Chapter 5

Discussion

Human immune response to perturbation is variable at numerous molecular and phenotypic levels. In this thesis, I profiled the transcriptomic response to *in vivo* vaccine and drug perturbations, established associations between expression and phenotypic response, and mapped changes over time in the genetic regulation of expression response. Chapter 2 focused on transcriptomic response to Pandemrix vaccination in the Human Immune Response Dynamics (HIRD) cohort, describing the transition from innate to adaptive immune response, and detecting associations between expression and antibody response. In Chapter 3, I considered the impact of host genetics on vaccine response in HIRD, identifying genetic variants associated with changes in expression post-vaccination, then exploring potential mechanisms explaining those associations. Finally, Chapter 4 applied similar analysis frameworks in a different context, response to anti-tumour necrosis factor (TNF) therapy in Crohn's disease (CD) patients in the Personalised Anti-TNF Therapy in Crohn's Disease (PANTS) cohort, finding distinct trajectories of expression between primary responders and non-responders to treatment. Each chapter presented its results and limitations in turn, but similarities in design and analysis qualify them for a joint deliberation. In this final chapter, I highlight shared themes, examine core limitations, and outline considerations for the design and analysis of future longitudinal *in vivo* perturbation studies to better our biological understanding of immune response to vaccines and drugs.

5.1 Strategies for detecting robust associations

In Chapters 2 and 4, I focused on identifying genes with differential expression after immune perturbation, or expression associated with phenotypic response variables—antibody titres and clinical anti-TNF response respectively. Vaccine and drug perturbation had strong effects on large proportions of the blood immune transcriptome, resulting in thousands of highly significant associations when comparing pre- and post-perturbation timepoints. In comparison, it was much more challenging to identify robust single-gene associations with response phenotypes. In Chapter 2, associations of day 7 expression with antibody response from Sobolev *et al.* [162] were replicated in my analysis of the original array data, but not in newly-generated RNA sequencing (RNA-seq) data, or in the meta-analysis. In Chapter 4, baseline associations with anti-TNF response from the literature—including at *TREM1*, previously reported by two independent groups [366, 367]—were not significant in my analysis of the PANTS cohort. The biological

effect size of a single gene's expression on phenotypic response is likely to be small, eclipsed by other sources of variation: measurement platform, difference in response definitions, sample characteristics, and noise. The idealistic suggestion is to increase sample size, but resource and ethical constraints will always exist. Rather than creating new cohorts, a logistically-efficient strategy is sampling from individuals enrolled in drug and vaccine trials, but care must be taken to ensure the trial is powered both for its primary endpoints, and for planned transcriptomic analyses. Power calculations for differential gene expression (DGE) are non-trivial and it is often unknown what a reasonable effect size to assume might be. Many experiments choose parameters like sample size and sequencing depth based on rules of thumb [196], or to be comparable to existing ones in the field. In cases where small effects are likely and high power is not guaranteed, one should be cognizant of winner's curse when reporting and interpreting associations determined to be significant based on some threshold [305].

Another consideration is how best to distribute a fixed sample size between depth (number of individuals) and richness (number of timepoints, phenotypes, data types). Some degree of longitudinal sampling is recommended for a phenotype as dynamic as immune response. Chapter 2 demonstrated a distinct jump from day 1 innate to day 7 adaptive immune expression profiles post-vaccination, but the kinetics of the transition are not clear. In hindsight, responses could have peaked earlier or later in different individuals, and variation in the speed of response cannot be examined without denser sampling. In Chapter 4, expression differences between anti-TNF responders and non-responders were apparent from week 14, but it is not known if differences actually appear much earlier. Future analysis of a (small) number of available PANTS RNA-seq samples from day 3 after initiating treatment may uncover associations in the early innate response.

