Chapter 5

Discussion

This dissertation describes three projects that explore different aspects of IBD genetics and pharmacogenetics of therapies used to treat IBD. In essence, all three were association studies of array genotyping or sequencing data. However, each posed unique challenges. The anti-TNF immunogenicity project, described in Chapter 2, required a non-standard genome-wide proportional hazards analysis followed by a scrupulous examination of the only significant association in order to understand which HLA allele it maps to and how it influences immunogenicity across different treatment regimes. In Chapter 3, I attempted to uncover the genetic variation associated with thiopurine-induced liver damage. While this project did not result in any robust associations, it underscores the importance of considering the quality of the dataset, data normalisation, and sample size when conducting GWAS. Lastly, in Chapter 4, I discuss the sample quality control and the initial association analysis of a large whole-genome sequencing dataset – IBD 15x. The scale of WGS datasets poses new computational challenges, which had to be addressed to enable further analyses. In addition, rare variant association studies require stringent quality control to avoid spurious genetic associations.

More than ten years since the first genome-wide association study [192], the genetics of IBD is far from being 'solved'. Large scale GWAS have demonstrated the genetic complexity of the disorder, underpinned by both the number of the associated loci and the complexity of resolving them down to a single variant and gene. Shortly thereafter, the GWAS techniques, and often the same datasets, were applied to study a variety of IBD-related traits, including disease progression, phenotype heterogeneity, and drug response. Below, I provide an

overview of several projects that use the techniques developed for identifying and elucidating trait-associated variation in order to understand novel aspects of the genetics of IBD.

5.1 Longitudinal studies for drug response leveraging expression data

Longitudinal studies follow the participants over a prolonged period of time, recording events of interest (e.g., treatment complication, remission, flareup). In addition, longitudinal studies often include a perturbation event at the beginning of the observation period (e.g., start of treatment). Such studies are widely used in epidemiological research and clinical trials, yet remain quite novel in the field of complex disease genetics.

The association between HLA-DQA1*05 and immunogenicity was, in part, established due to the longitudinal design of the PANTS study: in a purely case-control setting the association just about passed the genome-wide significance threshold. However, when performing a time to event analysis using the Cox proportional hazards regression, the association became much more robust. Additional statistical power allowed me to investigate the effects of the allele across different treatment regimes and to identify the non-additive nature of the association. As I will discuss in Section 5.3, for pharmacogenetic GWAS this approach is rarely used due to the difficulty of collecting longitudinal cohorts at scale.

However, smaller scale longitudinal studies (e.g., clinical trials) can leverage a combination of genotyping and gene expression or microbiome data to study the biological processes behind drug response.

Gene expression analysis is used to quantify the level of gene product across all genes. Gene expression was previously measured using microarrays, which have now largely been superseded by RNA-seq and a plethora of new single cell sequencing methods. Gene expression is measured at various points in time, including before the treatment is administered for the first time (base expression).

Once the outcome of the trial is known (e.g., responders versus non-responders, normal versus adverse drug reaction), longitudinal gene expression data can be analysed to derive response signatures that associate the expression at base level to the measured outcome. The

composition of the expression signatures can be investigated to understand which individual genes are up- and downregulated, providing additional insights into the biology of drug response.

Furthermore, eQTL mapping can be utilised to understand the role of genetic variation in drug response. Expression of individual genes is mapped to genetic variants in close (cis-EQTLs) or distant proximity (trans-eQTLs) from the gene. When the study utilises a longitudinal design, it is possible to map eQTLs at multiple time points; of particular interest are the eQTLs that exhibit a differential magnitude of effect across time points as they point out the likely mechanisms behind drug response.

RNA-seq at different time points was recently performed for around 400 individuals from the PANTS study – 200 responders and 200 non-responders. My colleague will shortly start analysing these data.

