Chapter 1

General introduction
Inflammatory bowel disease

Inflammatory bowel disease (IBD) is a chronic inflammatory disorder in the gastrointestinal tract. There are two main subtypes of IBD, Crohn’s disease (CD) and ulcerative colitis (UC). CD affects any segment of the entire gastrointestinal tract while UC mainly affects colon. Patients with IBD are suffering from gastrointestinal complaints and reduced quality of life. Most of the patients are treated with anti-inflammatory medication, however, the current treatment is often ineffective and a curative therapy is lacking. Therefore, numerous efforts have been made to explore the IBD risk factors which can facilitate to improve the disease prevention and clinical treatment. The incidence of IBD is varied across geographic regions with the highest rates in the North American and European countries. In the Netherlands, around 40 individuals per 100,000 are newly diagnosed as IBD in each year. In contrast to western countries, IBD was rare in East Asia decades ago, but with a rapid growth in recent years because of “westernized” life style, e.g. the average annual percentage change (AAPC) of CD and UC is more than 4% in the southern parts of China in the last 20 years.

Genetics of IBD

Genetic susceptibility is an important factor for the development of the disease. Early studies from familial aggregation and co-occurrence in twins of IBD, followed by genetic linkage studies identifying a small group of highly penetrant genomic mutations have highlighted the genetic component of IBD. With the advance of high-throughput genotyping and sequencing technologies, genome-wide association studies (GWAS) have become a routine approach to identify the associations between the genotype alleles and disease phenotype. The largest IBD GWAS (>25,000 cases) so far has identified more than 240 genomic loci which contains thousands of single-nucleotide polymorphisms (SNPs), which collectively explained 8~13% of IBD heritability. In addition, trans-ethnic association studies revealed that population ancestries enrich different genomic risk variants, for example, the variants of NOD2 are more dominant in Europeans while variants of TNFSF15 are more present in Asians. Collectively, these IBD genetics studies provides a good opportunity to dissect the population genetic susceptibility towards the disease etiology.

A well-known gene is NOD2 which encoding NOD-like receptors in the epithelial cell surface. Different genetic missense variants of NOD2 can lead to inability of the host to recognize the microbial sensing muramyl dipeptide (MDP) and subsequently cause dysregulation of innate immune response to the gut microbiota. An increased load of commensal intestinal bacteria has been observed in Nod2(-/-) mice, and the intestinal crypts isolated from these mice lose the ability to kill bacteria effectively. IL23 pathways play a critical role in T-helper cells development which contains a number of IBD associated variants coding the genes IL23R, IL12B, JAK2 and STAT3. For example, the SNP R381Q in IL23R has been shown to protect against IBD. The CD4+ and CD8+ T cells derived from R381Q human carries show decreased expansion which can potentially reduce the inflammation. IL10 is an anti-inflammatory cytokine in macrophages and T cells. Individuals with protein coding variants of IL10 have been associated with higher risk of CD (OR =1.46) in GWAS. In subsequent mice studies, IL10 gene-deficient mice have shown increased intestinal permeability and higher chance to develop colitis. IBD-associated variants in C1orf106, a gene participating cell adherence by interacting with cell-surface receptor ARF6, have also been confirmed by C1orf106−/− mice with defects in the intestinal epithelial cell barrier.

Black-box from genomic variants to IBD etiology

Despite the great achievement of genetics association studies, there is still a black-box of linking most of the identified genomic variants to the molecular mechanisms that underlie the disease etiology. Firstly, disease-associated loci identified from GWAS usually spans a genomic region. The region commonly contains many genes and genomic variants in strong linkage disequilibrium (LD), which makes it difficult to prioritize the most likely causal genes from the non-disease relevant genes only based on P values. This is even more problematic for validation purpose in downstream experiments. Secondly, less than ~30% of the variants from IBD GWAS are located in coding proteins regions, and therefore it is difficult to interpret their function. Even though genetic polymorphisms are located in non-coding regions, these can have regulatory effects by modifying the gene expressions. Thirdly, compared with monogenic disease, complex disease is assumed to be a result of the accumulation of multiple genetic risk factors. In addition, IBD etiology involves not only numerous factors but also interactions between each factor.

