CHAPTER 8

Conclusion, discussion and future perspectives
CONCLUSION

Our first goal of this thesis was to increase the knowledge of the genetic basis of the inflammatory bowel diseases (IBD) by identifying new genetic risk loci. We did this in multiple ways: at the time the first genome-wide association studies (GWAS) were published, there was a need for replication of their findings in an independent cohort. In chapter 2 we could confirm two out of three loci that were previously reported in an ulcerative colitis (UC) GWAS in our independent Dutch cohort, one locus harbouring HNF4A and one harbouring CDH1. We did not replicate the third locus, containing candidate gene LAMB1. We suggest this may be due to phenotype differences compared to the identification cohort, our cohort contained more severely affected patients. After generating more extensive data of this cohort, none of the single nucleotide polymorphisms (SNPs) in this locus turned out to be associated with UC, which underscores that this SNP covers the locus well and no technical error was made. The LAMB1 locus has been established in other cohorts in later years. All three genes play roles in the integrity of the intestinal barrier.

Next, in chapter 3 we performed a large trans-ethnic meta-analysis including almost 100,000 individuals from European, Indian, Iranian and East-Asian descent. We identified 38 new genetic risk loci for Crohn’s disease (CD), UC or IBD, increasing the number of independently associated SNPs to 231 in 200 genetic loci. Furthermore, we demonstrated that there is extensive sharing of genetic risk across populations, showing that combining multiple populations is helpful for increasing much needed statistical power. There are some striking exceptions for a few loci which can be attributed to differences in occurrence or effect size of these loci in different populations.

In the period before the large GWAS meta-analyses, when sample sizes were relatively small (meaning low statistical power) and costs for genotyping were relatively high, it was important to select SNPs for testing for association. In chapter 4 we developed a new method to prioritise SNPs that did not reach the genome-wide significant threshold in a published GWAS for replication. Our method is based on the known effect of these SNPs on gene expression, hypothesising that SNPs with an effect on gene expression were more likely to be involved in disease pathogenesis than SNPs with no such effect. In this way, we identified two new CD loci, harbouring candidate genes UBE2L3 and BCL3.

Because it is expected that genetic risk factors affect gene expression and we see an enrichment of eQTLs among disease associated SNPs we wanted to link genetic risk to gene expression by focusing on a defined pathway both genetically and functionally strongly implicated in IBD: the Th17/IL23 pathway. In chapter 5 we combined the genetic risk of 10 Th17/IL23 pathway associated SNPs and correlated that to gene expression of 9 representative Th17/IL23 genes in peripheral blood mononuclear cells (PBMCs) of 80 IBD patients and 40 controls. We observed a lower baseline expression of IL6 in IBD patients compared to controls. Upon stimulation of T cells, we saw a smaller increase in IL23A gene expression in cases compared to controls, and a larger decrease in RORC gene expression in IBD patients compared to controls. We did not find a correlation between genetic risk load and gene expression. One explanation for this lack of association would be that alterations in gene expression are cell specific and becomes overshadowed in a mixed collection of cells like PBMCs.

In order to expand the knowledge of the genetic background shared with other diseases, we first addressed the overlap between IBD and its extraintestinal manifestations (EIM), because we believe they share pathogenetic links. We demonstrated a large overlap in genetic risk loci between IBD and the EIM in chapter 6. Furthermore, based on co-expression of genes and protein-protein interactions, we showed that candidate genes residing in associated loci for the different diseases work together in pathways.

Finally, we aimed to identify genetic risk loci in a disease occurring after allogeneic
hematopoietic cell transplantation and clinically resembling Crohn’s disease: gastro-intestinal graft-vs.-host disease. We found no significant associations in our small cohort, but found suggestive signals in interesting loci like JAK2, IL2 and IL2RA. This work is presented in Chapter 7.

DISCUSSION AND FUTURE PERSPECTIVES

The wide variety of the IBD disease spectrum and the complexity of disease pathogenesis make the disease both a heaven and a hell for research purposes. On the one hand, we can study the disease from many perspectives and use multiple techniques that evolve along the way. On the other hand, we are hampered by the many different factors influencing disease pathogenesis, making it hard to distinguish the importance of each separate component and unravel how different components influence each other. Although genetic research in IBD has delivered many results, there are many more questions remaining in our research field regarding identification of genetic risk loci and causal variants, correlating genetic risk factors to biological processes and pathways that eventually lead to disease, and the influence of environmental factors and the microbiome on this process. My thesis has focused on different subsets of this field, therefore I will focus my discussion on several of these components, highlighting things I have learned or that have fascinated me over the years.