5.2 Host-microbiome interactions

IBD, being a disorder of the gastrointestinal tract, has long thought to be associated with changes in the gut microbiome. However, these changes are not well characterised. For example, dysbiosis (microbial imbalance) is frequently reported amongst IBD patients. However, no single microorganism has been consistently reported as being the one that dominates this imbalance [100]. Govers et al. describe the presence of dysbiosis amongst treatment-naive CD patients, suggesting that it is not entirely driven by microbiome alternations caused by therapies [66]. At the same time, it is unknown whether dysbiosis is caused by some of the symptoms of IBD themselves (e.g., diarrhoea), or driven by genetic variation, or is in fact itself a potential 'trigger' of the disease. Several approaches currently used in human disease genetics can be used to elucidate the causal relationship between the the host genetic variation, microbiome and IBD.

QTL mapping techniques can be used to identify genetic variants associated with the abundance levels of specific microorganisms. This analysis should be followed by a colocalization comparison with known IBD loci. Colocalization of the known IBD risk variants with the microbiome eQTLs would suggest that the bacterial makeup in the gut is partially driven by host genetics. One of the advantages of this approach is that the eQTLs can be mapped in the large-scale non-IBD microbiome datasets.

Sanna et al. [162] have used the two-sample bidirectional Mendelian randomisation technique to evaluate the causal relationship between host genetics, production of SCFA butyrate and insulin response, establishing a causal link. Similar approaches can be applied to IBD.

Recent work by Zimmermann et al. [199] indicates that microbiome-encoded enzymes can influence the metabolism of a broad variety of drugs. Future IBD pharmacogenetic studies can perhaps evaluate whether host genetics influence the microbiome composition, thereby modifying the drug metabolism, leading to drug non-response and adverse drug reactions.

5.3 Extending anti-TNF pharmacogenetic analysis

The association study of the PANTS cohort has resulted in the first known robust genetic association for immunogenicity to anti-TNF. While this was not the first attempt to do this, the previous studies were, arguably, subject to several methodological issues: small sample size ($N < 500$); or a targeted approach (e.g. specific HLA alleles, TNF); or the lack of self- or external replication.

The lessons learnt during the early stages of complex disease and trait GWAS still apply to pharmacogenetics – 'good enough' phenotyping combined with a decent sample size and adequate genotyping quality is more optimal than maximising either of the three components.

At the current stage of the field, pharmacogenetics appears to be limited largely by phenotype availability. While the GWAS of many complex diseases are now exceeding several hundreds of thousands of disease-affected subjects, genetic association studies of drug response for these conditions rarely exceed a thousand samples. The analysis described in the PANTS was based on a clinical trial that ran between 2013 and 2019 and was managed by 400 principal investigators and nurses across 150 research centres. While it is possible to meta-analyse these data with other anti-TNF trial cohorts in the future, it is evident that increasing the sample size by an order of magnitude would require finding alternative approaches for phenotype collection. One approach that could be worth exploring is to leverage the retrospective data available in bioresources and biobanks.

5.3.1 Analysing retrospective data from the NIHR IBD BioResource

The IBD BioResource project [143] has so far recruited 27,000 IBD patients. Based on the medication questionnaire, ∼8,000 of them have undergone treatment with anti-TNF and approximately an additional 800 have been treated with vedolizumab (a biologic drug targeting integrin $\alpha_4\beta_7$). The patients have consented to participate in research studies.

The BioResource participants are asked to fill out questionnaires on a variety of topics, including diet, lifestyle, and drug response. The drug response questionnaire captures the approximate dates (month and year) of when they have started and finished certain therapies – such as thiopurines, anti-TNF, and vedolizumab – and whether these have worked. The questionnaire responses can be used to reconstruct the approximate treatment timelines that can be analysed with a survival model. In addition, a blood sample is taken upon enrolment into BioResource. While this does not allow the measurement of antibodies longitudinally (like in PANTS), the analysis of the PANTS data indicates that patients who develop immunogenicity will maintain it for multiple observation time-points. The difference between the immunogenicity time-point and the blood draw may reduce the power of the time-to-even analyses. However, when I analysed the retrospective replication cohort used in the PANTS project (using the delta between start date and sampling date as time to event/censoring), the replication results were very consistent with the main prospective cohort. Therefore, I believe that the BioResource dataset can be used to study both the anti-TNF response and immunogenicity.

