THE UNIVERSITY
of EDINBURGH

This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.
A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.
This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.
The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.
When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.

Epigenetic biomarker discovery in Inflammatory Bowel Disease: unearthing clues for disease pathogenesis?

Nicholas T Ventham

THE UNIVERSITY
of EDINBURGH

Thesis to be submitted for the degree of Doctor of Philosophy

University of Edinburgh

2016

Abstract

Epigenetic alterations including DNA methylation and microRNAs may provide important insights into gene-environment interaction in complex immune diseases such as inflammatory bowel disease (IBD). An integrative genome-wide approach was used to analyse whole blood genetic, DNA methylation and gene expression data in 240 newly diagnosed IBD patients and 190 controls. Using the Illumina 450k array, differences in whole blood DNA methylation were observed in IBD cases versus controls including 439 differentially methylated positions (DMPs) and 5 differentially methylated regions (DMRs). The top DMP (*RPS6KA2*, discovery Holm adjusted p=1.22×10^{-16}, replication p=1×10^{-9}) and DMRs (*VMP1, ITGB2, TXK*) were replicated in an independent cohort using pyrosequencing. Paired genetic and epigenetic data allowed the identification of methylation quantitative trait loci (meQTL); two of the five DMRs (*VMP1*, *ITGB2*) demonstrated significant association with genetic polymorphisms. Methylation in the *VMP1/microRNA-21* region was significantly associated with two single nucleotide polymorphisms (cg18942579 -rs10853015 [meQTL FDR adjusted p=9.4 × 10^{-5}], cg16936953 - rs8078424 [meQTL FDR adjusted p=8.8 × 10^{-5}]), both of which are in linkage disequilibrium with a known IBD susceptibility variant (rs1292053). Separated leukocyte methylation data highlight the cell type of origin of epigenetic signals seen in whole blood. IBD-associated hypermethylation within the *TXK* gene transcription start-site negatively correlated with gene expression in whole blood and CD8$^+$ T-cells, but not other cell types, highlighting that cell-specificity and gene location-specificity of DNA methylation change is critical when associating methylation and gene expression. These data offer significant translational potential as diagnostic biomarkers. Least absolute shrinkage and selection operator (lasso) modelling identified 30 methylation probes can be used to accurately discriminate IBD cases from controls (Area under receiver operating characteristic curve = 0.898, sensitivity = 90.6%, specificity = 84.7%).

MicroRNAs (miRNA) are small non-coding nucleic acids that have the capacity to modulate gene expression. MiRNAs have been increasingly implicated in many of the important IBD pathogenic pathways including autophagy, intestinal epithelial barrier integrity and the Th17 pathway. In common with all epigenetic mechanisms, miRNA expression is dynamic and cell-specific. Small RNA sequencing (RNA-seq) was performed on RNA extracted from CD14+, CD4+ and CD8+ cells isolated from 8 newly diagnosed cases of ileal or ileocolonic CD and 8 age and sex matched controls. There was a median of 2.4 million reads per sample (range 132,800-12.8 million reads per sample). One microRNA was differentially expressed in CD compared with controls (hsa-miR-503-5p log fold change = 0.7, FDR adjusted p = 9.1×10^{-5}) in CD4+ lymphocytes, however this finding did not remain significant when alternative normalisation methods were used. The small number of cases used in microRNA analyses raises the possibility of both type I and II error, and limits the ability to draw firm conclusion from this series of experiments.

Site-specific differences in DNA methylation in IBD relate to underlying genotype and associate with cell-specific alteration in gene expression. This is the most detailed characterisation of the epigenome carried out in IBD to date. The findings strongly validate this approach in complex disease, are replicable, and provide clear translational opportunities

Lay summary

Inflammatory bowel disease (IBD) is a chronic condition that affects young people. Typical symptoms include bloody diarrhoea, abdominal pain, weight loss and fatigue. The condition can be controlled (but not cured) by drugs that target the immune system, most of which have significant side effects. Despite an ever increasing armamentarium of drugs, a large proportion of patients still require surgery. Uncontrolled chronic inflammation in IBD predisposes patients to bowel cancer.

Great progress has been made in identifying genes that are associated with developing the condition. However, advances in genetics have not provided all of the answers as once hoped. This project aims to look beyond the genetics, to study other ways that way in which genes can be switched on or off without a change in the underlying genetic code (Epigenetics). This project will mainly focus on DNA methylation; specific marks found on the DNA that can affect how easily the gene can be read. Another important facet of epigenetics looks at microRNA, small strands of RNA that can change the way genes are expressed.

In this project a large group of newly diagnosed patients (n=240) and controls (n=194) has undergone analysis across the genome identifying a number of genetic marks (DNA methylation) that are more or less common in IBD patients compared to healthy people. These data have been integrated with matched data on the underlying code (genetics) and whether genes are switched on or off (expression).

The overarching aim of this work is to identify a new blood test (biomarker) which can firstly help diagnose IBD, but importantly help identify which patients will have a severe disease course and require powerful drugs or surgery.

Declaration of Originality

I declare that all of the work presented in this thesis is the result of my own investigations, and that as part of a larger research group/collaborative, I have properly acknowledged the contributions of others in the section below as well in the appropriate part(s) of the thesis itself. The inclusion of text and/or results that have been published in peer-reviewed journals will be indicated. Work was conducted in the Gastrointestinal Unit (Level 2, Molecular Medicine Centre, Centre for Genomics and Molecular Medicine, University of Edinburgh) between January 2013 and June 2016. This work has not been submitted for professional degree or qualification elsewhere.

Patient recruitment and experimental work was conducted by Nicholas Ventham with the exception of:

- Patient recruitment (Edinburgh) – The vast majority of patients included in this thesis were recruited by myself. Additional patients were recruited by Dr Rahul Kalla, Dr Nick Kennedy, Dr Ray Boyapati and Ms Linda Smith. Retrospective stored samples used in this study have been recruited by Drs Anne Phillips, Dr Charlie Lees, Dr Richard Russell, and Dr Colin Noble.
- Sample processing and nucleic acid extraction was primarily performed by myself, with additional help from Dr Rahul Kalla, Ms Kate O'Leary, Dr Ray Boyapati and Dr Alex Adams.
- Flow Cytometry – Initial settings and laser frequencies on the BD FACS Aria II (Human Genetics Unit, CGEM) were programmed by Ms Elisabeth Freyer. I performed flow cytometric analysis on the majority of samples for flow cytometry myself with the remainder of samples being run by Ms Freyer.
- Agilent bioanalyzer nucleic acid quality assessment was performed by myself for microRNA experiments, but samples were also analysed by Angie Fawkes (WTCR, Western General Hospital) and Agnes Gallagher (Human Genetics Unit, CGEM).
- Microarrays were performed at the WTCRF by Louise Everden, Jude Gibson, and Tamara Gilchrist

- Pyrosequencing was performed at the WTCRF by myself. Additional pyrosequencing runs were performed by Mr William Hawkins and Dr Alex Adams.
- Small RNA sequencing experiments were performed by Edinburgh Genomics.
- Statistical analysis was performed by myself, however significant assistance was sought from Dr Nick Kennedy and some assistance from Dr Alex Adams.

Signed ……………………………………………………………………………………………………

Nicholas Ventham

Date …………………………………………………………………………………………………….

Dedication

This thesis is dedicated to my mother, Jennifer Ventham.

Acknowledgements

Supervisors – Professor Jack Satsangi & Dr Elaine Nimmo

Co-workers - Dr Nick Kennedy, Dr Rahul Kalla, Dr Alex Adams, Ms Nidhi Sharma, Ms Kate O'Leary, Ms Hazel Drummond, Ms Colette McColl, Dr Helen Newbury, Ms Stephanie Scott, Prof David Wilson, Dr Ray Boyapati, Dr Victoria Merrick

School of population studies - Dr Evropi Theodoratou

WGH – Drs Arnott, Noble, Shand, Lees, Ho, Shahida Din

WTCRF – Lee Murphy, Louise Everden, Jude Gibson, Angie Fawkes, Tamara Gilchrist, William Hawkins

King's buildings – Dr Amy Buck, Mr Juan Quintana

HGU – Ms Elisabeth Freyer, Ms Agnes Gallagher

Ludger – Dr Daniel Spencer, Dr Richard Gardner, Mrs Archana Shubhakar, Dr Daryl Fernandes

CNAG – Ivo Gut, Marta Gut, Matthew Ingham, Simon Heath

Thesis Committee – Profs Malcolm Dunlop and Harry Campbell

List of publications arising from this thesis

Prizes

Y-ECCO conference abstract award (European Crohn's & Colitis Organisation) 2016

John B Scrimgeour Medal (Western General Hospital) 2015

National Scholar Award for the UK (United European Gastroenterologists) 2015

Basic Science Travel award (United European Gastroenterologists) 2015

Oral Presentation of Distinction (Association of Surgeons of Great Britain and Ireland) 2015

Poster of Distinction (Digestive Diseases Week) 2015

Poster of Distinction (British Society of Gastroenterology) 2014

Ann Ferguson Prize (Scottish Society of Gastroenterology) 2013

Publications

List of publications arising directly from this thesis

1. Trbojević Akmačić I, **Ventham NT**, Theodoratou E, Vučković F, Kennedy NA, Krištić J, Nimmo ER, Kalla R, Drummond H, Štambuk J, Dunlop MG, Novokmet M, Aulchenko Y, Gornik O, Campbell H, Pučić Baković M, Satsangi J, Lauc G; IBD-BIOM Consortium Inflammatory Bowel Disease Associates with Proinflammatory Potential of the Immunoglobulin G Glycome. **Inflamm Bowel Dis** 2015 ; 21 (6): 1237-47

2. **Ventham NT.** Gardner RA. Kennedy NA, Shubhakar A, Kalla R, Nimmo ER; IBD-BIOM Consortium, Fernandes DL, Satsangi J, Spencer DI. Changes to serum sample tube and processing methodology does not cause inter-individual variation in automated whole serum N-glycan profiling in health and disease. **PLoS One**. 2015 Apr 1;10(4):e0123028.

3. **Ventham NT**. Kalla R. Kennedy NA. Satsangi J. Arnott IDR. Predicting outcomes in acute severe ulcerative colitis. **Expert Rev Gastro Hep**. 2014: 9(4):405-15

4. Kalla R. **Ventham NT**. Kennedy NA. Quintana JF. Nimmo ER. Buck Ah. Satsangi J. MicroRNAs: new players in IBD. **Gut** 2014 64(3):504-17

5. Kalla R. **Ventham NT**. Satsangi J. Arnott IDR. Crohn's disease. **BMJ** 2014 349:g6670

6. **Ventham NT**. Kennedy NA. Duffy A. Clark DN. Crowe AM. Knight AD. Nicholls RJ. Satsangi J. Nationwide linkage analysis in Scotland-Has mortality following hospital admission for Crohn's disease changed in the early 21st century? **J Crohns Colitis**. 2014. pii: S1873-9946(14)00267-0.

7. **Ventham NT**. Kennedy NA. Duffy A. Clark DN. Crowe AM. Knight AD. Nicholls RJ. Satsangi J. Comparison of mortality following hospitalisation for Ulcerative colitis in Scotland between 1998-2000 and 2007-09**. Aliment Pharmacol Ther** 2014; 39(12):1387-97

8. **Ventham NT.** Kennedy NA. Nimmo ER. Satsangi J. Beyond gene discovery in Inflammatory Bowel Disease: The emerging role of Epigenetics. **Gastroenterology**. 2013: 145 (2); 293-308

List of publications arising indirectly or from patient recruitment as part of this thesis

1. Liu JZ et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. **Nat Genet**. 2015 Sep;47(9):979-86. doi: 10.1038/ng.3359. Epub 2015 Jul 20.

2. Satsangi J, Kitten O, Chavez M, Kalla R, Prel N, Meuwis MA, Scott S, Bonetti I, **Ventham NT**, Louis E. How to apply for a secure EU funding for Collaborative IBD Research projects. **J Crohns Colitis**. 2016. 10(3); 363-70

3. Adams AT. Kennedy NA. Hansen R. **Ventham NT**. O'Leary K. Drummond H. Noble CL. El-Omar E. Russell R. Wilson D. Nimmo ER. Satsangi J. Two-Stage genome wide methylation profiling in childhood onset Crohn's implicates epigenetic alterations at the VMP1/MIR21 and HLA loci. **Inflammatory bowel diseases** 2014.

4. Theodoratou E. Campbell H, **Ventham NT**, Kolarich D, Pucic-Bakovic M, Zoldos V, Fernandes D, Pemberton IK, Rudan I, Kennedy NA, Wuhrer M, Nimmo E, Annese V, McGovern DP, Satsangi J, Lauc G. The role of glycosylation in IBD**. Nature Reviews Gastroenterology and Hepatology** 2014 2014 Oct;11(10):588-600

Abstracts

Oral presentations

- Comprehensive epigenome-wide DNA methylation profiling in IBD DDW San Diego 22/5/16 (Oral plenary presentation, unable to attend, presented by Prof Jack Satsangi)
- Comprehensive epigenome-wide DNA methylation profiling in IBD. ECCO Amsterdam 17/3/16 (Oral plenary presentation)
- Epigenome-wide DNA methylation profiling in IBD. UEG week Barcelona 26/10/15
- Epigenome-wide DNA methylation profiling in IBD. DDF London Excel 23/6/15
- Epigenome-wide DNA methylation profiling in IBD. ASGBI Manchester 24/4/15
- Epigenome-wide DNA methylation profiling in IBD. Royal Society of Medicine London 25/3/15
- Comparison of mortality following hospitalisation for ulcerative colitis in Scotland between 1998-2000 and 2007-2009. SSG Glasgow 6/12/13

Poster presentations

- IBD associates with proinflammatory potential of the IgG glycome. DDW Washington 2015
- IBD associates with proinflammatory potential of the IgG glycome. ECCO Barcelona 2015
- Comparison of mortality following hospitalisation for ulcerative colitis in Scotland between 1998-2000 and 2007-2009. ECCO Copenhagen 2014
- Comparison of mortality following hospitalisation for ulcerative colitis in Scotland between 1998-2000 and 2007-2009. BSG Manchester 2014
- Comparison of mortality following hospitalisation for Crohn's disease in Scotland between 1998-2000 and 2007-2009.BSG Manchester 2014

Abbreviations

ASUC = acute severe ulcerative colitis

CD = Crohn’s disease

CDXX = Cluster differentiation (e.g. CD8$^+$)

CpG = Cytosine-Guanine dinucleotide

CRP = C-reactive protein

DMP = differentially methylated position

DMR = differentially methylated region

ESR = Erythrocyte sedimentation rate

EWAS = epigenome wide association study

FC = Faecal calprotectin

GWAS = Genome wide association study

HLA = human leukocyte antigen

IBD = Inflammatory bowel disease

IBD-U = inflammatory bowel disease unclassified

ITGB2 = integrin beta-2 subunit

MAF = minor allele frequency

MHC = major histocompatibility complex

miR = microRNA

MRI = magnetic resonance imaging

NOD2= nucleotide oligomerization domain-containing protein 2

PBMCs = Peripheral blood mononuclear cells

QTL = quantitative trait loci

RPS6KA2 = ribosomal S6 kinase 2

SBNO2 = strawberry notch homolog 2

SNP = single nucleotide polymorphism

TXK = tyrosine kinase like tec receptor

UC = Ulcerative colitis

VMP1 = vacuole membrane protein 1

Contents

Figure Index

Table Index

Chapter 1. Introduction

This chapter includes text from the following publications for which I am a primary author:

1. Ventham NT. Kennedy NA. Nimmo ER. Satsangi J. Beyond gene discovery in Inflammatory Bowel Disease: The emerging role of Epigenetics. Gastroenterology. 2013: 145 (2); 293-308 (License Number: 3742451069879)
2. Ventham NT. Kennedy NA. Duffy A. Clark DN. Crowe AM. Knight AD. Nicholls RJ. Satsangi J. Comparison of mortality following hospitalisation for Ulcerative colitis in Scotland between 1998-2000 and 2007-09. Aliment Pharmacol Ther 2014; 39(12):1387-97 (Permission obtained from John Wiley and Sons License number 3741880854659)
3. Kalla R. Ventham NT. Satsangi J. Arnott IDR. Crohn's disease. BMJ 2014 349:g6670 (permission to use within thesis included in author copyright agreement with the BMJ)
4. Ventham NT. Kalla R. Kennedy NA. Satsangi J. Arnott IDR. Predicting outcomes in acute severe ulcerative colitis. Expert Rev Gastro Hep. 2014: 9(4):405-15 (Taylor & Francis is pleased to offer reuses of its content for a thesis or dissertation free of charge contingent on resubmission of permission request if work is published)

1.1 Incidence and prevalence studies

The inflammatory bowel diseases (IBDs) Crohn's disease (CD) and ulcerative colitis (UC) are an important health problem, with an incidence among European adults of 12.7 and 24.3 per 100,000 person-years respectively, and prevalence of 0.5%–1.0%.[1] Moreover, IBD incidence is increasing among adults and children and in the developed and developing world.[1–5] In the UK, IBDs cost the National Health Service approximately £720 million (about $1.1 billion) per annum, based on an average cost of £3,000 per patient per year with around half of the costs attributable to relapsing patients [6,7]

1.2 Mortality following IBD

The extent of disease-specific mortality in IBD remains contentious. Most population-based studies demonstrate a similar or slightly increased mortality in patients with CD and UC compared with the general population.[8] Subsequently other linkage analyses, focusing on patients requiring hospitalisation, have given cause for re-appraisal.[9] In particular, real concern was generated by record linkage studies from England, suggesting markedly increased 3-year mortality rates following admission for IBD, especially for in-patients treated medically.[10] Strikingly similar results were demonstrated in hospitalised patients in Scotland during the same time period.[11,12]

Using data from the ISD Scottish Morbidity Records database, the 3-year mortality in patients hospitalized for IBD between 1998-2000 and 2007-2009 has been compared.[13,14] The linked Scottish Morbidity Records database was used to identify patients admitted with CD and UC during two periods: Period 1(1998-2000) and Period 2 (2007-2009). Directly age-standardized mortality rates demonstrated a decreased mortality for UC (Period 1:373/10,000 person years [CI 309-437], Period 2: 264/10,000 person years [CI 212-316], p<0.0001) but were similar for CD (Period 1:338/10,000 person years [CI 282-394]; Period 2:333/10,000 person years [CI 276-390], p=0.2, Figure 1). A lower crude mortality was noted in patients following elective surgical admission when compared with emergency surgery and medical admission types (Figure 2). For both CD and UC, multivariable regression analysis demonstrated age and comorbidity to be significantly associated with mortality. These data demonstrates that overall 3-year mortality after hospitalization for IBD is high. While reductions in 3-year mortality have been observed patients admitted with UC, mortality was unchanged in CD.

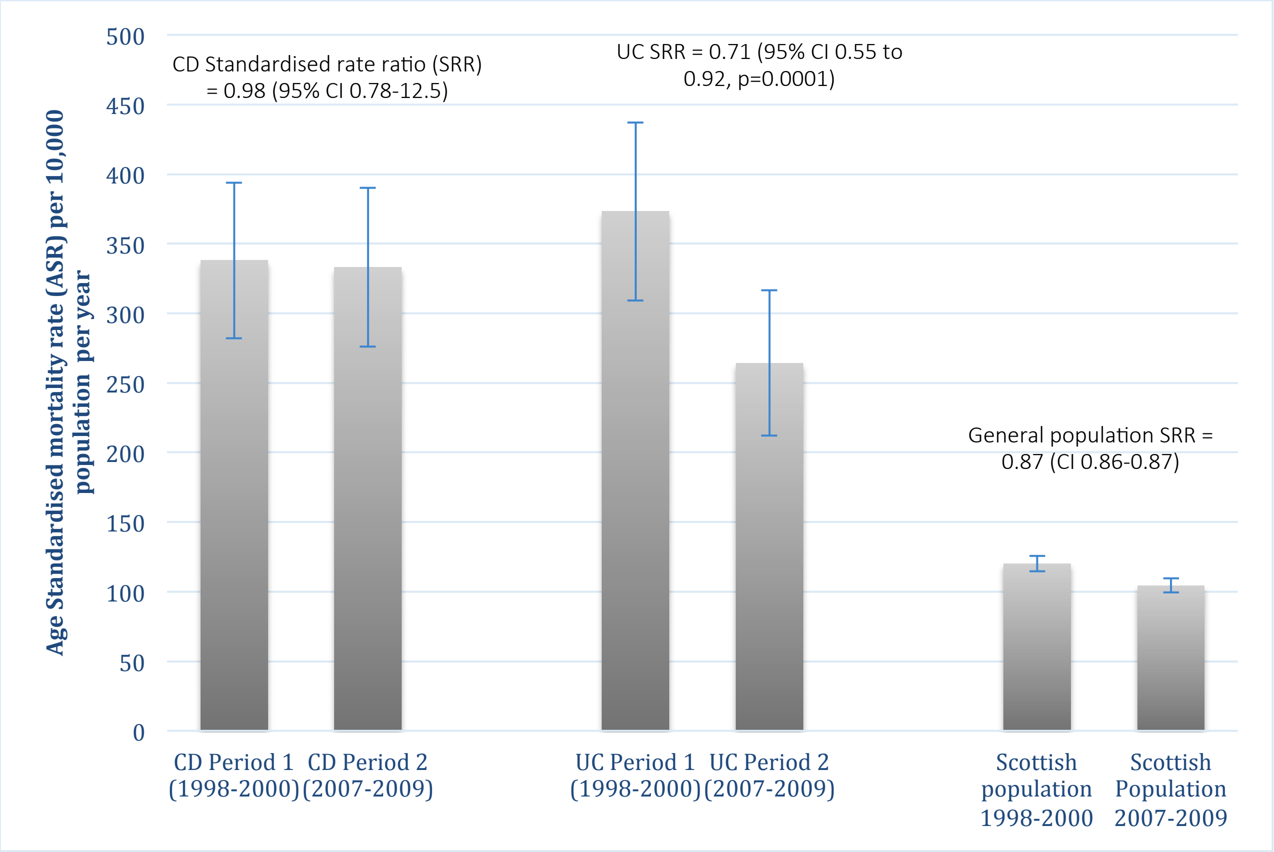


Figure 1 – Mortality following hospital admission for IBD. Age Standardised mortality rates (per 10,000 population per year) following first admission with ulcerative colitis or Crohn's disease between two periods: Period 1 (1998-2000) and Period 2 (2007-2009). *Standardised to 2003 general Scottish population. SRR = standardised rate ratio, CI = Confidence interval. Adapted from Ventham APT 2014[14]

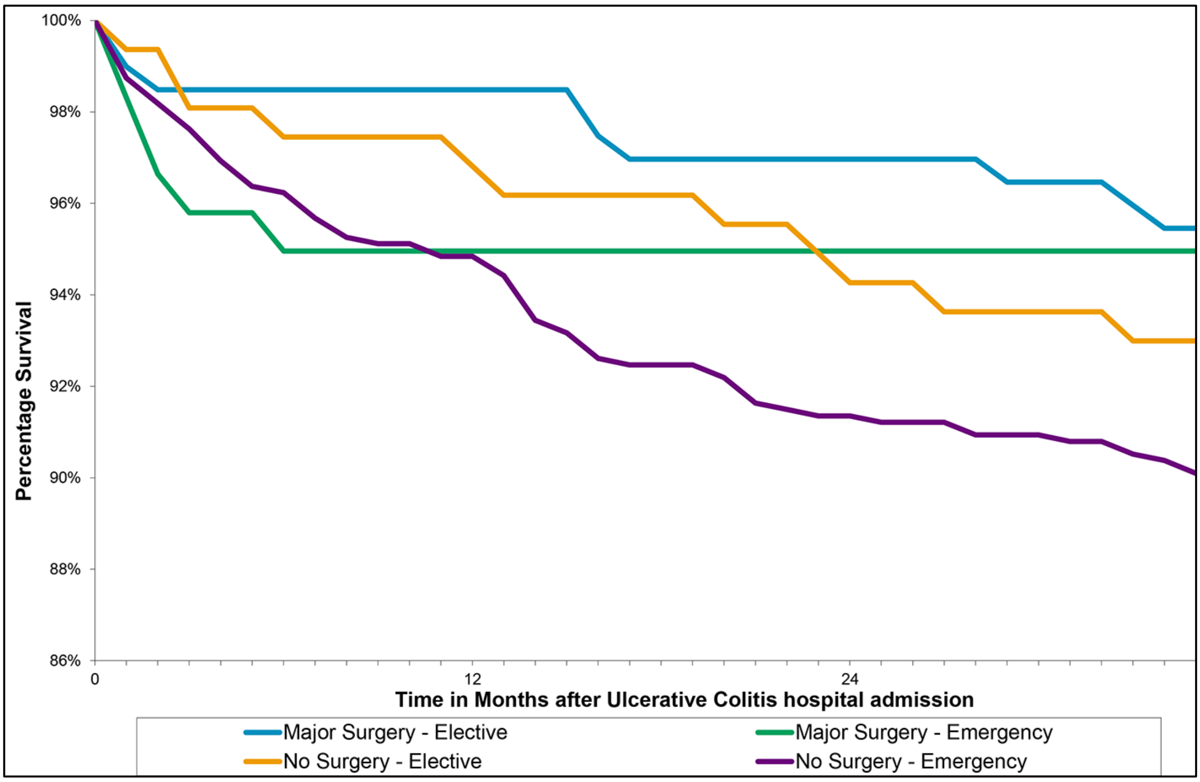


Figure 2 – Survival following admission for ulcerative colitis according to admission type. Unadjusted 3 year cumulative survival following admission for ulcerative colitis in Scotland between 2007-2009 according to admission type. Adapted from Ventham APT 2014[14]

1.3 Pathology and presentation

Diagnosing IBD can be challenging as a result of its widespread and often cryptic manifestations.[15] Distinguishing IBD from irritable bowel syndrome can be difficult, and there can be a significant delay before diagnosis of IBD.[16] The clinical features of IBD are related to the location of the bowel affected (Table 1) and include chronic diarrhoea (>4 weeks +/- blood and mucus), nocturnal defecation, abdominal pain and weight loss.[15,17] IBD may present with non-specific symptoms including malaise, fever and anorexia. Patients may also present initially with extra-intestinal manifestations (Figure 3) of IBD that include arthritis, aphthous ulcers, skin (pyoderma gangrenoum, erythema nodosum), liver (primary sclerosis cholangitis/primary biliary cirrhosis), or ocular manifestations (iritis, episcleritis). The natural history of IBD is one of relapse and remission. Severe bloody diarrhoea occurring over 6 times a day with systemic signs of toxicity (tachycardia, fever) may signify acute severe colitis warranting urgent admission.[18] Acute severe ulcerative colitis (ASUC) effects approximately a quarter of patients with UC.[19]

Crohn's disease				**Ulcerative colitis**			
Montreal location (* ***Paris modifier*****)**	Symptoms experienced	**CD – Montreal Behaviour (*** ***Paris modifier*****)**	Symptoms experienced	**Montreal Extent (*** ***Paris modifier)***	Symptoms experienced	**Montreal Severity(*** ***Paris modifier)***	Symptoms experienced
L1 – Ileal	abdominal pain and weight loss; diarrhoea may be absent; Malabsorption and nutritional deficiencies; acute ileitis disease can mimic acute appendicitis	**B1 – inflammation**	abdominal pain and weight loss; diarrhoea; nutritional deficiencies	**E1 – proctitis**	Diarrhoea +/- blood, tenesmus, urgency, incontinence	**SO: Remission (*Never severe)**	
L2 – Colonic	Bloody diarrhoea; can mimic acute severe ulcerative colitis; obstruction due to stricturing disease	**B2 – Stricturing**	Obstructive symptoms: distension, abdominal pain, vomiting, weight loss, reduced stool frequency, frank bowel obstruction	**E2 – Left sided disease**	As above	**S1: mild UC (* Ever Severe)**	≤ 4 stools/day
L3 – Ileocolonic	Right sided abdominal pain, diarrhoea, weight loss	**B3 - Penetrating**	Perforation & Intra-abdominal Abscess – Sepsis, abdominal pain, fever, peritonitis, symptoms arising from abdominal fistulae – enterovesical: recurrent UTIs/pneumatouria, enterovaginal: discharge, enteroenteral: bacterial overgrowth enterocutenous : discharge	**E3 – Extensive disease *(*proximal to hepatic flexure)***	As above Acute severe ulcerative colitis, abdominal pain, distension	**S2: moderate UC**	> 4 stools/ day, no systemic toxicity
L4 – Upper Gastrointestinal	Can mimic peptic ulcer disease; can present as chronic gastric outlet obstruction, mouth ulceration, oesophagitis	***B2B3 – both stricturing and penetrating, either at the same or different times ****		***E4 – pan colitis***	As above	**S3: Severe**	> 6 stools/ day plus systemic toxicity (pulse >90 bpm, temp > 37.5 ° C, ESR > 30, Hb 105 g/L
L4a*	Proximal to ligament of Treitz						
L4b *	Distal to ligament of Treitz to distal 1/3 ileum						
P - Perianal	Recurrent perianal abscesses & perianal fistulae						

Table 1 - The Montreal[20] and Paris[21](*) classifications of Crohn's disease and Ulcerative colitis respectively.

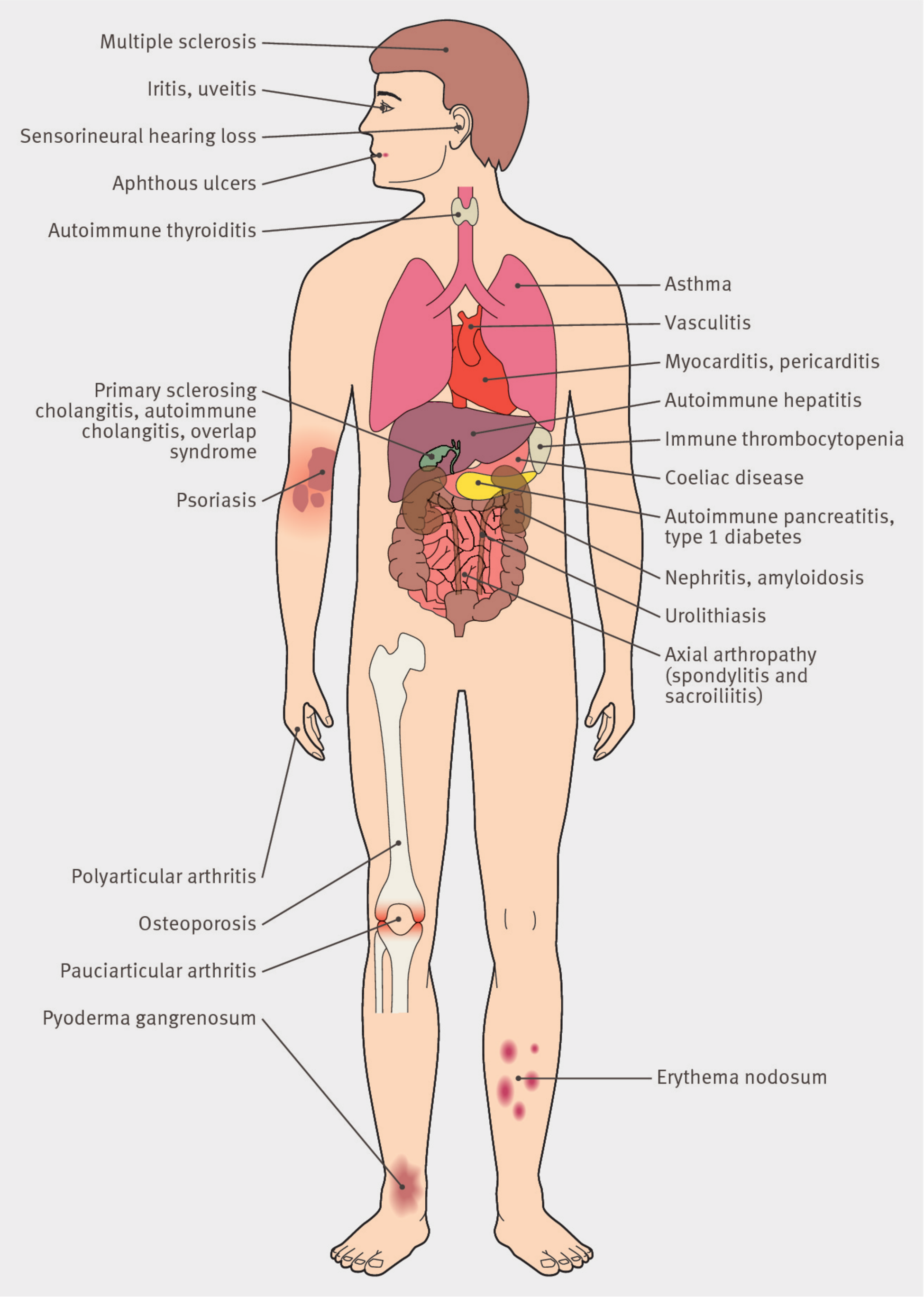


Figure 3 - Extra-intestinal manifestations of IBD. Taken from Kalla et al[15] with permission

IBD is usually diagnosed using a combination of clinical, laboratory, radiological, endoscopic and histological investigations. Biochemistry may reveal raised inflammatory markers (CRP/ESR), iron deficiency anaemia and nutritional deficiencies such as low B12 and folate. Stool cultures for clostridium difficile, ova and parasites should be performed in all patients with chronic diarrhoea.

Blood parameters may be normal in IBD and faecal calprotectin (FC), a neutrophil cytosolic protein, can determine the presence or absence of intestinal inflammation.

Ileo-colonoscopy and biopsies remains the cornerstone in diagnosing IBD and can help distinguish between CD and UC. CD can affect from the mouth to anus, and is associated with discontinuous colonic inflammation and ulceration, 'cobblestone' appearance and rectal sparing with histology demonstrating focal/ patchy chronic inflammation, focal crypt irregularity and granulomas.[22] UC is associated with continuous inflammation extending proximally from the rectum, but confined to the colon (with the exception of backwash ileitis). The histology in UC reveals ulceration and inflammation confined to the submucosa; a plasma-lymphocytoid cell infiltrate in the lamina propria, and crypt architectural distortion and crypt abscesses without granuolmata. Ileal biopsies can be useful in differentiating between CD and UC, but in ~5% of cases it is not possible to differentiate CD and UC; here the terms IBD-unclassified (IBD-U) or indeterminate colitis are used.[23,24] Whilst obtaining a tissue diagnosis is critical, this can be challenging when CD affects the small bowel, and small bowel magnetic resonance imaging (MRI) is becoming the imaging modality of choice. Other investigation modalities include CT for any extra-luminal complications of CD, small bowel ultrasound in specialist centres, small bowel capsule endoscopy and small bowel enteroscopy in whom the clinical suspicion for CD remains high despite negative first-line investigations.[25]

1.4 Management of IBD

1.4.1 Medical

Before considering medical therapy, patients diagnosed with IBD should undergo a thorough nutritional assessment and replacement of any deficient micronutrients including vitamin B12, folate, iron, calcium and vitamin D.

For CD, stopping smoking can be as effective as immunomodulatory therapy, and reduces the risk of relapse by 65% compared to continuing smokers and patients should be offered the full remit of smoking cessation services.[26–29] Non-steroidal anti-inflammatory drugs should be discontinued.[30–32] The choice of drug therapy is influenced by factors such as efficacy, the need for inducing or maintaining remission, side-effect profile, long-term risks and patient choice. Patients with a severe phenotype should have early, arguably combined, immunosuppressive therapy.[33]

1.4.2 Treatment of disease flare (Induction of remission)

Enteral nutrition

In adults, EEN can be poorly tolerated but is effective as first line therapy in CD. Guidelines recommend EEN to improve nutritional status or as first line therapy in those individuals who decline conventional drug therapies.[34] EEN is less effective than steroids in inducing remission in adults(odds ratio(OR): 0.33, 95% confidence interval (CI) 0.21-0.53).[35] EEN can be useful prior to starting immunosuppressive therapy whilst confirming a diagnosis of CD.

Corticosteroids

It is established practice to use corticosteroids to induce remission in CD and UC. A meta-analysis has demonstrated that glucocorticoids are effective in inducing remission in UC (relative risk (RR) = 0.65, CI 0.45-0.93, number needed to treat (NNT)=3) and probably also in CD (RR = 0.46, CI 0.17-1.28).[36] As a result of the short- and long-term side effects, corticosteroids should not be used to maintain remission.[37,38] The action of Budesonide is confined to the gut as a result of extensive first pass metabolism and consequently has fewer systemic side effects. Budesonide is indicated in patients with mild to moderate CD confined to the terminal ileum or the proximal colon, but is ineffective in maintaining remission.[39]

5-aminosalycilates (5-ASA)

5-aminosalycilates (5ASA) can be used to induce remission in mild and moderate UC, but are not now recommended in CD. Topical 5ASA (enemas, suppositories) can be used for distal disease and proctitis in UC,[40] whilst oral preparations (and combined oral and rectal therapy[41], NNT=5) are suitable for more extensive disease (NNT=6).[42]

Ciclosporin

Ciclosprorin can be used for the induction of remission in moderate to severely active UC.[43,44] Ciclosporin has been used as medical rescue therapy at centres as an alternative to Anti-TNFα therapies. The CONSTRUCT trial is due to report soon, and initial results demonstrate no difference between ciclosporin and infliximab for rescue therapy in ASUC.

Biological therapies

Anti-TNFα monoclonal antibodies (Infliximab, Adalimumab and Golilumab) are effective at inducing remission in moderate to severe CD, perianal Crohn's and have recently been approved (NICE ta329) for use in UC.[23,45] Early use of anti-TNFα agents (top-down approach) is associated with increased remission rates after 3 years of therapy.[46–48] NICE guidelines recommend the step-up approach: using anti-TNF agents only for patients who have failed conventional immunomodulatory therapies. Current practice is to undertake a rapid step-up approach for those with a severe disease phenotype.[49]

Although combination therapy with anti-TNFα agents and immunomodulators carries risks of non-melanoma skin cancer and other cancers compared to monotherapy (standardised incidence ratio (SIR), 3.46 [CI 1.08-11.06] and SIR 2.82 [CI 1.07-7.44] respectively), combination therapy is superior to monotherapy in maintaining steroid free clinical remission (56.8% vs 30% $p<0.001$) with evidence of better mucosal healing (43.9% vs 16.5% $p<0.001$).[50,51] The optimal time for anti-TNFα withdrawal is currently unknown but an expert panel review identified low risk groups where timed withdrawal may be considered.[52,53] Golimumab is a fully humanised anti-TNFα monoclonal antibody.[54] The PERSUIT study confirmed its efficacy for induction and maintenance of remission in moderate to severe UC.[55] Golimumab is currently approved in England and Wales for UC, but has not yet been approved the Scottish Medicines Consortium (SMC). The GoColitis trial (NTC02092285) is currently evaluating the use of Golimumab.

1.4.3 Maintenance of remission

Immunomodulators

Once in remission, maintenance therapy should be commenced, in order to avoid repeated steroid courses and to attenuate disease progression. Symptoms alone are a poor guide to the attainment of remission and clinical, biochemical (including FC) and endoscopic factors should be used to determine 'complete' remission and guide further treatment decisions.

Immunomodulatory drugs are effective at maintaining remission in moderate to severe CD and in those who are steroid dependent. The OR for maintenance of remission in CD with thiopurines, azathioprine (AZA) and mercaptopurine is 2.32 (CI 1.55 -3.49, NNT 6) and 3.32 (CI 1.40 – 7.87, NNT 4) respectively.[56] Azathioprine has an established role in the maintenance of remission in UC (HR 0.60, 0.37 to 0.95), and the NNT to prevent one relapse was 4.[57]

The onset of action of the thiopurines is slow (up to 17 weeks) and induction therapies (corticosteroids, anti-TNFα agents) are often needed for 'bridging'.[34] Methotrexate (MTX) is also effective at maintaining remission in CD (vs. placebo, 65% vs. 39% , NNT=4)[58], however it is teratogenic, often poorly tolerated and guidelines recommend their use only in patients intolerant or refractory to thiopurines or anti-TNF agents.[23,59] The optimal time for drug withdrawal has been debated, however expert opinion suggests drug discontinuation 4 years after remission.[53] Such decisions are often made on an individual basis taking into account the risk of relapse against the long term risks of therapy.[60]

5-aminosalycilates (5-ASA)

5-aminosalycilates (5ASA) are also used for maintenance of remission in UC. Oral 5ASA has a good NNT of 4 to prevent one relapse.[42] Topical 5-ASA can be used to prevent relapse in distal disease.[61]

1.4.4 Surgical

In the modern era of biologic therapies, the requirement for surgery is falling, but up to a quarter of patients still require surgery within 5 years of diagnosis.[62,63] Indications for resectional surgery often result from medical therapy failure and include treatment of fibrostenotic disease, penetrating disease (perforation, intra-abdominal abscess, abdominal fistulae) or evidence of colonic dysplasia at surveillance. Perianal CD may require surgery either to drain sepsis or to control fistulae. Ileoceacal resection is considered a first line treatment for discrete terminal ileal disease.[64–66] Anastomotic recurrence remains common. A wide stapled side-to-side anastomosis is slightly superior to hand sewn end-to-end anastomoses in preventing disease recurrence(OR 0.2, CI 0.07-0.55).[67] The role of medical therapy to prevent post-operative recurrence is currently being investigated (TOPPIC trial). The main principle of surgery in CD is to preserve bowel length in order to avoid short bowel syndrome and intestinal failure. Operations including stricturoplasty effectively treat strictures without the need for resection. Ileorectal anastomosis (IRA) is infrequently indicated due to the high risk of disease recurrence in proximal small bowel and the risk of anastomotic leaks.[68]

In UC, total colectomy provides a curative treatment and protects against the elevated risk of colorectal cancer in IBD patients. In an inception cohort, in the first 10 years following diagnosis the colectomy rate was 9.8%.[69] In the setting of acute severe colitis the higher colectomy rate of ~ 29% has been unchanged for many years.[70] In this setting of ASUC, subtotal colectomy is the procedure of choice, as an elevated mortality was demonstrated following total colectomy. Early results from the CONSTRUCT trial indicated no significant difference in quality of life scores in patients undergoing colectomy compared to those treated successfully medically. Bowel continuity can be restored in patients with UC following colectomy with ileal pouch to anal anastomosis (IPAA).

