Chapter 2

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 thank Adrian Hayday, for kindly extending the opportunity to collaborate on the HIRD cohort; Efstathios Theodoridis, for performing the RNA and DNA extractions; Sean O'Farrell and Anna Lorenc, for providing the HIRD clinical, FACS, and antibody titre data, and for providing advice on the data formats; and the Wellcome Sanger Institute Sample Management and Pipelines teams, for performing the RNA-seq library preparation and sequencing, and the array genotyping.

2.1 Introduction

2.1.1 Influenza

Influenza is an infectious respiratory disease caused by the influenza virus family (*Orthomyxoviridae*) in a variety of vertebrate hosts. Of the four virus types (A, B, C, D) defined by antigenic specificity of the viral nucleoprotein, human infections are primarily caused by influenza A and influenza B. Each year, seasonal epidemics result in ∼1 billion infections and 300 000–500 000 deaths worldwide. Peak seasonality is defined by low humidity, low temperature, and other climate factors. Risk factors for severe illness and death include extremes of age (infants <1 yr, elderly >65 yr), pregnancy, obesity, chronic illness, and host genetics (e.g. mutations in *IFITM3* and *IRF7*) [114, 115].

Influenza viruses are enveloped viruses with a negative-sense single-stranded RNA genome divided into segments (eight segments in influenza A and B), each encoding one or more viral proteins. Two glycoproteins occurring on the surface of the viral envelope are the main antigens targeted by the host immune system. Haemagglutinin (HA), with its characteristic head-stalk structure, facilitates viral entry by binding sialic acid-containing surface receptors on host cells. Neuraminidase (NA) facilitates viral release, cleaving sialic acids to prevent newly-synthesised viruses aggregating to each other—viral proteins can be sialylated post-translation—and to the dying host cell in the final stages of the viral life cycle. The gradual accumulation of mutations

in these surface protein genes is known as antigenic drift, and can lead to evasion of antibody-mediated immunity acquired during previous exposures. As the virus type with the greatest prevalence, host range, and genetic diversity, influenza A is classified into a number of subtypes based on the antigenic properties of its HA and NA. At least 18 HA subtypes and 11 NA subtypes exist [116]. Although these HA and NA subtypes are all antigenically-dissimilar, there can still be cross-reactivity between subtypes, and considerable antigenic drift within subtypes [117]. Influenza B viruses are less diverse, classified into two antigenically-distinct lineages: Victoria-like and Yamagata-like [114].

On occasion, reassortment of genome segments between viruses infecting the same cell can quickly generate new strains (antigenic shift). Antigenic shifts are associated with pandemics due to lack of pre-existing population immunity [114]. Pandemics have occurred four times in modern history: 1918 ("Spanish"), 1957 ("Asian"), 1968 ("Hong Kong"), and 2009 ("swine"). Each was caused by influenza A, involving either reassortment of human and animal strains or zoonotic transmission of animal strains [118]. For instance, the 2009 pandemic was due to an influenza A strain with HA subtype 1 and NA subtype 1 gene segments of swine origin [119]: A(H1N1)pdm09*. Pandemic strains tend to enter seasonal circulation post-outbreak, potentially replacing previously-circulating strains; A(H1N1)pdm09-like strains are now the predominant seasonal A(H1N1) strain [114].

2.1.2 Seasonal influenza vaccines

Vaccination is the primary method for prevention and control of influenza. Antigenic drift and decline of vaccine-induced immunity over time means annual vaccination is recommended. Seasonal vaccines are multivalent, usually formulated against three (trivalent) or four (quadrivalent) influenza strains anticipated to circulate in upcoming influenza seasons. The World Health Organization (WHO)-run Global Influenza Surveillance Response System (GISRS) makes recommendations on the most representative strains for the Northern and Southern hemispheres each year, about six months before the start of the respective seasons.

There are three classes of licensed vaccines against seasonal influenza: inactivated influenza vaccines (IIVs), live attenuated influenza vaccines (LAIVs), and recombinant HA vaccines [116, 121]. IIVs can be split virion, containing virions disrupted with detergent, or subunit, containing further purified HA protein. LAIVs contain low-virulence, cold-adapted viruses that replicate well only in the cool upper respiratory tract. Recombinant HA vaccines contain purified recombinant HA expressed in insect cell lines rather than relying on traditional viral propagation in embryonated chicken eggs; cell-based IIVs are also available. Cell-based vaccines are faster to manufacture in pandemic situations, not dependent on egg supply, and avoids egg-adaptation: mismatches between vaccine and circulating strains caused by adaptation to growth in eggs.

Licensed seasonal vaccines are effective and well-tolerated in healthy adults, but particular subclasses of vaccine are recommended for different demographics [122–125]. LAIVs are delivered via nasal spray and are more effective than IIVs at mitigating transmission. They are recommended for children—the major drivers of transmission due to high viral loads and prolonged shedding [114,

*The suffix "pdm09" distinguishes the 2009 pandemic strain from the circulating seasonal A(H1N1) strains at that time [120].

122]—but are contraindicated in young children <2 yr due to increased risk of wheezing, and also in immunocompromised individuals. Trials suggest LAIV has superior efficacy compared to IIVs in children. High-dose and adjuvanted IIV vaccines are recommended to enhance immunogenicity in the elderly. Cell-based and egg-free vaccines are suitable for people with egg allergies. No vaccines are licensed for use in infants <6 mo, but passive immunity can be conferred through vaccinating the mother.

Point estimates of seasonal vaccine efficacy range from 50–90 % in healthy adults in controlled trials. Real-world effectiveness can be as low as 10 %, depending greatly on vaccine class, choice of endpoint, the match between vaccine and circulating strains, and various host factors [115, 126]. In general, efficacy is comparable or better in children versus young adults, and lowest in the elderly due to immunosenescence. Females mount higher antibody responses than males to IIVs regardless of age, potentially mediated by sex steroid levels [115, 127]. Immune history has a major impact on vaccine response due to immune memory. Adults primed by past exposures to seasonal influenza strains have qualitatively different responses to unprimed adults or influenza-naive children. For example, influenza-naive children mount much higher serum antibody responses to seasonal LAIV than primed adults [123]; and antibody responses to IIV peak later in unprimed individuals, requiring two doses to generate optimal concentrations [122]. Immune history also affects response via antigenic seniority (a.k.a. immune imprinting), where the antibody response is biased towards recall against strains encountered in early childhood over generation of a *de novo* response. This is beneficial if strains with the same epitopes come back into circulation, and harmful against strains still similar enough to trigger immune memory, but with drifted epitopes [115, 128]. Finally, host genetic variation in cytokine genes, immunoglobulin genes, and the human leukocyte antigen (HLA) region are associated with antibody responses—reviewed in Section 3.1.1.

2.1.3 Quantifying immune response to influenza vaccines

The efficacy of IIVs is mostly mediated by induction of strain-specific anti-HA antibodies, although other antibodies (e.g. anti-NA) may also contribute in the case of non-purified vaccines. Antibody-secreting cells (ASCs) in peripheral blood peak around one week after vaccination, and serum antibodies peak around two to four weeks after vaccination. Antibody-mediated protection may last up to a year in healthy adults [122, 129]. The immunodominance of the HA head over the stalk means most anti-HA antibodies have epitopes in the head domain. Unsurprisingly, the resulting immune selection pressure concentrates antigenic drift in the head domain. The stalk domain is relatively conserved, hence anti-stalk antibodies are more likely to be broadly neutralising antibodies effective against multiple virus subtypes (heterosubtypic immunity) [130].

The haemagglutination inhibition (HAI) assay is an inexpensive method for quantifying serum anti-HA antibody concentrations. A serial dilution of serum is created and mixed with standardised concentrations of red blood cells (RBCs) and influenza virus. Without the presence of antibodies, the receptor site on the HA head binds to membrane-bound sialic acid on RBCs, agglutinating them into a lattice that appears as a cloudy red solution. Anti-HA antibodies inhibit agglutination, allowing the RBCs to settle, creating a clear solution with a dark red pellet. The titre value comes from the most dilute concentration of serum that completely inhibits

agglutination [131]. The value is relative to the concentrations of reagents, requiring standardised protocols for comparability. A standardised HAI titre of 40 (1:40 dilution) is deemed seroprotective, and is an accepted correlate of protection for IIVs, representing 50 % clinical protection rate against infection [122, 132]. Reliable correlates of protection are useful in vaccine trials to reduce resource requirements (e.g. time, sample size, cost) compared to disease or infection-based endpoints like clinical protection [133]. For seasonal IIVs, regulatory agencies define target criteria based on the minimum proportion of individuals achieving HAI seroprotection ($\geq$ 40 titre) and seroconversion ($\geq$ 4-fold increase in titre after vaccination, indicating the vaccine is immunogenic) [116, 130, 132].

An alternative method is the microneutralisation (MN) assay, which quantifies concentrations of serum antibodies capable of neutralising viral infectivity. Neutralising antibodies may be anti-HA antibodies quantifiable by HAI, but may also be anti-HA stalk antibodies or antibodies with non-HA targets not detectable by HAI [116]. The assay again involves a serial dilution of serum, which is incubated with standardised concentrations of virus. The serum-virus mixtures are inoculated into host cells *in vitro*. After incubation, virus-infected cells are quantified (e.g. enzyme-linked immunosorbent assay (ELISA) using antibodies against viral proteins), the lack of which indicates neutralising activity sufficient to suppress viral replication [131]. A MN assay value of 160 (1:160 dilution) is considered equivalent to the seroprotective HAI value of 40 [122].

IIVs primarily induce serum antibodies of the IgG isotype. The cellular response has not been extensively studied, but the induction of $CD8^+$ T cells by unadjuvanted subunit IIVs is considered poor [122, 134]. In contrast, LAIVs can induce serum IgG, but also efficiently induce mucosal IgA and T cell responses [123]. Protection may also have greater duration than that afforded by IIVs, although the longevity still pales in comparison to natural infection, which can grant strain-specific protection that is lifelong [116, 121–123]. Different facets of response play different roles in immunity: serum IgG is important for limiting severity of systemic infection, mucosal IgA in the upper respiratory tract inhibits initial infection and transmission, $CD8^+$ T cells promote viral clearance and recovery, and $CD4^+$ T cells help induce the humoral and $CD8^+$ T cell responses [114, 122, 130, 135]. Correlates of protection for LAIV have not yet been defined; licensed LAIVs have all been licensed on the basis of clinical protection. Their comparable efficacy to IIVs in adults despite low HAI titres and seroconversion rates are presumed to be mediated by mucosal and cell-mediated immunity [116, 123]. Clearly, a broader view of immunity than granted by serological antibody assays is needed to understand the mechanisms leading to efficacious influenza vaccine responses.

2.1.4 Systems vaccinology of seasonal influenza vaccines

Vaccinology has historically been driven by the "isolate-inactivate-inject" paradigm [136]. Many vaccines have been developed and licensed through expensive, large-scale, and largely empirical trials that deliver highly effective vaccines, but little understanding of the immunological mechanisms of protection. In response, the last decade has seen the rise of systems vaccinology, a subfield of systems immunology dedicated to the analysis of high-throughput data measured at multiple levels of the immune system to characterise response to vaccination [133, 137–143]. Traditional serological assays (e.g. HAI, MN) are complemented with a raft of other technologies

to give a broader view of immune response [137–139, 142, 143]. Flow (e.g. fluorescence-activated cell sorting (FACS)) and mass cytometry (e.g. cytometry by time-of-flight (CyTOF)) are used to quantify immune cell subpopulations by their surface markers using fluorescent and heavy metal tags. These technologies can also be used to quantify intracellular markers (e.g. cytokines) by cell staining. Frequencies of cells secreting specific proteins (e.g. cytokines or antibodies) can also be quantified (e.g. enzyme-linked immune absorbent spot (ELISPOT)), useful for monitoring activated cell populations involved in both humoral and cell-mediated immunity. The transcriptome of peripheral blood is extremely popular to assay (e.g. expression array, RNA sequencing (RNA-seq)), providing an accessible, global measure of gene expression in dozens of immune cell subtypes without the need to select specific genes of interest in advance. Recently, there has been a growing interest in targeted sequencing of B cell and T cell repertoires, responsible for the specificity of the adaptive immune system. Serum proteins can be quantified in a low-throughput (e.g. ELISA) or multiplex manner (e.g. Luminex). Modern proteomics platforms also embrace a global philosophy, simultaneously quantifying thousands of proteins (e.g. SOMAscan). Finally, although not often considered due to small cohort sizes, host genetic variation can be accurately measured by genotyping arrays and sequencing.

Longitudinal study design is key, not only to profile different stages of innate and adaptive immunity, but also for determining correlates of protection. Correlates are known for some but not all established vaccines [144, 145]. For novel and emerging diseases, there may be no prior knowledge of correlates for use in vaccine development. The term "molecular signature" was coined to refer to transcriptomic responses induced early after vaccination that correlate with, and importantly, are predictive of later immune phenotypes (e.g. antibody titres) [133], although non-transcriptomic data types can also be used to form signatures. The ultimate goal is baseline prediction, where the immune state immediately prior to vaccination predicts response, and could potentially be modulated to enhance response in a similar manner to adjuvanting the vaccine itself [146].

Work in the field has thus far focused on established vaccines. One can learn from the success of highly-efficacious vaccines like yellow fever vaccine (YF-17D); where interferon, complement, and inflammasome expression signatures measured 3–7 days post-vaccination predict $CD8^+$ T cell and neutralising antibody responses 60 days post-vaccination [138, 147]. Much has also been learnt from the study of vaccines with suboptimal efficacy in challenging populations: infants and the elderly, pregnant women, immunocompromised patients, ethnically-diverse populations, developing countries [148]. The field has not yet identified completely novel correlates for many vaccines, partially because protection itself can be difficult to measure. One promising system is the human challenge trial, applied by Vahey *et al.* [149] to identify genes in the immunoproteasome pathway associated with protection from malaria challenge after adjuvanted RTS,S malaria vaccination. If correlates for novel vaccine candidates could be routinely established based on shared immune mechanisms leading to efficacy and long-lasting protection derived from multiple successful vaccines, there is enormous potential for optimising trials to be fast and cost-effective [133, 150], and informing rational, mechanism-based design for diseases that have thus far proved intractable to empirically-designed vaccines (e.g. HIV, malaria, non-childhood tuberculosis) [136, 140, 143, 150].

Seasonal influenza vaccines have been well-studied by systems approaches. One of the earliest studies by Zhu *et al.* [151] found that expression of type I interferon-modulated genes at day 7 was more prominent for LAIV than trivalent inactivated influenza vaccine (TIV) in children (total cohort size $n = 85$). A subsequent study found that both LAIV and TIV could induce interferon-related genes in children ($n = 37$), but much earlier in TIV (day 1) than in LAIV (day 7) [152]. As serum antibody titres are an established correlate for TIV, studies have been carried out to identify its molecular basis. Bucasas *et al.* [153] ($n = 119$) reported a 494-gene expression signature (including *STAT1*, *CD74*, and *E2F2*) measured at day 1 and 3 that correlated with serum antibody titres measured 14 and 28 days after vaccination. Signatures including day 3 kinase *CaMKIV* expression were found to predict day 28 HAI antibody titres in independent trials over three consecutive influenza seasons ($n = 67$) [154]. Expression of gene sets related to B cell proliferation at day 7 were likewise predictive of day 28 HAI ($n = 15$) [155]. Work has also been conducted to understand the heterogeneity in response due to host factors like sex [127] and age [156–159].

Signatures can be derived from predictors measured pre-vaccination [146]. A gene module enriched in apoptosis-related genes measured at baseline was found to predict day 28 HAI response ($n = 89$) [160]. Tsang *et al.* [161] used not gene expression, but FACS measurements to establish signatures for day 70 neutralising antibody titres ($n = 63$). Frequencies of several B cell, myeloid dendritic cell (DC), $CD4^+$ memory T cell, and a number of other activated T cell populations were not only predictive, but also stable over a period of two months. Nakaya *et al.* [157] used data collected over five consecutive seasons ($n = 212$) to identify associations between day 28 HAI and baseline expression modules annotated to B cells (positive association), T cells (positive association), and monocytes (negative association). They were able to replicate these associations using published data from Franco *et al.* [94] and Furman *et al.* [160]. Another multi-cohort, multi-season study ($n > 500$) by the Human Immunology Project Consortium (HIPC) [159] found baseline expression of genes (*RAB24*, *GRB2*, *DPP3*, *ACTB*, *MVP*, *DPP7*, *ARPC4*, *PLEKHB2*, *ARRB1*) and gene modules to be associated with antibody response in young individuals. Again, the authors were able to validate the associations in an independent cohort.