Identification of genetic risk loci

Genetic studies in IBD have been extremely successful and the insights they have provided are undeniable. At the start of my PhD the first GWAS in IBD were published, and needed replication in independent cohorts, to which we contributed with the work presented in Chapter 2. Furthermore we provided a strategy to prioritise candidate SNPs based on eQTL effects to test for association (Chapter 4). Our biggest contribution to the identification of genetic risk loci was by identifying 38 new IBD risk loci in the transethnic meta-analysis, raising the total number to 200 genetic loci (Chapter 3). The disease variance explained based on these genetic risk factors went to 13.1% for Crohn’s disease and 8.2% for ulcerative colitis. Although we are still not able to explain the full genetic background of IBD, we provided biological meaning by reinforcing the importance of the autophagy pathway by identifying candidate gene ATG4B, which plays a crucial role in this process. Furthermore, we can now implicate all three components of T cell activation (TCR ligation, co-stimulation and interleukin-2 signalling) in disease pathogenesis by identifying LY75, CD28, CCL20, NFKBIZ, AHR and NFATC1.

One could think that the maximum number of associated genetic loci will once be reached. However, it is now clear that the number of IBD risk loci still has the potential to grow, as long as we are able to increase statistical power with extra samples to discover new loci with smaller effect sizes. The current number of IBD risk loci is 241 and a large new GWAS meta-analysis is underway within the Internation IBD Genetics Consortium. Although the genetic variance explained does not increase much with new loci, finding new (candidate) genes keeps enhancing our understanding of the pathogenesis of IBD by reinforcing existing or discovering new biological pathways that underlie disease pathogenesis.

Performing genetic studies has become more generally available over the last decade. Genotyping arrays that enable us to study multiple genetic loci at the same time have become cheaper, making it possible to include a large number of samples and making genetic studies more available for study groups that focus on non-Caucasian populations. Furthermore, the experience of performing genetic studies has grown by cooperating and sharing knowledge. Our trans-ethnic meta-analysis showcases the result of such cooperation. By combining multiple ethnic populations we were able to demonstrate for the first time that most of the genetic risk is shared across populations. Because we have more statistical power with this large dataset, we were able to identify new IBD risk loci. I believe this is great progress, but I
also think that the real added value of incorporating multiple ethnic populations will be beyond locus identification, and can be helpful in identifying causal variants and unravelling the effect of environmental factors on disease pathogenesis. It is clear that the overall focus of genetic research is no longer on locus identification but on discovering the implications of genetic risk factors.

The need for large sample sizes results from the fact that IBD has a complex pathogenesis with multiple genetic and non-genetic factors. To elucidate one of these factors with limited effect, studies require large numbers. This is not only the case for genetic studies focusing on locus identification, but it also hampers other fields of IBD research like identification of causal variants and gene expression studies. It might be rewarding to use patients with extreme phenotypes (severely affected individuals, individuals with early onset disease) where one would expect that individual risk factors play larger roles compared to mildly affected patients.

Furthermore, IBD has a wide variety of disease localisation and severity. Most of the genetic studies have applied a simple case-control strategy comparing CD or UC affected patients with healthy controls. Cleynen et al. went beyond this and undertook an interesting study to investigate the correlation between genetic risk factors and disease location and disease behaviour. They identified associations between SNPs in the MST1, NOD2 and MHC loci and age at diagnosis and disease location of CD and UC. They also showed that genetic risk scores differ between ileal CD, ileocolonic CD, colonic CD and UC. Moreover, they found that genetic risk loci associated with UC correlate better with CD localisation (colonic vs non-colonic) than risk loci associated with CD, suggesting that genetic variation contributes significantly to disease location. Also, genetic variation contributing to disease progression in CD has been described. Because of the wide variety in the phenotype, I believe it would be very interesting to investigate further how genetic factors lead to particular phenotypes or disease behaviour. With the sample sizes we are accustomed to, we should be able to have enough statistical power to detect correlations with a subset of individuals. The real challenge is to deeply phenotype individuals, which is a challenge many groups are willing to take on.