1.4.5 New Biologic treatments on the horizon for IBD

1.4.5.1 Biosimilars

Biosimilars are biologically similar medicinal products that share similar properties, in terms of efficacy and safety, to previously approved/licenced drugs. Of note, Infliximab has undergone 20-30 minor changes since it was licenced, and was rigorously assessed for pre- and post-modification comparability.[71] Monoclonal antibody proteins can be subjected to one of several post-translation modifications including glycosylation, phosphorylation, and ubinqutinisation.[71]

There are randomised data from other diseases and oberservational data in IBD supporting the use of CT-P13 (Remisa) as a biosimilar for Infliximab.[72,73]

1.4.5.2 Integrin/MAdCAM targeting monoclonal antibodies

Vedolizumab is a monoclonal antibody targeting α4β7 integrin that is variably expressed on the cell surface of some T-lymphocytes and B-cells. The drug acts by limiting trafficking and migration of leukocytes to inflamed sites within the gut. The great promise of the anti-mucosal addressin cell adhesion molecule (MAdCAM) antibodies over other biologics is the specificity of vedolizumab in targeting leukocyte recruitment within the gut, thereby limiting unwanted systemic immunosupression.[74] Natalizumab, which targets the α4 subunit of integrins α4β7 and α4β1, is efficacious in multiple sclerosis and CD, but is associated an unacceptable risk of PML (progressive multifocal leukoenchepalopathy)[75] and may be related to the off-target effects of non-specific inhibition of leukocyte traffic in other organs including the CNS. Vedolizumab is gut specific and there have been no reported cases of PML.

Two large studies have demonstrated Vedolizumab to be more effective than placebo for inducing and maintaining remission in UC (GEMINI-1)[76] and CD.[77] Vedolizumab has also been used in patients who have failed anti-TNF treatment, with benefit seen at 10 but not six weeks.[78] Ertrolizumab is another humanised monoclonal antibody directed at the β7 subunit has also been demonstrated to be effective in induction of remission in UC.[79]

1.4.5.3 Interleukin -12 and -23 pathway targeting monoclonal antibodies

The interleukin (IL)-12 and IL-23 pathways have been implicated in the pathogenesis of immune-mediated diseases including CD. Ustekinumab is an IgG monoclonal antibody targeting the shared p40 subunit of IL-12 and 23 prevents interaction with receptors. Ustekinumab was first used for the effective treatment for psoriasis.[80] Ustekinumab was beneficial in mild-moderate CD at 4 and 8 weeks.[81] CERTIFI trial has demonstrated benefit to those previously refractory to anti-TNF drugs.[82] The PSOLAR biologic registry in psoriasis of over 12000 patients found no evidence of increased risk of infection or malignancy in patients treated with ustekinumab.[83]

Other IL-17 and IL-23 targeting drugs (Briakunumab, Sucukinumab, Ixekizumab, Broadalumab) are under investigation.

1.4.5.4 Other promising therapeutics in the pipeline

Mongersen is an anti-sense oligonucleotide of SMAD7 ("Mothers against decapentraplegic") that has been developed for the treatment of CD. SMAD7 is an intracellular protein upregulated in CD that binds to TGF- β1, an immunosuppressive cytokine, to and prevents TGF signalling. [84] Orally administered Mongersen (pH modified release in the right colon) has been shown to effectively induce remission in a phase II trial.[84]

1.5 The Genetic architecture of IBD

IBD pathogenesis is believed to involve an aberrant immune response to intestinal microbiota in genetically susceptible individuals.[85] Genetic studies have provided many candidate loci in the last decade, and the innate and acquired immune responses have been implicated in pathogenesis. However, identified genetic factors account for only a modest proportion of the disease variance—13.1% in CD and 8.2% in UC.[86] These figures highlight the need for critical evaluation of genetic discoveries to date, and indicate the importance of the environmental factors in IBD pathogenesis; in addition the intriguing possibility arises that epigenetics may partially account for the 'hidden heritability' in IBD.

In the last 25 years there has been intense interest in identifying genetic, and more recently, epigenetic changes that relate to the pathogenesis of IBD. Few other complex diseases have been subjects of such extensive genetic and epigenetic research. National consortia and subsequently large international collaborative research groups, such as the International IBD genetics consortium (IIBDGC), have led the way in performing large-scale appraisals of the genome of patients with IBD (http://www.ibdgenetics.org/). The assumption-free approach of genome-wide association studies (GWASs) has helped to support established etiological roles of the innate and acquired immune system in IBD, and also identified interesting new mechanisms, such as autophagy.[87]

Findings from the last 25 years of genetic discovery in IBD have been put into context by a meta-analysis of the GWAS and Immunochip data.[88] The Immunochip was developed following GWASs of IBD and other immune disease; it contains 200,000 single nucleotide polymorphisms (SNPs) relevant to IBD and other immune-mediated diseases. The aims of the Immunochip experiments are to replicate and fine map the known IBD susceptibility loci, and to identify common links with other immune disorders. The meta-analysis comprised more than 75,000 cases and controls, and more than 1.23 million SNPs from several centres worldwide. It identified a further 64 loci, bringing the total number of IBD-associated loci to 163—significantly more than any other complex disease.[88] More recently, a trans-ancestry meta-analysis (including 9,846 people of Iranian, Indian and East Asian descent) has added 38 more loci, bringing the total number of genes up to 200.[86] The total proportion of disease variance now explained by genetics alone is 13.1% in CD and 8.2% in UC.[86]

1.5.1 Crohn's disease and ulcerative colitis: A Genetic continuum of disorders?

Early GWASs identified IBD loci common and unique to CD and UC.[89] More recent data demonstrate the increasing proportion of loci common to both diseases.[88] In Jostins et al, 110 of 163 loci were associated with both diseases, (30 CD, 23 UC specific),[88] in Lui et al 137 of 200 were associated with both forms of IBD (36 CD, 27 UC specific). [86]

Specific genetic variants have been known to associate with certain subtypes of IBD, nucleotide oligomerization domain 2 (*NOD2*) mutations have been associated with stricturing ileal CD, and *DRB1*01:03* has been associated with severe extensive UC.[90–92] A large collaborative genotype-phenotype study of CD and UC demonstrated three loci associated with disease location (*NOD2* - ileal, *MST1* 3p21 - ileal and MHC *DRB1*01:03*- colonic disease).[93] This study suggests that ileal and colonic CD are as genetically distinct from each other as they are from UC, and that IBD should perhaps be classified into three disorders rather than two: ileal CD, colonic CD and UC.[93] A genetic risk score was developed to assist classification of patients, and predict patients who were initially misclassified with a diagnosis of UC or CD.[93]

Studies of gene loci shared by UC and CD may provide insight into their common pathogenic mechanisms. The T-helper (Th)17 and interleukin (IL)12-23 pathway is well established in IBD pathogenesis, with susceptibility gene loci *IL23R*, *IL12B*, *JAK2*, and *STAT3* identified in both UC and CD.[94,95] Variants in *IL12B*, which encodes the p40 subunit of IL12 and IL23, have been associated with IBD and other immune disorders.

Defects in the function of IL10, an immunosuppressive cytokine, have also been associated with CD and UC. [96,97] A severe, childhood-onset, CD-like form of enterocolitis is associated with rare mutations in *IL10R*. However, this disorder could be a separate entity from idiopathic IBD.[98,99] Other susceptibility genes that regulate immune function include *CARD9*, *IL1R2*, *REL*, *SMAD3,* and *PRDM1*.[100] Interestingly, the well-established CD-risk variants of *NOD2* and *PTPN22* appear to protect against UC.

1.5.2 CD-Specific Susceptibility Loci and Pathways

Risk for CD has a greater genetic component than that for UC, and several CD-specific susceptibility loci have been delineated. Data from Jostins et al increasingly highlights the relationship between the host innate immune system and the intestinal microbiota in CD. GWASs have indicated that intracellular bacterial processing by autophagy is an important pathogenic mechanism. Importantly, the association between CD and *NOD2* has been consistently replicated, at the genome-wide significance level; [101] NOD2 has been mechanistically linked with autophagy.[102,103] Cigarette smoking, a strong environmental factor in CD risk, might affect NOD2 function.[104] Furthermore, the product of the CD susceptibility gene *ATG16L1* is recruited to the plasma membrane by NOD2, where it initiates bacterial internalization by autophagosomes.[97,101,103]

Another gene involved in autophagy-induced bacterial killing is Immunity-related GTPase family M (*IRGM)*. CD-associated polymorphisms in *IRGM* lead to reduced protein expression. A different SNP of *IRGM* protects against *Mycobacterium tuberculosis*.[105,106] The most recent data from Immunochip studies indicated an overlap between IBD loci and complex mycobacterial disease loci:[88] Seven CD susceptibility genes overlap with leprosy susceptibility genes, and 6 mycobacterium susceptibility genes overlap with IBD loci. However, for several of these diseases, the genetic associations have opposite effects.[88] Genes involved in the host response to mycobacteria that were previously associated with CD include *CARD9* and *LTA*. [89,97] Other CD-specific loci identified related to the immune system include *PTPN22*, *IL2RA*, *IL27*, *TNFSF11*, and *VAMP3*. [97,101]

1.5.3 UC-Specific Susceptibility Loci and Pathways

Although UC susceptibility loci have primarily included genes that regulate intestinal epithelial barrier function, there is recent evidence that human leukocyte antigen (HLA) variants are involved in development of UC.[88] *HLA-DQA1* was the locus most strongly associated with UC (odds ratio of 1.44),with no corresponding increased risk in CD.[88] The HLA class II genes are tremendously diverse and control antigen presentation to T cells; they have been implicated in other immune diseases.

Hepatocyte nuclear factor 4A (*HNF4A*) regulates expression of cell junction proteins in the intestinal epithelial barrier; variants have been associated with UC, and also colorectal cancer—a complication of chronic inflammation in patients with IBD.[107] Rare SNPs at the *HNF4A* gene locus, not implicated in UC, are associated with maturity onset diabetes (MODY), inherited in an autosomal dominant fashion.[108] Other UC-associated genes that affect epithelial barrier function include *CHD1*, which encodes E-cadherin, and *LAMB1*, which encodes the lamina β subunit 1. Many UC risk alleles encode cytokines and inflammatory mediators, including tumour necrosis factor (TNF) receptor superfamily members (*TNFRSF14*, *TNFRSF9*), interleukins, and interleukin receptors (*IL1R2*, *IL8Ra*/*RB*, *IL7R*).[100]

1.5.4 Relationships with Other Diseases

Jostins et al. reported that 70% of IBD loci overlap with loci associated with other complex immune diseases, such as *IL23R* variants associated with psoriasis and ankylosing spondylitis.[109–111] However, these polymorphisms sometimes have opposite effects in different diseases. For example a variant of *PTPN22* protects against CD, but is a risk factor for type-1 diabetes and rheumatoid arthritis.[112] Extra-intestinal manifestations of IBD also share common loci, which may explain their co-occurrence. For example variants of *REL*, *IL2*, and *CARD9* are associated with UC and primary sclerosing cholangitis.[89,113]

1.5.5 Current Agenda for IBD Genetic Studies

Many of the IBD loci identified so far have not been accurately characterized or fine-mapped, and the candidate genes commonly used to describe them are only putative. Moreover, the biological functions of their products, and their complex interactions, in most cases require delineation. Studies are underway to fine-map loci, and functional studies are needed. Further work is required to determine how specific variants affect levels of mRNA and consequently protein, which could provide further insight into mechanisms of pathogenesis. This is likely to take considerable time—*NOD2* was identified over 10 years ago and there is still uncertainty about its function.[114]

GWA studies have excelled in identifying moderate-risk genetic variants with at least 5% prevalence in the population. Novel approaches are needed to discover lower-prevalence variants with higher effect size. Whole-exome sequencing, which covers only coding areas of the genome, costs less than whole-genome sequencing and tends to afford higher depth coverage and therefore greater certainty about novel discoveries. It has been successfully used to identify single mutations in very early onset IBD,[115] and is perhaps most likely to produce results in individuals with a strong family history or early age of disease onset. However, many polymorphisms that affect disease susceptibility are located in non-coding areas of the genome; the ENCODE project has highlighted the importance of non-coding regions in disease risk.[116] Exome sequencing and whole genome sequencing are each underway, with large-scale endeavours at the Sanger Centre likely to report in the near future. (http://www.ibdresearch.co.uk/)

Most IBD genetic analyses have been performed in the white populations of Northern Europe and America. More recently, there has been a push to expand this work to other ethnic populations.[117] IBD-associated variants of *NOD2*, for example, are less prevalent in African American populations, and CD-associated mutations have not been detected in Asian or Sub-Saharan African populations.[118,119] The aforementioned trans-ancestry meta-analysis[86] has demonstrated population-specific effects caused by differences in risk allele frequency (RAF), for example, three coding variants of *NOD2* known to have a large OR in Europeans, had a RAF of 0 in East Asians.[86] Another difference was the relative risk contribution of susceptibility alleles; *TNFSF15/TNFSF8* variants have a small/modest effect size in those of European ancestry (OR~1.15) but a comparatively larger effect in people of East Asian ancestry (OR~1.75) despite similar RAF in both. [86]

Pharmacogenomics – the study of how genomic factors affect the efficacy, tolerability, and side-effects of a therapeutic agent– remains high on the research agenda. Patients are routinely evaluated for thiopurine S-methyltransferase (encoded by *TPMT*) genotype and phenotyping prior to initiation of thiopurine therapy is recommended by the US Food and Drug Administration.[120,121] There are ongoing attempts to predict patients' response to other agents, based on genetic factors. A notable recent success in this field is the association of an HLA variant (*HLA-DQA1-HLA-DRB1*, OR 2.59) with thiopurine-induced pancreatitis.[122] Studies supported by the Serious Adverse Events Consortium aim to predict 5-aminosalicyate-induced nephrotoxicity (http://www.ibdresearch.co.uk/5asa/) and serious complications of anti-TNF therapies (PANTS study).

A main goal of IBD research is to develop disease-specific therapeutics. Many researchers are developing reagents to alter activities of genes and pathways identified through GWAS—the IL12/23 signalling pathway is one promising target. Ustekinumab, a monoclonal antibody that binds to the shared p40 subunit encoded by *IL12B* described above, has undergone phase 2b induction and maintenance trials in patients with CD.[82] Apilimod mesylate, briakinumab (ABT-874), and SCH-900222) also target components of the IL12/23 signalling pathway are currently under evaluation.[123]

1.6 From genetics to environment via epigenetics

The challenge remains to measure patients' duration, intensity, and frequency of exposure to the many environmental factors that potentially could contribute to IBD (Table 2), making the environmental impact on disease difficult to disentangle.[124]

Exposure	**Evidence for causality**	**Role in natural history**	**Putative mechanisms**
Cigarette smoking[125]	Strong: +ve in CD, -ve in UC	Similar associations with flare	Effects on gut mucosa & vasculature
Appendicectomy[126]	Inverse ass. with UC, nil for CD	Uncertain	Removal of lymphoid tissue
Use of NSAIDS[127]	Inconsistent observational data	Associated with exacerbations	↓ prostaglandins and blood flow.
Oestrogens: OCP/HRT [128]	Data from observational studies	Uncertain	Reduction in mucosal blood flow.
Use of antibiotics[129]	Association with CD, nil for UC	↓ relapses for CD; nil in UC	Effect on gut microbiota
Deficiency of vitamin D[130]	Minimal epidemiological data	Not known	Immune-mediated
Polyunsaturated fatty acids High omega-6 : omega-3 ratio[131,132]	Emerging, but inconsistent data	No current studies	Effect via prostaglandin production
↓ of diet plant fibres[133]	Association with CD, nil for UC	Not known	↓ butyrate, ↓ energy for colonocytes.
↑ Physical exercise[134,135]	Associated with ↓ CD	Not known	↑ butyrate & ↑ microbiota diversity

Table 2 - Environmental factors associated with increased IBD-susceptibility

Epigenetic factors could mediate gene–environment interactions involved in pathogenesis. Epigenetic programming begins upon fertilization and continues throughout life. Studies of Agouti mice[136] and the offspring of post-World war II Dutch famine survivors[137] revealed how the environment may affect epigenetic factors. Dietary intake during pregnancy may affect the epigenetic reprogramming step in offspring during early development; an effect that may persist for up to 2 generations, although these data have caused some debate in the scientific community.[136,138–141] Moreover, there is evidence for acquired epigenetic changes with aging,[142] caused by a range of environmental factors.[143] Epigenetics could therefore play a central role in the pathogenesis of IBD and other diseases, affecting interactions among genetic and environmental factors such as the intestinal microbiome.

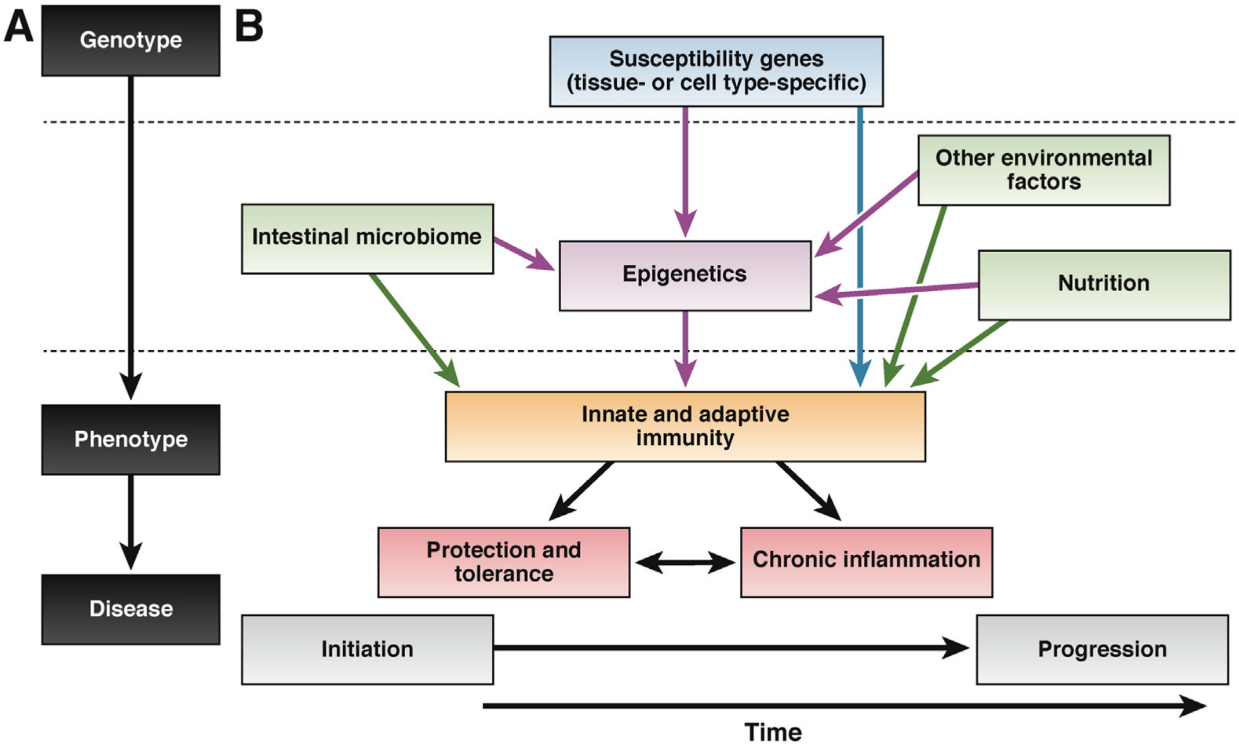


Figure 4 – Schematic diagram of the potential role of epigenetics in disease pathogenesis. The classic paradigm of genotype leading to phenotype and disease (A) is likely to be overly simplistic. In panel B, epigenetics may mediate between genetics (blue box), environmental factors (green boxes) and the immune system (orange) to contribute to disease initiation and propagation. Taken from Ventham et al.[144] License Number: 3742451069879

In IBD research, several key developments in molecular studies have led us from genetics to explore epigenetics. GWASs have identified key epigenetic regulatory enzymes, DNMT3a and more recently DNMT3b as CD-susceptibility genes.[88,97] Dendritic cells that express CD-associated variants of *NOD2* fail to upregulate microRNA clusters that regulate Th1- and Th17-cell mediated immune responses.[145] Epigenetic mechanisms have also been shown to regulate the immune system. For example, differentiation of Th2 cells requires epigenetic silencing of the *IFNG* locus.[146]

Epigenetics may be defined as mitotically heritable changes in gene function not explained by changes in the DNA sequence. Gene expression can be altered by changes to the structure and function of chromatin (Figure 5). The main epigenetic mechanisms include DNA methylation, histone modification, RNA interference, and the positioning of nucleosomes (which will not covered in depth).

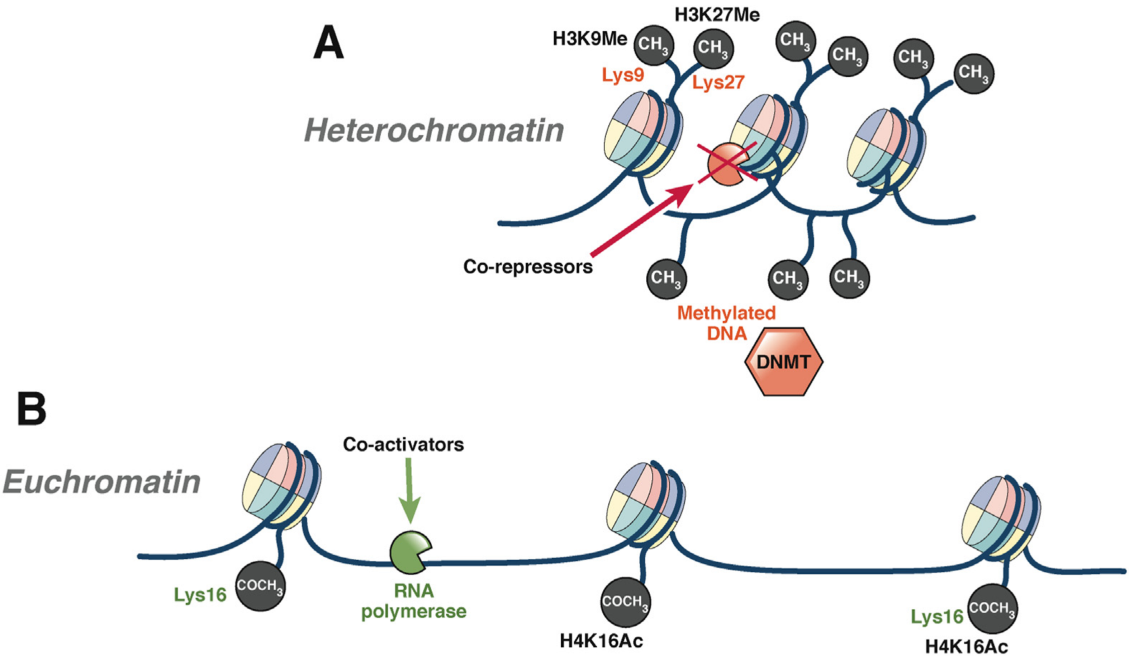


Figure 5 – Structure of chromatin. Chromatin can exist in a condensed form (A) – heterochromatin and is associated with transcriptional repression. Heterochromatin is associated with methylated DNA, and methylated histone H3 and low levels of histone acetylation. Heterochromatin. Euchromatin (B) is the relaxed form of chromatin that allows transcription. Here DNA is not methylated and specific histones are acetylated (H4). Taken from Ventham et al.[144] License Number: 3742451069879

The epigenome can be regarded as both stable and plastic. The epigenome can be regarded as stable as epigenetic marks are passed onto daughter cells during mitosis.[147] However, stochastic and environmental factors can cause dynamic changes to the epigenome over time.[148] During mitosis, the level of fidelity of epigenomic replication is much lower than that of genetic sequence (error rate of 1×10^6 for DNA sequence compared to 1×10^3 for DNA modifications), leading to an accumulation of epigenetic changes over time.[149,150] Similarly, several environmental factors produce epimutations (epigenetic changes associated with disease); factors relevant to IBD include smoking,[151,152] the microbiota,[153,154] and diet.[138] Epigenetic marks are reset during meiosis; the epigenome is established early in embryogenesis after undergoing several reprogramming steps, during which the epigenome is most subject to modification.[147,155] Given that the epigenome is reset during meiosis, it was believed that epigenetic marks were not passed between generations.

Although environmental exposures in utero can lead to epigenetic changes that persist for up to 2 generations, there is increasing interest in true epigenetic inheritance, which lasts multiple generations.[156–158] Transgenerational epigenetic inheritance has attracted both excitement and scepticism within the scientific community, and pertains to epigenetic marks resistant to the major reprogramming steps.[124,157] The most compelling evidence for transgenerational epigenetic inheritance comes from studies of plants: DNA methylation-mediated silencing of the *Lcyc* promoter causes variations in floral symmetry in *Linaria vulgaris* (toadflax) that are stably inherited over many generations.[159] Although additional examples have been reported from studies of plants, insects, and mammals, the concept is still met with some skeptacism.[160] Incomplete erasure of epigenetic mutations across generations could contribute to familial predisposition to diseases such as IBD.[124]

1.6.1 DNA Methylation

DNA methylation is the most widely studied epigenetic modification; in this process, a methyl group is covalently added to cytosines that are part of cytosine–guanine dinucleotides (CpG). Full methylation occurs when cytosine residues on both DNA strands are methylated. CpG dinucleotides are generally sparse within the genome (~1%), but are relatively concentrated in specific regions called 'CpG islands'. CpG islands are defined as a 200-base sequence containing greater than 50% CpG dinucleotides at an observed-to-statistically expected ratio of 0.6.[161] The areas where most tissue-specific methylation appear to border CpG islands and have been termed 'CpG shores'.[162]

Transcriptionally repressive activity generally occurs where a gene has methylation of CpG islands within promoter areas, and is an important mechanism of gene silencing.[163] DNA methylation may lead to transcriptional repression by hindering access of transcription factors to promoter regions, although many researchers believe the reverse is true: that gene silencing subsequently leads to DNA methylation.[164,165]

DNA is methylated by enzymes called the DNA methyltransferases (DNMTs). There are 5 members of the DNMT family: DNMT1 (maintenance of methylation), DNMT2 (involved in RNA methylation), DNMT3a, DNMT3b, and DNMT3L (involved in new methylation). There is evidence that these DNMTs interact, and that other epigenetic mechanisms can recruit DNMTs to specific gene loci.[161] Inherited deficiency of DNMT3B leads to immunodeficiency, centromeric instability, and facial anomalies (ICF) syndrome, whereas complete lack of DNMT enzymes leads to embryonic lethality. [161]

Initial DNA methylation studies in the context of IBD largely focused on IBD-related cancer predisposition. DNA methylation changes in colonic epithelial cells that normally occur with aging are accelerated in IBD, possibly related to a higher cell turnover in inflammation.[166] Increased age-related DNA methylation, observed in colon cells of patients with colitis, could lead to genetic instability and cancer development.[166] DNA hypermethylation has been demonstrated in dysplastic and the surrounding non-dysplastic colon tissues from patients with UC, compared to control subjects or UC patients without dysplasia.[166]

The increasing interest in the role of DNA methylation in IBD pathogenesis has occurred in tandem with advances in platform-based DNA methylation array technologies, which have superseded candidate gene methylation profiling techniques. Initial IBD epigenome-wide methylation association studies (EWASs) used platform-based arrays to analyse peripheral blood samples. Nimmo *et al*. analysed the methylation profile of peripheral blood from women and children with CD using the 27K Illumina microarray.[167] Fifty genes demonstrated significantly different levels of methylation between patients with IBD and controls, including some involved in immune system activation (*MAPK, RIPK3,* and *IL21R*). Ontology analysis highlighted several pathways associated with IBD, including immune system processes, immune response, and host response to bacteria, whereas canonical pathway analysis indicated the involvement of Th17 cell pathways.[167]

Another study demonstrated the tissue-specific nature of epigenetic marks.[168] No significant differences in DNA methylation were observed between children with IBD and controls, based on methylation-specific amplification microarray analysis of peripheral blood. However, they found that peripheral blood mononuclear cells (PBMCs) from the IBD patients demonstrated hypermethylation at the *TEPP* locus, which encodes testes, prostate and placenta-expressed protein and is of uncertain relevance in IBD.[168] A study that analysed DNA methylation in Epstein-Barr virus-transformed B cells from 18 patients with IBD vs non-affected siblings identified 49 differentially methylated CpG sites. More than half of the differentially methylated loci contained genes that regulate immune functions, including several (*BCL3, STAT3, OSM, STAT5*) involved in the IL-12 and IL-23 pathways.[169]

DNA methylation has also been studied in colonic tissue. An EWAS of intestinal biopsy samples from 20 monozygotic twins discordant for UC identified 61 differentially methylated loci, several containing genes that regulate inflammation (*CFI*, *SPINKK4*, *THY1*/*CD90*). This study had an interesting design, in that after the loci were identified in the analysis of discordant monozygotic twins (to exclude differences in genetic factors) they were validated an independent cohort.[170]

To overcome the heterogeneity of cell types in tissues, a methylation-wide profiling study of whole rectal biopsies from patients with active and quiescent UC and CD was validated using isolated epithelial cells from rectal biopsies.[171] Many differentially methylated genes were identified in whole tissue, encoding proteins including *DOK2* (involved in IL-4 mediated cell proliferation), Tap1 (an MHC class I transport molecule), and members of the TNF family (*TNFSF4* and *TNFSF12*). *ULK1* was methylated only in patients with CD; its product has a role in autophagy. Genes identified as being differentially methylated in this study, replicated findings from other EWAS,[167]and have also been identified as susceptibility genes in GWAS[94,107] including *CDH1, ICAM3, IL8RA,* and *CARD9*.

1.6.2 Histone Modification

Histones can undergo a range of complex modifications; the N-terminal amino acid histone tails can be modified by acetylation, methylation, ubiquitination, and phosphorylation.[172] Different post-transcriptional modifications to histone ends are thought to recruit different co-activators or co-repressors, which determines whether chromatin is in its relaxed or condensed form.[173]

Histone acetylation is the most well described post-translational modification and is regulated by the levels and activity of histone acetyl transferase (HAT) and histone deacetylase (HDAC).[174] In a simple model, chromatin is transcriptionally active when lysines on histones H3 and H4 are acetylated. Although it is not exactly clear how acetylated histones affect transcription, they might change the structure of chromatin (acetylation of lysine neutralizes the positive electrostatic charge of the histone, facilitating the opening of chromatin), and thereby reveal binding sites for important co-activators.[175] Overexpression or increased activity of HDACs can lead to hypoacetylation and gene silencing.

Patterns of colonic histone acetylation in rats with chemical-induced (dextran sulphate sodium (DSS) and 2,4, trinitrobenzene sulphonic acid) and in biopsies from patients with CD have been described. Inflamed tissue and Peyer's patches from rats with colitis and patients were found to have increased acetylation of H4.[176] Several mechanisms have been proposed to link histone modification with inflammation, involving the innate immune response to microbiota. Butyrate, an endogenous metabolite formed during fermentation of dietary fibres by the intestinal microbiota, is an HDAC inhibitor. Butyrate increases expression of *NOD2* by increasing histone acetylation in its promoter region.[154] Toll-like receptor 4 regulates intestinal homeostasis by preventing excessive inflammatory responses to commensal bacteria, and could be regulated by histone deacetylation. [153]

HDAC enzymes can be inhibited by a range of natural (e.g. lactate, butyrate) and synthetic compounds,[177] and much of our understanding of histone modifications in the context of IBD has come from experimental trials of histone deacetylatase inhibitors (HDACi). HDAC inhibitors have primarily been investigated in cancer research, but also have anti-inflammatory effects. [178] It is worth noting that the enzymes that affect histone acetylation status (HAT, HDAC), do not act exclusively on histones, but affect acetylation of a range of proteins, including p53, and STAT3.[179] Therefore, HDAC inhibitors act not only through epigenetic mechanisms, but on multiple histone-independent targets, including the transcription factor NFκB pathway, cytoskeletal proteins, and cell cycle and apoptosis regulators.[180,181] Butyrate enemas have been used to treat patients with colitis, although HDAC inhibition may not be their predominant mechanism of action. Butyrate has several effects of the gastrointestinal tract, including maintenance of barrier function and a homeostatic reduction in epithelial cell production of IL8.[182–184] Butyrate reduces the disease activity index of patients, as well as nuclear translocation of NFκB in lamina propria macrophages.[185] Other HDACis have also been shown to ameliorate DSS-induced colitis in mice.[183,186] Although an interesting field of research, histone modifications will not be a focus of this thesis.

1.6.3 RNA Interference

MicroRNAs (miR) are single-stranded non-coding RNAs typically 22-nucleotides and are highly conserved throughout evolution.[187] MiRs with members of the Argonaut (Ago) family form the RNA interference-silencing complex [RISC]. This complex regulates translation by binding to 3' regions of untranslated mRNAs, by directly inhibiting mRNA translation or by causing mRNA degradation (depending on the degree of complementarity between the miR and the mRNA target).[188]

Following their description in the mid-1990s, a large number of miRs have been described (>1600 in humans, see http://mirbase.org). miRs are transcribed by RNA polymerase II into hairpin structures called pre-miR. Pre-miR is processed in the nucleus (by enzyme Drosha) and then the cytoplasm by the Dicer enzyme.[189] After processing, each miR may demonstrate complementarity with many different mRNAs, and each mRNA may be targeted by many different miRs.[190,191] miRs regulate gene expression and thereby numerous biological processes, including cell proliferation, differentiation, and death.

Studies in animals have shown that intestinal miRs regulate gut homeostasis. Mice deficient in intestinal Dicer1, an miR-processing enzyme, have disorganized intestinal epithelial crypts with increased goblet cells, rapid jejunal epithelial migration, and accelerated apoptosis. Additionally, mice deficient in intestinal Dicer1 have increased inflammation and neutrophil and lymphocyte migration, and reduced epithelial barrier function, compared to mice not deficient in Dicer1.[192]

A number of studies have investigated differences in miRs between patients with and without IBD. Changes in miRs in human IBD were first described in 2008.[193] In sigmoid biopsies from patients with active UC, levels of 8 miRs were significantly increased and 3 were decreased, compared to samples from patients without UC. miR-192, normally expressed in colonic epithelial cells, was significantly reduced in tissues of patients with active UC.[193] miR-192 reduces expression of macrophage inhibitory peptide-2α, a CXC chemokine expressed by epithelial cells; its levels are increased in colon tissues of patients with UC.[193]

miR-150 is upregulated in mice with DSS-induced colitis and colon tissues from patients with UC; its levels correlate inversely with those of its target c-Myb, which has a role in apoptosis.[194] Upregulation of miR-21, which promotes inflammation, has been reported in several studies of patients with active UC and CD colitis (but not ileitis), along with miR-155.[195–197] miR-196 is overexpressed in the inflamed epithelium of patients with CD patients and may reduce IRGM-mediated autophagy.[198]

Distinct miR signatures have been identified in peripheral blood samples from patients with IBD, compared to controls, and between patients with CD vs UC.[199] Several miRs have been found to be significantly up- or down-regulated in 2 or more studies, including miRs-16, -21, -28-5p, -149, -151-5p, -199-a, and -532-3p.[199–201] Eleven miRs were also found to be differentially expressed between serum samples from paediatric patients with CD and healthy children.[200]

Further adequately powered studies are required to identify IBD-associated miR profiles in intestinal tissues and serum, plasma, and separated blood cells. Specific miR profiles might be able to predict IBD susceptibility, progression, and response to therapy. Moreover, identifying the targets of these miRs will provide additional insight into IBD pathogenesis.

1.7 Potentials pitfalls in epigenetic research

A major hurdle in interpreting results from epigenetic studies is to determine causality i.e. whether a particular epigenetic profile is cause or consequence of disease. Furthermore, many of these studies provide only a snapshot of the epigenetic profile after the disease process has been established, rather than describing a temporal relationship between an epigenetic alteration and subsequent disease development. The epigenetic profile changes over time, and apparent associations may be a consequence of the disease itself, or other environmental factors such as medications.[202,203] Conditional correlation models have attempted to determine how DNA methylation and genetic factors interact to cause diseases.[204]

Epigenetic marks are tissue and cell-type specific and therefore selection of a disease-relevant tissue type is crucial. In IBD research, it has been a major challenge to identify disease-relevant cell types. Currently, interest is focused on immune cells in the blood (such as CD4+ and CD8+ T cells) and the intestine (intraepithelial lymphocytes).

Results from studies of whole tissues, such as whole blood or colon biopsy samples, are difficult to interpret because of the heterogeneity among cells, each cell type possessing their own epigenetic signature. Early epigenetic studies of IBD were mostly performed with whole tissues, with some of the significant results likely to be arising from purely from differences in cellular proportions between cases and controls. Some researchers have used *in-silico* methods to adjust for differing cell proportions.[205]

1.8 Interaction between genetics and epigenetics

An intriguing field of investigation is the relationship between genetic and epigenetic factors. There is evidence of co-localization of differentially methylated CpGs at predisposing SNPs identified at GWAS. In our own EWAS of CD, we demonstrated enrichment of methylation changes within 50 kb from GWAS-identified susceptibility loci, including *IL-19, IL-27, TNF,* and

NOD2.[167] In a recent large methylation study of patients with rheumatoid arthritis, in 5 of 9 MHC genes, a specific genotype was associated with differential methylation.[206]

This phenomena has also been observed in studies of patients with type-2 diabetes mellitus, where a specific allele, rs8050136, within the obesity and diabetes susceptibility gene *FTO,* is associated with increased DNA methylation.[207] However, Toperoff et al. associated a different *SNP*, rs1121980, with hypomethylation of *FTO*.[208]

It is not clear how these SNPs affect methylation of the gene. They could increase the numbers of CpG dinucleotides, or alter the access of the methylation machinery to the gene.[207] Allele- or haplotype-specific methylation occurs more commonly with *cis*-acting polymorphisms.[209] A potential cofounder of the Illumina 450K HumanMethylation microarray, used in most DNA methylation studies, is that certain probes contain SNPs or repetitive elements that can affect methylation analysis.[210]

Variants in *STAT4* have also been reported to alter its methylation (Figure 6 A). Additional evidence of haplotype-specific methylation has been demonstrated in the promoter regions of *IL8RA* and *IL8RB*. Rectal biopsies from patients with IBD were shown to have increased methylation of the CpG island closest to the transcriptional start site of *IL8RA*—the proposed binding site of transcription factor PU.1 (SPI-1). [171] The risk allele rs11676348 alters a CpG, is located between *IL8RA* and *IL8RB* coding sequences, and contains a binding site for the transcription factor *STAT3.*[171] Although not specifically probed itself by the Illumina 27K microarray, differential methylation was observed on either side of rs11676348.[171]

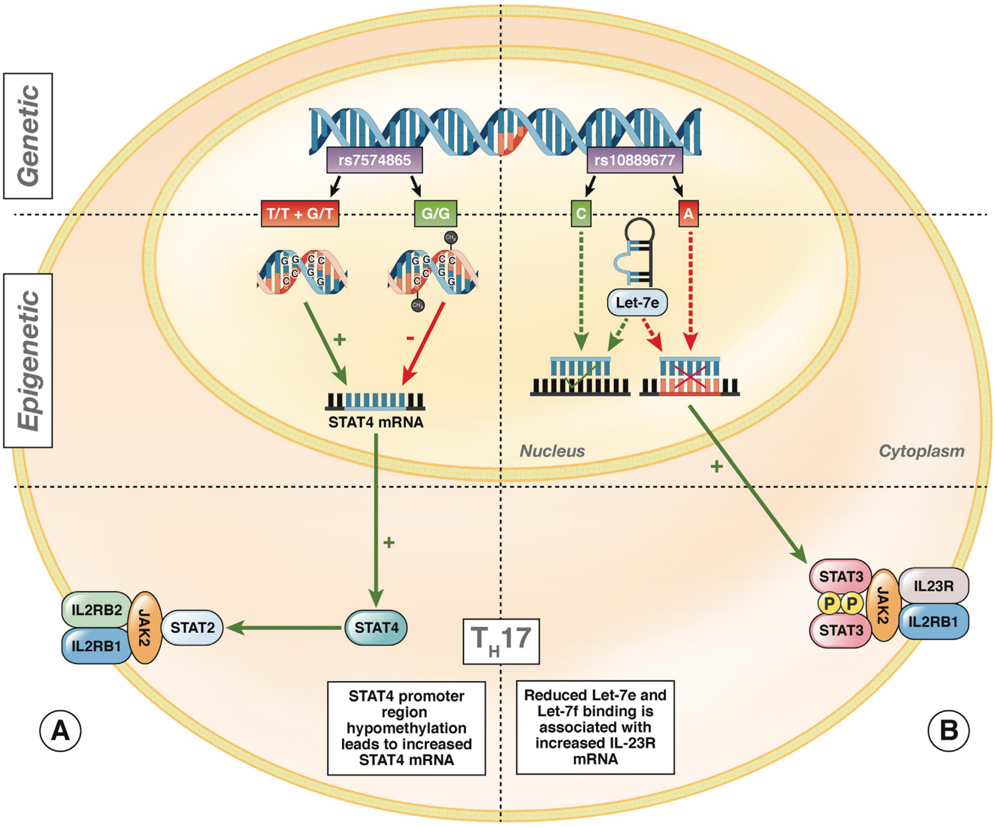


Figure 6 - The Relationship between Genetic Polymorphisms and Epigenetic Factors.