To conclude, it must be noted that the utility of molecular signature for predicting response to influenza or other vaccines in clinical trials has not yet been validated, and it is difficult to draw causal insights from studies that are largely descriptive or predictive. The existence of temporally-stable and replicable signatures is, however, encouraging.

2.1.5 The Human Immune Response Dynamics (HIRD) cohort

For studies of seasonal influenza vaccines in adults, responses are heavily influenced by immunological memory built by past vaccination or infection with antigenically-similar strains [137, 162]. Dependence on exposure is reflected in high variability of baseline vaccine-specific antibody titres and memory B cell numbers [161]. There have also been few systems vaccinology studies of adjuvanted influenza vaccines, known to have greater immunogenicity and efficacy than non-adjuvanted vaccines in children and the elderly [158, 163, 164]. The Human Immune Response Dynamics (HIRD) study conducted by Sobolev *et al.* [162] was conceived as a unique

opportunity to study response to an adjuvanted pandemic influenza vaccine (Pandemrix), where responses are more likely to be primary than recall, and variability due to prior exposure is minimised.

Pandemrix was one of several vaccines rapidly developed and licensed in response to the 2009 influenza pandemic [165]. It is a monovalent split-virion IIV against the pandemic influenza A/California/07/2009 (H1N1)pdm09 strain* developed by GlaxoSmithKline, containing 3.75 µg HA and adjuvant AS03 (oil-in-water emulsion containing DL-α-tocopherol, squalene, polysorbate 80). Subsequent studies estimated its effectiveness to be ∼80 % after a single dose [166]. As the H1N1 subtype had not circulated since the 1918 pandemic, the majority of the population was expected to be immunologically-naive at the time of study sampling (March 2010 to August 2011).

The study was a longitudinal, prospective cohort study. A total of 178 healthy adults in the UK were vaccinated with a single dose of Pandemrix. Clinical, transcriptomic, immune cell frequency, cytokine level, antibody titre, and adverse event phenotypes were collected. Genes associated with both myeloid and lymphoid effector functions had increased expression at day 1 versus baseline, most prominently for genes associated with the interferon response. Day 1 gene expression was impacted by age; significant global differences were observed in individuals older than 30–40 yr, considerably earlier than usually considered in studies of immunosenescence in old age. The early myeloid responses—increase in blood monocyte levels and cytokines associated with innate activation e.g. CCL4—were overall consistent with studies of unadjuvanted seasonal influenza vaccines. However, the early lymphoid responses—driven by a five-fold increase in serum interferon gamma levels at day 1—were unique to this adjuvanted pandemic influenza vaccine.

Vaccine (antibody) response was defined as a ≥4-fold increase in either HAI or MN titres after vaccination. Genes related to plasma cell development and antibody production were more highly expressed in 23 responders compared to 18 non-responders at day 7 post-vaccination. However, due to high variability among the vaccine non-responders in expression trajectory over time, a predictive model that segregated the two groups could not be built, even considering other predictors such as frequencies of immune cell subsets, and serum cytokine levels. There appeared to be many "routes to failure" [162], rather than any single determining factor leading to poor antibody response.

2.1.6 Chapter summary

Transcriptomic measurements in the original HIRD study were restricted to a relatively small number of individuals ($n = 46$), limiting power to detect expression associated with antibody response. In addition, the binary responder versus non-responder definition used does not account for variation in baseline titres, and dichotomisation of a continuous variable loses information and implies a discontinuity in response at the cutoff.

In this chapter, I combine the existing array data with newly generated RNA-seq data ($n = 75$)

*The WHO nomenclature for isolates specifies influenza type (A, B, C, D), host of origin (human if omitted), geographical origin, strain number, year of isolation, and isolate subtype (combination of HA and NA subtypes) [117].

on additional individuals from the HIRD cohort using Bayesian random-effects meta-analysis to account for between-platform heterogeneity. I also compute a baseline-adjusted, continuous phenotype of antibody response to vaccination, the titre response index (TRI) [153]. Leveraging the greater sample size, more statistically efficient definition of vaccine response, and greater sensitivity of rank-based gene set enrichment analysis over per-gene analysis, I identify gene expression modules associated with magnitude of antibody response. The strongest associations are seen at day 7, but significant module associations are also observed at baseline.

2.2 Methods

2.2.1 Existing HIRD data and additional data generation

The design of the HIRD study is described fully in Sobolev *et al.* [162]. In brief, blood samples were collected from each individual on each of six visits: two pre-vaccination (days -7 and 0), and four post-vaccination (days 1, 7, 14 and 63). A single Pandemrix dose was administered after blood sampling on day 0. Serum antibodies were measured for all individuals ($n = 178$) on days -7 and 63 using both HAI and MN assays. peripheral blood mononuclear cell (PBMC) gene expression was profiled for 46 individuals by expression array on days -7, 0, 1 and 7.

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

2.2.2 Computing baseline-adjusted measures of antibody response

There were 166/178 individuals with HAI and MN titres available at both baseline (day -7) and post-vaccination (day 63). Sobolev *et al.* [162] defined Pandemrix vaccine responders as individuals with ≥4-fold titre increases from day -7 to day 63 in either the HAI or MN assays. This is a typical threshold for HAI and MN seroconversion used to assess the immunogenicity of seasonal IIVs [116], and has also been recommended for pandemic IIVs [167]. However, they noted there was "a complete spectrum" of baseline titres of non-responders, citing "glass ceiling" non-responders whose high baseline titres made "enhancements by ≥4-fold harder to achieve". This may be referring to the dynamic range of the assays. In the full data, the range of HAI titres is 8–4096, and the range of MN titres is 10–5120 (Fig. 2.2a, Fig. 2.2b). In just the day -7 baseline titres, the range of HAI titres is 8–512, and the range of MN titres is 10–5120*. It is impossible for an individual with higher than 1280 MN at day -7 to achieve a 4-fold increase in MN after vaccination if the maximum MN value is 5120. This ceiling effect can been seen in Fig. 2.2d, where for a given baseline MN titre, there is a limit to the maximum observable fold change.

Another perspective is to consider that day 63/day -7 fold change is a change score on the log scale. It is well-known that change scores are usually negatively correlated to baseline. This can

*This indicates some individuals likely do have pre-existing antibodies to the pandemic strain (or cross-reactive antibodies), although the mean of the baseline titre distribution would still be expected to be higher if this were a seasonal vaccine.

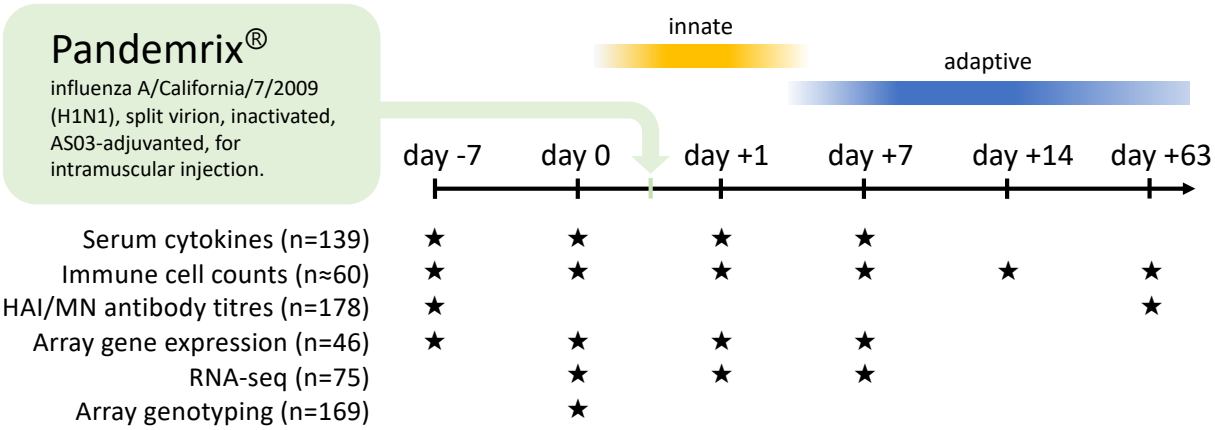

Figure 2.1: Overview of HIRD study data. The total cohort size was 178 individuals. Serum cytokines were quantified by 16-plex Luminex panel. Immune cell subsets were quantified by FACS. Serum antibodies were quantified by both HAI and MN assays. Array and RNA-seq gene expression were quantified in the PBMC compartment.

be due to individual-level regression to the mean*, the tendency for extreme observations to be followed by less extreme ones in the same individual [168], but is also due to the mathematical relationship between change score and baseline ("mathematical coupling" [170]). The correlation between change score and baseline is likely to be negative when the variance of the post-test score is much larger than the variance of the baseline and the correlation between baseline and post-test score is less than one [170, 171]. The negative correlation of titre fold change and baseline is visible in the HIRD data (Fig. 2.2c, Fig. 2.2d).

Additionally, dichotomisation of continuous variables can result in loss of information [172–175]. Cohen [172] presents a classic example where dichotomising a continuous independent variable reduces statistical power akin to throwing away a third of the samples—this being the optimal case when the cutpoint is the mean. A discontinuous cutpoint is also biologically implausible, implying that a 4.01-fold antibody titre change would be dramatically more protective than a 3.99-fold change.

To address these concerns, I computed the TRI as defined in Bucasas *et al.* [153]. For each assay, a linear regression was fit with the $\log_2$ day 63/day -7 titre fold change as the response, and the $\log_2$ day -7 baseline titre as the predictor. The residuals from the two regressions were each standardised to zero mean and unit variance, then averaged with equal weight. The TRI is a single variable that expresses a continuous measure of change in antibody titres averaged across both assays post-vaccination, compared to individuals with a similar baseline titre. It is no longer correlated with baseline (Fig. 2.2e, Fig. 2.2f), and remains qualitatively comparable to the original binary definition (Fig. 2.2g, Fig. 2.2h).

Descriptive statistics for the 114 individuals with both gene expression and antibody titre data are presented in Table 2.1. Although the proportion of responders between array (32/44) and RNA-seq (59/70) individuals is similar ($p = 0.16$, Fisher's exact test), the variance of TRI in array individuals is higher ($p = 2.10 \times 10^{-4}$, Levene's test), suggesting more extreme antibody response phenotypes are present (Fig. 2.3). The cause of this is unknown—there is a possibility

Cf. group-level regression to the mean, which is prominent if the baseline measurement is used as a selection criteria for follow-up [168, 169].

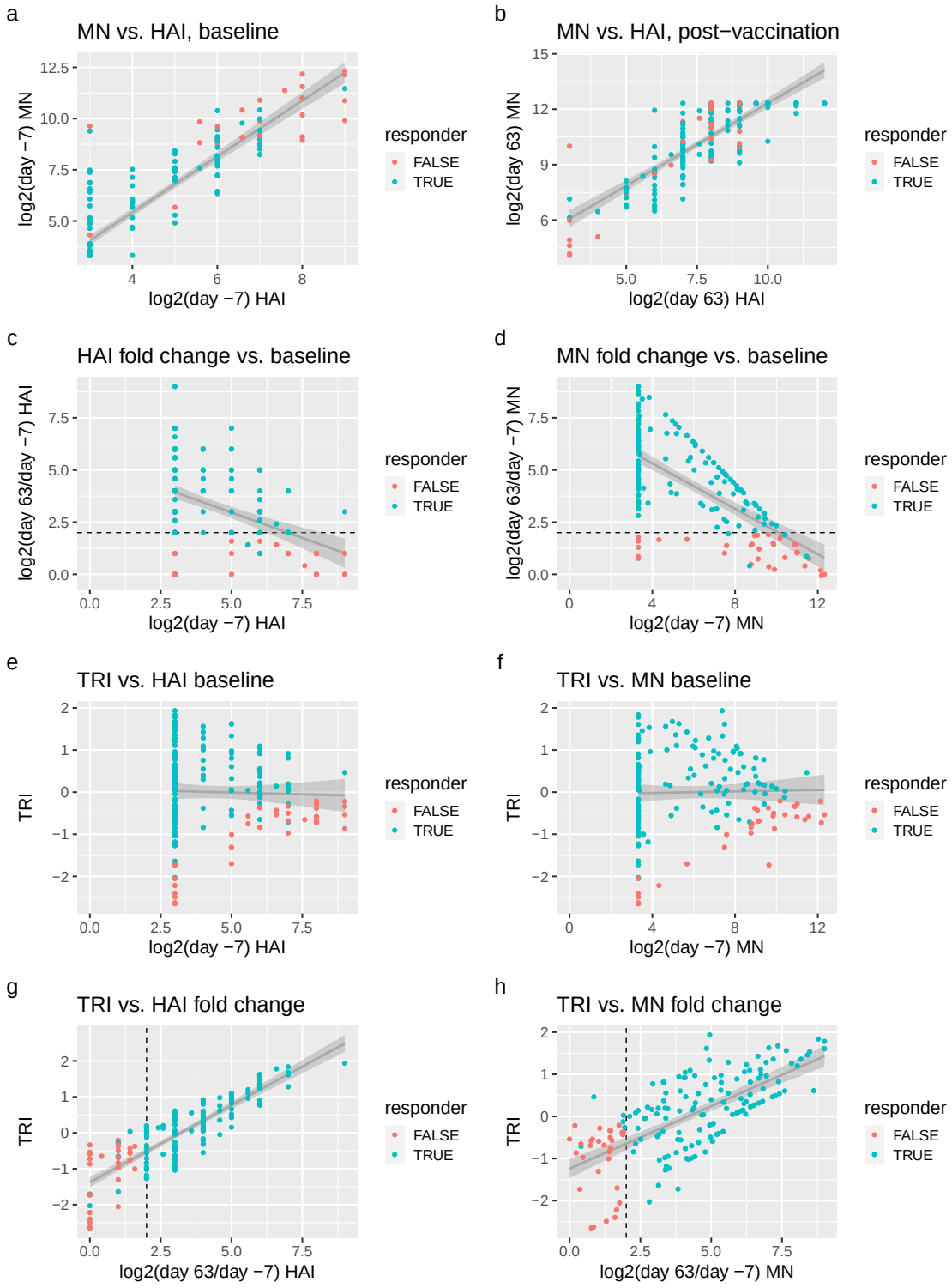

Figure 2.2: Antibody titre data and responder definitions. Titre values are on the $\log_2$ scale. Individuals are colored by binary responder status: ≥4-fold increase in either HAI or MN titres from baseline (day -7) to post-vaccination (day 63). Dashed lines show the ≥4-fold thresholds. **(a, b)** HAI and MN titres are correlated at baseline (**a**) and post-vaccination (**b**). **(c, d)** Baseline titres are negatively correlated to fold change. **(e, f)** TRI is computed from the standardised residuals from **c** and **d**, adjusting for baseline titre. **(g, h)** TRI remains comparable in ordering to binary response status.

that individuals with more extreme phenotypes were prioritised for array transcriptomics in the original HIRD study*.

2.2.3 Genotype data generation

DNA was extracted from frozen blood using the Blood and Tissue DNeasy kit (Qiagen), and genotyping was performed using on the Infinium CoreExome-24 BeadChip array (Illumina). In total, 192 samples from 176 individuals in the HIRD cohort—replicate samples were submitted for individuals where extracted DNA concentrations were initially low—were genotyped at 550 601 markers

2.2.4 Genotype data preprocessing

Using PLINK (v1.90b3w) [176], genotype data underwent the following quality control filters to remove poorly genotyped samples and markers:

- maximum marker missingness across samples <5 %;
- maximum sample missingness across markers <1 % (Fig. 2.4);
- sample heterozygosity rate within 3 standard deviations of the mean of all samples (threshold selected visually to exclude outliers, Fig. 2.4);
- sample sex mismatches based on X chromosome marker heterozygosity (`--check-sex` option);
- and marker deviation from Hardy-Weinberg equilibrium (HWE), an indication of genotyping or genotype calling errors [177–179] (`--hwe` option, p-value $<1 \times 10^{-5}$)†.

