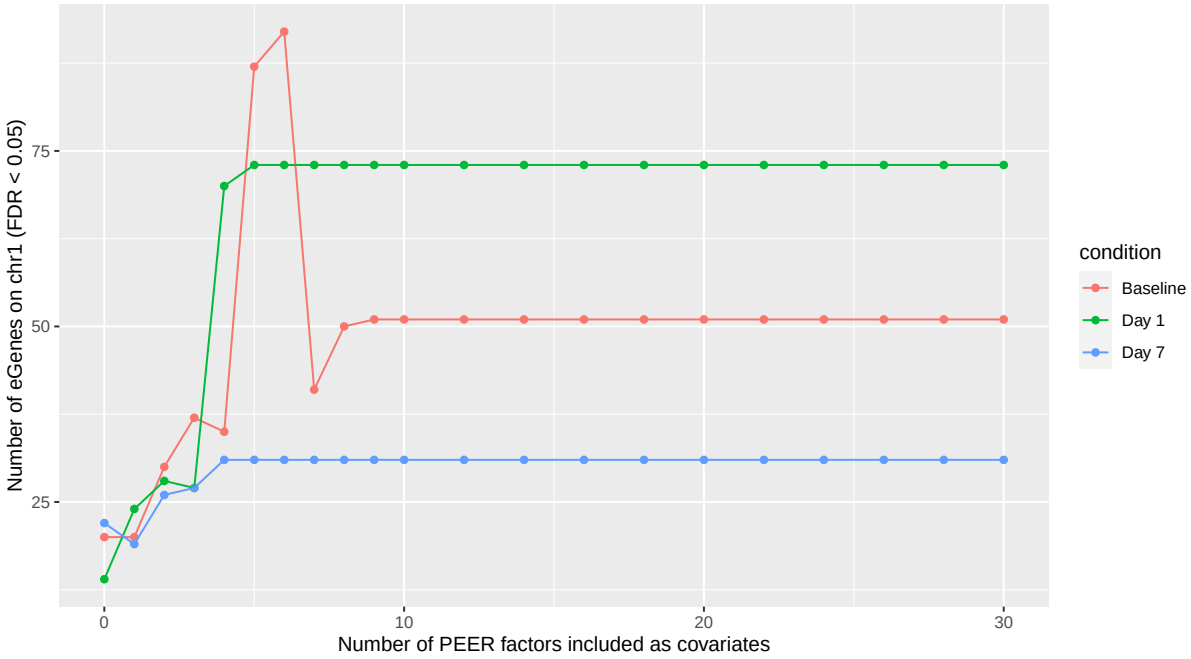


Figure 3.6: Number of significant genes with an eQTL detected on chromosome 1 as a function of the number of PEER factors included as covariates. FDR computed with hierarchical Bonferroni-Benjamini-Hochberg (BH) [305] with significance threshold set at 0.05. The number of eGenes stabilises since higher number PEER factors explain less and less variance in expression, thus having less and less influence on the regression.

to calculate data-driven covariance matrices.

The `mashr` model was trained on a random subset of 200 000 tests, using the exchangeable Z (EZ) model (assumes effects are independent of their standard errors, which performed better in GTEx data [281]). The correlation of null tests between conditions—critical to account for due to the repeated measures structure of the data—was estimated using `mashr::estimate_null_correlation`, which uses tests from the random subset that have small absolute z-scores. The fitted model was used as a prior to compute posterior effects and standard errors for all tests through shrinkage. A condition-specific Bayesian measure of significance is returned: local false sign rate (LFSR), which gives the probability that the declared sign of the effect is incorrect [238]. Note that `mashr` is the multiple-condition extension of `ashr` [238], previously used in Section 2.2.9.7 for computing posterior effects and their significance in DGE analyses.

3.2.8 Defining shared eQTLs and reQTLs

Many of the tested variants for each gene will be in high LD. To unambiguously select a lead eQTL variant per gene for comparison across timepoints, I selected the variant with the lowest LFSR over all conditions. If multiple variants had that same lowest LFSR value, ties were broken by highest imputation INFO, highest MAF, most upstream of the TSS, and finally genomic coordinate. Ties were not frequent. Sharing was then evaluated for that gene-variant pair across all three conditions.

Thresholding on the LFSR is not appropriate for determining sharing, as the difference

between significant and non-significant effect estimates in two conditions is not necessarily significant [306, 307]. Urbut *et al.* [281] provides a heuristic that two effects are shared by magnitude if they have the same sign, and are also within a factor of two, but this does not consider the posterior standard error of the estimates. Between a pair of effects in two conditions x and y, I computed a z-statistic for the difference in effects:

$$z = \frac{\beta_x - \beta_y}{\sqrt{\sigma_x^2 + \sigma_y^2 - 2\sigma_{xy}}} \tag{3.5}$$

This is a common strategy for comparing regression coefficients [279, 306] and has also been applied to call reQTLs by Kim-Hellmuth *et al.* [85]. Like Kim-Hellmuth *et al.* [85], I assume the posterior pairwise covariance of effects σ_{xy} is zero. This is conservative if the covariance is actually positive. A Wald test p-value for the difference can be computed, as under the null hypothesis of zero difference, asymptotically $z \sim \mathcal{N}(0, 1)$. I use nominal $p < 0.05$ as a heuristic threshold to separate shared and reQTL effects, and also computed the corresponding BH FDR per timepoint as a formal measure of significance. Note that even a nominal $p < 0.05$ threshold is still more stringent than calling sharing using the LFSR $= 0.05$ as a threshold (e.g. [74, 91]) or the 2-fold difference in magnitude threshold suggested by Urbut *et al.* [281].

Another statistic that quantifies the strength of an eQTL is the proportion of variance explained (PVE) by the variant. This was approximated using the following formula from Shim *et al.* [308] for variant X and expression Y:

$$\text{PVE} = \frac{\beta^2 \text{Var}(X)}{\text{Var}(Y)} = \frac{\beta^2 \text{Var}(X)}{\beta^2 \text{Var}(X) + \sigma_\epsilon^2} \approx \frac{\beta_p^2 2pq}{\beta_p^2 2pq + \sigma_p^2 2Npq} \tag{3.6}$$

where β is the beta from a simple linear regression of Y on X, σ_ϵ^2 is the residual error, β_p is the posterior beta from mashr, σ_p is its posterior standard error from mashr, N is the sample size, p is the sample MAF, and $q = 1 - p$. PVE was computed with the intention to allow for comparison of effect strength between timepoints, which have different sample sizes and different MAFs. In practice, this turns out to just be a monotonic transformation of the absolute posterior z-statistic $|\beta_p/\sigma_p|$, with more interpretable units.

3.2.9 Replication of eQTLs in a reference dataset

To validate the mega-analysis approach to eQTL mapping, I estimated the replication of significant eQTLs in a large independent reference. Due to the lack of large sample size eQTL maps specific to PBMC, I used the GTEx v8 whole blood dataset as my reference dataset ([54], $n = 670$, 51 % eGene rate). For lead variants called as significant at a given LFSR significance threshold in the HIRD dataset, for those variants that also exist in GTEx, I looked up their nominal p-values in GTEx. I then used `qvalue::qvalue_truncp` (v2.15.0*, implements theory from Storey *et al.* [309]) to estimate the proportion of those GTEx nominal p-values that are null (π_0), giving a measure of replication $\pi_1 = 1 - \pi_0$. The higher the π_1, the higher the proportion of HIRD eQTLs at this significance threshold replicating in GTEx. However, the higher the significance threshold,

*https://github.com/StoreyLab/qvalue

the fewer variants will have p-values meeting this threshold in HIRD, and thus fewer GTEx p-values will be available for computing π_1. More significant p-values in HIRD are also more likely to come from true eQTLs in general, so the higher the significance threshold, the lower the maximum nominal p-value from GTEx for those variants is likely to be. The π_1 procedure assumes a well-behaved p-value distribution with values over the full range $[0, 1]$, and reliability declines if the number of p-values is too small*, or the maximum p-value is much smaller than 1.

The mega-analysis had comparable replication rate to RNA-seq-only analysis for shared eQTLs at moderately stringent LFSR thresholds up to 10^{-5}, and better replication rate for very stringent LFSR thresholds (Fig. 3.7). The suggests the mega-analysis is not creating false positives due to technical effects from merging the expression data, and is preferred to either of the single-platform analyses. A caveat is this approach may be overestimating the replication rate as it does not take the direction or magnitude of eQTL effects into account. The numbers of reQTLs were too low to assess their replication using this method, and one might also not expect them to replicate in a baseline dataset such as GTEx whole blood, especially for those reQTLs significant only at post-vaccination timepoints.

3.2.10 Genotype interactions with cell type abundance

If the abundance of a particular cell type does truly modify the eQTL effect, then an interaction term between genotype and cell type abundance is required. As the additivity assumption no longer holds, a *ceteris paribus* interpretation does not make sense, as the effect of genotype holding cell type abundance constant depends on what value of cell type abundance you choose. One cannot adjust for modification just by including the main effect for cell type abundance; wrongly omitting a significant interaction term between cell abundance and genotype biases the estimation of the two main effects†. Given the modest sample size, I used a two-stage approach, where tests for interaction are only performed at a subset of tests. If the main effect estimates from main effect-only models (stage one) are used to filter SNPs for second stage testing, and are also independent from the interaction effect estimates in stage two, then the type I error can be controlled based on the number of interactions that are actually tested, rather the number of interactions that could have been tested [68, 311]. It is unclear whether this assumption holds in practice, as being able to detect a main effect at least implies that gene is sufficiently expressed for eQTL mapping. Nevertheless, the two-stage approach is often used for eQTL mapping with an interaction term [68, 71, 85, 96]. As the main purpose of my interaction analyses was scanning for cell type modification at detected reQTLs, I chose to test for interactions only at the lead eQTL variant for each gene with a significant main eQTL, controlling the FDR with BH—an approach used by Peters *et al.* [68] and Kim-Hellmuth *et al.* [85].

Models with interactions between genotype and other predictors were fit using `lme4qtl` [312]. The model specification was identical to Eq. (3.4), except the addition of three interaction terms between genotype and each `xCell` score. Significance was assessed using the likelihood ratio test

*In `https://github.com/StoreyLab/qvalue/pull/6#commitcomment-26277751` the developers suggest "you usually need a few hundred p-values" to reliably compute π_1.

†When a variable that is the function of another explanatory variable is omitted, this is known as functional form misspecification in the field of econometrics, a special case of omitted variable bias. Also see Mikucka *et al.* [310] for a review of bias caused by omitting significant interaction terms.