Epigenetic features of T cells in patients with IBD affect Th 1 and Th17 cell differentiation. A - STAT4 is associated with several immune diseases, acting as a transcription factor for IL12 and IL23 which leads to Th 1 and Th17 cell differentiation.[211] A SNP in *STAT4*, rs7574865, is associated with several immune disorders, including IBD, rheumatoid arthritis, type-1 diabetes, and lupus.[212–216] The rs7574865 risk variants (T/T +G/T) are associated with promoter region hypomethylation in colon tissues and PBMCs of patients with IBD. STAT4 promoter hypomethylation was associated with increases in STAT4 mRNA and could promote the Th1 phenotype and IFNγ production.[217] In T-cells from patients with asthma, STAT4 expression is also regulated by DNA methylation at promoter regions. Interestingly STAT4 expression was markedly increased following treatment with a DNMT inhibitor.[218]

B - An IBD-associated SNP in *IL23-R*, rs10889677, is associated with increased levels of IL-23R mRNA and protein. This could result from reduced binding of microRNAs Let-7e and Let-7f at the regulatory 3'UTR region of the rs10889677 risk variant (A) compared to cells from patients without IBD (C).[219] Reduced binding of Let-7e and Let-7f to rs10889677 is associated with increased levels of IL23R mRNA and protein, potentially leading to sustained activation of Th17 cells and the chronic inflammation associated with IBD.[219] Taken from Ventham et al.[144] License Number: 3742451069879

Similarly, SNPs can affect the complementarity of miR binding. IBD-associated variants of *IL23R,* which has a role in IL12 and IL23 signalling, may demonstrate altered binding with miRs Let-7e and Let-7f, leading to altered expression of IL23R and inappropriate Th17 activation (Figure 6 B). *IRGM* mediates innate immune defence against intracellular organisms, including mycobacterium tuberculosis.[106] Variants of *IRGM* alter the binding site for miR-196 (Figure 7 B).[220]

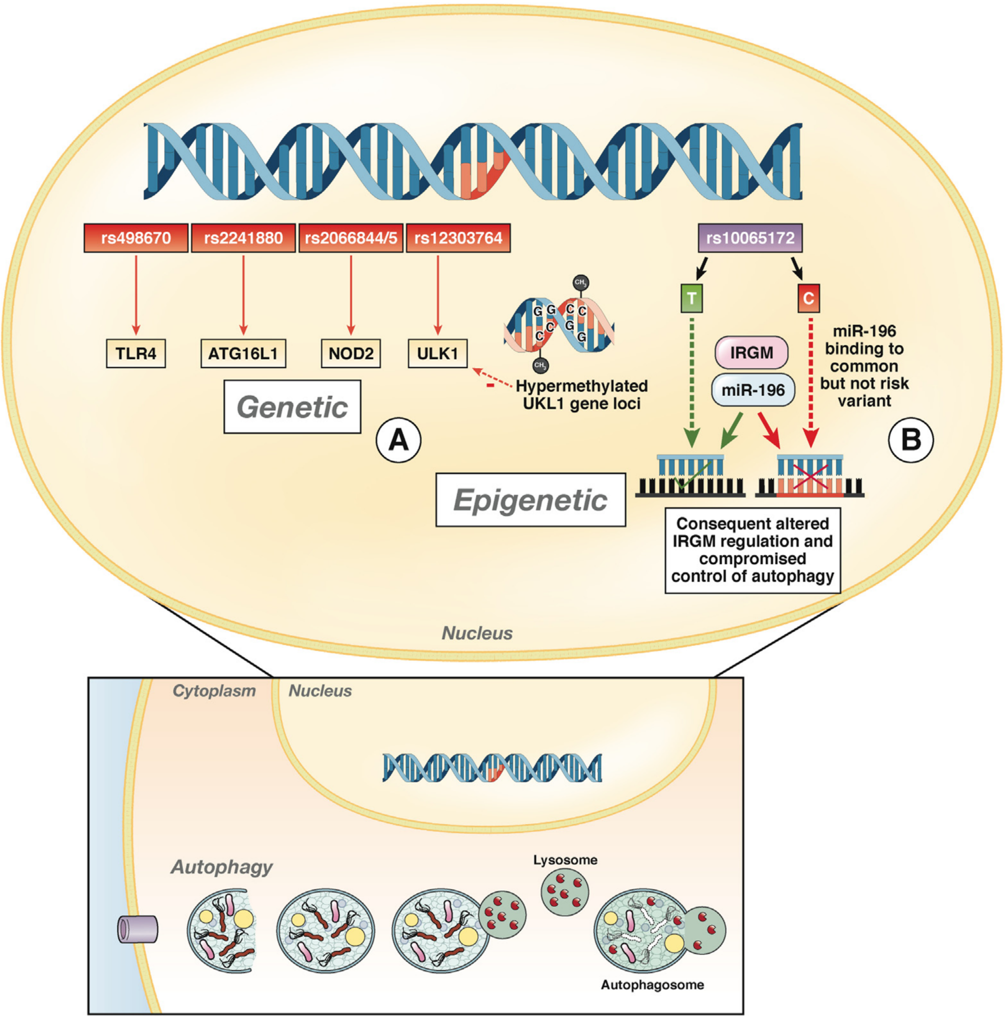


Figure 7 – Loci Identified in GWASs Indicating Roles for the Innate Immune Response to the Microbiota and Autophagy in the pathogenesis of CD.

A - Several CD risk alleles were identified in GWAS in genes that control autophagy, including *TLR4*, *ATG16L1*, *IRGM*, and *ULK1*. The *ULK1* locus is both a CD-susceptibility locus and has been shown to be hypermethylated in cells from patients with CD compared to controls. [171]

B - *IRGM* encodes a gene that regulates the innate response to intracellular organisms, including mycobacterium tuberculosis. The CD risk allele rs10065172 is associated with a deletion upstream of *IRGM*.[221] This SNP had been termed non-causative, due to an absence in alteration of protein sequence or splice sites. However, the risk variant has an altered binding site for microRNA-196. Individuals with this SNP downregulate *IGRM*. The consequence is a functionally reduction of autophagy and processing of the adhesive invasive *E coli*, which has been associated with CD.[220] Taken from Ventham et al.[144] License Number: 3742451069879

Another study evaluated the UC risk conferred by 3 common allelic variants of 3 pre-miRs (miR-146a, -196a and -499). Three SNPs (rs11614913, rs2910164 and rs3746444) were genotyped in 170 UC patients and 403 control patients. The AG heterozygous genotype of rs3746444, encoding miR-499, was significantly associated with an increased risk of UC (odds ratio of 1.51). The same genotype was also associated with older age of onset, left-sided colitis, hospitalization, and steroid dependence.[222]

1.9 Biomarkers in IBD for diagnosis and prognosis

A biomarker is defined as 'characteristic[s] that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention.'[223] In the context of IBD, biomarkers may be used to assist in establishing a diagnosis or to identify individuals who are likely to develop a more severe disease course. Many of the currently available biomarkers are not specific for IBD, and are general markers of inflammation. Whilst specificity is critical when attempting to diagnose IBD, specificity is less critical for monitoring disease course after a diagnosis has been established.[224]

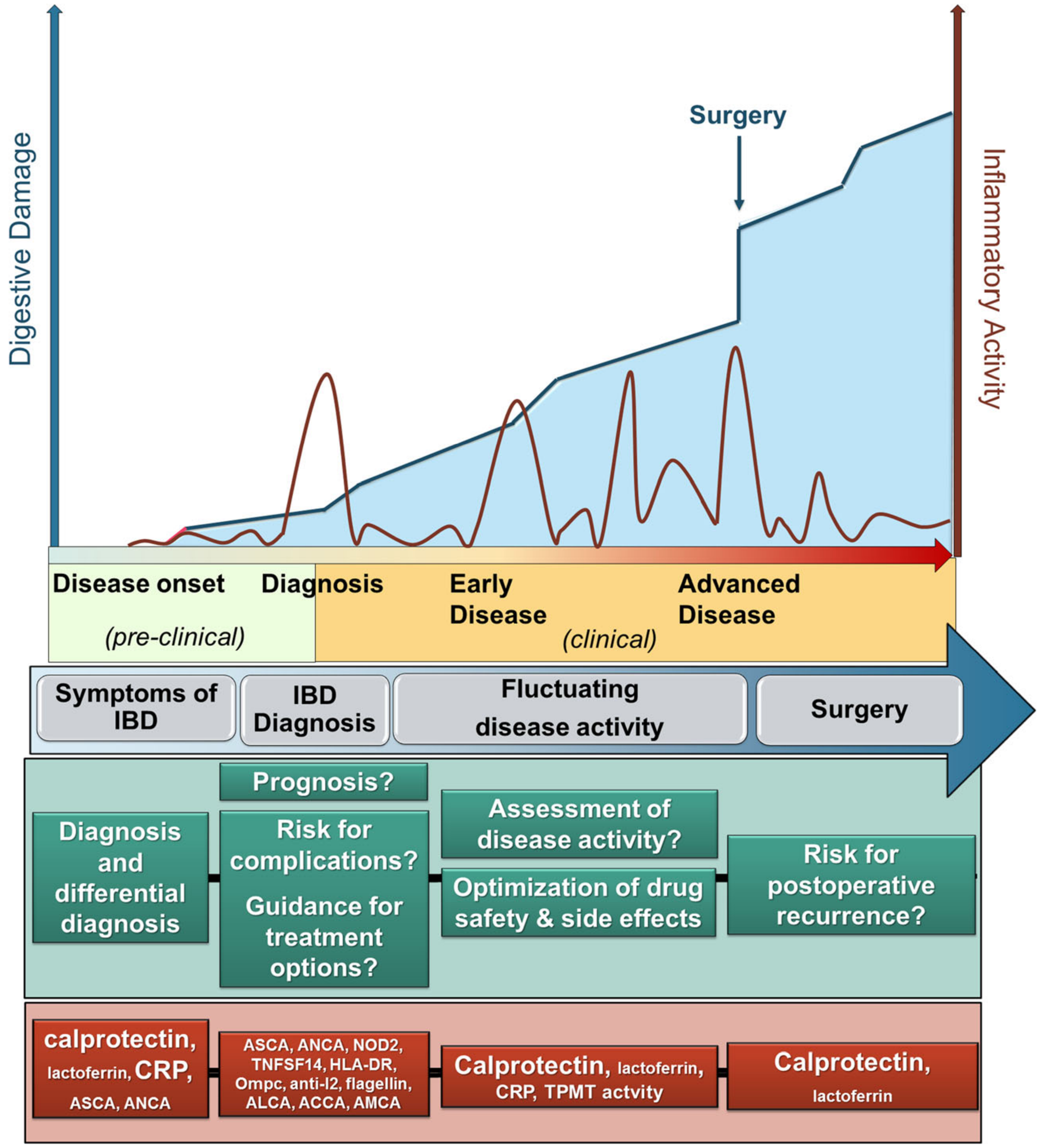

Digestive Damage
Surgery
Inflammatory Activity
Disease onset
Diagnosis
Early Disease
Advanced Disease
(pre-clinical)
(clinical)
Symptoms of IBD
IBD Diagnosis
Fluctuating disease activity
Surgery
Prognosis?
Diagnosis and differential diagnosis
Risk for complications?
Guidance for treatment options?
Assessment of disease activity?
Optimization of drug safety & side effects
Risk for postoperative recurrence?
calprotectin, lactoferrin, CRP, ASCA, ANCA
ASCA, ANCA, NOD2, TNFSF14, HLA-DR, Ompc, anti-I2, flagellin, ALCA, ACCA, AMCA
Calprotectin, lactoferrin, CRP, TPMT actvity
Calprotectin, lactoferrin

Figure 8 – Biomarkers for IBD diagnosis and prognosis in specific situations (green). Reproduced with permission from Rogler G Clinical Utility of Biomarkers in IBD. Current Gastroenterology Reports 2015 17 (26).[224] License Number: 3743150309751.

1.9.1 Diagnosis

Biomarkers are used in clinical practice to assist in the diagnosis of IBD. As detailed above, for a diagnosis of IBD to be established, patients require a combination of clinical, laboratory, radiological, endoscopic and histological investigations. However, histology obtained from ileocolonoscopy remains the most important single aspect of IBD diagnosis. Given endoscopy is an invasive test and requires considerable resources, a simple, easily assayed biomarker would assist in triaging those who require an endoscopy at first clinic attendance. Such a biomarker if available in primary care may also inform referral to secondary care. Lastly, such a non-invasive biomarker may be used in cases of small bowel Crohn's disease which can often be a diagnostic challenge prior to surgery or starting powerful imunomodulators. There are few biomarkers that can distinguish between the two forms of IBD.

1.9.1.1 Existing clinical and biochemical markers

Standard biochemical markers (C-Reactive protein (CRP), albumin) and haematological markers (thrombocytosis, leucocytosis, high erythrocyte sedimentation rate) are used to assist in the diagnosis of IBD.[224] These markers are non-specific markers of inflammation, and can be normal at the time of diagnosis, especially in UC.[224] A high CRP may be indicative of severe disease and/or complicated disease at the outset. CRP is thought to be superior to ESR, as the latter can be affected by other factors (haematocrit in anaemia and polycythaemia, pregnancy).[225] Other acute-phase proteins (ferritin, Interleukin-6, fibrinogen, α2 globulin, α1 antitrypsin) have been used, but have not superseded CRP.[224]

Serological antibodies directed against self, bacterial or fungal surface antigens have been used in the diagnosis of IBD and to differentiate CD and UC. The anti-*Saccharomyces cerevisiae* antibodies (ASCA) is more commonly expressed in CD compared to UC, especially in ileal disease.[226,227] The perinuclear anti-neutrophil cytoplasmic antibody (pANCA) is more commonly expressed in UC.[226,227] The combination of pANCA and ASCA has been used to differentiate UC and CD with a sensitivity of ~55%, which decreases further when attempting to distinguish colonic CD from UC.[228,229] Other commonly used antibodies include anti-*Pseudomonas flourescens* associated sequence I2 antibodies, anti-outer membrane of porin C (anti-OmpC) and antibodies against bacterial flagellin (CBri1). Such serological markers are relatively specific for IBD, but demonstrate lower sensitivity.

1.9.1.2 Faecal calprotectin

Faecal calprotectin (FC) is able to detect intestinal inflammation with a high degree of accuracy. A study from Edinburgh by Kennedy et al demonstrated that FC was able to discriminate IBD from functional disease with a high degree of accuracy (AUROC=0.97, with a threshold of >50 μg/g demonstrating a 97% sensitivity and 0.74% specificity).[230] FC also has excellent negative predictive value (0.99),[230] and therefore can be useful in deciding which patients require endoscopy.[231] A meta-analysis of 6 adult studies (n=670) demonstrated that FC had a pooled sensitivity and specificity of 0.93 (0.85-.097) and 0.96 (0.79-0.99) respectively for IBD and identified individuals requiring endoscopy for suspected IBD.[232] Notably, FC outperforms the other serological and biochemical markers (e.g. C-Reactive protein) in this context. An alternative faecal biomarker used in IBD is faecal lactoferrin.[233]

1.9.1.3 Using genetics to diagnose IBD

Given the high number of low penetrance genes associated with IBD and the low overall prevalence of the disease within the general population, it is unlikely that genetic testing could be used for population wide-screening of IBD.[227] A genetic risk score has been developed to predict IBD type and location, however the small effect sizes led to low predictive accuracy and (AUC 0.6) even when combined with clinical factors (e.g. smoking) only 6.8% of variance was explained in CD and even less in UC (1.1%).[93] The main utility of such a genetic risk score may be to re-classify misdiagnosed patients.[93]

1.9.2 Prognosis

As detailed above, faecal calprotectin is currently a highly useful currently available biomarker with a high specificity and sensitivity for identifying patients with IBD. Whilst FC is not without its issues (poor patient compliance, non-specificity for other forms of gut inflammation) it is currently positioned as a biomarker for use in the diagnosis of IBD. Currently there are fewer clinically available biomarkers to assist in predicting disease prognosis. This is especially relevant with an increasing armamentarium of new biologic medications in discriminating patients who require such powerful immunomodulation at the outset (top down approach) compared to patients who will have an essentially quiescent disease course in whom the possible side effects of such drugs would outweigh the benefits.

1.9.2.1 Existing clinical and biochemical markers

Standard biochemical markers such as CRP, white-cell count, ESR and albumin are used routinely to monitor disease course. CRP is known to predict patients with a more severe clinical phenotype, including patients at increased risk of requiring surgery.[234,235] Serological markers have been used for prognosticating. Positive ASCA antibodies may increase the risk complications in paediatric CD, including an increased risk of surgery by up to 10%.[224,236] An increasing number of positive serological antibodies to microbial antigens may also predict the likelihood of disease progression to fibrostenotic and penetrating disease.[237–239] Using a panel of four serological markers the number and magnitude of immune response was able to stratify risk of complicated disease from 44.7% to 82%.[240]

A specific scenario where significant research has been directed to predict prognosis in the short term is acute severe colitis (ASUC). In this context, clinical factors (particularly stool frequency) and biochemical markers have been used in scoring systems. Perhaps the best-known scoring system is the Travis Score in which high stool frequency (>3 to ≤8 or >8) and a elevate C-reactive protein(>45mg/L) at day 3 of corticosteroid treatment had an 85% chance of requiring colectomy.[241] A second prominent clinical scoring score is the Ho score where Stool frequency, colonic dilatation on day 3 and hypoalbuminemia on day 1 were defined as independent predictors of failure of corticosteroid therapy (i.e. colectomy). [242] Importantly these scores have both been validated in independent cohorts.

1.9.2.2 Faecal calprotectin

Faecal Calprotectin has been used to predict patients at risk of disease flare or relapse.[243] FC can be used for a surrogate marker of patients in deep remission.[244] FC has also been used in a prospective, protocoled fashion to step up treatment (5-ASA) to prevent relapse in UC (DEAR study).[245] FC may also be used to predict recurrence following ileal resection for CD.[246] Fecal calprotectin has been studied in the context of ASUC, with levels being significantly higher in patients requiring colectomy, and a trend toward higher levels in non-responders to corticosteroids (p=0.08) and infliximab (p=0.06) compared with responders.[247] A more recent analysis of biomarkers in 444 patients with ASUC from the same center failed to replicate the predictive ability of calprotectin (p=0.52).[248]

1.9.2.3 Biomarkers from –omic technologies

1.8.2.3.1 Genetic scores

As previously discussed above, several genetic loci are strongly associated with specific patterns of disease location (*NOD2* - ileal, *MST1* 3p21 - ileal and MHC DRB1*01:03- colonic disease).[93] Variants in the *CARD15/NOD2* loci are associated with earlier disease onset, ileal disease, fibrostenosing behaviour and increased likelihood of surgery.[249–251] A meta-analysis demonstrated that the risk of complicated disease increased when the number of *NOD2* mutations increased from one (8% risk) to two (41% risk).[252] In addition to *NOD2*, several other loci may also be predictive of a more aggressive disease course in CD, including *ATG16L1*, *IL23R* and *DLG5*.[253] In a large multicentre GWAS using the immunochip, *NOD2* was again demonstrated as the most important genetic factor predictive of ileal disease, fibrotic strictures, penetrating disease and need for surgery, with the frameshift mutation being particularly associated with deleterious outcomes.[254] Genetic loci were also able to predict fistulising disease (*IL23R, LOC441108, PRMD1, NOD2*), need for surgery (*IRGM, TNFSF15, C13ORF31, NOD2*) and stenosing disease (*NOD2, JAK2, ATG16L1*) .[254]

The *HLA-DRB*0103* variant has been known for some time to be associated with extensive UC, and increased colectomy requirement.[255–257] The Multidrug resistance (*MDR1, C3435TT allele, ABCB1*) gene locus is also significantly associated with extensive disease in UC.[258] A genome wide association study demonstrated genetic loci (including *HLA, IL12B and TNFSF15*) that were able to predict the need for colectomy.[259] This study generated a risk score based on 46 SNPs to account for 50% of the colectomy risk (AUROC 0.91), together with increased risk of colectomy at 3- and 5 years (Figure 9).[259]

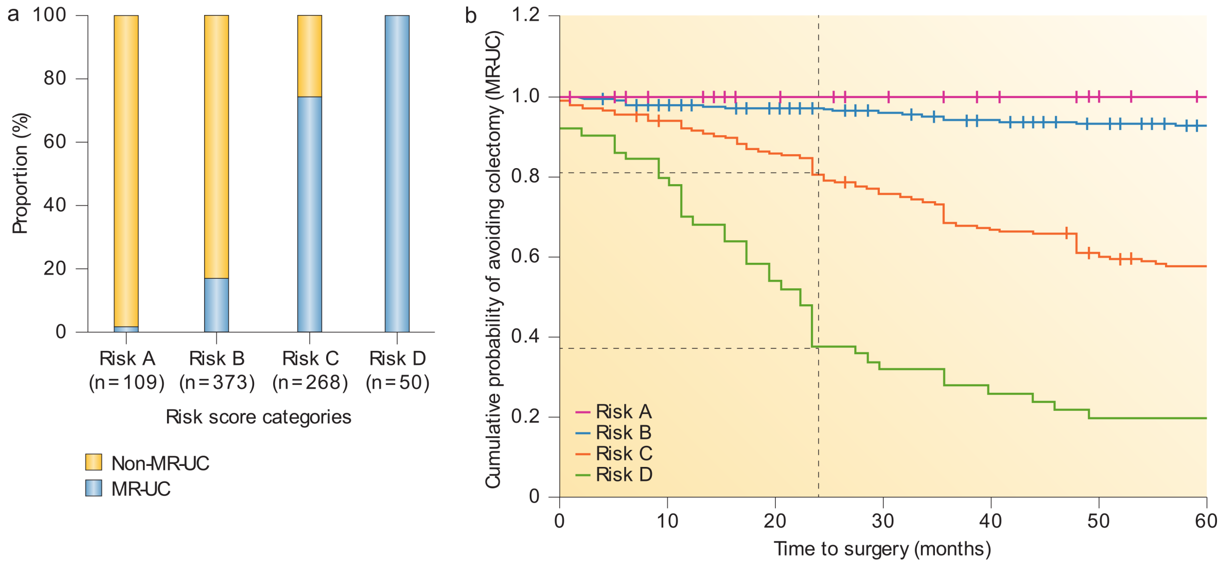


Figure 9 – Genetic risk score to predict IBD prognosis

Haritunians *et al* classified patients into medically refractory ulcerative colitis (MR-UC, n=324) and non-medically refractory UC (non-MR-UC, n=537)[259]. A genome wide association study was performed and a combination of 46 single nucleotide polymorphisms accounted for 46% of the variance of risk of colectomy. Based on genetic data, four genetic risk scores were devised (A to D, A-lowest risk, D-highest risk). The risk in each group of colectomy was 0%, 17%, 74% and 100%. (Reprinted by permission from Nature Publishing Group: Nature reviews Gastroenterology and Hepatology (Gerich, M. E. & McGovern, D. P. B. Nat. Rev. Gastroenterol. Hepatol. 11, 287–299 license number 3742031148208)

Given the large number of low penetrance genes associated with IBD susceptibility, together with the significant contribution of other factors including epigenetics, gut microbiota and the environment, genetic markers in isolation are unlikely to adequately predict disease course in ASUC.[227] Recent advances in other clinical aspects of IBD genetics give cause for optimism, particularly 'pharmacogenetics', for example *TMPT* genotyping prior to initiation of thiopurine medication and the identification of specific HLA variants linked with thiopurine-induced pancreatitis.[121,122]

1.9.2.3.2 -omic scores

Following on from genetics, transcriptomics, or the study of gene expression, is another emerging field in biomolecular research. An elegant study from Lee *et al.* in Cambridge (UK),

demonstrated that the gene expression profile of circulating CD8$^+$ T-lymphocytes is able to accurately predict a relapsing disease course from a stable one in patients with newly-diagnosed IBD.[260] The transcriptome has been studied in pediatric ASUC, with 41 genes being differentially expressed in those requiring second line medical therapy or colectomy.[261]
Several genes overexpressed in non-responders interacted with steroid treatment on pathway analysis (*CEACAM1*, *MMP8*).[261] Ten of the 41 genes were predictive of non-response with a sensitivity and specificity of 80%.[261] Interestingly, there was differential expression of *ABCC4*, a gene in the same superfamily as the aforementioned *MDR1* gene.[261]

Recently, there has been fervent interest in the role of the gut microbiome in disease pathogenesis. Microbial dysbiosis, including a reduced microbiological diversity has been noted in patients with IBD.[262] A study of pediatric ASUC, albeit in small numbers of children, indicate a reduced number of phylospecies compared with control.[263] A significant reduction in phylospecies was also seen in those who failed first line medical treatment.[263]

The other emerging exciting biomolecular disciplines such as metabolomics,[264–266] proteomics, glycomics,[267,268] and epigenetics[144] may also unearth promising markers, used to predict disease course in the future.

1.9.2.3 Composite scoring systems

Given the complex, multifactorial nature of IBD pathogenesis, it is unlikely that a single clinical or biochemical variable can be used in isolation to predict disease course. It is more likely that composite scores make up of a combination of clinical, genetic and other data will be most fruitful in assisting prognostication of disease course. In Crohn's disease a combination of clinical, genetic (*NOD2* status) and serological data(ASCA-IgA, ASCA-IgG, anti-OmpC, anti-CBir1, anti-I2, pANCA) more accurately predicted progression to stricturing or penetrating disease behavior compared with using any single parameter alone (AUC 0.8).[269] Another study used clinical and genetic data to predict requirement for surgery in CD, with progression to surgery faster in those with both clinical and genetic factors (IL12B).[270]

1.10 Thesis aims

The GWAS era has clearly demonstrated that IBD has a strong genetic association, however the proportion of disease variance explained by genetics alone is relatively small (13.1% in CD and 8.2% in UC.[86]). Attempting to understand this 'missing' heritability in IBD is high on the current research agenda. Together with the known genetic contribution, IBD has important environmental risk factors including smoking, diet and the gut microbiota. Epigenetics may provide an interface between genetics, the environment and disease. Two of the epigenetic mechanisms; DNA methylation and microRNAs will be the subject of investigation in this project. The purpose of epigenetic research will be twofold; firstly an attempt to gain greater understanding of IBD pathogenesis and secondly to identify potential biomarkers for IBD diagnosis and stratifying patients at risk of a more severe disease phenotype.

1.11.1 Hypothesis

The hypothesis of this thesis is that there are site-specific DNA methylation differences between patients with IBD and controls, and that these differences may be related to germline variation and these in combination may lead to differences in gene expression. A secondary hypothesis is that differences in DNA methylation at specific loci, may be related to disease pathogenesis either in a cause (through altered gene expression) or effect manner. Whilst implicating epigenetic mechanisms in disease pathogenesis is very difficult without obtaining samples prior to diagnosis of IBD (e.g. cord blood samples/Guthrie cards or using an approach adopted in the GEM study (http://www.gemproject.ca/)), specific DNA methylation and microRNA marks may provide compelling biomarkers. A major challenge confronting epigenetic research has been to integrate genetic and gene expression data. Disease-associated SNPs may lead to differential DNA methylation, and consequently DNA methylation may be the mechanism by which certain SNPs confer additional risk of IBD susceptibility. Additionally methylation quantitative trait loci (meQTLs) have been identified. Attempting to establish a link between germ-line variation and DNA methylation provides the rationale for studying genetics within this study. A second major hurdle for epigenetic research has been to establish a relationship with gene expression. DNA methylation in promotor regions is associated with gene silencing. However the relationship between DNA methylation and expression is likely to be complex.

1.11.2 Specific aims

The specific aims of this project are:

1) To compare genome wide site-specific DNA methylation in peripheral blood in IBD cases and controls (Chapter 3). Genome-wide methylation is examined in a large cohort of newly diagnosed cohort of IBD patients using the Illumina 450K platform. The 'methylome' of a subset of patients is characterised in detailed by performing cell separation of specific leukocytes (CD4+, CD8+ T-cells and CD14+ monocytes).

2) To validate and replicate the genome-wide DNA methylation results generated in Chapter 3 (Chapter 4). A targeted approach using pyrosequencing is performed to both technically validate the technique and replicate findings in an independent cohort of established IBD cases. Data from Chapter 3 are also correlated with our previous DNA methylation findings in childhood-onset Crohn's disease.

3) To perform genome wide profiling of individuals included in DNA methylation analyses and understand between genetic and epigenetic (DNA methylation) factors (Chapter 5). A genome-wide association study is performed. Furthermore, the association between quantitative traits (DNA methylation, gene expression) is assessed in relation to germ-line variation (meQTLs, eQTLs).

4) To perform genome wide gene expression profiling of individuals included in DNA methylation analyses and attempt to understand the relationship between DNA methylation and gene expression. Data from complementary genome-wide (microarray) and targeted (qPCR) approaches to assay gene expression in the most significant DMRs/DMPs identified in Chapter 3 is presented. Furthermore, an in-silico technique that correlates promotor region DNA methylation levels and gene expression within established gene networks is applied to further delineate this complex relationship between methylation and expression.

5) To utilise genome wide DNA methylation data to discriminate IBD cases from controls and to identify methylation profiles predictive of a more severe clinical phenotype (Chapter 7). Linear discriminant analysis is used to identify two-probe methylation markers that can discriminate between IBD cases and controls. An unsupervised clustering approach is

performed to identify specific methylation profiles associated with a more severe disease course (need for surgery and immunomodulatory drugs).

6) To compare microRNA expression levels in circulating peripheral blood leucocytes (CD4+, CD8+ and CD14+ cells) in IBD cases and controls (Chapter 8). Small RNA sequencing is exploited to provide a comprehensive overview of cell-specific miRNA expression in circulating leukocytes. Differential expression of miRNAs in patients with CD and controls is explored. MiRNA gene targets and downstream pathways are explored.

Chapter 2. Methods

2.1 Patients recruitment

2.1.1 Patient selection

Suitable newly diagnosed and treatment naïve IBD patients and controls were pre-identified prospectively from clinic and endoscopy lists. In order to recruit adequate numbers of treatment naïve patients, patients were targeted pre-diagnosis, with symptoms (bloody diarrhoea, weight loss) suggestive of IBD. Postal patient information sheets were sent prior to index hospital appointment. Patients whose investigations were normal were classified as non-healthy symptomatic controls. Additionally patients with an existing diagnosis of IBD were recruited using the same method. Healthy controls were first approached by email message to University of Edinburgh mailing list.

2.1.2 Ethics

Patient involvement was undertaken within the SAHSC BioResource framework with ethical approval from the Tayside committee on Medical Research B (10/S1402/33). At clinic appointment patients were counselled and written informed consent was obtained using the SAHSC BioResource consent form. In addition, the retrospective patient samples and controls used in this project were collected using the following ethical permission: Dundee ethics [Tayside Ethics committee 226/02] and Edinburgh Ethics [Lothian Ethics committee 2000/4/192].

2.1.3 Phenotypic data acquisition

A questionnaire was administered to each patient and venepuncture was undertaken at the same time as routine clinical blood sampling. Following recruitment, patients were followed to index investigations (e.g. colonoscopy), where a clinical, radiological, endoscopic and histopathological diagnosis of IBD was made according to Lennard-Jones criteria.[271] Disease location and behaviour was classified according to the Montreal-classification.[20] The following phenotypic data was collected on prospectively recruited patients:

Diagnosis, sex, date of birth, race, smoking status at diagnosis (non, current, ex), smoking start date, smoking stop date, number of cigarettes per day, diagnosis year, family history of IBD, joint association/arthritis, oral steroids since diagnosis, intravenous steroids since diagnosis, anti-TNFα or ciclosporin treatment since diagnosis, immunomodulator since diagnosis, Montreal disease location, behaviour (CD), Crohn's disease surgery, year of first surgery, number of operations, ulcerative colitis disease extent, ulcerative colitis surgery, surgery year, reason for UC surgery (dysplasia, acute, chronic disease), town of birth, rural/urban location, parents town of birth, where majority of childhood growing up, number of years in secondary education, people smoking in house growing up, partner/else smoking in home currently, alcohol exposure (units per week), type of alcohol consumed, vaccination history, pet exposure whilst growing up, growth delay/puberty delay, detailed IBD family history included of the following medical conditions (Psoriasis, ankylosing spondylitis, coeliac disease, colorectal cancer, any cancer, multiple sclerosis), other medical conditions, symptoms of IBD prior to diagnosis, duration of symptoms, weight loss, admissions with IBD, abdominal operations, operations for IBD, tonsillectomy, appendectomy, medication history, diet, pregnancy since diagnosis, contraceptive pill use, NSAID use, aspirin use, paracetamol use, antibiotic use, weight, height, bowel prep type, disease activity score (CDAI), health professional global assessment, medication/surgery/endoscopy/histology/radiology case note review.

Additional assistance with clinical phenotyping and follow up data was obtained from database manager Hazel Drummond.

2.1.4 Patient samples

Blood sample collection was undertaken at the same time for research and clinical samples to minimize patient discomfort from multiple venepunctures. A 21 gauge butterfly needle (Greiner , Frickenhausen, Germany) with 30cm safety tube with Luer lock device was used for venepuncture. Blood samples were taken in the following tubes: 9ml Z Serum clot activated vacuette (Greiner, Ger), 9ml K3 EDTA vacuette (Greiner), 9ml vacuette Tempus blood RNA (Greiner, Ger) and PAXgene blood RNA tubes (BD, NJ, USA). Serum tubes were taken first (including before clinical samples) to prevent reagent contamination from other blood tubes (e.g. EDTA).[272] Tempus and PAXgene tubes are effective in preventing RNA degradation.[273] To prevent backflow of potentially toxic AB- RNA stabilisation reagent from Tempus vacuette tubes and PAXgene tubes the following measures were implemented: butterfly with 30cm

safety tube; patients arm held in downward position; tube held with the cap uppermost; tourniquet release after blood seen to flow into the tube; and by avoiding contact of the reagents with the tube cap during venepuncture. When using Tempus vacuette tubes blood was filled to minimum collection line (draw volume) and inverted several times for 10-15 seconds immediately after collection. Blood tubes were stored at 4 °C prior to processing. EDTA tubes were stored at -80 °C prior to DNA extraction. The 9ml Tempus vacuette tube PAXgene tubes were kept upright at room temperature for 2 hours to ensure complete lysis of blood cells prior to long term storage at -80 °C later RNA extraction respectively.
Additional assistance with sample collection was obtained from clinical research nurse Linda Smith, and clinical fellows Drs Rahul Kalla, Ray Boyapati and Nick Kennedy.

2.1.5 Serum collection

The 9ml Z Serum clot activated vacuette tube (Greiner, Ger) was stored at 4 °C for at least one hour prior to centrifugation to allow blood coagulation. The tube was subsequently centrifuged at 2,500 × g/ RCF (relative centrifugal force) for 15 minutes at 4 °C. Serum was aliquoted into 1500 µL microcentrifuge tubes (Greiner, Ger) with care not to disturb the clot and stored at -80 °C. Heavily haemolysed samples were discarded. The collection method and times of blood draw, centrifugation and -80 °C storage were recorded contemporaneously on the Edinburgh IBD database.

2.2 Cell separation

2.2.1 Peripheral blood mononuclear cells

A Ficoll-Paque (GE healthcare, Bucks, UK) separation was used to obtain peripheral blood mononuclear cells (PBMCs) from whole blood.[274–277] Ficoll-Paque contains Ficoll PM400, a synthetic high molecular weight polymer (containing sucrose and epichlorohydrin) and sodium diatrizoate.

A between of 18 mL and 36 mL of EDTA buffered blood was used for Ficoll separation, and processed within 3 hours of venepuncture. Blood was diluted in a 50 mL falcon with cold (2-4 °C) 2nM EDTA and PBS in a 3:1 ratio. The higher the dilution, the greater the purity of PBMCs, as during aggregation some mononuclear cells became trapped in the clumps. More than one

Ficoll separation was performed if the blood volume was greater than 18mls to ensure adequate dilution (using several 50 mL falcon tubes). The diluted blood was carefully layered on top of 15 mL of Ficoll in a 50 mL conical falcon. A 25 mL pipette was used to gently layer the blood down the side of the tilted Ficoll containing falcon. The solution was centrifuged at 400 × g for 40 minutes at 20 °C in a swinging bucket rotor with no brake. Following this spin, the PBMCs were located at the interface between Ficoll and plasma, with the erythrocytes and granulocytes forming a pellet at the base of the tube (Figure 10).

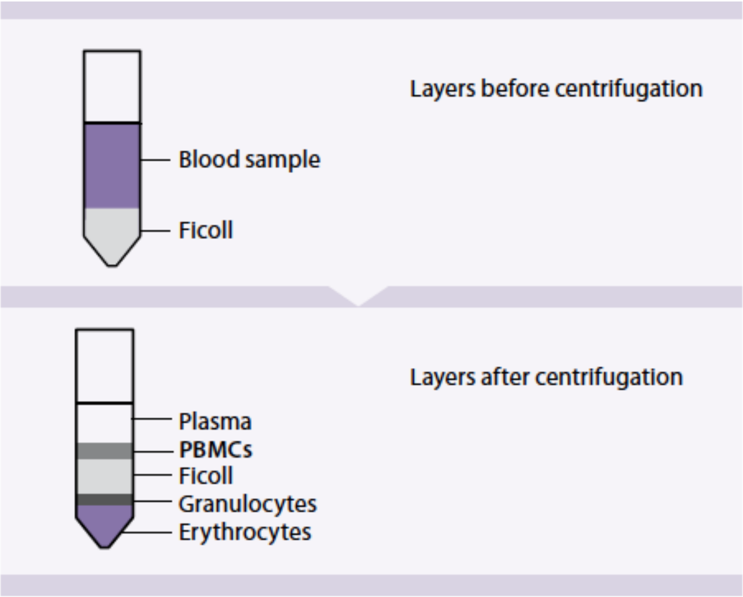


Figure 10 - Diagrammatical representation of Ficoll separation before and after centrifugation (taken from Greiner data sheet) [274]

The PBMCs were carefully aspirated using a Pasteur pipette leaving the interface undisturbed. The PBMCs were placed within a new 50 mL conical tube and topped up with PBS/EDTA buffer and centrifuged at 300g for 10minutes. The supernatant was removed and the pellet resuspended in 50mL of PBS/EDTA buffer (2 mL of 0.5nM EDTA in 500 mL of PBS). Two platelet-removing spins (200 × g for 10 minutes at 20 °C, supernatant [containing platelets] removed at each step) were performed to increase the purity of the PBMCs. The subsequent pellet was resuspended in 10 mL of PBS/EDTA buffer and 20μL was placed on a haemocytometer cover slip for cell counting. Cells were counted using the automated haemocytometer (PBMC fresh program (medium sized cells) at a dilution of 10). Following automated count, each quadrant was manually adjusted to include or exclude cells miscounted by the haemocytometer. Fifty microliters of PBMC solution was also reserved for flow cytometric analysis of lymphocyte differential counts. The PBMC suspension was again spun at 300 × g for 10 minutes to pellet the PBMCs, which were then used for magnetic labelling.

