Epigenetic marks of a stable host-microbiota association in the mammalian gut

Dissertation

zur Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultät

der Christian-Albrechts-Universität zu Kiel

vorgelegt von

Wei-Hung Pan

Kiel, Germany 2017

1st Examiner: Prof. Dr. Philip Rosenstiel

2nd Examiner: Prof. Dr. Thomas Roeder

Date of oral examination: 18.09.2017

Approved for print on: 24.05.2017

Parts of this dissertation are contained in the following manuscript:

Dynamic Methylation Signatures of Intestinal Epithelial Cells Reflect Postnatal Development

Wei-Hung Pan[1], Felix Sommer[1,2], Maren Falk-Paulsen[1], Thomas Ulas[3], Philipp Best[1], Priyadarshini Kachroo[1], Anne Luzius[1], Marlene Jentzsch[1], Ateequr Rehman[1], Fabian Müller[4], Thomas Lengauer[4,5], Jörn Walter[6], Stefan Schreiber[1,7], Joachim L Schultze[3,8], Fredrik Bäckhed[2,9], Andre Franke[1], Philip Rosenstiel[1*]

[1]Institute for Clinical Molecular Biology, University of Kiel, Rosalind-Franklin-Straße 12, 24105 Kiel, Germany

[2]The Wallenberg Laboratory, Department of Molecular and Clinical Medicine, University of Gothenburg, 41345 Gothenburg, Sweden

[3]Genomics and Immunoregulation, LIMES-Institute, University of Bonn, 53115 Bonn, Germany

[4]Max Planck Institute for Informatics, 66123 Saarbrücken, Germany

[5]Graduate School of Computer Science, Saarland University, 66123 Saarbrücken, Germany

[6]Department of Genetics, University of Saarland, 66123 Saarbrücken, Germany

[7]Department of Internal Medicine I, University Hospital Schleswig Holstein, 24105 Kiel, Germany

[8]Platform for Single Cell Genomics and Epigenomics (PRECISE) at the German Center for Neurodegenerative Diseases and the University of Bonn

[9]Novo Nordisk Foundation Center for Basic Metabolic Research, Section for Metabolic Receptology and Enteroendocrinology, Faculty of Health Sciences, University of Copenhagen, 2200 Copenhagen, Denmark

Table of Contents

Abbreviations and symbols

5fC	5-formylcytosine
5hmC	5-hydroxymethylcytosine
5mC	5-methylcytosine
A3SS	Alternative 3’ splice site
A5SS	Alternative 5’ splice site
AS	Alternative splicing
BS	Bisulfite sequencing
CD	Crohn’s disease
CONV-R	Conventional-raised
CRC	Colorectal cancer
CRP	C-reactive protein
DMP	Differential methylated position
DMR	Differential methylated region
DNMT	DNA methyltransferase
EPIC	Infinium MethylationEPIC BeadChip
ESR	Erythrocyte sedimentation rate
EWAS	Epigenome-wide association studies
FDR	False discovery rate
FPKM	Fragments per kilobase per million mapped reads
GF	Germ free
GO	Gene ontology
GWAS	Genome-wide association study
HM27	HumanMethylation27K BeadChip
HM450	HumanMethylation450 BeadChip
IBD	Inflammatory bowel disease
IBD-CRC	IBD associated colorectal cancer
IECs	Intestinal epithelial cells
IKMB	Institute of clinical molecular biology

ISCs	Intestinal stem cell
MDS	Multidimensional scaling
mRNA	Messenger RNA
MXE	Mutual exclusive exon
MZ	Monozygotic
OTU	Operational taxonomic units
PCA	Principal component analysis
QC	Quality control
RI	Retained intron
RNA-Seq	RNA sequencing
RRBS	Reduced representation bisulfite sequencing
SE	Skipped exon
SNP	Single nucleotide polymorphism
W1	One-week-old
W12/16	Age between 12-week-old and 16-week-old
W4	Four-week-old
WGBS	Whole genome bisulfite sequencing
RPKM	Reads per kilobase per million mapped reads
BAM	Binary format of Sequence Alignment/Map format
RPM	Reads per million
FPKM	Fragments per kilobase per million mapped reads
IIBDGC	International IBD Genetics Consortium

List of figures

List of tables

1. Introduction

1.1 Host-microbiota interaction in the gut

A tremendously complex and dynamic union of microorganisms inhabits the mammalian gastrointestinal tract and contributes to several aspects of host physiology including metabolism, maturation of the immune system, cellular homeostasis and behavior[1,2,3]. This diverse microbial community is composed of bacteria primarily, but also includes archaea, viruses, fungi and protozoa[2]. In a lately publication from Ron Sender and colleague, the number of microbes cells in colon is “only” 1.3 fold than human cell (3.8×10^{13} microbes cells in colon and 3.0×10^{13} human cells) with a wide uncertainty in a “reference man” (age: 20–30, weight: 70 kg, height: 170 cm)[4]. Furthermore, the genes of microbes that make up the microbiome in the whole body outnumber human genes at least two orders of magnitude, especially with over 3 million bacterial genes in the gut[5]. There are several benefits of gut microbiota to the host, including protection against enteropathogens[6], extracting nutrients and energy from our diets[7] and maintain the normal immune function[8]. However, the commensal microbial communities within the host also represent a potential danger due to their infection and overgrowth.

The dysbiosis in the gut might lead to obesity[9], malnutrition[10], inflammatory bowel disease (IBD)[11], neurological disorders[12] and even colon cancer[13]. The loss of microbiota diversity has been shown associated with several diseases (Figure 1). By this observation, loss of microbiota diversity is generally considered as the consequence of the disease instead of the cause. However, some recent studies suggested different concepts of the role of microbiota in gut inflammation. Michail *et al* found the increase of diversity in pediatric patients with ulcerative colitis (UC) who responded to corticosteroids than the non-responder paitents[14]. Additionally, Crohn’s disease (CD) can also be triggered by a dysbiotic gut microbiota in a mouse model[15]. Altogether, these results provide the strong argument for a causal effect of change of diversity in several human conditions[16].

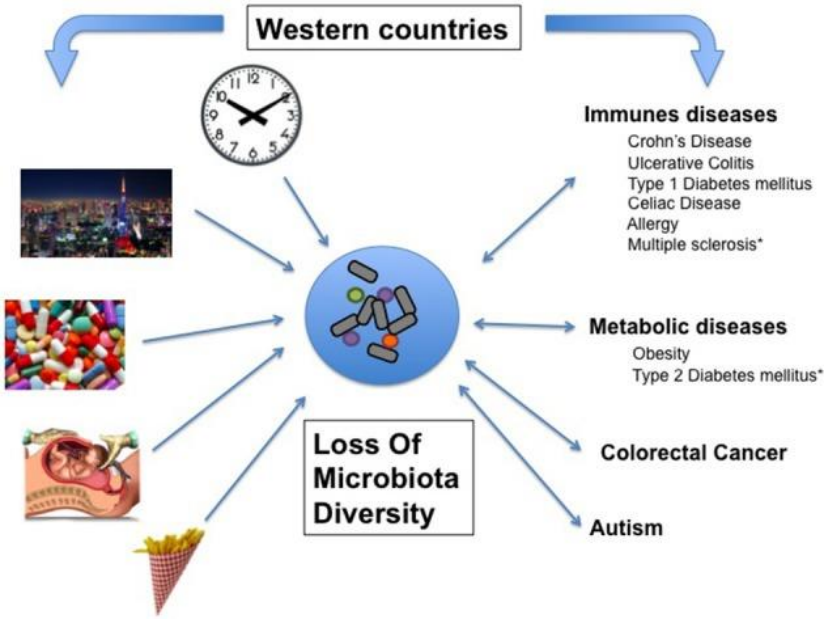


Figure 1 Associative links between Western lifestyle, Human conditions, and loss of microbial diversity

Most of the human diseases affecting westernized countries are associated with microbiota diversity on the right side of figure. However, some western lifestyle patterns cause the loss of microbiota diversity in the left side of figure. Microbiota diversity appears to play a prominent role to link western lifestyle and western chronic human conditions
Figure is from Mosca et al[16]

Intestinal epithelial cells (IECs) play a central role as they line the gastrointestinal mucosa and built a physicochemical and immunological barrier to restrain the microbiota and prevent invasion[17,18]. Interactions between the microbiota and the host, especially IECs, have therefore been studied intensively in the past decade[19]. Previous studies have shown that under normal homeostatic conditions, the gut microbiota regulates the expression of about 10% of host genes[19]. *NOD2* (Nucleotide-binding oligomerization domain-containing protein 2), a prototypic NLR gene, identified as a risk factor for Crohn's disease[20], was shown to specifically react to the intracellular presence of the bacterial cell wall component muramyl dipeptide[18]. Several mechanisms have been implicated in how the gut microbiota can drive these global changes in the host transcriptome. Transcriptional regulators such as NFκB (nuclear factor kappa-light-chain-enhancer of activated B cells) or *CEBPB* (CCAAT/enhancer-binding protein beta) may be engaged by the microbiota to modulate the expression of specific target genes[21].

1.2 Role of intestinal epithelium

The intestinal epithelium, composed of single layer of cells, builds a physical barrier to separate the external environment and host tissues of the gastrointestinal tract[22]. This single-cell layer consists of four major cell lineages including absorptive enterocytes, goblet cells, enteroendocrine cells and Paneth cells (Figure 2). The majority in IECs are absorptive enterocytes, which serve for nutrient absorption. They express many catabolic enzymes on their exterior luminal surface to break down proteins and sugars from the gut into smaller particles that are more easily absorbed. The function of goblet cells is to secrete the mucus layer that protects the epithelium from the luminal contents. Enteroendocrine cells are specialized hormone-producing endocrine cells in the gastrointestinal tract and Paneth cells are the main producers of antimicrobial peptides[23]. An intestinal stem cell (ISC) can be defined by two properties: self-renewal and multipotency[24]. ISC can maintain itself throughout long periods and differentiate into enterocytes, goblet cells, enteroendocrine cells and Paneth cells. During cell differentiation, enterocytes, goblet cells and enteroendocrine cells will migrate upward to the top of the villi where cellular apoptosis and epithelial shedding occur and Paneth cells will migrate downward to the bottoms of the crypts.

These various functions of subsets of IECs maintain intestinal homeostasis by separating the intestinal lumen from the underlying lamina propria and persevere symbiotic relationship with the host and intestinal bacteria. The colonic microbiota breaks down complex carbohydrates that cannot be metabolized by the host; bacterial fermentation generates short-chain fatty acids (SCFA) that serve as energy sources for colonic epithelial cells; colonic bacteria produce antimicrobial peptides that promote the maintenance of a symbiotic community[23]. Incompleteness of intestinal epithelium is a key pathogen of IBD[25]. Detrimental microbial composition changes will induce an inappropriate immune response causing damage to the intestinal epithelium.

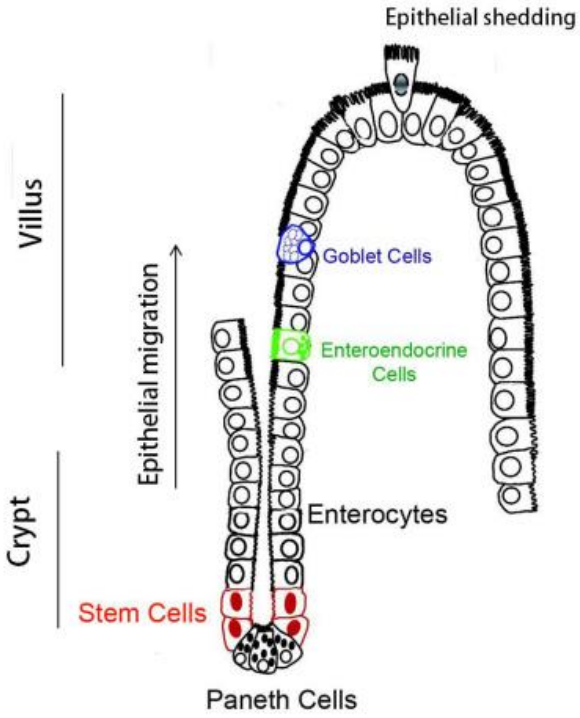


Figure 2 Intestinal epithelial cell lineages and the formation of the crypt-villus axis.

The intestinal epithelium consists of four major cell lineages that arise from a pluripotent stem cell progenitor located in the epithelial crypt region including absorptive enterocytes, goblet cells, enteroendocrine cells, and Paneth cells. The figure is modified from Yu et al.[26]

1.3 Intestinal inflammation and inflammatory bowel disease

The human digestive system is composed of the gastrointestinal tract plus the accessory organs of digestion that includes mouth, esophagus, stomach, small intestine, and large intestine. Gastrointestinal tract controls entire host metabolism and physiology and constitutes/represents the most diverse ecosystem of microbial interactions. Inflammation anywhere in the gut disrupts this normal process. The term IBD describes chronic inflammation of the intestine, which may affect different parts of the gut. IBD comprises two distinct subforms: UC and CD. The primary difference from UC and CD is the inflammation location where the digestive tract is affected. CD usually occurs in patches with granuloma formation and may affect any part of the gastrointestinal tract from mouth

to anus, though it primarily affects terminal ileum and colon. The areas of severe, persistent and transmural inflammation develop thickened sub-mucosal wall, strictures, fistulae, fissures, abscesses and fat deposits. However, UC shows that continuous inflammation is mainly limited to mucosa and submucosa of the colon and rectum which develop sores or severe ulceration[27] (Figure 3). The main symptoms of IBD are diarrhea, bleeding ulcers, stomach pain, weight loss and anemia. People with CD may get canker sores in their mouths, and the inflammation may also affect the skin, eyes, joints, and liver. In general, UC and CD have similar symptoms; hence, it is difficult to distinguish these two subtypes in the initial stage.

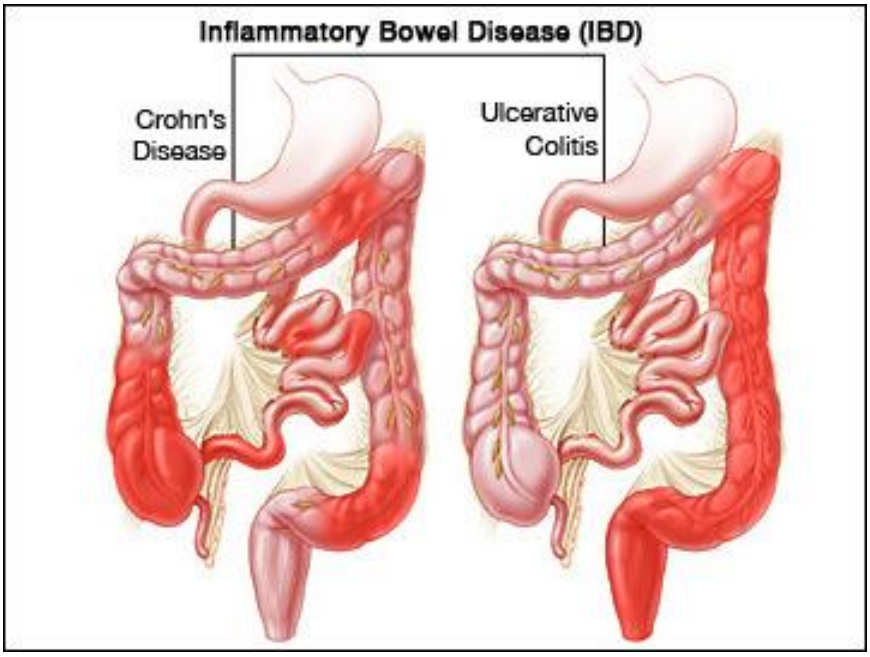


Figure 3 Comparison of CD and UC effects position

The red patches show active inflammation in intestines, and the white patched shows the normal part of the intestine. The figure is modified from Save Jon Blog: PSC and IBD: What's the Connection? (http://blog.save-jon.org/blog/psc-and-ibd-whats-the-connection)

Over the past decades, the prevalence of IBD increased significantly, especially in the developed countries[28]. In the United States, approximately 1-2 million people have IBD, with an incidence of 70-150 cases per 100,000 individuals[28]. CD and UC prevalence rate in European countries is around 322 and 505 per 100,000 individuals respectively[29]. Moreover, according to a survey from Hein and colleagues[30] in a large insurance cohort,

the IBD prevalence in Germany considerable increase since the 1980s. The IBD prevalence in 2010 was 744 per 100,000 (CD: 322, UC: 412 per 100,000 persons). These high incidence rates in Europe/Germany suggested that the lifestyle of western industrial societies is a contributing factor to an increased pathogenesis of IBD. The disease onset for CD and UC can happen at any age across pediatric and adult populations, while the rates are usually reported to be bimodal: the first peak is in the age group of 15-30 years and the second occurs in the aged group 50-70[31].

The exact cause of IBD remains unknown. However, genetics and gut microbiota have been associated with IBD. Genetic has been reported as an essential component long before for IBD, especially in CD[32,27]. This genetic effect can be seen from various familial aggregations, sporadic and twin studies[33]. Accordingly, the first-degree relatives IBD patients have risk 3 to 20-fold greater likelihood of developing the disease than the general population[34] and the likelihood increases further if both parents suffered from the disease[35]. International IBD Genetics Consortium (IIBDGC) recently identified totally 201 known IBD risk loci based on 75000 IBD (UC and CD) patients and healthy controls. Together, these loci explain 13.1% and 8.2% of the variance in disease liability in CD and UC respectively[36]. Most of the genes within the identified loci were also associated with key pathways involving both innate and adaptive immune system that are disturbed around IBD patients. Therefore, many IBD patients shared with other complex immune-mediated diseases like celiac disease, psoriasis or rheumatoid arthritis. However, genetic effect is not the dominant risk factor, considering the result of twin studies; CD shows a concordance of 20-50% in monozygotic (identical) twins and only 10% in dizygotic twins. The concordance would have been close to 100% in monozygotic twins and around 50% in dizygotic twins if the genetic components were fully responsible for IBD.

The aetiopathogenesis of inflammatory bowel disease has been considered from the host perspective for a long time, but more and more studies reveal the significant influence of host-gut microbiota interaction on IBD. Gut microbiota dysbiosis with a breakdown of host-microbial mutualism is probably the marking event in the development of IBD[37]. The reduced abundance of the phylum *Firmicutes* has been noticed in patients with IBD[11]. on the contrary, several studies reported the increased number of phylum *Bacteroidetes* in

the IBD patients[38]. These two dominated phyla make up for 90% of the phylogenetic categories in the normal gut microbiome ecosystem, and it is interesting to see the disparate ways where they are altered in IBD. Probiotics and prebiotics treatment have been applied in IBD therapy, this therapeutic strategy aims to restore the balance of the gastrointestinal microflora in order to reduce or prevent intestinal inflammation[39]. As microbiota are involved in IBD, fecal microbiota transplantation has been considered as a possible treatment for rebooting the gut microbiota composition in IBD patients. However, the evidence are insufficient to recommend fecal transplantation in IBD. This approach needs to be constructed the optimal design of delivery as well as randomized, placebo controlled trials to establish the effectiveness of fecal transplantation[40].

1.4 Transcriptome studies in intestinal inflammation

The transcriptome studies in intestinal inflammation provide us the evidence which describes the role of protein-coding and non-coding RNAs in modulating immune responses in IBD. Recently, more and more IBD-related genome-wide expression studies came out because of the NGS technique. That provides us a broad view of various study topics including site-specific comparison, IBD patient and healthy comparison, healthy tissue and inflamed tissue in IBD patients comparison, gene expression change in different developmental stages...etc. The understanding of transcriptome can help the clinical research for identifying the disease-associated gene that might serve as targets for therapeutic intervention. Furthermore, the targeted genes with the difference in transcription level can serve as the biomarkers to distinguish the subtype of IBD[41] or the period of the disease[42].

The transcriptome analysis for IBD started earliest in 1997 by using spotted cDNA arrays[43]. This study only compared 1000 pre-selected genes between tissue samples of rheumatoid arthritis and IBD due to the technology limitation. Nevertheless, still some genes were identified as differentially expressed genes with the function of tissue inhibitor of metalloproteinase, ferritin light chain, and manganese superoxide dismutase. Dieckgraefe and colleagues[44] next examined colonic mucosa samples using Affymetrix

Hum 6000 arrays with a coverage of ~6500 genes and expressed sequence tags. This study identified 74 differentially expressed genes between inflamed UC and normal mucosa, grouping in functional classes such as immunoregulation and tissue regeneration. For the diagnosis purpose, it is difficult to distinguish UC and CD precisely by gene expression pattern. Because inflamed mucosa from UC and CD are remarkably similar[45]. Although the similarity of transcriptome pattern in inflamed mucosa from UC and CD might not help for diagnosis, Olsen *et al.* used random forest modeling of genome-wide gene expression data for distinguishing quiescent and active UC colonic mucosa versus control and CD colonic mucosa[46]. Many following studies also discovered the list of significant genes in the different scenarios; however, the results are highly inconsistent even in the similar experimental setting.

There are two major reasons for the inconsistency: material resource and sample size (statistical power). Sample heterogeneity greatly changes the expression levels. Paneth cell metaplasia, leukocyte infiltration, crypt hyperplasia and ulceration with loss of epithelial cells are essential factors for gene expression levels[47]. Different cell types, different sample locations and even sequencing runs might end up to different results.This diversity often hinders the interpretation of the differences between sample groups. Moreover, the small sample size in each study increases the uncertainty and lower the reliability of the finding. With the small sample sizes (n<20), one need to very cautious for the false positive. Thus, validation process becomes a crucial step in IBD transcriptome study. One can either validate the finding in the independent cohort with different technique or benchmark with the former similar study. In addition, the meta-analysis, which merges the dataset from different studies, might be another alternative. Granlund *et al.*[45] employed the meta-analysis approach, by showing a similarity between the gene expressions in inflamed mucosa from UC and CD patients and furthermore confirmed by hierarchical clustering of 10 external data set from published article.

1.5 DNA methylation in inflammatory bowel disease

Many observations in humans family study and animal models all suggest that genetically determined factors contribute to IBD susceptibility[48]. Even though, genetics factor can only explain a small proportion of disease heritability (CD: 13.1% and UC: 8.2% of disease variance)[36]. To unravel other parameters which might be less strong than genetic effect become the further direction of IBD research. One of such newly created fields is epigenetics, particularly in DNA methylation. Epigenetic mechanism includes three major components: DNA methylation, histone modification, and non-coding RNA. DNA methylation is thought to inhibit gene transcription, but recent data indicates that the functional consequences may be more complex[49]. DNA methyltransferases (DNMTs) family enzymes catalyze DNA methylation. Methyl groups can be reprogramed via actions of DNA demethylases such as intestinal maturation process or disease onset of IBD. In the following content, the roles of DNA methylation will be discussed in the context of the IBD-related epigenome study.

1.5.1 DNA methyltransferases and IBD

There are three DNA methyltransferases, which have been proposed in the pathogenesis of IBD and IBD associated colorectal cancer (IBD-CRC): *DNMT1*, *DNMT3A*, and *DNMT3B* (Figure 4). *DNMT1* is a key maintenance methyltransferase that primarily methylates hemimethylated DNA in the genome during DNA replication[50]. *DNMT1* is highly expressed in active inflamed UC colon mucosa than in normal or quiescent UC colon mucosa[51]. *DNMT1* expression induces not only an elevation of genomic DNA cytosine methylation, but also CpG island methylation in promoter regions in particular targeted genes[52]. Signal transducer and activator of transcription 3 (*STAT3*) bind directly onto the *DNMT1* promoter in malignant T cell lymphoma that might induce DNMT1 expression[53]. *DNMT3A/B* and *DNMT1* regulate DNA methylation maintenance cooperatively. Besides, *DNMT3A/B* have additional roles in de novo DNA methylation and demethylation functions. *DNMT3A* involves in innate and adaptive immune responses. It suppresses interferon gamma (IFNγ) transcription in T cells by directly inhibiting

transcription factor binding[54]. DNA methylation in interferon gamma gene *(IFNG)* promoter region is correlated with *IFNG* expression and immune response against microbial antigens in UC patients[55]. *DNMT3B* expression showed upregulated in active UC colonic mucosa compared to normal colonic samples but relatively lower than that of *DNMT1*[51] (Figure 4). Huidobro *et al.* found the hypermethylation pattern on the distal *DNMT3B* promoter in human colorectal cancer cell lines (HCT15, DLD1, Col15, HT29, SW480 and RKO). On the other hand, low expression of DNMT3B results in hypomethylation of *FURIN* gene promoters[56]. Furthermore, *DNMT3* has been reported in GWAS study as a CD associated risk loci[57]. To summarize, the different expression level in DNMTs might serve as a biomarker for subtype diagnosis or monitor IBD and IBD-CRC progression in patients.

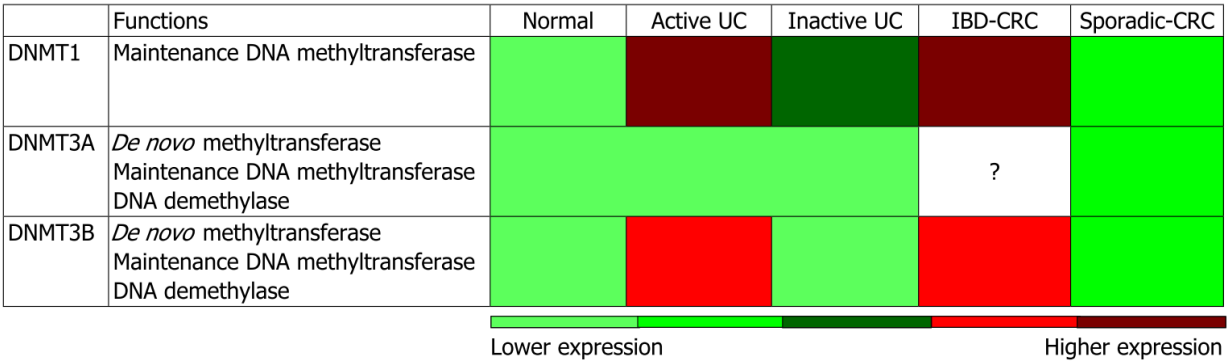


Figure 4 Potential relative expression levels of DNMTs in different IBD associated disease

The relative DNMTs expressions were normalized to healthy controls to display potential relative expression in different inflammatory bowel disease associated diseases: active-UC, inactive-UC, IBD-CRC and sporadic-colorectal cancer patient specimens consolidate from several studies. Figure is from Low D et al, 2013[50]

1.5.2 Whole methylome profile in IBD study

Microarray technique, especially Illumina infinium human 27k & 450K, has been widely employed to investigate the genome-wide epigenetic variants[58,59]. Moreover, the recent next generation sequencing[60,61] provides the comprehensive view for DNA methylome studies in different tissue/cell. One can apply these high-throughput data to connect between common diseases and complex traits or integrate methylome data with the

different layer of omics data (e.g. transcriptome, microbiome). Epigenome-wide association studies (EWAS) which connect epigenetic variant and disease has been applied in various clinical research fields, such as cancer, type 2 diabetes, obesity and other complex diseases. It aims to detect the variants that associated with the complex phenotype and determine the novel gene and the pathway in common disease. DNA methylation in the promoter region is thought to be associated with gene expression. However, the integration studies of methylome and transcriptome discovered the methylation modification for gene expression also in gene body, transcription start sites and enhancer region[62].

Nimmo *et al*[59] compared the methylation level from 21 ileal CD patients and 19 healthy controls in whole blood genomic,1174 CpG sites were identified as differentially methylated. Out of these CpG sites, 35 genes were found overlapped with published GWAS study in CD, including *NOD2*, TNFα and *CARD9* (caspase recruitment domain family, member 9). Häsler et al[49] examined three layers of genome-wide scans in 20 monozygotic twins discordant for UC, including transcriptome profiling, genome-wide differentially methylation positions (DMPs) and genome-wide differentially methylation regions (DMRs). They revealed 61 diseases associated gene with at least one DMP or DMR in the 50kb windows from the transcription starting sites. However, none of them overlapped with the identified UC GWAS loci[63]. Cooke *et al*[58] established multiple comparisons between inflamed/non-inflamed UC and CD. Interestingly, the methylation profile between inflamed UC and inflamed CD showed no difference, but the difference was found in 13 positions in non-inflamed UC and non-inflamed CD. That might imply the subtype of IBD can be distinguished by their methylome status. Additionally, the age-dependent methylation dynamics needs to be considered as an important risk factor in IBD[64]. One mouse study showed that 271 methylation loci underwent significant alteration during this developmental period in dextran sulfate sodium colitis model[64]. In conclusion, whole methylome screening will become a useful clinical diagnostic tool to detect the biomarker in IBD.

1.6 Epigenetic markers and host-microbiota interaction

The microbiota has the potential to modulate host epigenetic mechanisms and thereby regulate transcription more globally[65,66,67]. The microbially produced short-chain fatty acids (SCFAs) butyrate and propionate are potent inhibitors of histone deacetylase (HDAC) enzymes[68] and therefore may promote heterochromatin formation and increase transcriptional activity. However, global changes in the accessible chromatin landscape by the gut microbiota were not detected in previous study[21]. Additionally, the intestinal microbiota may modulate DNA methylation, since microbially produced folate is an essential methyl donor during DNA methylation[67].

Yu and colleagues have shown that during postnatal development, both the epithelial transcriptome and the DNA methylation landscape underwent fundamental reshaping[69]. The early neonatal period is a critical phase not only for the development of the intestinal tract but also for the establishment of the microbiota and proper maturation of the immune system[70,71]. Notably, colonization at a later stage fails to normalize these immunological defects. This persistence of microbiota-dependent regulatory signatures points to microbial imprinting through epigenetic mechanisms (possibly DNA methylation) that are long lasting once they are established[2,72]. However, whether microbial colonization early in life alters the DNA methylation pattern and alongside the epithelial transcriptome during postnatal development and maturation of the gut epithelium remains largely unknown.

1.7 Technological consideration on next generation sequencing

Discovering the full DNA structure and its role to answer the complex biological questions has been the dream of mankind since a longtime. The “original” DNA sequencing methodology, known as Sanger chemistry, uses specifically labeled nucleotides to read through a DNA template during DNA synthesis[73]. After a series of technical innovations, the Sanger method has reached the capacity[74]. In order to sequence longer sections of DNA, a new approach called shotgun sequencing was developed during the

establishment of Human Genome Project (HGP)[75]. Shotgun sequencing divides human chromosomes into DNA segments of an appropriate size and then subdivides these segments into smaller, overlapping DNA fragments for sequencing. Next step after sequencing is to fill in gaps and resolve DNA sequences in ambiguous areas which are not obtained during the shotgun phase[76]. The core philosophy of massive parallel sequencing used in next-generation sequencing (NGS) is adapted from shotgun sequencing[77]. It allowed the mass parallelization of sequencing reactions, greatly increasing the amount of DNA that can be sequenced in one run. This massively parallel sequencing technique revolutionized sequencing capabilities from the first sequencing techniques leading to the coining of the term “next-generation sequencing”.

Pyrosequencing method was the first sequencing technology established by 454 Life Sciences which was followed by Illumina/Solexa technology in 2007, SOLiD (Sequencing by Oligo Ligation Detection) by Life Technologies, and PGM (Personal Genome Machine) by Ion Torrent in 2010. 454 Life Sciences was purchased by Roche in 2007 and shut down in 2013 when its technology became noncompetitive[78]; likewise, Ion Torrent was bought by Life Technologies at the end of 2012[79]. Nowadays, Illumina occupies around 70% of the sequencing market. Apart of general sequencing method, Oxford Nanopore developed nanopore sequencing since 2008. The advantages of Oxford Nanopore are minimal sample preparation, sequence readout that does not require nucleotides, polymerases or ligases, and the potential of very long read lengths (>10,000-50,000 nt)[80]. Nanopore sequencing approach has long been a potentially strong competitor in sequencing market, but the company struggled to deliver a real-world commercial device. The basic workflow of NGS involves library preparation, cluster generation and sequencing; though library preparation might differ in different technologies. The generated data processing is followed by bioinformatics analyses of the sequencing data (2.2 Analytical pipelines).

1.7.1 16S rRNA sequencing

“What is the microbial composition there?” is one of the most frequent questions in microbial ecology. This question can be answered using various tools, but one of the long-lasting gold standards is to sequence 16S ribosomal RNA (16S rRNA) gene amplicons[81]. 16S rRNA sequence analysis, a culture-independent method, has been widely used to clarify the taxonomic affinities of bacterial taxa and as a powerful tool for assessing the diversity of environmental or clinical samples[82]. The 16S rRNA gene is a section of the prokaryotic DNA present in all bacteria and archaea. In contrast to the genes needed to make enzymes, mutations in 16S rRNA can be less tolerated since it may affect structures essentially (if a bacterium does not have the gene to make the enzymes needed to utilize lactose, it can use an alternative sugar or protein as an energy source). Thus, few other genes are as highly conserved as the 16S rRNA gene[83]. The consistency of these short regions (1542 nucleotides) increases its detection specificity and also the length allows us for the universal primer.

The 16S rRNA gene can be subdivided into highly conserved primer binding sites and nine variable regions (V1-V9 in Figure 5). The variable regions depict species-specific signatures. Universal primers are usually chosen as complementary to the conserved regions at the beginning of the gene and at either the 540-bp region or at the end of the whole sequence and the sequence of the variable region in between is used for the comparative taxonomy[83]. The 16S rRNA gene can directly be isolated by PCR with universal primers targeting the conserved regions. The library preparation is followed from the protocol of Illumina MiSeq 16S Metagenomic Sequencing Library Preparation[84]. Here , in this study, V3 and V4 region were amplified, both Illumina sequencing adapters and dual-index barcodes were added to the amplicon target. Paired 300-bp reads sequencing and MiSeq v3 reagents were employed. The ends of each read were overlapped to generate high-quality, full-length reads of the V3 and V4 region[84]. The data preprocessing detail are described in 2.2.1

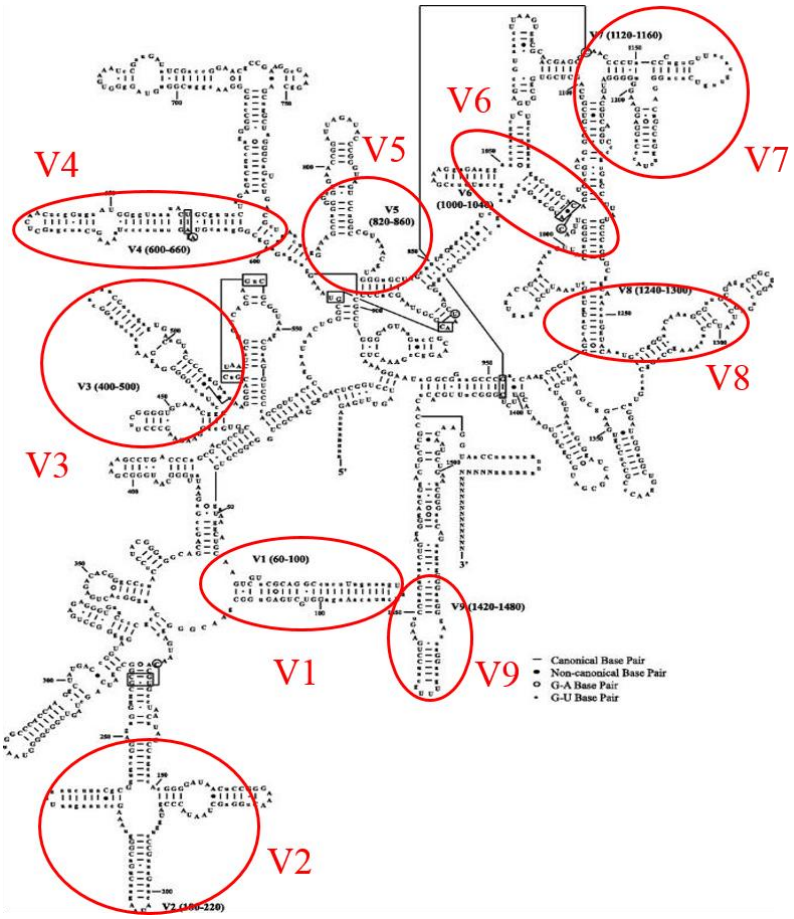


Figure 5 16S rRNA secondary structure for the Toll sequence.

Regions V1 to V9 are hypervariable regions as determined by Ashelford et al[85].; approximate nucleotide positions are given in parentheses. Numbering is unique to this structure. Modified from Justine R. Hall et al [86]

1.7.2 RNA-Seq

Genetic information from DNA to proteins is passed down via mRNA in a finely regulated fashion, wherein identity of each expressed transcript and its transcriptional levels make up the "transcriptome[87]." The definition of transcriptome can be stated as the complete set of messenger RNA (mRNA), which is produced by the genome in a single cell or a population of cells. Recently, RNA-Seq has been widely used for transcriptome profiling by deep-sequencing technologies which have several advantages over other existing approaches, especially low background noise and high reproducibility[88]. It provides a

precise measurement of gene expression levels from all transcripts and their isoforms[89]. It can detect even subtle changes in gene expression in response to environmental changes which are not captured by other methods. A typically established RNA-Seq workflow (Figure 6) starts with total RNA sample isolation and preparation, then followed by a ribosomal depletion step during purification to exclude ribosomal contamination. RNA is highly sensitive and easily degraded by RNAse enzymes; therefore, extreme care should be used while handling RNA samples.