**Identification of causal variants**

In order to fully comprehend how associated genetic risk loci contribute to disease pathogenesis it is crucial to identify causal variants. Multiple approaches have been applied. One of them is fine-mapping loci by using SNP data; however, this is complicated by the strong correlation structure of the genome (linkage disequilibrium – LD). The Immunochip accommodates fine-mapping by the dense coverage of relevant loci which allows researches to look for subtle differences of association within the locus. Huang et al. performed fine-mapping for 94 IBD signals in high density loci and found 139 independent signals. For each of these signals a set of SNPs (credible sets) was selected that was > 95% likely to contain the causal variant. For 18 signals, only one variant was selected, 23 had 2–5 variants that were likely to contain the causal variant, for only 23 signals the set contained more than 50 SNPs. There is an enrichment of functional elements. Furthermore, by identifying single or functional variants that are likely to underlie disease association, the number of candidate genes can be reduced. An approach based on conditioning analysis on association signals of (predominantly) classical HLA-alleles was used to fine-map the HLA region, a region with a large effect in predominantly ulcerative colitis but with very strong LD. In this study, Goyette et al. showed amongst others things that the disease variance explained can be increased by using causal variants (classical HLA alleles) instead of tag SNPs associations, which underscores the idea that part of the ‘hidden heritability’ phenomena is explained by incomplete covering of the association signal by the tag SNP. Another approach is sequencing interesting regions, for which multiple approaches can be used. Pooled targeted sequencing identified rare variants in for example MUC2, CARD9, IL18RAP and PTPN22.
Over the years, cost of sequencing has decreased dramatically, permitting researchers to increase sample size and genome coverage. Although some newly associated rare variants are being identified, pointing out the candidate gene in the locus relevant for follow up, the number of identified causal variants remains low and does not (yet) clarify the genetic base of the majority of associated loci.

I believe we can benefit greatly from the use of multiple ethnic populations in fine-mapping analyses. LD structures can differ between populations and especially populations with smaller LD blocks (e.g. African ancestry) can aid to downgrade large stretches of LD. This method has proven to be successful in traits like adiposity and fasting glucose and insulin levels\(^8\,^9\) and would be very interesting to perform on a large scale in IBD.

**Population specific loci?**

Interestingly, in the transethnic meta-analysis we have observed genetic differences between populations, some strongly associated ‘European’ loci don’t seem to play an important role in the Asian population and some loci contribute largely to the variance explained in the Asian population but have limited effect in the European population. Some of these can be explained by differences in occurrence of SNPs across populations and some by differences in effect size. A striking example is NOD2, the strongest associated locus in European Crohn’s disease. All three previously described causal variants are monomorphic, in other words non-existing, in the Asian population. Associations with other NOD2 markers have been described in Asian populations, for example JW1 and P268S; however, they are performed in small cohorts and/or could not be replicated.\(^10\,^11\) We could also not replicate the findings of these variants in our data, nor found any strong signal in the locus (lowest p-value 7.18 × 10\(^{-4}\)). The other autophagy loci (IRGM and ATG16L1) both show lower association signals in the Asian population compared to the European; for the IRGM SNP this can be attributed to a difference in effect size. Other SNPs in this locus show some convincing signals in our largest cohort, the East Asian, 138 kb downstream of the European top SNP (lowest p-value at rs2287720 7.66 × 10\(^{-8}\)). For the ATG16L1 SNP, the difference can be attributed to differences in both effect size and allele frequency. Furthermore, in the complete ATG16L1 locus only nominal signals are found (lowest p-value at rs6758317 0.0016). Of course it is too early to speculate that autophagy would not play a role in Asian IBD. We lack functional studies on autophagy in Asian populations, so we don’t know the role of autophagy or genetic variation in autophagy associated loci in Asian IBD, which would be very interesting to follow up.

In contrast, the locus containing TNFSF8/ TNFSF15 has a much larger effect size in the East Asian population compared to the European population (for example rs4246905 European OR = 1.15 and East Asian OR = 1.75). It might be more rewarding investigating a causal variant in this locus in the East Asian population rather than in the European population. Of course we need to take into account that possible (rare) causal variants might be population specific, but finding causal variants in a specific population could provide insight into overall disease pathogenesis.