Following removal of PBMCs from the Ficoll suspension, the plasma and Ficoll was aspirated leaving the red cell/granulocyte pellet. This pellet was resuspended using 50 mL of red cell lysis buffer and kept on ice for 10 minutes. The lysis buffer contained 1000 mL dH20, 8.3 g of NH_4Cl, 1.0 g of KHO_3 and 1.8 mL of 5% EDTA and was filtered sterilised following production (20μm Minisart (16532) single use filter unit). Lysis was considered complete when the solution turned translucent. The solution was the centrifuged at 300 × g for 10 minutes at 20 °C. The supernatant was removed and further spins were performed as above to remove platelets (2 × 200g for 10 minutes at 20 °C). The resultant pellet was resuspended in 10 mL of PBS/EDTA buffer for cell counting. Cell counting was performed in a similar manner as above. As the cell count usually was greater than 1×10^7, the 10ml solution was usually split to obtain a smaller volume containing 1×10^7 cells. This small volume was centrifuged (300 × g for 10 minutes) to obtain a pellet. The supernatant was removed and the pellet was covered in 40 μL of RNAlater (Qiagen) and stored at -20 °C for DNA and RNA extraction.

2.2.2 Nanobead Immunomagnetic cell separation

Cell separation was performed in 2 stages according to the algorithm in Figure 11. According to PBMC cell numbers, the PBMCs were split into two or three (corresponding to parts 1 and 2 on Figure 11). Cells were resuspended in 80 μL per 1×10^7 cells in AutoMacs running buffer (Miltenyi, Germany 130-091-221) [scaled proportionately according to predetermined cell number]. AutoMacs running buffer (Miltenyi, Germany 130-091-221) consists of phosphate buffered saline (PBS) [pH 7.2] and 0.5% bovine serum albumin. Cells were manually labelled using 20 μL per 1×10^7 cells with microbeads of appropriate CD antigen (human CD14 microbeads [130-050-201] or human CD8 microbeads[130-045-201]). Following mixing, cells and microbead solutions were incubated for 15 minutes at 4 °C. The cells were washed with 1 mL of running buffer per 1×10^7 cells and centrifuged at 300 × g for 10 minutes at 4 °C (aspirate supernatant). The cells were resuspended in 500 μL of running buffer before proceeding to automated cell separation using the autoMACS cell separator.

Immunomagnetic cell separation was performed using the autoMACs Pro cell separator (Miltenyi, Germany 130-092-545) using magnetic autoMACs columns (130-021-101), which were changed every 14 days (or 100 separations). The 'Possel' function was selected on the AutoMacs Pro to perform a positive cell selection. When the cell suspension passes through the magnetised column, the labelled cells (e.g. CD14+) are magnetically retained within the column. The unlabelled cells pass through the column and are eluted into the negative fraction, deplete of (e.g. CD14-) cells. The magnetic field is then removed from the column, allowing the positive fraction of labelled cells (e.g. CD14+) to be eluted. During the first stage, the CD14+ and CD8+ cells were eluted, and from both positive and negative fractions of each, 20 μL was placed on a haemocytometer cover slip for cell counting and 50 μL was reserved for flow cytometry. The positive and negative cell solutions both were centrifuged (300 × g for 10mins) to obtain a pellet. For the positive fraction (i.e. CD14$^+$ and CD8+ cells), the supernatant was removed and the pellet covered in 40 μL of RNAlater and stored at -20 °C for DNA and RNA extraction.

The negative fraction was used for stage 2 of the separation. In a similar manner to before, cells were counted and re-suspended in 80 μL per 1 × 10^7 cells in AutoMacs running buffer. From the CD14- fraction, cells were manually labelled with CD4+ microbeads (20 μL per 1 × 10^7 cells). The CD4 antigen is expressed on a proportion of monocytes (CD14+, CD4+), albeit at lower levels than CD4+ T helper cells, it is therefore important that the CD14 positive selection was performed first. Thus in the CD14- fraction, subsequently CD4 positive selection should contain almost exclusively CD4+ T helper cells (dendritic cells also express CD4 antigen at low levels).

CD19+ microbeads were used on the CD8- fraction. The process of incubation, washing, re-suspension and subsequent autoMACS separation was the same as above. The positive fractions containing CD4+ and CD19+ cells were counted, sampled for flow cytometry, centrifuged and the pellet was covered in RNAlater and stored at -20 °C.

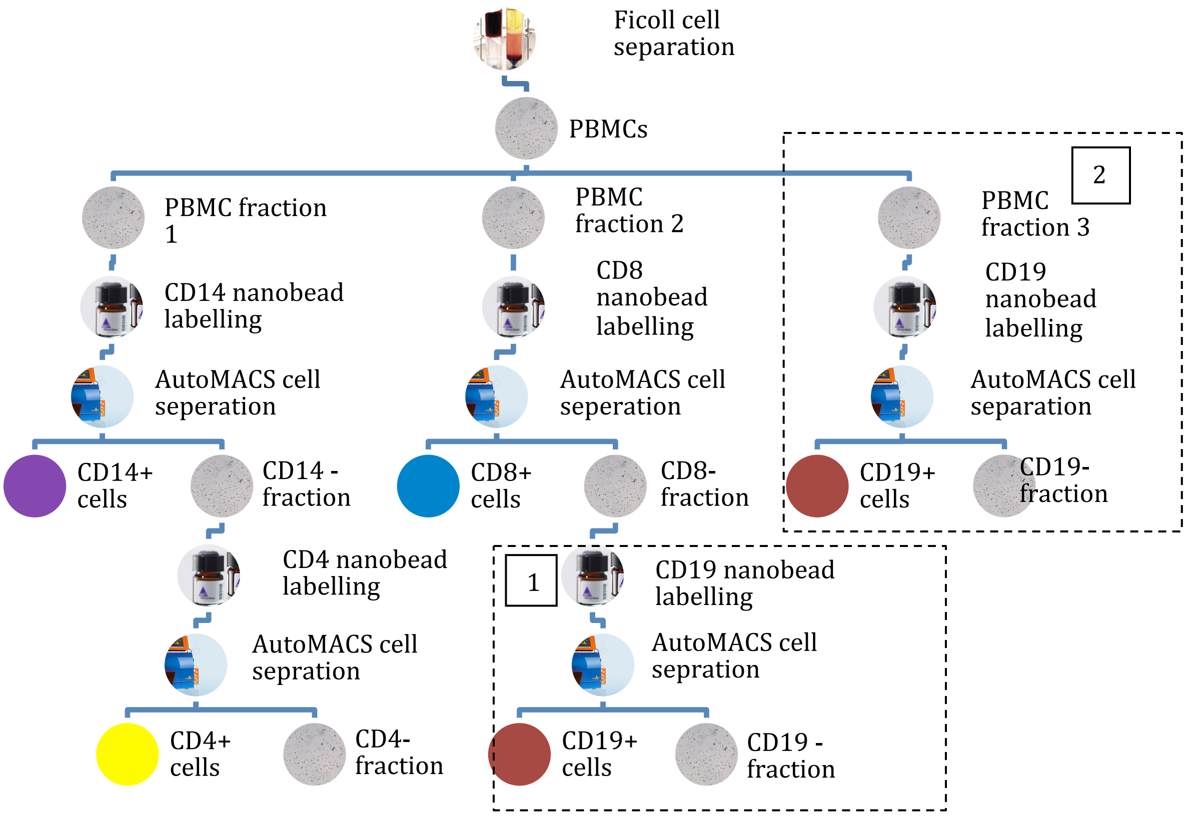


Figure 11 – Flow chart of immunomagnetic cell separation. CD19+ cells were either derived from the CD8- fraction (1) or by splitting the initial PBMC fraction into three (2). PBMC = peripheral blood mononuclear cells.

2.3 Flow cytometry

AutoMACS separated lymphocytes were evaluated using fluorescent antibody staining. PBMCs obtained from Ficoll separation were also evaluated for purity and lymphocyte cell count differential. Cells in 50 µL of buffer were stained with 5 µL of the appropriate antibody (see Table 3). After mixing, antibodies and cells were incubated for 10 minutes in the dark at 4 °C (in the refrigerator). Cells were washed with 1ml of buffer and centrifuged at full speed (14,000rpm) for 10 minutes, and the supernatant removed. Cells were re-suspended vigorously in 200 µL of 4% paraformaldehyde (to prevent clumping). Fixed cells were stored in foil in the dark at 4 °C for flow cytometric analysis within 7 days. CD45RO PE was additionally used for CD4+ cells to determine the proportion of activated CD4 cells. Flow cytometry was performed on FACS Aria II (BD) with assistance from Ms Elisabeth Freyer.

Table 3 – Flow cytometer FACS panel for separated cells

Cell type	Antibody
PBMCs stained	CD4 FITC, CD14 APCVio770, CD8 VioBlue, CD20 APC
PBMC unstained	None
CD4+	CD4 FITC, CD14 APC Vio770, CD45RO PE
CD14+, CD14-preCD4	CD4 FITC, CD14 APC Vio770
CD8+, CD8-preCD19, CD19+	CD8 VioBlue, CD20 APC

2.4 Nucleic acid extraction

2.4.1 Whole blood RNA extraction

Whole blood RNA was extracted using the QIAGEN QIAamp Blood mini kit (Qiagen), which extracts RNA greater than 200 nucleotides long (mostly mRNA). To minimize RNA degradation blood was processed within 3 hours and stored at 4 °C within this time. A total of 1.5 mL of fresh whole blood was taken from one 9ml EDTA vacuette tube and placed into a 15 mL falcon tube. As the columns are limited to 1 × 10^7 cells, less than 1.5ml of blood was used from samples of patients with a significant leucocytosis. A total of 7.5 mL erythrocyte lysis was added and stored on ice for 15 minutes until red blood cells were lysed and the solution looked translucent (vortexed up to twice to assist lysis). A Leucocyte pellet was recovered by centrifugation at 4,000 × g for 10 minutes. The supernatant was carefully removed with pipette. A further resuspension of the leucocyte pellet in 3 mL of EL lysis buffer and centrifuging the sample at 4,000 × g for 10 minutes was performed to lyse any remaining erythrocytes. The supernatant was again carefully removed. The leukocytes were lysed with 600 µL RLT buffer with 1:100 β Mecaptoethanol. The plasma membranes and organelles of the leucocytes are effectively lysed by this rapidly denaturing solution, which also inactivates RNases, and allows the recovery of intact RNA. The lysate was pipetted several times to disperse any clumps before being transferred to the QIAshredder spin column. The QIAshredder column was placed in the microcentrifuge and spun at maximum speed (14,000rpm) for 2 minutes. The QIAshredder column was discarded and 600 µL of 70% ethanol was added to the flow through to adjust the binding conditions. 600 µL of the sample was transferred to the QIAamp skin column. The total capacity of the column is 750 µL so usually two steps were required to process the whole sample through the column. The QIAamp column contains a silica-based membrane to which the RNA binds when centrifuged at 10,000 rpm for 15 seconds. The contaminants were washed with 700 µL of RW1 buffer (contains ethanol and a guanidine salt, centrifuged at 10,000 rpm for 15 seconds) and two 500 µL RPE buffer washes (removes salts, centrifuged at 10,000 rpm for 15 and 180 seconds). At each stage the flow through was discarded and a new collection tube was used, care was taken that the flow through did not come into contact with the column membrane. A further spin (maximum speed (14,000 rpm) for 1 minute) of the column was performed to dry the membrane. The RNA was eluted by adding 30 µL of RNase free water to the membrane and centrifuging at 10,000rpm for 1 minute. 1.5 microliters of total blood RNA was quantified using a nanodrop spectrophotometer. The absorbance is measured at 260 nm after blanking with RNase free water. The average of three readings was used.

2.4.2 Separated cell DNA, RNA, microRNA extraction

The separated cell nucleic acids were extracted using the QIAGEN Allprep DNA/RNA miRNA universal kit (Qiagen), which extracts RNA molecules greater than 18 nucleotides long. Cells were lysed using 600 µL RLT buffer plus with 1:100 2,β-mecaptoethanol together with mechanical disruption using vigorous pipetting. The lysis buffer, containing guanidine isothiocyanate, causes immediate denaturing of DNases and RNases, preserving DNA and RNA for isolation. If there was visible evidence of incomplete lysis (clumps etc. – occasionally the case for PBMC, granulocyte and CD14+ pellets), cell lysates were passed through a QIAshredder column, centrifuged at maximum speed (14,000 rpm for 1 minute).

2.4.3 Genomic DNA isolation

The lysate was passed through an AllPrep DNA minispin column (30 seconds at 14,000 rpm) which (together with the high salt concentration of the buffer) causes binding of DNA to the column membrane. The membrane was washed once with buffer AW1 (350 µL, 15 seconds, 14,000 rpm). Proteinase K was applied directly (20 µL) to the membrane together with 80 µL of AW1 buffer, causing digestion of cell proteins. A further wash was performed using 500 µL of buffer AW2. Elution buffer (EB, 100 µL), pre-heated to 75 °C, was applied directly to the column membrane, incubated for 1 minute, and centrifuged at 10,000 rpm for 1 minute to elute DNA.

2.4.4 Total RNA extraction

The flow-through from the AllPrep DNA column was used for subsequent Total RNA (including microRNA) extraction. Proteinase K (80 µL) was used for digestion of proteins in the flow-through, when incubated with ethanol (350 µL) for 10 minutes at room temperature. A further 750 µL of ethanol was added, and up to 700 µL of the solution was pipetted into a RNAeasy mini-spin column and centrifuged (maximum speed 14,000 rpm for 15 seconds) in three stages (as maximum capacity of RNAeasy mini-column is 700 µL). At this stage RNA and microRNA was bound to the spin column membrane (flow through discarded after each spin). Following a wash step (500 µL RPE for 15 seconds at maximum speed), an on-membrane digestion step was performed using DNAse I (10 µL) and RDD buffer (70 µL), which was incubated at room temperature for 15 minutes. The DNAse I was stored at -20 °C, and when thawed, was carefully mixed with RDD buffer, as it is liable to physical denaturing with over vigorous pipetting. DNAse I digests any residual DNA within the sample, helping to purify the RNA. Following incubation, 500 µL of buffer FRN (isopranalol added) was used to wash the column (maximum speed 14,000 rpm for 15 seconds) and also elutes microRNA. Unlike previous steps, where flow through was discarded, here the flow-through (containing small RNAs) was reapplied to the column membrane. A further spin (maximum speed 14,000 rpm for 15 seconds) binds small RNAs to the membrane. A further two washing steps were performed with buffer RPE (500 µL, maximum speed 14,000 rpm for 15 seconds) and 100% ethanol (500 µL , maximum speed 14,000rpm for 2 minutes). The column was transferred to an empty Eppendorf tube and centrifuged at maximum speed 14,000 rpm for 1 minute to dry the column and membrane. The RNA and microRNA was eluted in two stages using 30 µL of RNAse free water. The RNA sample was stored on ice until RNA amounts are quantified using the methods described in section 2.5 and stored at -80 °C.

2.4.5 RNA isolation from Tempus blood tubes

Total RNA (including microRNA) was extracted from 9ml vacuette Tempus blood RNA (Greiner) using the MagMax for Stablized Blood tubes RNA Isolation kit (Ambion, Life technology). Tempus tubes were thawed from frozen (stored at -80 °C) for 30 minutes on ice. The contents of the Tempus tube (9mL) were decanted into a 50 mL conical tube, with the remaining residue washed from the tube using 3 mL of 1X PBS, to make a total of 12 mL. If the contents of the Tempus tube were less than 9 mL (i.e. less than 3 mL of blood sampled), the difference was made up by adding extra PBS, to ensure the total final volume equalled 12 mL. The samples were vortexed at high speed (1800 × g) for 30 seconds, before centrifuging at 5000 × g for 15 minutes at 4 °C. The supernatant is removed and 4 mL of Pre-Digestion wash was added, mixed using the vortex, before pelleting the crude RNA (5000 × g for 10 minutes at 4 °C). The supernatant was again removed, and the 50 mL conical tubes were inverted on paper towel to remove any remaining liquid. The following steps were performed to digest protein and DNA within the sample. The sample pellet was resuspended with resuspension solution (117.5 µL) and proteinase (2.5 µL) was added. The sample was mixed using the vortex at gentle speed, and the sample was moved from the 50mL conical tube to a 48 well plate. Ten microliters of TurboDNase was added and mixed using an orbital shaker. The RNA binding beads were bound to the crude RNA pellet for magnetic capture. Twenty microliters of well mixed RNA binding beads were added to the sample well, together with 50 µL of binding solution concentrate. The 48-well plate was mixed on the orbital shaker for 1 minute, before adding 200 µL of 100% isopranolol. The sample plate was placed onto a magnet, which captures the RNA binding beads into a pellet (capture time 1-3 minutes) and the supernatant carefully removed with care taken not to disturb the magnetic bead pellet. The sample plate was removed from the magnet and washed twice with washing solution (2x 150 µL of Wash Solution 1, applied to each sample/well, mixed on the orbital shaker, before repeat magnetic capture and removal of supernatant, the same process was repeated twice with Wash Solution 2). The RNA binding beads were dried by placing the plate on an orbital shaker at room temperature for 2 minutes (or longer if there is remaining liquid/wash solution after 2 minutes). To elute the RNA, 80 µL of elution buffer was added to the sample well, mixed on the orbital shaker for 4 minutes, before magnetic capture (magnetic capture time 3 minutes). The eluted RNA solution was aspirated (with care taken not to disturb/aspirate the beads) and transferred to an RNase-free container, quantified using methods above, and stored at -80 °C.

2.4.6 Total RNA extraction from PAXgene tubes

Total RNA, including microRNA, was extracted form whole blood PAXgene tubes using the PAXgene blood miRNA kit (PreAnalytix, Switzerland). PAXgene tubes were thawed at room temperature and allowed to equilibriate at room temperature for up to 2 hours to allow complete cell lysis. The PAXgene tubes were centrifuged at 3200 × g for 10 minutes (centrifuging at 5000 × g led to tubes breaking). The supernatant was decanted and 4 mL of nuclease free water was added and the pellet resuspended using a vortex (New supplied haemaograd closure device was used to stop the tube). The sample was centrifuged again at 3200 × g for 10 minutes. The supernatant was decanted, and the top of the tube wiped with paper towel, and any remaining supernatant removed using a pipette (incomplete removal of supernatant results in diluted lysate and affect RNA binding to columns). To each sample, 350 µL Buffer BM1 was added and the pellet resuspended using the vortex. The sample was transferred to a 1.5 mL Eppendorf tube, and 300 µL of Buffer BM2 and 40 µL of Proteinase K was added. The sample was mixed using the vortex and incubated at 55 °C in a preheated shaker incubated at 1000 rpm. Following this incubation, samples were transferred into a QIAshredder spin column and centrifuged at max speed (15,000 × g) for 3 minutes. The supernatant was transferred to a new 1.5 mL microcentrifuge tube and 700 µL of isopranolol was added. Following mixing, 700 µL of the sample was transferred to a PAXgene RNA spin column and centrifuged for 1 minute at 15,000 × g. The flow through was discarded and the step repeated with the remaining sample, with the flow through being discarded at each step. The column was placed in a new 2 mL tube, and 350 µL of buffer BM3 was added to the column and centrifuged at 15,000 × g for 15 seconds, with the resultant flow through being discarded. The on-column DNA digestion was performed by adding DNAse I (10 µL per sample) to buffer RDD (70 µL) to the column filter and incubating at room temperature for 15 minutes. To the column, 350 µL of buffer BM3 was added and centrifuged at 15,000 x g for 15 seconds, and the resultant flow through being discarded. Two wash steps were performed: 500 µL of buffer BM4 (centrifuged at 15,000 × g for 15 seconds) and again 500 µL of buffer BM4 (centrifuged at 15,000 x g for 2 minutes). The column was transferred to a new tube and centrifuged at 15,000 × g for 1 minute to dry the column. The column is placed in a new 1.5ml nuclease free microcentrifuge tube and 40 µL of elution buffer (BR5) was added directly to the column membrane, and centrifuged at 10,000 × g for 1 minute. The previous step was repeated to elute in a total of 80 µL. To denature the RNA, the sample was incubated at 65 °C for 5 minutes, before nanodrop concentration estimation and stored at -80 °C.

2.5 Nucleic acid quantification and quality assessment

2.5.1 Nanodrop spectrophotometer nucleic acid quantification

Nanodrop quantification was performed using the NanoDrop 1000 (Thermo Scientific). Onto the open arm on the device, 1.5 µL of the elution medium (RNAse free water or elution buffer depending on protocol) was placed on the pedestal. When the arm is closed a column forms between the arm and pedestal as a result of surface tension between the two surfaces. The arm auto-adjusts the distance from the pedestal, creating two pathlengths for optimal measurement (measures @ 0.2mm and 1mm). A xenon flash lamp passes through the sample which is detected on the other side of the sample (CCD detector [charge coupled device]). Both surfaces were wiped before the application of the next sample. The process is repeated with RNAse free water to 'blank' the sample. DNA and RNA (1.5 µL) samples were the loaded in the same way and quantified accordingly.

2.5.2 QuBit flurometery nucleic acid quantification

The QuBit flurometer uses molecular probe dyes to quantify nucleic acid/protein concentration of a sample, in this case the dsDNA high sensitivity Assay dye. The DNA-specific dye fluoresces when bound to the specific molecule, in this case DNA (but not RNA or protein). The dye binds to DNA by an intercalation of bases, after which it confirms to a different (more rigid) shape and fluoresces intensely. The flurometer detects the level of fluorescent signal and coverts it into a concentration. The flurometer uses two standards (one low concentration and one high concentration) to create a standard curve. A master mix of 200 µL of buffer and 1 µL of dye per sample was prepared (4 pool samples and 2 standards = 1.2 mL), and 190 µL of this mix was used for each standard and 199 µL for each sample. Ten microliters of each standard and 1 µL of each pooled sample was added to respective tubes. The samples were incubated for 2 minutes before assessment on the QuBit flurometer.

2.5.3 Electrophrenography to assess RNA and microRNA quality

RNA integrity is determined using Agilent 2100 bioanalyzer (Agilent, Germany) and Pico LabChips (Agilent, Germany). The ladder was pre-prepared by denaturing the ladder at 70 °C, cooling and adding 90 µL of RNAse free water. Prior to running the assays, the electrodes were cleaned for 5 minutes using 350 µL of RNAse free water. Prior to preparing the gel, reagents were allowed to equilibrate at room temperature for 30 minutes.

2.5.4 Agilent RNA 6000 Pico chip

A 1.5 µL aliquot of the RNA sample was added to a small 0.7 mL tube and denatured by placing on the agitating heat block at 70 °C for 2 minutes and subsequently on ice. The ladder RNA (stored at -80 °C) is also denatured (once only) in the same way.

Preparing the Gel

65 µL of filtered gel was placed into a nuclease-free microcentrifuge tube. To make up the gel mix, 1 µL of Pico dye was added to aliquot of 65 µL of filtered gel. The sample was vortexed, taking care not to create bubbles (the solution is highly viscous) as the chip works by microfluidics and any bubbles can disrupt the running of the gel. The Pico chip was placed in the priming station. 9.0 µL of gel mix was pipetted into the appropriate well. The station was closed and a 1ml syringe plunder was pressed until positioned under the clip. This position was maintained for 30 seconds before the clip is released, allowing the plunger to recoil. The plunger was manually drawn to 1 mL, and the station opened. Two further wells were filled with 9.0 µL of gel mix.

Placing markers and samples into appropriate wells

The appropriate wells were filled with Pico Conditioning solution (9.0 µL, white), and the sample wells and ladder well were all filled with 5.0 µL of Pico marker (green). Following this, 1 µL of denatured RNA sample and 1 µL of the denatured ladder solution were added to each of the wells. The chip was placed in the chip vortex at 2400 rpm for 2 minutes. The chip was loaded into the Agilent BioAnalyzer for electropherenograpghy. Following loading of the gel with RNA samples, the sample was analysed within 5 minutes.

2.5.5 Agilent DNA 12000 chip

The DNA chip was performed in a similar way as the RNA pico chip above. Before starting the base plate was moved to position C on the priming station and the syringe clip was moved to the top position. Gel-Dye mix was prepared by adding 25 µL of the dye mix to the DNA gel matrix vial, which was mixed using the vortex. 9 mL of the gel-dye mix was added to the appropriate position on the DNA chip and pressurised with 1 mL of air in the syringe for 60 seconds. The syringe plunger was held in place by a clip which was released after the required time, and withdrawn back to 1 mL after 5 seconds. The priming station was opened and 9 µL of gel-dye mix was added to 2 other marked wells designated for dye. Five microliters of marker dye (green) were added to the remaining wells. 1 µL of DNA ladder was added to the appropriate well. 1 μL of DNA sample was added to each well as appropriate with empty wells being filled with the same volume of buffer (EB buffer in this case). The chip was the vortexed at 2000rpm for 1 minute before being analysed on the BioAnalyzer.

Chapter 3. Whole Genome DNA methylation in IBD

3.1 Abstract

3.1.1 Introduction

Epigenetic alterations including DNA methylation may provide important insights into gene-environment interaction in complex immune diseases such as inflammatory bowel disease (IBD). Whilst whole tissue methylation changes may provide clinically useful biomarkers, epigenetic changes are cell-type specific. This study aimed to characterise the circulating methylome in IBD, and relate changes seen in whole blood to the methylation profile in separated leucocytes, gene expression data, as well as our previous data in childhood-onset disease.

3.1.2 Method

The Illumina 450k array was used to assess whole blood leucocyte DNA methylation at over 485,000 CpG sites across the genome in 240 patients (121 Crohn's disease [CD], 119 ulcerative colitis [UC]) and 191 controls. Whole blood data was analysed after correcting for batch effects and for the cellular composition of the samples. Differentially methylated sites discovered in whole blood were also investigated in immunomagnetically separated leucocytes (CD4+ & CD8+ lymphocytes, CD14+ monocytes).

3.1.3 Results

There were 439 differentially methylated positions (DMPs) meeting epigenome wide significance as defined as a Holm corrected p value of <0.05 (uncorrected $p \leq 1.1 \times 10^{-7}$) in IBD cases versus control. No markers were significantly different between CD and UC following correction for multiple testing.

There were 5 differentially methylated regions (DMRs) with unidirectional methylation change in ≥3 adjacent markers each achieving Holm-adjusted significance of p<0.05.

There was significant enrichment of methylation alteration around known susceptibility loci. Established as well as novel pathways pertinent to disease pathogenesis are strongly implicated. The most significantly DMP in whole blood (RPS6KA2 [corrected $p=1.1 \times 10^{-16}$] was also hypomethylated in monocytes in UC (uncorrected $p=3.5 \times 10^{-6}$). The most significant DMR, VMP1/miR21 (most significant probe corrected $p=4.9 \times 10^{-14}$) strongly replicates the same finding in our previous study. The gene encoding Beta-2 Integrin (ITGB2) was a

hypermethylated DMR in IBD and more specifically CD (most significant probe corrected $p=4.3 \times 10^{-4}$) compared with controls.

3.1.4 Conclusion

This is the most detailed characterisation of the epigenome carried out in IBD to date and includes novel data exploring the circulating methylome in UC. The findings strongly validate this approach in complex disease, replicate and expand previous data, and provide clear translational opportunities.

3.1 Introduction

Inflammatory bowel (IBD) has a strong genetic contribution, and a meta-analysis of genome-wide associated studies (GWAS) of European and non-European patients with IBD has demonstrated 200 loci associated with developing the disease.[86,88] However despite this tremendous progress in delineating the genetic architecture of IBD, genetics alone explains only a small proportion of disease heritability (13.1% CD and 8.2% UC of disease variance).[86] A number of environmental factors are known to influence the development and course of disease; particularly smoking, diet and the gut microbiota.[278] This has led some investigators to look beyond genomics to investigate epigenetics, as a potential interface between genetics, environmental modifiers and disease.[144] Epigenome-wide association studies (EWAS) of complex diseases such as Rheumatoid arthritis,[204] type 2 diabetes mellitus[279] and obesity,[280] are now beginning to be published in high-impact scientific journals.

DNA methylation EWAS aim to determine the distribution of methyl groups at thousands of specific positions across the genome (CpG sites, cytosine-guanine dinucleotide) with the aim of identifying arrangements that are more common to certain disease traits compared to controls.[281] The biological significance of DNA methylation, is the association of DNA methylation (or hypermethylation) occurring within regulatory regions of genes (for example promotors or transcription start sites) and gene repression or gene silencing.[282] Epigenetic studies also have important confounding factors, significantly the varying epigenetic profile occurring in each different cell type.[283] Thus many of the early EWAS discoveries may have been more related to changes in differing cell proportions in cases and controls rather than disease-specific epigenetic changes.

In the context of IBD we have used the Illumina 450k platform to assess genome-wide DNA methylation patterns in treatment naïve children with Crohn's disease(CD).[284] This study demonstrated highly statistically significant differences in CpG sites that were replicable in a modest number of samples.[284] The most significantly differentially methylated position and regions highlighted genes implicated in disease pathogenesis. The same study also used two DNA methylation probes to accurately discriminate between cases and controls, indicating a strong translation potential.[284]

Many of the DNA methylation studies to date are limited by small numbers and a lack of cell-type specific information. This is the first major study of DNA methylation in IBD to include comprehensive integration of genetic and gene expression level data, and relate changes seen in whole blood to the methylation profile in separated cells. A well phenotyped prospective cohort of newly diagnosed patients allows the evaluation of DNA methylation data as a potential biomarker for diagnosis and stratification of disease progression.

This design of this experiment aims to address two of the major concerns current epigenetic research in complex diseases:

1. Epigenetic signatures arise from specific cells. Previous studies describe DNA methylation in heterogeneous tissues (e.g. blood, gut tissue), masking the individual epigenetic signatures that exist in each cell type. We aim to discover the specific cell type from which these epigenetic signals may arise.
2. Epigenetic marks may exist as cause or consequence of disease. Determining causality has remained a major barrier for epigenetic research. Recruiting patients at diagnosis will record the epigenome as near to disease onset as possible, and limit the impact of immunomodulating drugs and chronic inflammation on the epigenetic landscape.

3.2 Methods

3.2.1 Patient selection

Patients and controls were recruited according to methods 2.1 Patients recruitment. Patients within 3 months of diagnosis were selected in order to limit the potential effect of chronic inflammation and immunomodulatory drugs on the epigenetic profile. The patients were recruited prospectively as part of the IBD-BIOM inception cohort from gastroenterology and endoscopy appointments.

Symptomatic controls were recruited from gastroenterology clinics during the same period. A further control group consisting of healthy volunteers with no self-reported gastrointestinal symptoms were also recruited. Additional patients and control samples recruited retrospectively fitting the newly diagnosed criteria (<3 months) were also included.

3.2.2 Immunomagnetic cell separation

Sixty newly diagnosed, treatment-naive patients (20 CD, 20 UC, 20 controls) were selected for detailed cell separation analysis. DNA was extracted from whole blood and isolated lymphocyte subtypes (CD4+, CD8+, CD14+ and CD19+) separated using magnetic bead separation (AutoMacs Pro, Miltenyi)(Section 2.2).

3.2.2 Stratified randomisation

In Microsoft excel a column of random numbers was created using the rand() function. The samples were ordered according to (i) diagnosis, (ii) sex, (iii) smoking status, (etc. age, age of sample), (iv) age of patient, (v) age of sample and lastly the random number (e.g. low to high). Following ordering of samples, each sample was sequentially numbered 1 to X (where X=the number of arrays). The whole list was ordered on array number giving a stratified list of random samples with respect to position on array. The arrays themselves were randomised (otherwise bias toward the alphabetically first diagnosis, sex, smoking status on the first arrays) by sorting each array on the basis of a random number (again using the rand() function). Each array was checked manually for distribution of phenotypic characteristics, before ordering samples according to array, followed by random number (to ensure each sample had a random position on the array). There was no statistically significant difference in the age of patient (Kruskall Wallis rank sum test chi squared 21.7, df=31, p=0.9), age of sample (Kruskall Wallis rank sum test chi squared 21.7, df=31,p=0.98), gender (Kruskall Wallis rank sum test chi squared 5.1, df=31, p=1) and diagnosis(Kruskal Wallis rank sum test chi squared=2, df=31, p=1) between arrays (Figure 12).

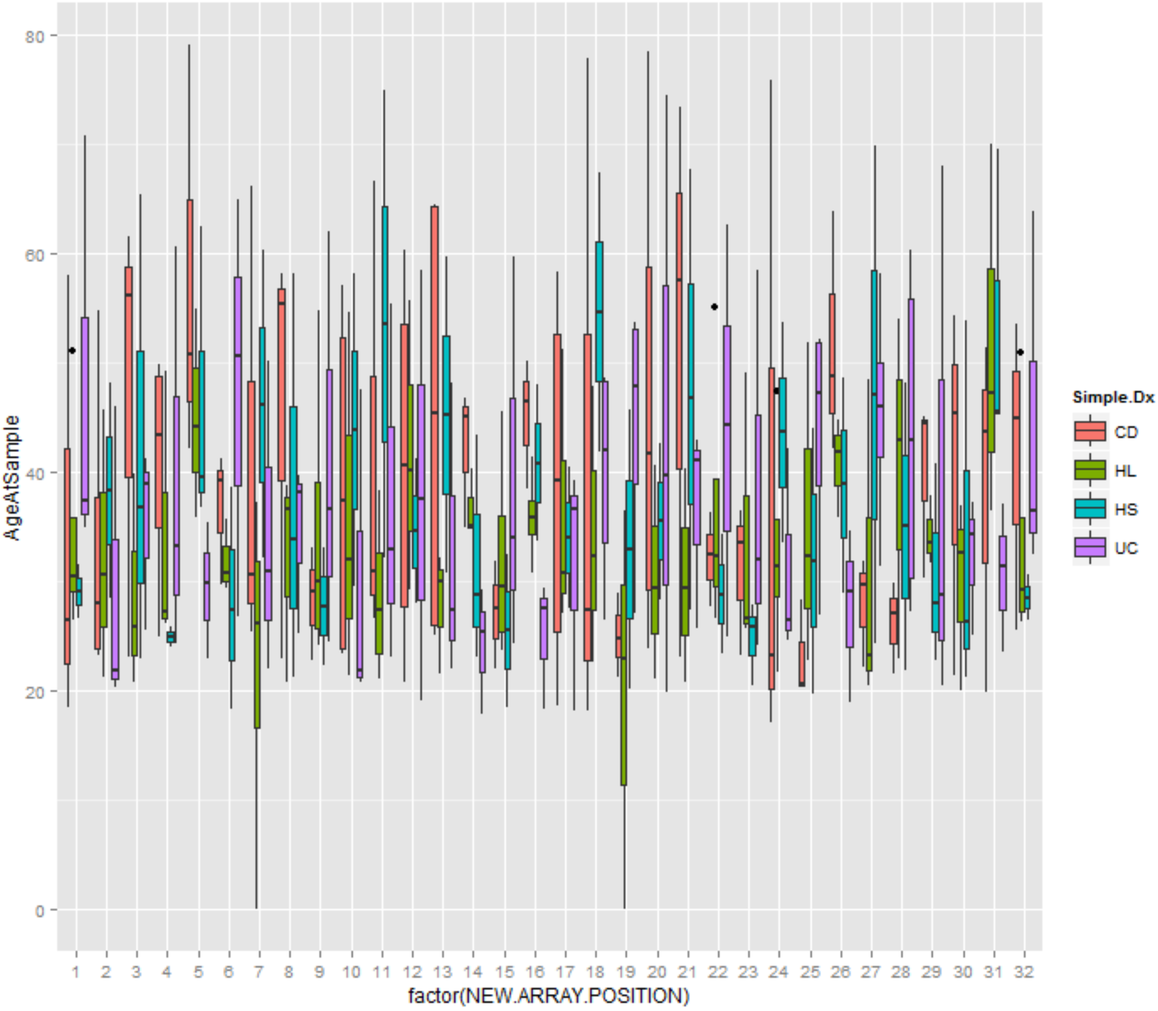


Figure 12 - Planning even distribution of samples across 450k microarrays. Age of Patients and Diagnoses across microarrays (y axis=age, x axis= array position, colour of boxplot corresponds to diagnosis, CD= Crohn's disease, UC= ulcerative colitis, HL = healthy lab volunteer, HS = symptomatic control)

3.2.3 DNA concentration

DNA was required for Illumina microarrays in a concentration of 50ng/ µL in 10 µL. Dilute samples were concentrated using the speed vac concentrator. The samples were placed in 1.5 mL tubes with lids off in a vacuum concentrator for a variable amount of time to concentrate the samples depending on the initial volume. Following concentration, dry samples were resuspended with 12 µL of buffer EB (Qiagen).

3.2.4 Picogreen DNA quantification

Picogreen dye binds to double stranded DNA, and when excited by light (485nm), fluoresces at 530nm. DNA were quantified by sphectroflurometer; fluorescent plate reader (Qubit, see methods 2.5.2 QuBit flurometery nucleic acid quantification).

3.2.5 Bisulphite conversion

The bisulphite conversion for the 450k microarrays was performed by the WTCRF. Genomic DNA was bisulphite converted using the Zymo EZ DNA methylation Kit (Zymo, USA [high profile plate kit used by WTCRF, low profile by MMC]). See also methods detailed in 4.2.2 Bisulphite conversion.

3.2.6 Sample size and Power calculation

From the previous paediatric dataset[284] the mean standard deviation across all probes and of the top 1000 differentially methylated probes (CD vs controls) was calculated. The variance across probes in general followed a log normal Gaussian distribution (Figure 13 Top). The median variance across all probes was 0.016 in cases and controls, however, there was a statistically significant difference between the variance of cases and controls within the top 1000 probes (Wilcox rank test p=0.02) and across all probes (p-value $< 2.2 \times 10^{-16}$). The median standard deviation was higher in the top 1000 probes in the paediatric dataset suggesting higher inter-individual variance in these probes (median 0.04, vs. probes outside top 1000, p-value $< 2.2 \times 10^{-16}$).