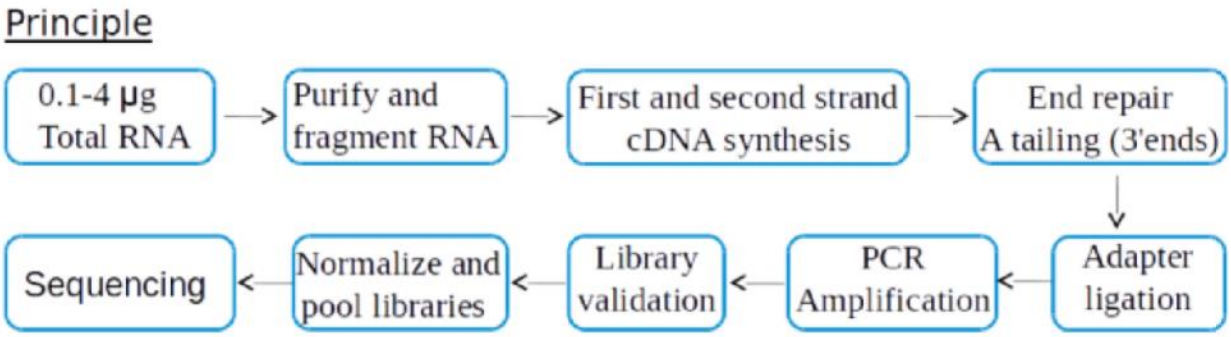


Figure 6 Principle sequential steps of an RNA-Seq workflow

1.7.3 RRBS

RRBS (reduced representation bisulfite sequencing) was first mentioned by Meissner *et al.*[90]. It is a restriction enzyme based (MspI, 5'-C|CGG-3') CpG enrichment method for DNA methylation analysis (Figure 7A). Bisulfite treatment converts unmethylated cytosine residues into uracil while methylated cytosines remain unchanged. The restriction enzyme (here is MspI) selects a size-range fragment for sequencing. This technique combines both bisulfite sequencing and restriction enzymes because MspI specifically recognizes CCGG sequence (Figure 7B). The reads are enriched with the high CpG content areas of the genome. Though RRBS captures only 1% of the whole genome, it can generate accurate methylation levels for most CpG rich and regulatory regions from sample input as low as 10ng[91] well suited for many clinical samples like tumors, sperm cell etc. These regions are typically CG islands or promoter regions[92], therefore, RRBS is a great way to

get high coverage methylation data on promoters. On the other hand, the detected sites might not be really representative. People need to be cautious about the conclusions of distal regions drawn simply based on RRBS. In my mouse study, I looked at a 'reduced and low-complexity' fraction of the genome at single nucleotide level. The other caution while using RRBS is the conversion efficiency of BS. In order to determine conversion efficiency prior to sequencing, lambda DNA were added in as an internal control (spike-in). The quality of conversion rates is based on the methylation status at non-CpG sites[93]. The biggest advantage of RRBS compared to WGBS (whole genome bisulfite sequencing), is cost-effectiveness since it focuses on the enrichment of CpG-rich regions in close proximity to the restriction enzyme's recognition sequence; however, these may exhibit a lack of coverage at intergenic and distal regulatory elements that are relatively understudied[94].

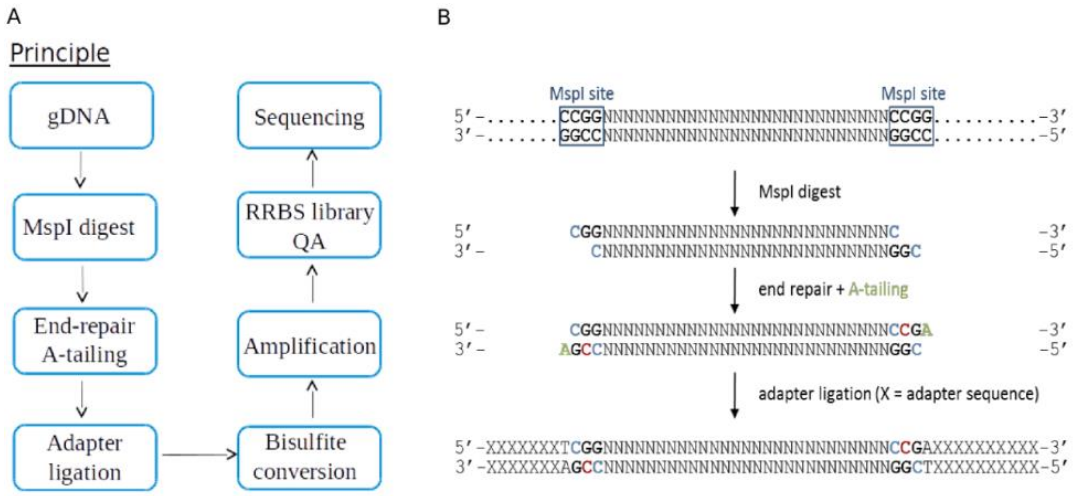


Figure 7 Principle and workflow for RRBS

(A) RRBS working flow; (B) Target cutting site of the MspI enzyme so that every read starts with either CGG or TGG depending upon the genomic state and will have information for at least one CpG dinucleotide. Figure is modified from: RRBS guide, Babraham[95]

1.8 Aims of this thesis

The change of host physical conditions might reshape the host-microbiota balance in the context of transcription and epigenetic. Although, the interactions between the microbiota

and the host are well established in previous studies, the cross talk between host epigenetic marker, transcription and gut microbiota in whole genome scale remains largely unknown. The aim of my thesis is to build a genome-wide map of the epigenetic marks (DNA methylation) with the different intestine conditions, and furthermore investigate host-microbe association in the murine gut. I expect the finding can provide a clear picture of the role of gut microbiota together with the epigenetic change in intestine inflammation.

- **Hypothesis**

I hypothesize that epigenetic marks (i.e. DNA methylation) are important biological master switches that contribute to the stability of the physiological host-flora association. In order to validate my hypothesis, I investigated the interplay of epigenetic marks, transcriptomal signatures and microbial communities in mice and in a human cohort of UC twins.

- **Aims**
    - The aim of the first study is to build a genome-wide map to present UC relevant effects on three layers: transcriptome, epigenome and gut microbiota.
    - In the second study, I furthermore sought to investigate the microbial effects on DNA methylation and the transcriptome of intestinal epithelial cells (IECs) during postnatal development.

2. Material and Methods

This section will provide the insights into sample preparation and various data analysis procedures including Methylation27K microarray, Affymetrix gene expression microarray, 16S rRNA gene sequencing, RRBS and RNA-Seq. Data analysis mostly includes in-house pipeline established for whole genomic screening.

2.1 Sample preparation

2.1.1 Twins study

There are two panels in this study: screening panel and validation panel. Screening panel consists of twenty monozygotic twins, discordant for ulcerative colitis (median age: 25, range 18-70). Biopsies were taken endoscopically from a defined area of the colon and immediately snap-frozen in liquid nitrogen. All biopsies used in this screening panel were primary tissues from the intestinal mucosa. DNA and RNA were extracted from sampled biopsies. DNA was extracted from biopsies using the QIAamp Tissue DNA preparation kit (Qiagen). Total RNA was extracted and processed by RNeasy mini-kit from Qiagen and quality controlled using an Agilent Bioanalyzer according to the manufacturer's protocol. These data were already published by Haesler and colleagues[49] in 2013.

Forty unrelated UC patients (n=20) and healthy control individuals (n=20) were recruited in the validation panel. The criteria for healthy participant includes age between 18 and 50, no antibiotic or antimycotic treatment in the previous 6 months, no probiotic based product consumption, hospitalization, and/or diarrhea in the previous 6 months, no known infection and sign of inflammatory markers. The biopsies in this cohort were sampled from the sigmoid colon. UC patients were selected for displaying an endoscopically active disease in the sigmoid colon at the time of sampling. The bioethical committee of the University of Kiel, where the patients were recruited, approved the study setup. DNA and RNA was extracted using same procedures as for screening cohorts. Validation of the observation was performed using real time PCR of transcript levels and microbiota as well as amplicon sequencing of DNA methylation.

2.1.2 Mouse study

C57BL6/N female littermate mice were maintained under standard specific pathogen free or GF conditions in the laboratory for experimental biomedicine at University of Gothenburg as described in the publication of Sommer et al 2015[19]. Mice were kept under a 12-h light cycle and fed autoclaved chow diet *ad libitum* (Labdiet, St Louis, MO, USA). They were sacrificed at different stages: week 1, week 4 and between week 12 and week16. Mice were killed by cervical dislocation and the small intestines were removed for isolation of IECs. All animal protocols were approved by the Gothenburg Animal Ethics Committee. IECs were isolated from small intestinal tissue using the Lamina Propria Dissociation Kit (Miltenyi BioTech, Bergisch Gladbach, Germany) according to the manufacturer´s protocol. In brief, intestinal epithelial cells were isolated by disruption of the structural integrity of the epithelium using ethylenediamine Tetraacetic acid (EDTA) and dithiothreitol (DTT). Purity of individual IEC fractions was analyzed by flow cytometry on a FACSCalibur flow cytometer (B&D, Heidelberg, Germany) with Cellquest analysis software from Becton Dickinson.

RNA was isolated from the purified small intestinal IECs using the TRIZOL method. Briefly, 1ml TRIzol was added to 50-75 mg pestle homogenized tissue followed by vortexing, five minutes' incubation at room temperature and addition of 200 µl chloroform. After mixing, incubation at room temperature for 2-3 min and centrifugation (12.000 g) at 4°C for five minutes was done. Further, the clear supernatant was mixed with 500 µl isopropanol followed by incubation at room temperature for ten minutes. After further centrifugation (12.000 g) at 4°C for ten minutes, the supernatant was discarded and the pellet washed with 1 ml cold 75 % EtOH followed by vortexing and centrifugation (7.500 g, 4°C, 5 min). The pellet was dried and dissolved in RNase-free water. RNA libraries were prepared using TruSeq v2 Kit (Illumina) according to manufacturer's instructions. All samples were sequenced using an Illumina HiSeq 2000 sequencer (Illumina, San Diego, CA) with an average of 23 million paired-end reads (2x 125 bp) at IKMB NGS core facilities.

RRBS methylome screening was employed in this study. In the protocol, purified DNA was well digested with MspI restriction enzyme. DNA oligos of known sequence and with known cytosine modifications were added to the digested DNA. PCR amplification of

bisulfite-converted reads was performed to generate the desired RRBS library. During this procedure, fragments bind on either side to a flow cell (solid glass surface) coated with specific oligonucleotides using their adaptors. The fragments hybridize with their complementary adapter and undergo bridge-amplification resulting in cluster synthesis of identical DNA fragments. The libraries were purified via magnetic beads (Ampure) employing acetonitrile instead of ethanol. The DNA was then converted with Cambridge Epigenetics (CEGX) TrueMethyl24-Kit according to the manufacturer's handbook. The final libraries underwent size selection of greater than 180 bp via magnetic beads (Ampure). This step made sure to remove adapter dimers. The DNA was end repaired and A-tailed, followed by the ligation of barcoded next generation sequencing adapters. Library pairs were pooled together and three pairs were sequenced per lane on an Illumina HiSeq 2500 platform (Illumina, San Diego, CA) at an average of 127 million single-end 50 bp reads at IKMB in Kiel, Germany. All RNA-seq and RRBS data have been uploaded to GEO with accession number [GEO:GSE94402].

2.2 Analytical pipelines

This thesis includes several types of high-throughput data and its analysis. In the first twins study, microarray data analysis and preprocessing procedures were mainly developed by Häsler and Feng[49] within Institute. For mouse study, the development and implementation of RRBS and RNA-Seq pipelines from the HiSeq sequencing platform were developed by Kachroo[96] and myself. All data analysis were executed on the high performance computing cluster (HPC) of Kiel University on Linux system and mainly packaged into shell scripts and some independent scripts as and when needed, which can be modified or run in parallel for other similar projects. Data preprocessing, analysis for microbiota, gene expression, DNA methylation and integrated analysis will be covered in the following text.

2.2.1 16S rRNA gene data analysis

The data preprocessing was followed the MiSeq standard operating procedure using software MOTHUR[97] as suggested by lab of Dr. Schloss for processing paired-end reads of 16S rRNA gene sequences (https://www.mothur.org/wiki/MiSeq_SOP). The modified pipeline was shown in Figure 8. After data preprocessing, the bacterial composition across samples was displayed as a table with columns indicating samples and rows representing OTUs. Based on the table, the statistical method can be applied to have further novel insights on the microbial communities. The analysis flow chart is displayed in Figure 9.

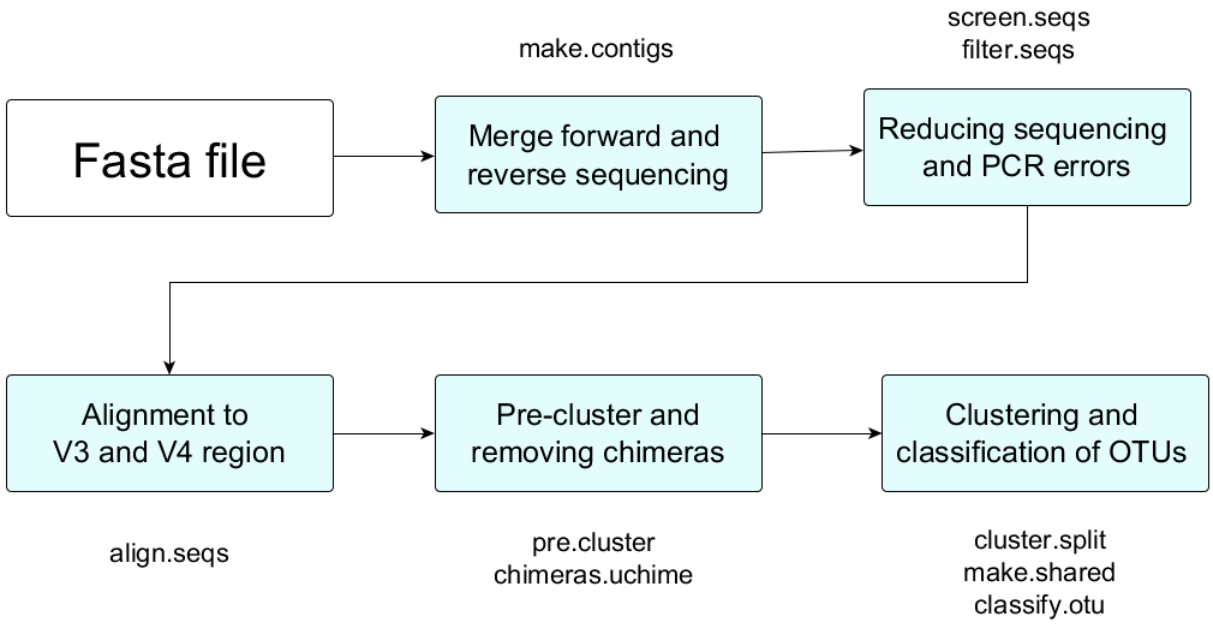


Figure 8 16S rRNA gene data preprocessing

Sequences merging, reducing sequencing and PCR errors

The first step of preprocessing is to match the paired sequences, combine the sequence data and also affiliating the sequences to corresponding sample. Mothur uses straightforward alignment algorithm for aligning the paired sequences (aka contigs). If a base has a disagreement, the one which has a quality score of 6 or more better than the other is chosen, otherwise the consensus base is set to ambiguous base (N). If one

sequence has a base and the other has a gap, the quality score of the base must be over 25 to be considered real. The fasta and qual files were generated after alignment. The reads were then filtered out with improper assembly, which had a single ambiguous base, or long homopolymers of equal and/or more than 8 bases.

Alignment, Pre-cluster and removing chimeras

Consequently, the sequences were aligned against Mothur curated Sliva reference database (https://www.arb-silva.de) in defined 16S rRNA gene variable region V3 and V4. The sequences were only kept if they aligned to defined V3-V4 region. In order to de-noise the sequences and reduce the computation loading, similar sequences which allowed only one difference for every 100 base were merged. After removing the sequencing errors by the above-mentioned procedure, next step was to remove chimeras. Chimeras are hybrid products between multiple parent sequences that can be falsely interpreted as novel organisms, thus inflating apparent diversity[98]. As suggested by Mothur pipeline, UCHIME algorithm[99] was employed to detect and remove these chimeras.

Clustering and classification of OTUs

For a deeper and more accurate analysis, the sequences were assigned into operational taxonomic units (OTUs). Because the sequence-based recognition of uncultivated microbial populations is not equivalent to the traditional taxonomic classification, Sequences having at least 97% of homology were clustered as an OTU. Individual OTUs were classified phylogenetically using RDP dataset (trainset9). Furthermore, equal number of sequences were subsampled (smallest sequence size) to normalize the sequencing depth. This subsample procedure is essential to compare all samples onto the same standard. The final OTU table as obtained after subsampling was used for downstream statistical analysis (Figure 9).

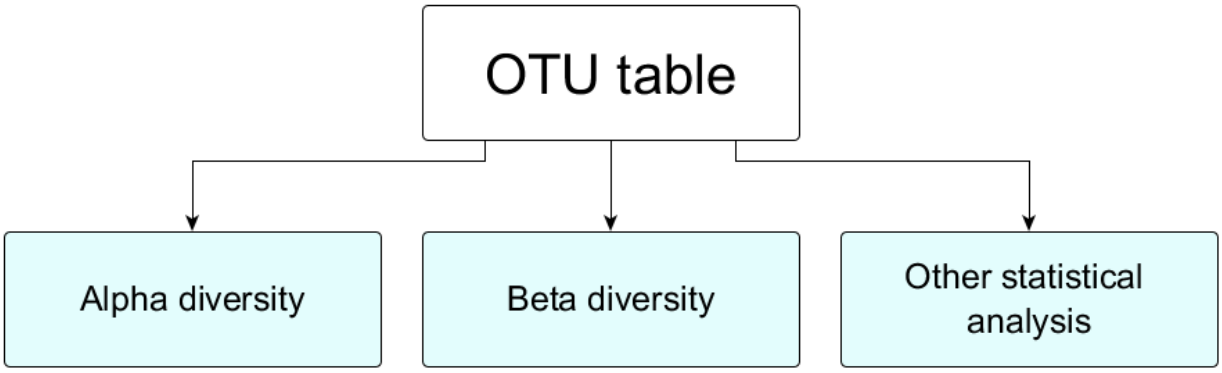


Figure 9 16S rRNA gene data analysis

The composition and structure of the microbiota in each sample can be represented through calculation of α-diversity and β-diversity metrics, or by other statistical analysis for different purposes (e.g. identifying differentially abundant taxonomic groups between sample groups).

Alpha diversity

From ecological point of view, alpha diversity is a measurement of the diversity of species within a sample. The sequencing depth (number of reads per sample) mainly affects the accuracy of the measured alpha diversity. Samples with a higher number of reads would show higher diversity than samples with a lower number of reads. This uneven sequencing depth might cause a bias in the interpretation of the results and lead to a misconstruction of the conclusions. Thus, read number normalization is important before going into any statistical analysis. Rarefaction analysis on the sample can find out whether the sequence depth is sufficient. If the sequence number is too low, one might discard the sample, or sequence it again to get better quality. The alpha diversity can also be divided into two categories: richness diversity and evenness diversity. Richness diversity estimators estimate the total number of species present in a community. Chao1[100] is the most widely used richness index in the microbial community. In contrast to richness diversity, evenness diversity measures the relative abundance of the different species within the sample. It takes into account the number of present species, as well as the abundance of each species. The Shannon index (entropy)[101] and Simpson index[102] are the two most popular evenness diversity indices in the microbial literature.

Beta diversity

The general definition of beta diversity is the distance, or dissimilarity, between each sample pair. Different beta diversity indices give different weights to rare species, in order to emphasize the role of rare species in the microbial composition between two sites or communities. Jaccard index is a useful measurement of calculating the overlapping species between two samples. It only takes into account the presence-absence of the species or OTU, giving the rare and abundant species the same weight in the index. Unlike the Jaccard index, Bray–Curtis dissimilarity quantifies the compositional dissimilarity between two samples based on the bacterial abundance. Beta diversity is bound between 0 and 1, where 0 means the two samples have an identical composition, and 1 means the two samples do not share any species. By using different pairwise matrix of the beta diversity metrics, one can visualize the relative distance between all the samples in different ways, such as a tree or graph. In this thesis, Jaccard and Bray–Curtis were employed as the different distance bases in principal coordinates analysis (PCoA) to explore the similarities for presence-absence and abundance of the microbial data.

Other statistical analysis

Additional statistical analyses, such as plot for bacterial abundances, PERMANOVA, non-parametric Wilcoxon test, and correlations with other attributes, were also included in the 16S rRNA gene data analysis. In this study, all statistical analysis were undertaken by the statistical computing software R (https://www.r-project.org).

2.2.2 RNA-Seq data preprocessing and analysis

RNA-Seq is the most widely used method for estimating gene expression. It reveals the presence and quantity of RNA in a biological sample and takes into account the transcriptional heterogeneity among cell types, as suitable for my purpose. The pipeline of RNA-Seq analysis is very similar to RRBS (section 2.2.4), but compared to RRBS, RNA-Seq is more straightforward (Figure 10).

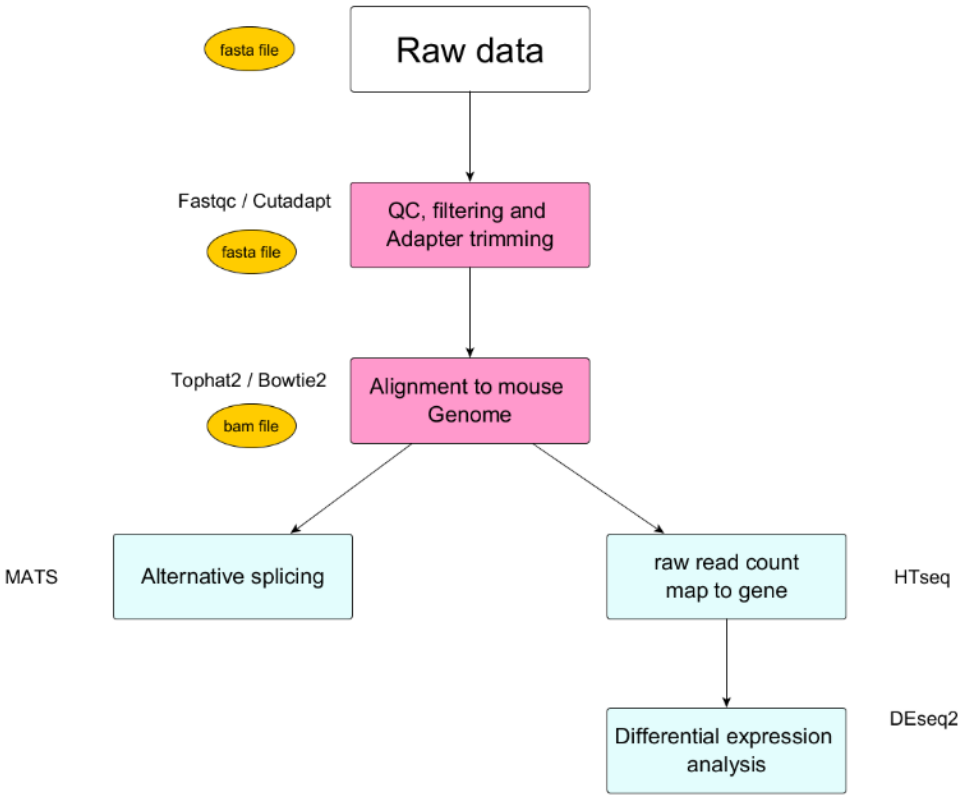


Figure 10 RNA-Seq data preprocessing and analysis pipeline

Starting from the raw reads, the pink colored components correspond to the preprocessing part of the pipeline and the blue colored components correspond to the downstream analysis of the pipeline.

Quality control

The quality control step can perform a quality check of the data to get an idea of whether or not the experiment worked as expected. FastQC checked the general sequencing information, such as the distribution of reads at each position, GC bias, ambiguous N bases, sequence duplication, adapter or primer contamination etc. FastQC is a widely accepted tool to visualize the quality of the sequences (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). In a normal situation, the duplication rate should not be higher than 80%. The other important stage is read trimming. If contamination was observed from the presence of adapters or primers in the tail of the

reads, trimming is necessary. To do this, Cutadapt (https://pypi.python.org/pypi/cutadapt/) was introduced for sequence trimming.

Mapping to reference genome

Alignment is a crucial step to find the location from which the reads originated. The greatest challenge of RNA-Seq transcriptome alignment comes from the eukaryotic gene structure, which contains introns, insertions, deletions, alternative splicing, presence of pseudogenes and gene fusions. These diverse situations might lead to incorrect alignments. Any mapping algorithm must be able to handle these huge gaps (splice sites) or other sources of error. Alignment or mapping to the reference genome for RNA-Seq data was performed using Tophat2[103], which incorporates Bowtie2[104]. Bowtie2 extends the full-text index–based approach of Bowtie for efficient alignment in RNA sequences. After alignment, BAM files (the binary format of Sequence Alignment/Map format) were generated automatically from Tophat2. BAM files from Tophat2 mapped with mouse genes from the reference genome mm10. Ambiguous bases (Ns) in the reference, and sequences marked as duplicates, were ignored for the calculation. HTseq[105] is then performed for the processed RNA-Seq alignments for differential expression calling. HTseq generates the counts for each gene and the number of aligned reads overlap its exons. These counts can then be used for gene-level differential expression analyses, such as DESeq2[106], which was used in this study.

Alternative splicing

RNA-Seq has revealed an enormous complexity of alternative splicing (AS) across diverse cell and tissue types. There are five main classifications of AS types: skipped exon , alternative 3’ splice site, alternative 5’ splice site, mutually exclusive exons, and intron retention[107](Figure 11). The differential alternative splicing tool: rMATS[108] (repeat multivariate analysis of transcript splicing), which is specific for replicate RNA-Seq data was employed in this study. It performs a hierarchical framework to model exon inclusion

levels, this hierarchical model estimates uncertainty in biological replicates and variability across replicates.

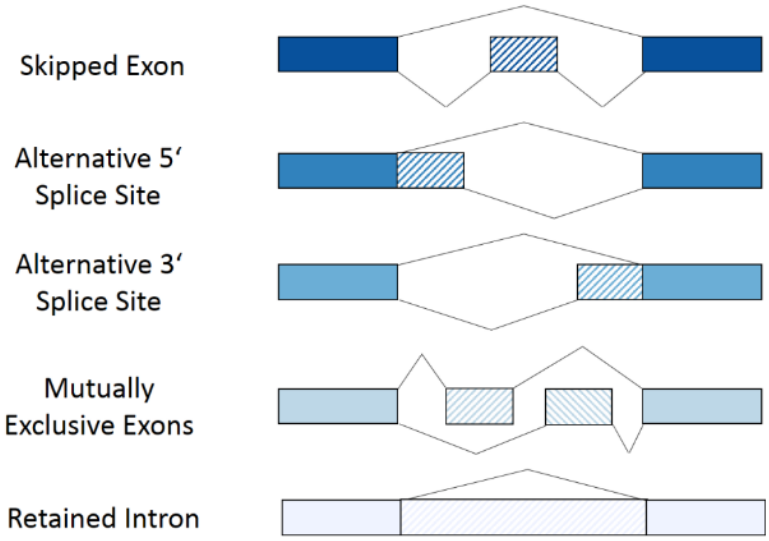


Figure 11 Five alternative splicing categories

The five main classifications of AS types demonstrated as exon skipping, mutually exclusive exons, alternative 5‘splice site, alternative 3‘splice site, and intron retention.

Differential expression analysis

With the improvement of high-throughput sequencing technologies in genomic studies, the need of statistical methods for differentiating genes increases. High uncertainty of within-group variance is the main challenge while detecting the differential gene expression. This can be overcome by measuring information across genes, specifically, by estimating the similarity of the variances of different genes measured in the same experiment. Several tools, such as edgeR[109], DSS[110] and DESeq2[106] all handle this issue in different ways. EdgeR moderates the dispersion estimate for each gene towards a common estimate across all genes by using weighted conditional likelihood. DSS uses a Bayesian approach to estimate the dispersion for individual genes, while considering the heterogeneity of dispersion values for different genes. DESeq2 not only detects dispersion estimates, but also corrects dispersion in low-read transcripts for averaging expression

strength over all samples[106]. For this analysis, DESeq2 was chosen based on a better false discovery rate (FDR) in the larger sample size and the outliers[111]. The expression counts were normalized by library size in DESeq2 for differential expressed analysis.

2.2.3 Affymetrix microarray data preprocessing and analysis

Affymetrix Human Genome U133 Plus 2.0 array was developed in 2003. A microarray contains oligonucleotide “probes” that bind to mRNA from a sample. There may be numerous probes from the coding regions of any given gene. This array provides comprehensive analysis of genome-wide expression on a single array. Quality assessment is essential in microarray and many probes need to be discarded during quality control processing. There are two major issues to be addressed during data preprocessing: background correction and normalization.

Background correction

Probe signal intensity is measured by auto fluorescence of the array surface, and also by some unspecific sources. Background correction methods estimate the background portion of the probe signals and subtract it accordingly[112]. Affymetrix chips are so dense that there is no space between two probes. Hence, it is not possible to measure the signal in the surrounding area, so the background noise must be estimated from the probe signals themselves. In the chip design, one can get two measurements: prefect match probes and mismatch probes. A perfect match probe matches a strand of cDNA, while the corresponding mismatch probe differs from the perfect match by a change in the central nucleotide. The MAS 5.0[113] method developed by Affymetrix averages over regions in the array for both perfect match and mismatch probe cells. This algorithm builds a hierarchal model, which is used to design robust estimators for comparative experiments.

Normalization

A chip effect is a bias of the raw probe signal measurement. This effect is chip specific and influences all probes on a given chip in a similar manner. It is caused by varying total RNA abundances, labeling and hybridization efficiency, scanner properties, and many other sources of variation. Non-biological factors can also contribute to the variability of the data. In order to compare data from multiple probe arrays reliably, differences of non-biological origin must be minimized. The purpose of normalization is to remove all non-biological effects. After normalization, variability between the different arrays is reduced and the changes in expression values becomes more reliable. R package GCRMA[114] is employed for normalization method in this study. The main function of GCRMA is to convert background adjusted probe intensities to expression measures using normalization and summarization methods, such as robust multi-array average[115].

Statistical test

In the twins study, non-parametric Mann-Whitney U-test was performed to determine the differential gene expression. Multiple testing correction was performed using the Benjamini-Hochberg[116] method. The criteria for transcripts to be categorized as differentially expressed for corrected was p-value ≤ 0.05.

2.2.4 RRBS data preprocessing and data analysis

In general, RRBS data preprocessing might be challenging compared to the other techniques. Firstly, removal of duplicate reads might not be suitable in RRBS data analysis due to the CpG enrichment in the genome. However, the difference between PCR duplicates and enrichment-based duplication cannot be distinguished. Secondly, the failure of the bisulfite conversion rate might affect the data quality, thus spike-in controls is necessary for measuring the bisulfite conversion efficiency and to adjust methylation levels for each sample. Thirdly, RRBS samples are very sensitive to any technical biases, such as the concentration of the sample, library preparation and sequencing batch, mice litter, etc. Thus, one should be very careful from the beginning of the experiment to avoid any possible bias, especially the batch effect. Last, but not least, the resolution of the

methylation level is down to each single CpG site. Therefore, annotation is a very essential process. One needs to annotate all the positions to the available genes (intron, exon, gene body etc.), and regions (CpG islands, enhancer, shores, shelves, other regulatory regions etc.). The known reference genome (human or mouse) from a reference genome browser, such as UCSC or Ensemble, could be used in annotation. Thus, RRBS might not be a proper method for the methylation level measurement of a species with an incomplete genome reference. The preprocessing pipeline is visualized in Figure 12. Here, several packages were employed for this large-scale analysis. Incidentally, with this sophisticated data structure, one should also consider the heavy computational work and machine capacity.

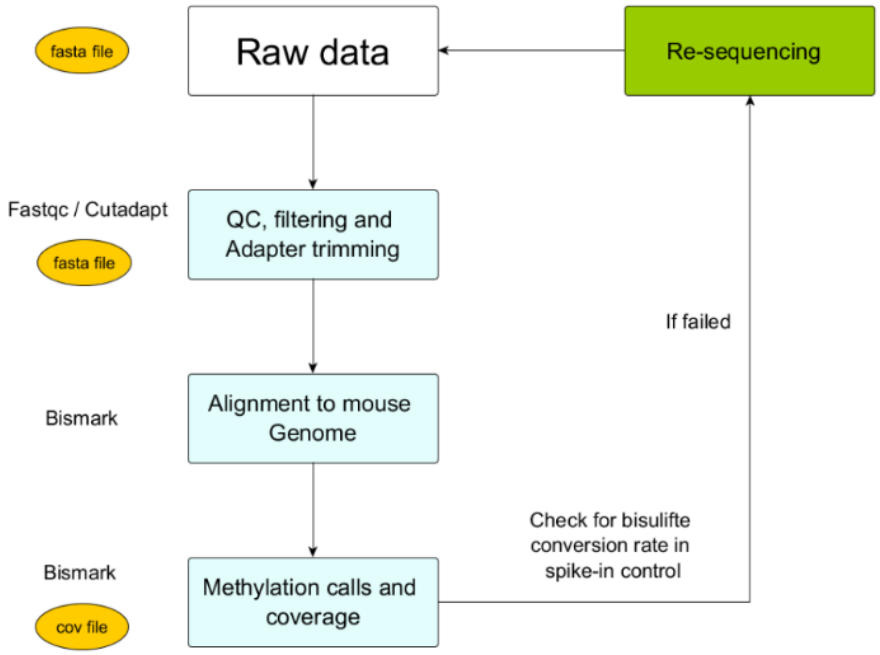


Figure 12 RRBS data preprocessing pipeline

Data preprocessing starts from the raw data (*.fasta) and ends as a *.cov file, which includes the coverage and the methylation rate.

Quality control

The quality check procedure for RRBS data is almost the same as the procedure used for RNA-Seq data. Here, the quality control process mainly followed the instructions from

Epigenesys, which was built by Felix Krueger and Simon R Andrews[117]. As mentioned above, the duplication rate should not be higher than 80% in a normal situation. An abnormal high duplication rate might due to the PCR duplication, which should probably be removed before commencing with downstream analysis. However, a high duplication level of 95% might be reasonable in RRBS library, since all fragments are expected to line up perfectly at exactly the same genomic location numerous times (there are only so many MspI recognition sites in a genome). Therefore, the duplication rate check could be temporarily ignored in RRBS library.

Alignment and methylation calling

Bisulfite converted reference data will have all unmethylated Cs converted to Ts, therefore it consists of three nucleotides: A, T and G. The aligned base T could be an unmethylated C or the original genomic base T. This characteristic further raises the complexity of alignment, and might create ambiguous results. In this study, Bismark[118] was selected, among several alignment approaches, to handle this challenge. In Bismark, the bisulfite reads are first transformed into a C-to-T and G-to-A (reverse strand) version. Then, each of them is aligned to equivalently pre-converted forms of the reference genome using four parallel instances of the short read aligner Bowtie2 (Figure 13). Thus, all possible outcomes can be considered in the alignment process. The first mapping output of Bismark contains one line per read and provides information on the mapping position, alignment strand, bisulfite read sequence, its equivalent genomic sequence, and a methylation call string. This initial output can be converted into BAM files, or imported to a genome browser for visualization for further research purpose by the users.

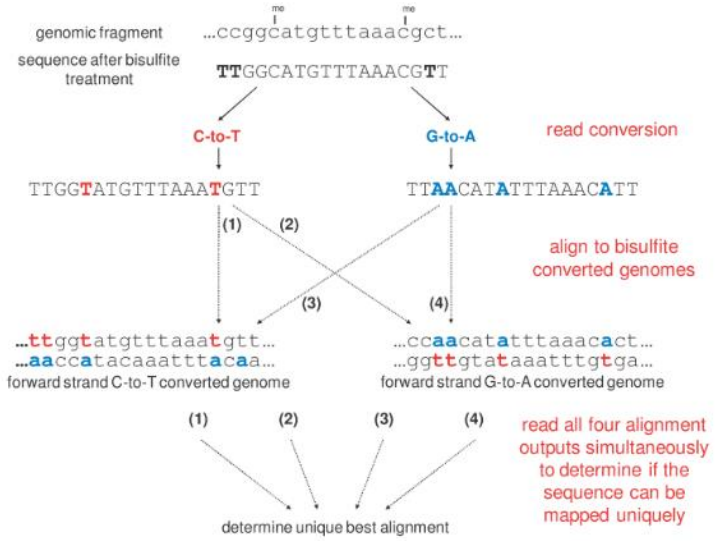


Figure 13 Bismark's approach to bisulfite mapping and methylation calling

BS reads are converted into a C-to-T and a G-to-A version and are then aligned to equivalently converted versions of the reference genome. The best unique alignment is then determined from the four parallel alignment processes. Figure modified from: Krueger, Felix and Andrews, Simon R[118]

After alignment, Bismark automatically calculates the methylated and non-methylated number in each detected CpG site. Ultimately, one can get the methylation-calling file with information on the CpG positions, methylated reads number, non-methylated numbers, and the methylation rate (percentage; number between 1 to 100) in each row. From this, one can calculate the coverage, which is a criterion for filtering the data with an alignment error in this initial step. There are three main steps in RRBS data analysis: sample merging and site filtering, differentially methylated site analysis and annotate to reference genome (Figure 14).