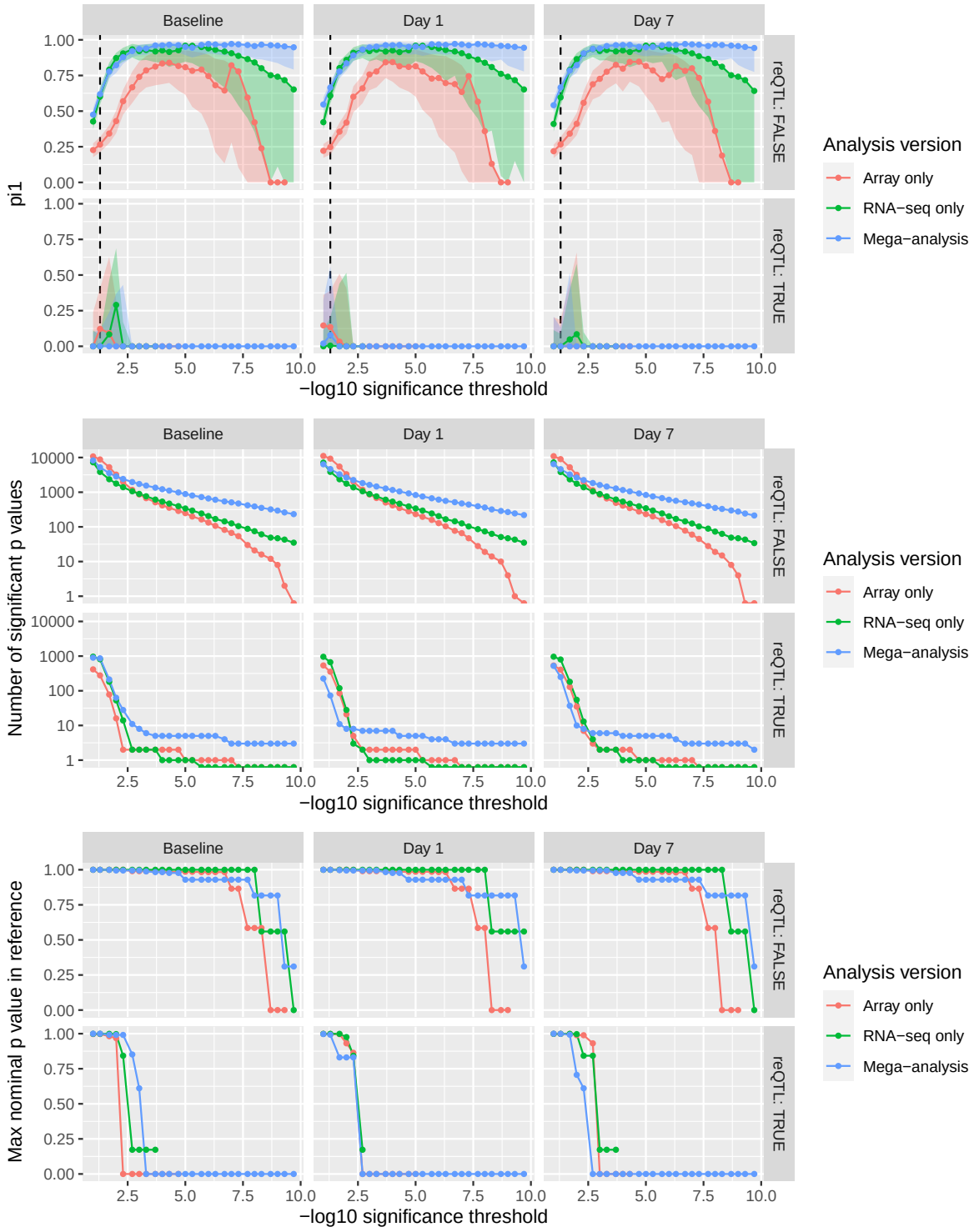


Figure 3.7: Replication rate π_1 of HIRD eQTLs in GTEx whole blood eQTL reference data. Three eQTL analyses were run in HIRD: array-only, RNA-seq-only, and a mega-analysis of the two datasets. The top panel shows π_1 replication in each analysis as a function of the significance threshold (LFSR), with the shaded region showing the 5th–95th percentile range of 1000 bootstraps. Vertical line shows LFSR = 0.05. The middle panel shows the number of significant HIRD eQTLs present in GTEx; this is the number of p-values available for computing π_1. The computation is more reliable when there are ∼1000 or more. The bottom panel shows the maximum nominal GTEx p-value for those variants use to compute π_1. The computation is more reliable when the maximum is near one. Each panel is stratified into HIRD shared eQTLs and reQTLs.

(LRT) versus the nested model with no interaction terms. Note that although PEER factors are correlated with `xCell` scores (Fig. 3.5), Kim-Hellmuth *et al.* [74] also used this approach, claiming that the interaction term between genotype and `xCell` scores should still be interpretable. I attribute their claim to the fact that there are no interaction terms between genotype and PEER factors, so the coefficients for the genotype-`xCell` score interactions still have their standard interpretation: $\beta_x + \beta_{cx}c$ increase in $\log_2$ expression per unit of effect allele dosage, where β_x is the main effect of genotype, β_{cx} is the interaction effect, and c is `xCell` score.

3.2.11 Gene set enrichment analyses

Ranked gene set enrichment analyses with `tmod::tmodCERNOtest` were conducted as described in Section 2.2.10, using blood transcription modules (BTMs) from Li *et al.* [240] (prefixed "LI"). Gene set overrepresentation analyses were run with `tmod::tmodHGtest` [241], which implements the hypergeometric test for enrichment in BTMs, controlling the FDR at 0.05 using the BH procedure. Gene set overrepresentation analyses were also run with `gprofiler2::gost` [313], which derives gene sets from Gene Ontology (GO), pathway databases (KEGG, Reactome, WikiPathways), regulatory motif databases (TRANSFAC, miRTarBase), protein databases (Human Protein Atlas, CORUM), and phenotype ontologies (HP). The default `gprofiler2` g:SCS method was used to control for multiple testing while accounting for the hierarchical structure of certain gene set databases like the GO. In both overrepresentation analyses, the 13 570 genes assayed by both array and RNA-seq were used as a custom background set.

3.2.12 Statistical colocalisation

Published GWAS and quantitative trait locus (QTL) summary statistics were downloaded for statistical colocalisation with per-timepoint HIRD eQTL summary statistics. Clinical blood count QTL maps generated by Astle *et al.* [303] in 173 480 European-ancestry participants were downloaded from `ftp://ftp.sanger.ac.uk/pub/project/humgen/summary_statistics/human/2017-12-12/hematological_traits/`. eQTL maps in fifteen FACS-sorted immune cell types generated by Schmiedel *et al.* [88] in a multi-ethnic cohort of 91 donors were downloaded from the eQTL Catalogue ([314], release 1, January 2020, `https://www.ebi.ac.uk/eqtl/`). The fifteen cell types included three naive innate immune cell types: classical monocytes ($CD14^{hi}CD16^{-}$), non-classical monocytes ($CD14^{-}CD16^{+}$), and NK cells; four naive adaptive immune cell types: B cells, $CD4^{+}$ T cells, $CD8^{+}$ T cells, and regulatory T cells (Treg); $CD4^{+}$ T cells and $CD8^{+}$ T cells stimulated with anti-CD3 anti-CD28 for 4 hours; and six $CD4^{+}$ memory T cell subsets: Th1, Th1/Th17, Th17, Th2, and memory Tregs. Inflammatory bowel disease (IBD) GWAS summary statistics generated by de Lange *et al.* [180] in a total of 59 957 European ancestry samples were downloaded from `https://www.ebi.ac.uk/gwas/studies/GCST004131`. Datasets were converted to GRCh38 coordinates with `rtracklayer::liftOver` (v1.46.0) [315] and harmonised to a standard format, matching variants between studies by genomic position and effect allele.

Multi-trait Bayesian colocalisation was performed using HyPrColoc [316]. HyPrColoc uses the pattern of per-variant summary statistics (betas and standard errors) from multiple traits in a locus to partition traits into clusters, where each cluster contains traits that share a causal variant. This can be seen as a multi-trait extension of pairwise Bayesian colocalisation methods

such as `coloc` [317]. Multi-trait colocalisation is more powerful than pairwise colocalisation for detecting causal variants shared between more than two traits, and large numbers of traits can be analysed simultaneously in a computationally efficient manner. The method formally assumes that studies generating the summary statistics for each trait are independent, but performs well even when there is complete sample overlap between traits [316]. If studies are non-independent, it is assumed the LD structure is the same across those studies (which holds in the case of multiple QTL maps generated from the same individuals). Each trait is assumed to have no more than one causal variant in the locus. Finally, it is assumed the causal variants for each trait are present in the input.

As with any Bayesian colocalisation method, the choice of priors and other algorithm parameters is influential. HyPrColoc implements variant-level priors where the prior depends on the number of traits a variant is causally associated with: `prior.1` is the prior probability that a variant is causal for one trait (default $= 1 \times 10^{-4}$), and $1 -$ `prior.2` is the prior probability that a variant is causal for an additional trait, given it is causal for one trait (default $= 0.98$). The prior for a variant being causal for a third trait given it is causal for two traits is $1 - (\texttt{prior.2})^2$, and so on. In the two trait case, the setup is identical to `coloc` [317]. `prior.2` tends to be more influential than `prior.1`, as it controls the probability of association with more and more traits.

The posterior probability of colocalisation for a cluster of traits is the product of regional association and alignment probabilities. The regional association probability is the probability there is a shared association region within the locus for all the traits in the cluster, containing one or more causal variants. The alignment probability is the probability that regional association is due to a single causal variant, rather than one or more variants in strong LD. A branch and bound algorithm is run, starting with all traits in one cluster, then recursively partitioning traits into subsets, assessing regional association and alignment probabilities for subsets at each iteration. The end result is clusters of traits sharing a causal variant, with each cluster having a distinct causal variant. Only clusters with more than one trait, and regional association and alignment probabilities above `reg.thresh` (default $= 0.5$) and `align.thresh` (default $= 0.5$) are reported.

In sensitivity analyses using the `sensitivity.plot` function, I fixed the less influential `prior.1` at the default of 1×10^{-4}, then iterated over combinations of four choices of `prior.2` (0.98, 0.99, 0.995, 0.999), five choices of `reg.thresh` (0.5, 0.6, 0.7, 0.8, 0.9), and five choices of `align.thresh` (0.5, 0.6, 0.7, 0.8, 0.9). Each range starts at the default value and becomes more stringent, requiring stronger and stronger evidence for clusters of colocalised traits to be identified.