**From genetic risk to biological implication**

In order to really understand disease pathogenesis we need to correlate genetic risk factors with biological processes to try and understand how these processes get disrupted. As mentioned before, this can be done in many different ways. One important factor is how genetic risk factors influence gene expression. Many GWAS associated loci harbour eQTLs, but for most GWAS hits an eQTL effect is not described. We have investigated the influence of multiple known IBD risk factors on gene expression of the Th17/IL23 pathway that where shown to be very relevant for disease pathogenesis from both a genetic perspective (enriched in genetic risk factors) and a biological perspective. Furthermore, one of the treatment options is directly targeting this pathway (Ustekinumab; targeted against IL23).
The results of this experiment are described in chapter 5. Although we found some alterations in gene expression between cases and controls in stimulated and unstimulated PBMCs, we found no correlation between genetic risk and altered gene expression. This was not the result we had hoped for. There are several possible reasons that would explain why we did not find more correlations, one of course being statistical power. One other important reason would be that alterations in gene expression are very cell specific and because we used PBMCs, which is a mix of cells, we were not able to pick up the relevant signal. This is underscored by the fact that we did not find any relation between genetic risk and gene expression in a larger database of PBMCs (not IBD patients). This was also found by Huang et al. when they searched for enrichment of eQTLs among their credible sets after finemapping: they did not find enrichment of eQTLs in PBMCs; however, when they analysed eQTLs in specific cell types (CD4+ cells or ileal cells) they did find enrichment of eQTLs (Huang). Therefore, I believe it is extremely important to use specific cells in relevant tissue (for example T-cells in inflamed colon) and in this way investigate how genetic risk factors influence gene expression. Luckily, with advances in techniques like cell sorting and single cell mRNA sequencing, such an effort can be made. Other ways to uncover functional consequences of genetic variation are identifying protein coding variants, disruption of enhancers, promoters and transcription-factor binding sites. In the last decade much has been discovered in this area.12

The next step is how genes work together in pathways. We know several relevant pathways for IBD pathogenesis such as the Th17/IL23R pathway and autophagy and we are currently at the point that we are able to identify pathways relevant for disease pathogenesis, but many questions remain unanswered: how does genetic variation influence the function of such a pathway? What is the contribution of a certain pathway to disease pathogenesis? Does this differ between subgroups of patients (certain phenotypes) or between populations? How do biological pathways work together? How do other factors (for example environmental factors) influence pathways? Which pathways are shared between CD and UC, and other diseases and which are specific for a certain trait? In order to answer all these questions we will eventually need functional studies incorporating multiple layers of data (multi-omics).

**Shared genetics and pathogenesis**

It has been known that the disease spectrum of IBD comprises extra-intestinal manifestations and IBD co-occurs with other immune mediated diseases. A logical explanation would be that these diseases share genetic risk factors and biological pathways in which they contribute to disease pathogenesis. It is estimated that around 70% of IBD risk loci are shared with other immune mediated diseases like celiac disease, ankylosing spondylitis, primary sclerosing cholangitis and type 1 diabetes mellitus.13 In chapter 6, we investigated the overlap between IBD and its extra-intestinal manifestations and found substantial overlap in genetic factors and biological pathways. This overlap was also described well in a recent review.14 On the other hand there must be factors that drive the occurrence of a specific EIM or co-occurrence of another immune mediated disease, and not lead to another EIM of immune mediated disease. In line with this is the co-occurrence of PSC and IBD: 75% of PSC patients also have a form of IBD; however, up until now the genetic overlap is limited.15 Therefore, there must be specific factors that drive the development of this disease phenotype in contrast to developing for example IBD only.

To further investigate the overlapping pathogenesis it is possible to take several approaches. On the one hand, we can combine IBD data with data from related disorders to search for overlapping signals. On the other hand, if we are able to deeply phenotype our IBD cases, we can focus on patients who develop (certain) EIM and investigate what genetic background or other factor drives a particular disease course.

In chapter 7, we focus our attention on the overlap with graft-versus-host disease after allogenic stemcell transplantation, a disease
in which the gastrointestinal form resembles Crohn's disease. Again we found substantial overlap. Interestingly, we found a suggestive signal at JAK2 in GVHD, a protein that is targeted by Tofacitinib, a recently registered treatment option in IBD. Because of the overlap between IBD and many other diseases, it is logical that we can learn and benefit from each other. For example, functional work done for another disease can also aid in our understanding of IBD. Furthermore, registered drugs for other diseases can also be interesting treatment options for IBD (drug repositioning).

**Exposome and microbiome**
As mentioned in the introduction a lot of research has been done on environmental factors influencing IBD (exposome) and more recently the role of the microbiome has been highlighted. The interaction between the genetic background, the exposome and microbiome is not yet understood. I believe that including multiple different populations can significantly aid in answering this question. Where environmental factors differ between populations, we have shown that genetic background greatly overlaps, providing great opportunities for further research.