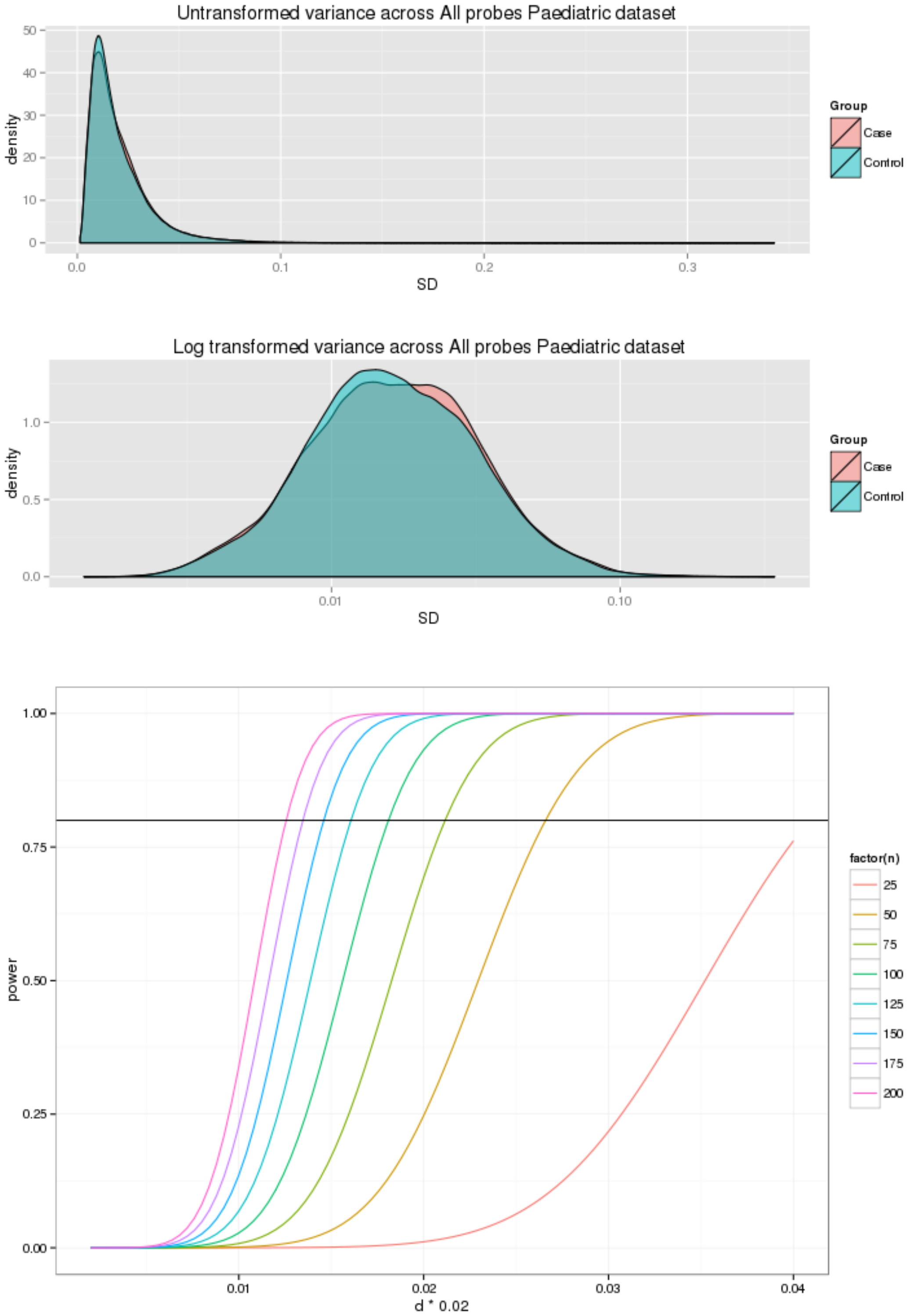


Figure 13 - Power calculation. Top - Distribution of beta values and log transformation. Bottom - Power curves used to determine numbers per group required to detect a difference of one standard deviation (y axis=power (ab line at 80% power), x axis = effect size (standard deviation of 0.02))

The power to detect various effect sizes for differing level group sizes can be observed on Figure 13 bottom panel. For this method, curves were modelled using an alpha error value of genome wide significance was set at p= 1 × 10^{-7} using the pwr.T.Test function with the beta error along the y axis (ab line at 80% power) and effect size (d) along the x axis. For an 80% power to detect an effect size of a difference in means of one standard deviation (median standard deviation of top 1000 probes in paediatric dataset=0.04) was 100 patients per group (Table 4). This was tested using 5 random probes in the top 100 differentially methylated probes with a difference in beta values between cases and controls of +/− 3% (cg27049094, cg01726890, cg22768358, cg21328643, cg27361520). Using an alpha of p=1 × 10^{-7}, beta of 80%, and an effect size of the observed difference in means the range of numbers per group was 20-45. The variance and magnitude of effect size differs from published estimation of sample size.[202]

delta	**Number per group**	**Power (%)**	Assuming SD of 0.04 the % difference in mean beta value
0.1	100	0.000171	0.4
0.2	100	0.003715	0.8
0.3	100	0.051914	1.2
0.5	100	2.767155	2
0.6	100	10.85059	2.4
0.7	100	29.01307	2.8
0.8	100	55.10669	3.2
0.9	100	79.09058	3.6
1	100	93.19707	4
1.5	100	99.99995	6
2	100	100	8

Table 4 – Power to detect varying effect sizes with the group size set at 100 patients per group

3.2.7 Data processing

Illumina 450K microarray data was analysed in R statistical environment (R version 3.0.2, Vienna) using the lumi[285], methylumi, minfi, wateRmelon and ChAMP packages and is summarised in Figure 14.

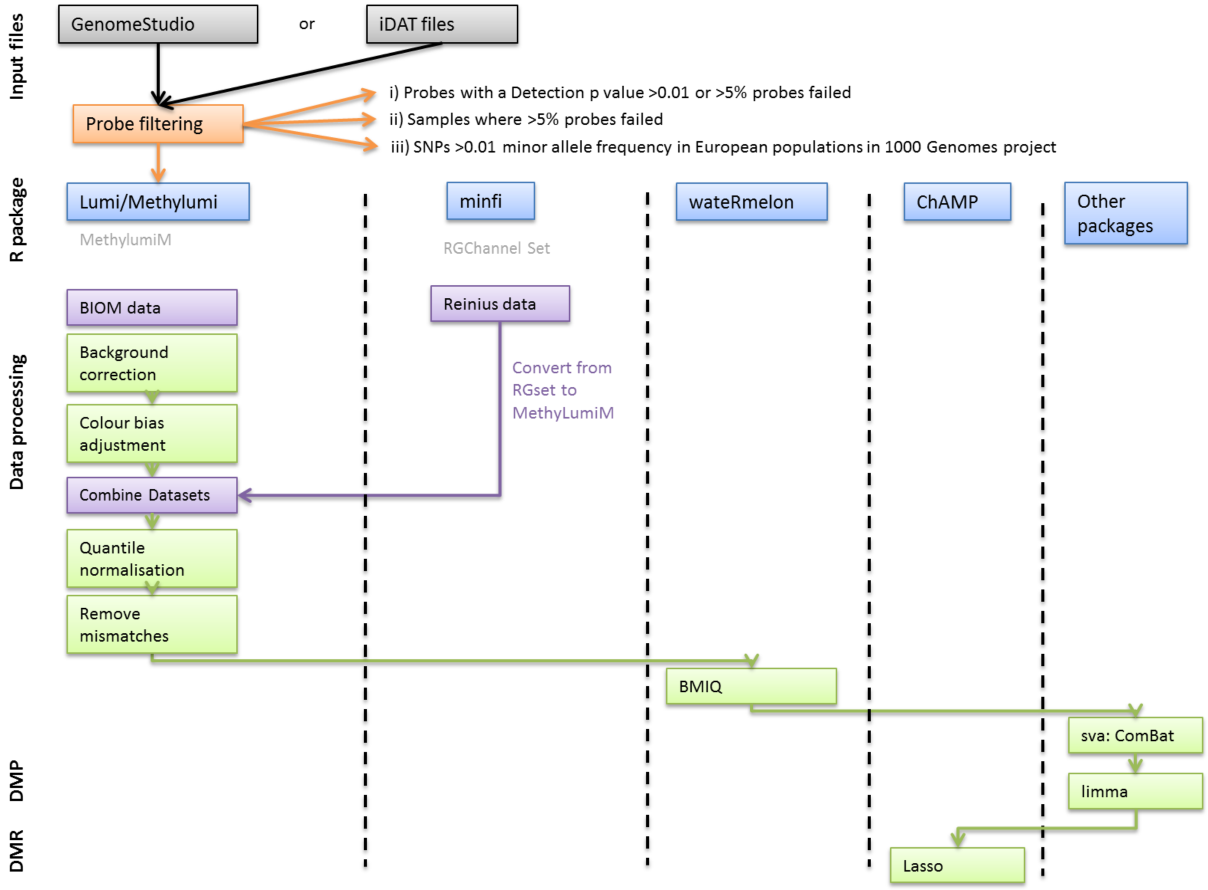


Figure 14 – Summary of Illumina 450k data processing. Note Houseman estimation of cell counts calculated independently.

3.2.8 450K array design

The 450K array measures the methylated and unmethylated intensity at each CpG site (the array measures over 485,000 CpG sites). Each sample is measured in two colour channels (Figure 15); red and green. [285,286] There are two distinct probe designs on the array:

- Type I probes: signals are measured in the same colour (red or green), one probe is used for the methylated signal and another is used for the unmethylated signal
- Type II probes: One probe is used, with the green signal measuring the methylated intensity and the red signal measuring the unmethylated intensity.

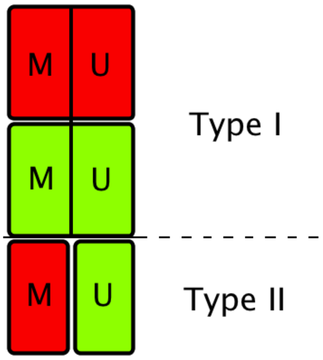


Figure 15 – Probe design on the Illumina 450k microarray. Taken from minfi and ShinyMethy bioConductor tutorial[287]

3.2.9 Data output from Illumina 450K microarray

Unprocessed data was outputted in iDAT files and GenomeStudio files. The iDAT files contain all raw bead-level information and are much larger, whereas the GenomeStudio have probe summary level information. Several downstream quality control measures require the iDAT files as inputs (i.e. background information); and the lumi package can import iDAT files using the importMethyIDAT function. GenomeStudio data are read directly into R using the lumiMethyR function. iDAT files can be read into the package mini using the command read.450k.sheet and read.450k.exp() to create an RGChannelSet object containing the raw data from the iDAT files and the phenotypic data.

3.2.10 Filtering of probes (methylumi, lumi)

The number of detected CpG sites should be >450,000 with 485,000 or more indicating 98% conversion. Probes with a detection *P* value of ≥0.01 were removed. Individual samples with >5% of probes failing were also removed. Probes containg a single nucleotide polymorphism (SNP) with a minor allele frequency of ≥0.01 (European population, 1000 Genomes Project) were also removed.[288]

3.2.11 Quality controls (methylumi)

Samples were removed if there was a sex mismatch (n=3) according to methylation on the X chromosome (Figure 16). In females with two X chromosomes, one X chromosome is methylated in the process of lyonisation. Consequently females will have just over 0.5 methylation of the X chromosome whereas males have much less. It is therefore easy to identify mismatches in sex using this method. Cell mismatchs (n=2) were identified using principal component plots and removed.

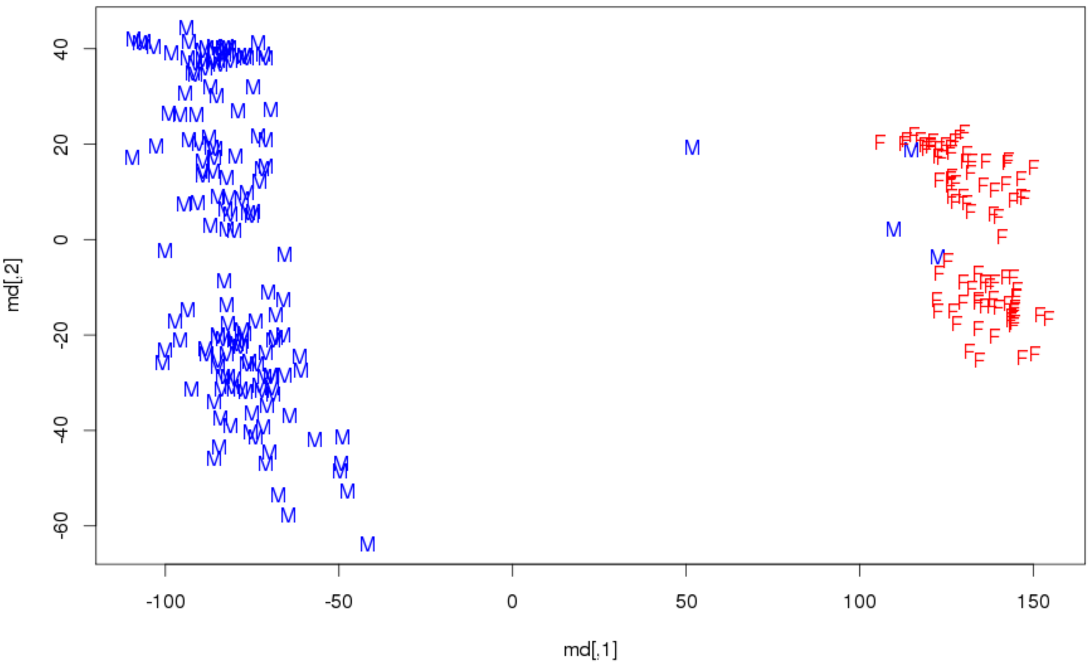


Figure 16 – Multidimensional Scaling plot to detect sex mismatches

3.2.12 Colour probe adjustment (lumi)

The Illumina 450K platform uses two colours (red and green) to label the final base based on a hybridisation of methylated and unmethylated probes.[285,286] The final extended bases ending in A or T are measured in the red channel and those ending in C or G are labelled in the green channel.[285,286] This can result in dye colour bias as a result in differences in scanning efficiency of the two colour channels. Colour adjustment was performed in lumi using the function (`lumiMethyC`). The colour balance was checked before and after colour adjustment Figure 17.

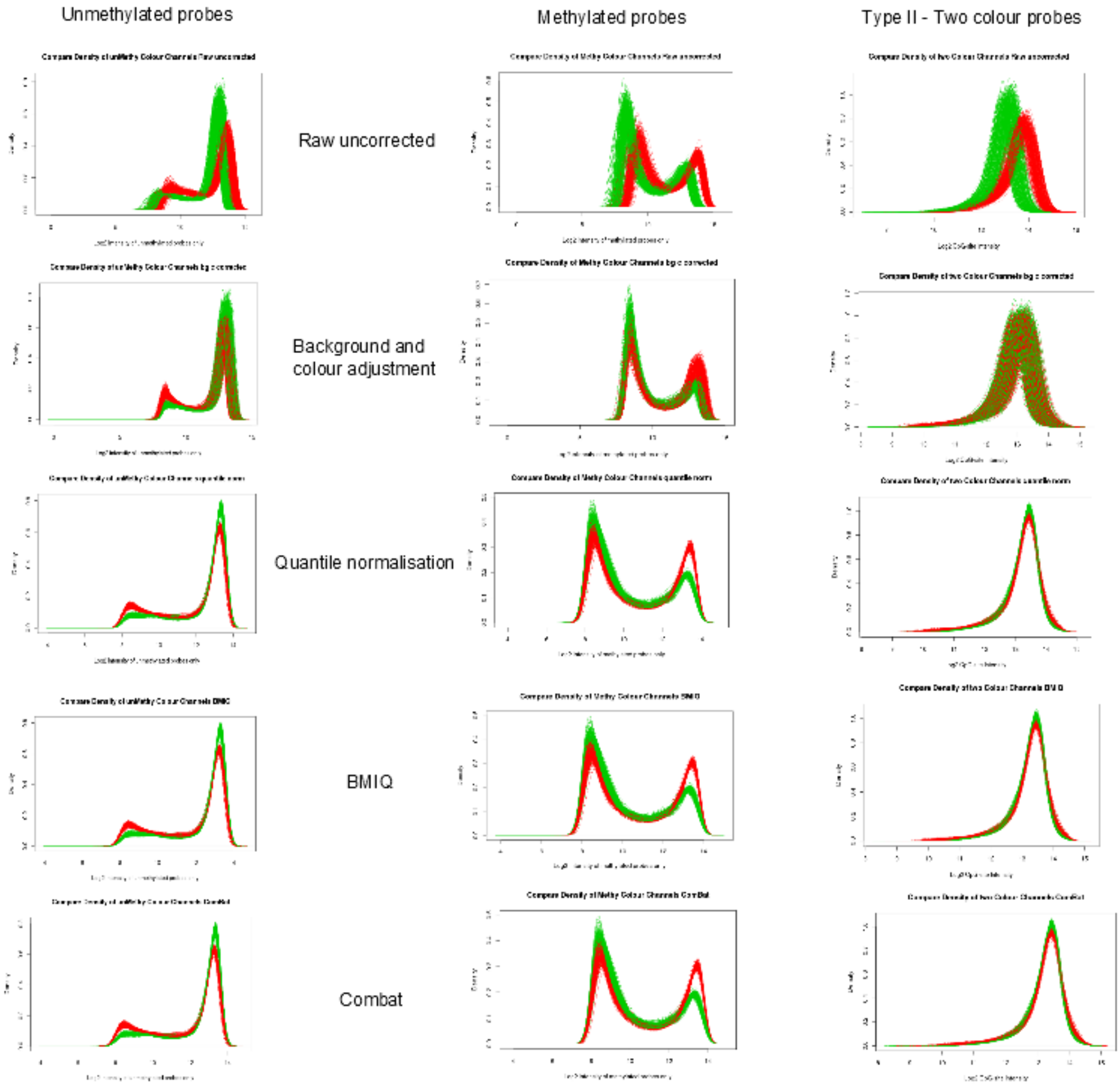


Figure 17 - Colour density plots following data processing. Y Axis = Density (0 to 1), X Axis = Log2 intensity of i) unmethylated probes ii) methylated probes iii) Two colour probes

3.2.13 Background adjustment

Control probes have been included on the Illumina 450K array allowing an estimation of background intensity. This background intensity is the median intensity of the negative control probes and the data are held within the controlData slot. As separate background intensity data is generated by the red and green control probes, the authors of the lumi pipeline suggest to correct for colour probe bias before background adjustment.[285,286] Background adjustment was performed using the function `lumiMethyB`, which first estimates the background level of each sample (`estimateMethylationBG`) and corrects based on the returned estimation (`bgAdjustMethylation`).[286]

3.2.14 Quantile normalisation

Methylation data from all samples were normalised to a common scale to allow comparison given the large inter-sample variation in total methylation. There are several normalisation techniques, and some debate within the literature regarding the most appropriate form of normalisation.[289] Quantile normalisation is performed by ordering the probes according to methylation intensity value in each sample, and then taking an average (usually mean) across all probes. The highest methylation intensity becomes a mean of all the highest methylation intensity values; the second highest value becomes the mean of all the second highest values etc. The new values are substituted back in for each sample according to the rank within that sample. The new normalised samples therefore have the same distribution and are more easily compared.[290] Quantile normalisation is used commonly when analysing array data, where expected changes are likely to be due to technical rather than biological variation. In the lumi package, the `lumiMethyN` function was used, with the X and Y Chromosomes excluded, using the quantile method.[286]

3.2.15 BMIQ

Beta-Mixture quantile dilation (BMIQ) is an intra-sample normalisation procedure to correct for intra-array technical variation caused by the two types of probe design.[291] The two probes have different methylation distributions and dynamic ranges. [291] The method used adjusts the beta values of type II probes to a statistical distribution of type I probes. The process of correction runs three stages; i) fitting a model to assign probes to one of a three methylation states; ii) transformation of the type II probe quantiles into those of type I probes and iii) a conformal transformation of hemi-methylated probes.[291]

The BMIQ function of the wateRmelon package was used for correction.[292]

```
CombinedQBMIQ <- BMIQ(CombinedQ)
```

3.2.16 Batch effects adjustment (ComBat)

Batch effects are artefactual differences in array data arising for technical reasons, most commonly when arrays are run on different days, at different sites or using different array lots. Given that the Illumina HumanMethylation450K array can assay 12 samples, any experiment larger than this is likely to incur batch effects. Combating Batch Effects When

Combining Batches of Gene Expression Microarray Data (ComBat) is an informed method where known batches are adjusted using an empirical Bayesian framework.[293] Combat uses the surrogate variables analysis package.[294] ComBat requires a model matrix with biological variables of interest (in this cases cell type, disease status), which is used to inform the algorithm, allowing correction of technical variability (batches) without removing the biological variability of interest. [294]

```
modMat <- model.matrix(~1+SimplifiedDiagnosis*CellType,newPheno)
```

ComBat is then performed using the default parametric Bayesian adjustment, using the chip number to correct for batch effects.[294]

```
NewCombat <-
ComBat(exprs(CombinedQBMIQ),CombinedQBMIQ$Chip,modMat)
```

The ComBat function returns an expression matrix that can be re-inserted into the MethyLumiM object.

3.2.17 Using multi-dimensional scaling plots to determine the effect of data processing

Multidimensional scaling, also known as principle coordinates analysis, is a useful method of visualising data at each stage of processing .[295] MDS plots present the degree of relatedness of samples as the proximity in p-dimensional space. MDS initially assigns samples to arbitrary co-ordinates in p-dimensional space. The Euclidean distance between each point is calculated and a stress function is calculated by comparing the input data and coordinates. The R function cmdscale() is used to create MDS plots.

3.2.18 Differentially methylated positions (DMPs)

Methylation status of cases vs. controls was compared using the R package limma.[296] Linear models of beta values were constructed with disease status as the coefficient and the measured or predicted cell proportions as covariates. Genes were annotated using the IlluminaHumanMethylation.db package.[297] Methylation beta values (ratio of methylated and total probe intensity (0 to 1, or % of methylation)) were used to identify DMPs. Although beta values demonstrate high heteroscedasticity at the extreme high and low methylation ranges, are more biologically intuitive to use that the M value (log2 ratio of methylated and unmethylated probe intensity).[298] The Holm methods was used to correct for multiple testing, with significance set at $p<0.05$.[299]

3.2.19 Differentially methylated regions (DMRs)

Differentially methylated regions (DMRs) were calculated using an adapted Lasso function from the CHaMP pipeline.[300] A DMR was defined as three or more probes reaching Benjamini-Hochberg False Discovery Rate (FDR)[301] adjusted significance ($p<0.05$) sharing the same direction of change (either all hypo- or hypermethylated) within a 2kb distance threshold.

3.2.20 Correcting for different cell type using the Houseman algorithm in Minfi

Given the potentially confounding effect of heterogeneous cell types on methylation data, bioinformatics methods have been used to estimate whole blood cell proportions using DNA methylation data.[205,302] The method uses a reference dataset generated by Reinius et al that studied the DNA methylation in healthy human blood cells (CD4+, CD8+, monocytes, NK cells, B cells, granulocytes, eosinophils) using immunomagnetic separation.[303] The raw methylation data was combined with Reinius reference data. The estimateCellCounts function in minfi using defult settings was used.[304,305] A matrix is returned giving the estimated relative proportions of pure cell types in a given sample. This cell count data was used as a variable in linear modelling to adjust results for cell count. It should be noted this step is done in separately to the normalisation steps described above and that the minfi function applies its own normalisation procedures after experimental and reference datasets are combined.

3.2.21 Gene Ontology analysis

Gene ontology analysis was performed on DMPs achieving Holm-corrected statistical significance using goSeq.[306] The same method as has been used elsewhere to correct for the number of probes per gene (The most statistically significant probe per gene) was performed together with probability weighted function.[284,307] A Benjamini-Hochberg correction for multiple testing was applied to GO term results.

3.2.22 GWAS co-localisation

The proximity of differentially methylated probes to the 163 known IBD-associated GWAS risk loci described in Jostins et al[88] within range thresholds of 25kb, 50kb, 100kb and 250kb was compared with 1000 randomly selected bins of the same size with matched probed density using Wilcoxon rank sum test. The same methodology was used previously by Adams et al in the paediatric dataset.[284] As a control, the IBD-associated differentially methylated positions were also tested for enrichment in the same way with GWAS data from 7 other diseases; rheumatoid arthritis, psoriasis, ankylosing spondylitis, TB, type I diabetes, Alzheimer's disease, IgG glycosylation, colorectal cancer and hair colour. The additional GWAS data was obtained from the GWAS catalogue (http://www.genome.gov/gwastudies/).

3.2.23 Epigenetic clock

DNA methylation data have been used to accurately predict the age at sampling of human tissue.[142] An algorithm developed by Hovrath was modified by Dr Nick Kennedy locally to compare the predicted age of samples based on DAN methylation data with the actual age of samples.[142] DNA methylation data were normalised using methods described by Hovrath.[142] The method used 353 CpGs to predict age.[142] Correlation was performed using Pearson's coefficient. The difference between actual and predicted age between cases and controls was compared using the Wilcoxon rank test.

3.3 Results

There were 440 individuals included in the main 450k DNA methylation experiment. Figure 18 details the DNA samples available for methylation experiments and the overlap of separated cell samples available. After removal of duplicates, mislabelled or sex mismatches, there were 431 participants included in whole blood 450K DNA methylation experiments.

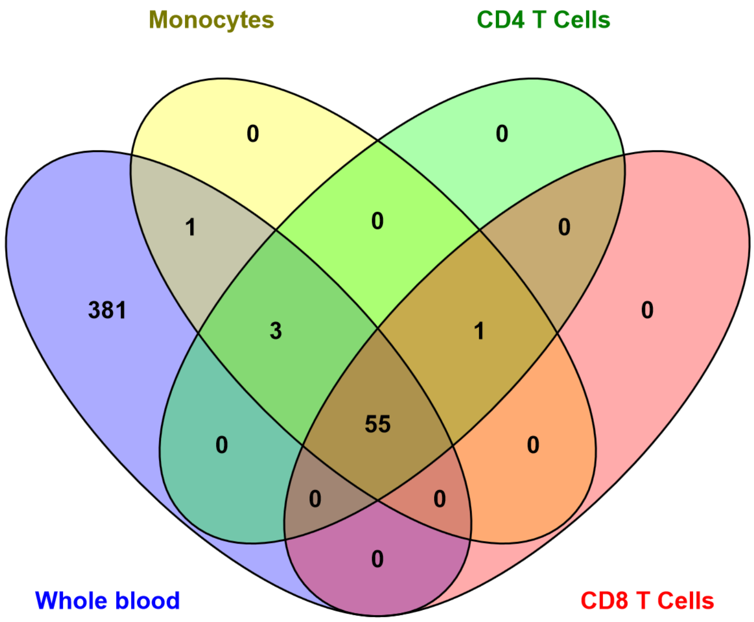


Figure 18 - Venn diagram[308] detailing samples derived from each individual participant used for whole genome 450K DNA methylation profiling (e.g. Of the 440 patients included, 55 patients had all 4 cell samples available for analysis, 381 patients had whole blood DNA alone. 3 patients had whole blood DNA, monocyte and CD4+ DNA but not CD8+ DNA etc)

3.3.1 Patient Demographics

Patient demographics for the whole blood DNA methylation cohort are outlined in Table 5 and Table 6. Cases were well-matched with controls for age and sex. There were more current or ex-smokers in the CD group, and a higher degree of inflammation in cases compared to controls as would be expected in a newly diagnosed cohort.

		CD (n=121)	UC (n=119)	Symptomatic controls (n=74)	Healthy controls(n=117)
Age [median(IQR)]		32.4 (24.9-50.7)	34.3 (25.5-47.8)	32.8 (26.4-45.5)	32.3 26.4-40.6)
Females (%)		58 (47.9)	51 (42.9)	39 (52.7)	59 (50.4)
Smoking status	Current	53	13	17	24
	Ex	29	45	17	32
	Never	39	58	40	56
	Unknown	0	3	0	5
CRP		8(2-23)	11.5 (2-31)	0 (0-3.5)	
ESR		18 (5-39)	5.5 (4.3-9.8)	6 (4.5-7.5)	
FC		495 (135-828)	760 (660-950)	19 (19-37)	

Table 5 - Patient Demographics for patients undergoing 450k analysis. Data presented as median and interquartile range except where specified. (CD= Crohn's disease, UC= Ulcerative colitis, SC= Symptomatic controls, HL=Healthy Lab volunteers, IQR=interquartile range, CRP=C-reactive protein, ESR=Erythrocyte sedimentation rate, FC= faecal calprotectin)

	CD (n=121)	UC (n=119)
Time between diagnosis and sample (days, median [IQR])	48 [7-90.8)	32 [1-71]
Treatment Naïve (denominator number with available data)	27/69	37/70
Oral/IV steroids at sample (duration of therapy in days, median [IQR])	10/69 (8[2-14])	10/70 (2[1-6.5])
Biologic at sample (duration of therapy in days, median [IQR])	4/69 (4 [2.75-5.25])	-
Aza/6MP at sample (duration of therapy in days, median [IQR])	10/69 (5[2-10])	4/70 (5[2-10])
Topical therapy at sample (duration of therapy in days, median [IQR])	1/69 (1 day)	11/70 (5[2.75-13.25])
Oral 5ASA at sample (duration of therapy in days, median [IQR])	3/69 (2[1-5])	16/70 (2[1-5)

Table 6 - Detailed baseline data on included patients. CD = Crohn's disease. UC = ulcerative colitis. IQR = interquartile range. 6MP= 6-mecaptopurine. IV = intravenous. Aza= azathioprine. 5ASA= 5 aminosalicylate.

3.3.2 Multidimensional scaling plots according to cell type

Based on the 450k methylation data, MDS scaling demonstrated that samples clustered closely according to cell type. MDS plots were used to visualise data following each of the data processing steps (Figure 19), and demonstrated closer clustering according to cell type following ComBat correction.

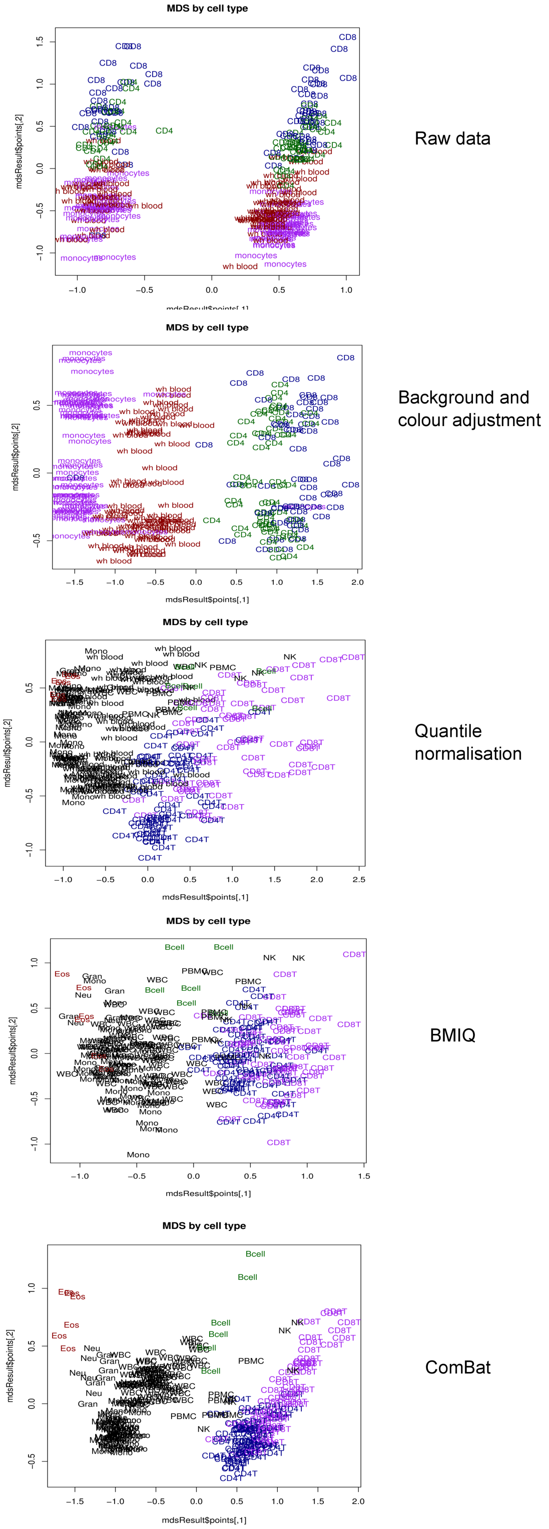


Figure 19 - multidimensional scaling plots according to cell type. Plots are demonstrated at each stage of data processing demonstrating good clustering following ComBat correction. This dataset includes Adult whole blood and separated cell data (CD4+, CD8+ and CD14+) combined with the Renius dataset.[303]

3.3.3 DNA methylation in whole blood samples

An epigenome-wide association comparison was made between IBD cases (both CD and UC) and controls (symptomatic and healthy controls). There were 439 differentially methylated positions in IBD cases compared with all controls achieving a significance of p<0.05 following Holm correction for multiple testing. The top ranking DMPs in IBD versus control are presented in Table 7 and the same information in the form of a Manhattan plot (Figure 20) and Volcano plot, (Figure 21) the latter displaying position of the methylation probe in relation to the gene and direction of methylation change. There were 412 DMPs when comparing CD (Table 44) to controls and 203 when comparing UC to controls (Table 45). CD-associated DMPs demonstrated a higher level of statistical significance than UC-associated DMPs. There were no DMPs that were differentially methylated between CD and UC following correction for multiple testing (Table 46), however there was significant overlap between IBD, CD and UC DMPs. Similarly, there were no DMPs that were differentially methylated between symptomatic controls and healthy volunteers (Table 47), as a result both control groups were combined as a single large control cohort.

Illumina 450k Probe id	Chr	Gene symbol	Δβ	P.Value	Holm adj.P.Val
cg17501210	chr6	RPS6KA2	-0.08	2.71E-22	1.22E-16
cg18608055	chr19	SBNO2	-0.07	2.02E-20	4.53E-15
cg16936953	chr17	VMP1	-0.09	1.33E-19	1.99E-14
cg09349128	chr22	NA	-0.04	3.11E-19	3.48E-14
cg25114611	chr6	NA	-0.04	1.10E-18	8.79E-14
cg12170787	chr19	SBNO2	-0.04	1.18E-18	8.79E-14
cg12992827	chr3	NA	-0.06	6.26E-18	4.01E-13
cg19821297	chr19	NA	-0.06	3.66E-17	1.98E-12
cg12054453	chr17	VMP1	-0.07	3.98E-17	1.98E-12
cg01059398	chr3	TNFSF10	-0.05	1.59E-16	7.13E-12
cg26470501	chr19	BCL3	-0.03	5.79E-16	2.29E-11
cg07398517	chr3	NA	-0.04	6.14E-16	2.29E-11
cg26804423	chr7	ICA1	0.04	6.84E-16	2.36E-11
cg18942579	chr17	VMP1	-0.05	1.17E-15	3.74E-11

Table 7 - Top table of differentially methylated positions (DMPs) between inflammatory bowel disease (IBD) cases and controls in whole blood. Δβ = difference in beta values (ratio of methylated and total probe intensity (0 to 1) between IBD cases and controls, positive value indicated increased methylation in cases compared to controls, negative values indicated hypomethylation in IBD cases vs. Controls

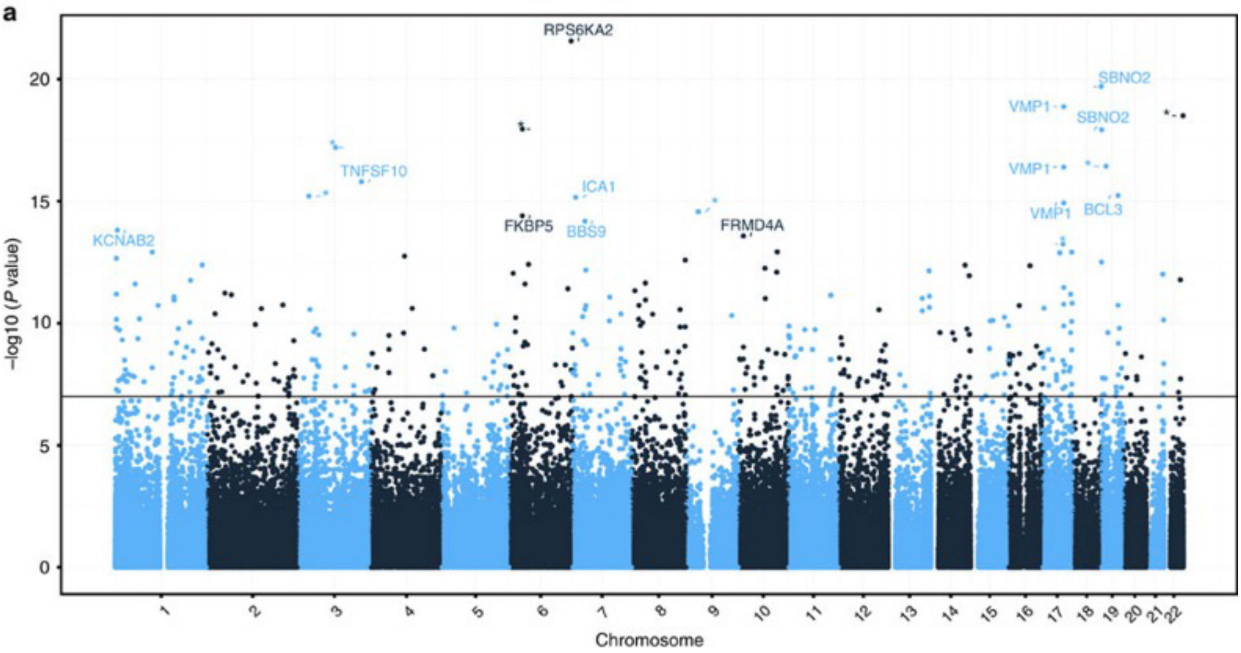


Figure 20 – Manhattan plot demonstrating differentially methylated positions in IBD versus Control

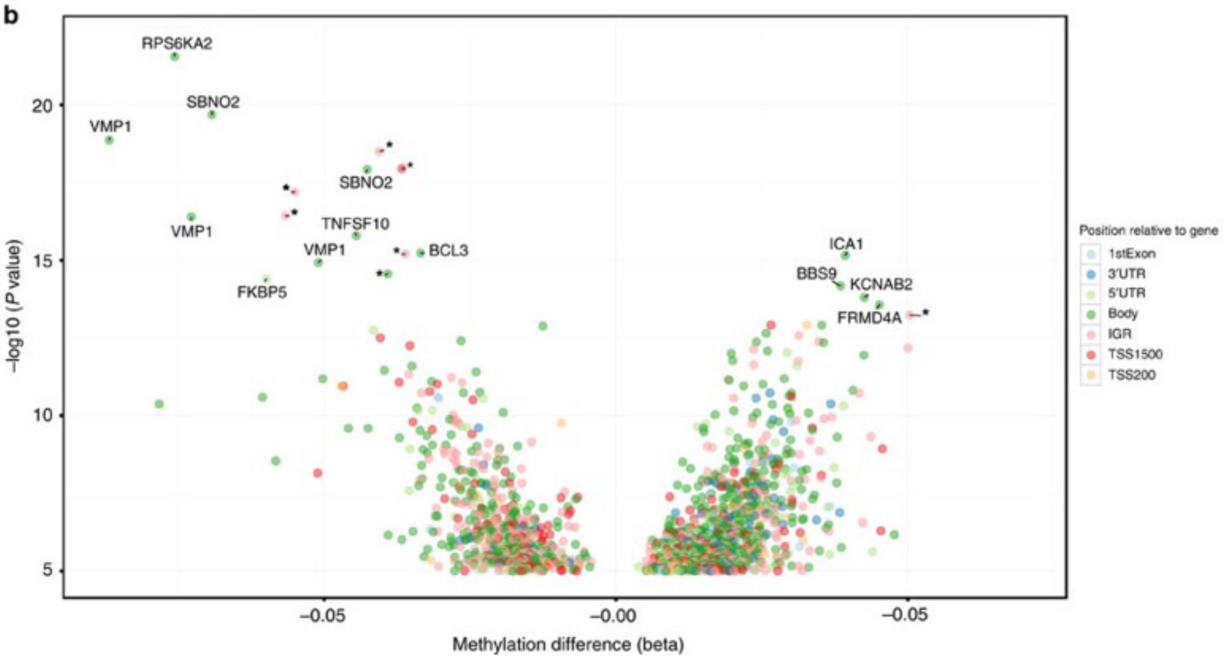


Figure 21 - Volcano plot of differentially methylated positions in IBD versus Control. Colour denotes position of methylation probe in relation to gene.

3.3.4 IBD-associated differentially methylated regions

Differentially methylated regions (DMRs) were defined as three or more contiguous probes all displaying the same direction of methylation change. Only statistically significant (Holm p <0.05) probes identified in the DMP IBD-case control analysis were include in this analysis. Five differentially methylated regions were identified in IBD cases versus controls and are listed in Table 8. There were 4 CD-associated DMRs (VMP1, ITGB2, WDR8, CDC4BPB), and 2 UC-associated DMRs (VMP1, WDR8) compared with controls.

Gene	Feature	CHR	Δβ	Min Holm adj.P.Val	DMR size	Probe Counts	Disease
VMP1	Body	17	-0.09	5.96E-14	1150	4	IBD, CD, UC
WDR8	Body	1	0.03	9.76E-08	1943	3	IBD, CD, UC
NA	IGR	1	0.04	1.83E-07	1997	3	IBD
ITGB2	5'UTR	21	0.04	3.28E-05	623	3	IBD, CD
TXK	5'UTR	4	0.02	0.00014	538	3	IBD

Table 8 - List of differentially methylated regions (DMRs) between inflammatory bowel disease (IBD) cases and controls in whole blood. Where a single p value or beta difference is presented, this represents the corresponding values from the most significant probe within the DMR. Δβ = difference in beta values (ratio of methylated and total probe intensity (0 to 1) between IBD cases and controls, positive value indicated increased methylation in cases compared to controls, negative values indicated hypomethylation in IBD cases vs. Controls.