3.3 Results

3.3.1 Mapping reQTLs in the HIRD cohort

To characterise the effect of common host genetic variation on expression response to Pandemrix, I mapped *cis*-eQTLs for each gene (±1 Mbp of the TSS) within each timepoint condition (baseline, day 1, and day 7), then conducted joint analysis of all three timepoints with `mashr` [281] to obtain per-timepoint posterior effect sizes (betas), posterior standard errors, and measures of

significance (the LFSR). At LFSR < 0.05, 6887/13 570 genes (50.8 %) were eGenes (genes with a significant eQTL) in at least one timepoint. The most significant tested variant over all timepoints was selected as the lead variant for each gene, then reQTLs were defined by comparing the effect size of this lead variant between each pair of timepoints. This guards against differences in effect size from differential tagging efficiency (of an assumed single causal variant), a potential issue if different variants are compared across timepoints. Fig. 3.8 shows patterns of sharing over timepoints for the lead variant for each of the 13 570 genes, illustrating the difference between calling reQTLs using a significance threshold and a difference in betas method. For example, there were 85 eQTL-eGene pairs significant only at day 1 post-vaccination (LFSR < 0.05); of these, only 40/85 were called as reQTLs by the difference in betas method. The difference in betas method is more strict because calling by significance alone would call a reQTL for an eQTL with LFSR $= 0.049$ at baseline and LFSR $= 0.051$ at day 1, even if the effect sizes are similar. Shared eQTLs were well-replicated in GTEx whole blood (Fig. 3.7).

The largest number of eGenes was detected at baseline, reflecting the larger sample size compared to other timepoints. Most eQTLs were shared across timepoints; these were also the strongest eQTLs in terms of both maximum absolute beta and PVE across timepoints, highlighting the power advantage for mapping shared effects granted by joint analysis. Based on difference in effect size between any pair of timepoints (nominal $p < 0.05$), 1154/6887 (16.8 %) eQTLs were classified as reQTLs. Of these, 690/1154 were reQTLs between both day 1 vs. baseline and day 7 vs. baseline, and only 23/1154 were unique to the day 7 vs. day 1 comparison, indicating most reQTL effects were differences between pre- and post-vaccination (Fig. 3.9).

3.3.2 Characterising reQTLs post-vaccination

To characterise the eGenes associated with post-vaccination reQTLs, I ranked eGenes by the increase in PVE for their associated reQTLs from baseline to day 1 and baseline to day 7, then performed ranked gene set enrichments with `tmod::tmodCERNOtest` [241]. The same four modules were significant at both post-vaccination timepoints: "immune activation - generic cluster" (LI.M37.0, day 1 FDR $= 1.28 \times 10^{-6}$, day 7 FDR $= 3.39 \times 10^{-6}$), "enriched in monocytes (II)" (LI.M11.0, day 1 FDR $= 4.69 \times 10^{-3}$, day 7 FDR $= 1.88 \times 10^{-2}$), "cytoskeleton/actin (SRF transcription targets)" (LI.M145.0, day 1 FDR $= 2.07 \times 10^{-2}$, day 7 FDR $= 2.04 \times 10^{-2}$), and "MHC-TLR7-TLR8 cluster" (LI.M146, day 1 FDR $= 2.07 \times 10^{-2}$, day 7 FDR $= 2.04 \times 10^{-2}$). These enrichments are weak, but consistent with immune activation driving post-vaccination reQTLs. Given that TLR7 and TLR8 are primarily expressed in monocytes, macrophages, and DCs [318], and SRF is a regulator of the cytoskeleton in macrophages [319], there is suggestive evidence reQTLs may be enriched in genes specific to these phagocytotic APCs.

Changes in PVE do not capture changes in allelic direction. I classified post-vaccination reQTLs into one of three effect types: magnified, where the beta increases after vaccination but remains the same sign; dampened, where the beta decreases after vaccination but remains the same sign; and opposite, where the allelic direction changes after vaccination. As LFSR quantifies uncertainty in the sign of the effect, I did not make this classification for reQTLs that were not significant both at baseline and post-vaccination—the effect type for these are unclear. The classifications are shown in Fig. 3.10, plotting all 6887 shared or reQTLs by their

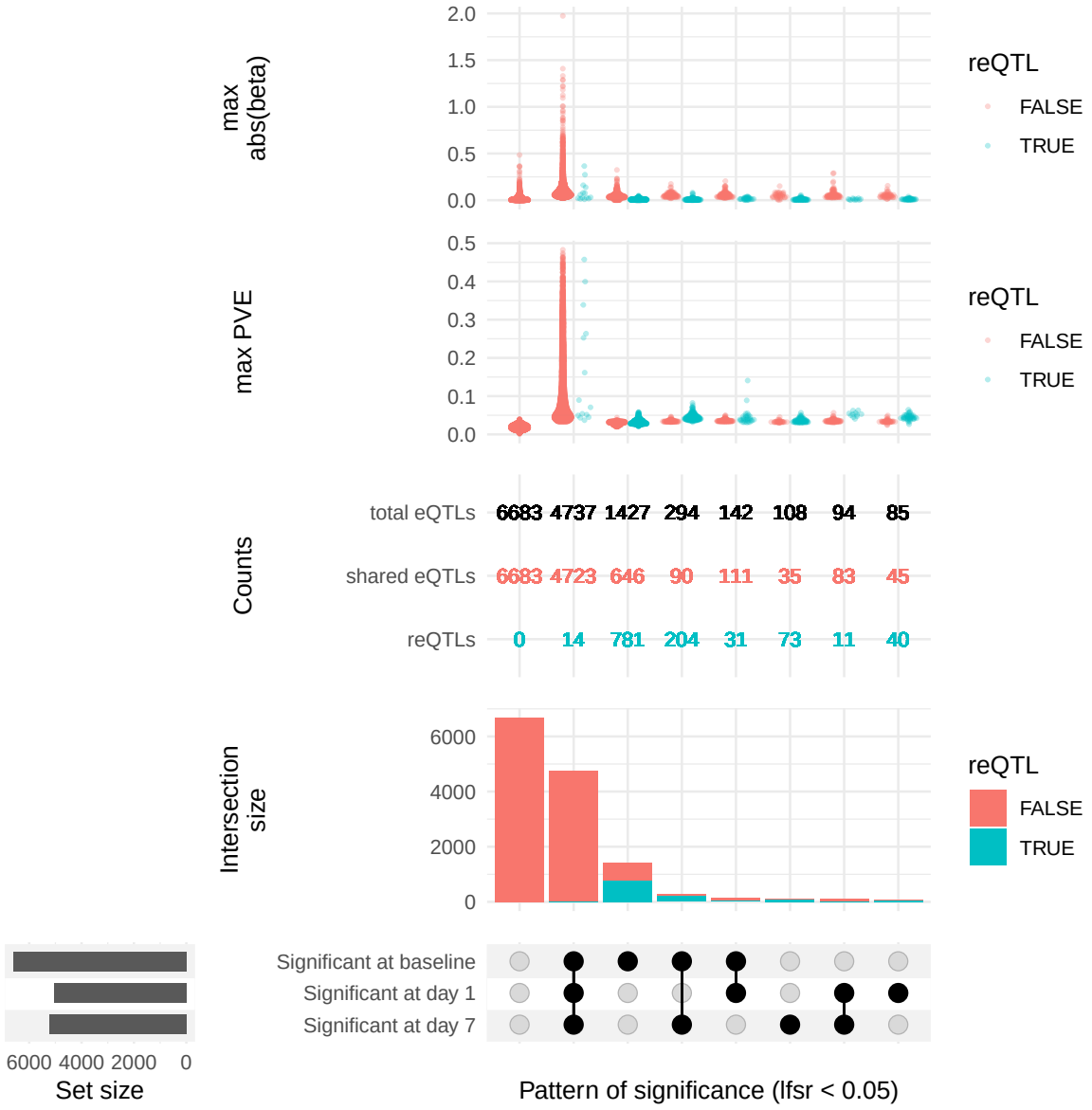


Figure 3.8: Summary of HIRD eQTL mapping from mega-analysis of array and RNA-seq expression data, binned by patterns of lead variant significance over the three timepoints. The most significant variant for each gene over all timepoints was chosen as the lead variant. Significant eQTLs (LFSR < 0.05) were found at 6887/13 570 eGenes. These were classified as reQTLs if there was a significant difference in beta (nominal $p < 0.05$) between any pair of timepoints, given that the eQTL was significant in at least one of those two timepoints. For variants in each bin, counts of shared and reQTLs, and distributions of maximum beta and PVE across timepoints are shown.

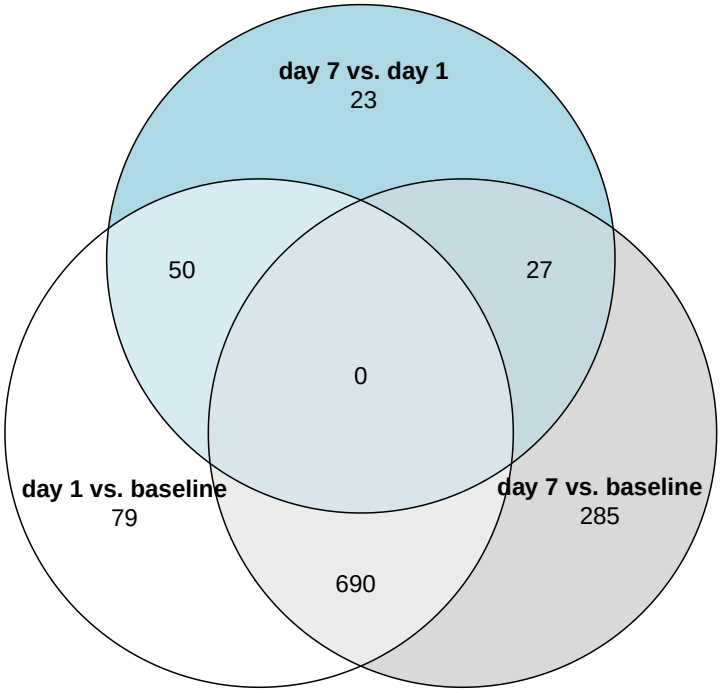


Figure 3.9: eGenes where the lead eQTL was a reQTL between a pair of timepoints. reQTLs were observed for 1154 unique eGenes, defined by a significant difference in beta between timepoints (nominal $p < 0.05$).

distance relative to the eGene TSS. Shared eQTLs have difference in beta z-statistics close to zero, and are concentrated close to the TSS as expected. reQTLs have a distribution of mostly negative z-statistics clearly separated from the shared eQTLs at both day 1 and 7, and these are mostly unclear or opposite rather than dampened effect types. Many of these unclear effects may actually be dampening, but as the sample size is greatest at baseline, dampening effects are hard to distinguish from drops in power at post-vaccination timepoints, whereas an opposite effect significant in both timepoints is unambiguous.