To exclude closely-related individuals and deduplicate samples from the same individual, pairwise kinship coefficients were computed using KING (v1.4) [181]. As rare variation is not generally required to determine relatedness, markers were filtered to minor allele frequency (MAF) >0.05 for computational efficiency. For each pair of samples with pairwise kinship coefficient >0.18 (first-degree relatives or closer), the sample with lower marker missingness was selected. After all filters, 169/176 samples and 549 414/550 601 markers remained.

2.2.5 Computing genotype principal components as covariates for ancestry

As shown in Table 2.1, the HIRD cohort is multi-ethnic. Large-scale population structure explains variation in gene expression [182, 183], so including genotype principal components (PCs) that reflect that structure as covariates can increase statistical efficiency for detecting associations with expression. I used HapMap 3 samples [184] as a reference population of unrelated individuals where the major axes of variation in genotypes are ancestry. Genotypes were first linkage

*Personal communication with Sobolev *et al.* [162] authors.

†A wide range of thresholds for the HWE marker filter in controls between 1.00×10^{-3} and 5.70×10^{-7} are reported in the literature [178]. The HWE threshold used here is from de Lange *et al.* [180]; since the HIRD cohort is two orders of magnitude smaller in size, this represents a relaxed threshold, so additional vigilance for genotyping errors downstream is required. In principle, it may be possible to select an appropriate threshold from the empirical distribution of HWE p-values [177].

Table 2.1: Descriptive statistics for HIRD individuals with both expression and antibody data. Values are count and percentage for categorical variables; mean and standard deviation for continuous variables.

		Platform	
	Total	Array	RNA-seq
	n = 114	n = 44	n = 70
Gender			
F	72 (63.2%)	27 (61.4%)	45 (64.3%)
M	42 (36.8%)	17 (38.6%)	25 (35.7%)
Age at vaccination (years)			
	29.2 (11.8)	32.9 (14.1)	26.8 (9.4)
Ancestry (self-reported)			
Asian	14 (12.3%)	5 (11.4%)	9 (12.9%)
Black/African	9 (7.9%)	4 (9.1%)	5 (7.1%)
Caucasian	82 (71.9%)	33 (75%)	49 (70%)
Latin American	2 (1.8%)	1 (2.3%)	1 (1.4%)
Mixed	5 (4.4%)	1 (2.3%)	4 (5.7%)
Other - Arab	1 (0.9%)	0 (0%)	1 (1.4%)
White Other	1 (0.9%)	0 (0%)	1 (1.4%)
log2 day -7 HAI			
	4.4 (1.8)	4.2 (1.6)	4.5 (1.9)
log2 day 63 HAI			
	7.6 (1.8)	7.4 (2.2)	7.6 (1.5)
log2 HAI fold change			
	3.2 (1.9)	3.2 (2.4)	3.1 (1.6)
log2 day -7 MN			
	6.2 (2.8)	5.4 (2.4)	6.6 (3.0)
log2 day 63 MN			
	10.4 (2.0)	9.5 (2.2)	10.9 (1.6)
log2 MN fold change			
	4.2 (2.3)	4.1 (2.6)	4.3 (2.1)
Responder (binary definition)			
FALSE	23 (20.2%)	12 (27.3%)	11 (15.7%)
TRUE	91 (79.8%)	32 (72.7%)	59 (84.3%)
TRI			
	-0.0 (0.9)	-0.2 (1.2)	0.1 (0.7)

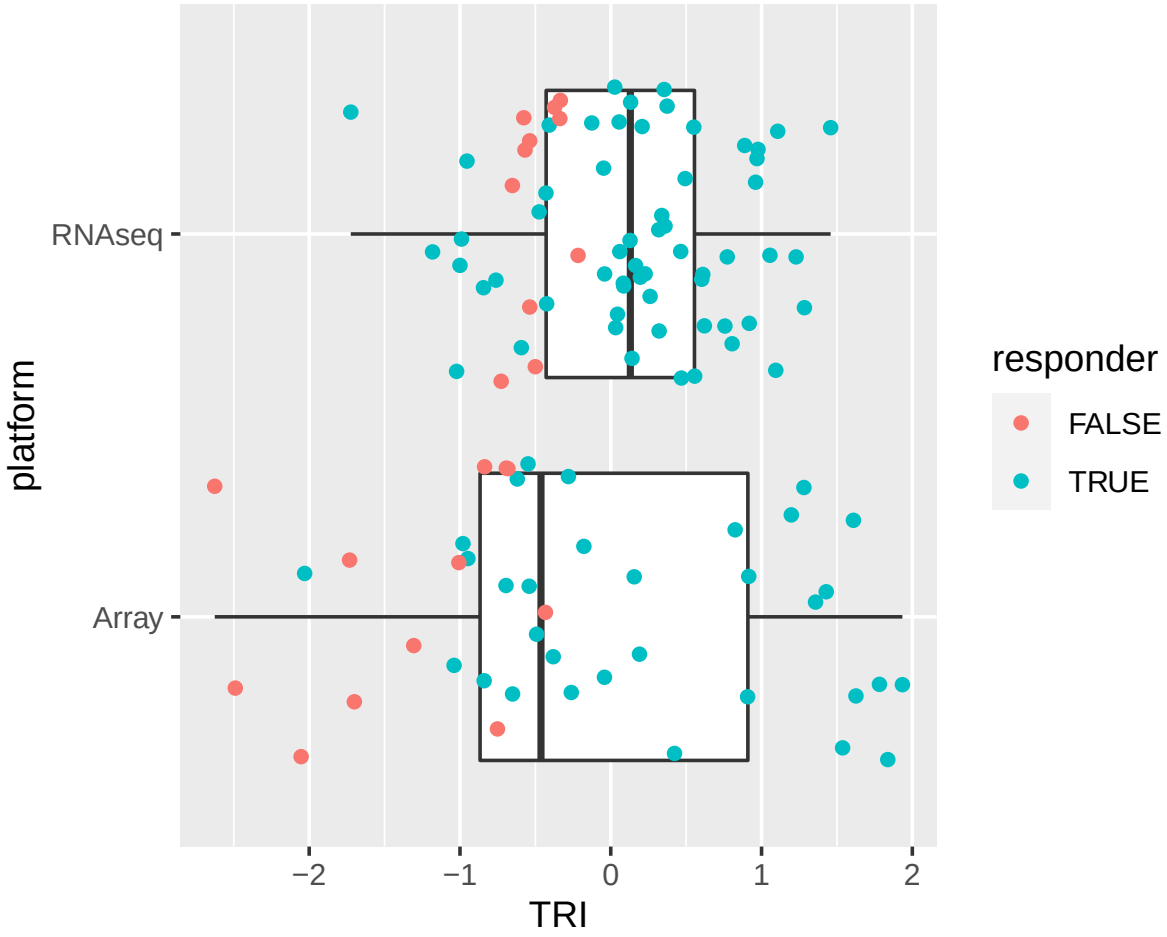

Figure 2.3: Distribution of patient TRIs, stratified by expression measurement platform. Points are colored by binary response status.

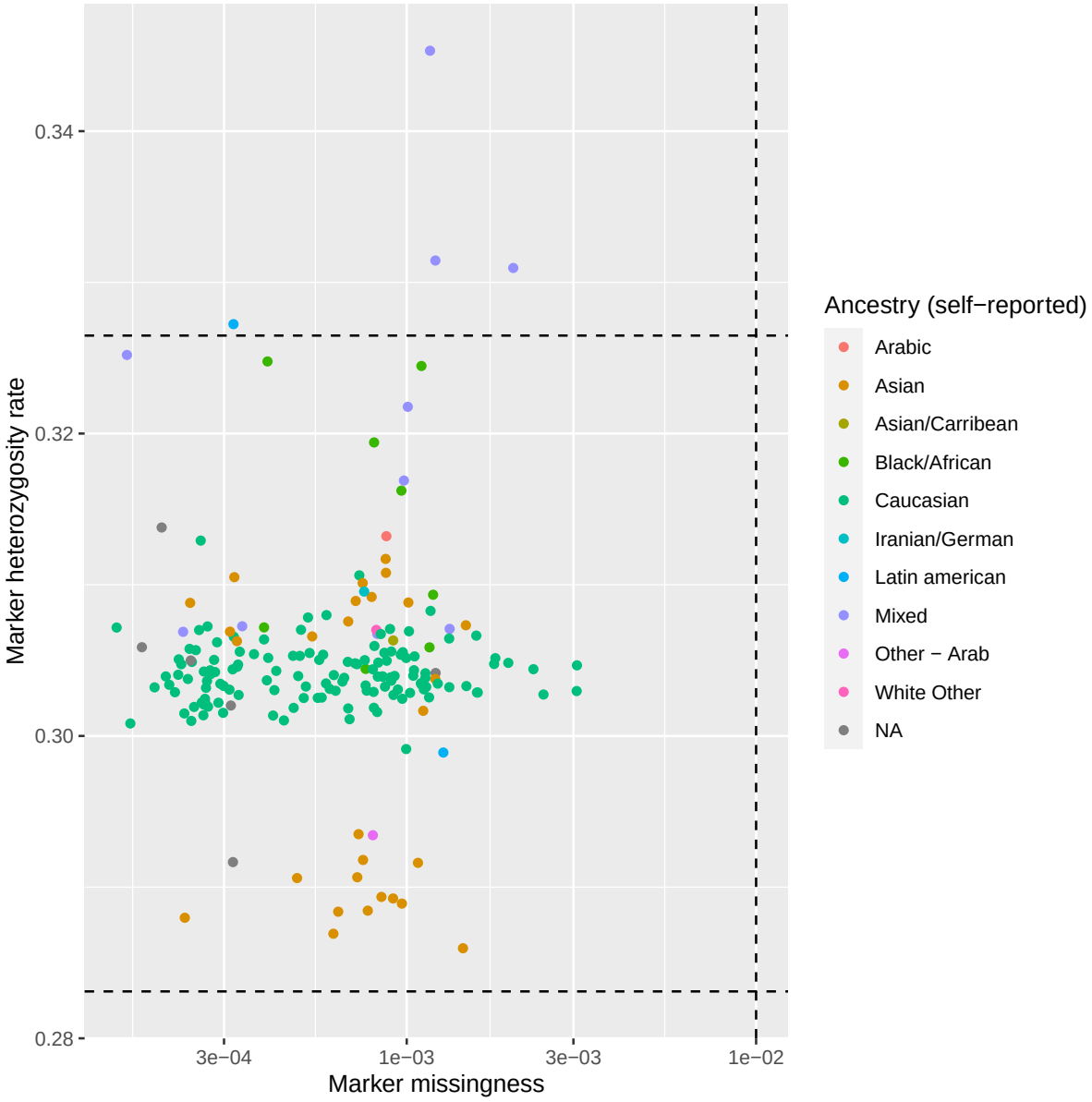

Figure 2.4: Sample filters for marker missingness and marker heterozygosity rate. Thresholds for missingness (1 %) and heterozygosity rate (mean ± 3 standard deviations) are shown by dashed lines.

disequilibrium (LD)-pruned (PLINK `--indep-pairwise 50 5 0.2` i.e. in a sliding window of 50 kbp, with a step size of 5 variants, remove variants at each step until no pair of variants has LD >0.2), to avoid regions with many redundant markers being overrepresented in the resulting PCs [185, 186]. Eighteen genomic regions with especially strong and/or long-range LD that contain many highly correlated markers were excluded, otherwise some PCs may reflect those just regions rather than genome-wide ancestry [185, 187]. Principal component analysis (PCA) was performed using smartpca (v8000) [185], then HIRD sample PCs were computed by projection onto the HapMap 3 PCA eigenvectors. A projection was used instead of in-sample PCA, as cryptic relatedness in HIRD may be reflected in the resulting PCs instead of ancestry [188]. For non-genotyped individuals with expression data, PC values were imputed as the mean value for all genotyped individuals with the same self-reported ancestry. The top PCs indeed separate HIRD samples by ancestry (Fig. 2.5). Significant PCs with large eigenvalues unlikely to be due to sampling noise were selected by Tracy-Widom test [189]. The fourth PC had an eigenvalue of 1.01 ($p = 0.02$), so the top four PCs were retained as continuous covariates for ancestry downstream.

2.2.6 RNA-seq data generation

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

RNA-seq quality metrics were assessed using FastQC* and Qualimap [190], then visualised with MultiQC [191]. Sequence quality was high, as measured by mean per-base Phred scores across sample reads (Fig. 2.6). The unimodal GC-content distribution suggested negligible levels of non-human contamination (Fig. 2.7).

2.2.7 RNA-seq quantification and preprocessing

Reads were quantified against the Ensembl reference transcriptome (GRCh38.p15) using Salmon [192] in quasi-mapping-based mode, which internally corrects for transcript length and GC composition by computing an effective length for each transcript. Relative transcript abundances were summarised to Ensembl (release 90) gene-level count estimates using `tximport` (`scaledTPM` method, which scales Salmon transcripts per million (TPM) values up to the library size [193, 194]) to improve statistical robustness and interpretability [193]. To combine technical replicates, as the sum of Poisson distributions remains Poisson-distributed, counts for technical replicates were summed for each sample. The mean number of mapped read pairs per sample after summing was 27.09 million read pairs (range 20.24–39.14 million), representing a mean mapping rate of 80.73 % (range 75.57–90.10 %). These meet sequencing depth recommendations for differential gene expression (DGE) experiments (e.g. diminishing returns after 10 million single-end reads

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

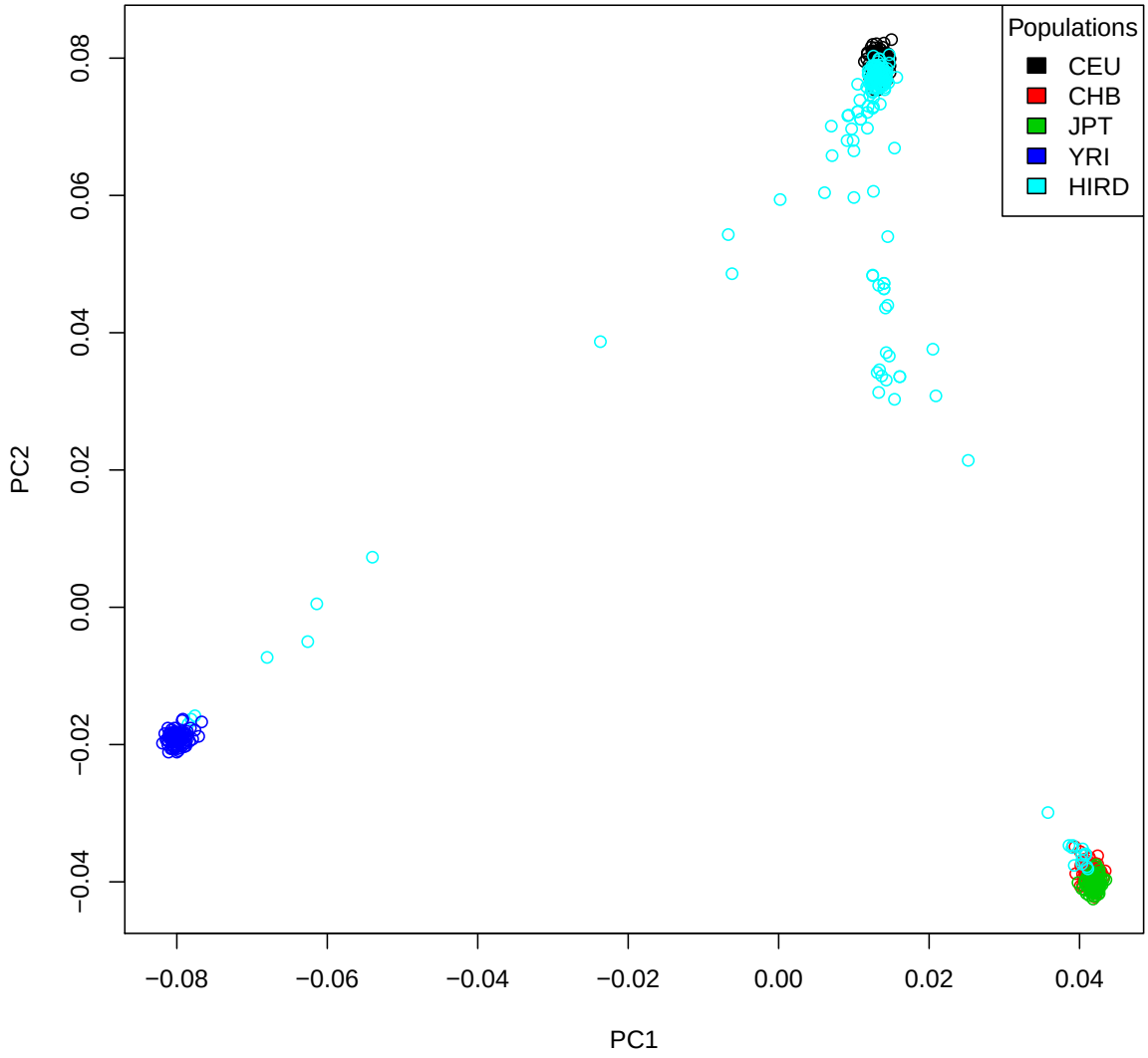

Figure 2.5: HIRD samples projected onto PC axes defined by PCA of HapMap 3 samples. The first two PCs separate individuals of European (CEU, upper-right) from Asian (CHB and JPT, lower-right) and African (YRI, lower-left).