Rich sampling also offers analytical advantages. Having repeated measures from the same individuals allowed modelling of within-individual covariance in Chapters 2 and 4, improving statistical efficiency. The spline model in Chapter 4 enabled separation of responders and non-responders based on expression trajectory over multiple timepoints. However, all those models only incorporated two data types: expression and phenotypic response. Studies in the systems vaccinology field have demonstrated how integrating networks of many data types identifies correlates and predictors of response not only in the transcriptome, but in multiple layers of the immune system [413]. In HIRD, longitudinal fluorescence-activated cell sorting (FACS) and cytokine measurements are available for this form of integrative modelling.

When transcription is quantified on a global scale, analyses should not consider genes in isolation. Genes in the immune system are not independent, and just as variation increases uncertainty, covariation reduces it*. In Chapter 2, imprecise estimates from multiple genes were used to build an informative empirical prior for between-platform heterogeneity. Throughout the thesis, I make extensive use of enrichment analyses with gene sets defined by prior biological knowledge, to detect subtle but coordinated changes based on the expression of multiple genes. General purpose gene sets may be less relevant in immune cells [157], so I used blood transcription modules (BTMs) [239, 240] tailored for immune gene expression in blood. Alternative databases

*Wickham, H. & Grolemund, G. Chapter 7: Exploratory Data Analysis. R for Data Science. https://r4ds.had.co.nz/exploratory-data-analysis.html

that provide immune-focused gene sets include InnateDB [414] and MSigDB [415]. Many significant module associations with vaccine antibody response and clinical drug response were identified in Chapters 2 and 4, and my expectation is that these should be more replicable than any single-gene associations I reported (e.g. *SIGLEC10* from Chapter 4). While the effect size of a single gene may vary from sample to sample due to noise, a summary measure computed from multiple genes should be more robust. Indeed, some module associations between baseline expression and antibody response found in Chapter 2 were reported in previous studies of seasonal influenza vaccines. Most systems vaccinology studies aiming to identify consistent associations with vaccine response over multiple cohorts and sampling years focus their analyses at the gene set level [146]. Gene set analyses cannot, however, be divorced from examining the genes within them, as the genes that drive set-level associations can differ between apparent replications, and the mapping between genes and gene sets is one-to-many.

5.2 Responder analysis

A key determinant of how well the models in this thesis might correspond to reality lies in the assumed model for phenotypic response. By encoding response as an independent variable, an assumption is made that it is a stable characteristic of an individual that is measured without error*. This may not be an accurate assumption. Imagine a hypothetical drug or vaccine where 60 % of a sample of individuals have an observed response phenotype: "60 % of the time, it works every time"†. This is compatible with a stable 60 % success rate in 100 % of individuals (variation in observed response is entirely due to chance), or a stable 100 % success rate in 60 % of individuals and a 0 % success rate in the other 40 % of individuals (response is highly personal)—most likely the truth is somewhere in between. In the first scenario, it is difficult to imagine identifying robust baseline associations with response. This has been extensively discussed in the context of randomised controlled trials [416], but similar issues pertain to response definitions in observational studies.

One needs to establish how correlated phenotypic response is over time within the same individual, and computing within-individual variation requires replication at the level of the individual. The same individual must be *perturbed and measured* more than once [417]. This is not always possible in practice; in Chapter 2, antibody response was defined based on a single measurement after a single vaccine dose, but measuring response after a hypothetical second dose would quantify a different phenotype: the secondary immune response based on vaccine-induced immune memory. In Chapter 4, patients did receive repeated anti-TNF doses interspersed with sampling timepoints, and the expression differences between clinical responders and non-responders seen at week 14—the timepoint where clinical response was assessed—were maintained at week 30 and week 54. This suggests the initial designation of non-responders is not entirely due to chance, but due to some characteristic of patient disease state.

Even if response is actually a stable personal characteristic, one still needs to select an appropriate mathematical definition. As discussed in Section 2.2.2, a binary definition of response

*Note that the regression framework can accommodate measurement error in the context of errors-in-variables models.