Fig. 3.10 also shows that reQTLs tended to be distributed evenly across the entire *cis* window, raising the question or whether they are enriched in false positives. A nominal $p < 0.05$ difference in betas threshold—although stricter than many other methods (see Section 3.2.8)—may still be too lax for calling reQTLs, so I applied a stronger BH FDR = 0.2 threshold. At this threshold, the only remaining reQTL was at day 1 was for *ADCY3* (nominal $p = 8.68 \times 10^{-6}$, FDR = 0.12)—the next smallest FDR value was 0.65. At day 7, 676 significant reQTLs had FDR < 0.2, of which 221 were opposite effects. Performing gene set overrepresentation analysis on the set of 221 eGenes to identify a shared biological signature was relatively uninformative, and revealed only one enrichment for genes *PRKACB*, *PRKACA*, *SAR1B*, and *APOE* in "Plasma lipoprotein assembly" (Reactome pathway identifer R-HSA-8963898, set size = 11, adj. $p = 0.01$). Since Fairfax *et al.* [59] found cases of opposite reQTL effects between B cells and monocytes, and B plasma cells but not monocytes were increased in abundance at HIRD day 7 [162], I also performed gene set overrepresentation analysis using BTMs to detect if eGenes related to B cells were enriched at day 7. No significant enrichments were identified.

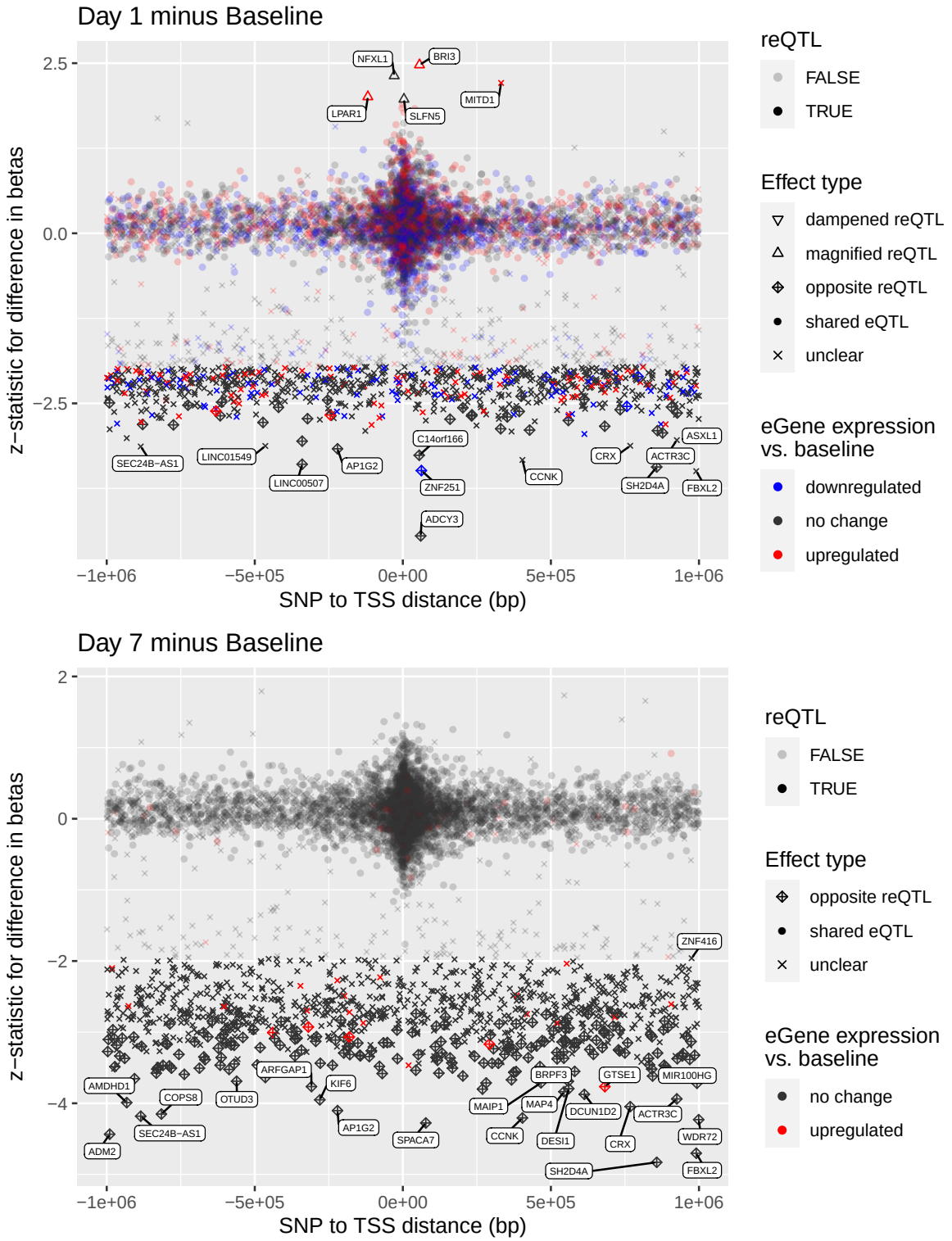


Figure 3.10: *z*-statistic for difference in beta post-vaccination versus baseline for shared and reQTLs, against distance from the eGene TSS. For each plot, all eQTLs significant in either timepoint are shown. Shared eQTLs can only have the shared effect type. An unclear effect type indicates the eQTL in question is not significant in both timepoints. Allelic direction of effect is aligned so that the beta at baseline is positive.

3.3.3 Exploring possible mechanisms generating reQTLs

3.3.3.1 Differential expression of genes with reQTLs

As gene set analyses based on the effect sizes of reQTLs at different timepoints had been largely uninformative, I considered whether reQTLs could be characterised by shared mechanisms. One mechanism that could generate reQTLs is differential expression, where an eQTL is not detected at baseline because the eGene is not expressed, and vaccine-stimulated upregulation reveals the effect post-vaccination.

Fig. 3.10 shows whether each eGene was up or downregulated at the timepoint based on the DGE analyses in Chapter 2. Visually, a large number of reQTLs occur without corresponding differential expression. Statistically, compared to genes without reQTLs, genes with reQTLs were less likely be differentially expressed at day 1 post-vaccination (26.5 % for genes with reQTLs, 42.3 % for genes without reQTLs, Fisher's test $p < 2.20 \times 10^{-16}$). This was also the case when restricting the scope to only eGenes (26.5 % for genes with reQTLs, 47.9 % for genes with shared eQTLs, Fisher's test $p < 2.20 \times 10^{-16}$). At day 7, no significant difference was observed comparing genes with and without a reQTL (2.2 % for genes with reQTLs, 1.4 % for genes without reQTLs, Fisher's test $p = 0.05$), but compared to genes with shared eQTLs, genes with reQTLs were more likely to be upregulated (2.2 % for genes with reQTL, 1.1 % for genes with shared eQTLs, Fisher's test $p = 0.01$). Twenty-two genes with both day 7 reQTLs and upregulated expression were strongly enriched within gene sets related to the cell cycle, such as "mitotic cell cycle" (GO biological process term GO:0000278, term size = 914, intersection size = 12, `gprofiler2::gost` [313] adj. $p = 1.42 \times 10^{-4}$), "cell cycle (I)" (LI.M4.1, module size = 137, intersection size = 12, `tmodHGtest` [241] FDR $= 1.41 \times 10^{-16}$), and "mitotic cell cycle in stimulated CD4 T cells" (LI.M4.5, module size = 33, intersection size = 3, `tmodHGtest` FDR $= 1.35 \times 10^{-3}$). However, these 22 genes previously appeared in Fig. 3.10 as having reQTL with decreased or opposite effect at day 7 versus baseline, making it implausible that the generating mechanism is increased detection power due to upregulation. The enrichment for cell cycle is likely driven by the DGE signal alone, especially as similar cell cycle gene modules were detected to be strongly upregulated at day 7 in Section 2.3.1.2.

The presence of reQTLs without DGE is exemplified by the strongest reQTL at each day. The only significant day 1 reQTL at difference in betas FDR < 0.2 was for *ADCY3* (Fig. 3.11). Computing the PVEs, this reQTL explained 1.9 % of *ADCY3* expression variation at baseline, increasing to 14.1 % at day 1, yet *ADCY3* was not differentially expressed from baseline to day 1 ($\log_2$ FC $= 0.10$, LFSR $= 0.26$). The strongest day 7 reQTL was at *SH2D4A* (difference in betas FDR $= 0.02$, Fig. 3.12). Here, the reQTL variant explained similar amounts of expression variation at baseline (PVE $= 8.2$ %) and day 7 (PVE $= 9.0$ %), with opposite directions of effect. Again, there was no differential expression. There is strong evidence that many post-vaccination reQTLs are generated by mechanisms unrelated to DGE.

3.3.3.2 Genotype by cell type abundance interaction effects

The presence of cell type-specific eQTL effects combined with changes in cell abundance between timepoints was considered as an alternate explanation generating reQTLs. Even if an eGene is

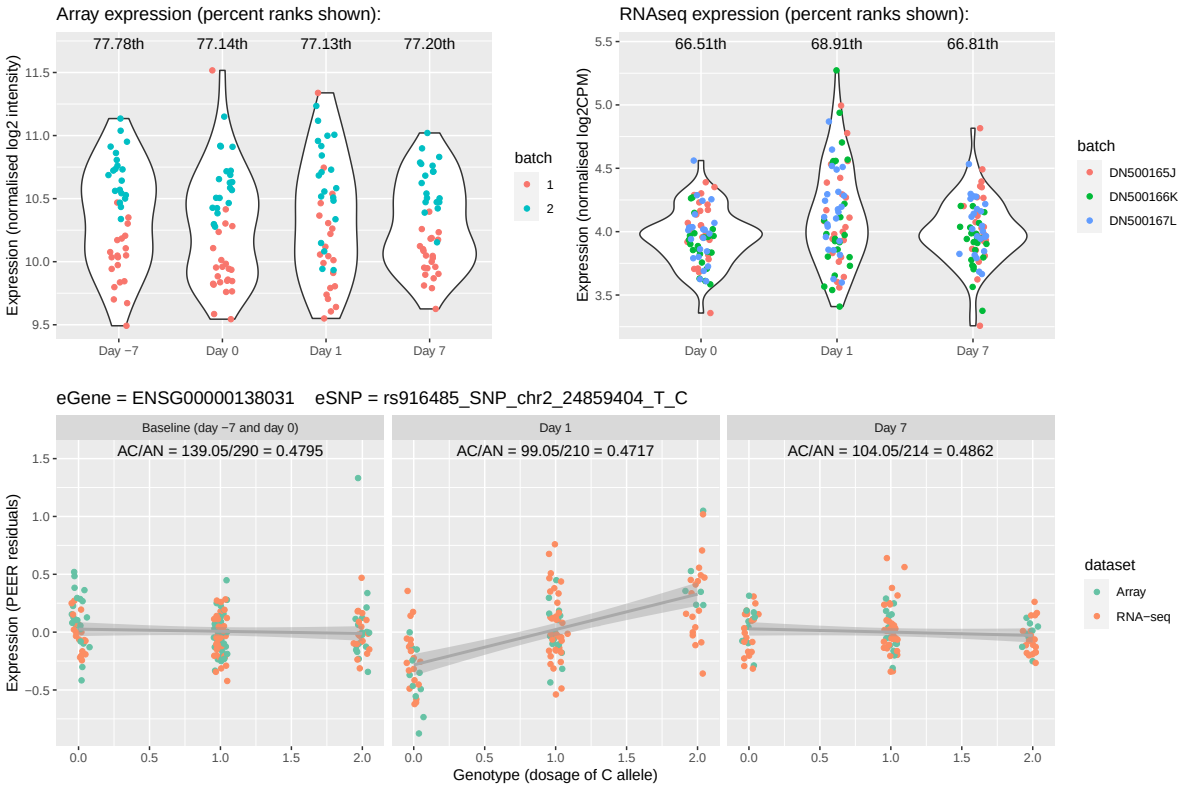


Figure 3.11: Expression and lead eQTL of *ADCY3* over study timepoints. Normalised array (top-left) and RNA-seq (top-right) expression before batch effect correction with ComBat and eQTL mapping. Bottom: eQTL effects at each timepoint condition in the mega-analysis of array and RNA-seq data.