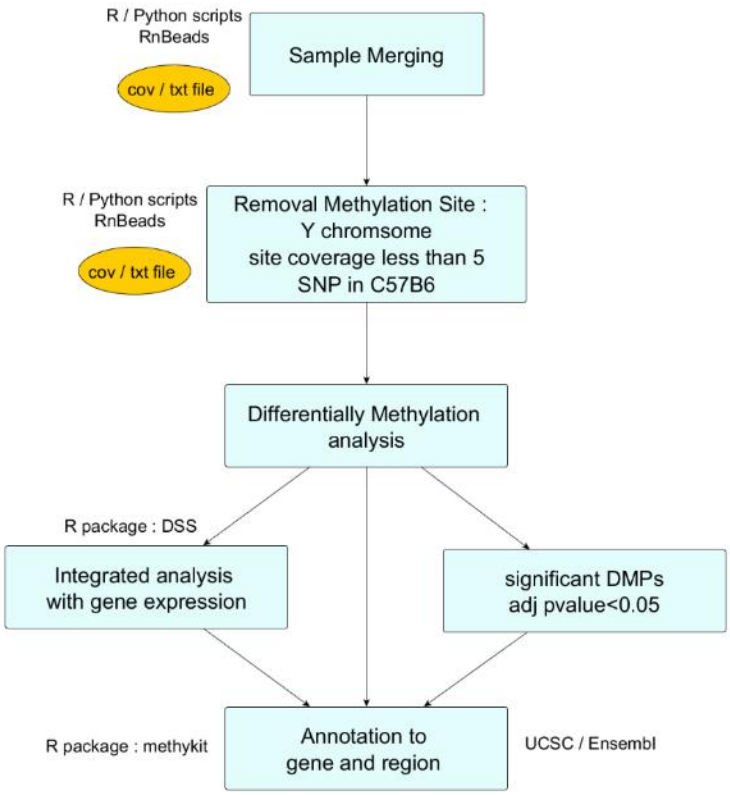


Figure 14 RRBS data analysis pipeline

Downstream analysis of the *.cov file from preprocessing (Figure 12).

Sample merging and site filtering

In order to have a global view of the data structure, methylation information of the CpG sites across all samples need to be merge as a matrix. Memory efficient R package RnBeads[119] provides the merge function specific for methylation data. Due to the huge number of missing value, multidimensional scaling (MDS) was performed for dimension reduction and pattern recognition problem. Unlike microarray data, the restriction enzyme Msp1 in RRBS only enriches the CCGG sequence; therefore, some CpG sites might be missing in certain conditions. Furthermore, the CpG sites, which aligned to the Y chromosome were removed by the female-only experiment design in this study. Additionally, the overall noise was greatly reduced with increasing coverage for each CpG site. Thus, the accuracy and statistical power of the results, strongly depends on the coverage of the methylation call[120]. However, the arbitrary cutoff depends on the quality

of data. Threshold of 5 was employed in this study, all CpG sites with coverage less than 5 were removed. The last stage of preprocessing was SNP removal. The presence of SNPs inside the data can have important consequences, or mislead the downstream analysis, therefore, SNPs of C57BL/6(N) mice strand were removed. The SNP information is from mouse genome project (http://www.sanger.ac.uk/science/data/mouse-genomes-project).

Differentially methylated site analysis

There are several free tools for identifying differentially methylated positions/region. RnBeads[119], which is already included in previous merge stage, computes p-values by using hierarchical linear models from the limma[121] package. It assumes that the number of reads from each single CpG site follow normal distribution. This assumption might not fit the real situation. The counts are binary (methylated or unmethylated) and sequenced with prior probability. Based on its character, the beta-binomial model is the first choice. Here, R package DSS[110] (Dispersion shrinkage for sequencing data) was recruited for detecting the DMPs. This method borrowed the dispersion shrinkage approach from differential expression analysis in RNA-Seq and microarray analysis, by taking information from CpG sites across the genome, to stabilize the estimation of the dispersion parameters in lognormal-beta-binomial hierarchical model.

Annotation to gene or genome

For gene category annotation, MethylKit[122] was performed. The methylation sites were annotated from UCSC mm10 mouse genome. The promoter region is defined as a 1,500 base pair window from the center of the transcription start site. The positions were annotated in four gene categories: exon, intron, promoter and intergenic.

Last but not least, it is always necessary to modify parameters of the pipeline according to the dataset, samples used and their quality, which can have some influence on the results.

2.2.5 HumanMethylation27 microarray data analysis

Beadchip is used for methylation probes on the Illumina 27k methylation array. Quantitative measurements of DNA methylation are determined for 27,578 CpG dinucleotides spanning 14,495 genes[123], including nearly 13,000 well-annotated genes in the NCBI CCDS Database (Genome Bulid 36) and over 1000 cancer-related genes, or targets.

Quality Control and methylation measurement

In order to get rid of the noise influence, the machine also calculates the “detection p-value” for each probe., There is no significant difference between the real biological signal and the background noise if the detection p-value is higher than 0.05. The probes with detection p-value > 0.05 will be discarded for downstream analysis. For an individual CpG site, a pair of bead-bound probes is used to detect the presence of T (unmethylated state) or C (methylated state). In default, a beta value is used to measure the methylation level of a single CpG site, and is subsequently determined by calculating the ratio of the fluorescent intensities of the methylated probe (M) and unmethylated probe (U) according to the following formula:

$$\beta = \frac{\max(M, 0)}{\max(M, 0) + \max(U, 0) + 100}$$

Here, max(M, 0) and max(U, 0) indicate the maximum value between M and 0, and U and 0, respectively. The number 100 in the denominator is a constant offset to standardize beta values when both methylated and unmethylated probe intensities are low. The Beta value has a direct biological interpretation - it corresponds to the methylation percentage of a single CpG site. However, because the Beta-value has the upper bound 1 and lower bound 0, this statistic violates the Gaussian distribution assumption. Furthermore, from

an analytical and statistical standpoint, the Beta-value method has severe heteroscedasticity outside the middle methylation range, which imposes severe challenges in applying many statistic models[124]. Thus, instead of the Beta-value, M-value was invented by Pan Du and colleagues[124].

$$\mathrm{M} = log_2\left(\frac{\max(M,0)+1}{\max(U,0)+1}\right)$$

Although the M-value statistic does not have an intuitive biological meaning, it is possible to provide an accurate estimation of the methylation status by modeling the distribution of the M-value statistic.

2.2.6 Integrated analysis in mouse study

With the rapid development of NGS in different omics scale, integrated analysis in multiple omics would reveal novel biological hypotheses involving complex interactions among the different conditions. For other genomic statistical methods (e.g. GWAS, eQTL, mQTL), multiple comparisons, or association issues, also exist in traditional transcriptome-methylome studies. Enormous comparisons with large noise weaken the statistical power. Hence, people use FDR (e.g. Bonferroni or Benjamin–Hochberg correction) or stricter standards (p-value<10^{-5} in GWAS) to filter out the signals with less significance. However, this approach ignores the fact that the two dimensions of multiplicity are not equivalent. Instead of massive calculations in the whole genome, the integrated analysis in mouse study focuses on the methylation patterns, which are close to the differentially expressed genes. The hierarchical testing approach[125] (Figure 15) was applied to identify interactions between the microbiota dependent alterations in the transcriptome and DNA methylation signatures. Hierarchical testing approach controls overall FDR at a set level, and also controls for mixed-directional FDR at the individual level. By using this two-step approach,

DMP and DMR were detected sequentially, and it can furthermore avoid FDR penalty in different biological signals especially for this small sample size comparison (5 vs 5).

All CpG sites of 5 kb up- and downstream of the transcription start of the microbially regulated genes were identified. Then, the neighborhood methylated positions were combined to methylated regions (maximum distance 200 bp). The regions with only a single point, or which contained less than 20% CpGs (p-value < 0.05), were excluded and all retained regions were considered as differentially methylated regions. In the end, the Benjamini–Hochberg procedure was performed to correct the FDR for retained regions (adjusted p-value<0.05). This filtering approach was well suited and beneficial for both DMR and DMP detection. Moreover, the essential biological signals can only be explored by avoiding the over-multiple correction.

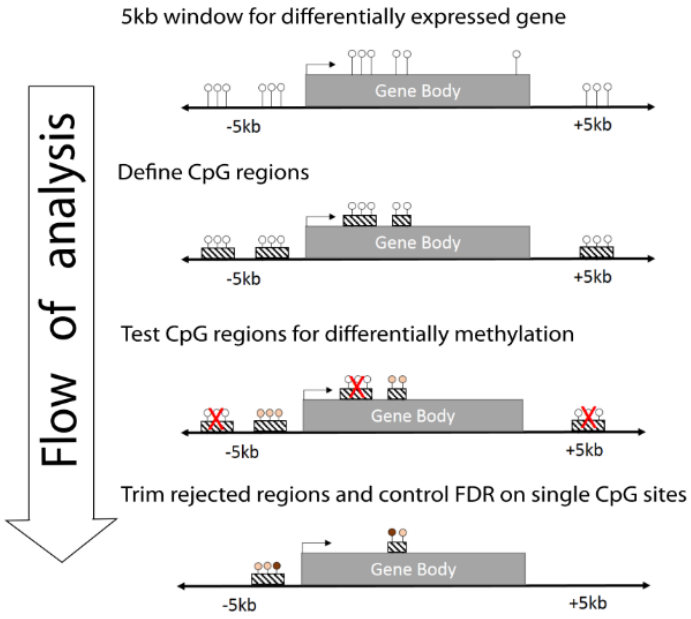


Figure 15 Hierarchical testing approach

Integrated analysis workflow. The hierarchical testing procedure was employed to detect the interactions between gene expression and methylation in three developmental stages.

3. Results

This thesis encloses the results of two studies driven and interconnected by the used technologies and pipelines. The main findings of these two studies are described as stated in the following manuscripts.

3.1 Twins study

3.1.1 Study design

The etiology is largely unclear in intestinal disease, however, genetics and environment has been considered to play the crucial roles for the pathogenesis. More and more studies[49] revealed the importance of epigenetic modifications, it represents a major interface between internal and external factors. **The main aim of the study was to present a high-resolution map of epigenetic modifications and host microbiota profile with potential disease relevant effects on the host transcriptome in ulcerative colitis**. The underlying hypothesis of the study was therefore, that epigenetic change of UC-relevant genes results in altered gene expression and host microbiota composition modification with pathological consequences, contributing to disease mechanisms. Integrating these three layers into the current picture of ulcerative colitis disease pathophysiology may close the gap between genetic susceptibility, missing heritability and disease manifestation. Therefore, the mucosal biopsies of colon were sampled from monozygotic discordant twins for UC. RNA and DNA were isolated to study gene expression and DNA methylation pattern in relation to healthy and disease status. Subsequently, the observed difference in gene expression and microbial groups were validated in an independent cohort consisting of 20 UC patients and 20 healthy controls (Figure 16).

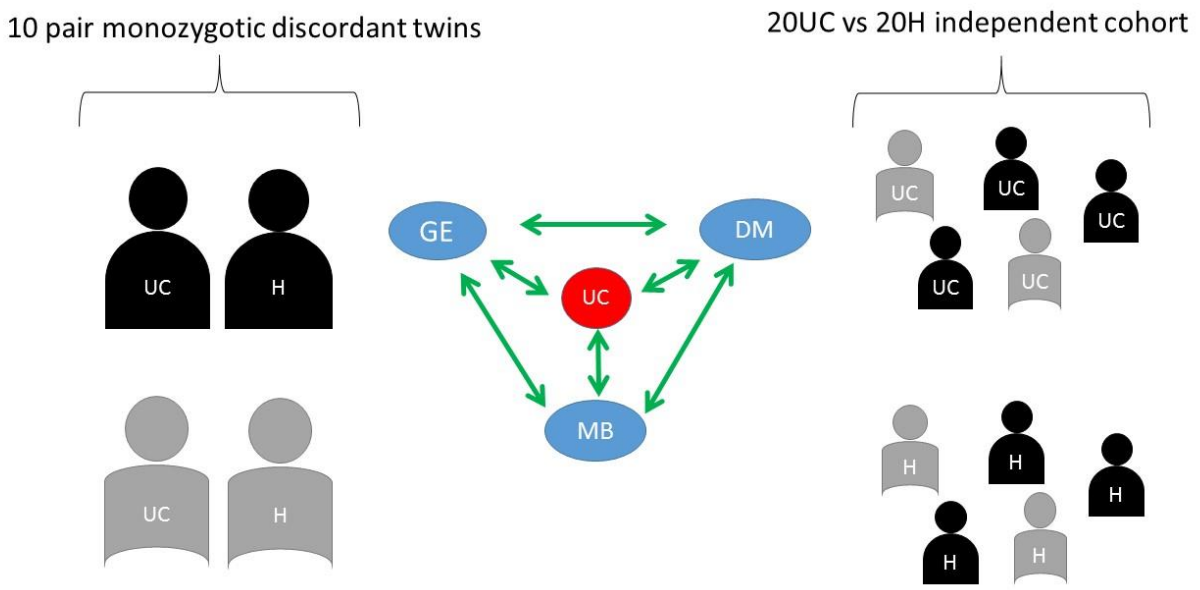


Figure 16 Study design of twins study

UC: ulcerative colitis patients; H: healthy control; GE: gene expression; MB: microbiota; DM: DNA methylation. The layout of the strategy was to investigate the gene expression, methylation and host microbiota interaction in ulcerative colitis.

3.1.2 Microbiota profile

After data preprocessing and normalization for sequencing depth (mentioned in 2.2.1), 292833 16S rRNA gene sequences resulted from five pair of twins. Analysis was performed in the host and microbial data which are available for both twin subjects. 211 OTUs were classified up to genus level by using Silva database (https://www.arb-silva.de). The relationship between age and alpha diversity was first investigated, the difference of Shannon diversities between healthy and UC discordant twin subjects increases along with the time (Figure 17).The diversity of richness and evenness between healthy control and UC patients were compared (Figure 18). Either in richness or evenness, the diversities in healthy individuals are generally higher than in UC (mean estimator of Healthy control vs UC patients, Chao1 index: 150.26 vs 109.24; Shannon index: 3.31 vs 2.77). Due to small sample size, the mean differences were not statistically significant,

although the diversity decrease pattern are obvious (Figure 18). PCoA plot was performed microbiota composition for observing the relationship between disease status and kinship. Bray-Curtis dissimilarly was used for PCoA analysis. This dissimilarly quantifies the compositional dissimilarity between two different sites based on read counts at each site (Figure 19). The first and the second principal components explained 22.89% and 16.47% variation of the composition respectively. The twin pairs were connected by solid lines in Figure 19. The pairs located in the proximity compared to the distance of unrelated individuals, this linking pattern showed the tight connection of the genetic effect on microbiota composition. SIMPER analysis[126] (Similarity Percentages analysis) was applied to find the contribution of each species between two conditions. It calculates the contributions of similarities among sample groups, and provides the variable importance percentage of contributing similarity between factors. 60 bacterial genera were selected as the cutoff of 90% variation.

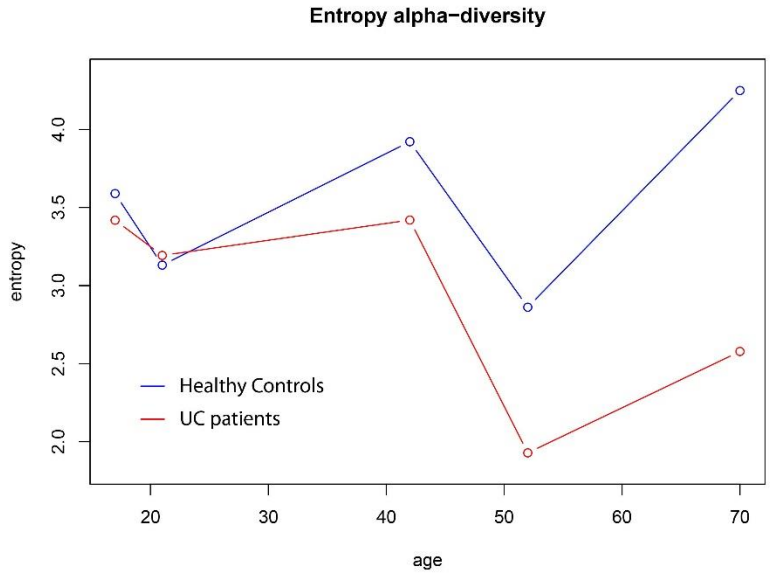


Figure 17 Shannon Entropy from twins microbiota

Every single circle indicates every individual. The difference between healthy and UC discordant twins increases along with the age.

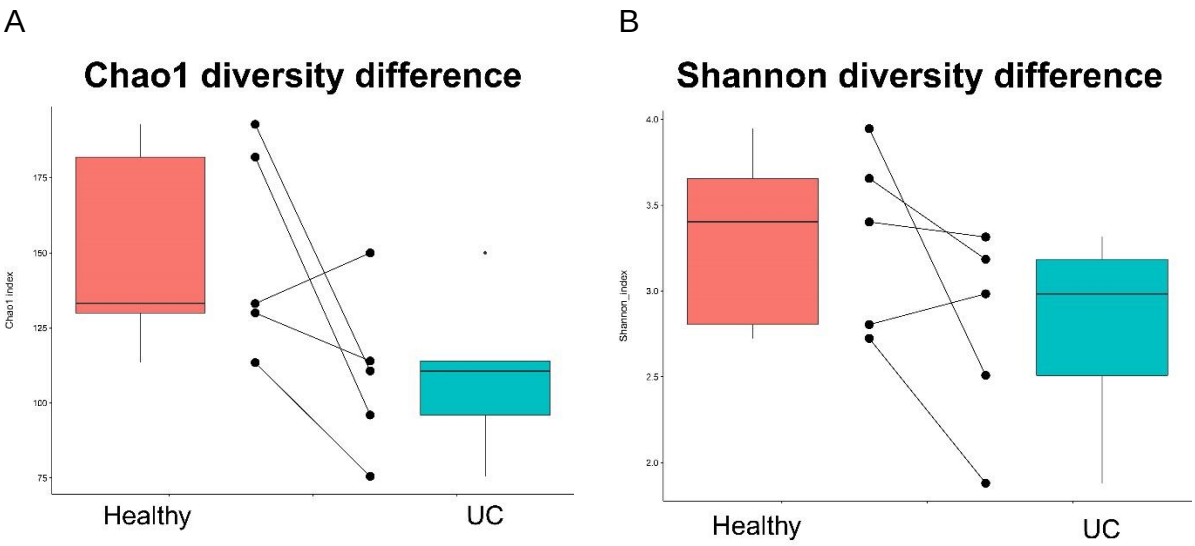


Figure 18 Alpha diversity for Healthy and UC in twins study

Here are boxplots of alpha diversity indices for healthy and UC discordant monozygotic twins. The connected points are from the same pair. (A) Chao1 index indicates the estimated bacteria species number of two conditions (B) Shannon index indicates the evenness of microbiota composition. Chao1 and Shannon index decreases generally in UC patients, except one pair.

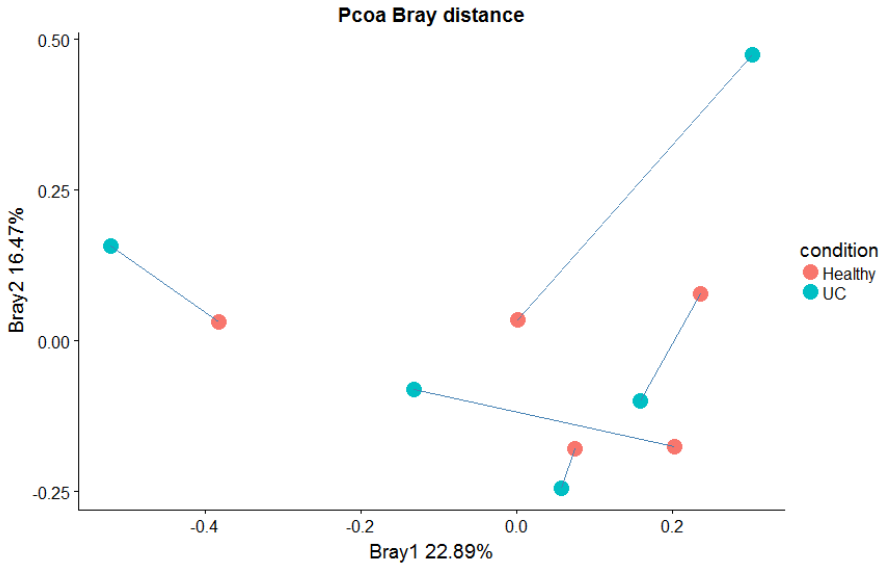


Figure 19 PCoA plot with Bray-Curtis distance in twins study

The connected points indicate the twin pair. The twin pairs in the figure located in the close area.

3.1.3 Gene expression and DNA methylation

The transcriptome analysis was performed from 20 monozygotic twins, discordant for UC. 19,025 (out of 54,675) transcripts were considered as expressed. This study focuses on the genes with particular functions, the genes with descriptions including keyword *Defense*, or including the words which start from *inflame* and *immune* from GO database (Geneontology.org) were selected for the following analysis. Based on this filtering criterion, 967 transcripts (584 genes) were kept for downstream analysis. Furthermore, 117 transcripts (97 genes) were differentially expressed (t-test p-value<0.05; fold change FDR<0.05) between healthy and UC subjects. GO analysis was performed to detect the encoded biological processes: regulation of inflammatory response (p-value=6.90E-4) and positive regulation of cell junction assembly (p-value=7.22E-4) were significant as the related biological process for the candidate genes. Hierarchical clustering was further performed to identify gene clusters, resulting in a heatmap (Figure 20). UC patients clustered together, as shown in the left side of the figure. The kinship is also partially evident in the figure, some twins also clustered together (UC_07a, H_07b & UC_12a, H_12b).

These selected genes were then compared with the result of similar experiment setting from Kugathasan S and colleagues[127]. Kugathasan's study investigated RNA expression from colon biopsy in unrelated 10 UC pediatric patients and 11 healthy controls by using microarray. Based on the same filtering criteria and the statistical test, 456 transcripts (337 genes) were identified as differentially expressed between healthy and UC subjects in Kugathasan's study (Figure 21). With comparing the results in both studies, there were 64 genes overlap between two studies.

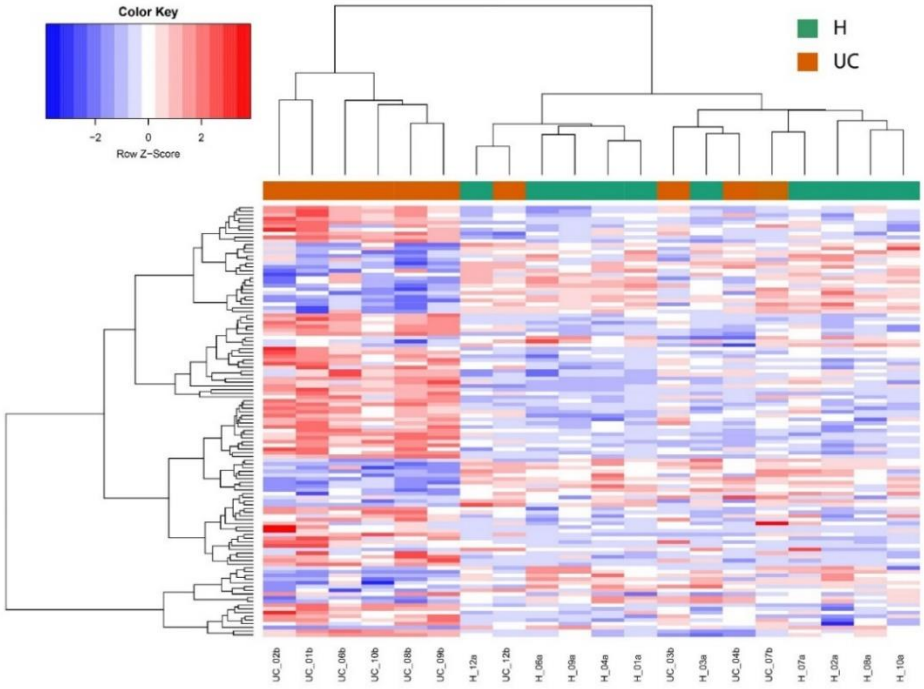


Figure 20 Heatmap of differentially expressed transcripts

H: healthy control groups, UC: ulcerative colitis patients. The numbers followed the disease status in the column labels were the label of twin pairs. The subjects that shared the same number were the twin pairs.

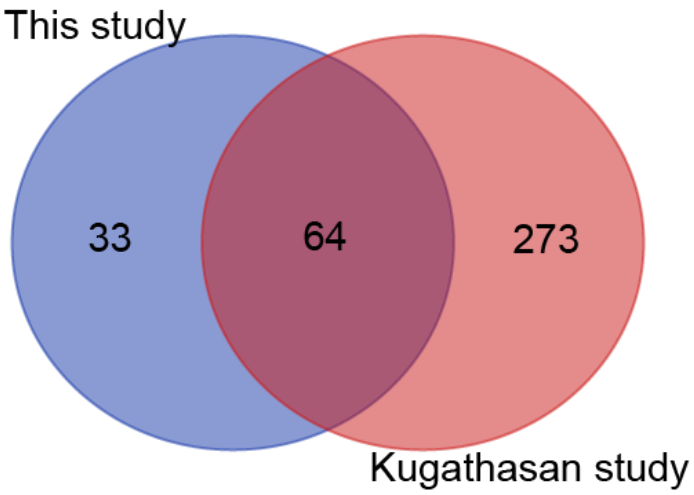


Figure 21 Differently expressed gene in both studies

The number in the circle means the number of differently expressed genes. This twins study recruited 10 pair discordant UC twins while Kugathasan recruited 10 UC and 11 healthy controls.

Furthermore, for the DNA methylation analysis, out of 27578 CpGs, 23085 were detected in this dataset. Among all informative CpGs, 382 methylation positions located in close proximity (±50 kb) to these 117 differently expressed transcripts. The correlations were calculated in methylation position and transcripts, to investigate the epigenetic modification influencing of gene expression. High correlation (spearman correlation |ρ| >0.6) was seen in 18 methylation positions with 17 transcripts (15 genes: *S100A8, AGT, IRAK2, TNFSF10, CXCL6, C2, CFB, CCL24, LYN, PRDX5, OAS1, ISG20, ABR, CCL11, PRDX2*). Regulation of inflammatory response showed statistical significant in GO analysis for these selected 15 genes.

Transcript ID	Gene name	cg number (methylation)	chromosome	cg position	correlation
202917_s_at	S100A8	cg02813121	1	151615535	-0.752941176
202834_at	AGT	cg06585893	1	228950175	-0.758823529
231779_at	IRAK2	cg20916523	3	10159584	-0.785294118
202688_at	TNFSF10	cg11979312	3	173725407	-0.602941176
206336_at	CXCL6	cg02029926	4	74953738	0.714705882
206336_at	CXCL6	cg22670329	4	74920939	-0.644117647
203052_at	C2	cg09583599	6	32033899	-0.629411765
202357_s_at	CFB	cg01883966	6	32046865	-0.638235294
221463_at	CCL24	cg05556717	7	75257240	-0.714705882
202626_s_at	LYN	cg03973663	8	56954130	-0.620588235
222994_at	PRDX5	cg11296937	11	63841542	0.708823529
1560587_s_at	PRDX5	cg11296937	11	63841542	0.685294118
1560587_s_at	PRDX5	cg13412615	11	63814383	0.723529412
205552_s_at	OAS1	cg19789466	12	111829306	0.605882353
33304_at	ISG20	cg08491125	15	86982946	-0.65
204698_at	ISG20	cg08491125	15	86982946	-0.694117647
212895_s_at	ABR	cg25374854	17	1030708	0.788235294
210133_at	CCL11	cg11860203	17	29606629	-0.785294118
39729_at	PRDX2	cg08694544	19	12807228	-0.638235294

Table 1 Highly correlated expression-methylation genes in twins study

3.1.4 Integrated with microbiota

The microbiota plays a fundamental role on the induction, training and function of the host immune system. After discovering the methylation-related genes, the next goal was to identify the potential bacterial genera which related to these candidate genes. The correlation between all 60 bacteria OTUs and 17 transcripts was calculated. The heatmap from hierarchical clustering was plotted to have an overview of the interaction between host transcript and microbiota (Figure 22). Genes and OTUs can be majorly divided to two groups (Gene group1: *ABR2, PRDX2, PRDX5*; Gene group2: *S100A8, AGT, IRAK2, TNFSF10, CXCL6, C2, CFB, CCL24, LYN, OAS1, ISG20, CCL11.* OTU groups: omitted). The OTU group1 on the top in the figure has the positive correlation with gene group1, but negative correlation with gene group2, vice versa for the down OTU group2.

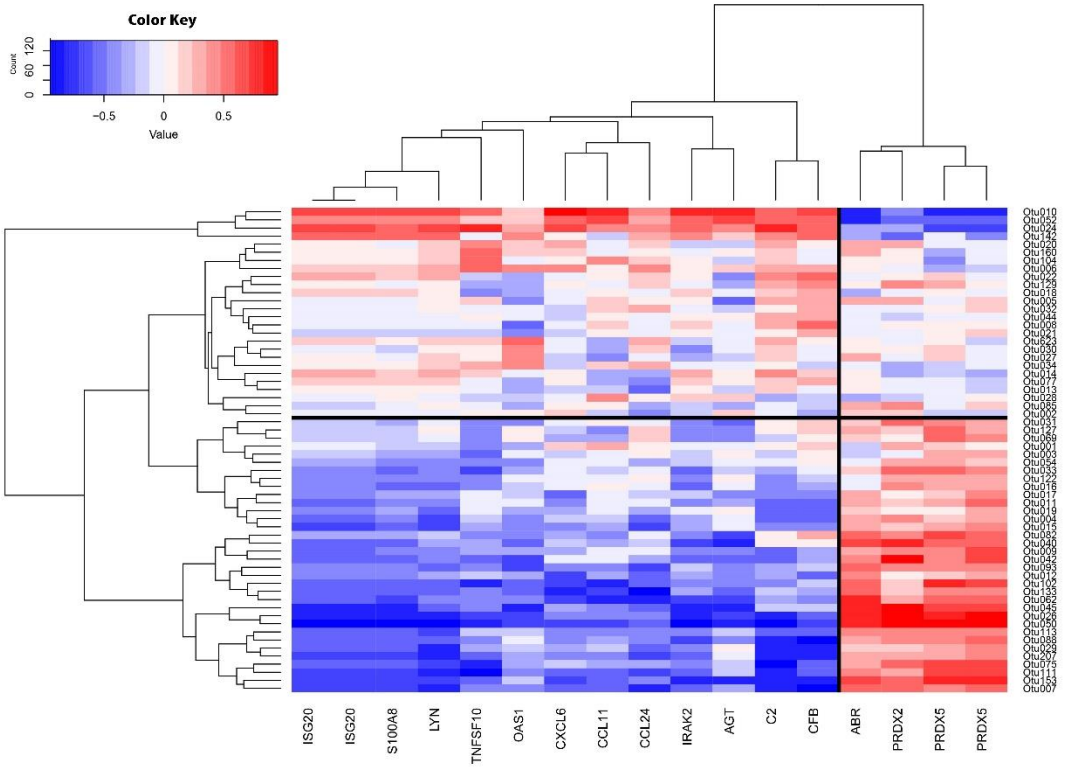
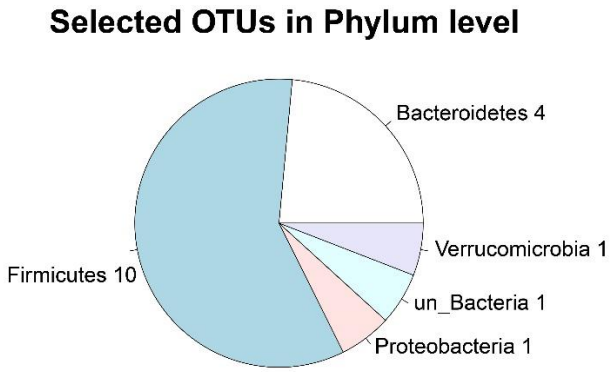


Figure 22 Heatmap showing the correlation between OTUs and transcripts

Each row represents one OTU and each column represents each targeted gene. The genes can be separated to two groups base on the correlation pattern.

Several OTUs were identified to have the stronger correlation patterns with methylation-related genes (more than 6 correlation values >0.5 or < -0.5.) compared to the other OTUs, 17 OTUs were chosen as potential candidates related to methylation mechanism. Ten of them belong to *Firmicutes* and four of them are phyla *Bacteroidetes* (Figure 23). The detail of these OTUs up to genus level was described in Table 2.

In summary, multi-omics approach (transcriptome, methylome and microbiota analysis) has been employed in this study. 17 immune function related transcripts (15 genes), 18 methylation positions and 17 OTUs were identified with the potential interacting with each other in UC. The next step was to validate these findings in independent cohort.



genus	number of OTU
Akkermansia	1
Alistipes	2
Bacteroides	1
Barnesiella	1
Blautia	1
Clostridium_XlVa	2
Dialister	1
Dorea	1
Desulfovibrio	1
Lachnospiracea _incertae_sedis	1
Ruminococcus	1
Sutterella	1
un_Bacteria	1
un_Lachnospiraceae	1
un_Ruminococcaceae	1

Figure 23 Selected OTUs in phylum level

Table 2 Selected OTUs in genus level

17 OTUs were selected as the potential disease and methylation related bacteria. These 17 OTUs belong to 5 phyla and 15 genera.

3.1.5 Validation

15 disease- and methylation-associated genes and 3 bacterial OTUs were subjected to further validation and replication in a larger collection of sigmoid colon biopsies from the intestinal mucosa. This validation for mRNA expression and microbiota were performed by TaqMan-based real-time PCR and methylation validation was performed by amplicon sequencing in unrelated UC patients (n=20) and healthy controls (n=20). This cohort was chosen by the same gender proportion as the twins cohort (20% female and 80% male), however, the age of validation cohort are relative younger than the original panel (age group between 20 to 40). Gene expression analysis was first examined in 15 candidate genes, the one which has the same expression pattern as the original panel then further kept for the methylation and microbiota validation (Figure 24).

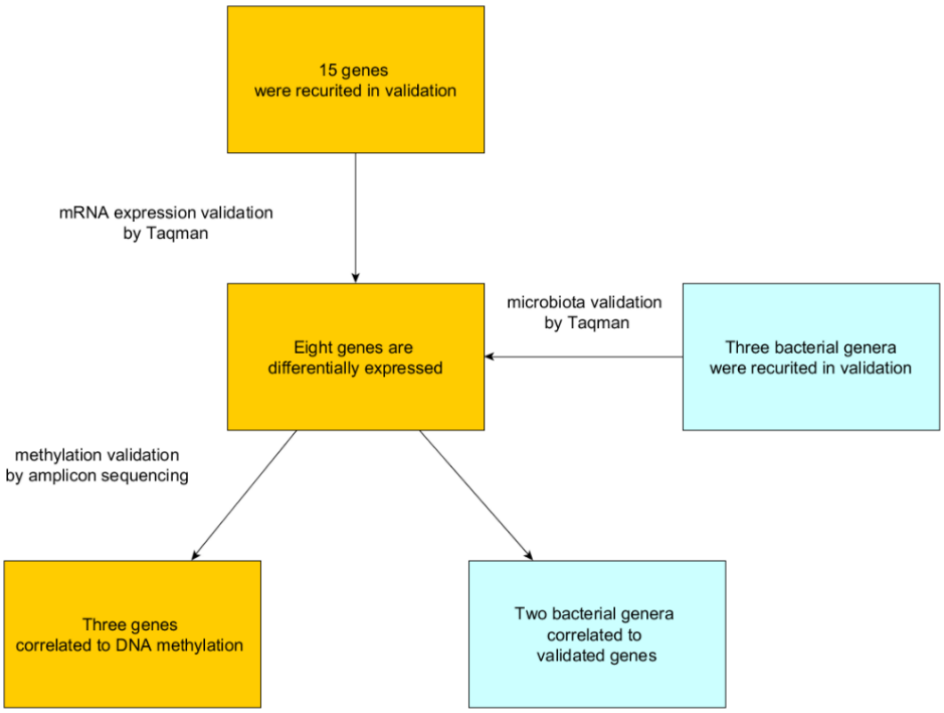


Figure 24 Validation working flow

Gene expression levels were measured by Taqman qT-PCR. After normalization from house-keeping gene ACTB, eight genes (*LYN, TNFSF10, OAS1, AGT, CFB, S100A8* and *CCL11*) out of 15 were validated as differentially expressed between UC and healthy

control (t-test p-value<0.05, same direction for up/down regulation, Table 3 & Figure 25). The expression level of these eight genes were all higher in UC (Figure 25). The amplicon sequencing was further employed to assess the DNA methylation level on the CpG sites around these eight genes. Methylation level were measured in the 200 base window around the microarray-identified CpG sites. *S100A8* was excluded because of the bad sequencing quality. Four CpG sites (in the adjacency area of three genes) were found differentially methylated between healthy control and UC (p-value < 0.05). Besides, the correlation between the methylation level and gene expression for these four CpG sites showed the same direction as previous twins panel (Table 4 & Figure 26). Microbiota information from this cohort were also measured by Taqman and went through the same preprocessing as mRNA expression. Correlation between three bacterial genera (Clostridium_XlVa, Bacteroides, Ruminococcus) and eight differentially expressed genes were calculated. Clostridium_XlVa and Bacteroides showed the similar correlation pattern with the previous panel (Table 5). Clostridium_XlVa had positive correlations with gene expression while Bacteroides showed the opposite direction.

In summary, eight out of 15 genes were validated as differentially expressed; three of them have been identified as epigenetic-associated genes from correlation analysis between gene expression and DNA methylation. Besides, two bacterial genera were found with highly correlation with validated differentially expressed genes in the technically and biologically validation cohort.