Link genetics to multi-omics: shedding the light on the black-box of IBD genetics

Recently, the advances in high-throughput data generation techniques and bioinformatics tools have brought new strategies and new insights in complex disease. The “multi-omics data” refers to multiple dimensions of...
large-scale biological data which allows us to comprehensively analyze the effect of disease risk factors\(^3\). Multi-omics data mainly include data reflecting gene expression, proteins, metabolites, gut microbiota and a variety of phenotypes (Box 1). Using multi-omics data can pinpoint how genomic variants affect downstream molecular traits\(^3\) (Figure 1A). A summary of linking genetic factors to multi-omics is the following:

<table>
<thead>
<tr>
<th>Categories</th>
<th>Biological perspective</th>
<th>Commonly used technologies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genomics</td>
<td>DNA sequence</td>
<td>Whole genome sequencing;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Whole exome sequencing;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Genotyping array</td>
</tr>
<tr>
<td>Transcriptomics</td>
<td>Gene expression</td>
<td>Bulk RNA sequencing;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Single-cell RNA sequencing</td>
</tr>
<tr>
<td>Proteomics</td>
<td>Protein level</td>
<td>Mass spectrometry-based</td>
</tr>
<tr>
<td></td>
<td></td>
<td>proteomics;</td>
</tr>
<tr>
<td>Metabolomics</td>
<td>Metabolites, molecules</td>
<td>mass spectrometry (MS)</td>
</tr>
<tr>
<td>Metagenomics</td>
<td>Microbiota</td>
<td>16s Sequencing</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Shotgun metagenomics</td>
</tr>
</tbody>
</table>

**Transcriptomics** refers to the whole RNA expression level of genes in a given biological sample and reflects the very beginning of the fundamental process from genome to final disease outcome. RNA sequencing technique on tissue (bulk RNA-seq) or cell level (single cell RNA-seq) enable us to associate genetic variants and gene expression levels, which is called expression quantitative trait loci (eQTL) mapping. eQTL method consists of two main categories based on the genomic distance between regulatory variants and affected genes. Variants located inside or closely (+/- 1 MB of the gene center) to target genes are called cis-eQTL while variants targeting distal genes called trans-eQTL. Studies have shown that almost one out of three IBD associated variants have a cis-eQTL effect which proved that eQTL mapping is an important approach to understand how genetic variants contribute to disease through gene expression level\(^3\).\(^2\).

**Proteomics** refers to the high-throughput profiling of proteins level. Most of studies focused on the proteins from blood circulations due to the stable cellular secretion which are usually explored as biomarkers in disease diagnosis. Proteins are produced in a more complex post-transcriptional process including translation and folding. Linking genetics to proteins provide an extra data layer of how genetic variants mediate the effect on disease which maybe not observed at gene expression level\(^3\). This method

**Figure 1.** A) The strategies to prioritize the disease associated SNPs/genes using multi-omics data. B) Illustration of context-specific eQTL effect. eQTL, expression quantitative trait loci; pQTL, protein quantitative trait loci; mQTL, metabolic quantitative trait loci; mbQTL, microbial quantitative trait loci.
is called protein quantitative trait loci (pQTL) mapping and also classified to cis-pQTL and trans-pQTL. Studying pQTL effect can identify genetics-dependent clinical biomarkers and novel drug targets and thereby adds key values in pharmaceutical applications\(^{35}\). Blood inflammatory proteins have been shown to be different between IBD and healthy controls\(^{34,36}\). However, there is little known of how genetics regulate proteins and its role in IBD mechanisms\(^{38}\).