not differentially expressed on average in bulk expression data, the composition of cell types that are the source of that gene's transcripts can change. `xCell` [298] enrichment scores were used to estimate abundance of seven PBMC cell types from the expression data. After pruning highly correlated cell types to avoid multicollinearity, standardised scores for monocytes, NK cells, and plasma cells were tested for interaction with genotype. Within each timepoint, full eQTL models including genotype main effect, the three cell type abundance main effects, and three cell type-genotype interaction terms, were fit using `lme4qtl` [312], then compared to a nested model excluding the three interaction terms with a LRT. Significant cell type interactions were detected at 16/1154 reQTL-gene pairs in at least one timepoint (BH FDR < 0.05). Fifteen were significant in only one timepoint: baseline (*SLAMF8*, *CSE1L*, *MAST1*, *DLGAP1*), day 1 (*ZNF519*, *LPAR1*, *ADCY3*, *NAA20*, *EPB41L5*), or day 7 (*APOL6*, *ADAR*, *ADAM17*, *UHRF2*, *MST1*, *CUL1*).

For *ADCY3* at day 1 (full vs. nested FDR $= 9.54 \times 10^{-5}$), although the genotype effect was 0.26 (standard error $= 0.03$) in the nested model; the estimate in the full model was -0.01 (0.07), with the three cell type-genotype interaction term estimates being 0.21 (0.05) for monocytes, -0.01 (0.04) for NK cells, and 0.02 (0.07) for plasma cells. The small magnitude of the genotype main effect in the full model compared to the nested model suggests the eQTL effect is driven largely by the monocyte score (or a cell type that is highly correlated with monocyte score in Fig. 3.2). In the case where the monocyte score is zero (representing an average abundance across samples, as scores were standardised), the effect of increasing genotype dosage on *ADCY3*

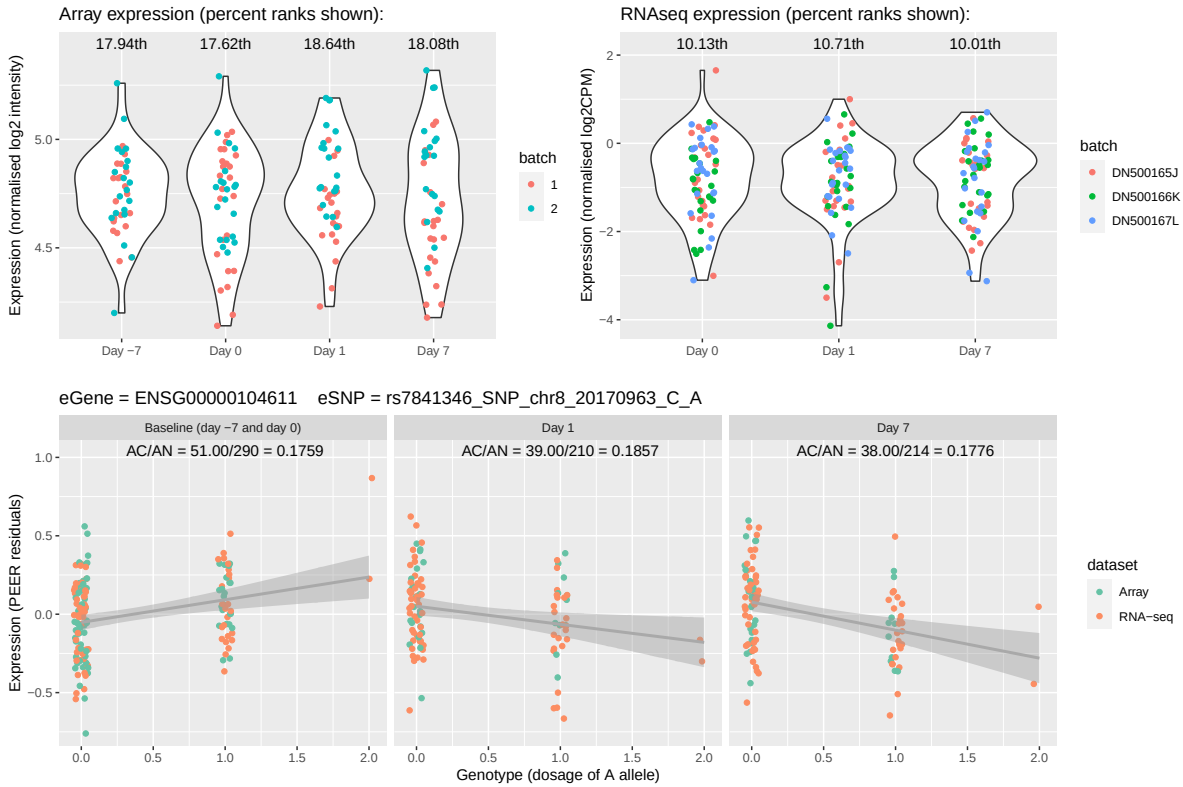


Figure 3.12: Expression and lead eQTL of *SH2D4A* over study timepoints. Normalised array (top-left) and RNA-seq (top-right) expression before batch effect correction with ComBat. Bottom: eQTL effects at each timepoint condition in the mega-analysis of array and RNA-seq data.

expression is minimal. Fig. 3.13 and 3.14 illustrate this effect. Monocyte abundance has no effect on expression at baseline, increases after vaccination, and modifies the effect of genotype on expression at day 1. It is feasible that the mechanism generating reQTLs at the remainder of these genes also involve cell type-specific eQTL effects, but unlike at *ADCY3*, I have not yet examined which of the three cell abundance scores have the greatest contributions.

3.3.3.3 Colocalisation with external QTL datasets at the *ADCY3* locus

The day 1 *ADCY3* reQTL is of particular interest, as reQTLs were also found for *ADCY3* in blood approximately 1 day after stimulation with TIV [94], rhinovirus [83], and *Mycobacterium leprae* [92]. The locus containing *ADCY3* has also been implicated in disease risk for immune-mediated inflammatory diseases (IMIDs) such as IBD [180], and *ADCY3* expression in immune cells in gut mucosa has been suggested to contribute to Crohn's disease (CD) risk (a subtype of IBD) [320]. Aside from monocytes, *ADCY3* is expressed in a wide range of immune cells: $CD4^+$ T cells, $CD8^+$ T cells, B cells, and NK cells (Fig. 3.15). Identifying cell type-dependent eQTLs through genotype-cell type abundance interaction terms cannot distinguish between cell types with highly correlated abundances [74]; the similar contributions to `xCell` score PC1 by monocyte, $CD4^+$ T cell, and $CD8^+$ T cell scores indicates cell types with *ADCY3* expression are indeed correlated in HIRD (Fig. 3.3). Given the *ADCY3* locus is associated with response to a wide range of immune stimuli, and also an IMID, I conducted colocalisation analysis to test if shared causal variants may be driving these associations, and to determine which among the

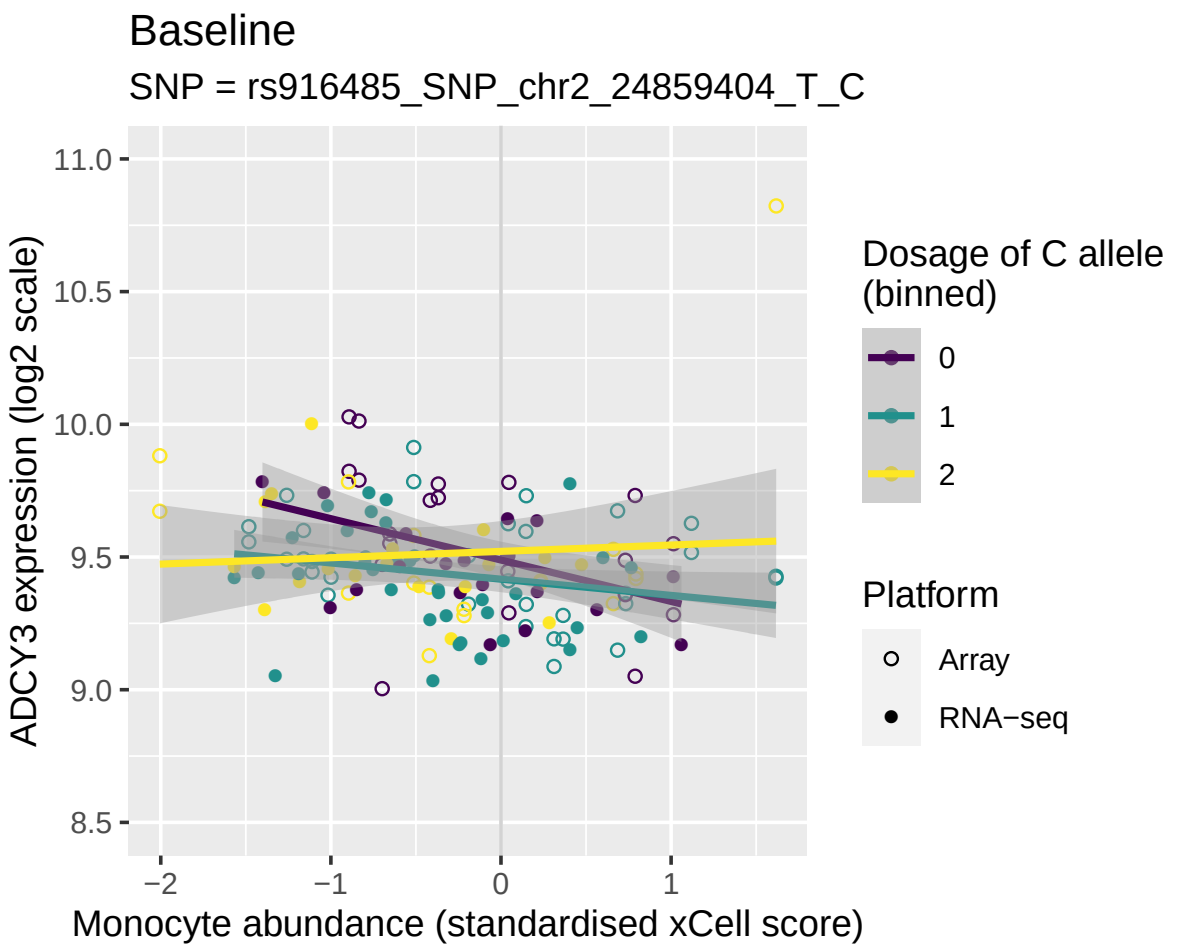


Figure 3.13: Effect of estimated monocyte abundance on *ADCY3* expression at baseline, stratified by genotype at a day 1 *ADCY3* reQTL.