Differentially methylated regions (DMRs)	
VMP1/ miR-21	Vacuole membrane protein1 (VMP1) encodes a transmembrane protein within the Golgi body, endoplasmic reticulum.[309] VMP1 is an inducer of autophagy via interactions with BECN1.[310,311] miR-21, a microRNA is encoded at the 3' end of VMP1,[312] where the change in methylation is concentrated at in IBD.[284] miR-21 has been linked with several types of cancer including colorectal cancer.[313] miR-21 has been widely implicated in pathogenesis of IBD.[314] miR-21 knockout mice have improve survival following chemically (Dextran sulphate sodium) induced colitis in two studies,[315] but exacerbation in TBNS (2,4,6-trinitrobenzenesulfonic acid) and T-cell transfer models of murine colitis.[316] Several microRNA screens have identified upregulation of miR-21 in IBD, notably in cases with active inflammation.[193,195–197,200]
WDR8/ WRAP73	WD (trp-asp) repeat protein family, antisense to Trp73. The WD gene repeat motif make up a large family of genes involved in several cellular and gene regulatory processes, including cell cycle progression, apoptosis and signal transduction.[317] Murine studies suggest a role for WRAP73 in the process of ossification.[317]
ITGB2 (CD18)	Integrin Beta 2 subunit. Cell adhesion and cell surface mediated signalling.[318] Heritable defects in this gene cause leukocyte adhesion deficiency type I characterised by severe recurrent infections.[319,320] Aberrant DNA methylation at the ITGB2 locus has previous been demonstrated in IBD[321] and other diseases.[322,323] Anti-integrin antibodies have attracted interest as therapeutics in IBD. Natalizumab is an anti-alpha4 integrin antibody is efficacious in multiple sclerosis and CD, but is associated an unacceptable risk of PML (progressive multifocal leukoenchepalopathy).[75] Vedolizumab is a gut-specific anti-α4β7 integrin antibody efficacious for inducing and maintaining remission in UC (GEMINI-1)[76] and CD.[77]
TXK	TXK is a member of the Tec family of tyrosine kinases.[324] T cells express TXK and the other tec kinases, which serve as modulators of T-cell receptor signalling and assist in cytokine production by CD4+ effector T-cells.[325] TXK may also have role in T-cell development in the thymus. TXK has been shown to be over expressed in the circulating leucocytes in Behcet's disease [326] including in Th1 lymphocytes accumulating within intestinal lesions.[327] An IBD-associated GWAS locus.[88]

HDAC4	Histone Deacetylase class II. Histone modifications are an important epigenetic mechanism that are associated with alternative conformations of chromatin. Acetylation of lysine on histones H3 and H4 is associated with transcriptional activity.[174] The extent of acetylation are regulated by the relative activity of the HDAC enzymes, together with the HAT (histone acetyl transferases).[174] Non-specific HDAC inhibitors such as butyrate are associated with amelioration of colitis.[182] Alternative methylation of HDAC4 has previously been described in other situations.[323,328,329]
Differentially methylated positions (DMPs)	
RPS6KA2 (RSK3)	Ribsomal S6 kinase A2 is a ribosomal kinase in the serine/threonine kinase family.[330] Acts on the intracellular MAP kinase signalling pathway (interacts with MAPK1 and 3).[331] Thought to have a role in cell growth, cell motility, proliferation[331] and cell cycle progression,[332] although may also act as a tumour suppressor gene in ovarian cancer.[333] Alternatively spliced isoforms exist.[333] An IBD-associated GWAS locus.[88]
SBNO2	Strawberry notch homologue 2 has an anti-inflammatory effect, by acting in the IL-10 downstream pathway.[334] IL-10 induced SNBNO2 expression was found to repress NF-κβ (but not IRF7) selectively within macrophages.[334] A susceptibility locus in IBD GWAS.[97]
TNFSF10 (TRAIL, CD253, Apo2L)	Tumour necrosis factor superfamily member 10/TRAIL acts as a ligand in the TNF family.[335] This widely investigated cytokine induces caspase-8-dependent apoptosis in tumour but not normal cells;[336] this property has led to extensive investigation as a chemotherapeutic agent.[337] TRAIL has been implicated in intestinal inflammation and demonstrated to be over-expressed in intestinal epithelial cell lines[338] and in human mononuclear cells in inflamed intestinal sections.[339]
BCL3	B-cell CLL/lymphoma 3 is a proto-oncogene acting as a co-activator through NF-κβ and is associated with translocation in a specific form of B-cell leukaemia (t(14;19)(q32;q13)) [340]
IL23A	Interleukin 23 subunit A. The T-helper (Th)17 and interleukin (IL)12-23 pathway is well established in IBD pathogenesis, with susceptibility gene loci IL23R, IL12B, JAK2, and STAT3 identified in both UC and CD.[94,95]

Table 9 - Selected candidate genes in differentially methylated regions and positions

3.3.5 Separated cell populations provide insight into cell-type of origin of methylation signals see in whole blood

Flow cytometric assessment demonstrated high purity of isolated cell populations following immunomagnetic cell separation (CD14+ median=92.4% (IQR 87-94.9), CD4=97.3% (93.8-98.9), CD8+=88.7 (80.5-93)). *In-silico* validation of immunomagnetic separation was also performed using the Houseman algorithm (CD14 median=98.8% (IQR 93.71-100.2), CD4=98.8 (93.7-101.2), CD8+=87.2 (75.9-91.5)). Based on methylation data, samples clustered according to cell type on MDS plots (Figure 19). The gene tables for each of the DMP case-control analyses for each of the separated cell types are presented in

Appendix 1.

Data derived from cell population provides insight into cell-type of origin of methylation signals see in whole blood; RPS6KA2, the top DMP in whole blood is also hypomethylated in CD14+ monocytes (Δβ -11.7%, p=5.8 × 10^{-8}, FDR adjusted p=0.009) whilst a probe from the VMP1 region, the top DMR was hypomethylated in CD8+ lymphocytes (cg20458044, Δβ - 8.3%, p=2.3 × 10^{-6}, FDR adjusted p=0.03). Cell specific data also demonstrates a DMR present in CD14+ monocytes, HDAC4 (3 hypermethylated probes, 1253 bases, minimum p value = 4.3 × 10^{-8}, minimum FDR adjusted p=0.009). This is particularly interesting given HDAC4 is a subclass of histone deacetylase enzymes, and may indicate interaction between epigenetic mechanisms.

3.3.12 Relationship between DNA methylation and smoking

The beta values of the top DMPs were plotting against smoking status as demonstrated in Figure 22. In IBD cases there was no difference in DNA methylation according to smoking status (current, ex, never) in the top DMPs (RPS6KA2 Kruskall Wallis p =0.3, VMP1 p = 0.3, SBNO2 p =0.3, TNFSF10 p =1). Interestingly, in controls there was a significant difference in DNA methylation according to smoking status for the top DMPs (RPS6KA2 p= 0.002, VMP1 = 0.01, SBNO2 = 0.03, TNFSF10 p=0.01). When smoking was additionally included as a covariate in linear models (along with age, sex, and estimated cell proportions) there was not a huge impact on the top DMPs (Table 48).

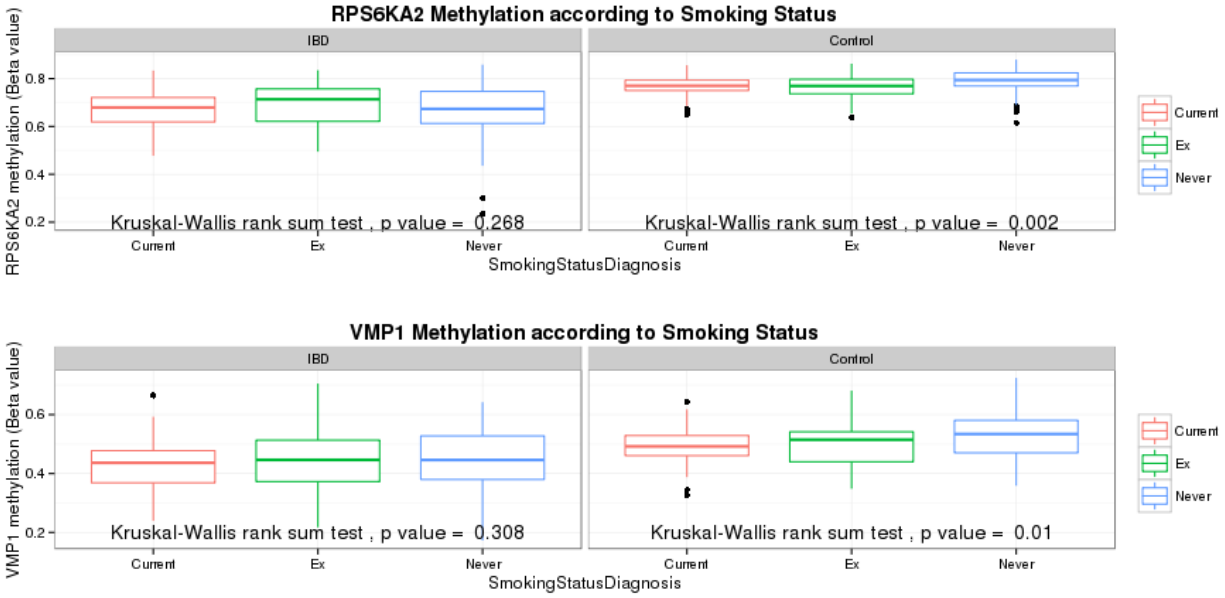


Figure 22 - Beta Methylation values for the top methylation DMPs according to smoking status

In a post-hoc analysis, the present IBD DNA methylation dataset was used to investigate the known effect of smoking on methylation status. A linear model was used to compare smokers and non-smokers amongst all cases and controls, using IBD status and cell proportions as co-variates. The differentially methylated probes were then compared with known smoking associated-probes published by Tsaprouni et al (Table 10).[152] The methylation difference also strongly correlated between the two datasets (Figure 23).

Table 10 - CpG probes correlated with Current Smokers and never smoked in Tsaprouni et al

					Current- Never		Adjusted for cell count P- values Current-Never	
					Tsaprouni	Ventham	Tsaprouni	Ventham
Type	Probe ID	Ch	Position	Gene Locus	Δβ	Δβ	p value	p value
Known Locus	cg05951221	2	233284402	ALPPL2	-13.2	-11.34	9.80E-30	5.04E-34
Known Locus	cg01940273	2	233284934	ALPPL2	-12.7	-8.46	2.33E-28	1.67E-29
Known Locus	cg21566642	2	233284661	ALPPL2	-17.6	-8.77	1.10E-25	2.52E-33
Known Locus	cg05575921	5	373378	AHRR	-27.8	-13.14	8.65E-25	2.13E-37
Known Locus	cg06126421	6	30720080	IER3	-12.5	-10.01	1.11E-22	2.44E-27
Known Locus	cg03636183	19	17000585	F2RL3	-13.4	-8.44	2.78E-19	1.49E-26
Known Locus	cg21161138	5	399360	AHRR	-8.7	-4.92	1.04E-14	2.48E-16
Known Locus	cg06644428	2	233284112	ALPPL2	-4.4	-2.38	3.17E-14	1.97E-07
Known Locus	cg19859270	3	98251294	GPR15	-3.5	-1.09	2.66E-13	4.74E-09
New Signals in Known Loci	cg03329539	2	233283329	ALPP	-5.7	-3.40	6.28E-12	6.00E-13
New Signals in Known Loci	cg24859433	6	30720203	IER3	-5	-2.57	6.33E-11	1.90E-09
New Signals in Known Loci	cg15342087	6	30720209	IER3	-4	-2.16	1.49E-10	1.95E-06
Known Locus	cg25648203	5	395444	AHRR	-6.7	-3.02	4.81E-10	2.57E-11
Known Locus	cg23480021	3	22412746	ZNF385D	13.8	1.94	6.86E-09	0.0561
New Signals in Known Loci	cg13193840	2	233285289	ALPPL2	-2.8	-1.28	5.54E-08	0.00042
New locus	cg22717080	6	166959505	RPS6KA2	-1.7	NA	6.42E-08	NA
New Signals in Known Loci	cg14817490	5	392920	AHRR	-7	-4.50	2.34E-07	2.59E-12
New Signals in Known Loci	cg24090911	5	400732	AHRR	-6	-2.68	2.48E-07	6.03E-09

New locus	cg11660018	11	86510915	PRSS23	-5.6	-4.04	2.66E-07	2.00E-11
Known Locus	cg19572487	17	38476024	RARA	-7.1	-3.41	6.91E-07	5.15E-11
New Signals in Known Loci	cg02657160	3	98311063	CPOX	-2.6	NA	1.23E-06	NA
New locus	cg20295214	1	206226794	AVPR1B	-2.4	-2.74	1.69E-06	1.04E-13
Known Locus	cg21611682	11	68138269	LRP5	-4.1	-3.01	4.12E-06	1.43E-12
Known Locus	cg27241845	2	233250370	(ALPP)	-7	-3.19	5.48E-06	4.53E-09
New locus	cg03547355	1	227003060	(PSEN2)	-2.6	-0.94	1.28E-05	0.01837
New locus	cg02451831	7	26578098	KIAA0087	-3.2	-1.38	1.34E-05	0.01598
Known Locus	cg25189904	1	68299493	GNG12	-7.9	-6.16	1.97E-05	4.24E-15
Known Locus	cg03991871	5	368447	AHRR	-6.2	-3.21	2.34E-05	4.55E-09
New locus	cg23079012	2	8343710	LINC00299	-5.3	-1.97	4.40E-05	1.11E-09
Confounded by Blood Cell Counts	cg17024919	3	21792248	ZNF385D	-6.3	-1.24	2.19E-03	0.13487

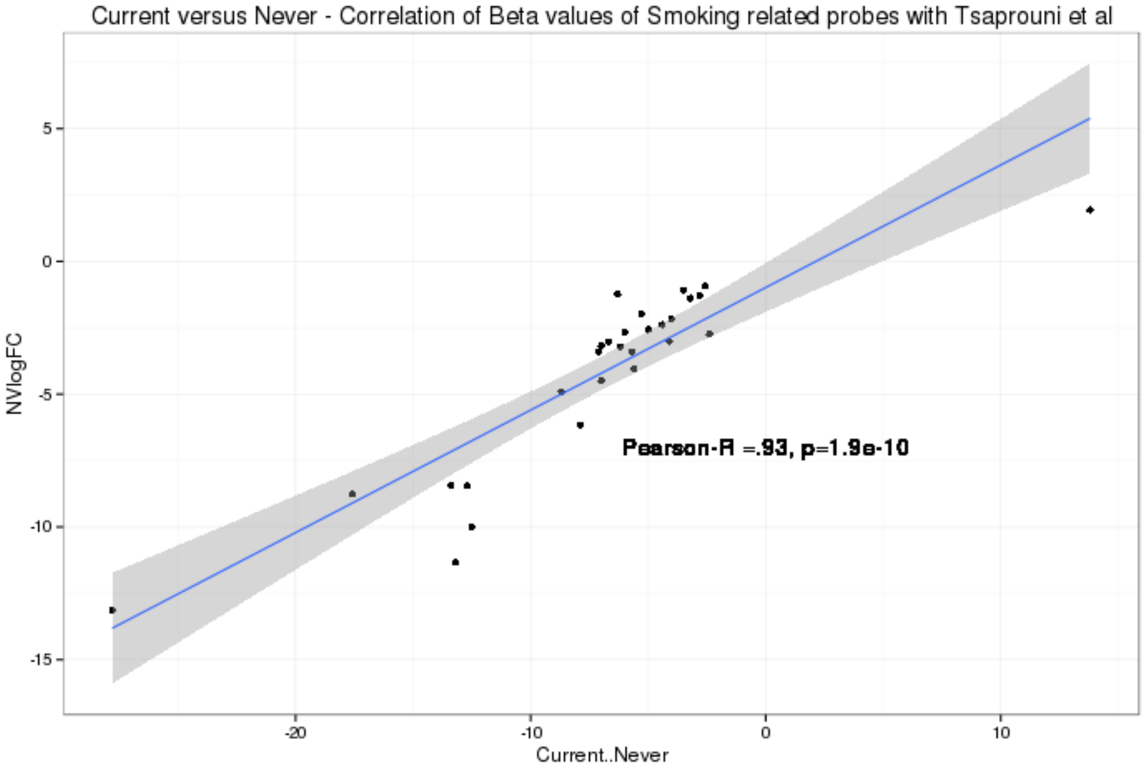


Figure 23 - Current Versus Never Smoked Correlation of Beta Values with Tsaprouni et al

3.3.13 Co-localisation with IBD GWAS loci and other disease

There was significant enrichment of the most highly differentially methylated positions with known IBD-associated GWAS loci. When compared to randomly generated bins with similar probe density, there was a statistically significant increase in the enrichment (Wilcoxon rank sum test, bin size 25kb p=0.0012, 50kb p=2.27×10^{-6}, 100kb p=4.85×10^{-11}, 250kb 1.7×10^{-20}, Figure 24). This effect appeared to be specific to IBD; there was no significant enrichment with GWAS loci in other related and non-related complex diseases.

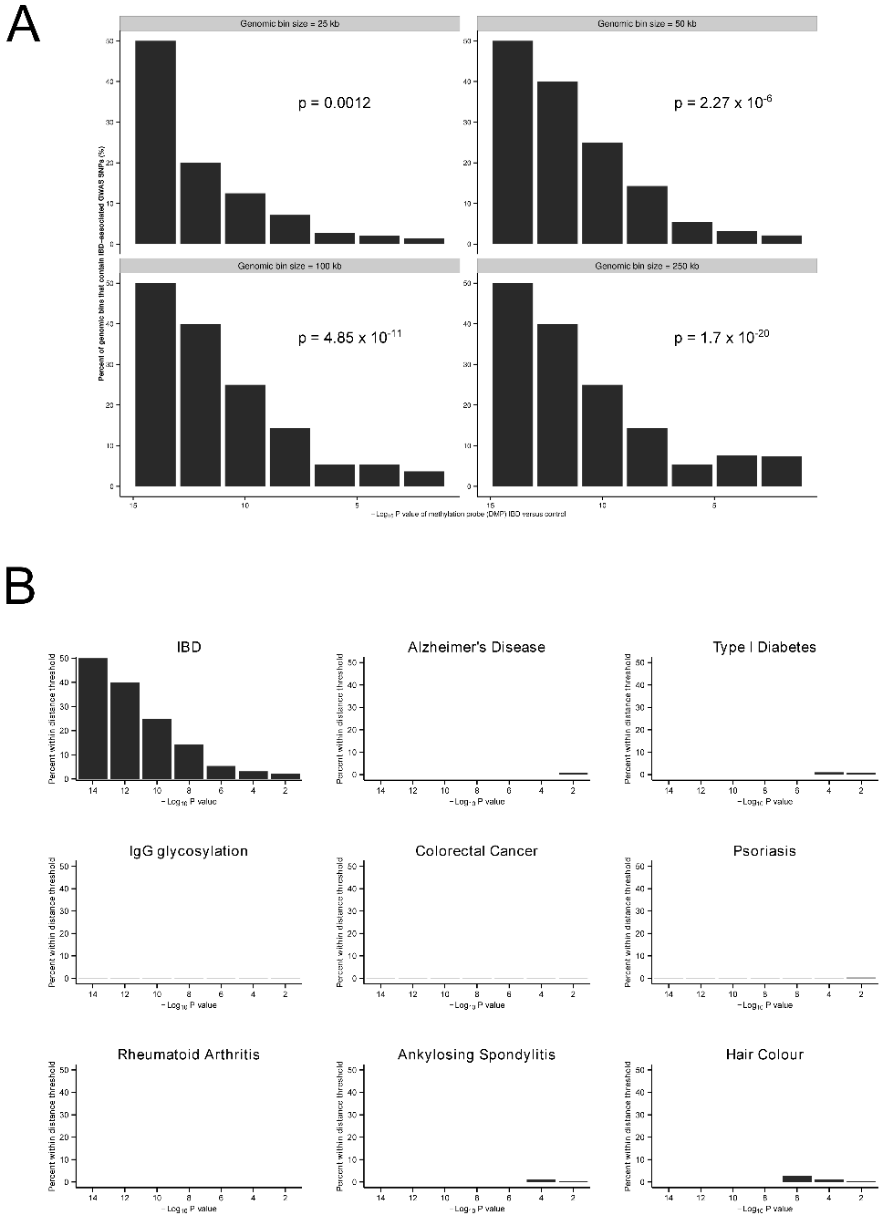


Figure 24- Co-localisation of statistically significant IBD-associated differentially methylated positions (DMP) and IBD GWAS loci. A -Each facet represents a different genomic bin size (25kb, 50kb, 100kb and 250kb). There was co-localisation of IBD-associated SNPs (y-axis) within genomic regions (i.e. defined bin) containing highly statistically significant DNA methylation probes compared those regions containing methylation probes with lower levels of statistical significance (x axis, -log 10 p value). P values denote Wilcoxon rank sum comparison with randomly generated bins with similar probe density. B – Enrichment of IBD-associated SNPs but not of SNPs associated with other complex immune traits and diseases within 50kb regions containing IBD-associated differentially methylated positions.

3.3.14 Epigenetic aging in cases versus controls

The epigenetic age of the samples was calculated using the method described by Horvath (modified by Nick Kennedy). [142] There was a high level of correlation between the actual age and calculated age based on the methylation data, this correlation was highest for whole blood and lowest for CD8 cells, but is likely to reflect the included number of samples in the analyses (Figure 25). There was no difference in the age acceleration between cases and controls (Figure 26)

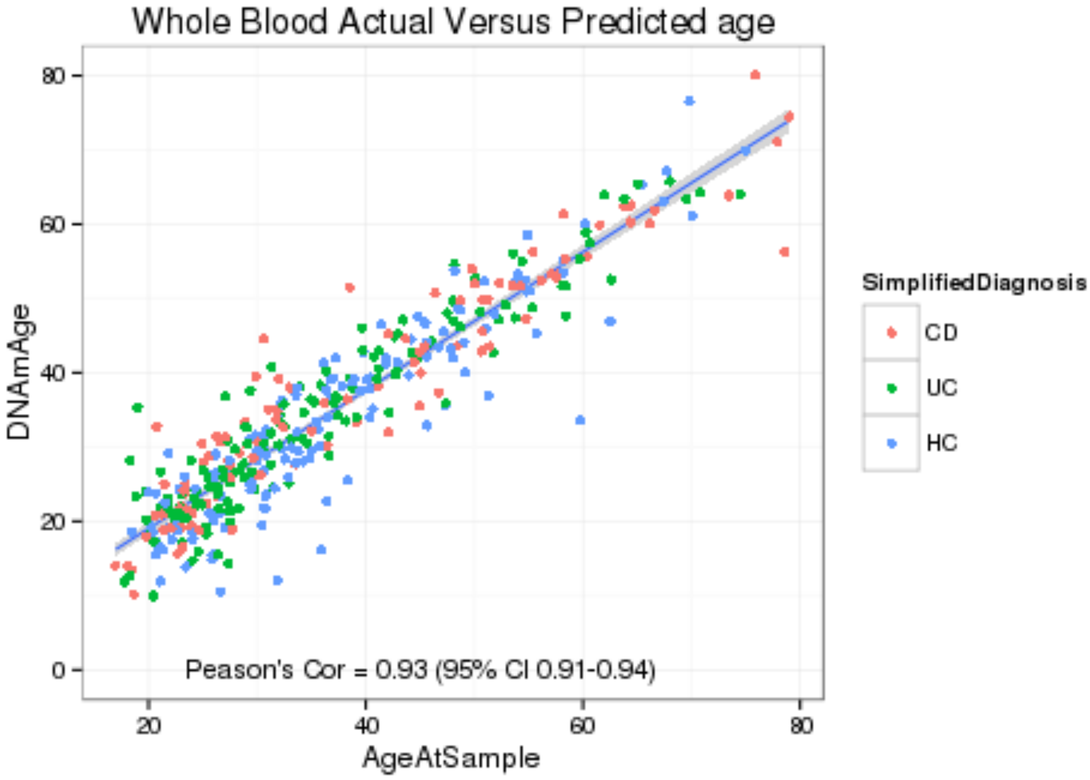


Figure 25 - Epigenetic aging correlation between actual age (X axis) and Predicted age (Y axis). Colour of points denotes diagnosis

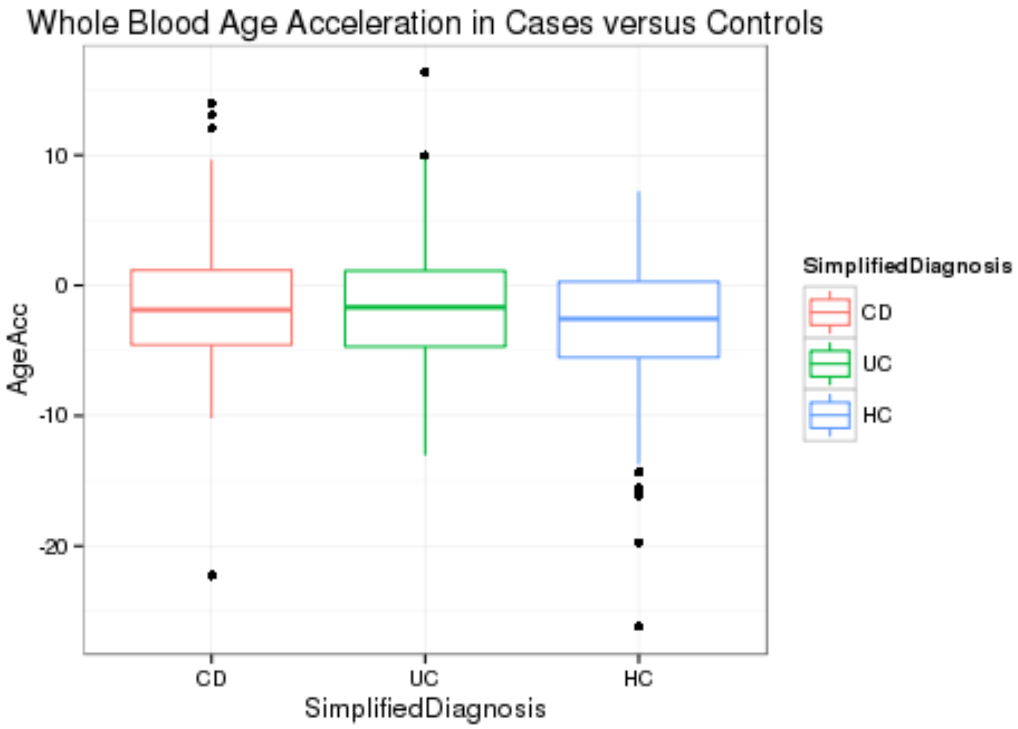


Figure 26 - Epigenetic Age Acceleration (y-axis) in Cases versus Controls

3.3.15 Gene ontology analysis

Gene ontology analysis was performed on the 439 DMPs identified in IBD versus controls in whole blood. After removing unannotated probes and duplicate probes with the same gene annotation 270 genes were included in the analysis. There were 54 significantly enriched GO terms following FDR correction for multiple testing (Table 11). Many of the included GO terms related to the immune response.

category	No. genes differentially expressed in category	Total No of genes in category	GO term	P Val	FDR Adj P Val
GO:0080134	47	1293	regulation of response to stress	1.27E-07	0.003
GO:0048518	119	4815	positive regulation of biological process	3.41E-07	0.003
GO:0044763	213	10900	single-organism cellular process	8.02E-07	0.005
GO:0044699	227	11972	single-organism process	1.02E-06	0.005
GO:0009607	32	816	response to biotic stimulus	1.61E-06	0.005
GO:0001775	34	853	cell activation	1.70E-06	0.005
GO:0002252	29	675	immune effector process	1.93E-06	0.005
GO:0048583	86	3212	regulation of response to stimulus	2.38E-06	0.005
GO:0048522	104	4164	positive regulation of cellular process	2.50E-06	0.005
GO:0051179	120	5016	localization	2.59E-06	0.005
GO:0031347	29	699	regulation of defense response	2.80E-06	0.005
GO:0002376	66	2336	immune system process	3.56E-06	0.006
GO:0043207	30	783	response to external biotic stimulus	5.01E-06	0.007
GO:0051707	30	783	response to other organism	5.01E-06	0.007
GO:0002366	13	171	leukocyte activation in immune response	5.27E-06	0.007
GO:0045321	27	635	leukocyte activation	5.86E-06	0.007
GO:0002263	13	173	cell activation involved in immune response	6.12E-06	0.007
GO:0002682	44	1365	regulation of immune system process	1.11E-05	0.012
GO:0006810	99	4044	transport	1.23E-05	0.013
GO:0051234	101	4156	establishment of localization	1.45E-05	0.014
GO:0070887	67	2444	cellular response to chemical stimulus	1.81E-05	0.017
GO:0006897	24	546	endocytosis	2.86E-05	0.026
GO:0030099	17	316	myeloid cell differentiation	3.18E-05	0.027
GO:0006952	47	1610	defense response	3.27E-05	0.027
GO:0045638	8	77	negative regulation of myeloid cell diff	3.93E-05	0.031
GO:0051049	48	1584	regulation of transport	4.45E-05	0.034
GO:0046632	8	75	alpha-beta T cell differentiation	4.72E-05	0.035
GO:0009891	49	1645	positive regulation of biosynthetic process	5.25E-05	0.037
GO:0009611	33	950	response to wounding	6.13E-05	0.040
GO:0010646	71	2666	regulation of cell communication	6.27E-05	0.040
GO:0050776	31	885	regulation of immune response	6.40E-05	0.040
GO:0002684	29	818	positive regulation of immune system	6.41E-05	0.040
GO:0044422	150	7205	organelle part	7.51E-05	0.042
GO:1902578	89	3663	single-organism localization	7.54E-05	0.042
GO:0002521	19	413	leukocyte differentiation	7.56E-05	0.042
GO:0044765	86	3512	single-organism transport	7.67E-05	0.042
GO:0051641	69	2638	cellular localization	7.72E-05	0.042
GO:0046637	6	43	regulation of alpha-beta T cell diff	8.18E-05	0.042
GO:0048534	27	706	Hematopoietic/lymphoid development	8.27E-05	0.042
GO:0016192	39	1186	vesicle-mediated transport	8.42E-05	0.042
GO:0050896	152	7392	response to stimulus	8.59E-05	0.042
GO:0048771	10	132	tissue remodelling	8.89E-05	0.042
GO:0033365	27	744	protein localization to organelle	9.73E-05	0.044
GO:0050778	23	587	positive regulation of immune response	9.78E-05	0.044
GO:0002274	10	143	myeloid leukocyte activation	0.000103	0.046
GO:0009605	61	2303	response to external stimulus	0.000108	0.046
GO:0045087	32	974	innate immune response	0.000109	0.046
GO:0009893	80	3185	positive regulation of metabolic process	0.000111	0.046
GO:0002697	17	371	regulation of immune effector process	0.000115	0.046
GO:0006955	42	1454	immune response	0.000117	0.046
GO:0023051	69	2620	regulation of signalling	0.00012	0.046
GO:0032101	27	763	regulation of response to external stimulus	0.00012	0.046
GO:1903707	9	119	negative regulation of hemopoiesis	0.000123	0.046
GO:0006950	85	3582	response to stress	0.000123	0.046

Table 11 – GO term analysis of DMPs in IBD versus control in whole blood

3.4 Discussion

This study has demonstrated several highly-significant site-specific methylation changes in IBD compared with controls. These findings are highly replicable both in an independent adult cohort and in our previous paediatric CD data. This study has used a large discovery cohort to demonstrate DMPs and DMRs and has formed a detailed characterisation of these sites in separated cells.

Disease-associated differentially methylated regions are perhaps the most compelling evidence of a methylation difference in IBD versus controls. The top IBD-associated DMR, *VMP1* (vacuole-membrane protein 1), was also one of the most significant DMPs. *VMP1* was also the principal finding in our previous paediatric dataset, and is validated here in a significantly larger cohort. The majority of methylation probes in the *VMP1* area are found towards the 3' end of the *VMP1*, in which the primary transcription site for microRNA-21 (pre-miR21) is located. This is an exciting finding given that this microRNA and has previously been implicated in colitis and IBD.[193,195–197,200,314–316] miR21 is considered by some to be a "pro-inflammatory microRNA", however is likely to have diverse actions within different pathways in different tissues.[316] Another notable IBD-associated DMR is *ITGB2* (integrin subunit beta 2), the gene of which has a role in leukocyte adhesion, activation and trafficking.[318] This is particularly interesting given the recent focus on strategies to therapeutically target leukocyte adhesion, namely vedolizumab, which targets a different integrin subunit (α4β7). [76,77] Aberrant DNA hypermethylation at the *ITGB2* locus has previous been demonstrated in IBD in mucosal[321] and peripheral blood leucocyte[307] samples as well as in other diseases.[322,323] Interestingly, the level of *ITGB2* expression, with three other genes, has been used to predict mucosal healing in ulcerative colitis.[341] Functionally, the subunits of *ITGB2* (CD18, CD11a and CD11b) have been knocked out in mice, with loss of CD18 and CD11a being associated with an attenuation of dextran-sulphate induced colitis.[342] An older study demonstrated that antibodies directed against CD11b/CD18 reduced gut inflammation in rats.[343] Clinically, an hereditary loss of CD18 function is known as leucocyte adhesion deficiency (LAD-1) and whilst not associated with gastrointestinal features, a Crohn's-like manifestation has been reported in the literature.[344] The other DMRs are also of great interest: *WDR8* or *WRAP73* (WD (trp-asp) repeat protein family, antisense to Trp73) which is involved in several cellular and gene regulatory processes,[317] and *TXK* a tec kinase in the tyrosine kinase family expressed in T-cells is also an IBD GWAS susceptibility locus (Chr 6, rs6837335) and is highly expressed in intestinal lesions found in Bechet's disease.[327]

Whilst DMRs have been considered as the hallmarks of differential methylation, DMPs should also not be overlooked, especially given the design of the 450k array may prevent certain DMPs from being identified as DMRs as a result of sparse coverage in certain genomic regions or the manner of DMR detection, the method has yet to reach a consensus within the literature. The top DMP was *RPS6KA2*, a ribosomal kinase in the serine/threonine kinase family [330] that regulate a diverse set of cellular processes including cell growth, cell motility, proliferation [331] and cell cycle progression.[332] *RPS6KA2* is found within an IBD-associated GWAS locus[88] (GWAS data from Jostins et al, rs1819333, p= 6.76×10^{21}, OR = 1.081) on chromosome 6, which also contains the genes *CCR6, RNASET2, FGFR10P, GPR31* and *TCP102L*. Differential methylation of *RPS6KA2* has previously been linked with smoking,[152] and although difference in smoking status does not explain the difference in *RPS6KA2* methylation in IBD cases and controls in this dataset, this may be an avenue for future research given smoking is a known environmental modifier of the disease. Another of the most DMPs was *SBNO2*, Strawberry notch homologue 2, which appears twice in the top 10 DMPs, and all 4 probes annotated to this gene demonstrate hypomethylation in IBD cases compared to controls. *SBNO2* is known to have an anti-inflammatory effect, by acting in the IL-10 downstream pathway.[334] *SBNO2* is found within a IBD-associated GWAS locus.[97] Other highly interesting DMPs implicated in well-known IBD pathogenic pathways include Interleukin 23 subunit A (*IL23A*), another IBD GWAS-susceptibility locus,[94,95] and Tumour necrosis factor superfamily member 10 (*TNFSF10/TRAIL*).

This is the largest study of DNA methylation in inflammatory bowel disease to date. Several other studies have investigated the DNA methylation pattern in IBD cases compared with controls. This includes our own study into paediatric treatment naïve Crohn's disease. In the present study again we have chosen to study newly diagnosed adult patients, where the impact of chronic inflammation and immunomodulation drugs and surgery on the epigenetic landscape will be minimised. Despite efforts to recruit newly diagnosed patients, only half of the included patients (with available data, Table 6) were treatment naïve, with the remainder having been exposed to a short period of immunosuppressive therapy (usually days). The extent of methylation change attributable to medication has not been quantified here, but this together with the temporal change in methylation profile with disease course would be worthy of further study.

Aside from the present study, the next largest blood based methylation study in IBD has been published by McDermott et al.[307] Whilst this study used a large cohort of both CD and UC patients, the patient cohort had established disease with varying degrees of disease activity and immunosuppressive duration. Further strengths of our own study over the previously published data is the inclusion of a large control cohort of both symptomatic and healthy controls and the large independent validation cohort.[307] Like previous studies,[307,321] the present study found no significant differential methylation between CD and UC. Some genetic loci that were initially thought to be CD- or UC- specific have since been found to associate with both diseases,[88] and IBD transcriptomics also demonstrates a largely shared gene expression profile in both diseases.[345]

The impact of cellular heterogeneity on DNA methylation data is a commonly cited limitation of EWAS studies conducted using whole tissue.[346,347] Statistical techniques such as the Houseman algorithm are now becoming increasingly accepted as a robust method for correcting for cellular heterogeneity.[205,302] A significant strength of the present work is the well-powered study of DNA methylation in a large cohort of whole tissues, with detailed characterisation of separated leukocytes in a subset with DNA methylation data. The small absolute number of samples included in the separated cell analyses limited the statistical power to determine significant differences in DNA methylation between cases and controls. Despite this limitation, these data provided tantalising clues in unmasking the cell of origin of the DNA methylation signals in whole blood. For example, the top DMP, *RPS6KA2* was differentially methylated in CD14+ monocytes in UC, whilst *VMP1* the top DMR was differentially methylated in CD8+ T-cells. There are few studies into DNA methylation using separated cells in the context of IBD.[348] Should the primary aim of a study be to explore biology, then performing a highly detailed characterisation of the methylome in separated cells is warranted, whilst the difficulty in enriching samples for these cell types makes them less attractive as biomarkers, and here whole tissue such as blood may be more useful. Blood has been used as the index tissue in this study for several reasons. IBD is an immune mediated disease with many of the currently used therapeutics targeting peripheral leukocytes. IBD has known extra-intestinal manifestations. Severe, treatment refractory CD can be treated using autologous stem cell transplant[349,350] and CD has been known to recur in transplanted intestinal tissue[351] indicating that IBD is not exclusively propagated at the gut level. Unlike gut

based markers, peripheral blood is easily accessible and thus attractive as a non-invasive biomarker. SEPT09, a commercially available blood-based biomarker for colorectal cancer demonstrating the feasibility of DNA methylation biomarkers in GI disease.[352,353]
Several methodological aspects have been explored in this chapter. There is some debate amongst the epigenetic scientific community regarding the most appropriate normalisation methods for Illumina 450k data. Marbita et al found that quantile normalisation and BMIQ was the most effective.[289] This study also suggested that it was necessary to correct for batch effects. [289] A study examining batch effects in gene expression arrays found that ComBat was the most effective method for correction.[354] In this chapter a combination of these methods (rather than using a complete 'pipeline' analysis) have been used. Principal coordinate analysis plots were used to visualise the effect of these normalisation methods on the data.

Reassuringly this dataset independently validates previous work demonstrating hypomethylation of specific probes related with smoking.[152] Strongly statistically significant findings were determined when comparing current- and ex-smokers with non-smokers and the most significant findings coincided with the previously published 'smoking-associated' probes. Moreover there was a strong correlation between the level of hypomethylation between the two datasets. Interestingly, one of the most significant DMPs when comparing IBD with control, *RPS6KA2* was defined by Tsaprouni et al as a 'smoking-associated' probe. However this was not significantly differentially methylated between smokers and non-smokers in this dataset, adding credibility that differential methylation of this probe is disease- and not smoking-associated.

The 'epigenetic clock' developed by Horvath that can predict age based on DNA methylation data, has demonstrated that diseased tissues from certain conditions have accelerated aging compared with controls. Such conditions include obesity and non-alcoholic fatty liver disease,[355] Down's syndrome,[356] and HIV infection.[357] Certain cancer tissues demonstrate marked age acceleration (e.g. luminal breast cancer) whereas others demonstrate negative age acceleration e.g. basal breast cancer).[358] This dichotomy may be explained by mutations to different pathways; cancers with few somatic mutations exhibit increased age acceleration, whereas cancers with p53 mutations demonstrated decreased age acceleration.[358] Whilst no difference in age acceleration was seen in this dataset, this may relate to tissue type, which is known to affect age acceleration within the same individual,[359] and sampling DNA methylation age in gut tissue may yield different results.

This is the most detailed characterisation of the epigenome carried out in IBD to date. These data are further explored in upcoming chapters in this thesis, with regards to validation of findings, application as biomarkers and the relationship between DNA methylation data and genetics and gene expression.

Chapter 4. Targeted replication of whole genome DNA methylation findings in IBD

Abstract

Introduction

Highly significant IBD-associated differences in DNA methylation have been presented in Chapter 3. Replication is a critical requirement of genome-wide and now epigenome-wide association studies in order to limit the impact of non-biological (technical) variation and to reduce false positives. The two aims of this chapter were firstly to perform technical validation of 450k microarray findings and secondly to replicate the findings from chapter 3 in an independent cohort.

Methods

Technical validation and replication of 450k array findings was performed using pyrosequencing. Pyrosequencing primers for the most significant DMP (*RPS6KA2*) and DMRs (*VMP1, IGTB2, TXK*, and *WRAP73*) were designed. Pyrosequencing was performed on the pyromark q24 platform. Technical validation was performed in a subset of patients included in 450k microarray analyses (n=231). Replication was performed in an independent cohort of patients with established IBD (n=240 [CD=121, UC=119]) and controls (n=98). Further replication of 450k methylation results was performed using previously published paediatric Crohn's disease data.