**Metagenomics** refers to the whole microbial genomes data from samples. In early time, 16S sequencing has been a main technology to detect microbiota, which aims to capture the taxonomy information in a given sample through sequencing the conserved genomic regions of Bacteria and Archaea. Metagenomics, or whole genome shotgun metagenomics sequencing detect the whole microbial community genome, detecting both microbial taxa and predicted functions. In the fecal samples from patients with IBD, a decrease in microbial diversity, an increased expansion of pathogenic and opportunistic microbiota has been observed compared with healthy controls\(^{39,40,41}\), indicating abnormal host-microbiota interactions in disease conditions. Microbial quantitative trait loci (mbQTL) studies aim to integrate host genomics with metagenomics/16S data and explore how the host genetic variants influence gut microbiota. These studies can fill the gap between IBD associated genes and the gut microbial taxa and functions, and help us further understand the role of host-microbiota interactions in disease pathology.

**Metabolomics** refers to measuring the metabolites from chemical processes which can be conducted by mass spectroscopy (MS). Blood metabolites mainly reflect host circulating cellular status. In patients with IBD, a range of blood metabolites related to lipid, amino acid and tricarboxylic acid has been reported to be altered compared with controls\(^{42}\). Fecal metabolites are more complex and it can be a reflection of interaction between host and gut microbiota, for example, microbes can modulate the primary bile acids for intestinal lipid digestion. In recent fecal metabolomics studies, dysregulation of short chain fatty acids and bile acids has been observed in IBD compared with controls, which indicates these fecal metabolites exert beneficial effects protecting against diseases\(^{43}\). Studying fecal metabolomics data is a complementary approach for metagenomics data and provides an extra window of the microbes’ actual activity. Fecal metabolites are also influenced by host genetics. By performing metabolic quantitative trait loci (mQTL) mapping in populations, SNPs in gene NAT2 are found to be associated with the abundance of metabolites derived from coffee intake\(^{44}\). Therefore, combining the metabolomics data with host genetics as well as gut microbiota gives an extra view of host-microbiota interactions.

Integrating genetics and omics data in IBD study, also known as systems genetics approach, has been a powerful method to unravel the molecular mechanisms from genotype to phenotype. However, there was few comprehensive studies using omics data in IBD before my PhD and they were limited by: 1) only one or two types of omics data, such as purely eQTL analysis which only assessed the genetic effects at gene expressions, 2) relatively small sample size, this particularly influenced the discrimination between disease heterogeneity effects, including disease locations, inflammation status, intestinal surgery and the use of medications, 3) inconsistent measurement across studies, for example, different proteomics measurement platforms lead to a relatively low replication rate ~65%\(^{45}\) in pQTL studies, 4) not right context, many studies have supported that genetics effect is context-specific\(^{46,47,48,49}\) (Figure 1B), for example, 24% of the cis-eQTLs identified in brain tissue have been reported to show opposite allelic regulation directions compared with those in blood tissue\(^{50}\). Other studies also demonstrated that the cis-eQTL effect only exists under certain environment exposures like medication usage\(^{51,52}\).

### Biobank is important for multi-omics study in IBD

To address those challenges discussed above, large disease cohort with well-recorded phenotype, well-collected of disease relevant tissues and uniform omics data measurement are important resources for IBD research.

The thesis includes five cohorts. One is the 1000IBD cohort from the University Medical Center Groningen which enrolled 1,215 IBD participants so far, mainly from the Northern provinces of the Netherlands. Multi-omics data layers have been generated and used here (Figure 2), including host genomics (whole exome sequencing, WES; global screen array, GSA on blood), transcriptomics (bulk RNAseq on intestinal biopsies), proteomics (Olink inflammation-panel on plasma), metagenomics (shotgun metagenomics sequencing on fecal samples), metabolomics (untargeted/targeted liquid chromatography–mass spectrometry, LC–MS on fecal samples) and extensive clinical records. Two cohorts are subsets of Lifelines biobank, a large population-based cohort with more than 160,000 participants in the Netherlands. One is Lifelines DEEP cohort (n=1,539) and the other one consists of 8,208 individuals from the Dutch
Microbiome Project (DMP) (Figure 2). The multi-omics data of these two population-based cohorts are mainly host genomics (GSA on blood), metagenomics (shotgun metagenomics sequencing on fecal samples) and detailed phenotypic data. We also included one healthy cohort from 300BCG (Bacillus Calmette-Guérin, n =148) established at Radboud University, Nijmegen, the Netherlands, with phenotypic and proteomics data (Olink inflammation-panel on plasma). In addition, a transplantation cohort from the TransplantLines biobank of University Medical Center Groningen was used to assess the role of gut microbiota in liver (n =415) and renal (n =672) transplant patients.