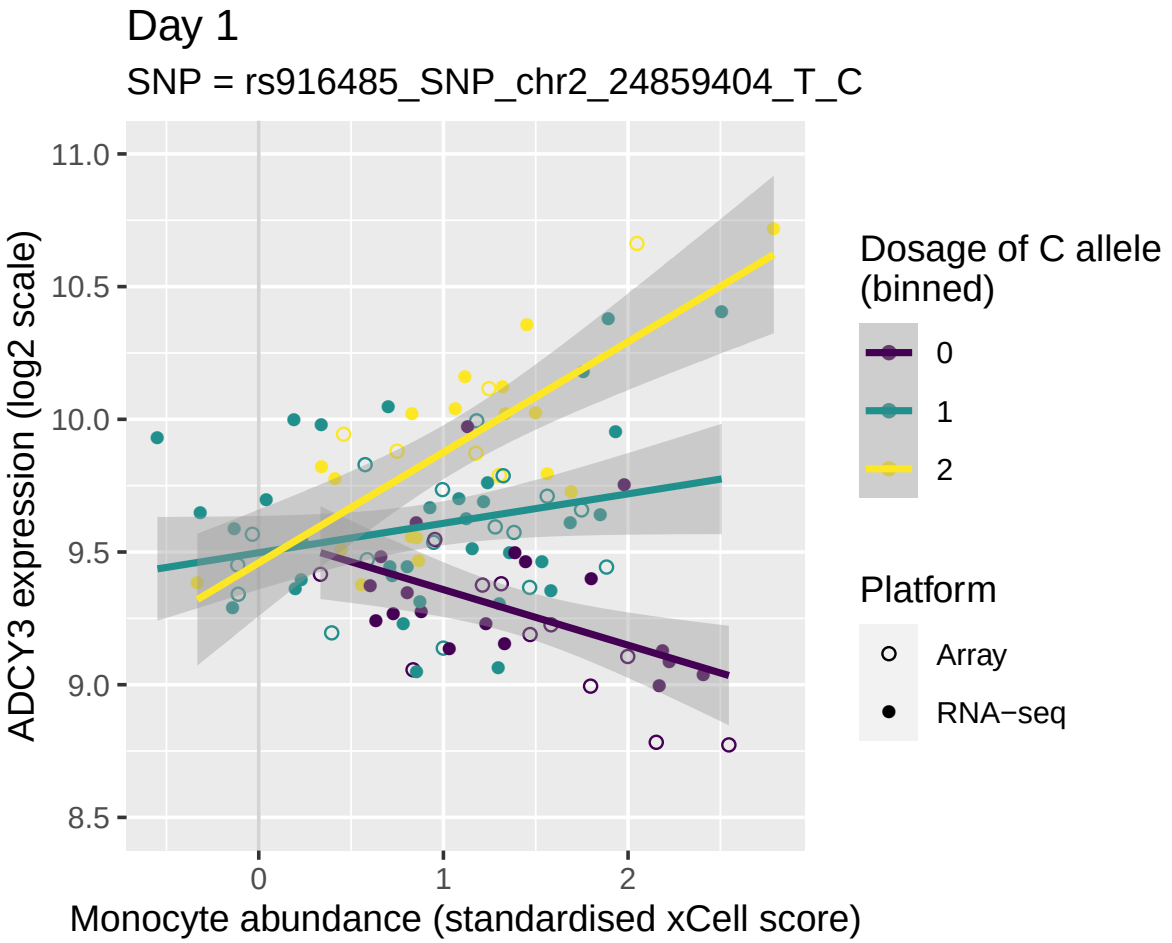


Figure 3.14: Effect of estimated monocyte abundance on *ADCY3* expression at day 1, stratified by genotype at a day 1 *ADCY3* reQTL.

correlated cell types expressing *ADCY3* are mostly likely responsible.

In a ±500 Mbp window around the lead reQTL variant rs916485, I performed Bayesian multi-trait colocalisation (HyPrColoc [316]) of the three per-timepoint *ADCY3* eQTL summary statistics with external summary statistics: *ADCY3* eQTL in 15 sorted immune cell populations from Schmiedel *et al.* [88], monocyte count QTL from Astle *et al.* [303], and IBD GWAS from de Lange *et al.* [180]. There were 1054 variants present in all 20 sets of summary statistics.

HyPrColoc identifies clusters of traits that colocalise at different causal variants in the locus. As Bayesian colocalisation can be sensitive to the choice of priors, I performed a sensitivity analysis iterating over configurations of priors and other algorithm parameters, ranging from default to more stringent parameter values. Two stable clusters were identified across 100 configurations of parameters (Fig. 3.16). A set of three traits—*ADCY3* expression at HIRD day 1, and in naive classical and non-classical monocytes—clustered in ∼65 % of tested configurations. A set of nine traits—IBD, and expression in eight naive and memory $CD4^+$ T cell subsets—clustered in ∼90 % of tested configurations. The remaining traits did not robustly cluster with any other traits over the tested configurations, except for the rare inclusion of HIRD baseline *ADCY3* expression into the larger cluster for less stringent configurations. The value of `prior.2` (the probability that a variant associated with at least one trait is not associated with any additional traits) was subsequently set to 0.99 (default = 0.98) to increase stringency, preventing this inclusion of baseline HIRD expression into the larger cluster. The values of other priors and algorithm parameters were left at their defaults, producing the final clustering shown in Fig. 3.17.

Under the final configuration, the posterior probability that all traits in the cluster share a causal variant was 0.94 for the smaller cluster (HIRD day 1 and monocyte expression), and 0.98 for the larger cluster ($CD4^+$ T cell expression and IBD). Distinct candidate causal variants were proposed for each cluster: for the smaller cluster, rs7567997, an intronic variant 45 kbp downstream of the canonical *ADCY3* TSS; and for the larger cluster, rs713586, a variant 15 kbp upstream of the TSS. In both cases, the variant explained all of the posterior probability for the cluster, but as the analysis was restricted to the 1054 variants present in all datasets, there is ample chance the true causal variants were not included. When it comes to fine mapping, it would be more appropriate to perform it using the dataset with the densest genotyping in each cluster. Nevertheless, the two main clusters being distinct from one another, and from non-colocalising traits across many configurations, still supports the existence of distinct causal variants, even if they may be unobserved. For HIRD day 1 expression of *ADCY3*, the more relevant cell type appears to be monocytes, not a correlated cell type like $CD4^+$ T cells—and vice versa for IBD. The clustering was robust despite the data containing no stimulated monocyte subsets. This eQTL effect is readily observable at baseline, and appears to be more significant in naive classical than non-classical monocytes in the Schmiedel *et al.* [88] data (Fig. 3.17). No colocalisation with blood monocyte count was observed, so the reQTL does not appear to affect monocyte abundance in general. I believe a variant that affects ability to increase monocyte counts post-vaccination can also be ruled out, as in that case the effect of genotype on expression is entirely mediated through the effect of genotype on monocyte abundance, so having cell abundance scores as covariates in the regression should eliminate that effect. Thus I hypothesise that a plausible mechanism generating the day 1 reQTL signal in HIRD is an increase in abundance of (classical)

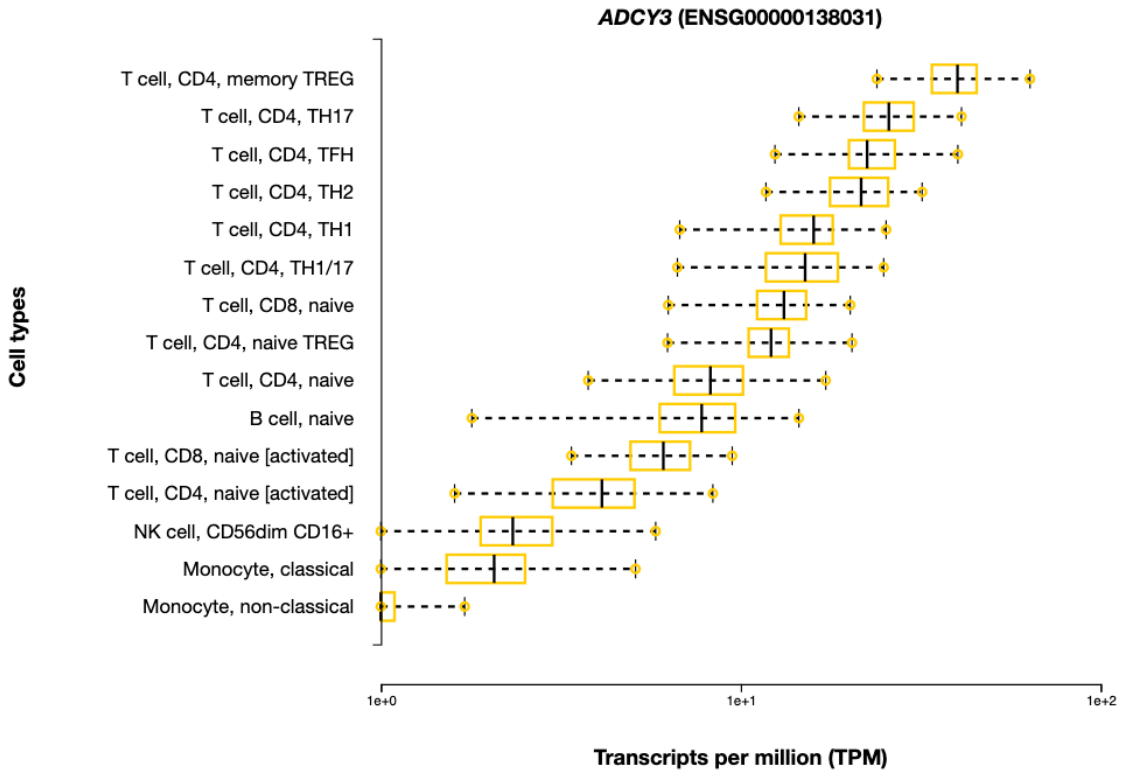


Figure 3.15: Expression of *ADCY3* in sorted immune cell subsets. Figure from Schmiedel *et al.* [88], *DICE (database of immune cell expression, expression quantitative trait loci, and epigenomics)*, https://dice-database.org/genes/ADCY3, accessed Nov 2020.

monocytes at day 1 post-vaccination, increasing the proportion of *ADCY3* transcripts in the bulk data originating from monocytes, thus making an eQTL specific to monocytes—not just stimulated monocytes—more readily detectable. This is the scenario where monocyte abundance modifies the effect of genotype on expression.