I prioritised the patients on anti-TNF for inclusion in our 15x and WES cohorts and ∼3,000 of them were already sequenced. In addition, efforts to genotype all 25,000 patients are ongoing. These data can be used for future anti-TNF pharmacogenetic projects. Increasing the sample size ∼7x (assuming all 8,000 anti-TNF patients get genotyped) will likely yield additional associations between genetic variation and response to anti-TNF, further improving our understanding of therapeutic response.

Some preliminary work was carried out to evaluate the feasibility of this project. I trialled imputing HLA alleles from sequencing data via HLA-LA. Compared to the genotype-based imputation techniques (like HIBAG), HLA-LA is a graph-based method that uses sequence-level data (e.g., aligned BAMs or CRAMs) to achieve higher imputation resolution and accuracy. This method imputes the HLA alleles at the G group-level resolution which matches the resolution of clinical HLA typing (compared to the more crude 2- and 4-digit

level imputation that I used in Chapter 2). Increased accuracy may enable the identification of novel associations within the HLA, though it appears that the immunogenicity is fully driven by the HLA-DQA1*05.

When examining an early release of the BioResource anti-TNF phenotype data, it became apparent that the real-world treatment histories are much more heterogeneous than those of the patients in the PANTS cohort: patients change the the types of biologic therapies, start and stop immunomodulators, and have gaps in the treatment history. During the analysis of the PANTS dataset, I used the right-censored Cox proportional-hazards regression to analyse time to immunogenicity. The model used a number of covariates, including immunomodulator usage, type of the anti-TNF drug (infliximab versus adalimumab), assuming they remain constant throughout the enrolment. In order to use the BioResource data to perform time-to-even analyses, one would need to control for changing covariates during the observation period. Time-varying survival regression can be used to achieve this [196]. To the best of my knowledge, no current GWAS survival analysis tools support it. In order to address this issue, I have written a prototype software package that is able to perform the regular Cox proportional-hazard and the time-varying regressions on the genotype data.

5.3.2 Prescription records from the UK Biobank

The UK Biobank (UKBB) is a population-scale cohort that has recruited 500,000 individuals from all across the United Kingdom. The participants were genotyped, and are currently being whole-exome (first 50,000 samples available) and whole-genome sequenced. Recently, anonymised GP prescription records of 222,000 individuals were released. Although the GPs are unlikely to prescribe anti-TNF therapy themselves, their records *should* contain records of all drug prescriptions that the patient receives. The average age of the UKBB participants was 57 years upon recruitment, meaning that the cohort is enriched for individuals suffering from IBD, rheumatoid arthritis, and other conditions for treatment of which anti-TNF is used.

It would be interesting to analyse whether there are any genetic variants associated with shorter prescription length of anti-TNF. A reasonable assumption can be made that the patients that are prescribed anti-TNF for a couple of months did not respond to the treatment. I acknowledge that this phenotype is quite heterogeneous and may result in spurious associations. If any associations for ‘short anti-TNF prescription time’ are uncovered, these should be replicated in a well-phenotyped clinical trial cohort like PANTS.

Similar to anti-TNF, the same approach can be adopted in the study of thiopurine-induced myelosuppression and liver injury, where the typical time of the adverse reaction is known (e.g., ten weeks for hepatocellular TILI).

Overall, I believe that future pharmacogenetic studies will have to leverage retrospective data from both specialised bioresourses and population-scale datasets. Here, I have described two of such datasets. In addition, one could leverage insurance records (especially, in the US) and national drug prescription databases (such as the one that operates in Finland). Validation of the results from such proxy phenotypes could be challenging, which means that clinical trial datasets (like PANTS) will be required for validation.