†Apatow, J., McKay, A. & Ferrell, W. Anchorman: The Legend of Ron Burgundy (2004).

based on dichotomisation is inefficient and biologically implausible. I instead used the titre response index (TRI) in Chapters 2 and 3, a continuous change score combining haemagglutination inhibition (HAI) and microneutralisation (MN) titres, residualised on the baseline titres. In Chapter 4, the binary clinical response phenotype is based on a complex decision tree with many inputs. Defining dichotomies based on multiple inputs can lead to discontinuities and non-monotonicity in response probabilities under small changes in inputs [173]. Pragmatism did come into play when choosing these definitions. For DGE, the most widespread models have expression as the sole dependent variable, and encode phenotypic response variables as independent variables. Both TRI and the PANTS clinical response definition provided that single independent variable. In hindsight, variation in response definitions likely contributes to difficulties in replicating associations between studies, so it may be more sensible to model on the component phenotypes themselves (e.g. log HAI and MN titres, C-reactive protein (CRP) levels and Harvey Bradshaw index (HBI) scores).

5.3 Challenges in the interpretation of bulk expression data

Bulk expression data is a mixture of cell types with heterogeneous expression profiles. One of the largest sources of variation in bulk blood expression data is variation in immune cell composition, generated both from true variation in composition and sampling effects. The more cell type-specific a gene's expression, the more its measurement in bulk is affected by cell composition [418]. Highly cell type-specific genes can be treated as marker genes, used in deconvolution methods to estimate cell proportions in bulk samples when they are not directly measured. In Chapter 3, `xCell` [298]—while not technically a deconvolution method—was used to estimate cell type enrichment scores from array and RNA-seq expression data. In Chapter 4, estimates of cell proportions were computed by deconvolution of matched genome-wide methylation data. When fit as covariates in linear regression, cell abundance estimates act as precision variables for sampling noise, but additionally as mediators of the perturbation's effect on expression. In Chapter 4, I chose to run two sets of models with and without including estimates of five major immune cell proportions, gaining some information on which effects are likely driven by cell abundance, and which are driven by per-cell up or downregulation of transcription.

Using major cell populations for correction misses the contribution of rare populations [389]. For *cis*-expression quantitative trait locus (eQTL) mapping in Chapters 3 and 4, where the main concern was maximising the number of eQTLs detected, hidden factors from PEER [182] were included into models in addition to cell abundance estimates from deconvolution. PEER factors were correlated with deconvoluted cell abundances, so it is likely they capture additional variation from rarer cell types. If having interpretable covariates for cell abundance is unimportant, methods like surrogate variable analysis [419, 420] can be used to adjust for cell composition and other unmeasured technical sources of variation in DGE analyses also.

Interpretable covariates for cell abundances are important for considering response expression quantitative trait locus (reQTL) effects in bulk data. As discussed in Section 3.2.10, it is model misspecification to omit genotype-cell abundance interactions if the effect of genotype changes depending on cell abundance. It is even popular to use such interaction terms between genotype

and cell abundance (or a proxy of cell abundance) to discover cell type-specific eQTLs [71, 74]. *In vivo*, cell abundances are causally affected by the perturbation due to active recruitment, differentiation, and proliferation of immune cells. Consider the case where vaccine perturbation causes active proliferation of a rare cell type that is near absent at baseline, but forms a greatly increased proportion of the bulk mixture after perturbation. Any baseline eQTL specific to this cell type will appear as a reQTL at the post-perturbation timepoint, because expression of that cell type contributes more to the bulk mixture. If the eGene is not cell type-specific in its expression, adjusting for abundance of the cell type will only offset the regression lines at each timepoint, but not change their slopes relative to one another. The eGene also does not have to be differentially expressed on average, as the effect of interest is not the pre-post difference in expression, but the effect size of genotype on that difference. In Chapter 3, I found that an increase in naive classical monocytes at day 1 revealing a non-stimulus-specific but monocyte-specific eQTL, for the non-monocyte-specific gene *ADCY3*, was a plausible mechanism underlying the strongest day 1 reQTL.