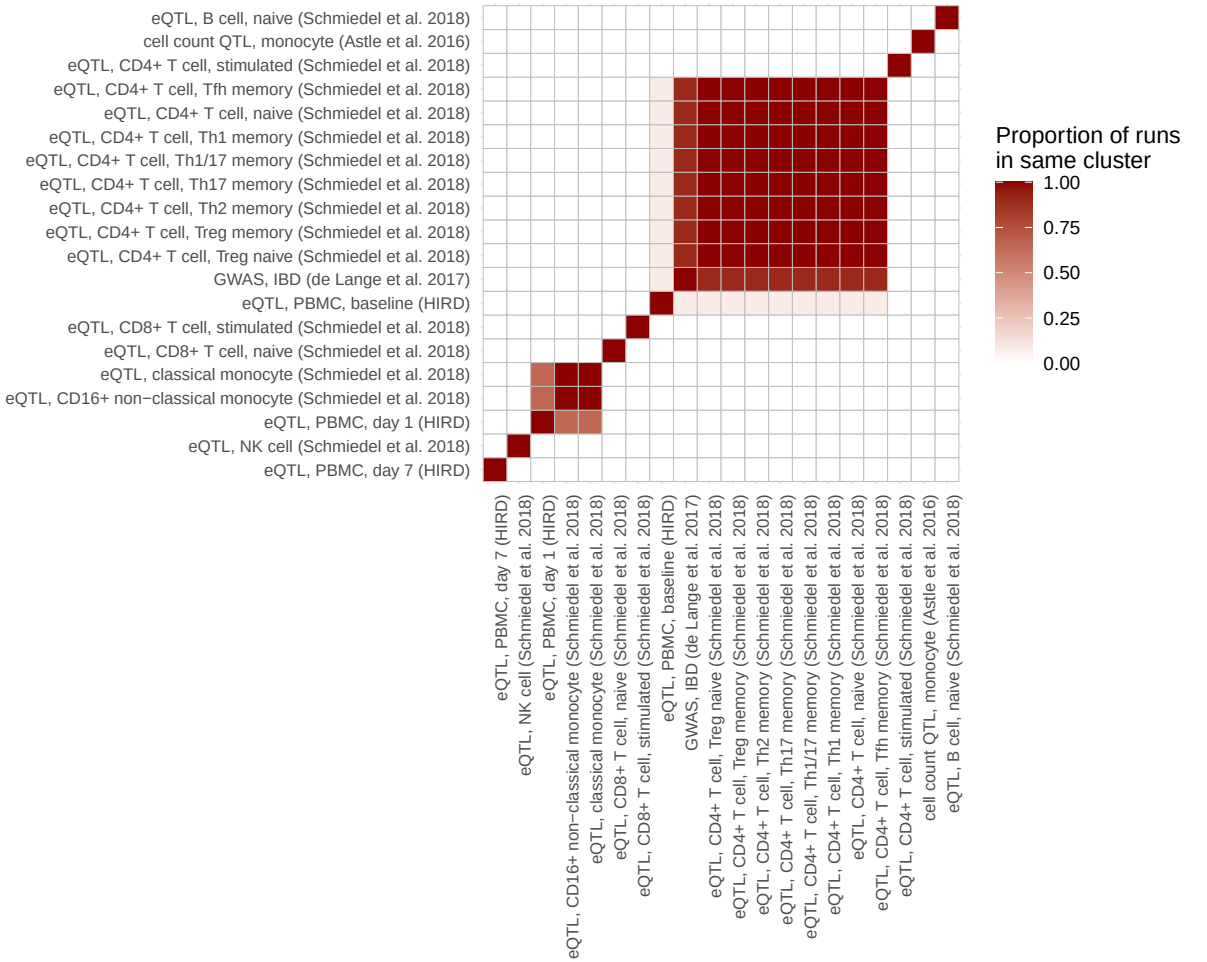


Figure 3.16: Sensitivity analysis for multi-trait colocalisation at the *ADCY3* locus. Colocalisation performed using HyPrColoc [316] in a ±500 kbp window around the lead variant for the day 1 *ADCY3* reQTL in HIRD, for trait datasets described in Section 3.2.12. Heatmap shows the proportion of configurations in which two traits colocalise in the same cluster over 100 configurations of algorithm parameters `reg.thresh`, `align.thresh`, and `prior.2` (range of values listed in Section 3.2.12). `prior.1` set at 1×10^{-4} (default). Rows and columns hierarchically-clustered.

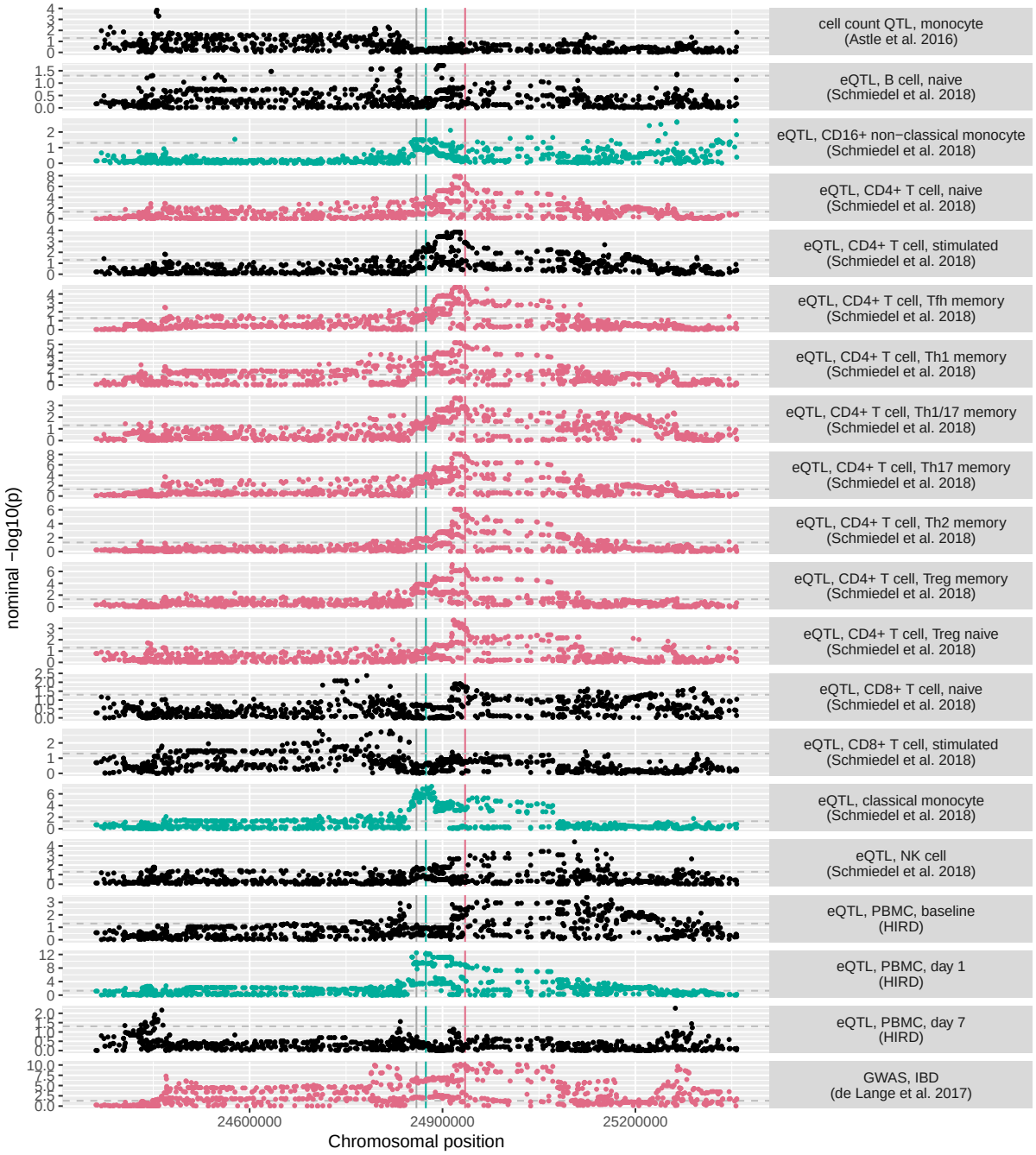


Figure 3.17: Multi-trait colocalisation at the *ADCY3* locus. Colocalisation performed using HyPrColoc [316] in a ±500 kbp window around the lead variant for the day 1 *ADCY3* reQTL in HIRD (vertical grey line). Traits are monocyte cell count (Astle *et al.* [303]), *ADCY3* expression in sorted immune cell subsets (Schmiedel *et al.* [88]), *ADCY3* expression at HIRD timepoints, and IBD (de Lange *et al.* [180]). Locus plots show summary statistics for 1054 variants present in all datasets. Traits in red and green represent two distinct clusters each hypothesised to be driven by a shared causal variant (vertical red and green lines). Non-colocalising traits are shown in black. Horizontal dashed lines show nominal $p = 0.05$. Default values for priors and algorithm parameters used, except `prior.2` $= 0.99$.