Results

Technical validation was achieved with pyrosequencing results strongly correlating with 450k array beta values for all DMP and DMRs tested. Using pyrosequencing, statistically significant differences in methylation were observed in cases and controls with the same direction of methylation change demonstrated on both platforms.

The 450k methylation results from chapter 3 were replicated in an independent cohort using pyrosequencing with significant results for the most significant DMP (*RPS6KA2*, IBD versus controls p=1 × 10^{-9}) and DMRs (*VMP1* p=1 × 10^{-6}, *IGTB2* p=2 × 10^{-7} and *TXK* p=4 × 10^{-10}).

There was a highly significant correlation between the difference in beta values for the top 5000 differentially methylated probes between CD cases and controls in the present dataset and the previously published early-onset CD methylation data (Pearson's correlation 0.77, 95% confidence interval 0.76-0.78, p-value < 2.2×10^{16}).

Conclusions

Persuasive differences in DNA methylation associated with IBD have been technically validated and replicated in more than one independent cohorts increasing confidence in the results.

4.1 Introduction

Highly significant IBD-associated differences in DNA methylation have been presented in Chapter 3. Whilst this represents the largest study of its kind in DNA methylation in IBD to date, it remains critically important to replicate these findings. Lessons learnt from GWAS highlight the importance of independent replication cohorts, in order to account for non-random technical biases and reduce the incidence of false positives.[360] Replication in epigenome-wide association studies (EWAS) is likely to be as important as in GWAS,[202,361] especially given the increased number of potential confounders in EWAS.[281] Many of the findings from early EWAS have yet to be replicated.[281] Pyrosequencing provides a logical platform for replication studies as large numbers of samples can be analysed at specific methylation sites in a targeted fashion (Figure 27).

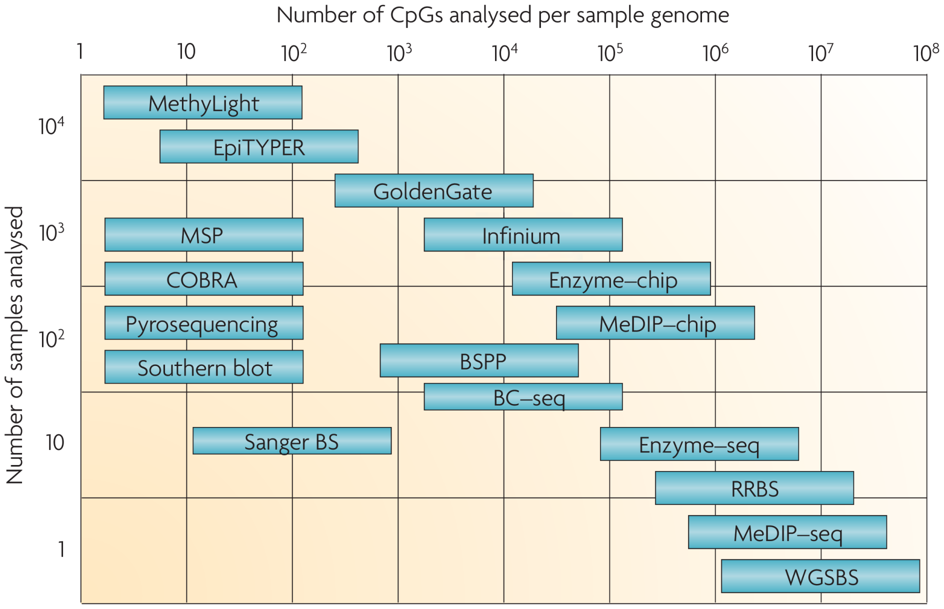


Figure 27 - Methods of DNA methylation analysis (Sample throughput versus genome coverage, reused with permission from Laird PW Nature Reviews Genetics 2010; 11:197 license number 3770261146857)

The two principal aims of this chapter were to

1. Perform technical validation of 450k microarray findings
2. Replicate the findings from chapter 3 in an independent cohort

4.2 Methods

4.2.1 Designing pyrosequencing probes

Pyrosequencing probes were designed using the pyromark assay design software (version 2.0.1.15, Qiagen, Dusseldorf, Germany). The genomic location of relevant CpG probes was extracted from Illumina 450k feature data and plotted in the UCSC genome browser. The DNA view function was used to generate raw sequence 500 bases around the CpG of interest. The assay design software was used to design primers (with the Allow Primer Over Variable Position mode turned OFF). Primer sets of a score greater than 70 were considered (Pyromark assay design own penalty system, scored between 100 and 0, zero being the worst). Primer sets were closely scrutinised for mispriming sites and checked for commonly occurring SNPs using ensemble and genome browser (excluded if minor allele frequency (MAF) >0.01). Primers were ordered from Sigma-Aldrich (St Louis, USA) with the reverse primer being HPLC purified and biotyinlated on the 5' end. The pyromark Q24 software was used to create pyrosequencing run files. Methylation ranges were increased from 0 to 100 % and bisulphite treatment controls were added. Primers were made up to 100 μM using a variable volume of TE as directed by the technical datasheet. A stock solution of 5 μM was made up with a twenty fold dilution (30 μL of 100 μM in 570 μL of TE). One microliter of 5 μM stock solution was used per 25 μL reaction to yield a final concentration of 0.2 μM.

4.2.2 Bisulphite conversion

The EZ-96 DNA methylation kit (D5003, Zymo Research, Irving CA USA) was used to bisulphite convert DNA for pyrosequencing. An input of 500ng of DNA was used (10 μL of 50ng/ μL). To the initial DNA, 5 μL of M-Dilution buffer and 35 μL of purified water was mixed. The 96-well conversion plate containing samples was incubated at 37 °C for 15 minutes. Following the incubation, 100 μL of CT conversion reagent was added and mixed. A foil lid was used to seal the conversion plate, and incubated on a thermal cycler overnight for 16 hours at 50 °C (lid heated to 100 °C), following which samples were held at 4 °C. The samples were incubated on ice for 10 minutes. The Silicon-A binding plate was placed above the reservoir plate and loaded with 400 μL of M-binding buffer, to which the samples were added and mixed. The Silicon-A binding plate was centrifuged at 3000 x g for 5 minutes. At each stage the flow through was discarded from the plate reservoir. Each sample was washed with 500 μL of M-wash buffer and centrifuged at 3000 x g for 5minutes. To each sample, 200 μL of M-desulphonation buffer was added and incubated at room temperature for 15 minutes. The samples were washed twice with M-wash buffer as previously described, with the second wash step using a centrifuge time of ten minutes. The sample was the eluted into a 96-well elution plate using 35 μL of M- elution buffer (centrifuged at 3000 x g for 3 minutes). Samples were stored at -20 °C until the PCR was performed.

4.2.3 Pre-pyrosequencing PCR

Prior to pyrosequencing the bisulphite converted DNA was amplified for target sequences using PCR (PyroMark PCR kit, Qiagen, Dusseldorf, Germany). A mastermix was made up of PyroMark PCR mastermix (12.5 μL per sample), CoralLoad concentrate (10x, 2.5 μL), RNAse free water (7 μL) and forward (1 μL) and reverse pyrosequencing primers (1 μL) were added. Within a 96 well plate was cut to 3x8 (24 well pyrosequencer) 1 μL of bisulphite converted DNA was added to each target reaction (in duplicate), but omitted from a no-template control reaction. The PCR protocol suggested by the manufacturer was used (95 °C for 15minutes, 45 cycles of 30 secs of 94 °C, 30 secs of 56 °C, 30 secs of 72 °C and a final extension of 72 °C for 10 minutes). Following PCR, the product was checked on an agarose gel. The agarose gel was produced by reserving 150ml of 800mls (780 mL water, 30 mL 0.5X TBE) and adding 2.25g of agarose to create a 1.5% gel (microwaved for 2 minutes until boiling, cooled under cold tap). Prior to setting (20 mins approx.), 15 μL of SYBR Safe gel DNA stain (Life Technology) was added and mixed. The Bioline hyperladder 1kb ladder (5 μL, London UK) was used and 5 μL of sample was added to each well. The gel electrophoresis was run at 150mV for 30 minutes.

4.2.4 Pyrosequencing

A mastermix was created composed of streptavidin-coated Sepharose beads (2 μL per sample, vigorously shaken to resuspend beads), binding buffer (40 μL) and high purity water (18 μL). 60 μL of the mastermix was added to each of the PCR wells containing 20 μL of PCR product. Strip cap tubes were applied to the 96 well plate and the plate was agitated on a vortex for 10 minutes at 1400rpm (room temp) to resuspend the beads. The PyroMark Q24 Vacuum workstation (Qiagen, Dusseldorf Germany) was used to separate DNA strands and clean samples prior to pyrosequencing. The sequencing primers were diluted to 0.3uM (1 μL of 100 μM in 333 μL of annealing buffer (cat no 979009, Qiagen)) and 25 μL of each diluted sequencing primer was added to the final pyroMark Q24 plate (corresponding to same position in PCR plate). The PyroMark Q24 Vacuum workstation was filled with the appropriate reagents in each of the trays (70%, ethanol, denaturation solution, wash buffer and high purity water). The vacuum pump was started and the filter probes were initially flushed with high-purity water. The PCR plate was removed from the vortex and within 1 minute placed on the workstation in the same orientation to the PyroMark Q24 plate. With the vacuum switched on, prongs of the vacuum head were inserted into the PCR plate to aspirate PCR bead solution with care taken not to knock beads from the tips of the vacuum tool. The vacuum tool is passed through each of the solutions (with vacuum on) containing 70% ethanol (5 seconds), denaturation solution (5 seconds) and wash buffer (10 seconds, timed rigorously using an electronic timer). The vacuum headset was held at a 90° vertical angle for 5 seconds to dry the beads. With the vacuum turned off the prongs of the headset were inserted into the PyroMark Q24 plate containing the sequencing primers and agitated to disperse the beads. The tool was rinsed in high purity water. The PyroMark Q24 plate was placed within the plate holder and positioned in a thermal cycler at 80° for 2 minutes. The plate was allowed to cool for 2 minutes before inserting the plate into the PyroMark Q24 pyrosequencer. The pyrosequencer cartridge was pre-filled with the appropriate volumes of enzyme solution, substrate solution and nucleotides in the respective slots within the cartridge. Samples were run in duplicate.

4.2.5 Statistical analysis

Analysis was initially performed using pyromark q24 software (2.0.26) using default quality assurance settings for peak height, width and bisulphite conversion controls. Samples that either passed or passed with caution on quality assurance pyromark software were included in analyses. Data were extracted from pyromark software and analyses in R (version 3.2.2, R statistical programming, Vienna Austria). The coefficient of variation (CV) calculated for duplicate samples and samples with a CV of ≤ 10% were included in analyses. Wilcox rank sum test was used to compare methylation values between cases and controls. Correlation between beta values (450k array data) and methylation percentages (pyrosequencing) was performed using Pearson's correlation coefficient.

4.3 Results

4.3.1 Technical validation of 450k array results using pyrosequencing

A subset of the complete adult 450k cohort was used to technically validate the most significant DMP (*RPS6KA2*) and DMRs (*VMP1, IGTB2, TXK*, and *WRAP73*) identified using the Illumina 450K microarray platform using pyrosequencing. The demographics of the subset of patients used for pyrosequencing are displayed in Table 12.

	IBD (n=130)	**Control** (n=101)	p Value
Female (%)	51 (50.5%)	47 (36.2%)	0.3◊
Age in years	29.2 (25.9-42.4)	30.7 (26 -38.9)	0.9†
Current Smokers (%)	36 (27.7%)	20 (19.8%)	0.2◊
C-Reactive Protein	8.5 (3-33.5)	2 (1-4.5)	1.7e-05†
Albumin	34 (26.3-38)	41 (38-43)	2.3e-07†
White cell count	8.3 (5.8-12.3)	5.7 (4.8-12.3)	6.1e-07†

Table 12 - Patient demographics of subset of adult cohort used for technical validation studies using pyrosequencing (Results are median and interquartile range unless stated, WCC=white cell count, CRP=C-reactive protein, IQR= interquartile range, IBD=inflammatory bowel disease, † = Wilcoxon rank sum test, ◊ = χ2 test)

The pyrosequencing results strongly correlated with 450k microarray beta values for all DMP and DMRs tested (Table 13, Figure 28 upper panel). Technical validation demonstrated statistically significant differences in methylation between cases and controls with the same direction of methylation change in pyrosequencing and 450k arrays was observed in the five DMRs and top DMP (Figure 28, lower panel).

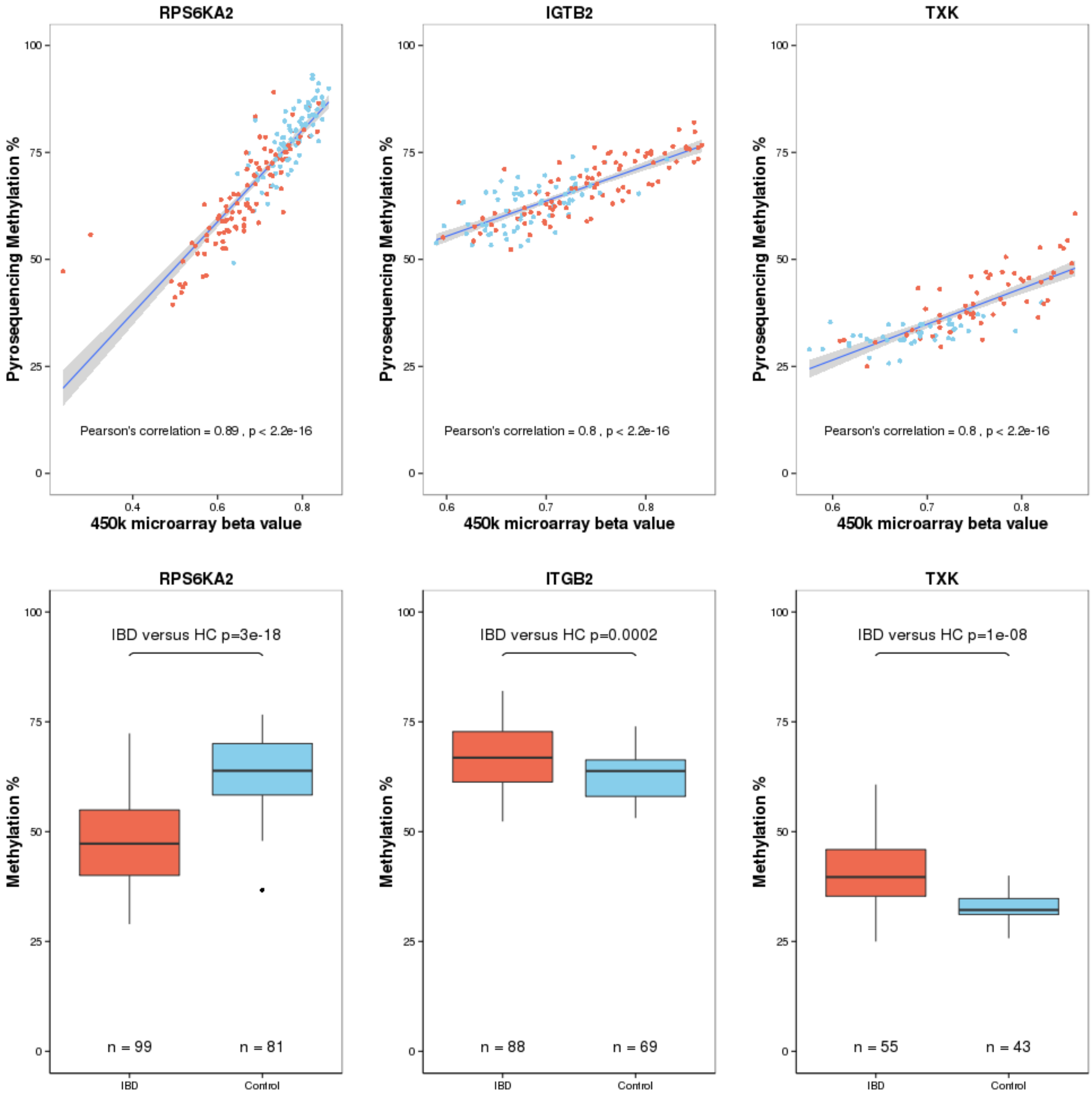


Figure 28 - Technical validation of 450k microarray results using pyrosequencing. The top panel demonstrates correlation between methylation percentage (pyrosequencing) and beta values (450k microarray). The red dots represent IBD cases, and the blue dots represent controls. The bottom panel demonstrates the methylation difference (%) between IBD cases (red) and controls (blue) using pyrosequencing.

	Pearson's correlation coefficient	Lower CI	Upper CI	P value
RPS6KA2	0.89	0.86	0.92	2.2e-16
IGTB2	0.8	0.73	0.85	2.2e-16
TXK	0.8	0.71	0.86	2.2e-16
VMP1	0.55	0.43	0.66	1.2e-12
WRAP73	0.75	0.68	0.81	2.2e-16

Table 13 - Technical replication: Correlation coefficient values between Methylation 450k microarray and pyrosequencing for same samples. CI = 95% confidence interval

4.3.2 Replication of 450k array methylation results in independent cohorts

Two methods were used to provide validation of differentially methylated positions and regions; validation using an independent previously published 450k cohort in paediatric Crohn's disease[284] and secondly using pyrosequencing in an independent cohort.

4.3.2.1 Previously published Paediatric Crohn's disease DNA methylation dataset

There was a strong correlation between the difference in beta values between the top 5000 differentially methylated probes in CD cases and controls in the present adult dataset and our previously published early-onset CD methylation data (Figure 29, Pearson's correlation 0.77, 95% confidence interval 0.76-0.78, p-value < 2.2×10^{-16}).[284] Table 14 demonstrates the top DMPs in the present adult 450k dataset with the corresponding probes beta methylation difference and level of statistical significance seen between CD and control in the paediatric dataset.[284]

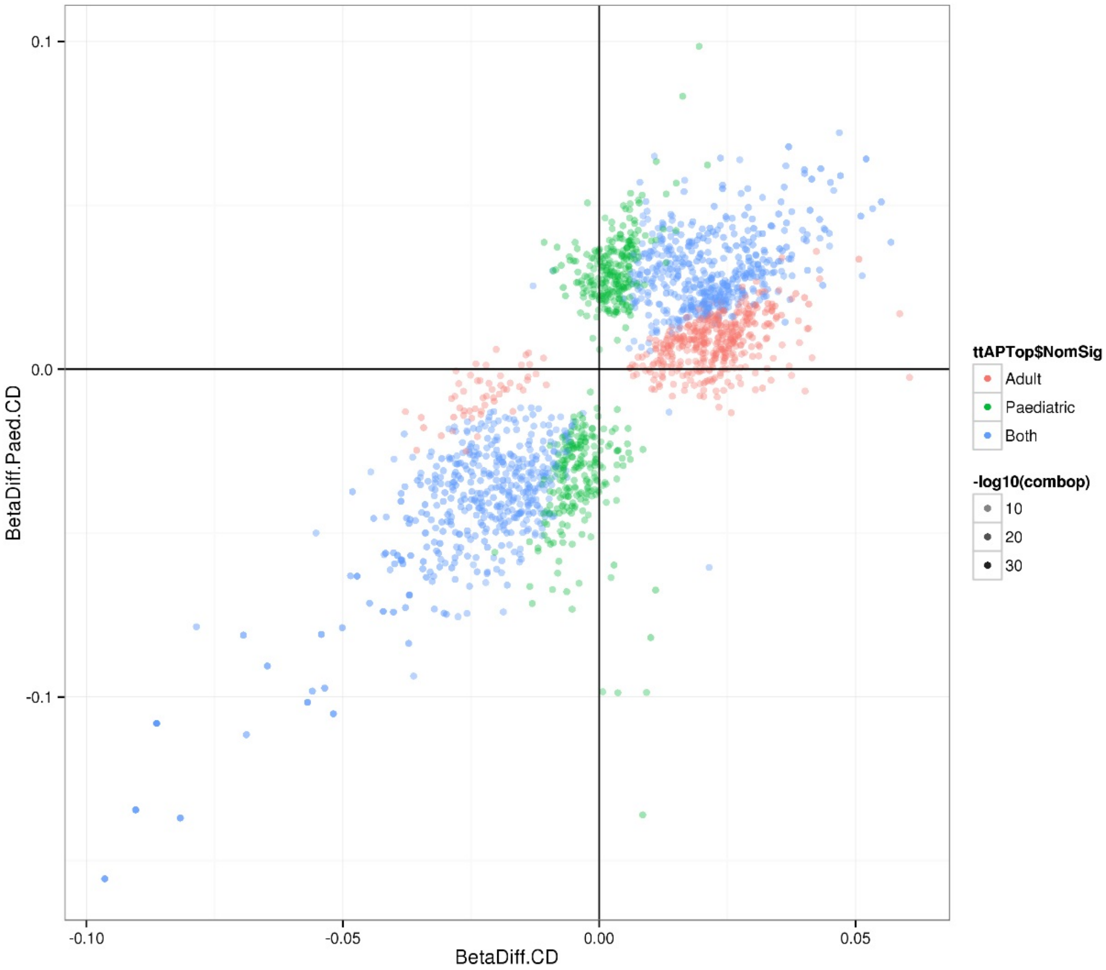


Figure 29 - Correlation between DNA methylation Beta values in Crohn's disease in the present adult 450k dataset and the paediatric Crohn's disease dataset (Appear in top 1000 most significant differentially methylated probes in both datasets (blue), adult but not paediatric (red) and paediatric but not adult (green).

Rank CD	Probe ID	Chr	Gene Symbol	Δβ CD	Holm adj P Value CD	Paed CD rank	Paed CD Δβ	Paed Holm adj P Value
1	cg17501210	chr6	RPS6KA2	-0.09	3.82E-19	2	-0.11	1.0E-09
2	cg18608055	chr19	SBNO2	-0.08	1.00E-16	29	-0.14	7.5E-05
6	cg09349128	chr22	NA	-0.05	7.82E-13	20	-0.06	6.3E-05
3	cg12170787	chr19	SBNO2	-0.05	2.91E-14	39	-0.11	1.9E-04
5	cg16936953	chr17	VMP1	-0.10	1.19E-13	4	-0.16	3.0E-08
7	cg12992827	chr3	NA	-0.06	7.74E-11	3	-0.10	2.4E-08
12	cg25114611	chr6	NA	-0.04	6.84E-09	102	-0.04	1.3E-03
10	cg02448796	chr1	KCNAB2	0.05	2.98E-09	2005	0.05	7.7E-02
4	cg12054453	chr17	VMP1	-0.09	1.15E-13	1	-0.13	8.9E-10
21	cg07398517	chr3	NA	-0.04	3.40E-08	27	-0.06	7.4E-05
19	cg13619623	chr7	BBS9	0.04	2.31E-08	13925	0.03	3.2E-01
65	cg16724148	chr1	AGL	0.03	1.47E-06	45887	0.01	5.2E-01
16	cg26804423	chr7	ICA1	0.04	1.25E-08	13175	0.03	3.1E-01
17	cg19821297	chr19	NA	-0.05	1.83E-08	9	-0.08	9.2E-06
60	cg22959742	chr10	FRMD4A	0.04	1.08E-06	6879	0.04	2.1E-01
13	cg03546163	chr6	FKBP5	-0.07	6.84E-09	59	-0.08	6.0E-04
64	cg01059398	chr3	TNFSF10	-0.04	1.44E-06	12	-0.07	1.2E-05
25	cg26955383	chr10	CALHM1	0.04	4.04E-08	38423	0.01	4.9E-01
32	cg10636246	chr1	AIM2	-0.04	1.41E-07	1736	-0.05	6.7E-02
24	cg18942579	chr17	VMP1	-0.05	4.04E-08	15	-0.10	2.5E-05

Table 14 - Table comparing top DMPs in present adult 450k dataset

(yellow, left of figure) with same probes in previously published paediatric Crohn's disease data from Adams et al (orange, right of figure). The hatched areas related to results for Crohn’s disease versus controls. Paed = Paediatric, adj = adjusted, CD = Crohn’s disease, IBD – Inflammatory bowel disease

4.3.2.2 Independent pyrosequencing cohort

The most significant DMP (*RPS6KA2*) and DMRs (*VMP1, IGTB2, TXK*, and *WRAP73*) occurring within annotated genetic regions were replicated in an independent cohort using pyrosequencing. The demographics of the replication cohort can be seen in Table 15.

		CD **(n=121)**	**UC** **(n=119)**	**Controls** **(n= 98)**
Age		36.5 (26.4-47.8)	36.6 (27.1-46.7)	34.9 (27.3-55.5)
Females *(%)*		68 *(57)*	74 *(62)*	65 *(66)*
Smoking status	Current *(%)*	18 *(15)*	7 *(6)*	14 *(14)*

Table 15 - Patient demographics of Independent pyrosequencing cohort (data presented are medians (interquartile range) except where specified)

Significant differences in methylation were seen in an independent cohort of IBD cases and controls, for the DMP (*RPS6KA2*, IBD versus controls p=1 × 10^{-9}) and DMRs (*VMP1* p=1 × 10^{-6}, *IGTB2* p=2 × 10^{-7} and *TXK* p=4 × 10^{-10}). The *WRAP73* was performed in a subset of the independent replication cohort, but was not performed in the entire cohort (IBD [n=32], control [n=17], p=0.05, Figure 70). Each assay demonstrated the same direction of methylation change (Figure 30).

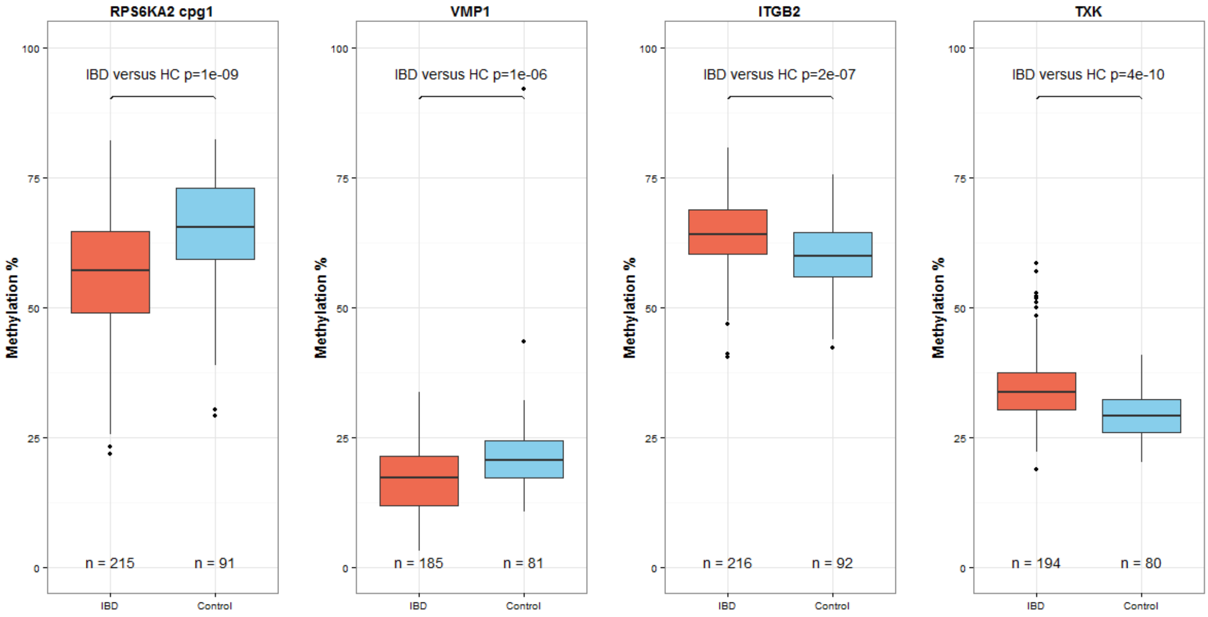


Figure 30 - Validation of top differentially methylated regions and position using pyrosequencing in an independent cohort. HC = healthy control. Wilcoxon Rank sum test used to compare groups.

When broken down in to individual disease types, both CD and UC demonstrated methylation differences in the same direction compared with controls (Figure 31).

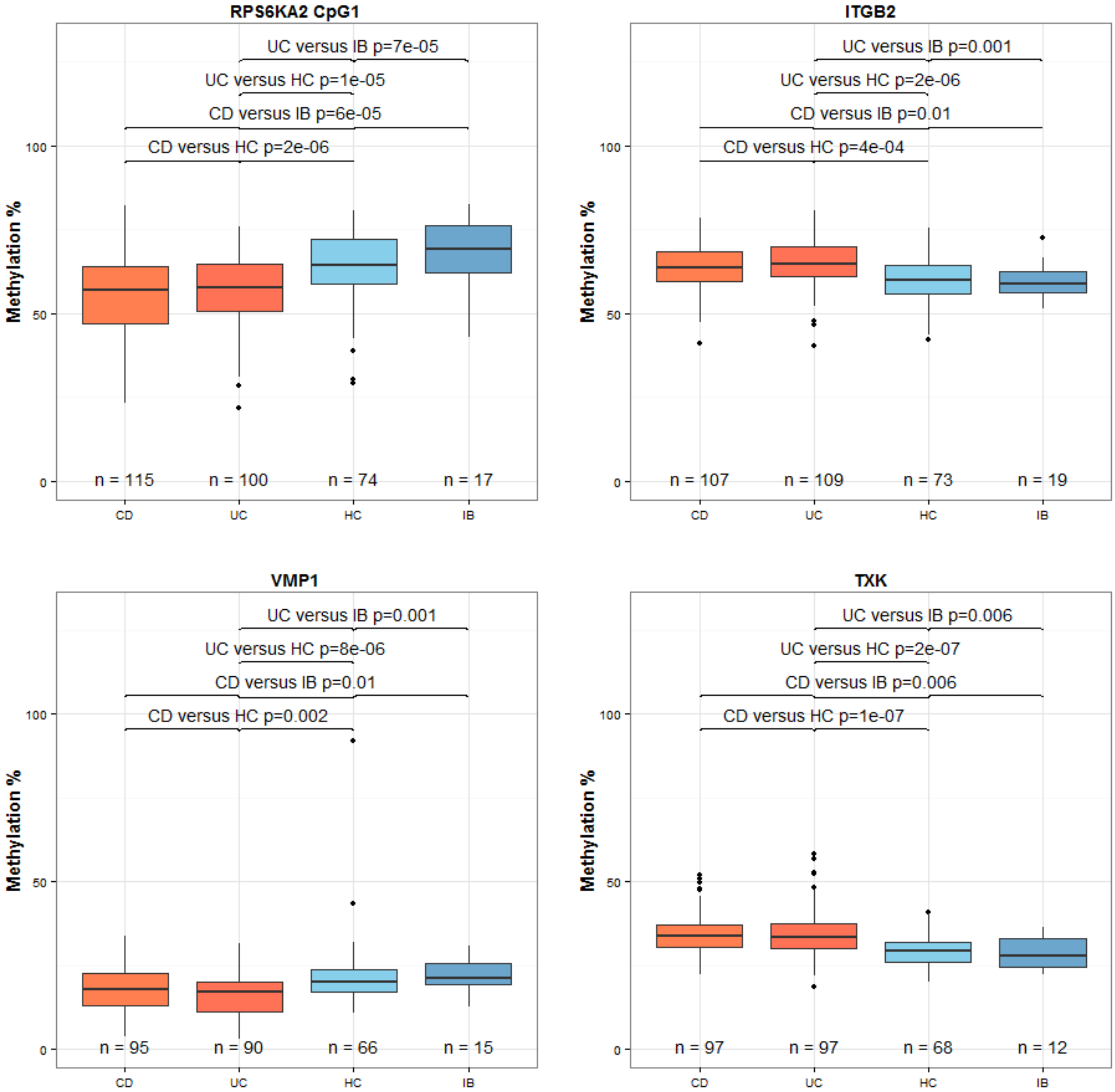


Figure 31 – Replication pyrosequencing in individual disease types. Methylation differences in Crohn's disease (CD) and ulcerative colitis (UC) compared to healthy controls (HC) and symptomatic controls (IB) for the top DMP (RPS6KA2) and DMRs (VMP1, ITGB2, TXK)

Discussion

The Illumina 450K has previously been extensively validated using pyrosequencing.[362] Like other previous 450k array-based DNA methylation studies[168,284,307,321], this study demonstrates the ability to technically replicate 450k array findings using pyrosequencing. Pyrosequencing and 450k array results demonstrated a high level of correlation for the top DMP (*RPS6KA2*) and most DMRs (*VMP1*, *IGTB2*, *TXK*, and *WRAP73*). The Illumina 450k platform is now considered a relatively mature technology and technical replication is probably not required, and more important is replication of findings in independent cohorts (Prof Stephan Beck, oral presentation, Infinium HumanMethylation450 3rd workshop – UCL April 2014). Aside from the present study, the next largest blood based methylation study in IBD has been published by McDermott et al.[307] Whilst this study used a large cohort of both CD and UC patients, the study lacked an independent replication cohort in adult peripheral blood samples. Instead this study used previously published data[321] from treatment-naïve mucosal samples form children with IBD and controls.

In this chapter replication in an independent cohort has been performed using two methods: replication using an independent adult cohort with pyrosequencing and secondly by utilising the previously published paediatric CD 450k data. The independent adult cohort used for pyrosequencing consisted of patients with established disease with varying degrees of active inflammation and exposure to therapeutics and/or surgery. This differs from the discovery cohort which consisted of newly-diagnosed patients. As previously has been noted, when designing EWAS, the power of the discovery cohort should be maximised by including all possible samples, rather than keeping samples back for replication.[202] Similarly, including patients at a different time points in the disease course demonstrates that methylation changes are not transient around the time of diagnosis. A limitation of this work is the limited clinical phenotypic information available for the samples used in the independent replication cohort. Given that these samples were obtained from patients with established disease with varying levels of inflammation at the time of sampling, it would have been interesting to correlate methylation data with routine clinical markers of inflammation such as C-reactive protein. An interesting subgroup analysis of DNA methylation changes in mucosal samples of paediatric treatment naïve IBD demonstrated that following treatment of active IBD, patients with disease in remission appeared to cluster with controls on principal coordinate analysis plots.[321]

A further limitation of this validation work is that all of the samples included in both discovery and validation cohorts all originate from the same geographic location. It remains to be seen if methylation changes demonstrated in Scottish samples can be reproduced in samples of similar genetic ancestry living in geographically remote areas. Newly-diagnosed IBD patients prospectively recruited as part of large consortia studies (IBD-BIOM/CHARACTER) may provide this opportunity. Should it not be possible to replicate these findings in other population, it may be possible that local environmental factors lead to specific epigenetic profiles, and this may contribute to differential geographical incidences of complex diseases such as IBD.[363]

Conclusion

Persuasive differences in DNA methylation associated with IBD have been technically validated and replicated in more than one independent cohorts increasing confidence in the results.

Chapter 5. Integrative analysis of Genetic and DNA methylation data

Abstract

Introduction

IBD has a strong genetic association and 200 loci have been associated with IBD. Most variants do not directly affect protein structure, but may exert an effect through alteration of expression (expression quantitative trait loci). It is also known the genetic variants control methylation and there has been increasing interest in identifying methylation quantitative trait loci (meQTLs).

Methods

All patients with 450k methylation data (Chapter 3) were genotyped using the Illumina Human CoreExome BeadChip microarray. The MatrixEQTL package was used to identify meQTLs and eQTLs using DNA methylation data and expression data from Chapters 3 and 6 respectively. A GWAS was performed using Plink. Mediation between genomics and epigenomics was investigated using the causal inference test.

Results

A GWAS did not demonstrate any significant associations with IBD following correction for multiple testing. Using the entire dataset, there were 424,880 significant associations between SNPs and methylation probes in *cis* following FDR correction for multiple testing. There were 220 IBD-associated meQTLs with an FDR adjusted p <0.05 (11,557 with uncorrected p<0.05) and 100 IBD-associated expression QTLs.

When considering only the 439 DMPs identified in Chapter 3, there were 326 meQTLs including 74 independent DMPs and 292 independent SNPs. Two of the five DMRs (*VMP1, ITGB2*) demonstrated significant association with genetic polymorphisms. Methylation in the *VMP1* region was significantly associated with two SNPs, both of which are in linkage disequilibrium with the IBD-GWAS SNP (rs1292053) (rs8078424, distance=13072bp, D' =1, r2=0.43 and rs10853015, distance=185198, D' = 0.93, r2=0.43).

Conclusions

In this chapter compelling associations have been made between germ line variation and DNA methylation and gene expression. *VMP1* a highly significant DMR is significantly associated with SNPs in linkage disequilibrium with a known IBD-susceptibility allele. Quantitative trait loci may be a mechanism by which genetics contributes to disease variance. Whilst IBD-associated meQTLs and eQTLs have been described, these appear to exist independently of each other.

5.1 Introduction

The method by which genetic variants associated with disease lead to functional changes that cause disease is still poorly understood. Relatively few IBD-associated genetic variants directly lead to changes in protein structure and even fewer have subsequently demonstrated biological effect in functional studies. Most variants described occur within intergenic or non-coding regions. Some of these variants occurring within gene promotor elements may exert their effects through alteration of gene expression.[364] Such loci are termed expression quantitative trait loci or eQTLs. Similarly, gene variants may exert their effect by alteration in DNA methylation levels,[365] by methylation quantitative trait loci (meQTLs). Methylation at some sites within the genome is almost entirely under genetic control, and is relatively common throughout the genome.[366] meQTLs have been described in a range of diseases including colorectal cancer,[367] bipolar disorder,[368] osteoarthritis,[369] rheumatoid arthritis[204] and type II diabetes mellitus.[366] In IBD to date, there has been no systematic study that investigates genetic control of DNA methylation and the assessment of disease risk for such loci.

Aims

The aims of this chapter were

1. to perform a genome-wide association study
2. to identify methylation quantitative trait loci (meQTLs) in *cis*
    a. In the entire dataset
    b. To identify disease-associated meQTLs
    c. To identify cis and trans meQTLs associated with DMPs identified in Chapter 3
3. to identify disease-associated expression quantitative trait loci (eQTLs) in *cis*
4. to determine whether methylation may mediate genetic risk

5.2 Methods

I had assistance from Dr Nick Kennedy for the processing and analysis of genotype data.

5.2.1 Genotyping

Genotyping was performed at the WTCRF by Louise Evenden and Tamara Gilchrist. In total 432 individual patient samples were genotyped. The Illumina Human CoreExome BeadChip microarray (HumanCoreExome 24v1-0_A) was used to assess genotype at 547,644 loci (HiScan H166).

5.2.2 Data processing

Genotypes were called using OptiCall[370] and plink[371] (excluding samples with less than 95% of genotyping complete, and SNPs with a genotyping rate of less than 95%). A sex check was performed in plink that uses the heterozygosity rates (Males have a homozygosity estimate of >0.8, whereas females have homozygosity rate of <0.2) on the X chromosome and flags samples where this does not match the reported sex in the PED file. Principal component analysis (PCA) was performed to assess the influence of ancestry on genotype. Samples were removed if there was evidence of relatedness (one pair removed when Identity By Descent > 0.1875).

5.2.3 Genome-wide association study

A genome-wide association study was performed using Plink (v1.07) using unimputed data (545732 snps) in 405 individuals.

5.2.4 meQTL analysis

The MatrixEQTL R package (version 2.1.0) was used to study associations between genotype and methylation.[372] For *cis* associations a distance threshold of 1Mb was used. SNPs with a minor allele frequency (MAF) of <5% were filtered from downstream analyses (<10% for DMP associated meQTLs, total SNPs included n=241,795). Associations with *trans* effects were investigated within the chromosome (to save on computational time) and across the genome for DMPs. *Cis* associations with all methylation probes (n= 448,363), and additionally both *cis* and *trans* effects were investigated amongst differentially methylated positions (n=439) identified in chapter 3. Age and sex were included as covariates, and for specific disease-associated meQTLs, IBD was additionally used as a covariate (using the modelLinear_cross option). A Benjamini-Hochberg False discovery rate (FDR) corrected p <0.05 was used as the threshold for statistical significance for disease associated meQTLs.[301]

5.2.5 Causal inference test

The causal inference test (CIT) described by Millstein et al[373] was used to attempt to identify DNA methylation as a potential mediator between genetic variants and IBD disease status. This technique has been employed before for linked genetic and DNA methylation in the context of allergy,[374] Type II diabetes[279] and Rheumatoid arthritis.[204] In order to determine mediation all of the following criteria must be met:

1) Genotype and disease are associated L→ T

2) Genotype is associated with methylation independent of disease L → G|T

3) Methylation is associated with disease independent of genotype G → T|L

4) Genotype is not independently associated with disease after adjusting for methylation L X T|G

A similar methodology was used as the previously published paper by Liu et al.[204] In the first step only methylation probes with an association with IBD were selected (i.e. DMPs from Chapter 3). In the second step, the DMPs were assessed for genetic association using matrixEQTL (as described above, FDR correction for multiple testing). Finally, only those DMPs and SNPs satisfying both of the previous criteria (i.e. meQTLs) were entered into the CIT algorithm using the R package cit.[373] The output p values of the CIT algorithm were corrected for multiple testing using Holm adjustment. Single nucleotide polymorphisms (SNPs) were assessed for linkage disequilibrium using SNAP using both r^2 and D'.[375]

5.3 Results

Using the GenomeStudio default output, total number of samples successfully genotyped was 410 (432 total, 94.90%). The locus success rate was 544,092 from a total of 547,644 (99.35%). The total number of genotypes returned for all samples was 98.2%. When genotypes were recalled using OptiCall, 19 samples failed genoyping (413 total, 95.6%) and 2 samples failed sex mismatch (Table 58 – Samples that failed quality assurance testing either by failing genotyping).