Gene name	twins cohort		validation cohort	
	p-value	Up/down regulation	p-value	Up/down regulation
ISG20	0.0044 / 0.0035	up	0	up
LYN	0.001	up	0.0001	up
TNFSF10	0.05	up	0	up
OAS1	0.0025	up	0	up
AGT	0.0027	up	0	up
CFB	0.0124	up	0.0004	up
S100a8	0.014	up	0	up
CCL11	0.002	up	0	up

Table 3 mRNA validation in independent cohort

Eight genes out of 15 were validated in the independent large cohort with the same direction of gene expression (up or down regulation). There are two transcripts in ISG20, both of them showed significant in original panel. Up means higher in UC, down means higher in Healthy.

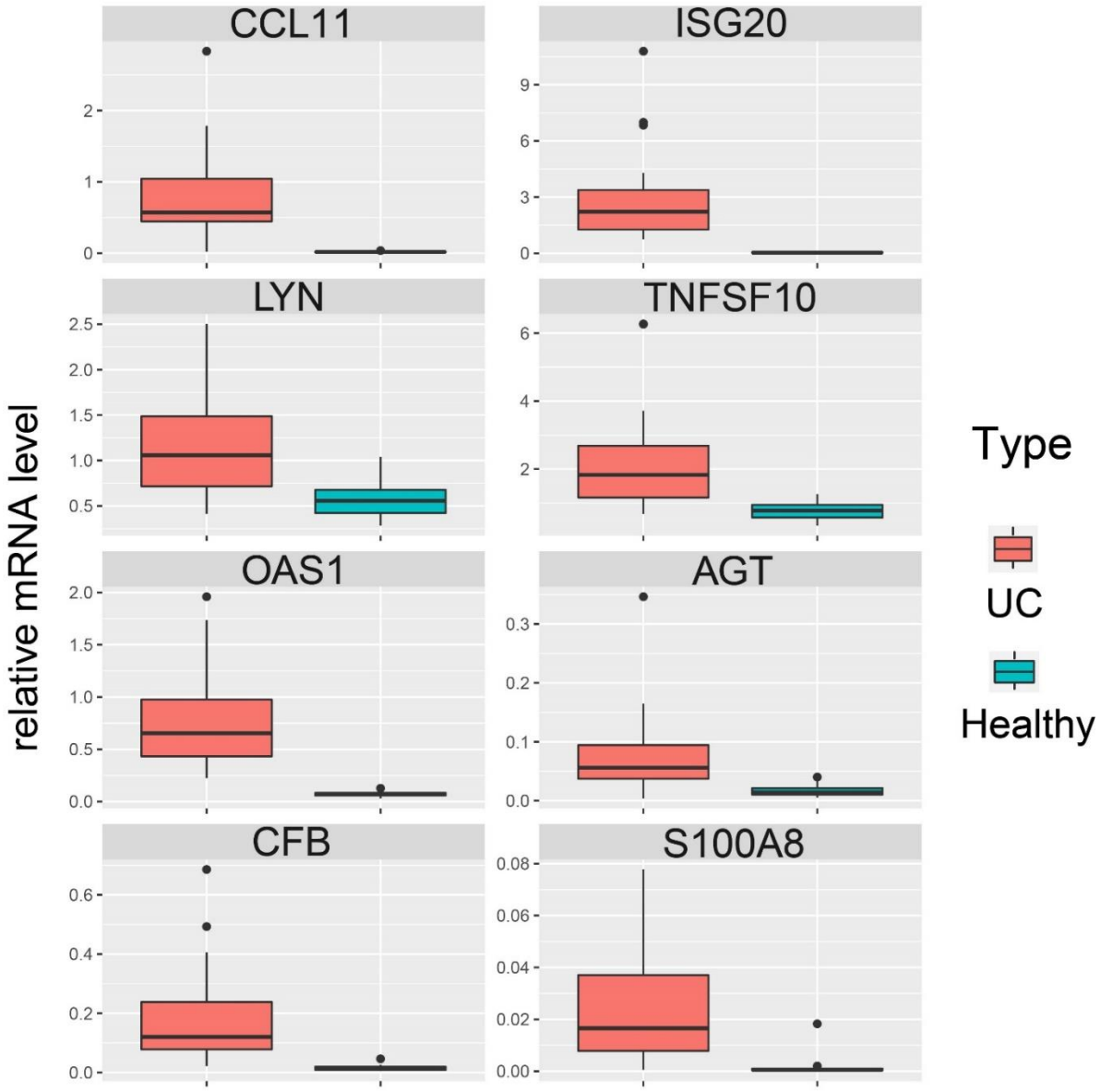


Figure 25 Differentially expressed gene in validation cohort

The gene expression of ISG20 (Interferon Stimulated Exonuclease Gene 20), LYN (LYN Proto-Oncogene, Src Family Tyrosine Kinase), TNFSF10 (Tumor Necrosis Factor Superfamily Member 10), OAS1 (2'-5'-Oligoadenylate Synthetase 1), AGT (Angiotensinogen), CFB (Complement Factor B), S100A8 (S100 Calcium Binding Protein A8) and CCL11 (C-C Motif Chemokine Ligand 11) are all higher in UC in both panels.

Gene name	CpG site	Correlation		P-value
		twins cohort	validation cohort	validation cohort
TNFSF10	cg11979312	-0.6029	-0.2375	0.0122
CFB	cg09583599	-0.5004	-0.5043	0.0431
CFB	cg01883966	-0.6382	-0.4255	0.0062
LYN	cg03973663	-0.6093	-0.2705	0.0122

Table 4 methylation validation

Three genes LYN, TNFSF10 and CFB were found as epigenetic-related genes with same correlation pattern in both twins and validation panels.

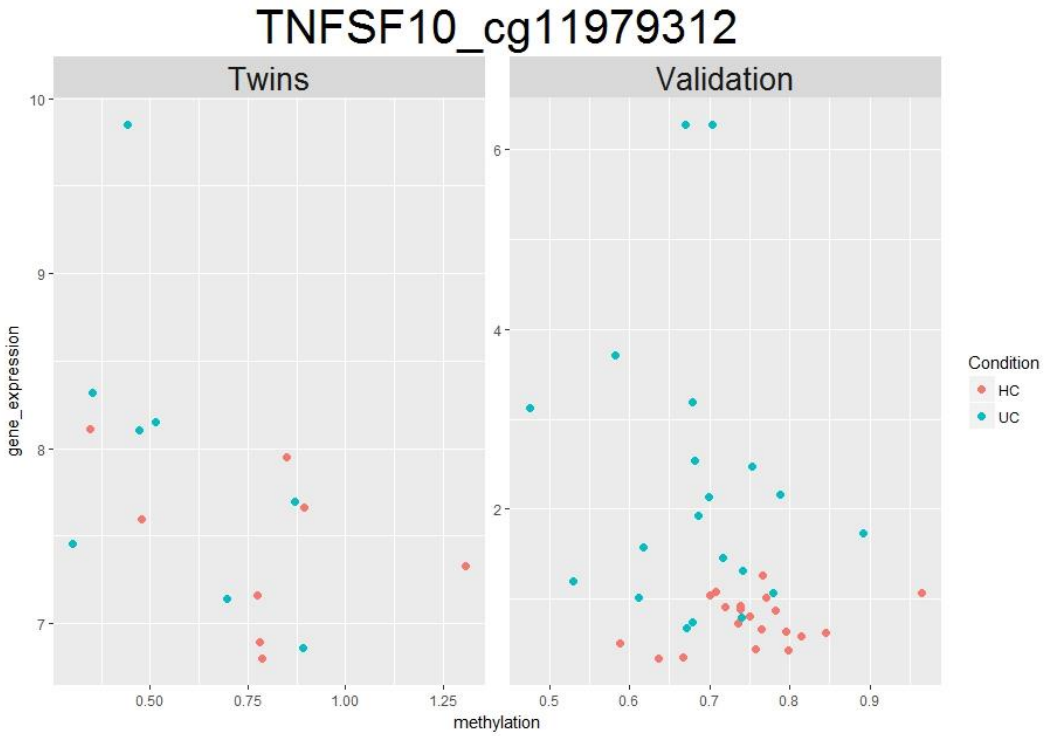


Figure 26 The scatter plot of TNFSF10 and the correspond methylation sites

The scatter plot of TNFSF10 and methylation site (cg11979312) in twins and validation panel. Both figure showed the negative correlation between gene expression and DNA methylation. The x-axis is the relative gene expression value (the value from microarray in twins cohort and from Taqman in validation cohort) and y-axis is the methylation level (M-value in twins cohort and beta-value in validation cohort)

bacteria genus	Gene name	Correlation	
		twins cohort	validation cohort
Clostridium_XlVa	S100A8	0.683792236	0.44145658
Clostridium_XlVa	TNFSF10	0.489319215	0.4051866
Clostridium_XlVa	CFB	0.59596571	0.4537696
Clostridium_XlVa	OAS1	0.213292991	0.4482985
Clostridium_XlVa	ISG20	0.664972266	0.5019149
Clostridium_XlVa	CCL11	0.777892085	0.4277273
Bacteroides	TNFSF10	-0.342649337	-0.3722326
Bacteroides	CFB	-0.756414575	-0.4741088
Bacteroides	LYN	-0.769344739	-0.5540338
Bacteroides	ISG20	-0.588322447	-0.4641651
Bacteroides	CCL11	-0.226277864	-0.3525328

Table 5 OTUs validation

Two genera, Bacteroides and Clostridium_XlVa were validated with the same trend correlating with the disease and methylation –related genes.

3.2 Mouse study

3.2.1 Study design

IECs study is substantial in investigating intestinal physiology and pathology[17]. IECs in the small intestine are focused in this study because of importance for several human disorders like inflammatory bowel disease (IBD) or cancer[17]. **Our aim was to explore the bacterial effect on dynamic host epigenetic markers along with gene expression changes during the postnatal duration.** Therefore, IECs were collected from conventionally-raised and germ free C57BL6 female mice at three different postnatal stages: week 1, week 4 and week 12/16 (W1, W4, W12/16), which represented the infant, juvenile and adult gut (Figure 27A). RNA and DNA were isolated for the purpose of gene expression and DNA methylation analysis. RNA and DNA were subjected to fuencing and reduced representation bisulfite sequencing to assess global mRNA expression and methylation, respectively (Figure 27B). After data pre-processing, there were 21619 genes and around 12 million methylation sites employed in the downstream analysis.

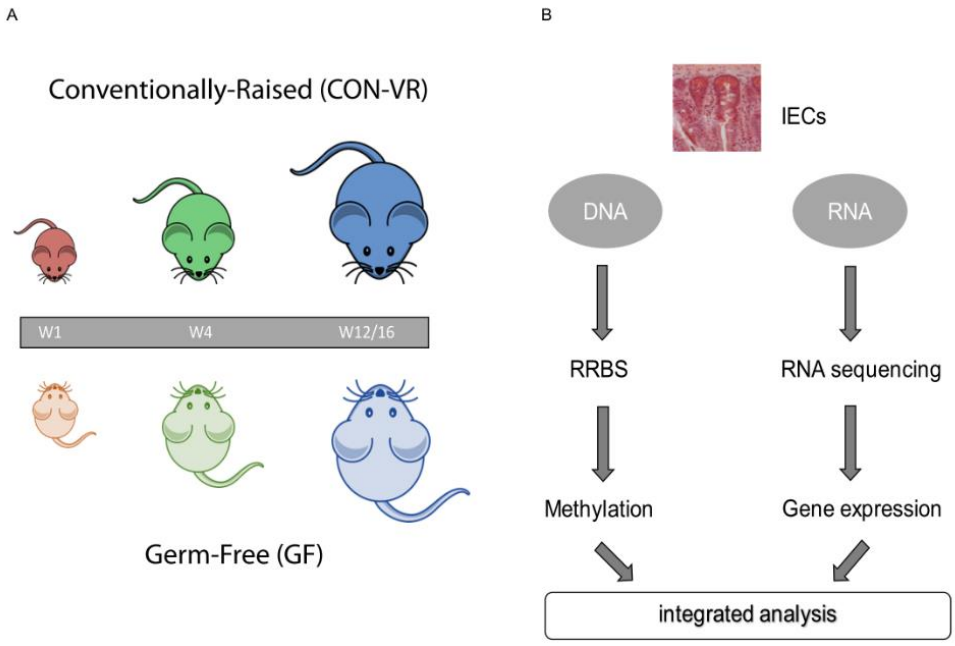


Figure 27 Study design of mouse study

(A) Conventional-raised (CONV-R) and germ free (GF) mice were chosen respectively and sacrificed in three developmental stages: week1, week4 and between week12 to 16. (B) The intestinal epithelial cells (IECs) from small intestine were collected. Though the next generation sequencing, isolated DNA and RNA are used for measuring gene expression and DNA methylation.

3.2.2 Microbiota Profile

Data Quality

In order to investigate the microbial composition in small intestine of CONV-R mice, the 16S rRNA genes were amplified and sequenced by Illumina MiSeq (2 X 300 bp). The resulting 119603 sequences were cleaned through data preprocessing by MOTHUR[128] v 1.37.6 (details were described in 1.7.1 One sample was discarded from subsequent analysis because of low number of sequence reads. The remaining samples were further normalized by subsampling the reads number down to 1691 reads (the smallest reads number for the remained samples). The sequences were then clustered into Operational

Taxonomic Units (OTUs) level, based on 97% similarity. In this study, 1392 OTUs were classified down to genus level by Silva database (https://www.arb-silva.de).

Alpha diversity and composition

Alpha diversity measurements were applied to determine the aspects of within-sample bacterial diversity that may be influenced by different conditions. The alpha diversity generally describes the species composition within the sample. Three measurements of alpha diversity are commonly used: rarefaction curves, species richness estimators, and community evenness diversity indices.

Rarefaction curves were used to estimate and compare bacterial richness among different conditions with a distance cutoff level of 97% similarity (Figure 28). All amplified rarefaction curves increased rapidly from 0 to 2000 sequences, indicated that sequence-derived diversity and richness in this study were sufficient to characterize the species in each samples. The rarefaction curves showed that the bacterial richness were higher in the later time points compared to W1.

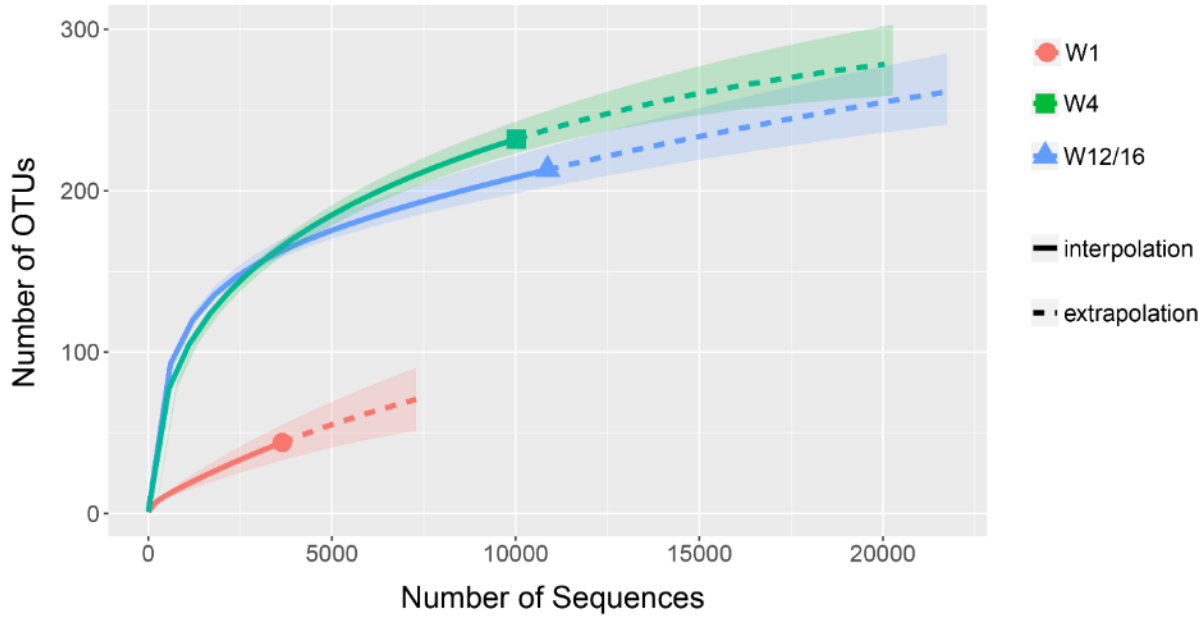


Figure 28 Rarefaction curves for 16S sequencing

The rarefaction curve, plotting the number of observed OTUs as a function of the number of clean sequences

The OTU were classified to have the deeper knowledge of intestinal bacterial phyla. Figure 29 showed the relative abundance of eight bacterial phyla (*Actinobacteria, Bacteroidetes, Firmicutes, Fusobacteria, Proteobacteria, Tenericutes, TM7* and *unidentified Bacteria*). These phyla were found within the small intestine samples across all time points. *Firmicutes* almost dominated all small intestine ecosystem in newborn mice and the *Bacteroidetes* percentage increased with ontogenies period. Apart of these two main bacterial phyla, the percentage of *Proteobacteria* also raised up along with the time points. The individual variation for biological replicate within subgroups is obvious in W4 and W12/16. One sample in W4 and the other two subjects in W12/16 showed the clearly drop in *Bacteroidetes* compare to the other members within the group.

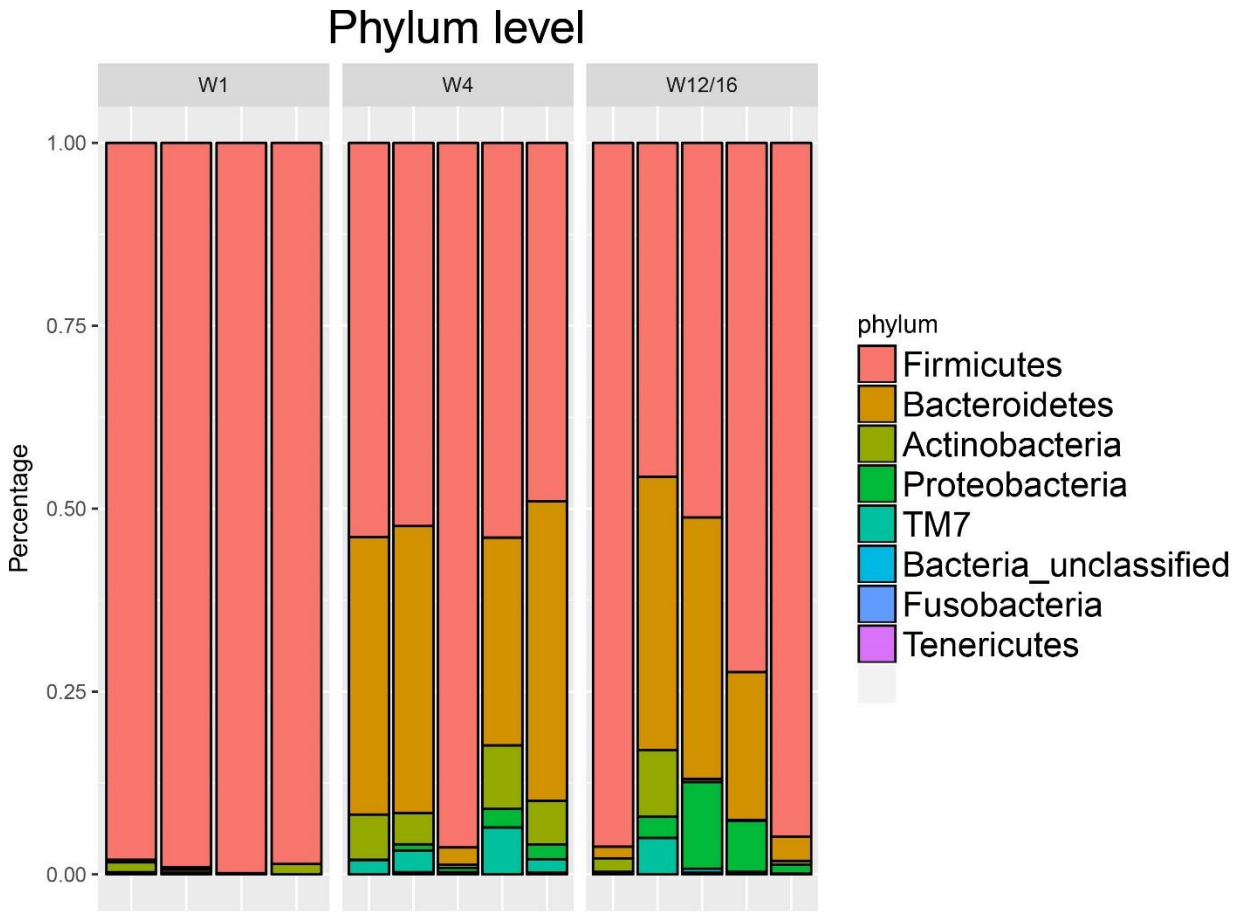


Figure 29 Phylum distribution of the microbiota of the 14 mouse samples at all-time point

The distribution of the microbiota composition in different development stage. *Firmicutes* and *Bacteroidetes* were two dominated phyla in gut microbiota composition

Species richness index estimates the number of species and evenness diversity index measures the relative abundance of different species in the given sample. Chao1 index and Shannon entropy index were employed as richness and evenness diversity in this study, respectively. Both diversity indices in W1 were lower than W4 or W12 (Figure 30). The diversity indices increased significantly from W1 to W4 or W12/16 (Table 6). However, there is no statistical difference between W4 and W12/16 in both indices. These results implied that the species number and composition structure of intestinal microbiota changed subsequently from early postnatal stage (W1) and then remained relatively stable from W4 to W12/16.

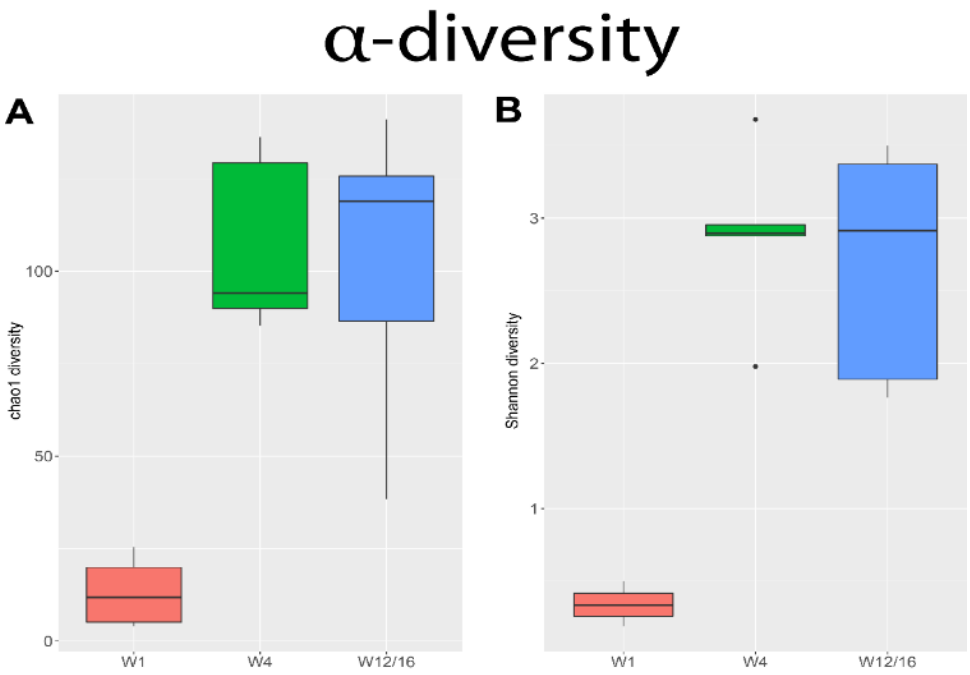


Figure 30 Chao1 richness and Shannon evenness diversity

(A)Chao1 richness estimator estimates the total number of OTUs, and (B)Shannon index estimates the entropy of data composition. Both figures showed that either the richness or evenness change in the earlier period, and getting stable from W4

	P-value of Chao1 index		P-value of Shannon index	
Timepoints	W4	W12/16	W4	W12/16
W1	0.00323	0.01413	0.01442	0.02412
W4		0.826		0.6888
W12/16				

Table 6 P-value for each comparison between time points in alpha diversity index

We use two tailed t-test for testing the mean difference between groups. The significant values ($p<0.05$) are highlight as red.

Beta diversity

The other goal of microbiological study was to compare the similarity/dissimilarity between bacterial composition in different conditions, such as antibiotic treatment or at different

time points during development. Beta diversity is the distance-base measurement for both presence/absence and abundance data in ecological studies. In this study, Jaacard and Bray-Curits were employed for calculating the similarity distance between samples. The structure of the relative distances between samples was projected in 3-dimensional plot by using a Principal Coordinate Analysis (PCoA) with first three principal coordinates. This method can focus on the most important axes and investigate the relationship between figure pattern and experimental factors.

Figure 31 showed the PCoA plot of Bray-Curtis distance of the mouse samples. Each dot represents one individual. First three principal coordinate (Bray1, Bray2, Bray3) explained 80% variation of total data structure. The first principal coordinate (Bray1) separated W1 with other two time points and explain 38% variation in data structure. This separation matched the observations which were found earlier in alpha diversity indices. The separation between W1 and W4 can be explained by the plane of Bray1 (38%) and Bray 2 (26%). Moreover, the samples from W12/16 contributed around 16% variation (Bray3) in microbial comoposition. The pattern of Jaccard based PCoA was similar to Bray-Curtis based PCoA (Figure 32). The first three principal coordinates of Jaccard distance (Jaccard1, Jaccard2, Jaccard3) explained only 52% variation. Samples from W12/16 spilled out in this 3-dimensional space, on the other hand, W1 and W4 were clustered on the plane of Jaccard1 (24%) and Jaccard3 (12%). The results of either presence/absence or abundance distance based PCoA suggested that microbiota composition pattern in matured mammalian (W12/16) gut and developing stage (W1 and W4) were very distinct.

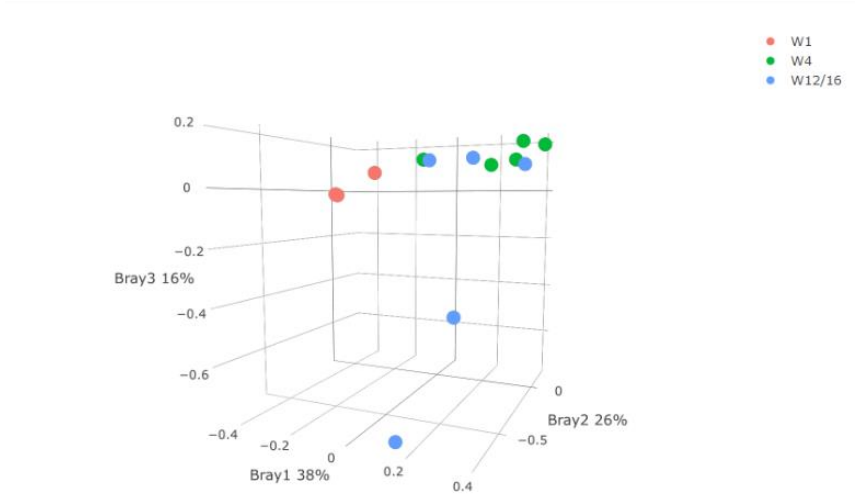


Figure 31 PCoA plot with Bray-Curtis distance

The first and the second principle component explained around totally 64% variation, and the third component which is mainly from W12/16, explained 16% of variation in the data.

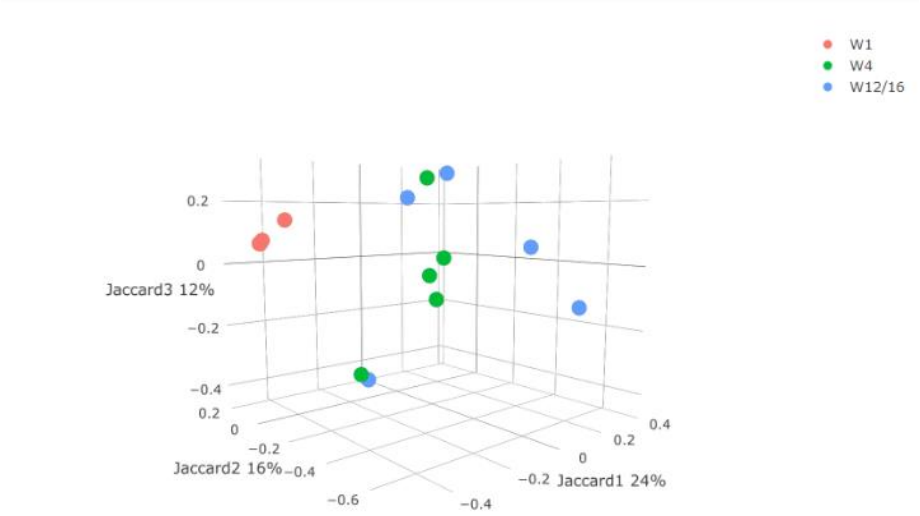


Figure 32 PCoA plot with Jaccard distance

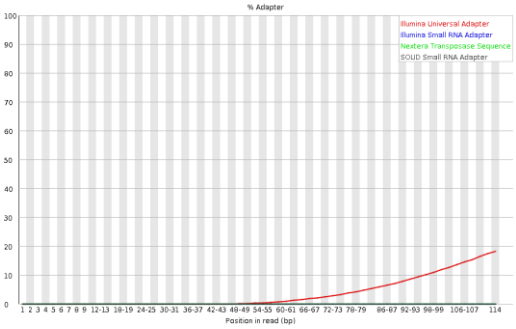
The pattern is similar with Figure 31, and the first and third component explained around only 36% variation, and the second component which is mainly from W12/16 explained 16% of variation in the data.

3.2.3 Gene expression

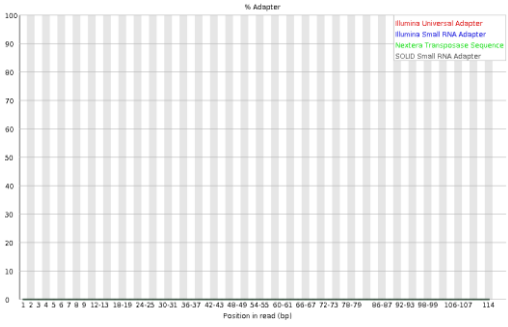
Data Quality

The pre-processing pipeline in 2.2.1 (Figure 10) was applied in this study. As a first step in processing, the illumina adapters were trimmed from the raw reads. The trimming of adapters leaded to the improvement of read quality (Figure 30A,C). Reads were mapped to the mouse genome (MGI assembly version 10) using Tophat2[103] program. On an average, more than 16 million reads (570 mio. total reads) were mapped per transcriptome library to mouse genome (Figure 34A). The average mapping rate was 83.3% (Figure 34B). Subsequently, HTSeq was used to generate the read counts of 21,619 Mouse genes using the following parameters (mode: intersection-strict., minaqual: 20). The read counts of these 21,619 genes were employed in further downstream analyses.

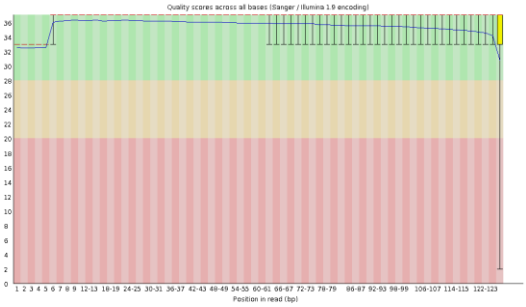
(A) Adapter content before trimming



(B) Adapter content after trimming



(C) Per base quality before trimming



(D) Per base quality after trimming

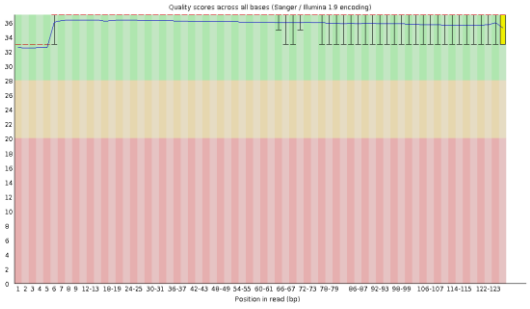
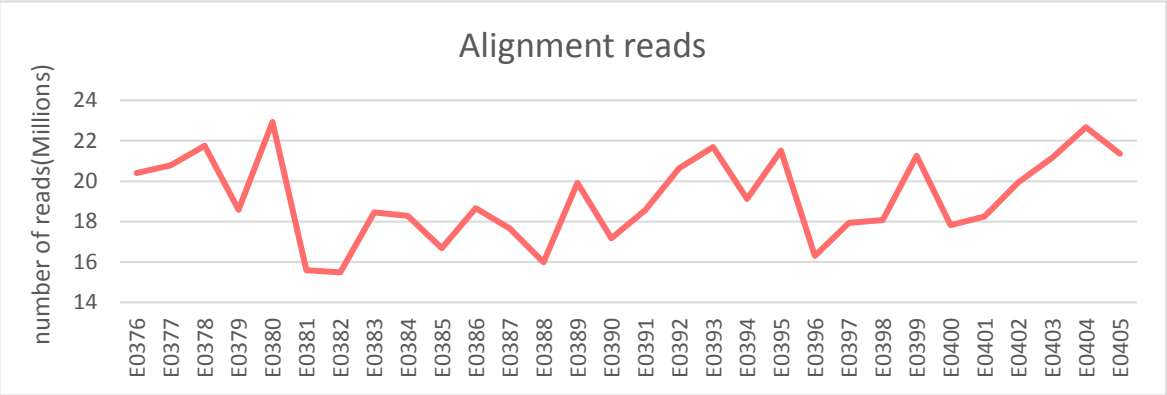


Figure 33 Quality check before and after trimming

Sample E0376 forward sequence was taken for the trimming example (A,B) Adapter content plot before and after trimming from FastQC. Before trimming, the percentage of illumina universal adapt is abnormally higher in the tail. After trimming, the adapter content is disappear. (C,D) Per base quality plot before and after trimming from FastQC.

(A) Number of aligned reads



(B) Mapping rate

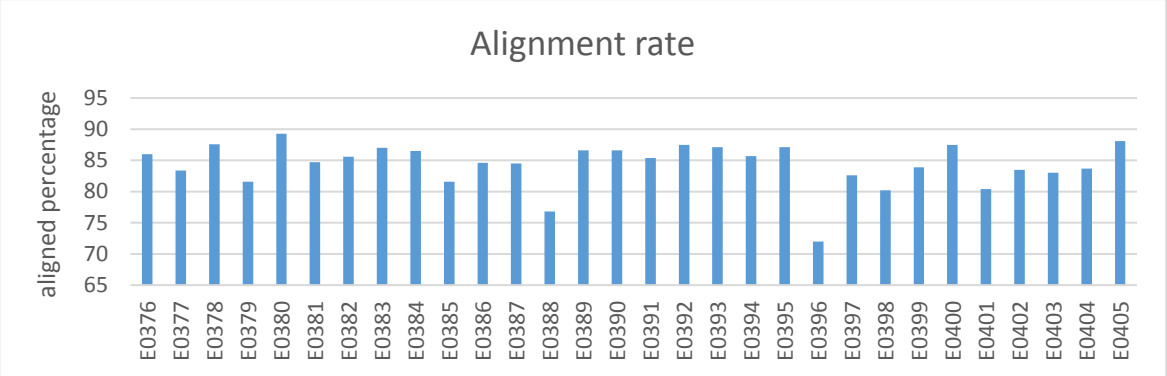


Figure 34 Alignment reads number and mapping rate

Average number of aligned reads for RNA-Seq is 19155018 (15490672-22929203, median = 18627290); Average alignment rate for RNA-Seq was 83.3% (73.3%-89.9%, median = 85.7%)

PCA and differentially expressed analysis

Principal Component Analysis was performed to visualize the sample clustering based on the expression data of the 21,619 genes. The gene expression levels were normalized by library size. Samples clustered both according to the developmental stage and microbial status (Figure 35). The first principal component explained 63% variation and separated samples from W1 and the other two stages W4 and W12/16 indicating that gene expression changed dramatically during IECs maturation, especially after the very early postnatal period. Whereas, the second principal component explained 8% of variation and separated W4 and W12/16 but also CONV-R and GF within a single developmental stage.

Notably, the distance between CONV-R and GF samples increased along with the time from W1 to W12/16.

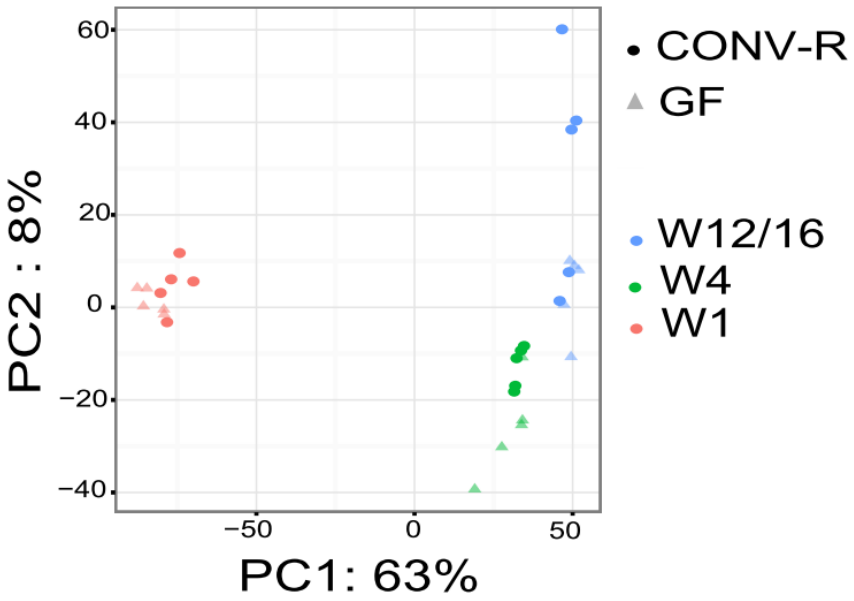


Figure 35 Principal component analysis displaying the overall gene expression profiles across all samples.