Figure 2.6: FastQC per-base sequence quality (Phred scores) versus read position for RNA-seq samples. Visualised with MultiQC [191].

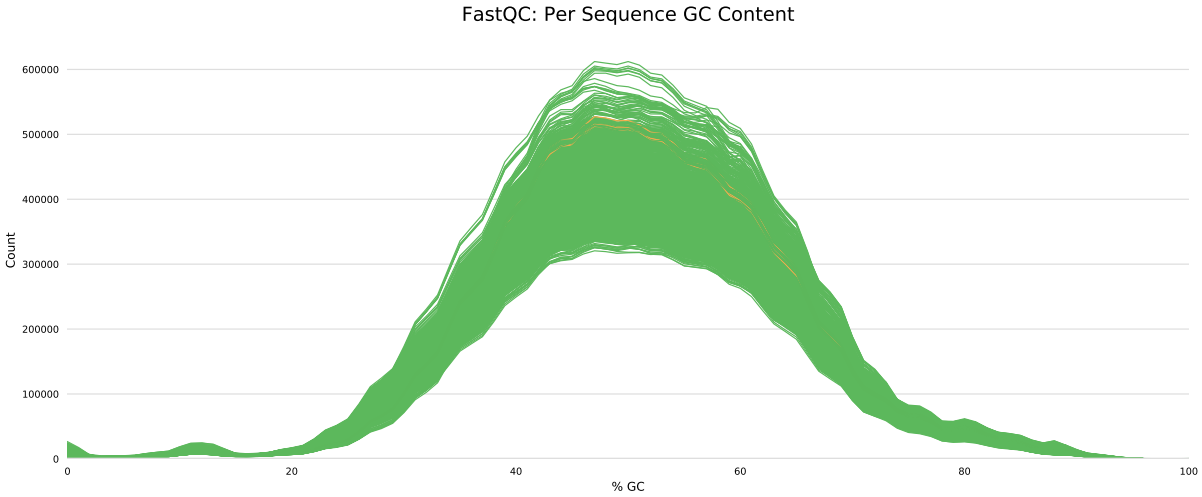

Figure 2.7: FastQC per-read GC distributions for RNA-seq samples. Visualised with MultiQC [191].

[195]) and mapping rate expectations (e.g. 70–90 % [196]).

Genes with short non-coding RNA (ncRNA) biotypes* were filtered out. These are generally not polyadenylated, depleted by size selection during library preparation, and shorter than the 75 bp read length, so expression estimates for these genes can reflect misassignment of counts from overlapping protein-coding or long ncRNA genes [197]. Globin genes, which are highly expressed in RBCs and reticulocytes—cell types expected to be depleted in PBMC [198]—were also filtered out. Given the proportion of removed counts at this stage was low for most samples (Fig. 2.8), poly(A) selection and PBMC isolation were deemed to have been efficient.

Many of the genes in the reference transcriptome were not detectably expressed in PBMC (Fig. 2.9), and many genes were expressed at counts too low for statistical analysis of DGE. Genes were filtered to require a minimum of 0.5 counts per million (CPM) in at least 20 % of samples. The 0.5 CPM threshold was chosen to correspond to approximately 10 counts in the smallest library, where 10–15 counts is a rule of thumb for considering a gene to be robustly expressed [199, 200]. Genes were further filtered to require detection (non-zero expression) in at least 95 % of samples to lessen the impact of low-expression outliers. The change in the distributions of gene expression among samples before and after filtering shows a substantial number of low expression genes are removed (Fig. 2.10).

RNA-seq produces compositional data due to sequencing a fixed number of reads per library; if one gene's expression goes up in a library, another's must go down. In order for expression values to be comparable between different libraries (samples), it is important to account for composition bias: the dependence of expression estimates on the expression properties of other genes in each library [201]. Effective library sizes were computed as between-sample normalisation factors using the trimmed mean of M-values (TMM) method [201, 202] from `edgeR::calcNormFactors` [203]. Precision weights for each (gene by sample) observation were computed with `limma::voom` [204] to account for the mean-variance relationship in RNA-seq data; `limma::voom` also transforms expression values to the $\log_2$ CPM scale using effective library sizes.

Finally, 15 samples were excluded for having missing HAI or MN data. After the application of all filters, expression values were available for 21 626 genes over 208 samples (70/75 individuals on day 0, 68/75 on day 1, and 70/75 on day 7).

2.2.8 Array data preprocessing

Single-channel Agilent 4x44K expression array data (G4112F, 60-mer oligonucleotide probes) for 173 samples from Sobolev *et al.* [162] were downloaded from ArrayExpress (`https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-2313/`). These arrays were originally processed in two batches, the effect of which can be seen in the raw foreground intensities (Fig. 2.11).

`VSN::normalizeVSN` [205] was used for simultaneous background correction, between-array normalisation (affine transformation, centers and scales each array to control for systematic experimental factors), and variance-stabilisation of intensity values (generalised logarithm, similar to $\log_2$ with better performance for small values), resulting in expression values on a $\log_2$ scale. As systematic experimental factors might differ between batches, requiring different centering

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

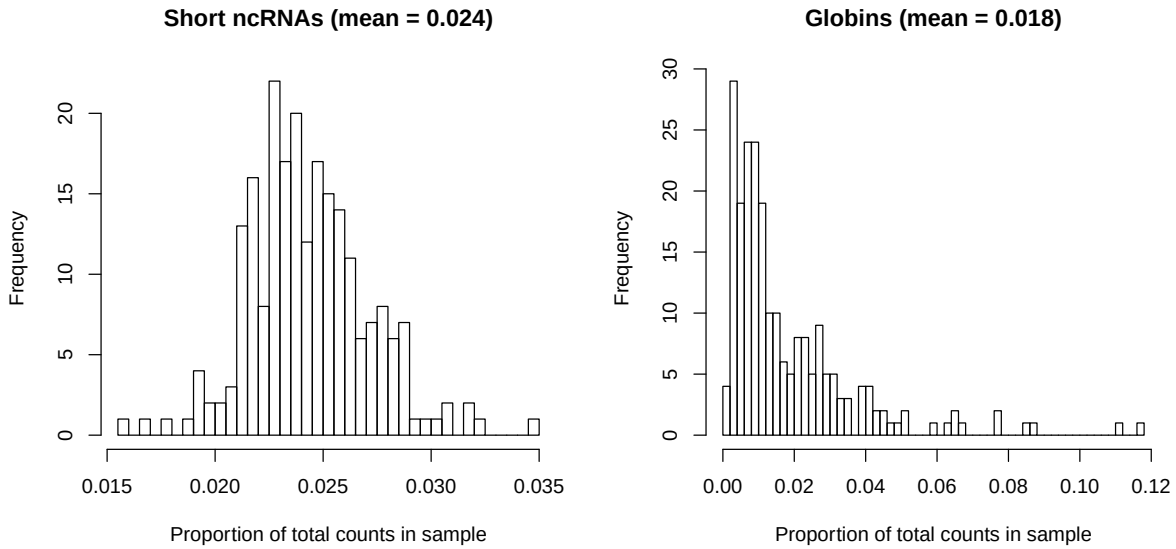

Figure 2.8: Distributions of removed short ncRNA and globin counts as a proportion of total counts in RNA-seq samples.

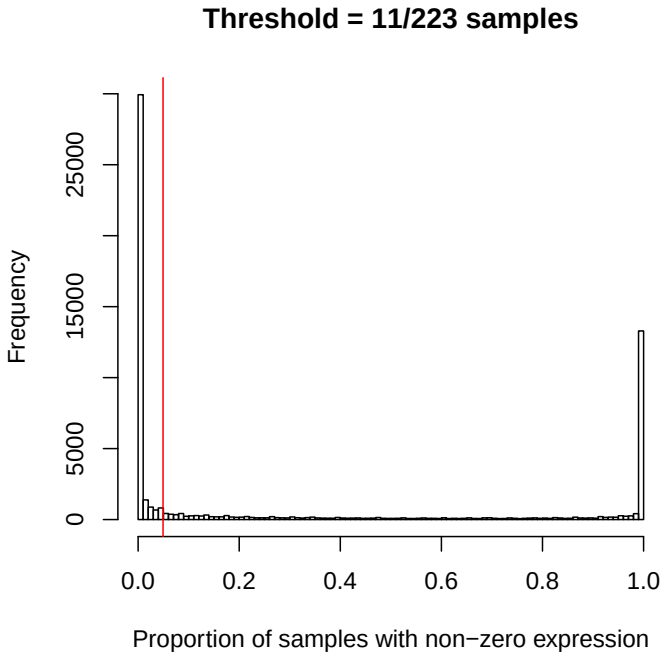

Figure 2.9: Distribution of the proportion of samples in which genes were detected (non-zero expression). Many genes are not detected in any samples (left-hand side). Vertical line shows 5 % threshold below which genes were discarded.

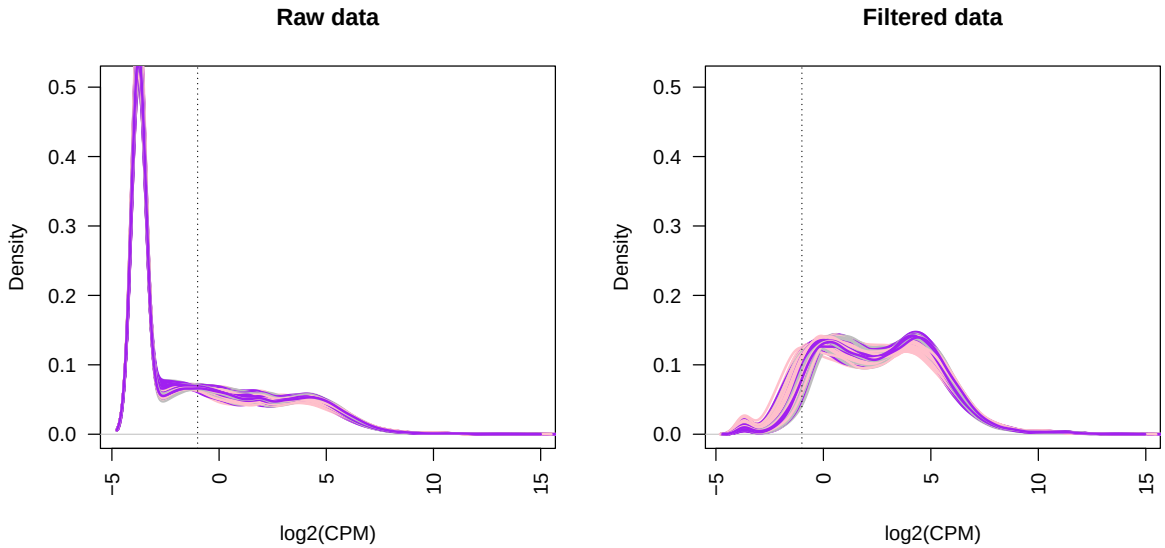

Figure 2.10: Distributions of gene expression for RNA-seq samples before and after filtering low expression and non-detected genes. Vertical line shows CPM = 0.5 threshold.

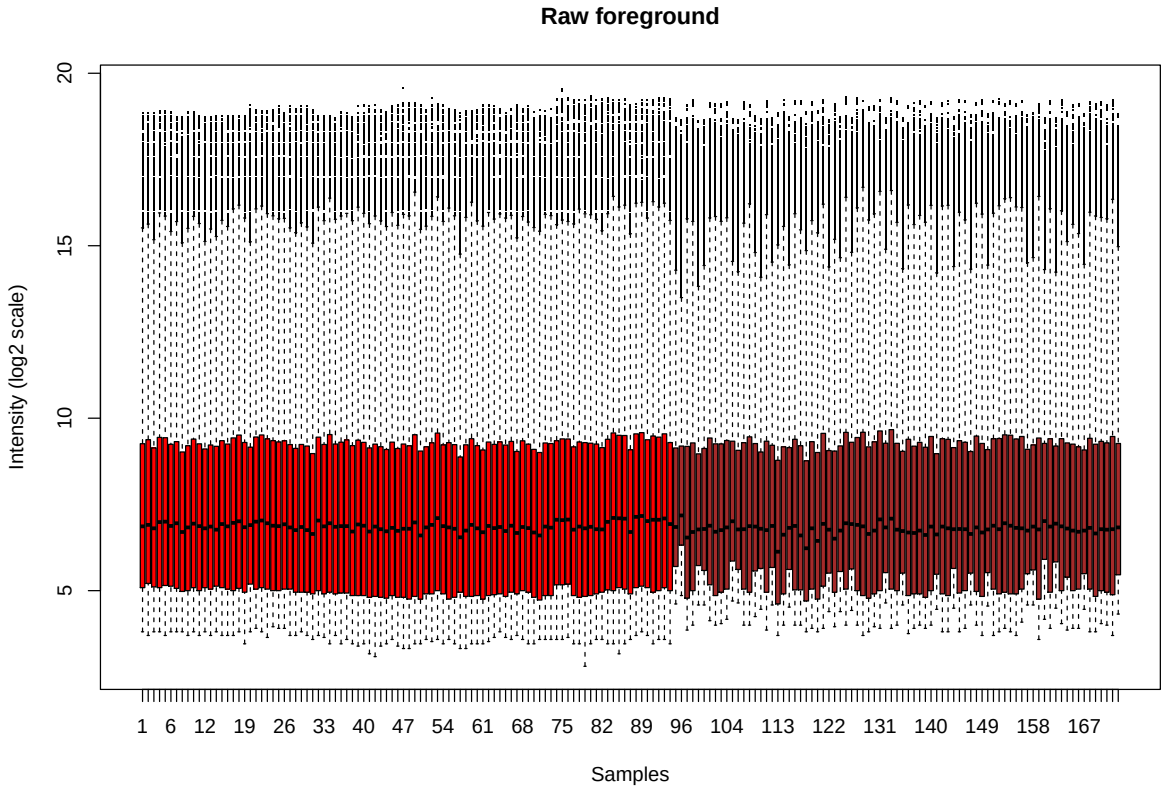

Figure 2.11: Distribution of raw foreground intensities for HIRD array samples ($n = 173$). Colored by array processing batch.

and scaling factors, normalisation was performed per-batch, then the two batches were merged.