An aim of the *in vivo* reQTL design is to find host genetic variants with a causal effect on response to perturbation. The crux of the issue is whether such an interpretation is justifiable: whether a difference in group-level eQTL regression slopes between baseline and post-perturbation necessarily entails a causal effect of genotype on change in expression from baseline to post-perturbation at the individual level. For the specific case of the *ADCY3* day 1 reQTL, I believe so. As the variant is an eQTL in monocytes, individuals with more copies of the effect allele have higher *ADCY3* expression per-monocyte on average. If you were to change the genotype of an individual from homozygous non-effect to homozygous effect, you *would* change their post-vaccination increase of *ADCY3*, because the exact same increase in monocyte abundance from baseline to day 1 would provide more *ADCY3* transcripts. It is less clear in the general case, as there are many possible mechanisms for an observed reQTL: a gene with an eQTL not expressed at baseline becoming detectable (power), a cell type with a cell type-specific eQTL increasing in proportion (recruitment or proliferation), the effect of a cell type-specific eQTL increasing within that cell type (activation, the canonical scenario assumed for *in vitro* stimulation), a genotype-dependent increase in cell abundance creating a reQTL for a gene with cell type-specific expression, *et cetera*. Not all of these can be ruled out just by including cell abundances as covariates in the eQTL model. Even if a large number of reQTLs can be detected by statistical interaction, as in Chapter 3, the challenge is distinguishing between these mechanistic scenarios and forming causal hypotheses. It is also unclear whether *in vivo* reQTLs provide additional utility over *in vitro* reQTLs for gene prioritisation at genome-wide association study (GWAS) loci. Theoretically, there may be effects unobservable without *in vivo* interactions in the immune system, but a systematic comparison of reQTLs detected with *in vivo* and *in vitro* stimulation has not been performed. *In vivo* reQTL studies are certainly not ineffectual at their stated goals, but cell composition does add considerable complexity to their interpretation. Although insights into the biological mechanism of the stimulation response is easier to gain when cell type abundance is controlled *in vitro*, one basic utility of *in vivo* stimulation is allowing the detection of additional cell type-specific eQTL effects in bulk data using genotype-cell type abundance interaction terms, a methodology already well-established in non-stimulated bulk

samples (e.g. [71]).

To truly control for cell composition, the best option is to control it at the study design stage. Adjusting for cell abundance in regression only attempts to estimate the effect of other predictors if cell abundance were held constant. It does not change the cell abundances to be equal—it is a change of viewpoint, not a change of data. Adjusting for abundance also cannot distinguish cell types with correlated abundance estimates. Single-cell RNA sequencing (scRNA-seq) after *in vivo* perturbation would quantify per-cell expression and cell abundance simultaneously. The technology is emerging as an alternative to bulk sequencing of FACS-sorted cells, having comparable cost, and the additional advantage of not requiring pre-defined marker sets [421]. There is flexibility in choosing to conduct DGE analysis and eQTL mapping within each cell type cluster, or to pool clusters to mimic bulk data. Paired designs that leverage the power of bulk reQTL mapping and the cell type resolution of single-cell data have been explored, using eGene expression in clusters to annotate bulk reQTLs to likely cell types [90]. As an emerging technology, scRNA-seq still faces many limitations, such as low coverage of the transcriptome due to dropout, smaller sample sizes due to cost, difficulties in defining robust cell type clusters, and sample processing effects on the transcriptome, but progress in the field has been nothing but rapid.

5.4 From association to prediction

In the DGE regression models I used to test for association of expression with phenotypic response, expression was always placed as the dependent variable, and response as an independent variable. In a clinical setting, a more relevant concern is prediction of patient response from expression (ideally baseline expression), reversing the roles of expression and response in the model. In Chapters 2 and 4, I observed few significant single-gene associations with response at baseline. It is first useful to consider what implications this has on the move from association to prediction in these datasets.