The first dimension explained 63% variation and separated W1 and the other two stages. The second dimension explained 8% variation and separated both W4 versus W12/16 and samples of a stage for their microbiota status.

We then test for differentially expression between CONV-R and GF in three fixed time points. We detected 56 microbially regulated genes in W1 (differentially expressed in CONV-R vs. GF comparison with adjusted p-value < 0.05 and absolute log2-fold change > 1), 614 in W4 and 1084 in W12/16 (Table 7, Figure 36A). Moreover, the expression differences between CONV-R and GF (fold change) of the microbially regulated genes increased simultaneously with time (Figure 36B). Thus, mainly ontogeny (developmental stage) and to a lesser extend bacterial status determined the epithelial transcriptional profile during postnatal development. To gain insights into the biological functions of the microbially regulated genes during postnatal development, we employed Gene Ontology (GO) enrichment analysis on the differentially expressed genes in the three developmental stages. GO terms were mainly enriched in immune response related or

metabolic functions. Biological functions such as cellular response to interferon-beta, defense response to another organism, immune response were enriched in both W4 and W12/16. Positive regulation of NF-kappaB transcription factor activity and MyD88-dependent toll-like receptor signaling pathway were enriched in W12/W16.

	Number of DE	Up-regulated	Down-regulated
W1	56	47	9
W4	614	348	266
W12/16	1084	613	471

Table 7 number of differentially expressed gene in three fixed time points

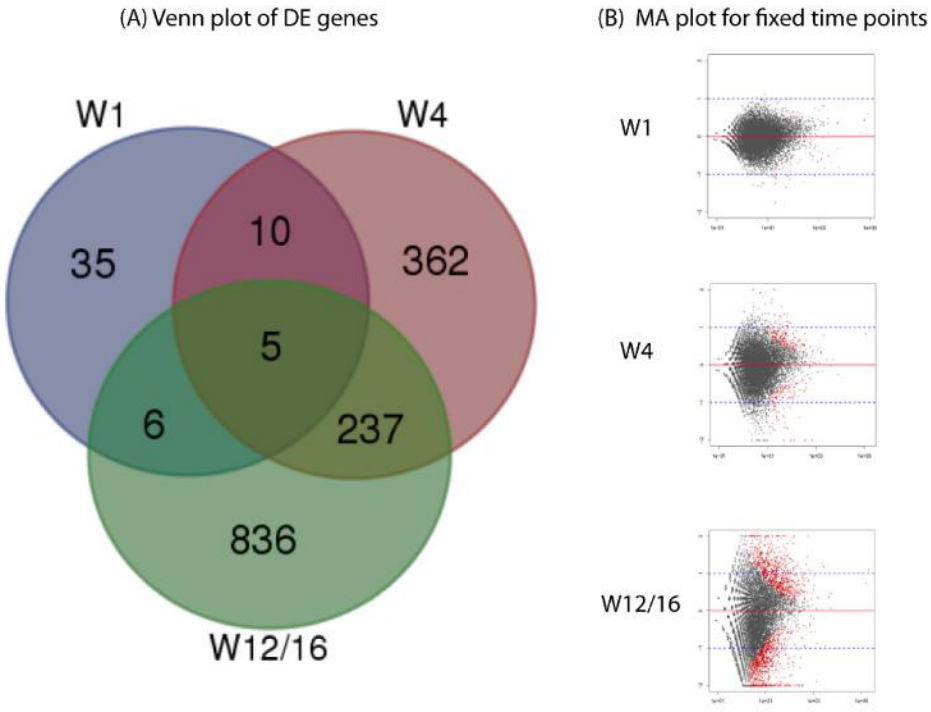


Figure 36 Differential expressed gene

(A) Numbers of differentially expressed genes intersection, adjusted p-value<0.05, fold change >2 (B) MA plot of CONV-R vs GF for three different time points. Every dots represent one gene, x-axis is the mean expression value and y-axis is the fold change between CONV-R vs GF. The red dots indicated the statistically significant genes.

To identify gene clusters modulated by the microbiota we selected the 200 most significant bacterially regulated genes individually from the three pairwise comparisons of the developmental stages with the GF mouse (W1: CONV-R vs GF; W4: CONV-R vs GF; W12/16: CONV-R vs GF), created the union of these genes (n = 547 genes), and performed hierarchical clustering which resulted in a heatmap of clusters of microbially regulated genes (Figure 37A). A similar analysis was performed based on the selection of developmentally regulated genes for the two bacterial conditions CONV-R and GF (CONV-R: W1 vs W4; CONV-R: W4 vs W12/16; GF: W1 vs W4; GF: W4 vs W12/16) to create a heatmap with clusters of developmentally regulated genes (n = 553 genes, Figure 37B). Both analyses reveal a microbial effect (e.g. clusters 2, 3, 4, 8, 11 in Figure 37A), and a developmental effect (e.g. clusters 8, 10 in Figure 37A). However, while the developmental effect is clearly visible in the heatmap of microbially regulated genes (Figure 37A), the microbial effect is not obvious in the heatmap of developmentally regulated genes (Figure 37B). This might be beause of the fact that the developmental effect on the epithelial transcriptome is greater than the microbial effect (Figure 37B). Cluster 8 in Figure 37A contains microbially responsive genes that mainly have functions in immune responses and are induced by the microbiota and the effect increases during development. Notable genes of this cluster include Duox2 (dual oxidase 2), Reg3g (regenerating islet-derived protein 3 gamma), Nos2 (inducible nitric oxide synthase), Saa1 (serum amyloid A-1) and Saa2, which have been reported previously as microbially induced in IECs[19]. The clusters 3 and 4 in Figure 37A contain genes such as Sdr16c6 (short chain dehydrogenase/reductase family 16C, member 6) or Fn3k (fructosamine 3 kinase), which are associated with metabolic functions, and expression of these genes is repressed by the microbiota.

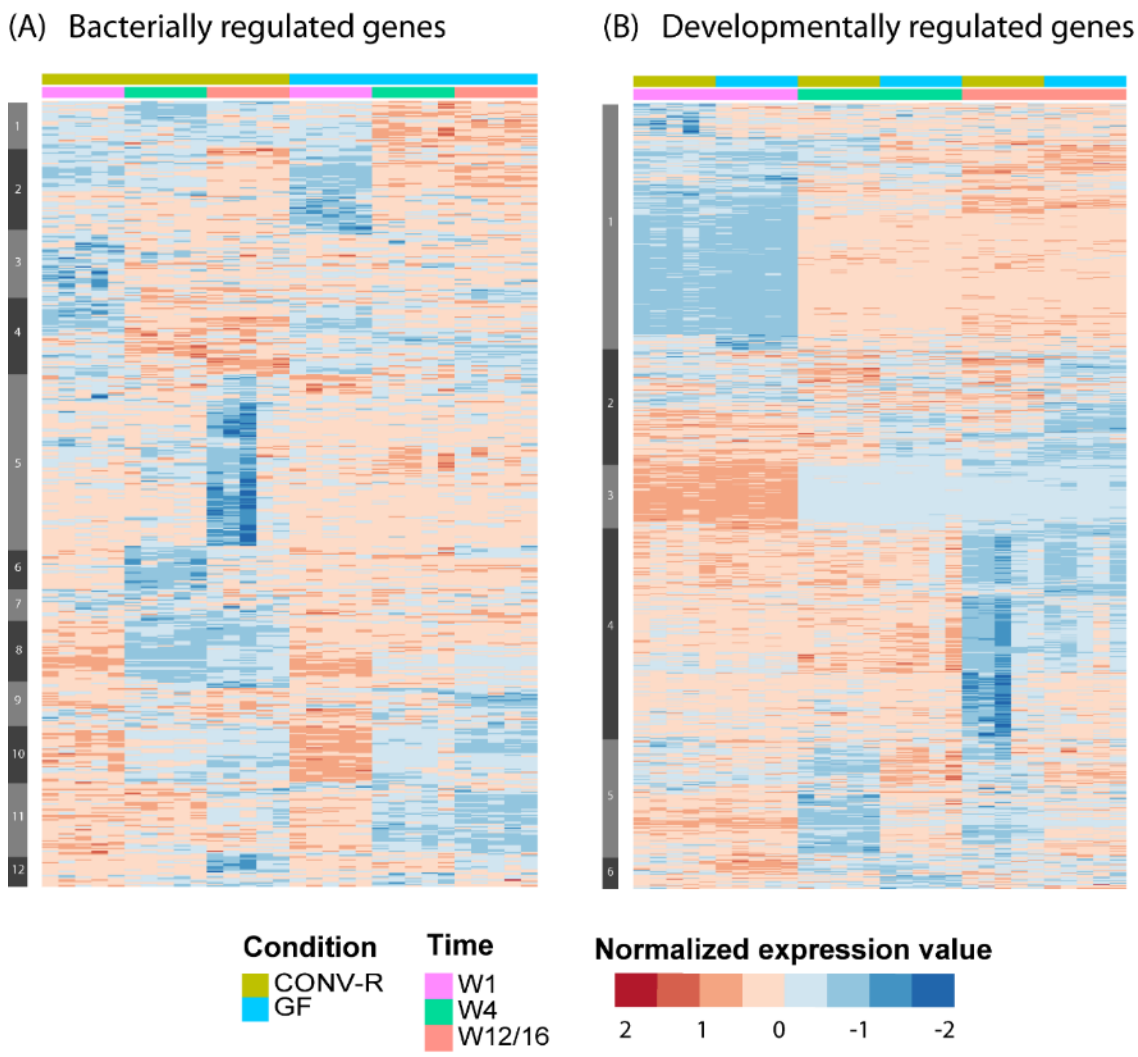


Figure 37 Heatmap of bacterially and developmentally regulated genes

Transcription factor binding site analysis

Transcription factor binding sites prediction and identification throughout genomes are integral for understanding the details of gene regulation and for inferring regulatory networks[129]. We employed transcription factor binding site enrichment analysis from innateDB database among the promoters of microbially regulated genes to investigate the regulatory networks that underlie the microbiota induced transcriptome alterations. Interestingly, the transcriptional regulators most enriched among promoters of microbially regulated genes were unique to W1 whereas W4 and W12/16 shared several transcription factors (Figure2B). For example, in W1 the transcription factor XBP1, which functions in

ER stress, cellular proliferation and differentiation and protects from intestinal inflammation[130,131], was enriched in the promoters of genes upregulated by the microbiota. Egr1 transcription factor was enriched in W4 upregulation genes. Egr1 induction in animal models implies complex responses such as inflammation, and fibrosis[132]. Raised Egr-1 was also found in mice model in the gut with chronic experimental colitis[133].In W4 and W12/16 the transcription factor HIF1, which functions in mediating hypoxia effects and regulates metabolism and immune responses[134,135,136], was enriched among downregulated genes.

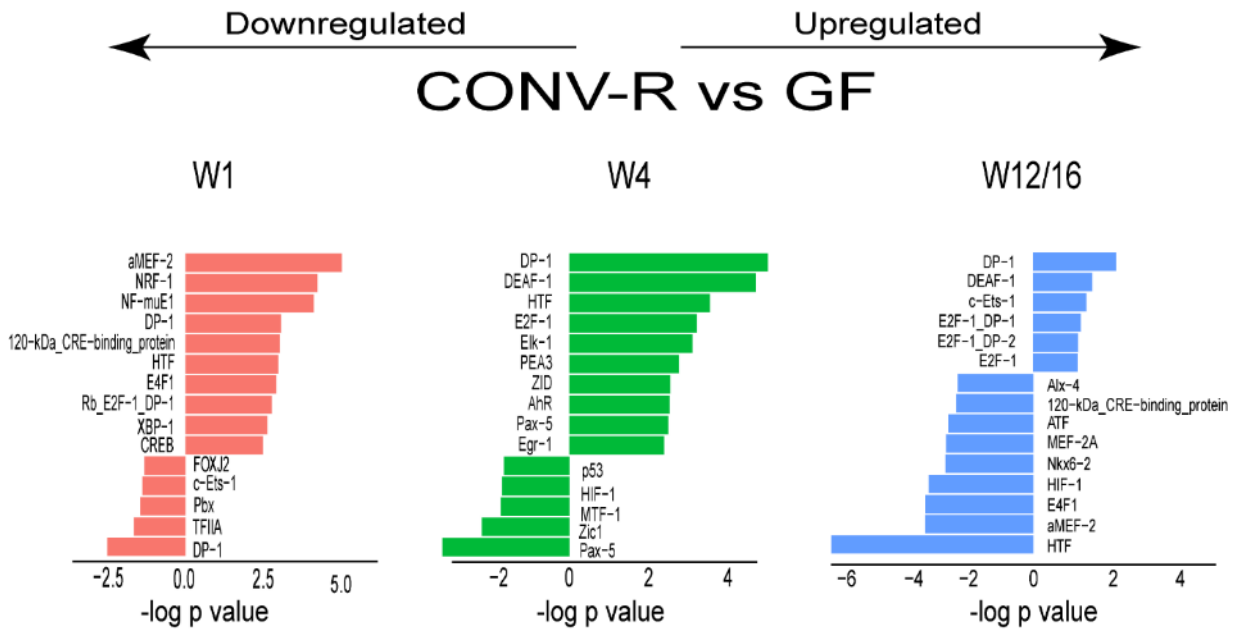


Figure 38 Transcription factor binding site analysis

Transcription factor binding sites enriched among microbially regulated genes (differentially expressed in CONV-R vs. GF) for each of the three developmental stages. The bar plot depicts the 15 most significantly enriched transcription factors of either up- or downregulated genes.

Co-expression network analysis

The genes that have similar expression patterns across the different condition are assumed to have the functional relationship[137]. These co-expression genes do not necessarily have causal relationship between each other, but more and more studies

showed the co-expression genes might related to some biological processes or pathways [137]. I further investigated the micrcobiota influence during postnatal development by co-expression network analysis[138]. 970 co-expressed genes were selected based on correlation cutoff of 0.8, normalized by their expression level and tested for up/down regulation compared to the average expression. Differential nodes were then highlighted and GO analysis was performed to identify the encoded biological processes (Figure 39). This co-expression network analysis was done by the cooperation partner Dr. Thomas Ulas. At the W1 stage, we did not detect a microbiota-dependent node but a cluster of enriched genes in both CONV-R and GF. Genes of this group A were involved in cell differentiation and basic epithelial maintenance. At the W4 stage, group B genes involved in innate immunity were upregulated in CONV-R and group C genes encoding metabolic functions were upregulated in GF. During W12/16 group D genes functioning in adaptive immunity displayed strong microbiota dependency.

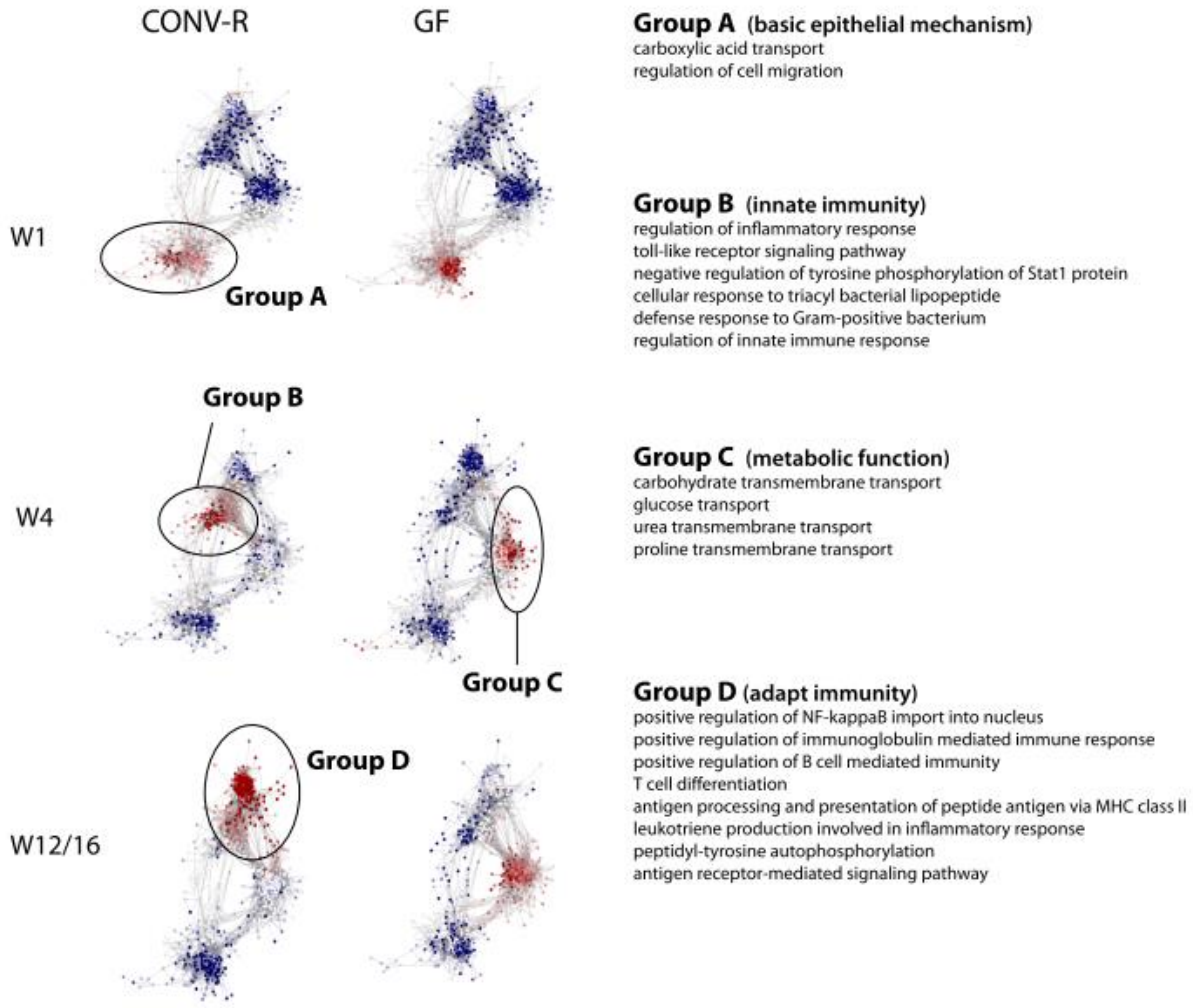


Figure 39 The microbiota modulates distinct functional expression nodes during postnatal development.

Co-expression network analysis (CENA) was performed based on 970 co-expressed genes (correlation factor greater than 0.8 across all conditions). Each dot represents a gene and the color indicates its expression compared to the average gene expression level (red = up, blue = down). This analysis was from a collaborative approach with Thomas Ulas, Bonn.

3.2.4 Alternative splicing

Alternative splicing is a key molecular mechanism that creates diverse RNA isoforms from a single gene, potentially increasing protein variety. More and more evidence suggests that alternative splicing has the relation with colon cancer progression[139,140]. However, the

connection between gut bacteria and alternative splicing events has not been well discovered. In this study, the microbial effects on alternative splicing events in postnatal period was investigated. Five categories alternative splicing were classified by rMats[108] software: exon skipping, mutually exclusive exons, alternative 5′ splice site, alternative 3′ splice site, and intron retention (Figure 11). The significant differences was not observed in the overall alternative splicing patterns between CONV-R and GF in the three developmental stages (Chi-square test, p-value = 0.99, Figure 40 & Table 8) , that means the global alternative splicing trend might not directly connect to gut microbiota. Then the comparison was performed for the different alternative splicing event between CONV-R and GF in three developmental statuses, few specific events were significantly different, for example, the number of microbiota-dependent intron retention events was 2.3-fold higher in W1 compared to W4 or W12/16 (Figure 41). Intron retention has been thought to be a result of mis-splicing, but more and more studies recently revealed the importance of developmental stages and occurs in notably cancer[141]. Here the microbiota effect for intron retention was discovered, especially in early gut development.

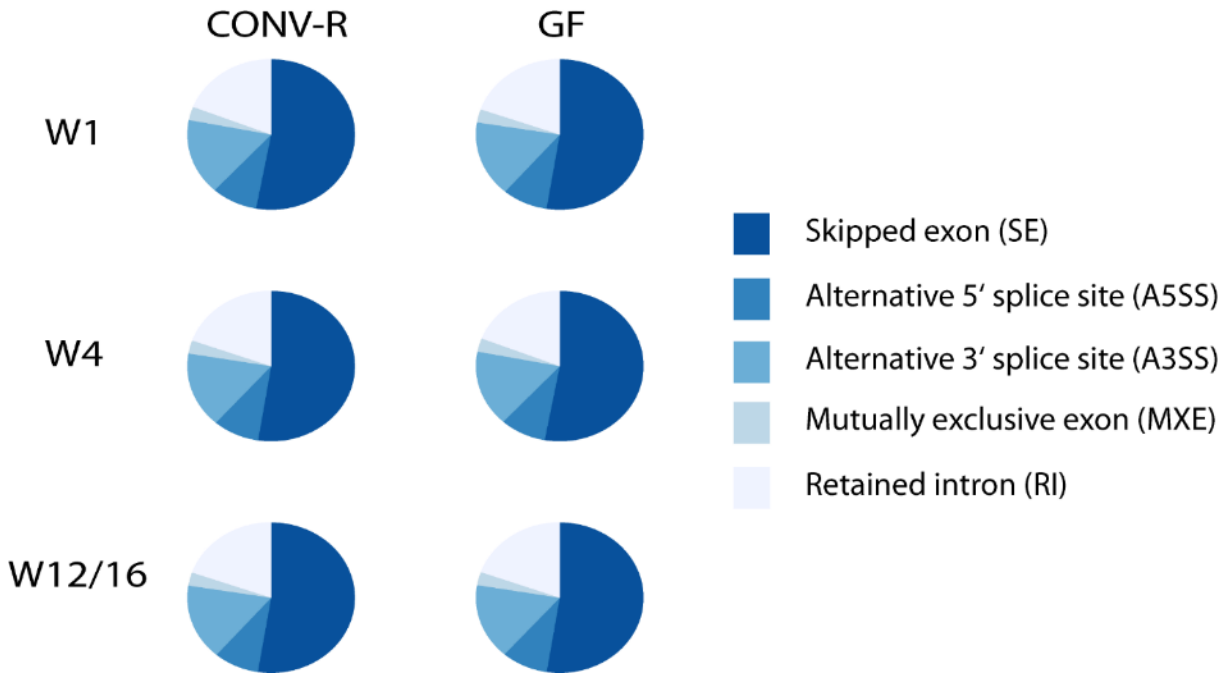


Figure 40 The composition of AS events in all conditions

	CONV-R			GF		
	W1	W4	W12	W1	W4	W12
SE	6690	6422	6423	6884	6614	6527
A5SS	1152	1100	1119	1158	1139	1118
A3SS	2037	1978	1972	2071	2019	1998
MXE	351	348	333	353	345	335
RI	2212	2198	2185	2236	2204	2197

Table 8 AS events in all condition
The distribution of 5 different categories alternative splicing events in all condition. There is no significant difference between CONV-R and GF (Chi-square test, p-value = 0.99)

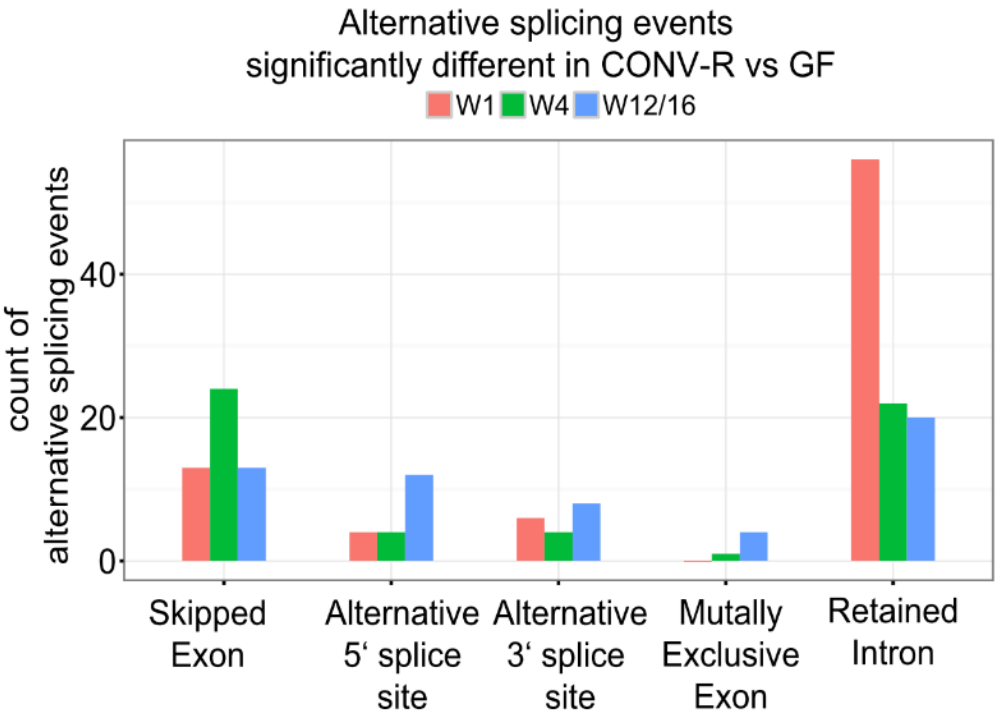


Figure 41 Differentially AS events in fixed time points

The significant event between CONV-R and GF in five alternative splicing categories. The number of significant retained intron events in W1 were significant higher than the other weeks.

3.2.5 DNA methylation

To investigate whether postnatal development or the microbiota affect the DNA methylation pattern of IECs, RRBS was applied to measure the methylation level of isolated IECs from CONV-R and GF mice at W1, W4 and W12/16.

Data Quality

TrueMethyl®Seq Kit from Cambridge Epigenetix was used for RRBS library preparation. The sequencing controls were spiked into the genomic DNA sample during NGS-library preparation (prior to adapter ligation). Each duplex contains C, 5mC, 5hmC and 5fC bases

at known positions, which can be interrogated after sequencing to give a quantitative assessment of the efficiency of conversion. The conversion efficiency was calculated after adapt trimming and alignment (Figure 12). Among all 30 samples, 7 of them had higher conversion rate (> 10%) in 5hmC which showed the failure of bisulfite conversion in 5hmC. Thus, failure samples resequencing is essential in this case. After resequencing, there was still one sample (in the group of W1/CON) can't pass the quality control in the second run. The rest 29 samples were employed in the downstream analysis. SNPs (C57BL/6 in dbSNP dataset) were filtered out and masked the CpG site which the coverage lower than 5. Totally 1,296,536 CpG sites were detected across all samples (Table 9).

Sample groups		Average mapping efficiency	Average CpG sites	Average CpG sites >5x coverage
	W1	69.46%	2385443.5	1136155.2
	W4	72.09%	3383208.4	1016776.4
	W12/16	71.88%	2862948.6	1022393.4
	W1	69.94%	1928314.6	1123435.6
	W4	72.50%	2827844.4	981752.4
	W12/16	71.92%	2490268.4	1061716.4

Table 9 Mapping efficiencies and CpG coverage of libraries

NMDS and differentially methylation analysis

The overall methylome pattern (1,296,536 CpG sites) was examined by using multidimensional scaling analysis (NMDS) (Figure 42). As for the transcriptome analysis, samples separated according to the developmental stage (Figure 42) and the methylation level increased with time (Figure 43), indicating a strong effect of postnatal development on DNA methylation. However, in contrast to the transcriptome, the microbiota did not affect the global methylation pattern. By comparing the methylome of CONV-R and GF in each time point, we identified 1499, 137 and 220 differently methylation positions (DMPs, false discovery rate <0.05) in W1, W4 and W12/16 respectively (Figure 44A). Surprisingly, the number of DMPs from the early stage was about 10x higher that of the later stages indicating that the microbiota acted stronger on DNA methylation during W1. The detected

DMPs were equally hypo- and hypermethylated (Figure 44B). The DMPs were classified according to their genomic location: exon, intron, intergenic or promoter. Notably, in W1 DMPs located in gene promoter regions (within 1500 base pairs upstream and downstream of transcription starting sites) were enriched (175 DMPs or 11.67%) compared to W4 (1 DMP or 0.73%) and W12/16 (15 DMPs or 6.81%) (Figure 45).

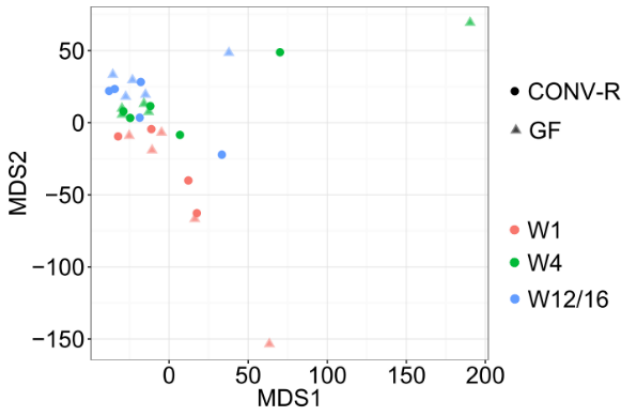


Figure 42 Multidimensional scaling analysis plot

MDS displaying the overall methylation profiles. The developmental factor pattern can be recognized in MDS plot, in contrast, bacterial effect is not visible.

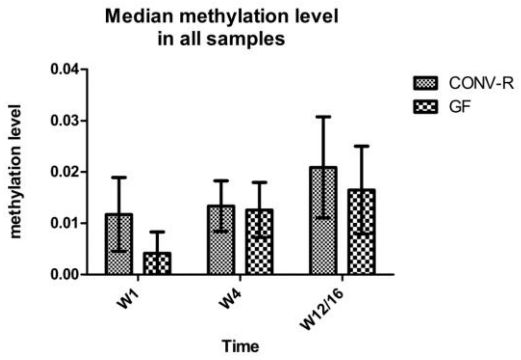


Figure 43 Overall methylation level across all samples

The median methylation level was calculated in detected sites. The methylation levels generally increase along with the time, regardless of GF or CONV-R.

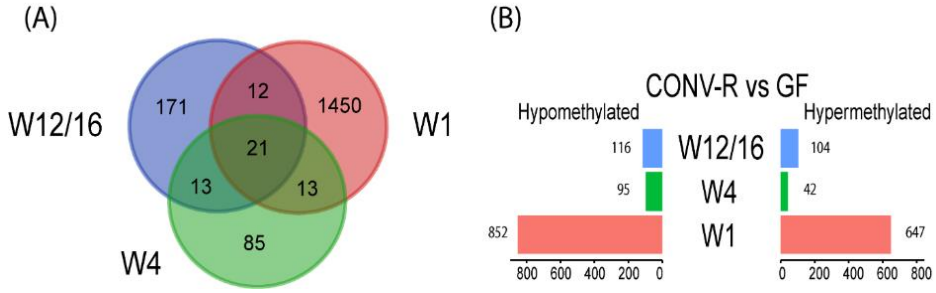


Figure 44 Differentially methylated positions

(A) Venn plot of DMPs in three time points. (B) Hypomethylated and hypermethylated number of DMPs in three time points. The number of DMPs in W1 is 10 times higher than W4 an W12/16.

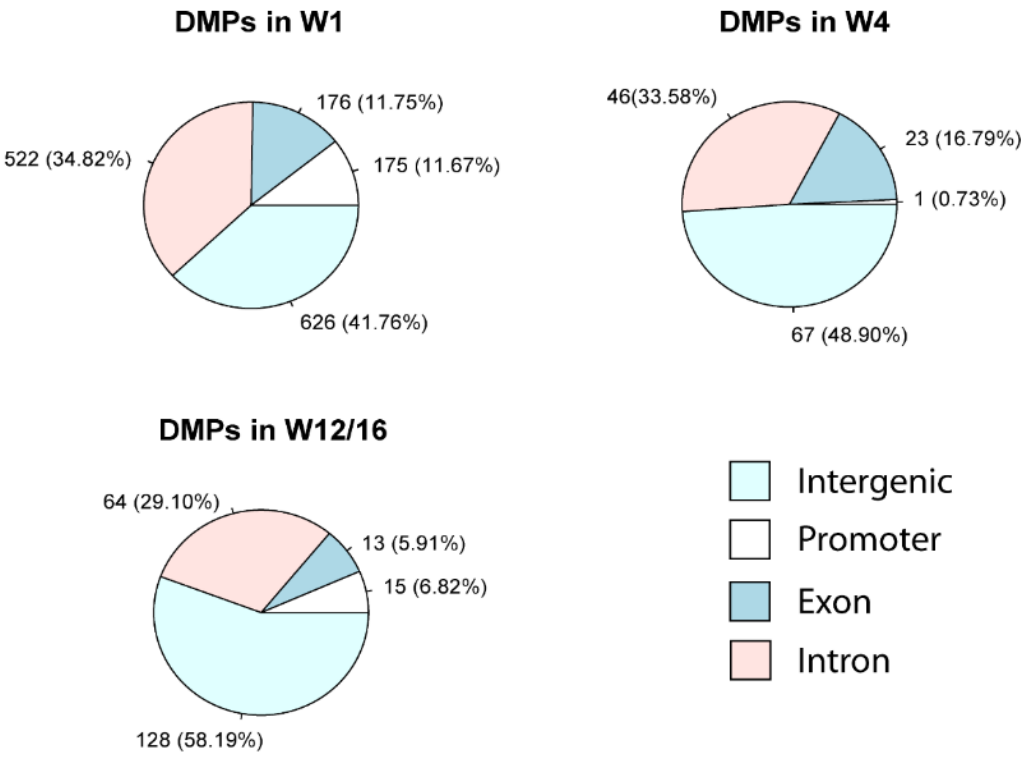


Figure 45 DMPs location

DMPs in promoter were enriched in W1, compared to W4 and W12/16.

Gene expression of methylation related gene

Given the observation of methylation, the gene expression value of some known genes which can alter the DNA methylation were further checked (DNA methyltransferase 1: *DNMT1*; DNA methyltransferase 3A: *DNMT3a*; DNA methyltransferase 3B: *DNMT3b*; Tet methylcytosine dioxygenase 1: *TET1*; Tet methylcytosine dioxygenase 2: *TET2*; Tet methylcytosine dioxygenase 3: *TET3*; Ubiquitin Like With PHD And Ring Finger Domains 1: *UHRF1*; Ubiquitin Like With PHD And Ring Finger Domains 2: UHRF2; Methyl-CpG Binding Domain Protein 2: *MBD2*; Methyl-CpG Binding Domain Protein 3: *MBD3*; Forkhead box O3: *FOXO3* in Figure 46). Expression of *DNMt3a* and *TET3* were highlight as significantly altered by the microbiota in W1 and W12/16. *DNMT3A* is important for *de novo* methylation[142], whereas TET3 is essential for demethylation[143].