Probes were matched to genes using `hgug4112a.db`*. Most genes were targeted by multiple array probes; 31 208 probes were collapsed into 18 216 Ensembl genes using by selecting the probe with the highest mean intensity for each gene (`WGCNA::collapseRows(method = "MaxMean")`, recommended for probe to gene collapsing by Miller *et al.* [206]). While it would be optimal to select a collapsing method to maximise the concordance between array and RNA-seq expression values, there were no samples assayed by both platforms in the HIRD dataset. The final normalised $\log_2$ intensity values for these 18 216 genes over 173 samples is shown in Fig. 2.12. Finally, `limma::arrayWeightsQuick` [207] was used to compute per-sample quality weights used to downweight unreliable arrays (samples) in the DGE analyses.

2.2.9 Differential gene expression (DGE)

2.2.9.1 Platform and batch effects

Combining the normalised array and RNA-seq data resulted in expression values for 13 593 genes assayed in both platforms for a total of 374 samples. PCA revealed that although samples separate by experimental timepoint along PC3 (Fig. 2.13e), measurement platform is by far the largest source of variation (Fig. 2.13a). Normalisation was also not able to completely remove the batch effect within the array data (Fig. 2.13a). The large platform effect likely stems from systematic technological differences in how each platform measures expression. RNA-seq has a higher dynamic range, resulting less bias at low expression levels, but estimates are more sensitive to changes in depth than array estimates are to changes in intensity [208]. Agreement between the two platforms is poor at extremes of expression [209, 210]. The preprocessing steps for the two platforms (Sections 2.2.7 and 2.2.8) were also vastly different.

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

Regarding the batch effect within the array data, a popular adjustment method is ComBat [211], which estimates per-gene, per-batch centering and scaling parameters, which are shrunk towards the per-batch mean parameters over all genes using empirical Bayes to improve robustness. ComBat was the method used by Sobolev *et al.* [162]. In comparisons of array batch effect adjustment methods, ComBat performed favourably (versus five other adjustment packages) [212] or comparably (versus fitting batch as a fixed or random effect in the linear model, which are centering-only corrections) [213]. However, where batches are unbalanced in terms of sample size [214] or distribution of study groups that have an impact on expression [215], ComBat can overcorrect batch differences or bias estimates of group differences respectively. In our data, sample size and timepoint groups are fairly balanced between the two array batches (Table 2.2). The proportion of responders is not, but response status does not have as prominent an impact on global expression as timepoint (Fig. 2.13). For the DGE analyses in this chapter, I chose to

*https://bioconductor.org/packages/release/data/annotation/html/hgug4112a.db.html

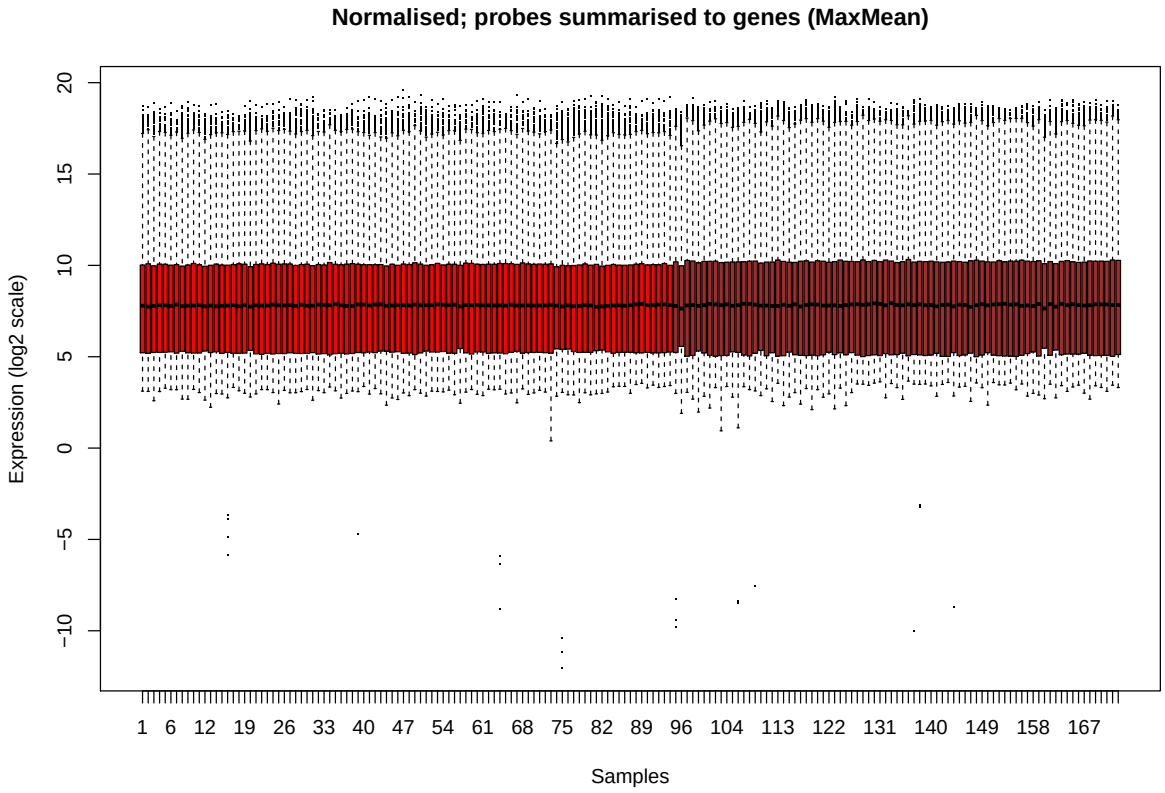

Figure 2.12: Distribution of per-sample expression estimates after normalisation and collapsing of probes to genes. Colored by array processing batch.

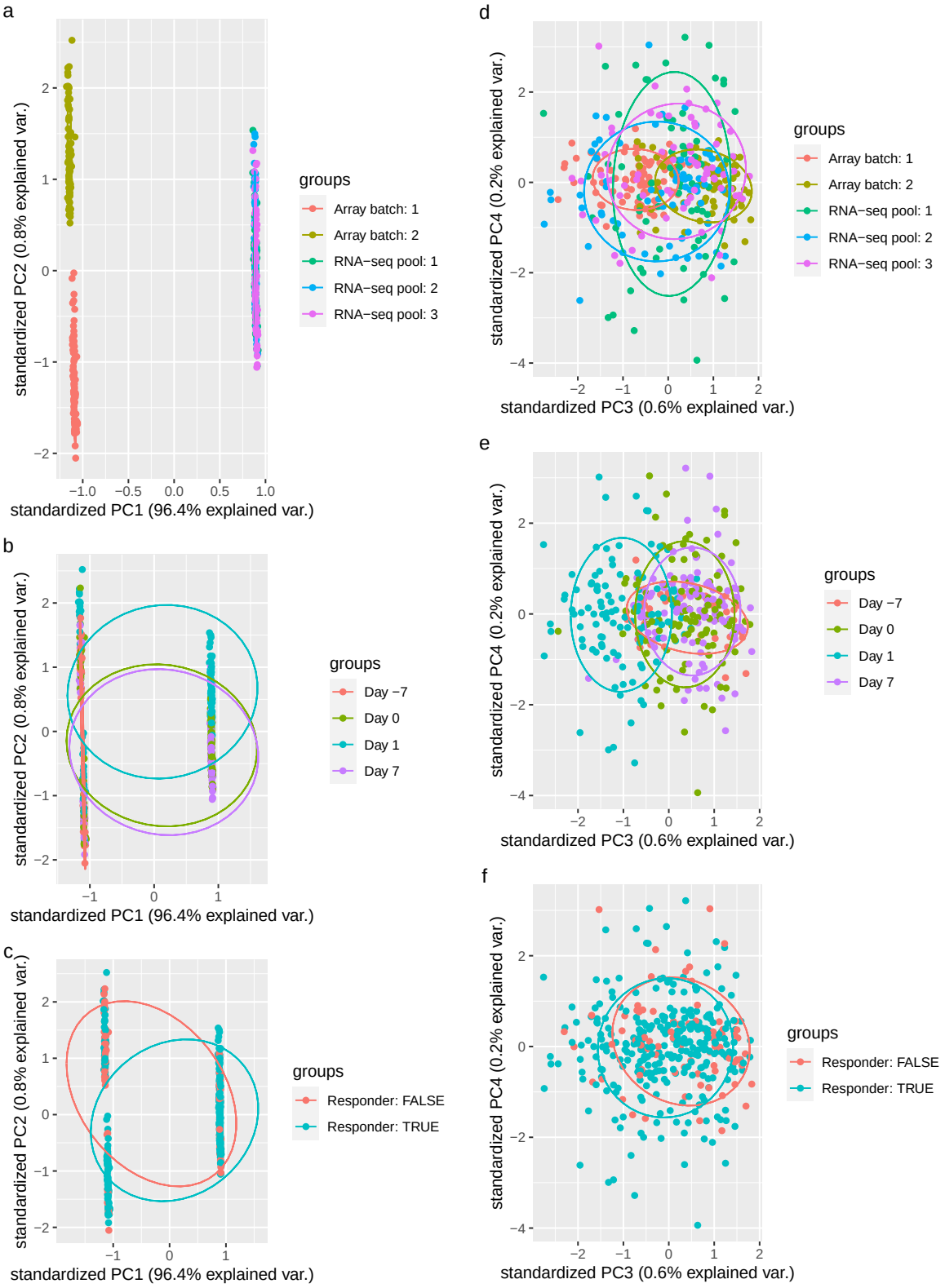

Figure 2.13: First four standardised PCs in the expression data, colored by array batch/RNA-seq pool (a, c), timepoint (b, d), and binary response status (e, f). Expression of each gene was standardised across samples within each platform before PCA.

model batches (array batch and RNA-seq pool) as fixed effects rather than pre-adjusting with ComBat in a separate step, ensuring the degrees of freedom (df) in the DGE model were correct. In practice, results from the analyses were not substantially affected by the choice of whether to use a ComBat pre-adjustment or a fixed effect.

2.2.9.2 Per-platform DGE model

As a meta-analysis was performed, DGE analyses were restricted to the 13 593 genes assayed by both the array and RNA-seq platforms. Linear models were fit using `limma` [216], which is computationally fast, performs well for sufficiently large ($n \geq 3$ per group) sample sizes [217], and internally considers the precision weights computed for RNA-seq observations in Section 2.2.7, and the array quality weights computed for array samples in Section 2.2.8. As Sobolev *et al.* [162] already found there was no global dissimilarity in array expression between day -7 and day 0, for the DGE analyses in this chapter, array day -7 and day 0 are treated as repeated measurements taken at a single "baseline" timepoint.

For each gene and platform, I fit a model (model 1) with expression as the response variable; with an intercept, timepoint (baseline, day 1, day 7), TRI, array batch/RNA-seq pool, sex, age, and the first 4 genotype PCs as fixed-effect predictors; and individual as a random-effect predictor. Within-individual correlations for the random effect were estimated using `limma::duplicateCorrelation`. A second model (model 2) was also fit, the only difference being two additional predictors for the multiplicative interactions between day 1 and day 7 with TRI. Model 1 was used for testing differences in expression between pairs of timepoints, and for testing association between TRI and expression with timepoints pooled. Model 2 was used for testing association between TRI and expression at specific timepoints.

Contrasts were defined, testing if linear combinations of estimated model coefficients are different from zero. From model 1, I defined contrasts for day 1 vs. baseline, day 7 vs. baseline, day 7 vs. day 1, TRI, sex, and age. For example, to test for association between TRI and expression, I used a contrast where the weight for the TRI coefficient was 1, with all other coefficient weights set to 0; to test for differences between day 7 vs. day 1, I used a contrast where the weight for the day 7 coefficient was 1, the weight for the day 1 coefficient was -1, and all other coefficient weights were 0. From model 2, I defined contrasts for the TRI, TRI-day 1, and TRI-day 7 interaction terms, which respectively test for association between TRI and expression at specifically at baseline, day 1, and day 7. Corresponding coefficients and standard errors for the contrasts were extracted from the `limma` models, which represent effect size in units of $\log_2$ expression fold change per unit change in predictor value.

2.2.9.3 Choice of DGE meta-analysis method

Two popular frameworks for effect size meta-analysis are fixed-effect and random-effects [218, 219]. The fixed-effect model assumes a single true effect size θ common to all studies. Given k studies ($i = 1, \ldots, k$), the observed effect size in the ith study is commonly assumed to be $y_i \sim \mathcal{N}(\theta, \sigma_i^2)$, where observed variation is explained only by within-study sampling error σ_i. In meta-analysis, the effects are combined with some weighting, commonly the inverse variance (precision) $1/\sigma_i^2$.

Table 2.2: Distribution of HIRD samples among timepoint and responder groups in the array batches and RNA-seq pools. Values are count and percentage for categorical variables; mean and standard deviation for continuous variables.

		Array batch/RNA-seq pool				
	Total n = 374	Array 1 n = 87	Array 2 n = 79	RNA-seq 1 n = 70	RNA-seq 2 n = 69	RNA-seq 3 n = 69
Day						
-7	40 (10.7%)	20 (23%)	20 (25.3%)	0 (0%)	0 (0%)	0 (0%)
0	114 (30.5%)	24 (27.6%)	20 (25.3%)	24 (34.3%)	23 (33.3%)	23 (33.3%)
1	109 (29.1%)	21 (24.1%)	20 (25.3%)	22 (31.4%)	23 (33.3%)	23 (33.3%)
7	111 (29.7%)	22 (25.3%)	19 (24.1%)	24 (34.3%)	23 (33.3%)	23 (33.3%)
Responder						
FALSE	80 (21.4%)	12 (13.8%)	36 (45.6%)	11 (15.7%)	9 (13%)	12 (17.4%)
TRUE	294 (78.6%)	75 (86.2%)	43 (54.4%)	59 (84.3%)	60 (87%)	57 (82.6%)
TRI						
	-0.1 (1.0)	-0.1 (1.0)	-0.4 (1.4)	0.1 (0.6)	-0.0 (0.8)	0.2 (0.6)

The random-effects model assumes a distribution of true effects centered around a common mean μ. Each of the k studies estimates its own study-specific true effect size θ_i. These are distributed around μ with variance τ^2 (standard deviation τ), representing an additional source of variation: the between-study heterogeneity. Then we have $y_i \sim \mathcal{N}(\theta_i, \sigma_i^2)$ for the first level of variation, $\theta_i \sim \mathcal{N}(\mu, \tau^2)$ for the second level of variation, and assuming these distributions, we have a normal-normal multilevel model [220]. Study weights include both within- and between-study variance $1/(\sigma_i^2 + \tau^2)$, reducing to the fixed-effect model when $\tau = 0$.

The choice of fixed or random effects depends on whether it is tenable to assume studies are identical enough that they all estimate a common effect*. In the HIRD data, there are $k = 2$ studies: array and RNA-seq. The between-platform differences described in Section 2.2.9.1 represent considerable sources of between-study heterogeneity. For DGE effect sizes, arrays also suffer from ratio compression of fold change estimates due to cross-hybridisation and probe saturation [210, 222, 223]. The assumption of $\tau = 0$ is unrealistic, so a random-effects model is more appropriate.

Unfortunately, there is no optimal solution for directly estimating τ in random-effects meta-analyses with small k [224], and especially in the case of $k = 2$ [225]. Many estimators are available [226], but lack of information with small k causes estimation to be imprecise, and often results in boundary values of $\tau = 0$ that are incompatible with the assumed positive heterogeneity [227, 228]. In such circumstances, the most sensible approach may be to incorporate prior information about hyperparameters μ and τ in a Bayesian random-effects framework [226–229]. For this study, I used the implementation in `bayesmeta` [220].

*A common misinterpretation is that random-effects meta-analysis assumes studies *themselves* are sampled from a population of studies. This is rarely appropriate since the design of new studies is influenced by existing studies [221]. The required assumption is exchangeability of study *effects*, which informally states effects are neither completely identical nor completely independent, but "similar" [221].