Prediction from genome-wide transcriptomic data is often a $p \gg n$ prediction problem, where the number of potential predictors p dwarfs the sample size n. Efron [422] provides a fascinating case study on predicting prostate cancer status from expression array data ($p = 6033$ genes) in samples from 52 prostate cancer patients and 50 controls ($n = 102$). After randomly splitting the data into training and test sets, each with 26 cancer patients and 25 controls, a random forest used to predict cancer status from gene expression recorded a 2 % test set error. Repeating over many random splits showed this high predictive performance was not an outlier. Random forests have embedded feature selection, assigning their predictors an importance score, with a positive importance score indicating that a predictor was utilised by the model. After removing all 348 genes with positive importance scores in the trained model from the dataset, then repeating the process with remaining $p = 5685$ by $n = 102$ matrix, another model was produced where a set of 364 genes with positive importance—completely disjoint from the first 348—predicted cancer status with a similar error rate. This process could be repeated multiple times, each time producing a model with similar error rate, using none of the "important" genes from the previous models. Although these error rates come from internal validation, which have an optimistic

bias, the performance of pure prediction models does appear to be dominated by the confluence of many weak predictors. Therefore it is still feasible to consider prediction in datasets where attribution of significance to individual strong predictors may be impossible.

A large part of systems vaccinology in the last decade has been building models to predict vaccine-induced antibody and cellular responses from high-dimensional data. The methods used span the full gamut of statistical and machine learning algorithms, including classification to nearest centroid (ClaNC) [249], discriminant analysis via mixed integer programming (DAMIP) [154, 157, 249, 423], nearest shrunken centroid algorithm (e.g. PAM [149]), linear regression [153, 413], logistic regression [127, 155, 424], linear discriminant analysis (LDA) [161, 425], elastic net [160], partial least squares (PLS) [161], artificial neural networks (ANN) [157], naive Bayes [426], lasso regression [427], sparse partial least squares (SPLS) [428], and logistic multiple network-constrained regression (LogMiNeR) [429, 430]. The choice of methodology can be daunting. Fortunately (or unfortunately), an extensive survey of transcriptomic prediction models by the MicroArray Quality Control Consortium [431] found that the choice of algorithm was not as influential on predictive performance as the endpoint itself, with some endpoints being inherently difficult to predict. There is also no need to restrict oneself to a particular method; ensemble models that combine multiple algorithms consistently have the best performance and robustness [432]. It is hard to say *a priori* whether antibody response in HIRD and anti-TNF response in PANTS are "difficult" endpoints. The existence of predictive signatures for seasonal influenza vaccine response using baseline expression—validated over multiple cohorts, years, and geographical locations—does set an encouraging precedent for the former [159].

Oncology was one of the earliest fields to adopt predictive gene signatures into clinical practice. Despite the first commercial tests launching in the early 2000s (e.g. MammaPrint, a 76-gene signature for breast cancer prognosis), only a handful are in use today [433–435]. There are multiple hurdles to clinical implementation, requiring that a signature not only have validated accuracy, but provide sufficient incremental value on top of existing clinical markers in a cost-effective manner [434]. Feature selection is of particular importance when building models for the clinic; cost-effectiveness entails that most expression tests are qPCR-based tests that measure at most a few dozen genes. There is an interesting tension between the sparsity assumed by feature selection methods (that most predictors have no effect) and the observation that prediction algorithms depend on many weak predictors. A balance between predictive performance and measurement cost will likely need to be struck. The ability to predict individual response to anti-TNF treatment would be revolutionary due to the treatment cost and quality of life impact of taking ineffectual biologic therapy. The case for personalised vaccinology lies mostly in building understanding of the best vaccine type, dose, and timing for vaccination of challenging populations [124].