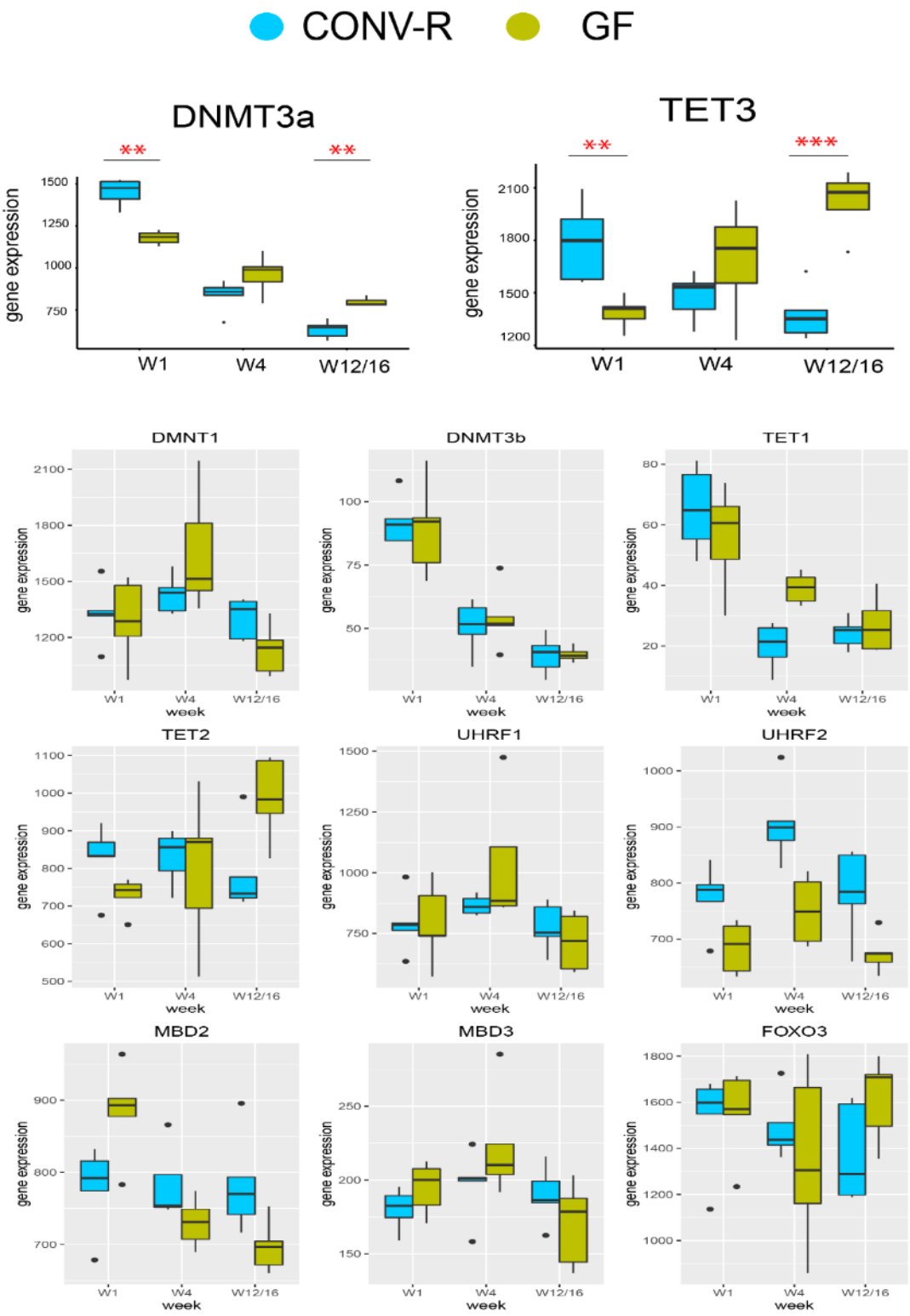


Figure 46 Gene expression value for methylation related genes

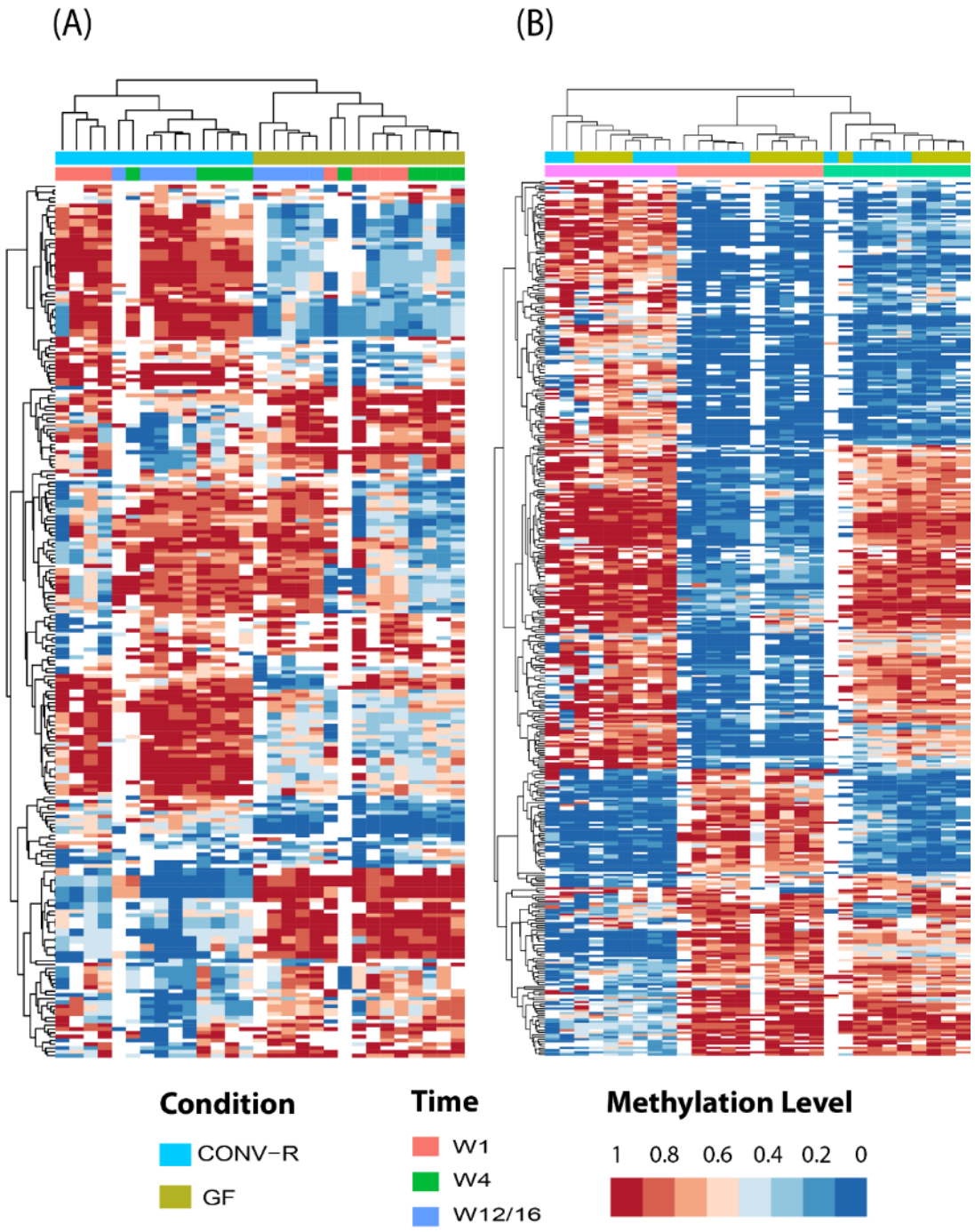


Figure 47 Methylation level of selected methylation sites

As for the transcriptome analysis, all DMPs were ranked based on their p-value and the most significant microbiota or developmentally regulated DMPs were chosen for each

comparison to generate heatmaps by hierarchical clustering (Figure 47). For the microbiota related DMPs, samples clustered according to microbial status and developmental stage (Figure 47A) except for few samples with too many missing values due to insufficient sequencing depth. For the developmentally related DMPs, samples clustered only by developmental stage but not according to microbial status (Figure 47B).

3.2.6 Whole genomic map

The hierarchical testing approach[125] was employed to identify interactions between the microbiota dependent alterations in the transcriptome and DNA methylation signatures (Figure 15). To that end, all differentially expressed genes (CONV-R vs. GF) for differential methylation positions (DMPs) within a 5kb window up- and downstream were screened. There were 17, 34 and 79 microbially regulated genes with both an altered expression and a differential methylation in W1, W4 and W12/16 respectively (Figure 48). *Fry* (FRY Microtubule Binding Protein) occurred both in W1 and W4, while *Cd59a* (CD59 Antigen), *Sorcs3* (Sortilin Related VPS10 Domain Containing Receptor 3) and *Pik3c3* (Phosphatidylinositol 3-Kinase Catalytic Subunit Type 3) were reported both in W4 and W12/16 (Figure 48). Tracking both the transcriptome and DNA methylation during postnatal development allowed us to identify specific changes in the DNA methylation signature that may underlie the microbiota dependent transcriptome alterations. For example, expression of *Camk2b* (calcium/calmodulin-dependent protein kinase II), which is involved in calcium-dependent signaling[144], was only altered by the microbiota at W12/16 but not at the younger stages W1 or W4 (Figure 49A). Interestingly, nearby CpG sites were not differential methylated at W1, whereas in week W4 we detected three DMPs and another four DMPs at W12/16 (Figure 49A). Therefore, either the complete demethylation of all DMPs or only the four downstream DMPs may be required to mediate the microbial induction of *Camk2b* expression at W12/16. Another example is *Neurl1b,* expression of *Neurl1b* (Neuralized E3 Ubiquitin Protein Ligase 1B) which highly expressed during embryonic development of the brain and several non-neural tissues[145], downregulated in CONV-R by the microbiota at W12/16, and the nearby DMPs were

hypomethylated only at same time periods. Thus, these four DMPs might also essential for microbial induction of *Neurl1b* (Figure 49B).

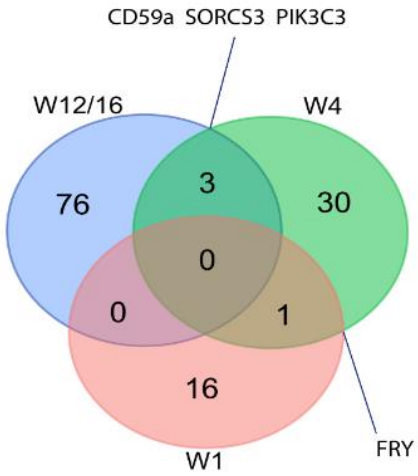


Figure 48 Genes with DMPs in 5kb window.

Potentially methylation moderiated genes in three different stage (hierarchical testing, adjust p-value < 0.05)

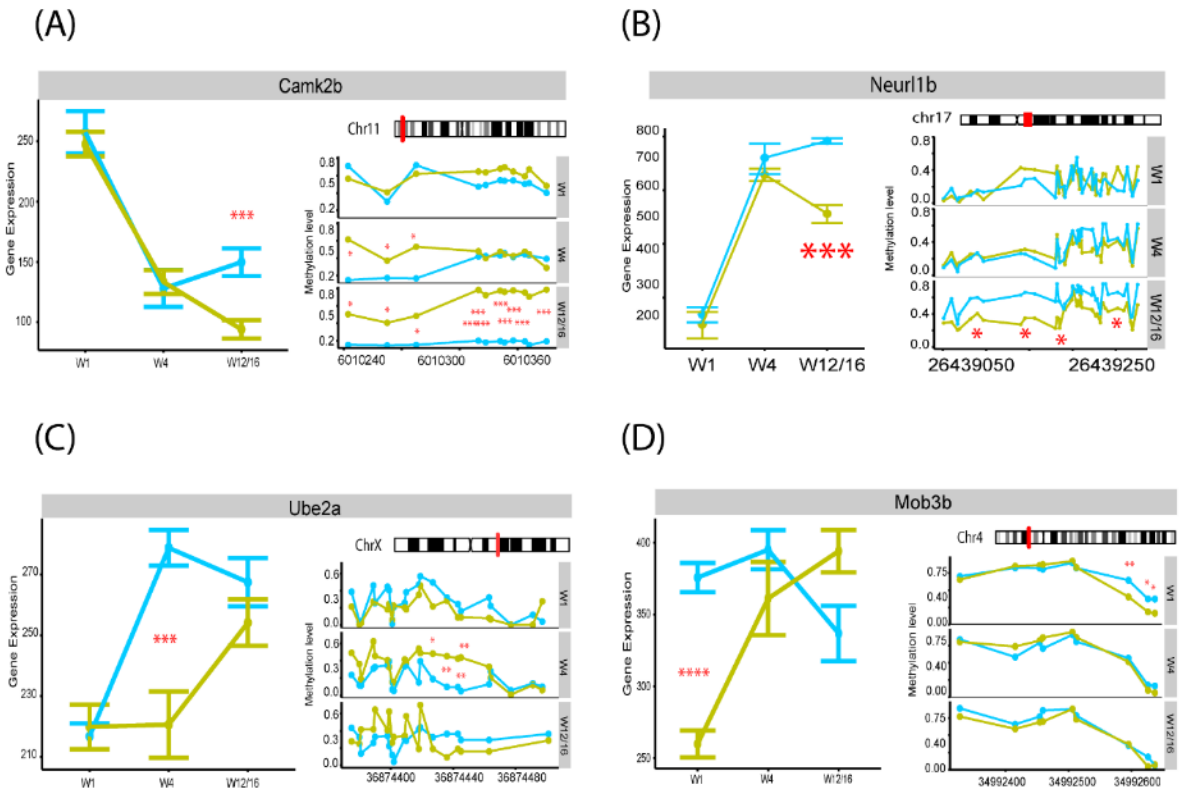


Figure 49 Gene expression and DNA methylation change

The gene expression and methylation change along with time points in four different genes.

The regional plots showed the DMPs and differently gene expression for *Pik3c3* in W4 (Figure 50). The DMPs in *Pik3c3* were hypomethylated (CONV-R < GF) in promoter region and upregulated (CONV-R > GF) in gene expression. Genome-wide mapping of the host-microbiota interactions for gene expression and DNA methylation during the three development stages revealed equal distribution among chromosomes (Figure 51). Genes belonged to W12/16 highlighted the GO biological process categories, such as regulation of multicellular organismal development, positive regulation of multicellular organismal process (Table 10).

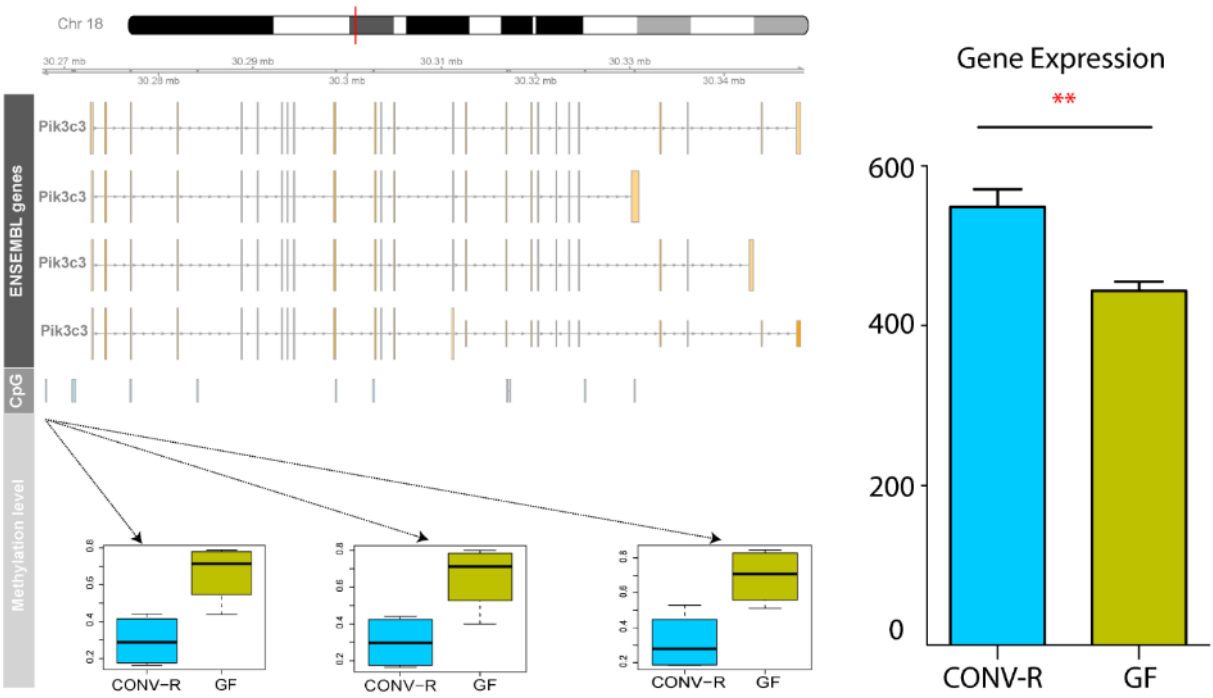


Figure 50 Regional Plot of Pik3c3 in W4

DMPs in promoter region showed the hypormethylation in CONV-R methylation level and up-regulated in gene expression. This might potentially point out the relationship between methylation and gene expression in Pik3c3

W1

GO term	Description	P-value	FDR q-value
GO:0043576	regulation of respiratory gaseous exchange	0.000212	1
GO:0090311	regulation of protein deacetylation	0.000631	1

W4

GO term	Description	P-value	FDR q-value
GO:0043271	negative regulation of ion transport	0.0000765	1
GO:0051051	negative regulation of transport	0.000251	1
GO:0002294	CD4-positive, alpha-beta T cell differentiation involved in immune response	0.000595	1
GO:0042093	T-helper cell differentiation	0.000595	1
GO:0002293	alpha-beta T cell differentiation involved in immune response	0.000737	1
GO:0002287	alpha-beta T cell activation involved in immune response	0.000814	1
GO:0021533	cell differentiation in hindbrain	0.000894	1

W12

GO term	Description	P-value	FDR q-value
GO:0052697	xenobiotic glucuronidation	3.90E-06	5.57E-02
GO:0045074	regulation of interleukin-10 biosynthetic process	1.02E-04	7.31E-01
GO:0052696	flavonoid glucuronidation	1.05E-04	4.98E-01
GO:0052695	cellular glucuronidation	1.05E-04	3.74E-01
GO:0009813	flavonoid biosynthetic process	1.05E-04	2.99E-01
GO:0043412	macromolecule modification	1.15E-04	2.74E-01
GO:0006063	uronic acid metabolic process	1.27E-04	2.58E-01
GO:0019585	glucuronate metabolic process	1.27E-04	2.26E-01
GO:0009812	flavonoid metabolic process	1.51E-04	2.40E-01
GO:0032879	regulation of localization	2.36E-04	3.37E-01
GO:0006464	cellular protein modification process	3.53E-04	4.58E-01
GO:0036211	protein modification process	3.53E-04	4.20E-01
GO:0042036	negative regulation of cytokine biosynthetic process	4.14E-04	4.54E-01
GO:0016569	covalent chromatin modification	4.57E-04	4.66E-01
GO:0050864	regulation of B cell activation	4.58E-04	4.36E-01
GO:0001782	B cell homeostasis	5.22E-04	4.66E-01
GO:0006325	chromatin organization	5.38E-04	4.52E-01
GO:0051239	regulation of multicellular organismal process	6.67E-04	5.29E-01
GO:0050869	negative regulation of B cell activation	7.15E-04	5.37E-01

Table 10 GO analysis for DE-DM gene

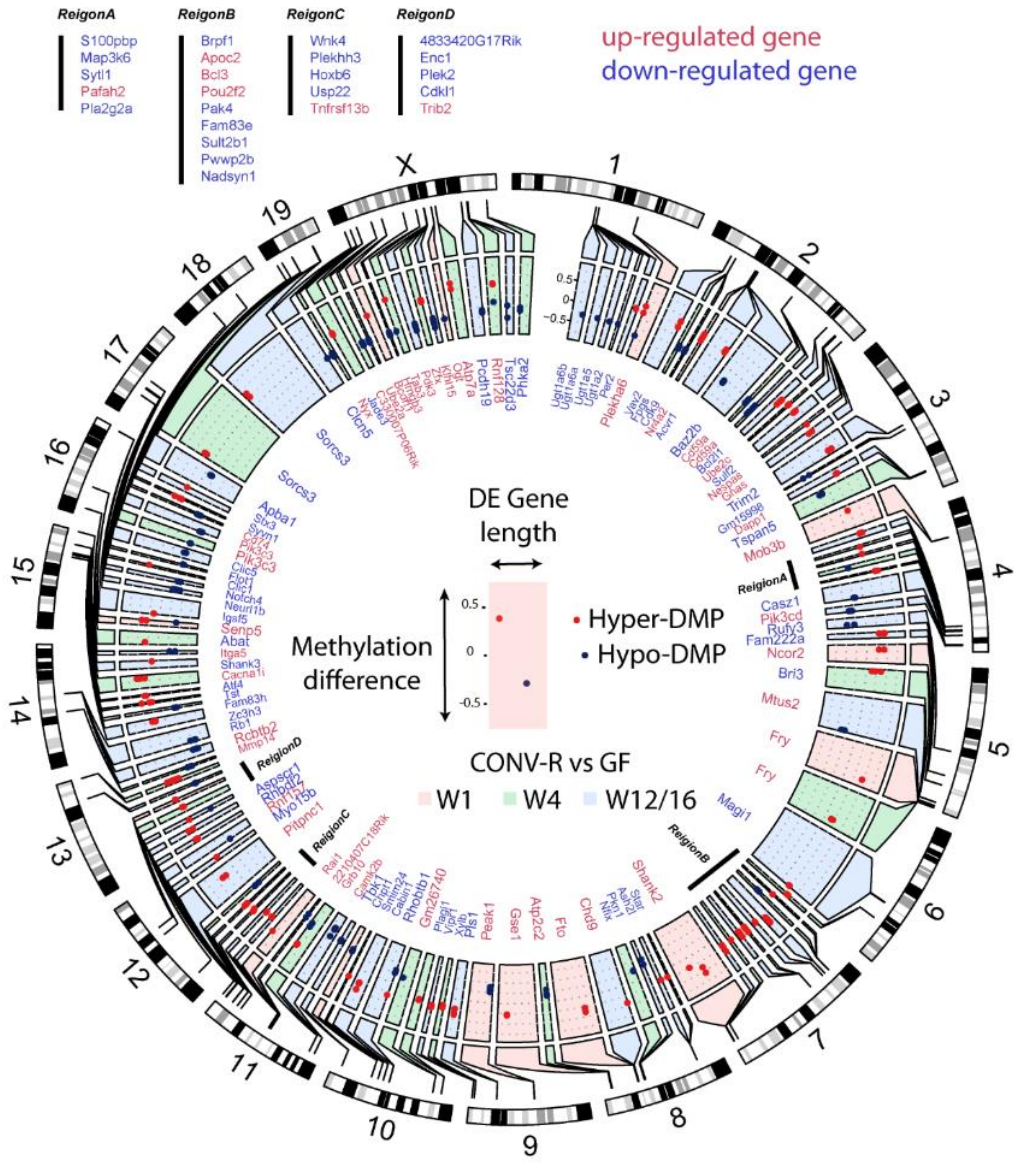


Figure 51 Genomic map of all methylation-transcription interactions

Genomic map of all methylation-transcription interactions dependent on the microbiota and postnatal development. The boxes in the outer circle depict the mouse chromosomes. The boxes in the inner circle represent genes that were both differentially expressed and methylated. The gene name is colored according to the expression difference in CONV-R vs. GF comparison (red = up-regulated, blue = down-regulated). Box coloring corresponds to the developmental stage, in which a significant difference was detected (red = W1, green = W4, blue = W12/16). Length of the boxes indicate the methylation difference in CONV-R vs. GF comparison. Red and blue dots within the gene boxes represent hyper- and hypomethylated CpG sites respectively.

4. Discussion

4.1 Cross-talk of transcriptome, epigenome and microbiota in intestinal inflammation

This study mainly aims for understanding the interaction between three omics layer in UC: transcriptome, epigenome and gut microbiota. Monozygotic (MZ) discordant twins were recruited in this study. The identical genes in twin subjects contribute equally genetic effect in UC. Thus, the setting of this experiment can farther investigate the other disease factors apart of genetic. Besides the shared genes, the twin subjects might also be raised in the same home, and experienced equally similar environments in their earlier life. One might argue that shared environment assumption might not be hold exactly, however, some researches suggest that parents, teachers, peers and others may treat identical twins more similarly than fraternal twins[146]. Based on this design, this study can focus on the risk contribution of non-shared environment, which are the epigenetic marker and epigenetic-microbiota interaction in this case.

4.1.1 Microbiota status in intestinal inflammation

In this study, kinship and disease effect were observed as the determining factors for the gut microbiota composition. First, Shannon entropy difference between the twin pairs increases along with the age (Figure 17). Twin pairs were raised in the same environment during their early life, thus the microbiota compositions pattern were very similar in the early age. This dissimilarity of the microbiota composition increases because the change of life style and surrounding areas. This reflects the importance of environment factor on microbiota. In the best of my knowledge, this finding have not been reported in other discordant MZ studies. One might further investigate the non-shared environment with more information in MZ study, such as household or smoking. Furthermore, the coming study can also look deeper in to bacterial taxonomy, the variety of specific bacteria phylum or genus might link to metabolic function for intestine inflammation.

Surprisingly, the kinship effect was still visible in microbiota composition, even in the elder twin pairs (Figure 19). Inter-individual distances between unrelated subjects were higher

than the distances between pairs. To note, one sibling was disease and the other was healthy, thus we can claim that the kinship effect exist in disease scenario. My finding in this study was also reported similarly in previous studies. One Chinese study showed that the microbiota composition in infant MZ is similar than dizygotic (DZ) twin or non-twins.[147] They also mentioned the age represent the strong factor to shape the microbiota composition, even before one year of age. Furthermore, the connection of microbiota composition and age has also been reported in adult twins. Dicksved, J. *et al.* found the microbiota composition in CD MZ discordant twin pairs is less similar than the healthy twins and CD concordant MZ twins[148]. However, the dizygotic healthy twin pairs were very young (7–8 years old), and were still living in the same household. This could also contribute to their high similarities in profiles apart of genetic effect. In conclusion, there is a clear association between microbiota and kinship.

Furthermore, disease status was also associated tightly with dysbiosis. In Figure 18, the diversity differences between UC and healthy were visible and lower in UC compared to healthy partners. The study from Lepage et al[149] mentioned that the bacterial abundance of *Bacteroidetes* and *Firmicutes* in unaffected siblings from UC discordant pairs is even closer to healthy individual than the affected siblings. Furthermore, this pattern was also observed in CD twins. CD patients with ileal involvement clustered separately from all others[148]. All these consistent studies confirmed the association with dysbiosis and intestine inflammation.

4.1.2 Epigenome-transcriptome interaction in ulcerative colitis

This study followed the study from Haesler et al[49], but only focused on the genes with specific immune or defense related function. The aim of this study was to detect the epigenetic linked immune-related genes as well as microbiota change in UC. Obviously, this approach cannot demonstrate the causality of epigenome-transcriptome interaction. Even though, this targeted disease-associated transcripts might support the hypothesis that pathophysiological events are a reflection of—and potentially controlled by—epigenetic modifications with consequences on transcriptional changes. 15 genes were

differentially expressed between UC and healthy control and highly correlated with the adjacent CpG sites. These genes were considered as epigenetic-related genes for UC. Furthermore, the gene expression of eight genes (*ISG20, LYN, AGT, CFB, S100A8, OAS1, TNFSF10* and *CCL11*) were then confirmed differentially expressed in unrelated UC and healthy cohort in the validation cohort. These eight genes directly link to regulation of inflammatory response (GO analysis, p-value=9.85E-5). This is consistent with previous findings on functional genomics of UC[44]. The consistency potentially attributes to the lower technical and/or biological variance in inflammation-associated mRNA patterns.

By all of these findings, certain genes have been directly/indirectly associated with chronic intestinal inflammation or gut microbiota; *ISG20* (Interferon Stimulated Exonuclease Gene 20) and *OAS1* (2'-5'-Oligoadenylate Synthetase 1) are related to Immune response IFN alpha/beta signaling pathway. *OAS1* and the other IFN pathway related genes were found increased from ileum in indoor-housed piglets compared to outdoor-housed piglets, indicating that the IFNα/β pathway is directly affected by the housing environment[150]. They further suggested that microbial composition influences Type 1 IFN signaling during early colonization and development. *TNFSF10* (Tumor Necrosis Factor Superfamily Member 10) has been identified as IBD associated gene, it disrupts the intestinal epithelium integrity by induction of epithelial cells apoptosis and possible contribution to development of fistulas and strictures in CD patients[151]. *CCL11* (C-C Motif Chemokine Ligand 11) which is a eosinophil-specific chemokine gene has been associated with IBD pathogenesis[152]. Waddell *et al* suggested that $Ly6C^{high}CCR2^{+}$ inflammatory monocyte/macrophage-derived CCL11 mediated DSS-induced colonic eosinophilia[153].

UC-relevant epigenetic modifications as well as their interaction with environmental factors was first reported in 1996 by Gloria *et al*[154]. Environmental factor regulated epigenetic markers have been reported as a contributor to disease susceptibility, manifestation, and progression[155]. There are around 100 genes whose methylation have been related UC[156] in previous studies. A recent methylome study of UC patients reported three genes (*FAM217B*, *KIAA1614* and *RIBC2*) were found to be significantly enhancing the promoter methylation levels if compared to normal controls[157]. In this study, 15 genes were found with strong association between methylation and transcription. Three (*AGT,*

TNFSF10, CFB) genes then further validated as epigenetic related in independent cohort. The methylation level in the promoter region of TNFSF10 (cg number: cg11979312) showed differentially methylated pattern between UC patients and healthy control, and furthermore associated with gene expression of TNFSF10. *TNFSF10* is also reported as epigenetic marker in the previous intestine inflammation study. The methylation level in the gene body of TNFSF10 (cg number: cg01059398) is able to discriminate between disease and control in UC[60] accurately.

Conclusively, the variation of the outcome in different studies are likely from the sample collected location, different cell/tissue and technical process. Although there are amount of evidence supporting the role of DNA methylation in regulating gene transcription, however, the functional relation within these gene has not been fully revealed. Determining the causative relationship between an epigenetic marker and gene expression is one of the major challenges in the IBD study. Furthermore, this finding might be limited by the microarray design; HM27 array which obviously cannot cover all the regions around the targeted genes. One might use the up-to-date EPIC or NGS for further research.

4.1.3 Transcriptome-microbiome interaction in ulcerative colitis

The commensal microbiota is well known for shaping the immune system and is involved in many host physiological functions including the digestion of nutrients. Furthermore, the transcriptional changes associated with IBD has been shown in different studies. However, only few studies have addressed the connection between the human mucosal transcriptome and the gut microbiota[158,159]. For the best of my knowledge, this study is the first to correlate the gut microbiota and epigenetic linked transcripts in UC patients. There were 17 bacterial OTUs identified with the strong correlation between the epigenetic markers and transcripts in twins cohort. Furthermore, two (*Clostridium_XlVa* and *Bacteroides*) were validated in the independent cohort. This finding suggests the presence

of these bacteria might be due to a defect in the barrier function of the epithelium in UC and potentially acted as the environment factor for modulating epigenetic marker.

Clostridium cluster XIVa, butyrate-producing species, specifically colonize mucins in gut model[160]. *Clostridium* cluster XIVa were enriched in the mucosal environment, and the butyrate-producing bacteria from these clusters had higher abundances in the luminal content. Butyrate is a short chain fatty acid (SCFA) derived from the microbial fermentation of dietary fibers in the colon[161]. One study suggests that probiotics induced epigenetic mechanisms through butyrate[162]. In another study, butyrate is also able to modulate intestinal microflora through regulation of lumen pH and to exert many beneficial extraintestinal effects through epigenetic mechanisms[163]. In a review paper from Canani et al[164], they made the connection between diet and epigenetic modulation through butyrate and listed the effect for some complex disease (Figure 52). Deeper study of butyrate might help to develop improved strategies for regenerative medicine by promoting epigenetic remodeling and the expression of pluripotency-associated genes. In this context, discovery of mucosal butyrate producers may lead to a novel therapy for IBD, which are characterized by an impaired butyrate transport to the colonocytes. Gever et al investigated the gut microbiota composition from 447 children and adolescents (< 17 years) with newly diagnosed CD. They investigated samples from multiple gastrointestinal locations collected both prior and after antibiotic treatment. The increased levels of *Bacteroides* and *Clostridiales* were found in patients who are non-CD afterwards compared to those who maintain in CD[165]. Consistently, the increase of *Bacteroides* abundance in healthy controls also found in the discordant MZ UC twins cohort as well as in the validation panel.

In conclusion, this human twin pairs study provides us the unique opportunity to discriminate between the contribution of genetic and environmental factors to phenotypic variance. One might keep following the finding for more biological functional analysis, and discover the direct/indirect functional host-microbiome interaction in intestine inflammation.

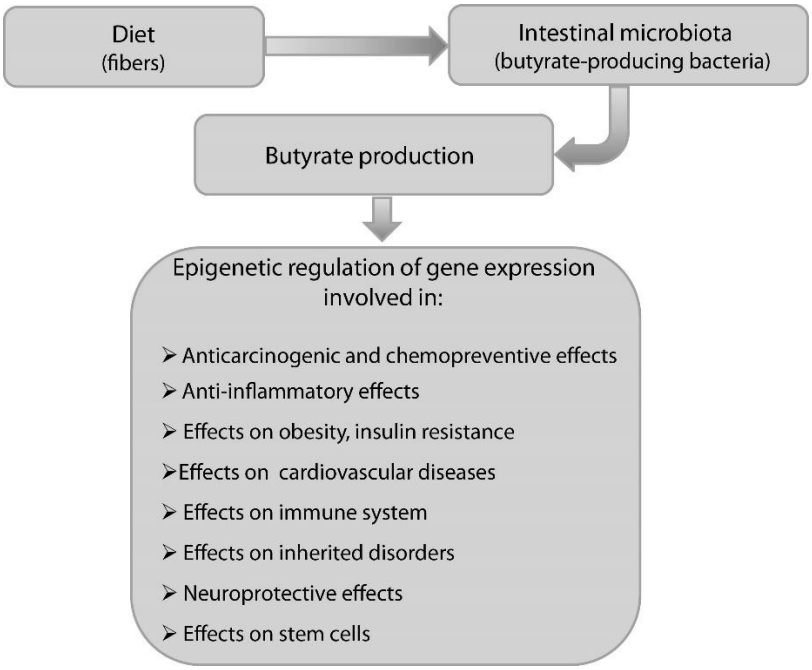


Figure 52 Diet influences intestinal microbiota

Diet influences intestinal microbiota composition. The balance of intestinal microbiota has an important role in the fermentation of dietary. Butyrate exerts multiple beneficial effects at intestinal and extraintestinal level, linked to the epigenetic regulation of gene expression. Figure is modified from Canani et al[164]

4.2 Cross-talk of transcriptome, epigenome and microbiome in intestinal development

This study systematically investigated the regulatory effects of the microbiota on the transcriptome and the genome-wide DNA methylation status of IECs from the small intestine of infant, juvenile and adult mice, which were raised either in the presence or in absence of a microbiota. This analysis revealed that both the IEC ontogeny and the microbiota affect the epithelial transcriptome signature along with the DNA methylation status and that the microbial effect increases during postnatal development. Furthermore, the microbial impact on the interplay of DNA methylation and the epithelial transcriptome were stage-specific as we detected almost no overlap between the genes that were regulated by the microbiota and also displayed an altered DNA methylation status for the

three developmental stages. Our data provide groundwork to dissect the endogenous developmental and microbial effects on the host's transcriptional and epigenetic program on a mechanistic level.

4.2.1 Dynamic transcriptome and epigenome pattern during development

This study validated that many developmentally regulated genes such as *Pigr* (polymeric immunoglobulin receptor), which was reported to have an increasing expression from infant to juvenile or *Tet1* (Ten-eleven translocation methylcytosine dioxygenase 1) having a decreasing expression from infant to juvenile[166], in addition also were differentially methylated and therefore appeared to be epigenetically regulated during postnatal development. Moreover, several genes were previously reported as microbially regulated in the adult[19,167], and were also regulated transcriptionally during postnatal development. For example, the glycolysis regulator *Pfkfb3* (6-phosphofructo-2-kinase) was not only induced by the microbiota in the adult as reported[19,167], but is already microbially regulated in the infant.

Surprisingly, about ten times more DMPs in W1 were detected compared to W4 or W12/16. Since methylation levels did not differ between the developmental stages, the increased number of DMPs in W1 did not seem to be simply due to higher overall methylation activity. Instead, the microbiota may differentially modulate *de novo* methylation and demethylation in the neonate mice. First, I detected generally higher levels of *Dnmt3a* (DNA-methyltransferase 3A) during W1 compared to W4 or W12/16 and increased expression in CONV-R compared to GF mice. As DNMT3 (DNA-methyltransferase 3) mediates *de novo* methylation and parental imprinting[168], this temporal and microbiota dependent expression pattern of *Dnmt3a* may therefore relate to the increased number of hypermethylated DMPs in the newborn mice. Conversely, *Tet3* expression was induced by the microbiota in W1 and since TET3 possesses hydroxymethylation activity[169,170] and therefore mediates demethylation[143], the time- and microbiota-dependent expression pattern of *Tet3* may thus contribute to the increasing number of hypomethylated DMPs

with increasing age. However, the maternal imprinting effect can not be rule out, which may be dependent on the presence of microbiota in the mother before birth.

4.2.2 Microbiota modified epigenome-transcriptome interaction in intestinal development

The value of this experimental approach is demonstrated by the finding that although several previous studies established that the microbiota modulates the expression of more than 2,000 genes in the intestinal epithelium[19,167,171] , only a subset of these microbiota-responsive genes appear to be regulated by the epigenetic process of DNA methylation. Using this approach, the microbiota was found to inversely affect DNA methylation and gene expression throughout postnatal development. Whereas the number of differentially expressed (CONV-R vs. GF) genes increased with postnatal development, the number of DMPs decreased from W1 to W12/16. The number of genes that are regulated by the microbiota both in their transcription and DNA methylation (differentially expressed and DMPs within 5kb window) increased with time. Together these observations indicate that the microbial effect on modifying the epithelial DNA methylation and transcriptional status increased during maturation and postnatal development of the intestine. However, the microbiota did not seem to engage DNA methylation to regulate transcriptional responses globally, but instead only seemed to target a specific subset of microbially responsive genes through their DNA methylation status. This unexpected finding is not caused by inherent differences in this and published datasets as, for example, our transcriptome sequencing data and the list of microbially regulated genes from the adult stage overlapped significantly with our previous data obtained from microarray analysis of laser-dissected ileal IECs[19]. This observations are further supported by a study by Camp et al., which reported that the microbiota did not globally alter the chromatin architecture to drive gene expression but only for specific genes[21]. Thus, host epigenetic mechanisms do not seem to be employed by the gut microbiota to drive transcriptional responses on a global scale.

Future studies are needed to functionally validate the involvement of methylation modifying enzymes during early postnatal development and in relation to the microbiota. For example, tracking the changes in intestinal microbiota composition along with epithelial DNA methylation and transcriptome signatures of DNMT or TET-deficient mice during postnatal development would be a promising approach. Together our data suggests that the microbiota seems to engage components of the DNA methylation machinery, which may at least partially translate into the observed epigenetic and transcriptional differences through postnatal development.