2.2.9.4 Prior for between-study heterogeneity

The choice of prior for between-study heterogeneity τ is influential when k is small [229]. Gelman [230] considers the case of $k = 3$, showing that a flat prior places too much weight on implausibly large estimates of τ, and recommends a weakly informative prior that acts to regularise the posterior distribution, constraining it away from implausible values. Since I assumed zero estimates for τ are unrealistic, I used a weakly informative gamma prior, as recommended by Chung *et al.* [227], which has zero density at $\tau = 0$ but increases gently as τ increases (a positive constant derivative at zero). This constrains τ to be positive, but still permits estimates close to zero if the data support it. This is in contrast to priors used in other studies from the log-normal (e.g [231, 232]) or inverse-gamma (e.g. [233]) families that have zero density at zero and derivatives of zero close to zero, ruling out small values of τ no matter what the data suggest; and in contrast to half-t family priors (e.g. [229, 230]), which have their mode at zero, and do not rule out $\tau = 0$.

Instead of constraining the value of τ for a gene's effect size to be a single unreliable estimate from $k = 2$ data points, assuming a prior distribution recognises that other genes may be informative of the range of plausible values for between-platform heterogeneity. To estimate the appropriate shape and scale parameters for the gamma empirically, a frequentist random-effects model using the restricted maximum likelihood (REML) estimator for τ (recommended for continuous effects [226]) was fit for each gene using `metafor::rma.uni` [234]. Depending on the contrast, over half of resulting per-gene τ estimates were boundary values of zero. Small estimates of $\tau < 0.01$ were excluded, and a gamma distribution fit to the remaining estimates using `fitdistrplus` [235].

2.2.9.5 Prior for effect size

While the choice of prior on τ is influential when k is small, there is usually enough data to estimate the effect size μ such that any reasonable non-informative prior can be used [228, 230]. `bayesmeta` implements both flat and normal priors for μ. Assuming that most genes are not differentially expressed with effect sizes distributed randomly around zero, I selected a normal prior with $N(\mu = 0, \sigma^2)$, over a flat prior. As in the section above, to determine an appropriate scale, a normal distribution with mean $\mu = 0$ was fit to the distribution of effect sizes from the per-gene frequentist models to empirically estimate σ.

Heavy-tailed Cauchy priors have been proposed for effect size distributions in DGE experiments to avoid over-shrinkage of true large effects in the tails [236]. Since `bayesmeta` does not implement a Cauchy prior, to avoid over-shrinkage, I flatten the normal prior considerably by scaling up the standard deviation by a factor of 10: $N(0, (10\sigma)^2)$. This places a 95 % prior probability that effects are less extreme than approximately 20 times the observed σ, sufficient to allow for extreme fold-changes.

2.2.9.6 Example of priors

An example of the empirically estimated hyperparameters for the priors for the day 1 vs. baseline contrast are shown in Fig. 2.14 (for τ) and Fig. 2.15 (for μ). For τ, the final prior used was Gamma(shape $= 1.57$, scale $= 0.06$). This is comparable to the default recommendation from

Chung *et al.* [227] of a Gamma(shape $= 2, \text{scale} = \lambda$) prior where λ is small. For μ, the final prior used was $N(0, (0.324 \times 10)^2)$. The tails of the non-scaled normal fit (black) are light compared to the Cauchy fit (red), which may lead to over-shrinkage, especially since there are many genes with high positive fold changes for the day 1 vs. baseline effect.

2.2.9.7 Multiple testing correction

For per-platform DGE, false discovery rate (FDR) was controlled with `limma::decideTests` using the Benjamini-Hochberg (BH) procedure. For the frequentist random-effects meta-analysis, nominal per-gene p-values were converted to FDR estimates using `p.adjust(method = "BH")` in R. For the Bayesian random-effects meta-analysis, the effect sizes and standard errors from the per-gene meta-analysis output from `bayesmeta` were supplied to `ashr` [238], which models the distribution of effects under the assumption of unimodality. `ashr` applies empirical Bayes shrinkage to improve the accuracy of effect estimation (e.g. against winner's curse), returning posterior effect sizes, posterior standard errors, and their significance (local false sign rate (LFSR)). LFSR is analogous to FDR, but quantifies the probability, given the data, of calling the wrong sign for an effect, rather than the confidence of a non-zero effect [238]. Unless otherwise stated, FDR and LFSR were controlled at the 5 % level separately for each contrast, as control is for the proportion of positives expected to be false positives, which is scalable to multiple contrasts.

2.2.10 Ranked gene set enrichment using blood transcription modules

The gene sets used were blood transcription modules (BTMs) from Chaussabel *et al.* [239] (prefixed "DC") and Li *et al.* [240] (prefixed "LI"). Modules are sets of genes with transcriptional and functional similarities across a variety of healthy, diseased, and stimulated conditions. The 260 modules from Chaussabel *et al.* [239] were constructed by unsupervised clustering of 239 PBMC transcriptomes from multiple disease datasets, then annotated by data mining of gene names in PubMed abstracts. The 334 modules from Li *et al.* [240] were constructed from coexpression analysis of approximately 30 000 blood transcriptomes, then annotated making use of Gene Ontology (GO) terms, cell type-specific markers, pathway databases, and manual literature searches. These datasets are particularly suitable for systems vaccinology studies, given their focus on the blood transcriptome. Li *et al.* [240] modules are better annotated in general, and were used for the majority of gene set enrichments in this chapter.

Gene set enrichment analyses were conducted using `tmod::tmodCERNOtest` [241], which assesses the enrichment of small ranks within specific sets of genes compared to all genes, after the genes are ranked by some metric—here I used effect sizes from `bayesmeta`. The CERNO statistic for a gene set is:

$$-2\sum_{i=1}^{n}\ln\frac{r_i}{N} \sim \chi^2(2n) \tag{2.1}$$

where n is the number of genes in the set, N is the number of measured genes in the experiment, and $r_i \in 1, 2, ..., N$ is the rank of the ith gene in the set among all measured genes. CERNO is relatively robust to the ranking metric [242]. FDR control for the number of gene sets tested was performed using BH, again separately for each contrast. The χ^2 test is one-sided, so `tmod::tmodCERNOtest` only considers enrichment of small ranks when computing significance.

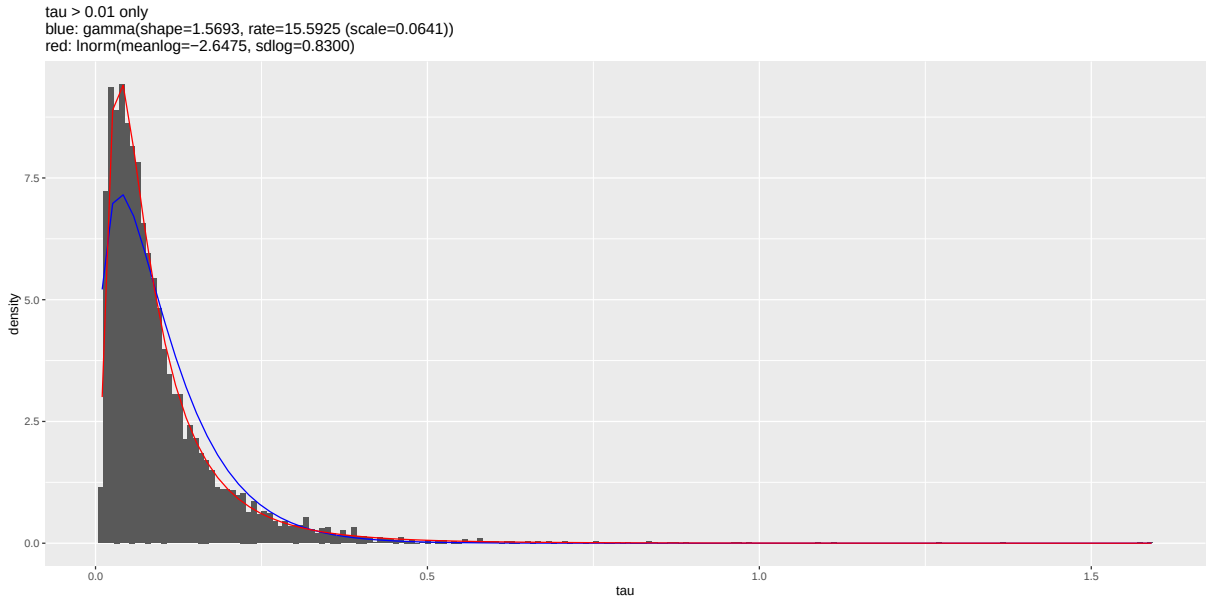

Figure 2.14: Gamma prior for τ (blue) used for `bayesmeta` analyses of the day 1 versus baseline effect, compared to the empirical distribution of per-gene frequentist `metafor::rma.uni` estimates for τ. Genes with small estimates of $\tau < 0.01$ were excluded before distribution fitting. Empirical log-normal fit also shown (red). Distribution parameters are listed.

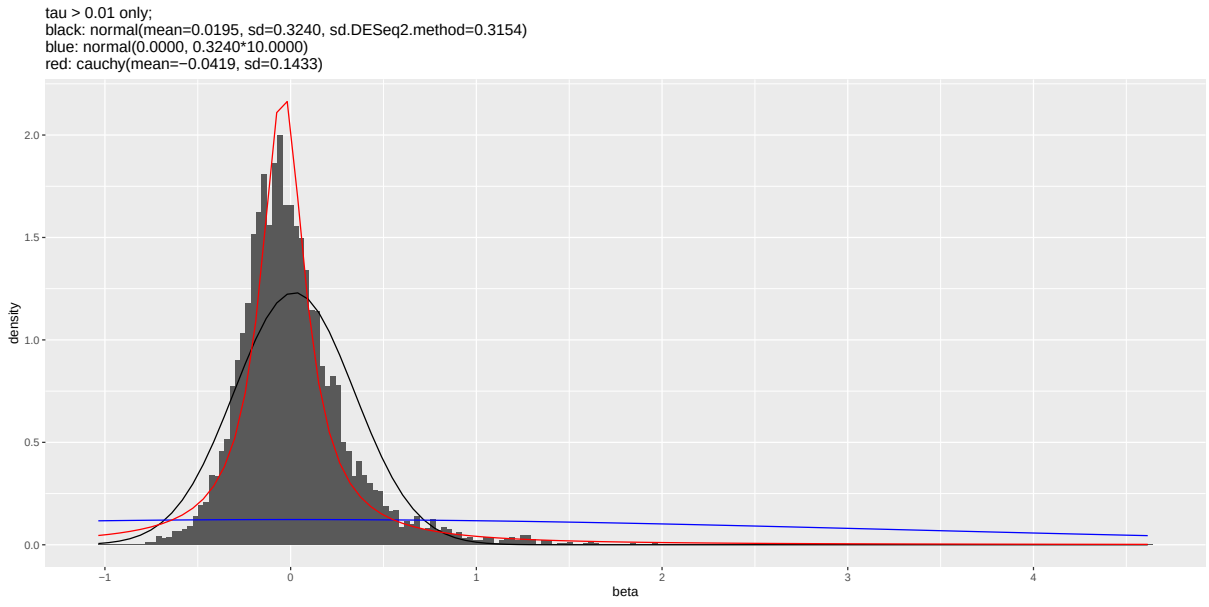

Figure 2.15: Normal prior for μ (blue) used for `bayesmeta` analyses of the day 1 versus baseline DGE effect, compared to the empirical distribution of per-gene frequentist `metafor::rma.uni` estimates for τ. Genes with small estimates of $\tau < 0.01$ were excluded before distribution fitting. The original non-scaled normal fit is shown (black), as well as a Cauchy fit (red). Distribution parameters are listed. An alternative estimate of the Normal standard deviation more robust to outliers using a quantile matching method from `DESeq2` [237] is also given. In this case, it was comparable to the maximum likelihood (ML) estimate from `fitdistrplus`.

As genes can be down or upregulated, separate tests were performed sorting genes in ascending and descending order, and the more significant result was used to determine the overall direction of effect for each gene set. As the approach is rank-based and considers all measured genes, no filters based on the ranking metric were necessary.

The effect size of a gene set enrichment can be quantified with the area under the curve (AUC), computed from U, the test statistic from a Mann-Whitney U test (also known as the Wilcoxon rank-sum test):

$$U = n(N-n) + \frac{n(n+1)}{2} - \sum_{i=1}^{n} r_i \tag{2.2}$$

This is a non-parametric test for whether genes in the set have smaller ranks than genes not in the set on average. Then AUC $= U/(n(N-n))$, which takes values from 0 to 1. Significant results from the one-sided `tmod::tmodCERNOtest` will always have AUC > 0.5.

2.3 Results

2.3.1 Extensive global changes in expression after vaccination

To gain an overview of how the transcriptome changes after vaccination, linear models were fit to identify genes differentially expressed at day 1 or day 7 compared to baseline (day -7 and day 0) in the HIRD array and RNA-seq expression data, accounting for covariates such as batch effects, sex, age, TRI, and ancestry. At the 13 593 genes with expression measured by both platforms, models were fit within each platform. A frequentist random-effects meta-analysis was initially run to generate plausible values for DGE effect size and between-platform heterogeneity. These were used to form empirical priors for a Bayesian random-effects meta-analysis, producing final posterior estimates of effect size and standard errors.

Vaccination induced changes in a large proportion of the PBMC transcriptome; 6257/13 593 genes were differentially expressed between any pair of timepoints (LFSR < 0.05). Applying an absolute FC > 1.5 cutoff identified 857 genes with the strongest effects. Their expression clustered into three general patterns: upregulation from baseline to day 1, then downregulation from day 1 to day 7 back to baseline levels; upregulation from baseline to day 1, sustained at day 7; and downregulation from baseline to day 1, then upregulation from day 1 to day 7 back to baseline levels (Fig. 2.16).

2.3.1.1 Innate immune response at day 1 post-vaccination

Consistent with global expression at day 1 being markedly different from expression at other timepoints (Fig. 2.13), the highest numbers of differentially expressed genes were observed at day 1, with 644 genes differentially expressed vs. baseline. The majority of these (580/644) were upregulated. The gene with the highest FC increase at day 1 compared to baseline was *ANKRD22* ($\log_2$ FC $= 4.49$), an interferon-induced gene in monocytes and DCs involved in antiviral innate immune pathways [243]. Other key genes in the interferon signalling pathway [244] such as *STAT1* ($\log_2$ FC $= 2.17$), *STAT2* ($\log_2$ FC $= 0.95$), and *IRF9* ($\log_2$ FC $= 0.82$) were also upregulated at day 1. Rank-based gene set enrichment analysis using `tmod` [241] revealed that genes with