5.5 From association to causality

Knowing the causal mechanisms of immune response to perturbation is crucial for conceiving of possible interventions. For example, assuming the baseline association of *SIGLEC10* with anti-TNF response identified in Chapter 4 is a true association, would intervening on baseline

SIGLEC10 expression affect probability of response? The study designs used in this thesis are uncontrolled, but still provide useful guarantees against reverse causality. Post-perturbation phenotypic or expression measurements cannot cause baseline gene expression. Post-conception phenotypic or expression measurements cannot cause genotype. To estimate causal effects of expression on phenotype, models are needed that encode causal relationships as testable hypotheses. There are several families of such methods; as I shall now describe, they should be used in combination.

Mendelian randomisation (MR) is a form of instrumental variable (IV) analysis that uses genetic variants as IVs to estimate the causal effect of an exposure on an outcome. Three assumptions define a valid genetic IV [436–438]. In the case where the exposure is gene expression, and the outcome is some phenotypic response such as antibody titre, the first assumption (IV1) is that the variant should be associated with the exposure as an eQTL. The term MR comes from an analogy to randomised controlled trials; meiotic segregation is largely independent of environmental confounders, so different eQTL alleles can be thought to randomly assign different "doses" of expression [436]. The second assumption (IV2) is that the variant is not associated with unmeasured confounders of the expression-phenotype association (e.g. population structure). The third assumption (IV3) is that the variant has no association with phenotype except through expression. Combined, these assumptions place expression as a complete mediator (vertical pleiotropy) of the effect of the eQTL on phenotype (Fig. 5.1). The effects of variant on expression and expression on phenotype can be estimated in the same sample, or in non-overlapping samples (two-sample MR [437, 438]). Two-sample MR can leverage existing large eQTL catalogues and helps mitigate weak-instrument bias, where eQTLs with weak effects on expression are used as IVs. The direction of bias in estimating the expression-phenotype effect is away from the null in single-sample MR, but towards the null in two-sample MR [439, 440]. A related family of methods, transcriptome-wide association studies (TWASs) [441], train predictive models of expression from eQTL data, then apply those models in GWAS cohorts to test the association of genetically-predicted expression with phenotype. TWAS methods have methodological similarities to two-sample MR [442].

Violating the assumptions of MR will likely lead to biased causal estimates. The most troublesome assumption is often IV3. If there is no temporal ordering of exposure and outcome, IV3 can be violated by reverse causation. For example, if evaluating the causal effect of day 7 post-vaccination gene expression on day 7 $CD4^+$ T cell abundance, an association between variant and expression might be mediated by the phenotype, cell abundance. If this is suspected, one can perform MR in the reverse direction if there are available instruments for the phenotype (bi-directional MR), or perform a statistical test of the directionality (MR Steiger) [436–438, 443]. IV3 can be violated by linkage if the eQTL does not actually have any effect on the phenotype at all, but simply is in linkage disequilibrium (LD) with another variant that does; and can also be violated by the existence of horizontal pleiotropy, where the effect of the variant on expression and phenotype are independent (Fig. 5.1). Colocalisation methods, as used in Chapter 3, can be used to test whether the same causal variant affects expression and phenotype, distinguishing pleiotropy from linkage. However, colocalisation is necessary but not sufficient for mediation, thus it does not distinguish mediation (vertical pleiotropy) from horizontal pleiotropy [437].