We do need to take into account that the influence of environmental factors might be different in other populations compared to European studies. For example, cigarette smoking increases the risk of Crohn's disease in Western countries, but not in Eastern Asian countries and among immigrants from the Middle East. In developing countries, measures of hygiene have not demonstrated the inverse association reported in the West and have in fact been associated with an increased risk of UC.

Intriguing is the overall first occurrence of UC and later occurrence of CD, both in the past in European and now in industrialising countries. This leads to the question why: do environmental factors drive this difference? In European populations the heritability of UC is lower compared to CD, if this is the same for industrialising populations, a greater contribution of environmental factors to disease pathogenesis of UC compared to CD would explain this. We don’t know the exact mechanisms how environmental factors contribute to disease pathogenesis, and we know even less about differences in exposome for UC and CD. One exception is smoking, which, in European populations increases the risk of CD and quitting smoking increases UC. Part of industrialising is an increasing number of smokers; in our Asian cohort we also saw a higher occurrence of smokers. One would expect to see a higher incidence of CD. Another hypothesis is that the genetic background of UC in for example Asian populations would resemble European UC more than in CD. Other possible explanations would be a better recognition of UC than CD, because clinical aspects are more pronounced. Nonetheless, a thorough investigation of the genetic basis and the interaction with the exposome between populations will be needed.

There is still limited access to non-European cohorts, so efforts will have to be made to include more non-European cases and controls. There are some factors that we have to account for when using different populations: (I) genotyping platforms were primarily designed to cover the ‘European genome’; therefore genetic variation in other population might not be covered well. A striking example is the Immunochip, where one fifth of the technically successful genotyped SNPs are monomorphic in our East Asian cohort. (II) imputing genotypes is a very successful method of improving genotyping coverage, but because we are dealing with population specific variation, the reference genomes used for imputation need to originate from the same population. (III) the statistical method used needs to take into account genetic differences between populations. (IV) differences in cohort size. (V) differences in clinical aspects

**Systems biology**
It is clear that in order to provide answers related to disease pathogenesis, behaviour and response to therapy we will need many different types of data (omics): different types of genetic data (genotype data, sequence data), gene expression data in relevant tissues, microbiome data and detailed phenotype data including
information about the disease, other diseases, family history and detailed environmental exposure. Ideally, all data should originate from the same samples. In our own centre (UMCG), such data collection is already being undertaken: for 1,000 IBD patients, different forms of genetic data (GWAS, Immunochip, whole exome sequence data), gene expression data (RNAseq on intestinal biopsies) and microbiome data (16S and full sequence data) are being collected in conjunction with clinical, exposome and dietary data. Furthermore, we will need robust statistical approaches that can deal with all these layers of data and provide answers to the question how the different omics work together and lead to the disease. Moreover, over the last few years different molecular profiles have been associated with a particular IBD subphenotype, like a specific localisation or disease behaviour. Such a dataset provides opportunities to fully investigate such questions.

**Genetic information currently used in clinical practice**

In addition to understanding the pathogenesis of IBD, we can use genetic information in daily practice. Genetic risk models are currently not useful in order to distinguish IBD affected individuals from non-IBD affected individuals. This is probably due to the fact that many more non-genetic factors contribute to disease pathogenesis. If we get a better grasp of these other factors, as stated above, we might be able to better predict IBD occurrence, and perhaps disease behaviour and progression and response to therapy.

Another purpose genetic information can serve is found in the field of pharmacogenetics, where genetic factors predict the effect or side effects of drug therapies in IBD, as has been described for the development of thiopurine induced myelotoxicity with certain TPMT and NUDT15 genotypes, the development of thiopurine induced pancreatitis with certain HLA variants and anti-TNF antibody development associated with both genetic variants and gene expression profiles.

**Final remarks**

After successfully identifying numerous IBD genetic risk loci, the research focus will shift towards finding underlying causal variants and investigating the functional consequences for disease pathogenesis. Furthermore, the influence of environmental factors and the microbiome will have to be clarified. I believe characterising multiple ethnic populations and incorporating information from other related (immune mediated) diseases can aid in achieving this. The integration of multiple omics will be necessary to start to comprehend the complex interaction of the different components. I also believe that the attention should focus on subphenotypic aspects like disease localisation and behaviour. In order to do this, detailed phenotype information will be needed. The scale of this challenge is incredible, making close collaboration necessary. However, the goal of understanding and predicting disease behaviour and offering a personalised therapeutic strategy for each IBD affected patient makes this tremendous effort worthwhile.

**REFERENCES**