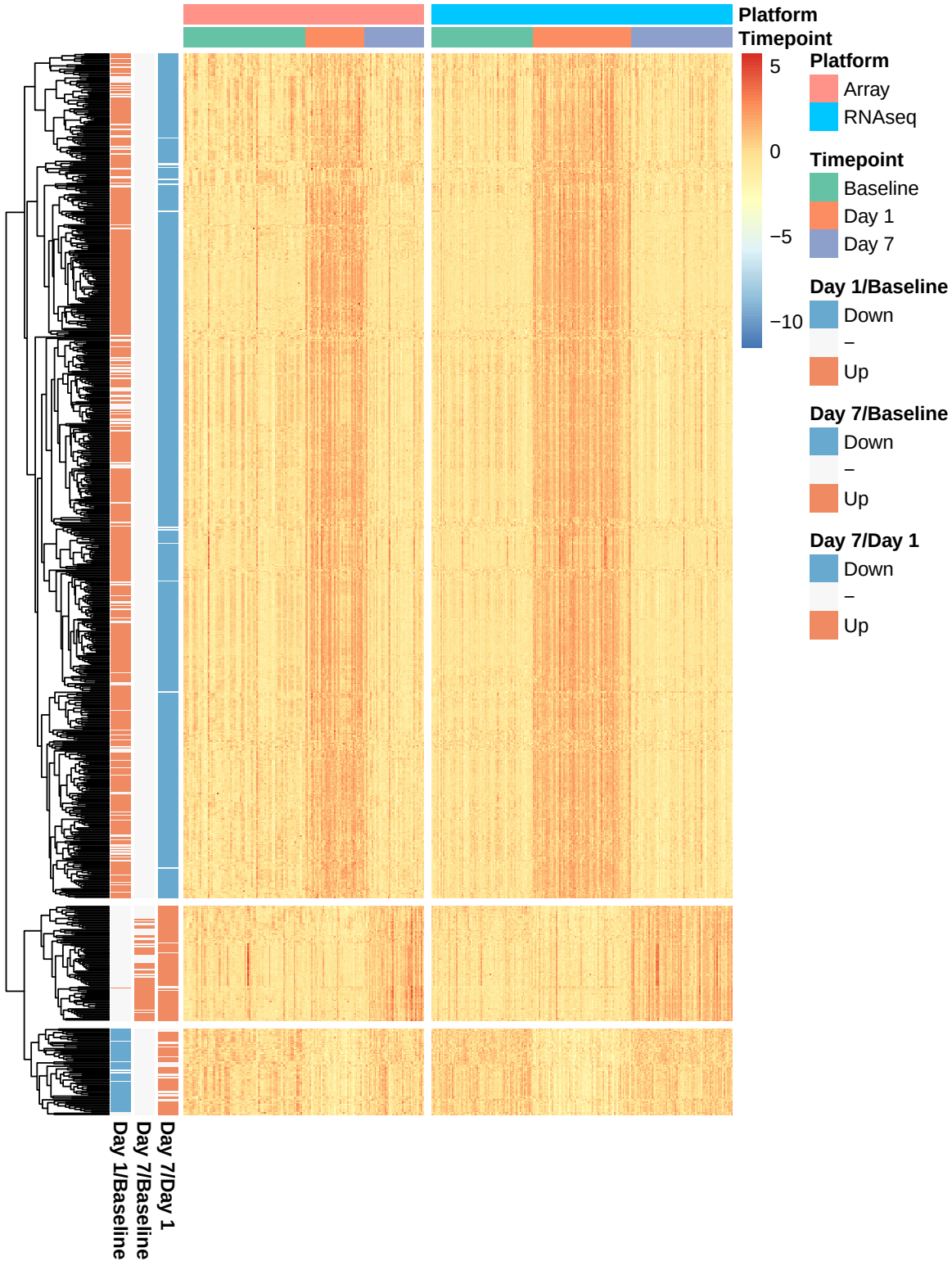

Figure 2.16: Normalised gene expression for 857 genes differentially expressed between any pair of timepoints (lfsr $<$ 0.05, |FC| $>$ 1.5). Rows are genes; columns are samples. Genes were standardised within-platform, then hierarchically-clustered by Manhattan distance. Baseline timepoints are days -7 and 0. Row annotations show DGE between pairs of timepoints. Column annotations show sample platform and timepoint.

the large FC increases at day 1 were enriched in modules associated with interferon, activated DCs, monocytes, and toll-like receptors (TLRs) and inflammatory signalling (Fig. 2.17). Sobolev *et al.* [162] reported only a 1.6-fold ($\log_2 1.6 = 0.68$) increase in blood monocytes from baseline to day 1, as measured by FACS, so these changes reflect active, per-cell upregulation as well as proliferation.

Sixty-four genes were downregulated at day 1, enriched in modules associated with T cells and natural killer (NK) cells. The largest absolute fold change was observed for *FGFBP2* ($\log_2$ FC $= -0.91$), which encodes Ksp37, a secretory protein unique to $CD8^+$ T cells and NK cells [245]. Again, the fold changes in expression were of greater magnitude than observed for the abundance of these cell types, suggesting active downregulation Sobolev *et al.* [162].

As can be seen in Fig. 2.16, there was a general tendency for expression to return to baseline levels by day 7. This was the case for 566/644 upregulated genes and 44/64 downregulated genes, indicating the innate phase of response likely peaks in the first few days.

2.3.1.2 Adaptive immune response at day 7 post-vaccination

Fifty-nine genes were differentially expressed at day 7 vs. baseline. The genes with the highest upregulation were genes associated with B cell differentiation and maturation: *TNFRSF17* (marginal zone B and B1 cell specific protein, $\log_2$ FC $= 1.75$) and *MZB1* (B-cell maturation antigen, $\log_2$ FC $= 1.74$). Genes specific to plasma cells, including *SDC1* (which encodes CD138, required for plasma cell maturation [246]) ($\log_2$ FC $= 1.37$) and *ELL2* (which mediates antibody secretion [247]) ($\log_2$ FC $= 0.87$) were also prominently upregulated. This matches an almost 5-fold increase in plasma cell abundance at day 7 compared to baseline [162]. Strongly enriched modules at day 7 were related to mitosis and cell proliferation, particularly in $CD4^+$ T cells (Fig. 2.17). Both the $CD4^+$ T cell and plasma cell response are indications of a shift toward an adaptive and primarily humoral immune response by day 7.

2.3.2 Expression associations with antibody response

2.3.2.1 Between-platform heterogeneity hinders detection of gene-level associations

Using only array expression data, Sobolev *et al.* [162] identified a set of 62 genes with day 7 expression associated with antibody response, where response was defined as a binary phenotype based on 4-fold increases in HAI or MN titres from day -7 to day 63. Many of these genes were related to plasma cell development and antibody production. I aimed to find genes similarly associated with antibody response in the meta-analysis of array and RNA-seq expression data, and assess the replication of the 58/62 genes that fell into the set of 13 593 genes measured by both platforms.

I computed a baseline-adjusted, continuous measure of antibody response, the TRI [153]. The TRI is comparable to the binary definition in ranking (Fig. 2.2g, Fig. 2.2h), but as a continuous phenotype, it improves statistical efficiency to detect associations. Within just the array data, 51/58 genes were replicated (FDR < 0.05), confirming TRI and the binary response phenotype were comparable. However, using only the RNA-seq data replicated 0/58 genes.

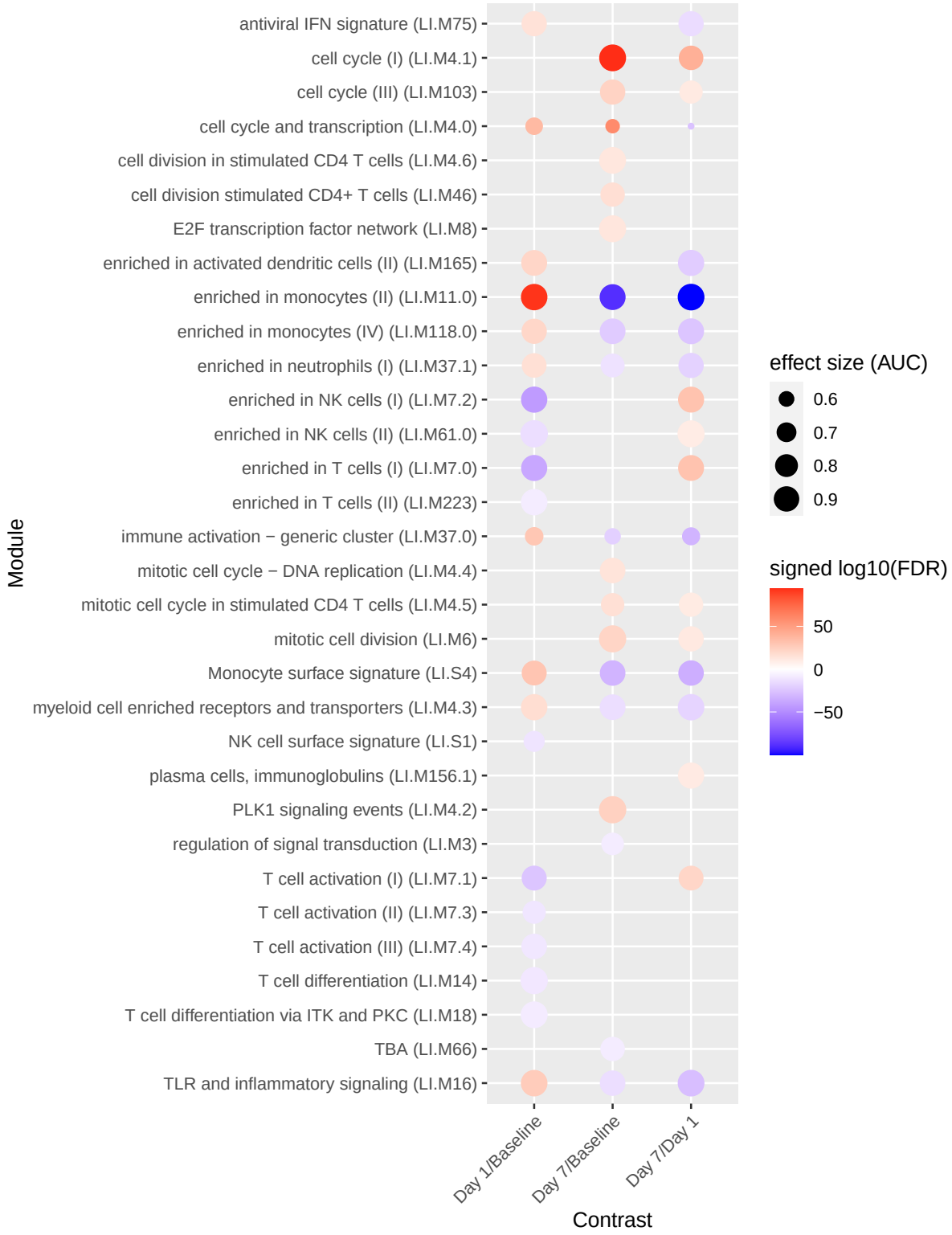

Figure 2.17: Transcriptomic modules up or downregulated between pairs of timepoints. The top ten most significant modules for each contrast are shown. Size of circle indicates absolute effect size (AUC). Color of circle indicates significance (FDR < 0.05) and direction of effect (red = upregulation, blue = downregulation). Absence of circle indicates non-significance.

In the initial frequentist random-effects meta-analysis, with a significance threshold of FDR < 0.05, 6 genes had expression associated with TRI at baseline (Fig. 2.18f), 55 at day 7 (Fig. 2.18h), and 11 pooling samples over all timepoints (Fig. 2.18e). Of the day 7-specific associations reported by Sobolev *et al.* [162] (circled in Fig. 2.18h), 15/58 replicated, all with the same positive direction of effect (high expression with high TRI). However, almost all significant results displayed higher effect sizes in the array compared to RNA-seq (13/15 genes). This was in contrast to the associations identified with timepoint, where significant genes had more consistent effects between platforms along the diagonal (Fig. 2.18b–d). The likely cause is the presence of more extreme antibody response phenotypes (higher TRI range) in the array versus the RNA-seq dataset (Fig. 2.3). This represents an additional source of between-platform variation not due to technical factors, but inherent to the samples themselves.

The Bayesian meta-analysis pipeline more robustly models between-study heterogeneity due to platform and sample-specific effects. Due to shrinkage of effects, few genes with effects closer to the dense center of the effect distribution were called as significant, and significant genes tended to have outlying effect sizes in both platforms (compare Fig. 2.18b–d with Fig. 2.19b–d). No single gene was detected as significantly associated with TRI at LFSR < 0.05 for any contrast: not at any single timepoint, nor when pooling samples across all timepoints (Fig. 2.19e–h). The frequentist meta-analysis is likely to use poor estimates of the between-platform heterogeneity, as there are only two data points to estimate it from. Indeed, all 15 significant genes with day 7 expression associated with TRI in the frequentist meta-analysis had unrealistic between-platform heterogeneity estimates of exactly zero (Fig. 2.20).

2.3.2.2 Module-level associations with antibody response

Using effect sizes from the Bayesian meta-analysis, significant enrichments were detectable at the gene set level. The strongest effects were seen at day 7, where expression of modules related to the cell cycle, $CD4^+$ T cells, and plasma cells were positively associated with TRI—"cell cycle (I)" (LI.M4.1, FDR $= 6.81 \times 10^{-54}$), "Plasma cell surface signature" (LI.S3, FDR $= 1.78 \times 10^{-12}$), and "cell division stimulated CD4+ T cells" (LI.M46, FDR $= 5.54 \times 10^{-10}$) (Fig. 2.21).

Associations with TRI were also detected at baseline. A diverse set of set of modules had positive associations, including "chemokines and inflammatory molecules in myeloid cells" (LI.M86.0, FDR $= 2.25 \times 10^{-11}$), "platelet activation - actin binding" (LI.M196, FDR $= 1.71 \times 10^{-8}$), "enriched in B cells (I)" (LI.M47.0, FDR $= 2.40 \times 10^{-7}$), "cell adhesion" (LI.M51, FDR $= 1.22 \times 10^{-10}$), "myeloid, dendritic cell activation via NFkB (I)" (LI.M43.0, FDR $= 4.68 \times 10^{-7}$), and "proinflammatory dendritic cell, myeloid cell response" (LI.M86.1, FDR $= 4.11 \times 10^{-7}$). Monocyte modules "enriched in monocytes (II)" (LI.M11.0, FDR $= 3.53 \times 10^{-4}$) and "Monocyte surface signature" (LI.S4, FDR $= 1.17 \times 10^{-3}$) were negatively association with TRI. Negative associations for these same modules were also maintained at day 1 (LI.M11.0, FDR $= 1.41 \times 10^{-10}$; LI.S4, FDR $= 1.74 \times 10^{-6}$) and at day 7 (LI.M11.0, FDR $= 5.54 \times 10^{-10}$) (Fig. 2.21).

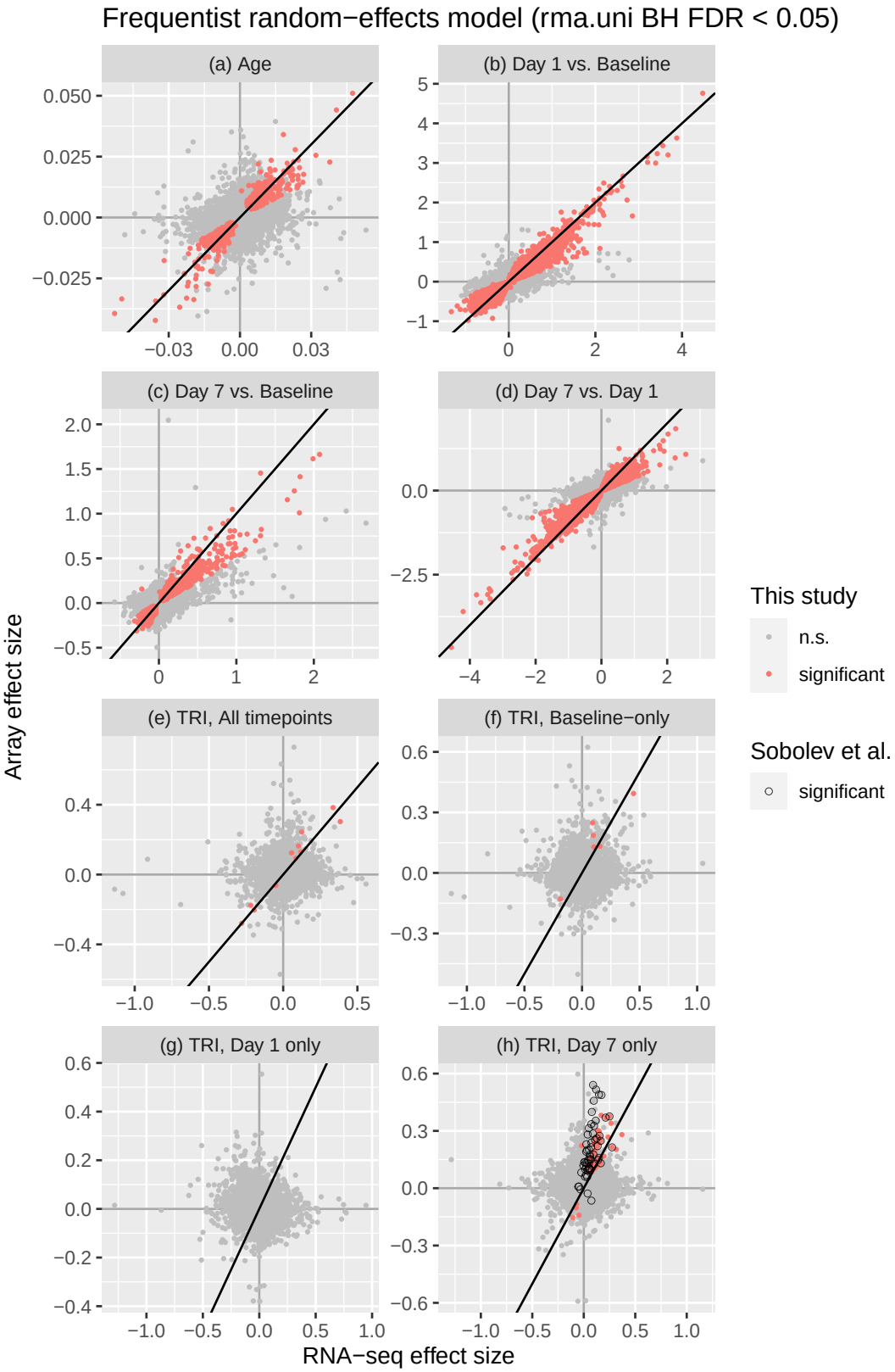

Figure 2.18: DGE effect sizes ($\log_2$ FC) estimated in array versus RNA-seq samples, colored by significance in frequentist random effects meta-analysis using `rma.uni` at BH FDR < 0.05. Genes with day 7 expression associated with binary responder/non-responder status in Sobolev *et al.* [162] are circled for that contrast.