4.3 Methodological considerations and pitfalls

The genomic scale data from this study are based on microarray and NGS technologies. With the application of these technologies, a higher resolution for epigenome and transcriptome status of a cell can be obtained. With the evolution and advancement of technologies, enormous amounts of data has been generated. How to manage and use the proper statistical method to get meaningful results remains an issue and challenge. In this context, the pros and cons of microarray and NGS with regards to statistics and data processing will be discussed

4.3.1 Improvement of genome-wide screening technique

The concept and methodology of microarrays was first introduced by Tse Wen Chang in 1983[172], and commercialize in 1995[173]. The invention of microarray opened new field in genomic research. Microarrays have been applied in various fields of biology (e.g. gene expression, genotyping and DNA methylation) and have yielded numerous significant findings in clinical and basic research. However, arrays suffer from their fundamental “pre-select” design. With the radical decline in sequencing costs and the greater improvements in NGS systems, a large number of studies are now performed using NGS. Both microarray and NGS techniques were performed in this study to analyze transcriptome and methylome. The processed twins study data was generated in 2009 using microarray as it was the popular whole genome screening technique in that era. Even NGS already

existed in 2009, but the price was higher, and accuracy was not well established and accepted in scientific community. In the second mouse study, all data were generated in 2014, the protocol of library preparation and analysis pipeline for NGS are all well built up in-house. Therefore, NGS methodology was employed for mouse study.

Transcriptome

There are several obvious benefits to encourage the researcher switching from microarray to RNA-Seq in gene expression detection. The most important reason is that RNA-Seq does not require any prior knowledge the species and genes under investigation. Furthermore, RNA-Seq allows the detection theoretically in whole transcriptome and analyses of novel transcripts, splice junctions and noncoding RNAs as well. These characteristic of RNA-Seq can identify the novel associated genes which are not included in microarray. Microarray was employed in the twins study and RNA-Seq was used in mouse study. There is a significant difference of discovered gene numbers between two studies. 11544 expressed genes were found in twins study while almost double gene numbers (21619 genes) were detected in mouse study by using RNA-Seq. This unbiased feature allows researchers to have broad view for the gene expression modification.

Methylome

HumanMethylation27K BeadChip (HM27) for whole methylome was used in twins study. It claims to contain 25,578 probes predominantly targeting CpG sites within the proximal promoter region of 14,475 consensus coding sequence (CCDS) genes[174]. However, after quality control, only 23,477 methylation sites were remained available for analysis. Obviously, HM27 cannot cover whole methylome, thus the company developed new microarray Infinium HumanMethylation450 BeadChip (HM450) in 2011 and Infinium MethylationEPIC (EPIC) BeadChip in 2015, which can detect over 485,000 and over 850,000 methylation sites per sample at single-nucleotide resolution, respectively. HM450 and EPIC are still the popular tools for EWAS studies, and widely embraces by the epigenetics research community. However, there is no well-developed mouse genome-

wide DNA methylation array equivalent to the human methylation array. Dr. Richard Saffery's group from the University of Melbourne got a creative idea to use human DNA methylation arrays on mouse samples[175]. Their idea was proven and the measurements were replicated by a different assay. Even though, only 13,715 uniquely mapping probes in bisulfite space of mm9 were found on the HM450. After comparing the price and the amount of information, NGS for methylome study was chosen in the mouse study. In this study, 1,296,536 methylation positions were discovered by RRBS. The number of methylation site is 55 times than in twins study.

Two popular NGS methods, WGBS and RRBS, are already mentioned in 1.7.3 . In best of my knowledge, RRBS provides the best cost-benefit trade-off compared to WGBS, which yields 50-fold more reads per sample and is therefore comparatively more expensive and computationally intensive. For the following, one might use oxidative RRBS which can distingue hydroxyl-methylation and methylation within the sample. It reveals a complete picture of genomic state of methylation, and provides the board view of the function of hydroxyl-methylation.

4.3.2 Statistical and bioinformatics concern in genome data science

With the rapid improvement of whole genome screening technique, the enormous amount of genomic data have been generated in past decade. However, processing and analysis the various type of large dataset in different biological layer is a bottleneck. In this paragraph, the statistical and bioinformatics approach that were employed in this thesis will be discussed.

Microarray

Nowadays, microarray has been considered as an out-of-date method, especially for transcriptome studies. However, there are two main reasons for researcher to continue choosing microarray for transcriptome study: price and easy bioinformatics analysis

process. Take the transcriptome analysis as the example, the price of Illumina NextSeq 500 NGS is around 5 to 10 times than Affymetrix GeneChips in different experiment setting[176]. The researcher might choose microarray under the limit budget. For the analysis, Affymetrix provides its own software for data pre-processing. Herein researcher can easily click the button for choosing different filter criteria. Besides, one can also use well-established bioinformatics and statistics practices with free software packages that can almost do everything for people not familiar with computational work. The downside of array is that it can only be applied in organism with good reference sequence (e.g. human). It is useless for researcher who are interested in organism whose genome is either not sequenced or little information is available about their genetic components.

In contrast to the downside of microarray in transcriptome, microarray in DNA methylation is still active in recent research. HM450 covers methylation positions in a good degree and the newest EPIC chips can even cover 850K methylation sites. Even one can argue the selecting bias for coding and promoter regions, arrays still provide a good initial overview for methylation study. My research group benefited from the light computation work of HM27 in twins study which give us a good start point. For the following study, one could continue with deeper insight in to methylation pattern by using EPIC or NGS. Regarding to the analysis tool, Rnbeads[119] is an all-inclusive package for data quality check and statistical analysis. It is initially designed for microarray, and then further supports the entire high-throughput methylation platform (e.g. HM27, HM450 and EPIC). It generates a lot of presentable figures and the reliable statistical results for microarray.

RNA-Seq

After data preprocessing, RNA-Seq provides the number of reads that map to each transcript sequence. The higher coverage across genes is a statistical advantage. However, it biased towards transcript length i.e. longer transcripts or genes will have more reads mapped to it compared to shorter ones power[177]. According to this matter, many studies employ FPKM[178] (fragments per kilobase per million mapped reads) for paired-end reads RNA-Seq or RPKM[178] (reads per kilobase per million mapped reads) for single-

end reads RNA-Seq as the quantitative measurements. For calculating the RPKM value, the row reads were first divided by 1,000,000 and then normalized for sequencing depth in each sample to get reads per million (RPM). Then the RPM values were divided by the length of the gene in kilobases to get RPKM. Comparably, the calculation of FPKM is very similar to RPKM, the only difference is that FPKM takes into account that two reads can map to one fragment. FPKM and RPKM can be generated by Cufflink[179] and serve as an input to Cuffdiff2[180] for differential expression. However, this approach changes the data variance according to the gene length normalization and adds to a new source of bias[177]. For this study, the reads were only normalized by the sequencing depth. The focus of the differentially expression comparisons in this study is on the difference between biological conditions in same gene, not the difference between genes. Thus, the gene length normalization is not necessary for this study.

Unlike the intensity measurement from microarray, RNA-Seq generates discrete count data. The general statistical approaches like student t-test or Wilcoxon non-parametric test might not be the proper methods for differentially expression detection in RNA-Seq. These methods are too conservative so that the biological signals couldn't be found because of the small sample size and the measurement noise. Thus, negative binomial model[181], poisson model[180] and other non-parametric approaches[182] have been introduced to fit RNA-Seq data distribution for differentially expression analysis. For the current study, the differential expression gene calculation was done using DESeq2. It assumes that the count data follows the negative binomial distribution, and uses generalized linear model to find the significant expressed genes. DESeq2 is chosen due to the low FDR and lower computational requirements compared to other methods[183]. One should always select the proper method based on data heterogeneity, sample size, experimental setting and the computational loading. Suitable statistical models would help in making either the right conclusions or achieving results that might be used to generate new hypothesis for testing them in functional experiments.

RRBS

In order to achieve the right statistics and coverage of CpG sites, I modified several computational parameters starting from preprocessing until the downstream analysis in RRBS data. RnBeads helps me to handle memory issues efficiently and accounts for many covariates. It provides me a good overview and descriptive statistic of data quality. However, the differential methylation site analysis in Rnbeads is based on R package LIMMA[184] (Linear Models for Microarray Data) which is designed for gene expression microarray data. For the statistical point of view, RRBS data does not fit this model assumption. The processed RRBS data gives the number of methylated C and un-methylated C. The proportion of unchanged Cs regarded as the absolute DNA methylation level. The proportion usually follows the bimodal distribution instead of normally distributed data assumption. Additionally, methylation data is limited within the range of 0 and 1, and therefore variability is much smaller at the extreme values. Thus, another tool instead of RnBeads was chosen for differently methylation site detection.

There are several common ways to analysis methylation data DMP (differentially methylated positions) and DMR (differentially methylated regions). In this regard, Bayesian-Beta binominal model seems to be the best-fit model for DMP and DMR detection. Many methods are built by this assumption, like Biseq[125], MOABS[185] and DSS[110], usually, the strong effect is not sensitive to these chosen method. DSS (Dispersion shrinakge for sequencing data) was chosen because of the user-friendly interface and the character for processing the sparse data like RRBS. Any statistical method that tests for millions of CpGs needs to face the multiple correction issues to avoid many false positives results. Only the strongest single-CpG site differences would remain significant after correction for multiple testing. However, this conservative strategy might ignore the biological effect with less strong signal. In order to, not to over-correct the multiple testing, hierarchal testing approach (2.2.6) was applied in this study to identify the methylation signals.

4.4 Outlook for clinical applications in intestinal inflammation

4.4.1 Detection of Biomarkers for diagnosis or monitoring of IBD

Biomarker in medicine refers to the measurable physical, functional, biochemical indicators of which can identify the physiological changes, or disease processes. Several biological measurements have been identified as biomarkers of IBD[186]. C-reactive protein (CRP) is produced by hepatocytes in response to inflammation, stimulated by certain cytokines. CRP levels increase significantly during IBD, but the rise of CRP might also be due to infection, autoimmune conditions, other inflammatory conditions, and malignancy as well as cell necrosis[187]. Erythrocyte sedimentation rate (ESR) is another biomarker for IBD[188]. ESR is an indirect measurement of plasma acute phase protein concentration. Like CRP, ESR is a generally detection index of systemic inflammation, not entirely specific to IBD. The correlation between ESR and UC is good, but less accurate with CD. Certainly, genetic variants are also important biomarkers in IBD. With the rapid improvement of sequencing techniques, a number of genome-wide association study (GWAS) has discovered susceptibility genetic locies in IBD. PRDM1 (PR domain zinc finger protein 1) and NDP52 (Nuclear Domain 10 Protein 52) were determined to increase susceptibility to CD[189]. A latest study from the Wellcome Trust Sanger Institute and their collaborators have identified the genetic variant of ADCY7 (Adenylate Cyclase 7) that doubles an individual's risk of developing UC[190].

Changes in gut microbiota profiles and methylation patterns also strongly associated with IBD. Both changes are considered as the potential important biomarkers for the disease. Hypermethylation in several gene promoter regions including *APC* (Adenomatous polyposis coli), *TIMP3* (Tissue Inhibitor Of Metalloproteinases 3) were found aberrant in IBD-related colorectal cancer (CRC) patients[191]. Moreover, Carmona and colleague discovered the increase methylation pattern in several genes in tumor samples compared to normal tissue from CRC patient biopsies. Three of their findings (*AGTR1:* Angiotensin II Receptor Type 1*, WNT2:* Wnt Family Member 2 and *SLIT2:* Slit Guidance Ligand 2)

were validated in stool DNA with same hypermethylation pattern of affected CRC patients (with a detection sensitivity of 78%)[192]. A review paper from Karatza *et al* listed the suspicious methylation position and related genes as potential biomarkers in IBD[156]. Among all the biomarker in previous studies, the methylation site around TNFSF10 which can discriminate between disease and control in UC was found both from Ventham *et al* [60] and this twins study. TNFSF10 has been shown to be subject to methylation-dependent silencing in cancer cells[193] and might serve as an UC-detection biomarker in the future. To summarize, the methylation status in these candidate genes from stool or tissues can serve as biomarkers to intestinal inflammation diagnose.

Comparable to epigenetic signatures, dysbiosis of microbiota in IBD can also serve as a clinical biomarker. In a pediatric IBD study, researcher used anti-TNF therapy which increased relative abundance of Gram-positive bacteria (especially Clostridium clusters IV and XIVa). These bacteria were found associated with patients responding to anti-TNF therapy[194]. Interestingly, Clostridium clusters XIVa was identified in my twin study and showed strong correlation with methylation modulated genes (Section 3.1.4). Hence, the change of Clostridium clusters XIVa aboudance might be a potential index for IBD diagnose. Furthermore, the decrease of Faecalibacterium prausnitzii population in the resected ileum correlated with increased risk of recurrence from a post-operative recurrence cohort in CD[195]. The information of the change in candidate bacteria abundance can be extracted very easily from stool sample in routine health check. Especially for the people with IBD family history, this non-invasive test can monitor the potential disease status progress.

Although more and more biological signals from omics studies (microbiome, epigenome…etc) have been identified as statistically different in intestine inflammation scenario, only few of them are useful for clinical application in IBD practice[196]. One reason is the inconsistent of population stratification and patient materials between researches. IBD is a genetic associated disease, the population stratification is very essential. The significance of prevalence exists between different populations and contries[33]. Moreover, the collection of patient material is also matter. In IBD studies, some of them had biopsy, other groups employed blood sample and another took the stool. This inconsistent might

lead to different results, even gene expression pattern differs in different location of intestine in the same indiviual[19]. The biomarkers, which were found differentially between healthy control and patients, really depended on the material of sample collection. The other reason is the limitation of the experiment design; mainly studies so far were cross-sectional studies. Researcher aiming for biomarker detection in earlier diagnose of IBD should design as prospective study to monitor disease progression in the same individual. Because both the gut microbiome and the epigenome are very sensitive to environmental factors and the biological variation is huge in every individual. The longitudinal study could enhance the reliability for the target groups. With the drop of sequencing price nowadays, the study design in the future could be more flexible.

In conclusion, the results in my study might bring some hints for understanding the biological function of IBD and the raise of omics-integrate studies could provide better insight of biomarker detection.

4.4.2 Environmental effects as risk factors in intestinal inflammation

This research focused on host-microbiota interactions in intestinal inflammation and the dynamic pattern in during intestine development. Apart from the microbiota, many other environmental factors also play a role in intestinal inflammation. These factors not only contribute to intestine status individually but also interact with the other factors (Figure 53). The interplay between genetics, immunology, environment and microbiome has been shown in several studies. IBD develops at the intersection of genetic predisposition (leading to immunological abnormalities), dysbiosis of the gut microbiota and environmental influences[33]. Some environmental factors such as diet, have been identified as crucial factors for IBD and widely studied. However, there are several factors also contributed to IBD have not yet generally investigated. In the following part, the impact of other risk factors of IBD including smoking, vitamin D intake and sleeping sleep disturbance will be discussed. These factors might potentially become the further research direction of IBD.

The relationship between smoking and IBD is complex. Smoking seems to have opposite effects in UC and CD. Surprisingly, smoking is a risk factor for CD, but protective for UC[197]. There was a Swedish twins follow-up study showing that twin who smoked might develop CD whereas the other non-smoking twin might develop UC[198]. A meta-analysis study compared the IBD risk of current smoking and ex-smoking people[199]. They found that current smoking had a protective effect on the development of UC when compared with controls. Smoking was shown to affect T cells which express the α7 nicotinic receptor causing production of T helper (TH) 1 cytokine interferon-γ which has been associated with CD but not to UC[200]. From the epidemiology point, interestingly, highest adult male smokers countries (60-70%) like China and Mongolia with low prevalence of IBD, whereas countries like Sweden and Canada with overall high IBD rates have lower percentage of male smokers (17-28%)[201]. Even the smoking has the protective effect for UC statistically, one should be aware that the evidence only showed the association not causality. More functional study need to be established for verifying the role of smoking in IBD. To conclude, smoking as the IBD-related environmental factor is neither necessary nor sufficient to cause or protect IBD[202].

Vitamin D insufficiency can be found in up to 60% to 70% of IBD patients[203], but because of the well-known chicken-or-the-egg–type dilemma it cannot be described as causative for the disease outbreak. Nevertheless, vitamin D deficiency is a significant component in the development of IBD. The Vitamin D level in the body has strong connection with sunlight (ultraviolet B rays) exposure. During exposure to sunlight, 7-dehydrocholesterol in the skin absorbs ultraviolet B radiation and then converted to pre-vitamin D3 which in turn isomerizes into vitamin D3[204]. By using national health insurance databases, researcher found that high residential sunlight exposure was associated with a significant decrease in risk of CD, but not UC[205]. Vitamin D can be considered as a hormone with a number of effects on the immune system that are responsible for mediating susceptibility to infections[206]. Furthermore, one latest GWAS study suggested that the genetic variation at Vitamin D receptor (VDR) locus significantly influences microbial co-metabolism and the gut–liver axis[207]. In the future, it might be interesting to perform a mouse colitis experiment with DSS mice model in different UV exposure setting to discover the direct/indirect effect for IBD and furthermore investigate the change of vitamin D level as

well as gut microbiota composition. Further understanding of the relationship between IBD and vitamin D might be helpful for developing personalized therapies, e.g. UV exposure.

Sleep disturbance has also been identified as a risk factor of IBD. Patients with IBD are at increased risk for altered sleep patterns[208]. Shift workers with disrupted sleep pattern get higher risk in some gastrointestinal diseases (e.g. gastroesophageal reflux disease[209] and peptic ulcer disease[210]). Thaiss *et al.* showed that the diurnal oscillations of intestinal microbiota shape leads to time-specific compositional and functional profiles over the course of a day in both mice and humans[211]. Fecal transplantiation of human stool from donors with jet lag to mice resulted in glucose intolerance and obesity[211]. It might be interesting to investigate the effect of sleep-disturbance dysbiosis in intestine inflammation. One could first design a mouse study with two genotype in two different conditions: wild type vs IBD associated knock-out gene (NOD2, ATG16…etc) genotype and normal sleeping pattern vs sleeping disturbance. Through this study design, one could observe changes of gut microbiota pattern and discover the potential impact on sleeping disturbance. The future knowledge of identified pathways in pathophysiology and course of IBD may lead to the most appropriate therapies applying in an individual approach. For some patients, curing IBD via sleep pattern adjustment might be more efficient than invasive medical procedures[212] which can prevent the waste of the medical resource and reduce the pain of the patient during the treatment.

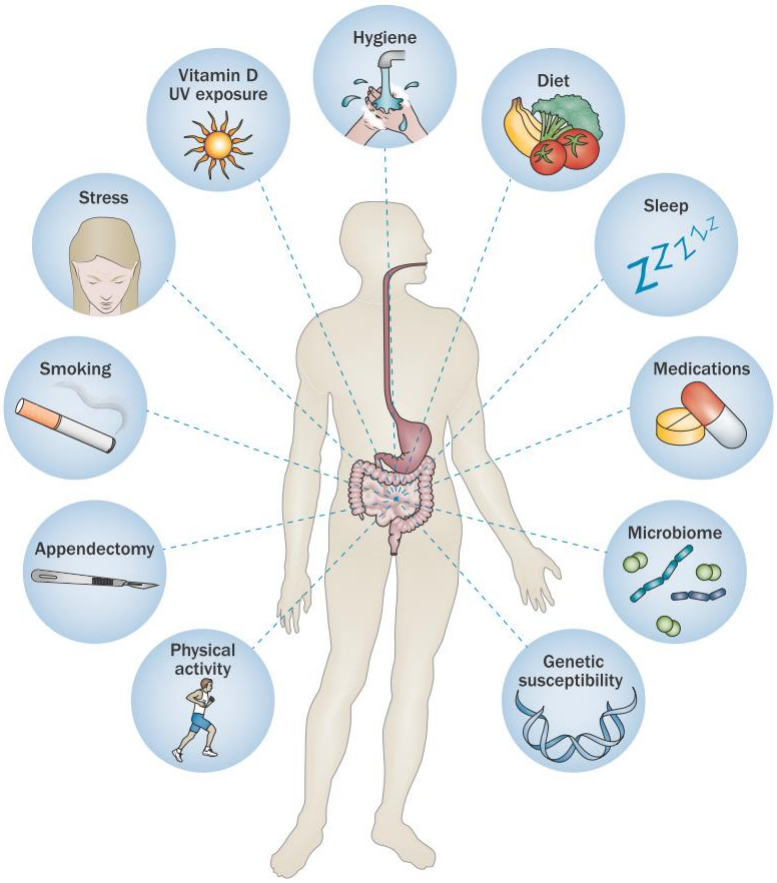


Figure 53 The important factors for IBD development

None of the risk factors alone are sufficient for the development of the disease and complex interactions between each factor occur before IBD break-out. Figure is modified from Ananthakrishnan et al[33]

4.5 Conclusion

With the advantage of omics research, it became clear that most complex diseases such as IBD arise not due to a single factor in a single layer. DNA methylation as the non-genetic medium may influence host gene expression through other environmental factors. A dysfunction of gene-environmental interaction in each biological layer might trigger intestinal inflammation. Besides, the balance of gut microbiota composition maintains the function of the host immune system of the gut. In order to understand and moderate IBD development, it is important to study interaction between DNA methylation and gut microbiota as well as the interplay with other factors.

This thesis mainly identified the DNA methylation marks and the corresponding bacterial effect that could contribute to the UC pathophysiology and the dynamic pattern during the intestinal maturation. Biological validation of this study is important to confirm the finding here and then investigate the pathway for further clinical application.

5. Summary

The role of epigenetic alterations and the interplay with the intestinal microbiota and inflammation is still not fully understood. I herein employed high throughput genomic screening technique to investigate the influence of interaction between host transcriptome, host epigenome and intestinal microbiota in human and mice. In my first twins study, a three-layer epigenome-wide association study (EWAS) is reported, using intestinal biopsies from ten monozygotic twin pairs discordant for the manifestation of UC by employing NGS (16S rRNA gene sequencing) and microarray (HM27 and Affeymetrix U133). Furthermore, the finding are validated in independent cohort with UC and healthy control (n=20 in two groups). The identified candidate genes have been functionally implicated in regulation of inflammatory response, and the identified bacterial genera have the potential impact of methylation modification. The targeted genes and bacteria could be taken further for technological and biological validation to identify their associations with IBD disease etiology and metabolic disorders.

The second mouse study showed the bacterial effect during intestine maturation process in intestinal epithelial cells (IECs) in GF and CONV-R mice in three different development stages. RNA-Seq and RRBS were employed for measuring the dynamic pattern of transcriptome and methylome. Postnatal development was observed to affect DNA methylation signatures and expression in IECs indicating that epigenetic processes contribute to developmental transitions largely driven by endogenous programs, independent of microbial cues. However, this data clearly shows that the gut microbiota influences specific modules of the epithelial transcriptional network during postnatal development and targets only a subset of microbially responsive genes mainly functioning in IEC proliferation and immune responses through their DNA methylation status.

To summarize, the results shown here confirms that the host-microbiota interaction is a critical check pint for intestinal inflammation and development. Though, it is still a debate whether the interaction is a cause or consequence of the disease, the results indicate a potential role of epigenetic modification in disease manifestation of UC or postnatal development. The finding might be helpful to support the combinational epigenetic and microbiota based therapies of intestine inflammation.

6. Zusammenfassung

Der Einfluss des Mikrobioms auf epigenetische Muster und Differenzierungsprozesse in Zellen der intestinalen Mukosa ist noch immer weitgehend unverstanden. Die vorliegende Studie unternimmt den Versuch, diesen Zusammenhang in verschiedenen Modellen näher zu beleuchten. Ich habe hierbei genomische Hochdurchsatzanalysen, in Mäusen und Menschen, verwendet, um parallel Transkriptom- und Epigenomsignaturen (DNA Methylierung) in der Mukosa als auch die phylogenetische Diversität des intestinalen Mikrobioms zu untersuchen. Im ersten Teil der Arbeit wird eine drei-stufige, epigenomweite Assoziationsstudie vorgestellt. In dieser Studie wurden intestinale Biopsien von zehn eineiigen Zwillingspaaren, welche diskordant für die Krankheitsausprägung Colitis ulcerosa (UC) sind, mittels moderner Sequenzierungsverfahren (16S rRNA Gensequenzierung) und Microarray (HM27 und Affeymetrix U133) analysiert. Die Ergebnisse aus dieser Studie wurden mit Hilfe einer unabhängigen Kohorte an UC-Patienten und gesunden Individuen (n = 20 in zwei Gruppen) validiert. Ein Schwerpunkt der Analyse lag hierbei auf der Gruppe der entzündungsregulierenden Gene. Es konnten weiterhin Beziehungen zwischen bestimmten bakteriellen Taxa und DNA-Methylierungsmuster des Wirts nachgewiesen werden. Die Befunde bilden eine interessante Grundlage, um den Zusammenhang von Mikrobiota als Umweltfaktor mit der Entstehung von chronisch entzündlichen Darmkrankheiten weiter funktionell zu charakterisieren.

Der zweite Teil der Arbeit beschäftigte sich mit dem Einfluss von kommensalen Bakterien auf Differenzierungsprozesse intestinaler Epithelzellen in Mäusen. Hierbei wurden Darmepithelzellen (IECs) aus GF und CONV-R Mäusen in drei verschiedenen Entwicklungsstadien entnommen. Zur Messung der dynamischen Muster des Transkriptoms und des Methyloms wurden RNA-Seq und RRBS (reduced representation bisulfite sequencing) angewendet. Es wurde beobachtet, dass die DNA-Methylierung und Genexpression in IECs durch das postnatale Entwicklungsstadium beeinflusst wird. Dies deutet darauf hin, dass epigenetische Prozesse zur Weiterentwicklung beitragen und die zugrundeliegenden endogenen Programme weitestgehend unabhängig von mikrobiellen Einflüssen funktionieren. Die Daten zeigen aber auch, dass die Darmflora, im postnatalen

Entwicklungsverlauf, spezifische Teile der epithelialen Transkriptmuster beeinflussen kann. Dabei wird nur eine kleine Fraktion von mikrobiellen Reaktionsgenen, welche im Bereich des IEC Wachstums und der Immunantwort wirkt, durch deren DNA-Methylierung beeinflusst.

Zusammenfassend bestätigen die gezeigten Ergebnisse, dass die Interaktion von Wirt und Mikroflora einen wichtigen Kontrollpunkt der intestinalen Entwicklung, aber auch von Entzündungsprozessen darstellt. Die Diskussion, ob die Interaktionen ein Krankheitssymptom oder eine Krankheitsursache darstellen, steht noch immer im Raum. Dennoch deuten die Ergebnisse auf eine potentielle Rolle der DNA Methylierung sowohl beim postnatalen Wachstum und funktionellen Differenzierung der Mukosa als auch bei der Entstehung von chronisch-entzündlichen Erkrankungen hin. Diese Ergebnisse könnten bei der Suche nach kombinierten, epigenetisch und mikrobiell basierten Therapien gegen Darmentzündungen hilfreich sein.

7. Reference

1. Hooper, L. V *et al.* Commensal host-bacterial relationships in the gut. *Science* **292,** 1115–8 (2001).

2. Sommer, F. & Bäckhed, F. The gut microbiota — masters of host development and physiology. *Nat. Rev. Microbiol.* **11,** 227–238 (2013).

3. Lozupone, C. A., Stombaugh, J. I., Gordon, J. I., Jansson, J. K. & Knight, R. Diversity, stability and resilience of the human gut microbiota. *Nature* **489,** 220–230 (2012).

4. Sender, R., Fuchs, S. & Milo, R. Revised estimates for the number of human and bacteria cells in the body. *bioRxiv* (2016).

5. Weinstock, G. M. Genomic approaches to studying the human microbiota. *Nature* **489,** 250–6 (2012).

6. Candela, M. *et al.* Interaction of probiotic Lactobacillus and Bifidobacterium strains with human intestinal epithelial cells: Adhesion properties, competition against enteropathogens and modulation of IL-8 production. *Int. J. Food Microbiol.* **125,** 286–292 (2008).

7. Sonnenburg, J. L. *et al.* Glycan Foraging in Vivo by an Intestine-Adapted Bacterial Symbiont. *Science (80-.).* **307,** (2005).

8. Olszak, T. *et al.* Microbial Exposure During Early Life Has Persistent Effects on Natural Killer T Cell Function. *Science (80-.).* **336,** (2012).

9. Ley, R. E., Turnbaugh, P. J., Klein, S. & Gordon, J. I. Microbial ecology: Human gut microbes associated with obesity. *Nature* **444,** 1022–1023 (2006).

10. Kau, A. L., Ahern, P. P., Griffin, N. W., Goodman, A. L. & Gordon, J. I. Human nutrition, the gut microbiome and the immune system. *Nature* **474,** 327–36 (2011).

11. Frank, D. N. *et al.* Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proc. Natl. Acad. Sci. U. S. A.* **104,** 13780–5 (2007).

12. Gonzalez, A. *et al.* The mind-body-microbial continuum. *Dialogues Clin. Neurosci.* **13,** 55–62 (2011).

13. Lupton, J. R. Microbial degradation products influence colon cancer risk: the butyrate controversy. *J. Nutr.* **134,** 479–82 (2004).

14. Michail, S. *et al.* Alterations in the gut microbiome of children with severe ulcerative colitis. *Inflamm. Bowel Dis.* **18,** 1799–1808 (2012).

15. Schaubeck, M. *et al.* Dysbiotic gut microbiota causes transmissible Crohn's disease-like ileitis independent of failure in antimicrobial defence. *Gut* **65,** 225–37 (2016).

16. Mosca, A., Leclerc, M. & Hugot, J. P. Gut Microbiota Diversity and Human Diseases: Should We Reintroduce Key Predators in Our Ecosystem? *Front. Microbiol.* **7,** 455 (2016).

17. Peterson, L. W. & Artis, D. Intestinal epithelial cells: regulators of barrier function and immune homeostasis. *Nat. Rev. Immunol.* **14,** 141–153 (2014).

18. Rosenstiel, P. Stories of love and hate. *Curr. Opin. Gastroenterol.* **29,** 125–132 (2013).

19. Sommer, F. *et al.* Site-specific programming of the host epithelial transcriptome by the gut microbiota. *Genome Biol.* **16,** 62 (2015).

20. Hampe, J. *et al.* Association between insertion mutation in NOD2 gene and Crohn's disease in German and British populations. *Lancet* **357,** 1925–1928 (2001).

21. Camp, J. G. *et al.* Microbiota modulate transcription in the intestinal epithelium without remodeling the accessible chromatin landscape. *Genome Res.* **24,** 1504–16 (2014).

22. Stadnyk, A. W. in *Encyclopedia of Life Sciences* (John Wiley & Sons, Ltd, 2009). doi:10.1002/9780470015902.a0003816.pub2

23. Frantz, A. L. The Role of Intestinal Epithelial Cells and the Regulation of the Polymeric Immunoglobulin Receptor in Homeostasis and Inflammation (Doctoral dissertation). *Univ. Kentucky* (2012).

24. Umar, S. Intestinal stem cells. *Curr. Gastroenterol. Rep.* **12,** 340–8 (2010).

25. Maloy, K. J. & Powrie, F. Intestinal homeostasis and its breakdown in inflammatory bowel disease. *Nature* **474,** 298–306 (2011).

26. Yu, L. C.-H., Wang, J.-T., Wei, S.-C. & Ni, Y.-H. Host-microbial interactions and regulation of intestinal epithelial barrier function: From physiology to pathology. *World J. Gastrointest. Pathophysiol.* **3,** 27–43 (2012).

27. Khor, B., Gardet, A. & Xavier, R. J. Genetics and pathogenesis of inflammatory bowel disease. *Nature* **474,** 307–317 (2011).

28. Loftus, E. V. Clinical epidemiology of inflammatory bowel disease: incidence, prevalence, and environmental influences. *Gastroenterology* **126,** 1504–1517 (2004).

29. Molodecky, N. A. *et al.* Increasing Incidence and Prevalence of the Inflammatory Bowel Diseases With Time, Based on Systematic Review. (2012). doi:10.1053/j.gastro.2011.10.001

30. Hein, R., Köster, I., Bollschweiler, E. & Schubert, I. Prevalence of inflammatory bowel disease: estimates for 2010 and trends in Germany from a large insurance-based regional cohort. *Scand. J. Gastroenterol.* **49,** 1–11 (2014).

31. Loftus, E. V *et al.* Ulcerative colitis in Olmsted County, Minnesota, 1940-1993: incidence, prevalence, and survival. *Gut* **46,** 336–43 (2000).

32. M. Orholm, V. Binder, T. I. A. Søre, V. B. T. I. A. S. L. P. R. K. O. K. Concordance of Inflammatory Bowel Disease among Danish Twins: Results of a Nationwide Study. *Scand. J. Gastroenterol.* **35,** 1075–1081 (2000).

33. Ananthakrishnan, A. N. Epidemiology and risk factors for IBD. *Nat. Rev. Gastroenterol. Hepatol.* **12,** 205–217 (2015).

34. Probert, C. S. *et al.* Prevalence and family risk of ulcerative colitis and Crohn's disease: an epidemiological study among Europeans and south Asians in Leicestershire. *Gut* **34,** 1547–51 (1993).

35. Liu, J. Z. & Anderson, C. A. Genetic studies of Crohn's disease: past, present and future. *Best Pract. Res. Clin. Gastroenterol.* **28,** 373–86 (2014).

36. Liu, J. Z. *et al.* Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* **47,** 979–86 (2015).

37. Hold, G. L. *et al.* Role of the gut microbiota in inflammatory bowel disease pathogenesis: what have we learnt in the past 10 years? *World J. Gastroenterol.* **20,** 1192–210 (2014).

38. Neut, C. *et al.* Changes in the bacterial flora of the neoterminal ileum after ileocolonic resection for Crohn's disease. *Am. J. Gastroenterol.* **97,** 939–946 (2002).

39. Damaskos, D. & Kolios, G. Probiotics and prebiotics in inflammatory bowel disease: microflora 'on the scope'. *Br. J. Clin. Pharmacol.* **65,** 453–67 (2008).

40. Moayyedi, P. Fecal transplantation : any real hope for inflammatory bowel disease? *Curr. Opin. Gastroenterol.* **32,** 282–286 (2016).

41. Bjerrum, J. T., Rantalainen, M., Wang, Y., Olsen, J. & Nielsen, O. H. Integration of transcriptomics and metabonomics: improving diagnostics, biomarker identification and phenotyping in ulcerative colitis. *Metabolomics* **10,** 280–290 (2014).

42. Cardinale, C. J. *et al.* Transcriptome profiling of human ulcerative colitis mucosa reveals altered expression of pathways enriched in genetic susceptibility loci. *PLoS One* **9,** e96153 (2014).

43. Heller, R. A. *et al.* Discovery and analysis of inflammatory disease-related genes using cDNA microarrays. *Proc. Natl. Acad. Sci. U. S. A.* **94,** 2150–5 (1997).

44. Dieckgraefe, B. K., Stenson, W. F., Korzenik, J. R., Swanson, P. E. & Harrington, C. A. Analysis of mucosal gene expression in inflammatory bowel disease by parallel oligonucleotide arrays. *Physiol. Genomics* **4,** 1–11 (2000).

45. Granlund, A. van B. *et al.* Whole genome gene expression meta-analysis of inflammatory bowel disease colon mucosa demonstrates lack of major differences

between Crohn's disease and ulcerative colitis. *PLoS One* **8,** e56818 (2013).

46. Olsen, J. *et al.* Diagnosis of ulcerative colitis before onset of inflammation by multivariate modeling of genome-wide gene expression data. *Inflamm. Bowel Dis.* **15,** 1032–1038 (2009).

47. Noble, C. L. *et al.* Regional variation in gene expression in the healthy colon is dysregulated in ulcerative colitis. *Gut* **57,** 1398–1405 (2008).

48. Satsangi, J., Jewell, D. P. & Bell, J. I. The genetics of inflammatory bowel disease. *Gut* **40,** 572–4 (1997).

49. Hasler, R. *et al.* A functional methylome map of ulcerative colitis. *Genome Res.* **22,** 2130–2137 (2012).

50. Low, D., Mizoguchi, A. & Mizoguchi, E. DNA methylation in inflammatory bowel disease and beyond. *World J. Gastroenterol.* **19,** 5238–49 (2013).

51. Saito, S. *et al.* DNA methylation of colon mucosa in ulcerative colitis patients: Correlation with inflammatory status. *Inflamm. Bowel Dis.* **17,** 1955–1965 (2011).

52. Foran, E. *et al.* Upregulation of DNA Methyltransferase-Mediated Gene Silencing, Anchorage-Independent Growth, and Migration of Colon Cancer Cells by Interleukin-6. *Mol. Cancer Res.* **8,** 471–481 (2010).

53. Zhang, Q. *et al.* STAT3 induces transcription of the DNA methyltransferase 1 gene (DNMT1) in malignant T lymphocytes. *Blood* **108,** 1058–1064 (2006).

54. Jones, B. & Chen, J. Inhibition of IFN-γ transcription by site-specific methylation during T helper cell development. *EMBO J.* **25,** 2443–2452 (2006).

55. Gonsky, R. *et al.* Distinct IFNG methylation in a subset of ulcerative colitis patients based on reactivity to microbial antigens. *Inflamm. Bowel Dis.* **17,** 171–178 (2011).

56. Huidobro, C. *et al.* A DNA methylation signature associated with aberrant promoter DNA hypermethylation of DNMT3B in human colorectal cancer. *Eur. J. Cancer* **48,** 2270–2281 (2012).