Figure 2. Illustration of multi-omics data layers that were analyzed in the current thesis from the 1000IBD, LifeLines DEEP and Lifelines Dutch Microbiome Project cohorts.

Aim and outline of this thesis
The aim of this thesis is to extend the current knowledge on the effects of genetic factors on downstream molecular mechanisms, with a particular focus on IBD associated genomic loci. Using multi omics data, I investigated how the IBD associated genetic factors influence gene expressions, plasma proteins, fecal metabolites and the gut microbiota.

Blood eQTL studies have been largely performed in samples from IBD, however, blood is not the most disease-relevant tissue and eQTL effects are highly tissue-specific. In chapter 2, we integrated genomics (WES+GSA) and transcriptomics (bulk RNAseq) data of the intestinal tissues from 1000IBD and proved that a large part of IBD associated genetic risk loci exert an eQTL effect. Moreover, we also showed these eQTLs are inflammation- and cell type-dependent which provided new insight on the IBD genetic mechanisms regarding the disease heterogeneity.

The eQTL studies only focus on genetic effect on gene expression levels without emphasizing how genomic variants influence the post-transcriptional process. In Chapter 3 we investigated the genetic regulations on plasma proteins measured using the Olink panel, which contains 92 inflammation- or IBD-related biomarkers. These data were derived from the 1000IBD and 300BCG cohorts. The findings suggested using these biomarkers for clinical purpose should take the patients genetic and exposure background into account. To construct more complete molecular pathways, we further combined intestinal eQTLs and gut mbQTLs using bulk RNAsesq and metagenomics data.

In chapter 4 I presented a literature review on the two main factors of IBD, the host genetics and the gut microbiota. We described the host genetics-microbiota interactions identified in microbial GWAS in human populations, providing a view on how these interactions might be involved in IBD. We also pointed out that current microbial GWAS is limited by the sample size and lacking further investigation in patients with IBD.

To fill the gap of host genetics-microbiota interactions in disease conditions, in chapter 5 we carried out a mbQTL meta-analysis in the 1000IBD cohort and population-based LifeLines DEEP cohort combining whole exome sequencing (WES) and metagenomics sequencing (MGS) data. This study revealed that IBD associated genes and rare mutations can potentially regulate the microbial composition and functions. Moreover, we also highlighted the disease-specific host genetics-microbiota crosstalk due to the microbial alterations in disease.

Sample size is critical for mbQTL identification. Therefore, in chapter 6 we performed the largest mbQTL study so far in the population-based Dutch Microbiome Project cohort. We associated the genomic data from global screen array (GSA) with the microbial taxa and functions from MGS. The potential gut microbial associated genomic loci are enriched in the genes reported to be involved in human metabolic and immune diseases including IBD.
Human fecal metabolites are heritable and closely linked to the activity of gut microbial community. In chapter 7, we integrated host genetics, medication, diet, fecal microbiota and metabolites data in a subset group of 1000IBD patients and LifeLines DEEP controls. We revealed a general gut metabolic dysbiosis in IBD and the moderating effect from host genetics.

Chapter 8 describes another example combining multi-omics data in disease study. We assessed the role of gut microbiota in liver and renal transplantation patients from TransplantLines biobank. The gut microbial alteration was identified to be associated with post-transplantation mortality which indicated a novel potential microbial biomarker related to organ transplantation survivals.

In the last chapter, I summarized the main conclusions of this thesis and discussed in a broader perspective. I presented my thoughts on the limitations and future directions of the use of multi-omics data in IBD research.

References