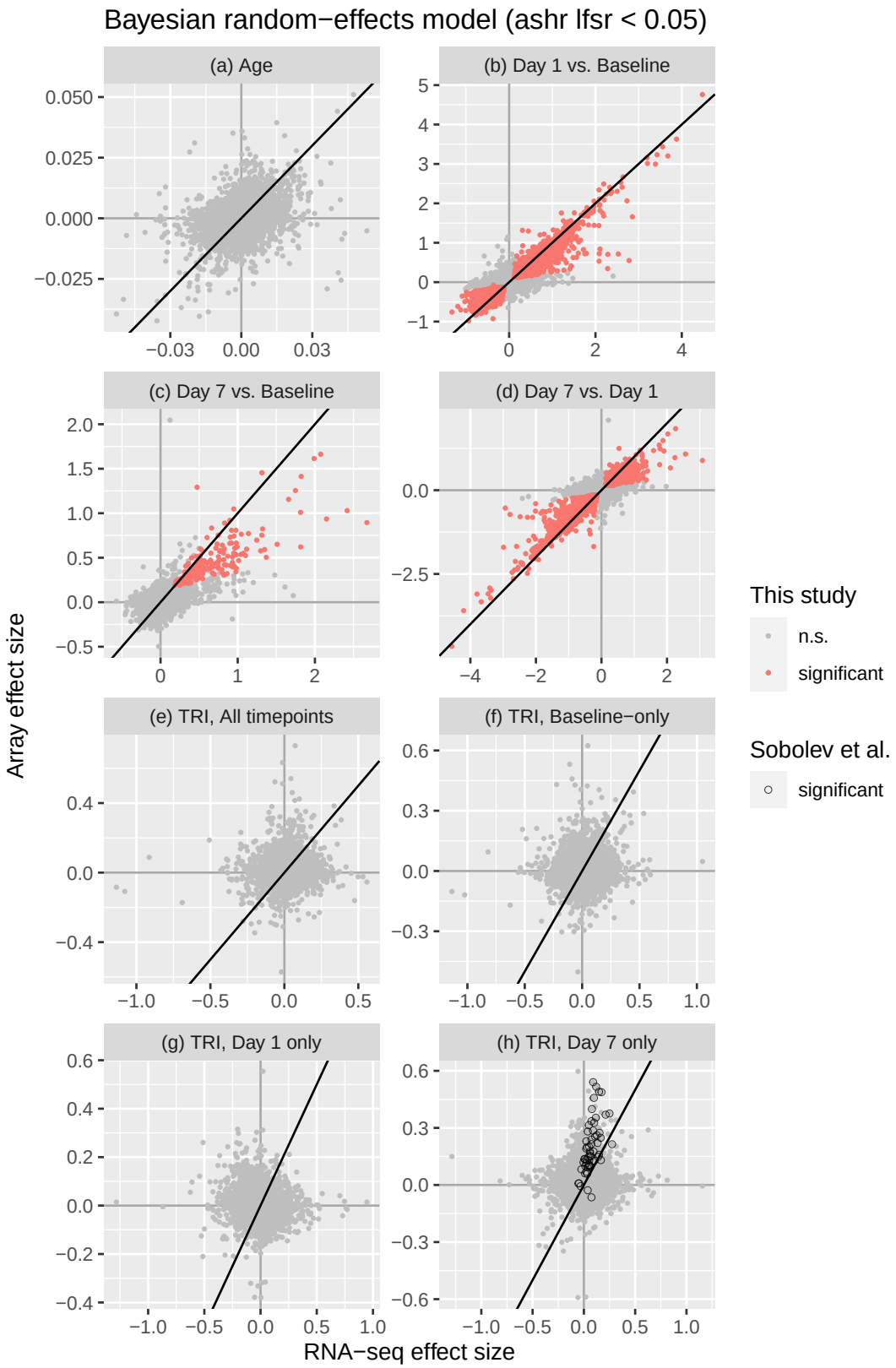

Figure 2.19: DGE effect sizes ($\log_2$ FC) estimated in array versus RNA-seq samples, colored by significance in Bayesian random effects meta-analysis using `bayesmeta` at `ashr` LFSR < 0.05. Genes with day 7 expression associated with binary responder/non-responder status in Sobolev *et al.* [162] are circled for that contrast.

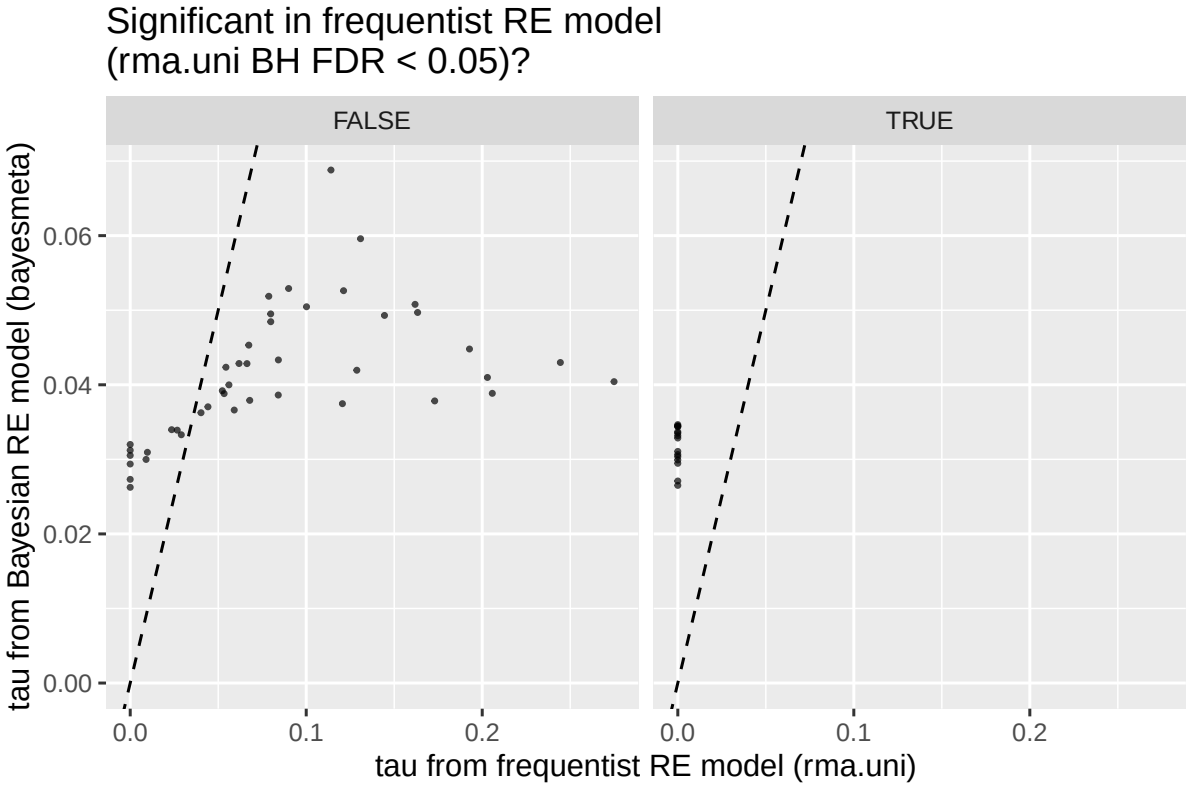

Figure 2.20: Estimates of between-platform heterogeneity τ from frequentist and Bayesian meta-analysis, for the 58 genes with a significant association between day 7 expression and binary responder/non-responder status in Sobolev *et al.* [162]. Dashed line is the identity line. Estimates from the frequentist method cover a wide range and can be zero. For this contrast testing association between day 7 expression and TRI, 8563/13 593 of per-gene τ estimates are zero, including all 15/58 significant results (right). Significant results are array-driven, with 13/15 having higher effects in array than RNA-seq (54/58 genes overall). Estimates of τ from the Bayesian method are in a narrower range and constrained away from zero by the prior.

Figure 2.21: Gene expression modules associated with antibody response (TRI). Enrichments were performed with all timepoints pooled, and at each timepoint specifically. The top ten most significant modules for each contrast are shown. Size of circle indicates absolute effect size (AUC). Color of circle indicates significance (FDR < 0.05) and direction of effect (red = expression positively correlated with TRI, blue = negatively correlated). Absence of circle indicates non-significance.

2.4 Discussion

A meta-analysis of array and RNA-seq data revealed extensive transcriptomic response to Pandemrix vaccination in the HIRD cohort. At day 1, there was upregulation of genes and modules related to monocytes, interferon signalling, and the inflammatory response; and downregulation of T cell and NK cell genes and gene modules. Concordant changes in these gene modules were also reported by Nakaya *et al.* [158] at day 1 after MF59-adjuvanted seasonal TIV in young children, but changes in these modules were not as consistent in children who received non-adjuvanted TIV. The AS03 adjuvant in Pandemrix is thought to act by promoting chemokine secretion, predominantly targeting monocytes and macrophages [163, 248], which concurs with the strong upregulation of monocyte and DC modules observed at day 1 after Pandemrix. A large component of the expression response at day 1 may reflect response to the adjuvant. Most genes differentially expressed at day 1 returned to baseline expression by day 7. Nakaya *et al.* [158] saw a similar trend comparing day 0 and day 3 for MF59-adjuvanted TIV. Unadjuvanted seasonal TIV also causes peak transcriptomic induction at day 1 [153]. Although the timepoint resolution here is coarse, the early innate response to Pandemrix is transient, peaking less than 7 days, and likely less than 3 days post-vaccination. Upregulation of cell cycle, proliferating $CD4^+$ T cell, and B (plasma) cell genes and modules were detected at day 7. This indicates a shift to the adaptive immune response, likely involving $CD4^+$ T cell-supported differentiation and proliferation of ASCs.

Both day 1 and day 7 expression module changes were concordant with changes in cell populations seen in the HIRD FACS data. The greater magnitude of expression fold change of individual genes compared to cell abundance fold changes suggests the influence of both mechanisms [162]. Statistical adjustment for measured or estimated cell composition is one possibility; I explore these methods in Chapter 3 and Chapter 4. An experimental approach would be *in vitro* stimulation of PBMCs with vaccine, ruling out cell migration, but not shifts in cell subtype composition [249].

The overall patterns of expression over time were consistent between array and RNA-seq, with the meta-analysis identifying genes with outlying effects in both platforms. In contrast, I was not able to replicate the 58 gene-level associations reported by Sobolev *et al.* [162] between day 7 expression and antibody response that were assessable in my meta-analysis. The difference was not wholly due to response definitions, as within the array data alone, switching from binary response status to TRI still replicated the majority of reported associations, but using either binary response status or TRI in the RNA-seq data alone found no significant associations. Initially, 15/58 signals replicated using frequentist random-effects meta-analysis to combine per-platform estimates. I do not consider these hits as robust, as the estimated between-platform heterogeneity was zero for all 15 of these signals. None of these signals replicated in the Bayesian random-effects meta-analysis, where prior information about τ could be incorporated, discouraging unrealistic estimates of zero heterogeneity. The Bayesian meta-analysis was in general more conservative, calling fewer differentially expressed genes compared to the frequentist analysis for all contrasts. Most of the 58 genes also had larger effects in the array dataset than in the RNA-seq dataset, possibly because the array data contains more extreme TRIs. At a single-gene level, significant associations with timepoint are robustly detectable, but associations with TRI

have effects too modest relative to the noise introduced by platform-dependent technical effects and dataset-dependent phenotype distributions.

Expression associations with antibody response were, however, observed at the gene set level, at modules associated with TRI as a whole. The strongest effects were observed at day 7, where modules related to adaptive immunity (cell cycle, stimulated $CD4^+$ cells, plasma cells) were positively associated with TRI. These same modules were upregulated at day 7 compared to baseline; it seems that those individuals with the greatest antibody response to vaccination are most able to induce these modules by day 7 post-vaccination.

Module associations with TRI were also observed pre-vaccination with both positive (e.g. chemokines, proinflammatory DCs, B cells, platelet activation) and negative (e.g. monocytes) directions of effect, suggesting baseline immune state has influence on long-term antibody response to Pandemrix. Some of the positive associations have been previously reported for unadjuvanted seasonal influenza vaccines in multiple independent cohorts. The same B cell modules were reported by Nakaya *et al.* [157], and similar DC, inflammatory, and platelet activation modules were found to be predictive of antibody response in young adults [159]. The negative association of monocyte modules with antibody response at baseline was also reported by Nakaya *et al.* [157]. Interestingly, I detected the same negative associations at day 1 and day 7. Monocyte modules were one of the most upregulated modules at day 1, and although the module annotations do not separate monocyte subsets, abundance of $CD16^+$ inflammatory monocytes was particularly increased at day 1 in the FACS data [162]. This lends some support to the hypothesis that chronic baseline inflammation or excessive/prolonged post-vaccination inflammation—specifically driven by monocytes—can be detrimental to the humoral response [157, 250, 251].

There are several caveats to consider when drawing comparisons to the systems vaccinology literature. Most studies are of unadjuvanted multivalent seasonal vaccines; HIRD used an adjuvanted monovalent pandemic vaccine. Most studies measure post-vaccination antibody response around the expected peak of day 28; HIRD measured later at day 63, which may attenuate the signal. The specific genes within modules driving associations may also differ between studies. Nevertheless, the ability to observe module-level associations with TRI also reported in previous studies with diverse populations, measurement platforms, influenza seasons, and analysis pipelines, is a stark contrast to difficulty of replicating single-gene associations even within the HIRD cohort itself. When the effect of individual genes on phenotype is expected to be subtle, module-level analyses are not only more sensitive, but appear to be more generalisable.

The next step is to explore the utility of the identified associations for prediction. Although I have identified highly significant associations between expression modules and antibody response, that does not imply the ability to accurately predict response from expression [161]—that is, the existence of molecular signatures. Some exploration can be done within HIRD using cross-validation, or by setting aside a subset (e.g. the array data) as a test set, but having an independent test set is especially important for prediction to guard against overfitting. Matched expression and antibody data are rare for adjuvanted and pandemic vaccines, so an initial effort would likely draw on published seasonal vaccine datasets (e.g. [157, 159]), with the aim of identifying shared molecular signatures.

The fundamental question of why gene expression and antibody responses vary between

HIRD individuals also remains. Which genes, if their expression were to be modulated, would lead to a change in antibody response? This is a critical question in the move from identifying correlates of protection and molecular signatures, towards targeted interventions to improve vaccine outcomes [146]. The descriptive design of the HIRD study does not lend itself to exploring causation between expression and antibody titres without a causal anchor. Interindividual genetic variation could play such a role; Chapter 3 will examine the impact of common host genetic variants on expression response in the HIRD cohort.