57. Franke, A. *et al.* Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat. Genet.* **42,** 1118–1125 (2010).

58. Cooke, J. *et al.* Mucosal genome-wide methylation changes in inflammatory bowel disease. *Inflamm. Bowel Dis.* **18,** 2128–2137 (2012).

59. Nimmo, E. R. *et al.* Genome-wide methylation profiling in Crohn's disease identifies altered epigenetic regulation of key host defense mechanisms including the Th17 pathway. *Inflamm. Bowel Dis.* **18,** 889–899 (2012).

60. Ventham, N. T. *et al.* Integrative epigenome-wide analysis demonstrates that DNA methylation may mediate genetic risk in inflammatory bowel disease. *Nat. Commun.* **7,** 13507 (2016).

61. Sadler, T. *et al.* Genome-wide analysis of DNA methylation and gene expression

defines molecular characteristics of Crohn's disease-associated fibrosis. *Clin. Epigenetics* **8,** 30 (2016).

62. Jones, P. A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.* **13,** 484–92 (2012).

63. Anderson, C. A. *et al.* Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. *Nat. Genet.* **43,** 246–252 (2011).

64. Kellermayer, R. *et al.* Epigenetic maturation in colonic mucosa continues beyond infancy in mice. *Hum. Mol. Genet.* **19,** 2168–2176 (2010).

65. Alenghat, T. *et al.* Histone deacetylase 3 coordinates commensal-bacteria-dependent intestinal homeostasis. *Nature* **504,** 153–7 (2013).

66. Kellermayer, R. *et al.* Colonic mucosal DNA methylation, immune response, and microbiome patterns in Toll-like receptor 2-knockout mice. *FASEB J.* **25,** 1449–60 (2011).

67. Mischke, M. & Plösch, T. More than just a gut instinct–the potential interplay between a baby's nutrition, its gut microbiome, and the epigenome. *Am. J. Physiol. - Regul. Integr. Comp. Physiol.* **304,** (2013).

68. Arpaia, N. *et al.* Metabolites produced by commensal bacteria promote peripheral regulatory T-cell generation. *Nature* **504,** 451–455 (2013).

69. Yu, D.-H. *et al.* Postnatal epigenetic regulation of intestinal stem cells requires DNA methylation and is guided by the microbiome. *Genome Biol.* **16,** 211 (2015).

70. Van den Abbeele, P. *et al.* The host selects mucosal and luminal associations of coevolved gut microorganisms: a novel concept. *FEMS Microbiol. Rev.* **35,** 681–704 (2011).

71. Rodríguez, J. M. *et al.* The composition of the gut microbiota throughout life, with an emphasis on early life. *Microb. Ecol. Heal. Dis.* **26,** (2015).

72. Celluzzi, A. & Masotti, A. How Our Other Genome Controls Our Epi-Genome. *Trends Microbiol.* (2016). doi:10.1016/j.tim.2016.05.005

73. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* **74,** 5463–7 (1977).

74. Zhang, J., Chiodini, R., Badr, A. & Zhang, G. The impact of next-generation sequencing on genomics. *J. Genet. Genomics* **38,** 95–109 (2011).

75. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).

76. Heidi Chial. DNA Sequencing Technologies Key to the Human Genome Project. *Nat. Educ.* (2008).

77. Venter, J. C. *et al.* The Sequence of the Human Genome. *Science (80-.).* **291,**

(2001).

78. Mark Hollmer. Roche to close 454 Life Sciences as it reduces gene sequencing focus | FierceBiotech. (2013). Available at: http://www.fiercebiotech.com/medical-devices/roche-to-close-454-life-sciences-as-it-reduces-gene-sequencing-focus.

79. Ion Torrent bought by Life Technologies. (2013). Available at: https://www.wired.com/2010/08/ion-torrent-bought-by-life-technologies/.

80. Branton, D. *et al.* The potential and challenges of nanopore sequencing. *Nat. Biotechnol.* **26,** 1146–53 (2008).

81. Sanschagrin, S. & Yergeau, E. Next-generation Sequencing of 16S Ribosomal RNA Gene Amplicons. *J. Vis. Exp.* e51709–e51709 (2014). doi:10.3791/51709

82. Baker, G. C., Smith, J. J. & Cowan, D. A. Review and re-analysis of domain-specific 16S primers. *J. Microbiol. Methods* **55,** 541–555 (2003).

83. Clarridge, J. E. & III. Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. *Clin. Microbiol. Rev.* **17,** 840–62, table of contents (2004).

84. Illumina. 16S Metagenomic Sequencing Library. *Illumina.com* 1–28 (2013).

85. Ashelford, K. E., Chuzhanova, N. A., Fry, J. C., Jones, A. J. & Weightman, A. J. At Least 1 in 20 16S rRNA Sequence Records Currently Held in Public Repositories Is Estimated To Contain Substantial Anomalies. *Appl. Environ. Microbiol.* **71,** 7724–7736 (2005).

86. Hall, J. R. *et al.* Molecular characterization of the diversity and distribution of a thermal spring microbial community by using rRNA and metabolic genes. *Appl. Environ. Microbiol.* **74,** 4910–22 (2008).

87. Licatalosi, D. D. & Darnell, R. B. RNA processing and its regulation: global insights into biological networks. *Nat. Rev. Genet.* **11,** 75–87 (2010).

88. Costa, V., Angelini, C., De Feis, I. & Ciccodicola, A. Uncovering the Complexity of Transcriptomes with RNA-Seq. *J. Biomed. Biotechnol.* **2010,** 1–19 (2010).

89. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10,** 57–63 (2010).

90. Meissner, A. *et al.* Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res.* **33,** 5868–77 (2005).

91. Ziller, M. J. *et al.* Genomic Distribution and Inter-Sample Variation of Non-CpG Methylation across Human Cell Types. *PLoS Genet.* **7,** e1002389 (2011).

92. Deaton, A. M. & Bird, A. CpG islands and the regulation of transcription. *Genes Dev.* **25,** 1010–22 (2011).

93. Garrett-Bakelman, F. E. *et al.* Enhanced reduced representation bisulfite sequencing for assessment of DNA methylation at base pair resolution. *J. Vis.*

Exp. e52246 (2015). doi:10.3791/52246

94. Yong, W.-S. *et al.* Profiling genome-wide DNA methylation. *Epigenetics Chromatin* **9,** 26 (2016).

95. Andrews, S. Reduced Representation Bisulfite-Seq – A Brief Guide to RRBS. *Babraham Bioinforma.* 1–12 (2013).

96. Kachroo, P. Genome-wide mapping of the inflammation methylome (Doctoral dissertation). (Kiel university, 2015).

97. Kozich, J. J., Westcott, S. L., Baxter, N. T., Highlander, S. K. & Schloss, P. D. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the miseq illumina sequencing platform. *Appl. Environ. Microbiol.* **79,** 5112–5120 (2013).

98. Haas, B. J. *et al.* Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Res.* **21,** 494–504 (2011).

99. Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C. & Knight, R. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27,** 2194–200 (2011).

100. Chao, A. Nonparametric Estimation of the Number of Classes in a Population. *Biometrical Journal* **30,** 265–270 (1984).

101. Shannon, C. E. A Mathematical Theory of Communication. *he Bell Syst. Tech. J.* **27,** 379–423 (1948).

102. E. H. SIMPSON. Measurement of Diversity. *Nature* **163,** 688–688 (1949).

103. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14,** R36 (2013).

104. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9,** 357–359 (2012).

105. Anders, S., Pyl, P. T. & Huber, W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31,** 166–9 (2015).

106. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15,** 550 (2014).

107. Ghigna, C., Valacca, C. & Biamonti, G. Alternative splicing and tumor progression. *Curr. Genomics* **9,** 556–70 (2008).

108. Shen, S. *et al.* rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc. Natl. Acad. Sci. U. S. A.* **111,** E5593-601 (2014).

109. McCarthy, D. J., Chen, Y. & Smyth, G. K. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* **40,** 4288–97 (2012).

110. Feng, H., Conneely, K. N. & Wu, H. A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. *Nucleic Acids Res.* **42,** 1–11 (2014).

111. Tarca, A. L., Bhatti, G. & Romero, R. A comparison of gene set analysis methods in terms of sensitivity, prioritization and specificity. *PLoS One* **8,** e79217 (2013).

112. Freudenberg, J. M. Comparison of background correction and normalization procedures for high-density oligonucleotide microarrays (Diplom dissertation). *Univ. LEIPZIG* (2005).

113. Hubbell, E., Liu, W.-M. & Mei, R. Robust estimators for expression analysis. *BIOINFORMATICS* **18,** 1585–1592 (2002).

114. Wu, J. & Irizarry, R. Package 'gcrma' Background Adjustment Using Sequence Information. *R Packag.* (2016).

115. Irizarry, R. A. *et al.* Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4,** 249–264 (2003).

116. Author, T., Benjamini, Y., Hochberg, Y. & Benjaminit, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *Source J. R. Stat. Soc. Ser. B J. R. Stat. Soc. Ser. B J. R. Stat. Soc. B* **57,** 289–300 (1995).

117. Krueger, F. & Andrews, S. R. *Quality Control , trimming and alignment of Bisulfite-Seq data (Prot 57)*. *Epigenesys* (2012).

118. Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27,** 1571–2 (2011).

119. Assenov, Y. *et al.* Comprehensive analysis of DNA methylation data with RnBeads. *Nat. Methods* **11,** 1138–1140 (2014).

120. Bock, C. Analysing and interpreting DNA methylation data. *Nat. Rev. Genet.* **13,** 705–719 (2012).

121. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43,** e47–e47 (2015).

122. Akalin, A. *et al.* methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol.* **13,** R87 (2012).

123. Weisenberger, D. J. *et al.* Comprehensive DNA Methylation Analysis on the Illumina ® Infinium ® Assay Platform.

124. Du, P. *et al.* Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics* **11,** 587 (2010).

125. Hebestreit, K., Dugas, M. & Klein, H. U. Detection of significantly differentially methylated regions in targeted bisulfite sequencing data. *Bioinformatics* **29,** 1647–

1653 (2013).

126. CLARKE, K. R. Non-parametric multivariate analyses of changes in community structure. *Austral Ecol.* **18,** 117–143 (1993).

127. Kugathasan, S. *et al.* Loci on 20q13 and 21q22 are associated with pediatric-onset inflammatory bowel disease. *Nat. Genet.* **40,** 1211–5 (2008).

128. Schloss, P. D. *et al.* Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* **75,** 7537–7541 (2009).

129. Hannenhalli, S. Eukaryotic transcription factor binding sites--modeling and integrative search methods. *Bioinformatics* **24,** 1325–31 (2008).

130. Kaser, A. *et al.* XBP1 links ER stress to intestinal inflammation and confers genetic risk for human inflammatory bowel disease. *Cell* **134,** 743–56 (2008).

131. Hasegawa, D. *et al.* Epithelial Xbp1 is required for cellular proliferation and differentiation during mammary gland development. *Mol. Cell. Biol.* **35,** 1543–56 (2015).

132. Bhattacharyya, S., Fang, F., Tourtellotte, W. & Varga, J. Egr-1: new conductor for the tissue repair orchestra directs harmony (regeneration) or cacophony (fibrosis). *J. Pathol.* **229,** 286–97 (2013).

133. Fichtner-Feigl, S. *et al.* IL-13 signaling via IL-13R alpha2 induces major downstream fibrogenic factors mediating fibrosis in chronic TNBS colitis. *Gastroenterology* **135,** 2003–13, 2013–7 (2008).

134. Benizri, E., Ginouvès, A. & Berra, E. The magic of the hypoxia-signaling cascade. *Cell. Mol. Life Sci.* **65,** 1133–1149 (2008).

135. Formenti, F. *et al.* Regulation of human metabolism by hypoxia-inducible factor. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 12722–7 (2010).

136. Glover, L. E. & Colgan, S. P. Hypoxia and metabolic factors that influence inflammatory bowel disease pathogenesis. *Gastroenterology* **140,** 1748–55 (2011).

137. Lee, H. K., Hsu, A. K., Sajdak, J., Qin, J. & Pavlidis, P. Coexpression analysis of human genes across many microarray data sets. *Genome Res.* **14,** 1085–94 (2004).

138. Xue, J. *et al.* Transcriptome-based network analysis reveals a spectrum model of human macrophage activation. *Immunity* **40,** 274–88 (2014).

139. Gardina, P. J. *et al.* Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array. *BMC Genomics* **7,** 325 (2006).

140. Bisognin, A. *et al.* An integrative framework identifies alternative splicing events in colorectal cancer development. *Mol. Oncol.* **8,** 129–141 (2014).

141. Wong, J. J.-L., Au, A. Y. M., Ritchie, W. & Rasko, J. E. J. Intron retention in mRNA: No longer nonsense. *BioEssays* **38,** 41–49 (2016).

142. Fatemi, M., Hermann, A., Gowher, H. & Jeltsch, A. Dnmt3a and Dnmt1 functionally cooperate during de novo methylation of DNA. *Eur. J. Biochem.* **269,** 4981–4984 (2002).

143. Shen, L. *et al.* Tet3 and DNA replication mediate demethylation of both the maternal and paternal genomes in mouse zygotes. *Cell Stem Cell* **15,** 459–470 (2014).

144. Yamauchi, T. Neuronal Ca2+/Calmodulin-Dependent Protein Kinase II—Discovery, Progress in a Quarter of a Century, and Perspective: Implication for Learning and Memory. *Biol. Pharm. Bull.* **28,** 1342–1354 (2005).

145. Rullinkov, G. *et al.* Neuralized-2: expression in human and rodents and interaction with Delta-like ligands. *Biochem. Biophys. Res. Commun.* **389,** 420–5 (2009).

146. Saudino, K. J. Behavioral genetics and child temperament. *J. Dev. Behav. Pediatr.* **26,** 214–23 (2005).

147. Zhou, S. *et al.* Diversity of Gut Microbiota Metabolic Pathways in 10 Pairs of Chinese Infant Twins. *PLoS One* **11,** e0161627 (2016).

148. Dicksved, J. *et al.* Molecular analysis of the gut microbiota of identical twins with Crohn's disease. *ISME J.* **2,** 716–727 (2008).

149. Lepage, P. *et al.* Twin Study Indicates Loss of Interaction Between Microbiota and Mucosa of Patients With Ulcerative Colitis. *Gastroenterology* **141,** 227–236 (2011).

150. Balzola, F., Bernstein, C., Ho, G. T. & Lees, C. Environmentally-acquired bacteria influence microbial diversity and natural innate immune responses at gut surfaces: Commentary. *Inflamm. Bowel Dis. Monit.* **10,** 134 (2010).

151. Tomasz, J. U. & T, Z. K. Tumour Necrosis Factor Superfamily Members in the Pathogenesis of Inflammatory Bowel Disease. **2014,** (2014).

152. Waddell, A. *et al.* Intestinal CCL11 and eosinophilic inflammation is regulated by myeloid cell-specific RelA/p65 in mice. *J. Immunol.* **190,** 4773–85 (2013).

153. Waddell, A. *et al.* Colonic Eosinophilic Inflammation in Experimental Colitis Is Mediated by Ly6Chigh CCR2+ Inflammatory Monocyte/Macrophage-Derived CCL11. *J. Immunol.* **186,** 5993–6003 (2011).

154. Glória, L. *et al.* DNA hypomethylation and proliferative activity are increased in the rectal mucosa of patients with long-standing ulcerative colitis. *Cancer* **78,** 2300–6 (1996).

155. Ventham, N. T., Kennedy, N. A., Nimmo, E. R. & Satsangi, J. Beyond gene discovery in inflammatory bowel disease: The emerging role of epigenetics. *Gastroenterology* **145,** 293–308 (2013).

156. Karatzas, P. S., Gazouli, M., Safioleas, M. & Mantzaris, G. J. DNA methylation changes in inflammatory bowel disease. *Ann. Gastroenterol. Q. Publ. Hell. Soc. Gastroenterol.* **27,** 125–132 (2014).

157. Kang, K. *et al.* A Genome-Wide Methylation Approach Identifies a New Hypermethylated Gene Panel in Ulcerative Colitis. *Int. J. Mol. Sci.* **17,** (2016).

158. Häsler, R. *et al.* Uncoupling of mucosal gene regulation, mRNA splicing and adherent microbiota signatures in inflammatory bowel disease. *Gut* gutjnl-2016-311651 (2016). doi:10.1136/gutjnl-2016-311651

159. Haberman, Y. *et al.* Pediatric Crohn disease patients exhibit specific ileal transcriptome and microbiome signature. *J. Clin. Invest.* **124,** 3617–3633 (2014).

160. Van den Abbeele, P. *et al.* Butyrate-producing Clostridium cluster XIVa species specifically colonize mucins in an in vitro gut model. *ISME J.* **7,** 949–61 (2013).

161. den Besten, G. *et al.* The role of short-chain fatty acids in the interplay between diet, gut microbiota, and host energy metabolism. *J Lipid Res* **54,** 2325–2340 (2013).

162. Licciardi, P. V, Wong, S.-S., Tang, M. L. & Karagiannis, T. C. Epigenome targeting by probiotic metabolites. *Gut Pathog.* **2,** 24 (2010).

163. Canani, R. B. Potential beneficial effects of butyrate in intestinal and extraintestinal diseases. *World J. Gastroenterol.* **17,** 1519 (2011).

164. Berni Canani, R. *et al.* The epigenetic effects of butyrate: potential therapeutic implications for clinical practice. *Clin. Epigenetics* **4,** 4 (2012).

165. Gevers, D. *et al.* The treatment-naive microbiome in new-onset Crohn's disease. *Cell Host Microbe* **15,** 382–92 (2014).

166. Kraiczy, J. *et al.* Assessing DNA methylation in the developing human intestinal epithelium: potential link to inflammatory bowel disease. *Mucosal Immunol.* **9,** 1–12 (2015).

167. Larsson, E. *et al.* Analysis of gut microbial regulation of host gene expression along the length of the gut and regulation of gut microbial ecology through MyD88. *Gut* **61,** 1124–31 (2012).

168. Okano, M., Bell, D. W., Haber, D. A. & Li, E. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell* **99,** 247–257 (1999).

169. He, Y.-F. *et al.* Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science* **333,** 1303–7 (2011).

170. Kang, J., Kalantry, S. & Rao, A. PGC7, H3K9me2 and Tet3: regulators of DNA methylation in zygotes. *Cell Res.* **23,** 6–9 (2013).

171. El Aidy, S. *et al.* Temporal and spatial interplay of microbiota and intestinal

mucosa drive establishment of immune homeostasis in conventionalized mice. *Mucosal Immunol.* (2012). doi:10.1038/mi.2012.32

172. Tse-Wen Chang. Binding of cells to matrixes of distinct antibodies coated on solid surface. *J. Immunol. Methods* **65,** 217–223 (1983).

173. Maeda, H. *et al.* Quantitative real-time PCR using TaqMan and SYBR Green for Actinobacillus actinomycetemcomitans, Porphyromonas gingivalis, Prevotella intermedia, tetQ gene and total bacteria. *FEMS Immunol. Med. Microbiol.* **39,** 81–86 (2003).

174. Bibikova, M. *et al.* Genome-wide DNA methylation profiling using Infinium ® assay. *Epigenomics* **1,** 177–200 (2009).

175. Wong, N. C. *et al.* Exploring the utility of human DNA methylation arrays for profiling mouse genomic DNA. *Genomics* **102,** 38–46 (2013).

176. Pricing » Microarray and Sequencing Resource | Boston University. Available at: http://www.bumc.bu.edu/microarray/pricing/. (Accessed: 24th March 2017)

177. Oshlack, A. & Wakefield, M. J. Transcript length bias in RNA-seq data confounds systems biology. *Biol. Direct* **4,** 14 (2009).

178. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5,** 621–628 (2008).

179. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28,** 511–515 (2010).

180. Trapnell, C. *et al.* Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat. Biotechnol.* **31,** 46–53 (2012).

181. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11,** R106 (2010).

182. Li, J. & Tibshirani, R. Finding consistent patterns: A nonparametric approach for identifying differential expression in RNA-Seq data. *Stat. Methods Med. Res.* **22,** 519–536 (2013).

183. Seyednasrollah, F., Laiho, A. & Elo, L. L. Comparison of software packages for detecting differential expression in RNA-seq studies. *Brief. Bioinform.* **16,** 59–70 (2015).

184. Smyth, G. K. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor, chapter 23 Limma: Linear Models for Microarray Data*. (Springer, 2005).

185. Sun, D. *et al.* MOABS: model based analysis of bisulfite sequencing data. *Genome Biol.* **15,** R38 (2014).

186. Soubières, A. A. & Poullis, A. Emerging role of novel biomarkers in the diagnosis of inflammatory bowel disease. *World J. Gastrointest. Pharmacol. Ther.* **7,** 41–50 (2016).

187. Pepys, M. B. & Hirschfield, G. M. C-reactive protein: a critical update. *J. Clin. Invest.* **111,** 1805–1812 (2003).

188. DESAI, D., FAUBION, W. A. & SANDBORN, W. J. Review article: biological activity markers in inflammatory bowel disease. *Aliment. Pharmacol. Ther.* **25,** 247–255 (2006).

189. Ellinghaus, D. *et al.* Association between variants of PRDM1 and NDP52 and crohn's disease, based on exome sequencing and functional studies. *Gastroenterology* **145,** 339–347 (2013).

190. Luo, Y. *et al.* Exploring the genetic architecture of inflammatory bowel disease by whole-genome sequencing identifies association at ADCY7. *Nat. Genet.* **49,** 186–192 (2017).

191. Ashktorab, H. & Brim, H. DNA Methylation and Colorectal Cancer. *Curr. Colorectal Cancer Rep.* **10,** 425–430 (2014).

192. Carmona, F. J. *et al.* DNA Methylation Biomarkers for Noninvasive Diagnosis of Colorectal Cancer. *Cancer Prev. Res.* **6,** (2013).

193. Dylan Hunter, Leanne Edson, W. C. Loss of tumor necrosis factor superfamily genes in breast cancer cell lines. *FASEB J.* **28,** 1047.8 (2014).

194. Kolho, K.-L. *et al.* Fecal Microbiota in Pediatric Inflammatory Bowel Disease and Its Relation to Inflammation. *Am. J. Gastroenterol.* **110,** 921–930 (2015).

195. Sokol, H. *et al.* Faecalibacterium prausnitzii is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients. *Proc. Natl. Acad. Sci.* **105,** 16731–16736 (2008).

196. Boyapati, R. K., Kalla, R., Satsangi, J. & Ho, G. Biomarkers in Search of Precision Medicine in IBD. *Am. J. Gastroenterol.* **111,** 1–9 (2016).

197. Thomas, GA; Rhodes, J; Green, JT; Richardson, C. Role of smoking in inflammatory bowel disease: implications for therapy. *Postgr. Med J* **76,** 273–279 (2000).

198. Halfvarson, J., Bodin, L., Tysk, C., Lindberg, E. & Järnerot, G. Inflammatory bowel disease in a Swedish twin cohort: a long-term follow-up of concordance and clinical characteristics. *Gastroenterology* **124,** 1767–73 (2003).

199. Mahid, S. S., Minor, K. S., Soto, R. E., Hornung, C. A. & Galandiuk, S. Smoking and Inflammatory Bowel Disease: A Meta-analysis. *Mayo Clin. Proc.* **81,** 1462–1471 (2006).

200. Kikuchi, H., Itoh, J. & Fukuda, S. Chronic nicotine stimulation modulates the immune response of mucosal T cells to Th1-dominant pattern via nAChR by

upregulation of Th1-specific transcriptional factor. *Neurosci. Lett.* **432,** 217–221 (2008).

201. Kaplan, G. G. The global burden of IBD: from 2015 to 2025. *Nat. Rev. Gastroenterol. Hepatol.* **12,** 720–727 (2015).

202. Bernstein, C. N. Why and where to look in the environment with regard to the etiology of inflammatory bowel disease. *Dig. Dis.* **30 Suppl 3,** 28–32 (2012).

203. Narula, N. & Marshall, J. K. Management of inflammatory bowel disease with vitamin D: Beyond bone health. *J. Crohn's Colitis* **6,** 397–404 (2012).

204. Wacker, M. & Holick, M. F. Sunlight and Vitamin D: A global perspective for health. *Dermatoendocrinol.* **5,** 51–108 (2013).

205. Nerich, V. *et al.* Low exposure to sunlight is a risk factor for Crohn's disease. *Aliment. Pharmacol. Ther.* **33,** 940–945 (2011).

206. Ananthakrishnan, A. N. Vitamin D and Inflammatory Bowel Disease. *Gastroenterol. Hepatol. (N. Y).* **12,** 513–515 (2016).

207. Wang, J. *et al.* Genome-wide association analysis identifies variation in vitamin D receptor and other host factors influencing the gut microbiota. *Nat. Genet.* **48,** 1396–1406 (2016).

208. Swanson, G. R., Burgess, H. J. & Keshavarzian, A. Sleep disturbances and inflammatory bowel disease: a potential trigger for disease flare? *Expert Rev. Clin. Immunol.* **7,** 29–36 (2011).

209. Demeter, P. *et al.* Severity of gastroesophageal reflux disease influences daytime somnolence: a clinical study of 134 patients underwent upper panendoscopy. *World J. Gastroenterol.* **10,** 1798–801 (2004).

210. Segawa, K. *et al.* Peptic ulcer is prevalent among shift workers. *Dig. Dis. Sci.* **32,** 449–53 (1987).

211. Thaiss, C. A. *et al.* Transkingdom control of microbiota diurnal oscillations promotes metabolic homeostasis. *Cell* **159,** 514–529 (2014).

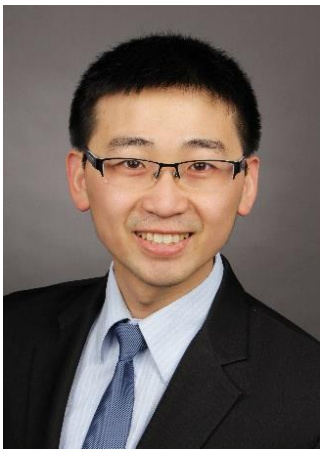
212. Sobolewska-Włodarczyk, A. *et al.* Circadian rhythm abnormalities – Association with the course of inflammatory bowel disease. *Pharmacol. Reports* **68,** 847–851 (2016).

213. Lee, W. & Living Stream Ministry. *The world situation and God's move*. (Living Stream Ministry, 1991).

9. Supplements

9.1 Curriculum Vitae

PERSONAL DATA



Name	Wei-Hung Pan
Address	Harmsstr. 48, 24114, Kiel Germany
Mobile	+49 17684212817
E-mail	n124080@gmail.com
LinkedIn	https://de.linkedin.com/in/Wei-Hung-Pan-30319479
Skype ID	joshuapanwh
Date / Place of Birth	16 December 1984 in Changhua City, Taiwan
Citizenship	Taiwanese

EDUCATION

03/2013 – 05/2017 (expected time) **Christian- Albrechts-Universität**, Kiel, Germany

Institute of Clinical Molecular Biology / University Hospital Schleswig-Holstein

- PhD Candidate in Bioinformatics / Hands-on data analysis experience
- Focus: Machine Learning, Data Integration, Clinical data analysis.

02/2008 – 01/2010 **National Tsing Hua University**, Hsinchu City, Taiwan

Institute of Statistics

- Master of Science in Statistics (GPA 3.8 = Top 15%)
- Focus: Statistical Simulation, Regression Model, Biostatistics
- Thesis Topic: Confidence Interval of Biodiversity Index (Adviser: Dr. Anne Chao)

09/2002 – 06/2006 **National Sun-Yet San University**, Kaohsiung City, Taiwan

Department of Applied Mathematics

- Bachelor of Science in Mathematics (GPA 3.54 = Top 10%)

WORK EXPERIENCE

01/2012 – 01/2013 **National Health Research Institute**, Miaoli County, Taiwan

Research assistance/Data analyst in institute of Cancer

11/2006 – 12/2007 Republic of China Air Force, Tainan and Penghu, Taiwan

PUBLICATION

11/2016	**Journal: Nature Genetics**
	Genome-Wide Association Analysis Identifies Variation in Vitamin D Receptor and Other Host Factors Influencing the Gut Microbiota – Wang J, Thingholm LB, Skiecevičienė J, Rausch P, Kummen M, Hov JR, Degenhardt F, Heinsen FA, Rühlemann MC, Szymczak S, Holm K, Esko T, Sun J, Pricop-Jeckstadt M, Al-Dury S, Bohov P, Bethune J, Sommer F, Ellinghaus D, Berge RK, Hübenthal M, Koch M, Schwarz K16, Rimbach G, Hübbe P, **Pan WH**, Sheibani-Tezerji R, Häsler R, Rosenstiel P, D'Amato M, Cloppenborg-Schmidt K, Künzel S, Laudes M, Marschall HU, Lieb W, Nöthlings U, Karlsen TH, Baines JF, Franke A
07/2017	**Journal: Genome Medicine (submitted)**
	Dynamic Methylation Signatures of Intestinal Epithelial Cells Reflect Postnatal Development – **Pan WH**, Sommer F, Falk-Paulsen M, Ulas T, Best P, Kachroo P, Luzius A, Jentzsch M, Rehman A, Müller F, Lengauer T, Walter J, Schreiber S, L Schultze J, Bäckhed F, Franke A, Rosenstiel P

PROGRAMING SKILLS

Proficient	• R (ggplot2,dplyr), Linux/UNIX shell, Python (pandas, numpy, scipy), Perl
Good	• Mathematica, SAS, C++

LANGUAGE SKILLS

Chinese	• Native Speaker
English	• Working Fluency
German	• Conversational Proficiency (Telc Zertifikat B1 : Gut)

AWARD AND SCHOLARSHIPS

09/2016	**Selected Talk and Travel Award** in Symposium 'The indigenous Microbiota', Mainz, Germany
12/2015	**Selected Talk** in Life Sciences Student Conference, Kiel, Germany
12/2009	**Scholarship**: Recipient of Chinese Budget-Account-Statistics Coordination and Development Society, Taiwan
06/2006	**Award for Excellent Graduate**, National Sun-Yet San University, Kaohsiung City, Taiwan
2004, 2005	**Award for Best in Class**, National Sun-Yet San University, Kaohsiung City, Taiwan

VOLUNTEER EXPERIENCE

02/2010 – 12/2011	**Bible Truth and Church Service Training** in the Church in Taipei, Taipei, Taiwan
	• Thesis Topic: Watchman Nee's View of New Jerusalem
08/2011 – 10/2011	**Christian Volunteer after Tsunami in Japan**, Sendai, Japan

9.2 Declaration

I hereby confirm that this thesis is exclusively the result of my own work. Apart from the advice of my supervisors, all sources are listed in the references. All articles also list the individual author contributions. This thesis has not been submitted elsewhere. It has been prepared in strict accordance with the rules of good scientific practice of the Deutsche Forschungsgesellschaft.

Erklärung

Hiermit versichere ich, dass diese Dissertation nach Inhalt und Form das Resultat meiner eigenen Arbeit ist. Abgesehen vom Rat meiner akademischen Lehrer sind sämtliche Quellen in den Referenzen aufgeführt. Der geleistete Beitrag eines Autors ist in der jeweiligen Publikation vermerkt. Diese Arbeit lag und liegt nirgends sonst im Rahmen eines Promotionsverfahrens vor. Die Arbeit wurde unter Einhaltung der Regeln guter wissenschaftlicher Praxis der Deutschen Forschungsgesellschaft verfasst.

Kiel, ...

.....

Wei-Hung Pan

9.3 Acknowledgements

“He who goes forth and weeps, Bearing seed for scattering”—Psalm 126:5

In March. 2013, I made a crucial decision in my life. I said goodbye with my family and my country, went alone to a foreign land, to pursue my PhD dream in a northern harbor in Germany. After four and half working hard years, I will finish the last academic degree of my life in Kiel. The change during this period is hard to explain in a few words. From single to married; from a totally non-German speaker to be able to handle the daily conversion; from a layman in bioinformatics to a professional and qualified PhD in this research field. I believe every tear during these years will become the nutrient in my life journey. Of course, I am not alone. I would like to thank the people below, without you, the research work can not be done.

To my mentor and Doctor father (Doktorvater), Philip: for giving me this great chance to enter in this interesting field, a lot of patience and tolerance for my immature idea and behavior; professional advice and helpful tips when I got stuck in the research direction; always support and trust me for being a great supervisor.

To the other postdocs and co-workers, Ateeq, Rob, Felix and Maren: for the valuable suggestions in the cooperated projects, and the correction of my thesis. I can not get this progress in the project without your help.

To good friends from RTG, Priya and Pankaj: for being together 3 years in RTG and support each other mentally and practically.

To all the colleagues in IKMB, Richa, Anupam, Phili, Jacqueline, Frauke, Anna, Steffi, Antonia, Alejandro, Go, Daniela…etc: for the alcohol/cookies/cake restoration on the red coach and sharing the funny story in the Room 4.20.

To all the brothers and sisters in the church: for non-stop encouraging each other, either in prayer or in the divine supply from God’s word. This mutual and spiritual encouragement helps me not to forget the initial calling of God to be here.

To my family, Dad, Mom, Brother, Aunts and all the other family members: for all the caring and understanding in my difficult moment, not to stress me but comfort me.

To My Dear Wife Ching-Ting: Being with me all the time especially in my last year of PhD. Holding my hands, standing together and growing up together in the grace of God. You are the best gift to me from God.

Sorry for those who didn't list above, I hope you also receive my appreciation for all your help in my research work.

Last and the most important, to my dear Lord Jesus: Thanks for bringing me here in Germany, to be you testimony and part of the divine history. For remind me not to forget the responsibility as a Christian in the world.

I want to quote some sentences from the book "The World Situation and God's Move" as my ending: *(As a Christian) What is our responsibility? We must bear the testimony of Jesus…We must be witnesses to Him... We must bring forth fruit by abiding in Him …We live Him according to this view we have of Him. We live the all-inclusive, extensive Christ who is now the life-giving Spirit as the ultimate expression of the Triune God after many processes. And we meet together according to locality as the church, the church which is not only an assembly but also the Body, the new man, the lampstand, and the bride. We also practice the genuine oneness in every locality—one Body, one Spirit, one city, one church. Such a living is our ultimate responsibility*[213]*.*

21.04.2017

Wei-Hung in Kiel

Chinese version

「流淚灑種的，必歡呼收割」--詩篇一百二十六篇第五節

2013 年三月，在眼淚中揮別家人與故土，踏上了未知的旅程，在遙遠德意志的小城，吹著熟悉又陌生的海風，開始了新的生活。不意外的，在四年多的殷勤工作後，即將在基爾拿到我人生中最後一個學位。四年半的變化，一言難盡。從單身變已婚；從一句德文都不會說，到現在能與路人隨意地聊兩句；從對生物資訊一無所知，到現在能看懂生物資料處理的細節。我相信，過程中的每一滴眼淚，都是幫助我成長的養分。當然，一路走來並非單獨，我接受過許多人的扶持與鼓勵，沒有你們，就沒有今天的這本論文。

首先的感謝，給當初給我機會的 Philip Rosenstiel 教授。謝謝您的許多耐心，包容我這個生物領域的外行人，讓我在這邊的四年裡，能沒有壓力的做研究。您的專業想法還有處事態度，都是我人生的導師。

再來要感謝在這四年中指導我的許多 Postdoc: Ateeq, Rob, Felix and Maren. 在我們一起合作的計畫中，給了我很多寶貴的建議和指教。在論文的撰寫與計畫的執行上，沒有你們的幫忙，不可能有這麼順利的進展。

除此之外，要謝謝 Priya 和 Pankaj，一起在 RTG 三年多的時間，我們一起經歷許多有趣的事。也謝謝你們在論文撰寫跟找工作的規劃上給我許多正面的影響。

當然，IKMB 對我來說，不只是工作的地點，更像是彼此關心的群體。要謝謝同一個辦公室的每一位，在遇到挫折時一起的打氣與鼓勵，互相吐吐苦水。當工作疲累時，在紅沙發上聊一聊有趣的事，再繼續往前。Richa, Anupam, Phili, Jacqueline, Frauke, Anna, Stefii, Antonia, Alejandro, Go, Daniela…

還要謝謝在這幾年一起成長，教會的弟兄姊妹們，不論是 SKYPE 的交通禱告，或是聚會中的神聖供應。讓我在繁忙的工作中，不忘記當初神呼招我們來到此地的異象。

給我親愛的家人，謝謝爸爸媽媽，總是關心我，讓我在德國能夠無後顧之憂的打拼奮鬥。弟弟，姑姑們，給我許多的支持。最重要的是我親愛的老婆靖婷，陪伴我最後一年的博士生涯，有妳真好，我們在這裡一起努力，一起經過人生變動的階段，我們有主有召會。在德國的生活，還要繼續互相扶持，去門徒化萬民直到這世代的終結。

最真實的感謝，給我最愛的主耶穌。是神帶領我來到這裡，並且是他願意我們留下來。節錄一段話從「世界局勢與主的行動」書中出來 ：***(當一位基督徒) 我們的負擔是什麼？我們的責任是什麼？…我們必須背負耶穌的見證，…我們必須是祂的見證人，…我們必須借著住在祂裏面而結果子。…我們的責任乃是活基督，在我們日常的生活中為祂作活的見證。…我們照著對祂所有的看見來活祂。我們活這位包羅萬有、延展無限的基督，祂現今乃是賜生命的靈，作經過許多過程的三一神終極的彰顯。並且我們按著地方聚集一起成為召會，不僅是會集，也是基督的身體、新人、燈台和新婦。我們在各地實行真正的一一一個身體、一位靈、一個城市、一個召會。這樣的生活乃是我們終極的責任。***

21.04.2017

暐弘于基爾港