3.4 Discussion

Just as Pandemrix vaccination was found to induce extensive changes in the transcriptome in Chapter 2, it also induces changes in the regulatory architecture of gene expression. In a mega-analysis of array and RNA-seq datasets, *cis*-eQTL were detected for 50.8 % (6887/13 570) of genes in at least one timepoint, with the majority replicating in the much larger GTEx whole blood dataset. This is a substantial eGene rate given the modest per-timepoint sample size in HIRD, reflecting the gain in effective sample size from joint mapping over multiple conditions. Defining reQTLs by a significant difference in beta of the same eQTL between two timepoints, 1154/6887 (16.8 %) of lead eQTLs were classified as reQTLs. This is comparable to estimates of a 3–18 % reQTL rate between monocytes in different stimulation conditions by Kim-Hellmuth *et al.* [85], who also used a beta comparison method. The method is relatively stringent for calling reQTLs, avoiding both threshold effects where significant and non-significant eQTLs may have very similar betas, and discovery power biases caused by sample size differences between conditions. Indeed, had reQTLs been called by significance alone, 1427 reQTLs would have been detected with effects specific to baseline, the timepoint with the largest sample size in HIRD. There is growing consensus in the literature that most eQTLs are shared between conditions such as tissue and cell-type [74, 281, 321, 322], and that high estimates of >50 % condition-specificity based on significance thresholds (e.g. [276]) are overestimated. A counter-argument is that many studies overestimate sharing by calling condition-specific effects in LD as shared [322]. Here I compared the same gene-tag SNP pair across timepoints, but distinct causal, condition-specific variants may be tagged in such a way that the effect size of the tag SNP ends up similar—multi-trait colocalisation would be required to truly confirm a shared eQTL.

Gene set enrichment analyses to identify shared biological processes among target genes for reQTLs were generally uninformative. Genes targeted by reQTLs that explained more variation in expression post-vaccination were enriched for immune activation, with weaker enrichments related to APCs. This misses the full picture, as many of the strongest reQTLs were those with opposite sign effects at baseline and post-vaccination, but little change in PVE. Prevalence of opposite sign effects between pairs of conditions has been previously described in multi-tissue studies: in Fu *et al.* [76], the proportion of opposite sign effects among all reQTLs between five tissues was 4.4 %. In HIRD, I found an unexpectedly high proportion: 39/819 (4.8 %) for day 1 reQTLs, and 211/1002 (21.1 %) for day 7 reQTLs (Fig. 3.10). Given the global change in expression versus baseline was larger at day 1 than at day 7 (as described in Chapter 2), the larger number of strong reQTLs at day 7 was also unexpected. The genes with these opposite sign effects were not significantly enriched in any of the gene sets or BTMs I tested.

Post-vaccination DGE was considered as a mechanism that might generate reQTLs. As in Kim-Hellmuth *et al.* [85] and Davenport *et al.* [96], the overlap between differentially expressed genes and genes with reQTLs in HIRD was poor. Only at day 7 were genes with reQTLs more likely to be differentially expressed than genes without reQTLs—specifically after excluding genes without an eQTL from the analysis. The genes with both day 7 reQTLs and day 7 upregulation were enriched in cell cycle GO terms, but it is unclear how this may lead to generation of opposite sign effects, and the enrichment may largely have been driven by the DGE signal rather than the reQTL one. As described in Section 3.1.2, to define genes important to TIV response, Franco

et al. [94] made heavy use of the overlap of genes with DGE and reQTLs, followed by gene set enrichment. Unfortunately, their filtering before enrichment selected genes with either DGE *or* reQTL, making it difficult to assess which criteria contributed more to the significant enrichments they observed in the antigen presentation pathway. As noted by Davenport *et al.* [96] and Cuomo *et al.* [323], it may be that DGE and reQTLs are generated by different mechanisms, and focusing on the overlap is an unnecessarily narrow view.

An unappealing thought is that opposite sign effects are enriched in false positives, especially as they seem to show no positional enrichment near the TSS. While it is known that stimulation-specific reQTLs are more distal than baseline eQTLs [79], the HIRD reQTLs are evenly spread across the *cis* window. Some reQTLs may be statistical artifacts of the shrinkage of effects by `mashr`. Small and opposite effects generated by noise may be frequent enough for `mashr` to consider them a "pattern" of effects. This might explain the clear separation of the distribution of z-statistics for difference in beta between reQTLs and shared eQTLs. Conversely, it may be that small and opposite effects are more prevalent than expected, and combining `mashr` and a difference in betas test is the best framework for detecting them. To confirm either way, it may be necessary to repeat the reQTL calling without the influence of `mashr` shrinkage in a different modelling framework, such as one using a timepoint-genotype interaction term [96]. A complementary approach for validating these opposite sign reQTLs using the existing RNA-seq data might be within-individual allele-specific expression (ASE) (e.g. RASQUAL [324], PLASMA [325]). One would expect a true opposite sign reQTL effect to be recapitulated as opposite directions of allelic expression imbalance between timepoints. ASE may also provide more interpretable effect sizes than eQTL betas [326], for purposes such as clustering effect sizes to determine patterns of effects across timepoints [323].

At least one reQTL signal was plausibly not a false positive. The strongest reQTL detected at day 1 was for *ADCY3*, a membrane-bound enzyme that catalyses the conversion of ATP to the second messenger cAMP [327]. *ADCY3* is upregulated after the differentiation of monocytes—induced by beta-glucan—into macrophages in a state of trained immunity: a state in which they are more responsive to future immune stimuli [328]. GWASs have implicated the *ADCY3* locus in diseases such as obesity [327] and IBD [180]. *ADCY3* has also been identified as a post-stimulation reQTL in other studies involving stimulated blood immune cells: in PBMCs 24 h after *in vitro* infection with rhinovirus [83], *in vivo* in whole blood at day 1 after vaccination with seasonal TIV [94], and in whole blood after *in vitro* stimulation with *M. leprae* antigen for 26–32 h [92]. Given the diversity of stimulations and tissue types, the effect is likely a consequence of general immune activation, rather than a Pandemrix-specific response.

The strength of the *ADCY3* reQTL at day 1 was found to be modified by `xCell` estimates of monocyte abundance. The `xCell` scores are imperfect. Compared to FACS measurements in a cohort subset, the `xCell` scores were only weakly correlated, and the signatures used by `xCell` may be less accurate after vaccine stimulation. Fortunately, statistical colocalisation confirmed that the day 1 *ADCY3* reQTL signal is likely to be a monocyte-specific effect—and independent to the IBD signal in the locus, which colocalises with $CD4^+$ T cell eQTL datasets. The proportion of monocytes in the PBMC compartment increases at day 1, supported by both FACS [162] measurements and an increase in monocyte `xCell` score. Expression of *ADCY3* in HIRD is not

monocyte-specific: despite the increase in monocyte proportion, no upregulation was observed at day 1. Colocalisation was also not restricted to stimulated monocytes. The probable mechanism is an increased proportion of the bulk sample taken up by monocytes at day 1 providing more monocyte-derived *ADCY3* transcripts, rather than an upregulation-driven increase in detection power, or a vaccine-induced activation of the locus at day 1. Although multi-trait colocalisation proved to be the crucial piece of evidence suggesting the effect is not related to T cells, only 15 immune cell types were included in the analysis, so it is possible the reQTL is not entirely monocyte-specific.

Overall, cell type interactions were only detected at 16/1154 reQTLs. Although power to detect significant interactions is lower than power to detect main effects—not helped by the unclear reliability of `xCell` scores—it is still likely that mechanisms other than shifts in cell abundance underlie a large number of the detected reQTLs. One type of mechanism by which *cis*-eQTLs affect expression is through their impact on transcription factor (TF) binding affinity to motifs in promoters and enhancers [329]. Immune cells, including monocytes, are heavily regulated by cell type-specific TFs [330]. Cell type-specific expression of TFs has been proposed as a model for explaining magnifying, dampening, and opposite sign reQTL effects; for example, opposite sign effects could result from different TFs regulating the same gene via the same regulatory element, with activating effects in one cell type and suppressive effects in another [76]. There is evidence that TF activity is important for *in vivo* immune reQTLs: Çalışkan *et al.* [83] found rhinovirus reQTLs in PBMCs were enriched in ENCODE chromatin immunoprecipitation sequencing (ChIP-seq) peaks for the TFs *STAT1* and *STAT2*, and Davenport *et al.* [96] found interferon and anti-IL6 drug reQTLs likely disrupt *ISRE* and *IRF4* binding motifs. Rather than condition-specific expression of the eGene, reQTL effects could be generated by condition-specific expression of TFs whose activity is affected by the reQTL variant. A genomic feature enrichment for TF binding sites and other regulatory elements among (fine-mapped) HIRD reQTL variants could expose shared regulatory factors that explain subsets of the remaining reQTLs. This would also help evaluate if the even distribution of reQTLs across the *cis* window is a cause for concern.

Not only are the mechanisms at many detected reQTLs unknown, there may be many more reQTLs yet to detect in HIRD. Multiple independent eQTLs are present for a large fraction of eGenes [331]. As a single lead variant for reQTL assessment was chosen per gene to avoid reQTLs caused by differential tagging efficiency, I could not detect secondary reQTLs masked by a stronger shared eQTL for the same gene. This is not expected to be uncommon, as the effective sample size for shared eQTLs is usually large due to borrowing of information across conditions. Secondary eQTL signals tend to be weaker, more distal to the TSS, more likely to be enriched in enhancers rather than promoters, and—importantly—more context-specific [55, 85, 332, 333]. The proportion of genes with reQTLs I detect based only on the lead signal likely represents a lower bound. Stepwise conditional analyses at each lead variant will be required to uncover secondary associations, which then can be compared across timepoints in the same manner as the primary associations. These associations, although weaker on average, may actually have more variable effects between timepoints. I also did not consider *trans*-eQTLs due to sample size, which are more likely to be condition specific than *cis*-eQTLs [56, 59, 79].

Finally, I address the prospect that common genetic variation may explain some variation

in antibody response to Pandemrix. I have indirectly demonstrated genotype-dependent effects on expression response by identifying reQTLs with differing effect size between timepoints, but have yet to determine resulting genotype-dependent differences in antibody phenotypes. Some of the identified reQTLs will undoubtedly affect genes whose expression or post-vaccination expression change correlates with antibody response, but correlation is not transitive [334], so correlation of genotype with expression and expression with antibody response does not imply a correlation between genotype and antibody response. Formal tests such as the CIT [263] will be required to distinguish mediation of genotype-antibody response associations through gene expression from competing models. Franco *et al.* [94] realised this, but concluded that they had insufficient power for the CIT, with a greater sample size and comparable study design to HIRD. The HIRD cohort is also too small for a direct GWAS of Pandemrix antibody response. An approach for prioritising reQTLs that contribute to the antibody response to Pandemrix may need to leverage external genetic associations with similar phenotypes found in larger cohorts; for example, colocalisation with existing GWAS summary statistics for antibody response to other vaccines (ideally adjuvanted and inactivated vaccines). However, due to the number of possible generating mechanisms for bulk reQTLs *in vivo*, careful interpretation will be required to glean any insight into the biology of the stimulation in question. Chapter 5 will continue the discussion on the methodologies, experimental designs, and upcoming technologies required to complement the *in vivo* reQTL study design.