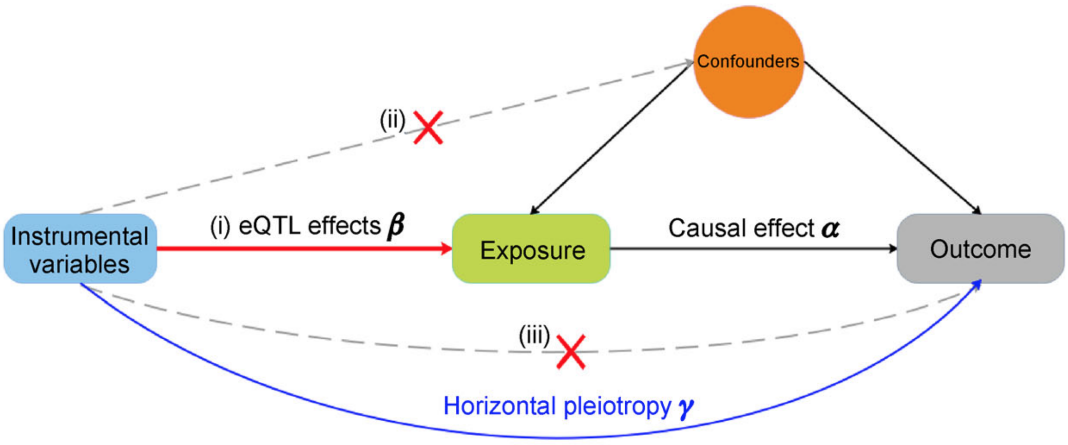

Figure 5.1: The three assumptions of MR. MR uses genetic IVs to estimate the causal effect α of an exposure (here, gene expression) on a phenotypic outcome, under three assumptions: (i) IV1: the variant is associated with the exposure (here, an eQTL with effect size β); (ii) IV2: the variant is not associated with any unmeasured confounders; (iii) IV3: the variant is not associated with the outcome except through exposure. The directionality of the arrows in the causal diagram are also assumed to hold. The blue arrow shows a horizontal pleiotropic effect of the variant on outcome, a violation of the IV3 assumption. Figure reprinted by permission from Springer Nature: Springer Nature, Quantitative Biology, Zhu *et al.* [442], © 2020.

Mediation analysis methods (e.g. CIT [263], Findr [444]) can be used to test for violations of IV3 by horizontal pleiotropy. They distinguish mediation from horizontal pleiotropy using comparison of causal models with different structures, but require individual level data, and are more susceptible to measurement error than MR [437, 443].

5.6 Triangulation

Triangulation refers to the use of methods that address the same question, but with different assumptions, biases, and limitations [445]. An example from this thesis appears in Chapter 3, combining DGE, between-individual reQTL mapping, and colocalisation—and pending validation by within-individual allele-specific expression (ASE)—to propose mechanisms behind changes in the genetic architecture of immune gene expression after vaccination. As discussed above, MR, colocalisation, and mediation analysis can be seen as complementary methods for triangulating the causal relationships between variant, exposure, and outcome. Taylor *et al.* [446] and Zheng *et al.* [447] exemplify how these methods can be combined in practice for genetic instruments, molecular exposures, and molecular outcomes. A combination of methods addresses limitations that cannot be solved by increasing sample size. Triangulation will be critical in moving from a descriptive to a mechanistic understanding of immune response to perturbations.

5.7 Concluding remarks

It has now been almost two decades since the completion of the Human Genome Project and the conception of systems biology, and almost fifteen years since the first GWASs and systems immunology studies. High-throughput profiling, complex algorithms, and big data are the new normal, yet the classical principles of perturbation and observation are alive and well. The

projects in this thesis come in the wake of these monumental achievements, yet still lie at the beginning of a long road leading towards a fuller understanding of our immune system.

The goal must be to not only observe the immune response to perturbation, but to be able to predict it, and to understand the causal relationships within the immune system that will ultimately guide the rational design and administration of vaccines and drugs. For this, we need study designs and analysis strategies for detecting robust and replicable associations with sensible response phenotypes. We need technologies that quantify the immune system with great richness and resolution, yet remain affordable enough to do so without sacrificing sample size. We need triangulation via multiple lines of evidence, requiring both confluence of methodology and collaboration of minds. The road from perturbation to understanding is a long one indeed, but it shall be a road paved by good science.