5.3.1 Genome-wide association study

A GWAS was performed in 228 cases and 177 controls (212 males, 193 females) using data directly obtained from genotyping microarrays (non-imputed data). The total genotyping rate was 0.999574. Following correction for multiple testing there were no SNPs significantly associated with IBD compared with control (Table 16).

CHR	SNP	UNADJ	GC	BONF
20	rs1009737	2.98E-06	2.34E-05	0.999
16	rs9926160	5.87E-06	4.10E-05	1
16	rs9930603	6.66E-06	4.55E-05	1
9	rs7875833	1.65E-05	9.62E-05	1
18	kgp225519	1.69E-05	9.82E-05	1
16	rs12709171	1.78E-05	0.000103	1
1	rs6691048	1.85E-05	0.000106	1
1	rs1894702	1.85E-05	0.000106	1
4	rs724454	2.32E-05	0.000128	1
15	rs2342120	3.06E-05	0.000161	1
18	rs4530229	3.23E-05	0.000168	1
4	rs17030327	4.12E-05	0.000206	1
16	rs2346254	4.61E-05	0.000226	1
14	exm-rs8022503	5.37E-05	0.000256	1
1	rs2205895	5.96E-05	0.000279	1
1	rs6028	6.15E-05	0.000286	1
6	rs946351	6.19E-05	0.000288	1
23	exm2268471	6.30E-05	0.000292	1
6	rs3806095	6.82E-05	0.000312	1
3	rs1461820	7.08E-05	0.000322	1
14	rs2178785	7.84E-05	0.00035	1
20	rs1381100	7.99E-05	0.000356	1
23	rs5971983	8.03E-05	0.000357	1
6	rs7764657	8.35E-05	0.000369	1
3	rs7646881	8.59E-05	0.000378	1
18	rs1561823	8.64E-05	0.00038	1
23	rs4421510	0.000102	0.000436	1
15	rs12899976	0.000105	0.000446	1
5	exm-rs31489	0.00012	0.000498	1
5	rs31489	0.00012	0.000498	1

Table 16 – Genome wide association study of IBD versus control (CHR=chromosome, SNP=single nucleotide polymorphism, unadj= unadjusted p values, GC=genomic-control corrected p value, BONF=Bonferroni adjusted p values)

5.3.2 Methylation quantitative trait loci (meQTLs)

meQTLs were identified in *cis* using matrixEQTL using the entire cohort (regardless of cases status, cases and controls combined together).[372] A *cis* distance of 1×10^6 bp (1 megabase) and a p value threshold of 1×10^{-4} was used. A minor allele frequency threshold of >10% was used. There were 424,880 meQTLs in cis that were statistically significant below a threshold of $p < 1 \times 10^{-4}$ (FDR $p<0.05$), of these there were 85,727 independent SNPs. The top meQTLs are presented in Table 17 and Figure 32. As a positive control, the SNP probes included on the Illumina 450K strongly associated with the appropriate SNP on the genotyping array (e.g. rs715359 and rs715359). The disease status was not included as a covariate, and therefore these meQTLs are not disease-specific (Figure 33). A list of 574 (550 independent SNPs) meQTLs derived from whole blood has previously been published by Van Eijk.[376] Of 550 independent SNPs, 140 (25.5%) overlapped with whole blood meQTL derived from the present dataset.

SNP	SNP chr	SNP pos	Methylation probe	meQTL_dist	pvalue	FDR
rs17015259	chr1	209982923	cg26035071	516	0.00E+00	5.80E-302
rs17266366	chr6	157199844	cg23603995	1196	0.00E+00	5.80E-302
rs11239157	chr10	45078170	cg02113055	5650	0.00E+00	5.80E-302
rs9535274	chr13	50194394	cg08779649	-160	0.00E+00	5.80E-302
kgp10984609	chr17	43657257	cg22968622	-6322	0.00E+00	5.80E-302
exm1331231	chr17	44248837	cg22968622	585258	0.00E+00	5.80E-302
exm2253037	chr17	44249096	cg22968622	585517	0.00E+00	5.80E-302
rs17585974	chr17	44249199	cg22968622	585620	0.00E+00	5.80E-302
kgp10190983	chr17	44288281	cg22968622	624702	0.00E+00	5.80E-302
rs2957297	chr17	44368212	cg22968622	704633	0.00E+00	5.80E-302
rs12974071	chr19	41641134	cg20242889	323297	4.6E-302	1.13E-295
rs1044516	chr1	209959614	cg26035071	-22793	3.5E-299	8.00E-293
rs877707	chr4	7792662	cg25817503	4320	5.2E-299	1.16E-292
rs4963867	chr12	25454578	cg25134647	-412	3.0E-292	6.37E-286
rs1044013	chr1	154243115	cg14859874	4850	1.6E-284	3.22E-278
rs652243	chr11	107470916	cg22355889	9331	1.6E-283	2.98E-277
rs1939900	chr11	107471983	cg22355889	10398	1.6E-283	2.98E-277
exm2271812	chr12	129288534	cg09035930	6477	8.8E-277	1.60E-270
rs7611945	chr3	125677514	cg05084668	22133	1.1E-271	1.97E-265
rs3762352	chr1	38156902	cg24088508	440	1.4E-267	2.35E-261

Table 17 – Top table of Methylation probes that are significantly associated with SNPs (meQTLs) in *cis* in entire cohort (cases and controls combined into one cohort). Only the top association for each SNP is shown. There are a number of methylation probes labelled with an ‘rs’ number and denote the SNP probes included by Illumina on the 450k microarray that have been removed from this table. A *cis* distance of 1 × 10^6 bp (1 megabase) and a p value threshold of 1 × 10^{-4} was used. A minor allele frequency threshold of >10% was used.

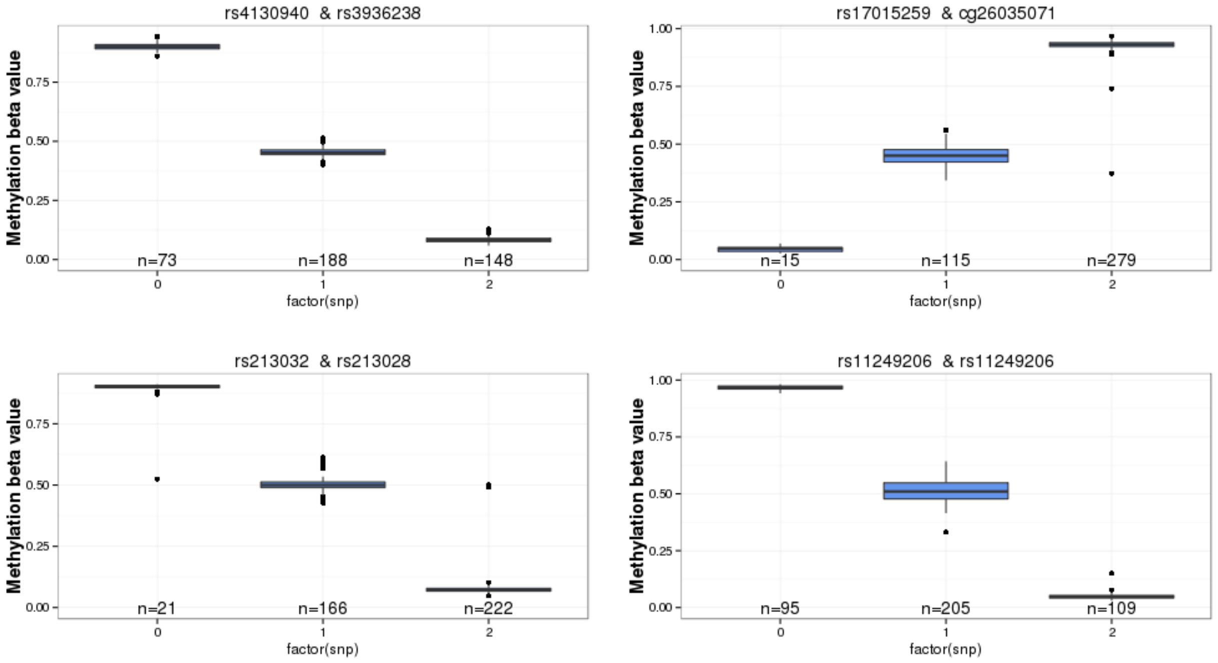


Figure 32 – A selection of the top meQTLs (disease non-specific) in *cis* in entire cohort (cases and controls combined into a single cohort). The title of each panel denotes SNP (first) followed by methylation probe. There are a number of methylation probes labelled with an 'rs' number and denote the SNP probes included by Illumina on the 450k microarray. A *cis* distance of 1 × 10^6 bp (1 megabase) and a p value threshold of 1 × 10^{-4} was used. A minor allele frequency threshold of >10% was used.

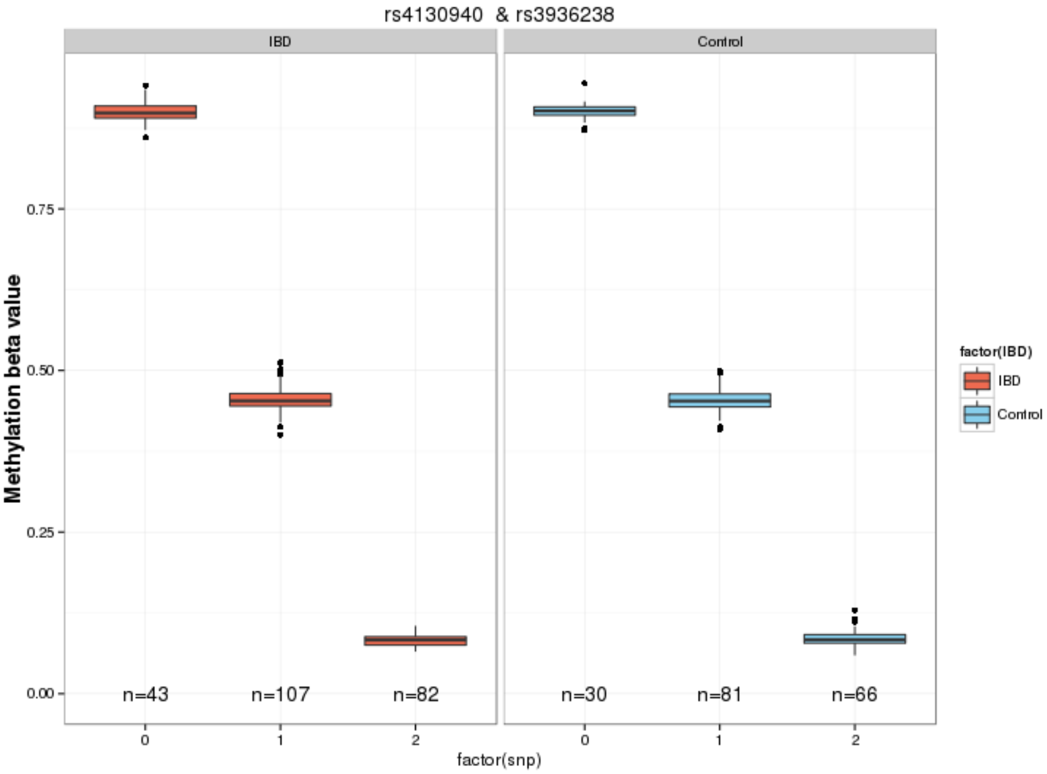


Figure 33 - Top meQTL in cis (rs4130940 and rs3936238) are similar in IBD cases and controls. The title of the panel denotes SNP (first) followed by methylation probe (in this case an Illumina SNP probe). A *cis* distance of 1 × 10^6 bp (1 megabase) and a p value threshold of 1 × 10^{-4} was used. A minor allele frequency threshold of >10% was used.

5.3.3 IBD-associated meQTLs

meQTLs were identified in *cis* using matrixEQTL. A *cis* distance of 1 × 10^6 bp (1 megabase) and a p value threshold of 1× 10^4 was used. Age, sex and IBD status were used as covariates. A minor allele frequency threshold of >5% was used. The MAF threshold was reduced for this analysis in an attempt to be more inclusive of potential IBD-associated SNPs. There were 220 IBD-associated meQTLs with an FDR adjusted p <0.05 (11,557 with uncorrected p<0.05)(Table 18). A Manhattan plot of IBD-associated meQTLs is presented in Figure 34. There was no overlapping SNPs in the 220 IBD-associated meQTL and the 163 IBD-associated SNPs described by Jostins et al.[88]

snps	SNP chr	SNP pos	Meth Probe	meQTL_dist	pvalue	FDR
exm1617809	22	46644177	cg10231785	-48263	3.74E-20	6.19E-14
exm-rs9261403	6	30069525	cg23934075	135828	6.91E-21	1.14E-13
rs16995069	22	46643774	cg10231785	-48666	9.76E-19	8.09E-13
kgp8465547	22	46614274	cg10231785	-78166	9.99E-18	5.52E-12
exm-rs3891157	6	29988439	cg23934075	54742	1.76E-16	5.78E-10
exm-rs7770505	6	30028913	cg23934075	95216	1.76E-16	5.78E-10
exm-rs9261285	6	30036083	cg23934075	102386	1.76E-16	5.78E-10
exm-rs6923832	6	30062058	cg23934075	128361	1.76E-16	5.78E-10
exm-rs9261257	6	30022425	cg23934075	88728	2.54E-16	6.96E-10
exm1617795	22	46643023	cg10231785	-49417	2.04E-15	8.44E-10
exm-rs9260934	6	29957982	cg23934075	24285	2.11E-15	4.95E-09
rs16995069	22	46643774	cg07012999	-48672	2.01E-14	6.65E-09
rs3796352	3	52913279	cg24616795	-119612	2.18E-15	9.34E-09
rs17331151	3	52844534	cg24616795	-188357	4.14E-14	6.29E-08
exm2256083	3	52551010	cg24616795	-481881	4.41E-14	6.29E-08
exm-rs6923856	6	29967529	cg23934075	33832	5.79E-14	9.63E-08
exm-rs7758512	6	29970589	cg23934075	36892	9.36E-14	9.63E-08
exm-rs6905157	6	29971548	cg23934075	37851	9.36E-14	9.63E-08
exm-rs6926792	6	29985849	cg23934075	52152	9.36E-14	9.63E-08
exm-rs6919617	6	29991699	cg23934075	58002	9.36E-14	9.63E-08

Table 18 - Top Table of IBD-associated meQTLs in cis

(age and sex as covariates). A *cis* distance of 1×10^6 bp (1 megabase) and a p value threshold of 1×10^4 was used. Age, sex and IBD status were used as covariates. A minor allele frequency threshold of >5% was used. There were 220 meQTLs in total associated with IBD status.

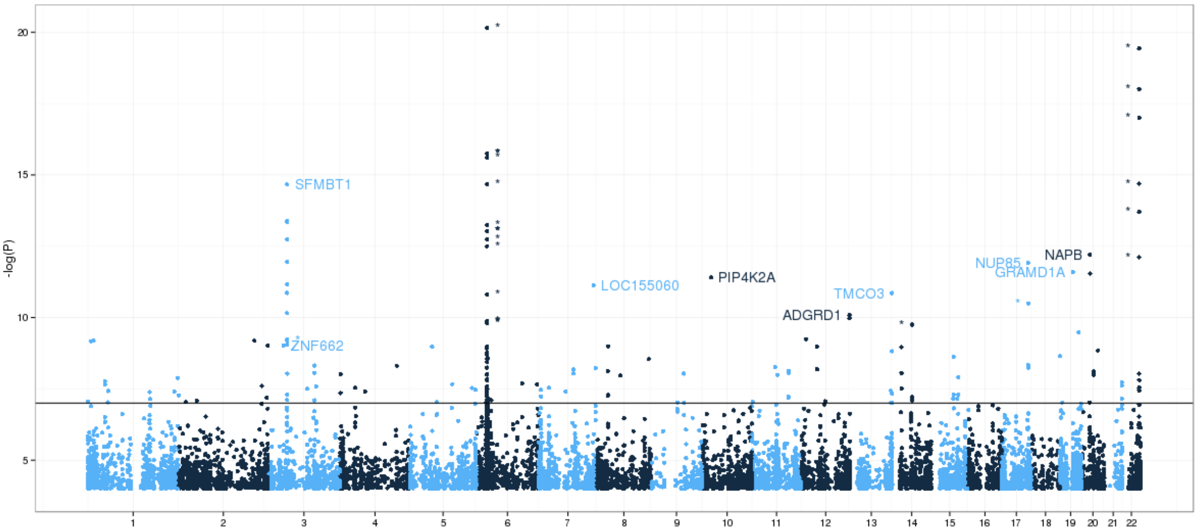


Figure 34 - Manhattan plot demonstrating IBD-associated meQTLs in *cis* (gene annotations refer to the methylation probe, *=methylation probe without names gene annotation). X-axis denotes SNP chromosomal location.

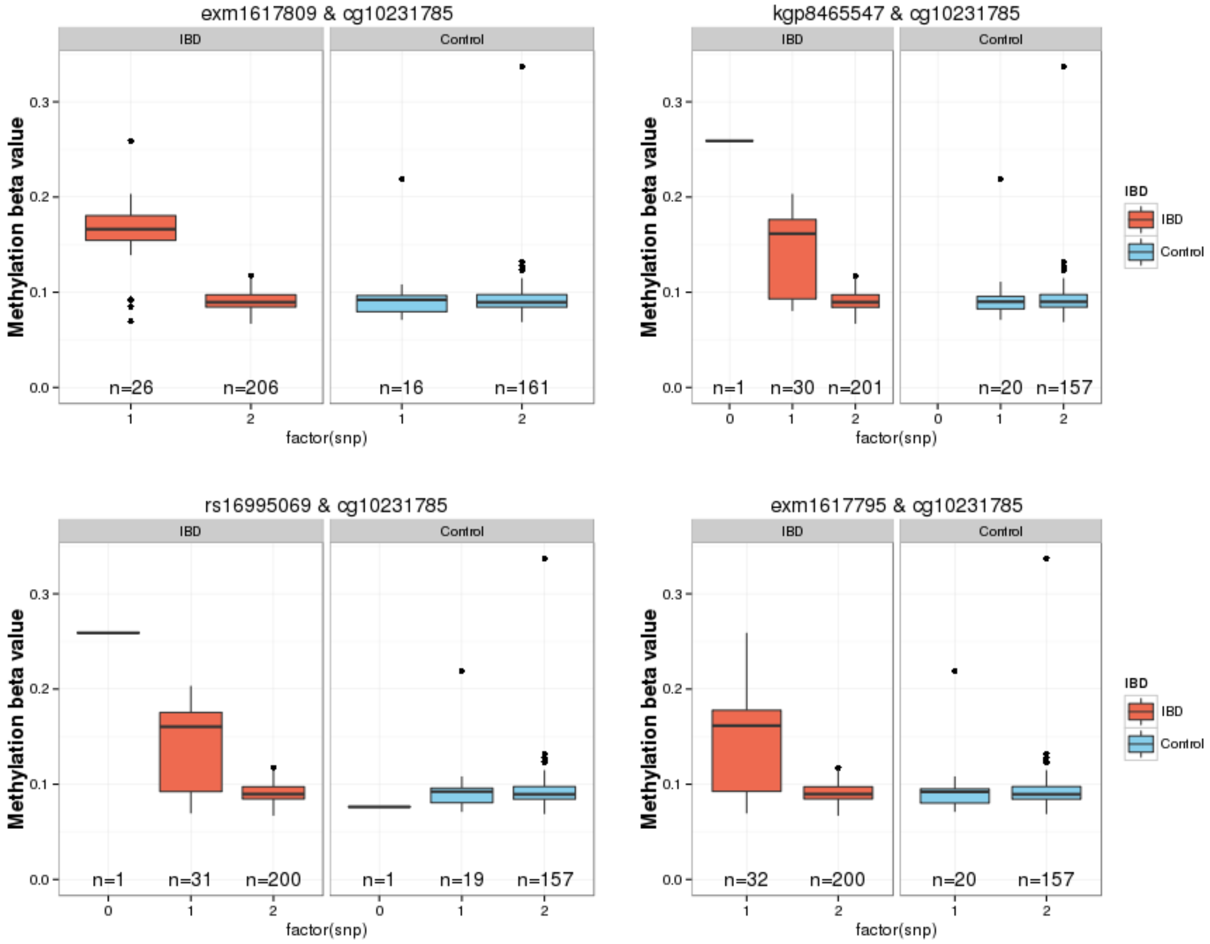


Figure 35 - cg10231785 cis meQTL in IBD cases and controls on chromosome 22. The title of each panel denotes SNP (first) followed by methylation probe.

5.3.5 Expression quantitative trait loci (eQTLs)

eQTLs were identified in *cis* using matrixEQTL. A cis distance of 1×10^{-6} bp (1 megabase) and a p value threshold of 1×10^{4} was used. A minor allele frequency threshold of >10% was used. There were 3,518 eQTLs in *cis* that were statistically significant below an a threshold of $p < 1 \times 10^{-4}$, with 1,975 eQTLS with an FDR p<0.05. The top eQTLs are presented in Table 19 and Figure 36. The disease status was not included as a covariate, and therefore these eQTLs are not disease-specific (Figure 36). Of the 431 independent whole blood eQTLs published by Van Eijk et al,[376] 20 SNPs replicated with the eQTLs identified in the present dataset (5%).

snps	Gene expression probe	SYMBOL	CHR	beta	p value	FDR
exm535799	ILMN_1697499	HLA-DRB5	6	-5.52	2.91E-60	1.84E-53
rs8676	ILMN_2399463	VAV3	1	-1.49	1.96E-33	6.19E-27
rs11150882	ILMN_1707137	C17orf97	17	-0.80	3.45E-32	7.27E-26
rs9271170	ILMN_1715169	HLA-DRB1	6	-3.01	7.50E-30	9.81E-24
rs7143764	ILMN_1798177	CHURC1	14	-1.50	7.76E-30	9.81E-24
exm-rs3135388	ILMN_1697499	HLA-DRB5	6	-3.88	1.08E-28	9.77E-23
rs3129889	ILMN_1697499	HLA-DRB5	6	-3.88	1.08E-28	9.77E-23
rs14139	ILMN_1753164	IPO8	12	-0.60	1.41E-27	1.11E-21
rs10760117	ILMN_3236498	LOC253039	9	-0.80	1.77E-27	1.25E-21
rs9270986	ILMN_1697499	HLA-DRB5	6	-3.90	4.77E-27	2.74E-21
exm-rs9271366	ILMN_1697499	HLA-DRB5	6	-3.90	4.77E-27	2.74E-21
rs11150882	ILMN_1713803	C17orf97	17	-0.48	5.30E-27	2.79E-21
rs521802	ILMN_2209115	MAK	6	-0.73	1.35E-26	6.58E-21
rs7316477	ILMN_1753164	IPO8	12	-0.60	2.52E-26	1.06E-20
rs6487927	ILMN_1753164	IPO8	12	-0.60	2.52E-26	1.06E-20
rs3135005	ILMN_1697499	HLA-DRB5	6	-3.52	3.34E-26	1.26E-20
rs11158568	ILMN_1798177	CHURC1	14	-1.49	3.38E-26	1.26E-20
rs10771752	ILMN_1753164	IPO8	12	-0.60	3.58E-26	1.26E-20
rs7197	ILMN_1697499	HLA-DRB5	6	-3.47	3.96E-26	1.32E-20
rs1968871	ILMN_2205322	TREML4	6	-0.52	1.14E-25	3.60E-20

Table 19 – gene expression probes associated with SNPs (eQTLs) in *cis* in entire cohort regardless of case status (n=67, cases and controls combined). A cis distance of 1 × 10^{-6} bp (1 megabase) and a p value threshold of 1 × 10^{4} was used. A minor allele frequency threshold of >10% was used.

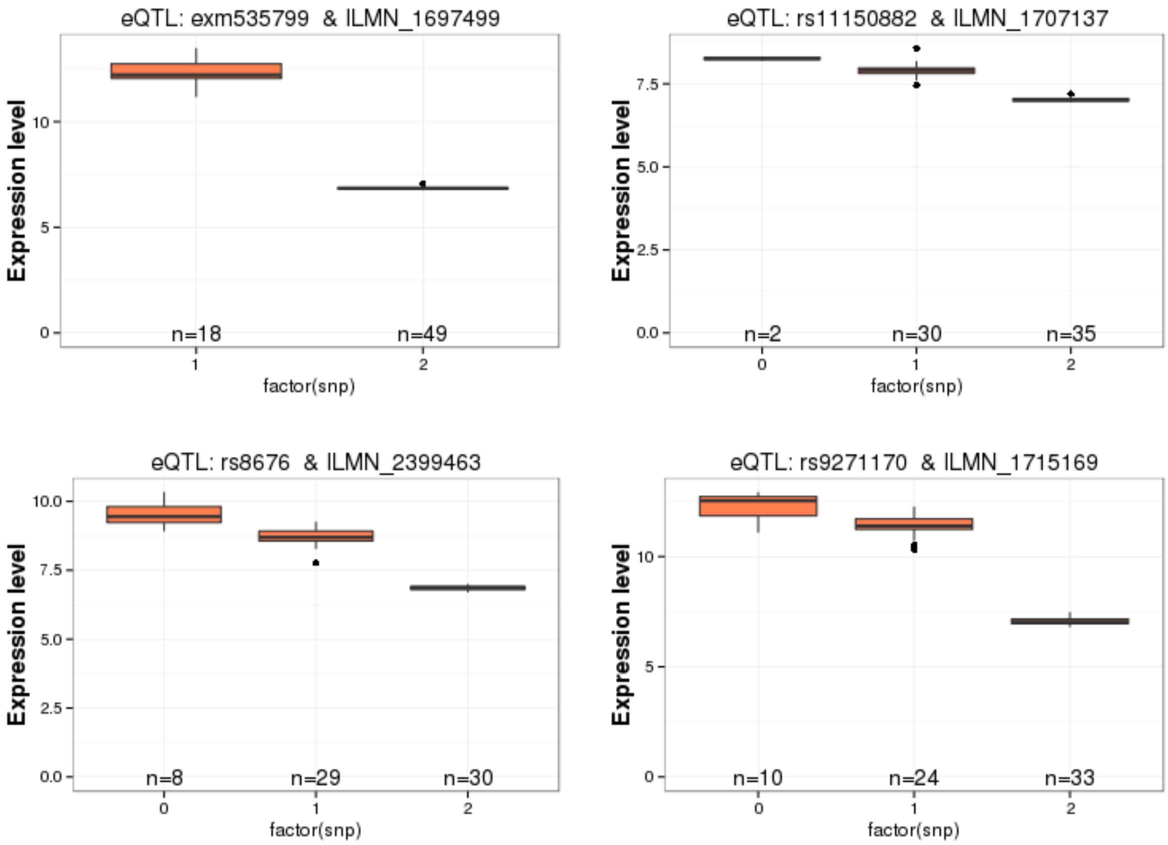


Figure 36 - Top eQTLs (disease non-specific) in *cis* in entire cohort

5.3.6 IBD-associated expression quantitative trait loci (eQTLs)

There were 109 IBD-associated eQTLs in *cis* in whole blood (Table 20). There were no overlapping SNPs in the 109 IBD-associated eQTLs and the 163 IBD-associated SNPs described by Jostins et al.[88]

snps	Gene expression probe	SYMBOL	CHR	beta	p value	FDR
rs118774	ILMN_1654663	LOC642843	17	1.11	2.69E-20	1.70E-13
rs1841955	ILMN_1668521	LIM2	19	1.01	1.34E-18	4.23E-12
rs16913885	ILMN_1747911	CDC2	10	1.43	1.76E-16	3.71E-10
rs1530947	ILMN_1774336	POLE2	14	1.42	1.06E-13	1.67E-07
rs11063582	ILMN_1670353	RAD51AP1	12	0.53	2.15E-12	2.58E-06
rs12809264	ILMN_1700337	TROAP	12	0.93	2.81E-12	2.58E-06
rs1569579	ILMN_1720114	GMNN	6	1.26	2.85E-12	2.58E-06
rs2214526	ILMN_1700337	TROAP	12	0.91	4.39E-12	3.47E-06
rs4888984	ILMN_1720526	CENPN	16	1.56	9.62E-12	6.28E-06
rs3788994	ILMN_1720114	GMNN	6	1.22	9.93E-12	6.28E-06
rs2071917	ILMN_1732198	UTS2	1	0.48	1.62E-11	9.30E-06
rs4733058	ILMN_1673673	PBK	8	0.69	2.30E-11	1.15E-05
rs4545047	ILMN_1673673	PBK	8	0.68	2.36E-11	1.15E-05
rs11778759	ILMN_1673673	PBK	8	0.68	2.84E-11	1.28E-05
rs6598008	ILMN_1711032	IFITM5	11	0.43	4.55E-11	1.84E-05
rs7927267	ILMN_1711032	IFITM5	11	0.43	4.65E-11	1.84E-05
rs11754578	ILMN_1720114	GMNN	6	0.60	5.58E-11	2.08E-05
rs3774473	ILMN_1767523	IL17RB	3	0.47	7.37E-11	2.59E-05
rs6542248	ILMN_2202948	BUB1	2	1.26	2.31E-10	7.68E-05
rs17637922	ILMN_1760247	CD70	19	0.75	3.42E-10	0.000105

Table 20 – Top list of IBD-associated eQTLs in whole blood in *cis* (age and sex as covariates). A cis distance of 1 × 10^{-6} bp (1 megabase) and a p value threshold of 1 × 10^{4} was used. A minor allele frequency threshold of >10% was used.

5.3.7 Overlap between genetic variants associated with both DNA methylation and gene expression

When considering the 220 IBD-associated meQTLs, there were 145 unique SNPs and for the 109 IBD-associated eQTLs there were 97 unique SNPs. There were no overlapping SNPs in the list of meQTLs and eQTLs in whole blood. For non-disease associated QTLs (85,727 meQTLs, 1627 eQTLs) there were 1384 overlapping SNPs.

5.3.8 Causal inference test

A similar methodology was used as the previously published paper by Liu et al.[204]

Step 1 – identify IBD-associated DMPs

In the first step the 439 Bonferroni-corrected IBD-associated DMPs identified in Chapter 3 were carried forward for meQTL analysis.

Step 2 – Genotype dependent DMPs

To identify genotype-dependent DMPs, the 439 significant DMPs were analysed using matrixEQTL (MAF>10%, cis distance 1 megabase, covariates age, sex, cell proportions, no disease covariate, model_linear) to identify 326 meQTLs (Table 60, 74 independent DMPs, 292 SNPs).

Table 21 - Top list of differentially methylated positions (DMPs) with genetic association (meQTL). Only the top SNP association is shown for each methylation probe. The variables used to search for meQTLs were as follows: MAF>10%, cis distance 1 megabase, covariates age, sex, cell proportions, no disease covariates. The table is ordered according to the significant of the DMP in the IBD vs. Control methylation comparison (DMP rank, P.Value and Holm Adjusted P.Value correspond to results of linear modelling carried out in Chapter 3) rather than the significance test of the association between SNP and methylation probe (meQTL rank). The results of the significance test of the association between SNP and methylation probe are presented in columns marked meQTL P value and FDR corrected as meQTL FDR P Value.

ProbeID	Chr	Meth symbol	Δβ	P Value	Holm adj P.Val	DMP rank	Top SNP	meQTL. P Value	meQTL. FDR P Value	meQTL. rank
cg16936953	17	**VMP1**	-0.09	1.3E-19	6.0E-14	3	rs8078424	2.9E-07	8.8E-05	265
cg12054453	17	**VMP1**	-0.07	4.0E-17	1.8E-11	9	rs8078424	4.4E-07	1.2E-04	284
cg18942579	17	**VMP1**	-0.05	1.2E-15	5.2E-10	14	rs10853015	3.1E-07	9.4E-05	267
cg02448796	1	KCNAB2	0.04	1.5E-14	6.9E-09	18	rs546526	2.2E-13	2.5E-10	71
cg12582317	17	NA	0.05	5.7E-14	2.5E-08	20	rs886926	7.4E-35	1.0E-30	6
cg16724148	1	AGL	0.03	1.2E-13	5.4E-08	22	rs2640911	3.4E-24	1.4E-20	20
cg01409343	17	**VMP1**	-0.04	3.4E-12	1.5E-06	45	rs10853015	9.1E-07	2.3E-04	322
cg16755922	17	FOXK2	0.04	1.5E-11	6.8E-06	61	rs11658011	1.3E-08	5.9E-06	176
cg27023597	17	**MIR21**	-0.03	1.6E-11	7.4E-06	62	rs10853015	9.0E-07	2.3E-04	320
cg02508743	8	LYN	0.03	4.3E-11	1.9E-05	80	rs2719236	2.3E-08	1.0E-05	184
cg24469729	7	HOXA3	0.03	5.3E-11	2.4E-05	82	rs2465276	7.1E-16	1.3E-12	45
cg14722693	8	CSGALNACT1	-0.03	6.8E-11	3.0E-05	86	rs10107533	1.3E-07	4.5E-05	232
cg24707889	21	ITGB2	0.03	7.3E-11	3.3E-05	88	rs2070946	2.6E-11	2.0E-08	108
cg08423142	15	MYO1E	-0.02	7.7E-11	3.4E-05	89	rs17236536	2.1E-07	6.7E-05	252
cg12807764	5	NA	0.04	1.1E-10	4.9E-05	95	rs17106769	3.4E-11	2.5E-08	110
cg02719954	8	NA	0.04	1.2E-10	5.5E-05	97	rs1438455	7.8E-11	5.6E-08	114
cg02782634	17	**VMP1**	-0.03	1.3E-10	5.8E-05	99	rs10853015	2.3E-07	7.3E-05	258

Step3 – Perform CIT

The causal inference test was performed using both genotype and methylation values as the potential causal mediator.

Methylation as potential causal mediator

When using methylation as the potential causal mediator with genetics as the instrumental variable there were no meQTLs that met the criteria for CIT following correction for multiple testing. For the SNP rs9789054 on chromosome 17, the corrected omnibus CIT was p=0.06 for methylation probes cg12582317 and cg12229367 (Figure 37).

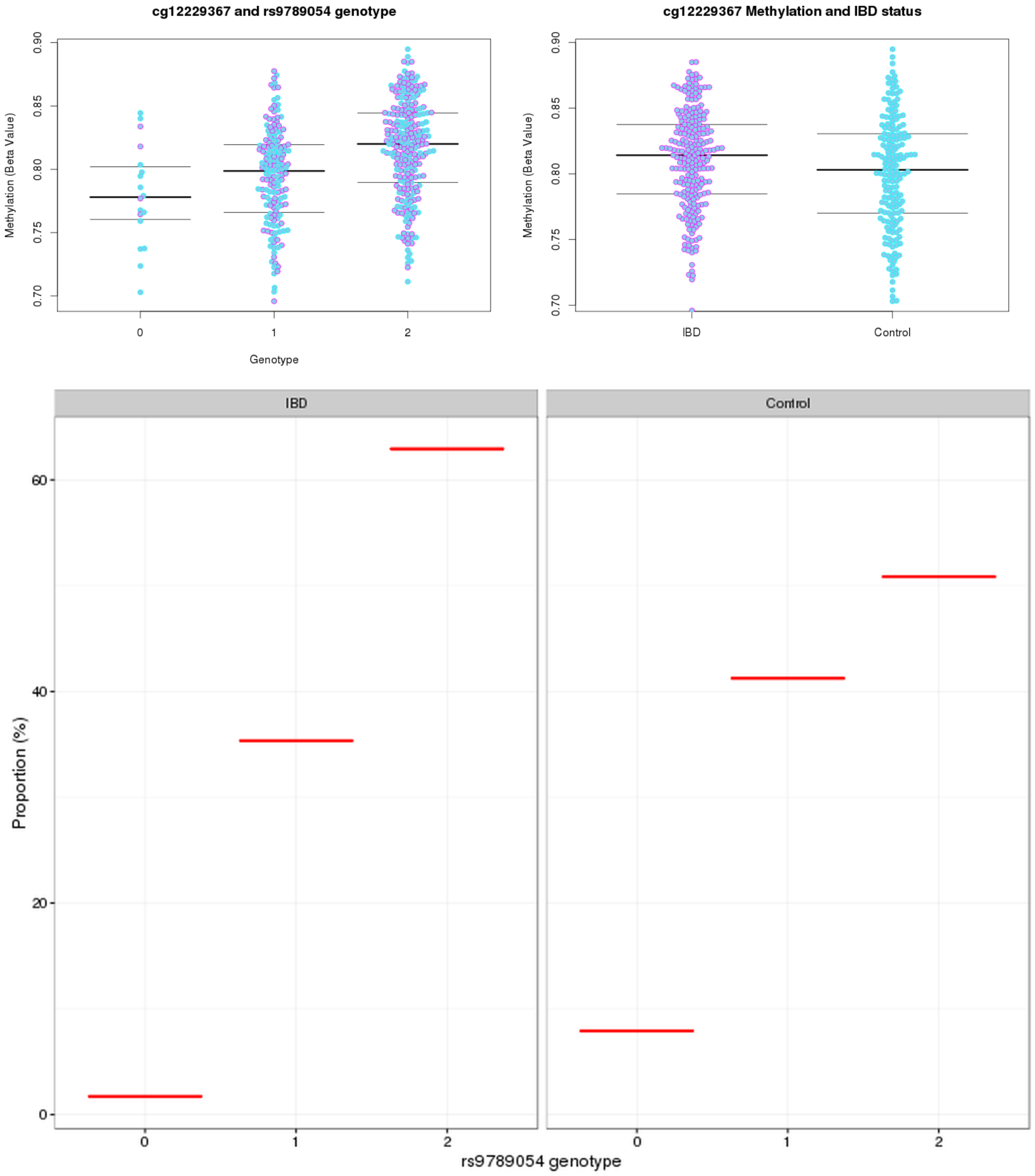


Figure 37 - rs10853015 association between genetics, methylation and disease. Top left panel demonstrates the association between genotype (X-axis) and cg12229367 methylation (beta-values, y-axis). Top right panel demonstrates the association between methylation at cg12229367 and IBD case status. Bottom panel illustrates the proportion of IBD cases (left) and controls (right) with each rs9789054 genotype.

Genetics as potential causal mediator

When using genetics as the potential causal mediator with methylation as the instrumental variable, there were two meQTLs that demonstrated a statistically significant CIT. The methylation probe cg03951877 (*PHACTR1*, phosphatase and actin regulator 1) and SNP exm-rs1332844 on chromosome 6 demonstrated a significant meQTL association (FDR p = 4.3×10^{-36}) and that phenotype was significantly associated with both methylation (Holm p=1.9×10^{-6}) and genotype (Holm p=0.003) when taking methylation into account. The CIT test also suggested that methylation was independent of phenotype given the genotype. The methylation probe cg26126879 and SNP rs678839 pair demonstrated a significant association (meQTL, FDR p = 6.2×10^{-53}) and significant CIT omnibus (Holm p=0.01). Phenotype was associated with both methylation (Holm p =1.3×10^{-5}) and genotype given methylation (Holm p=0.01).

VMP1/miR-21 locus

Five of the VMP1/miR-21 methylation probes were significantly associated with two SNPs (rs8078424, rs10853015) on chromosome 17 (Figure 38). Using the CIT with genotype as the potential mediator, the VMP1 methylation probe (cg16936953) and SNP rs8078424, there was overall causal inference (p=0.0007). A significant association between phenotype and methylation (Holm p= 2.2×10^{-13}), genotype and methylation (FDR p= 8.8×10^{-5}) and genotype and phenotype given methylation (Holm p=0.0007). Furthermore methylation is independent of IBD after adjustment for genotype (p=0). Using CIT with methylation as the potential mediator, the overall CIT was non-significant.

SNPs	Methylation probe	meQTL FDR p value	Meth Probe	CIT omnibus P.val	Pheno. assoc. w. Meth	Pheno. assoc. w. Geno. Given .Meth	Geno. assoc. w. Meth. given. Pheno	Meth. indep. of. Pheno .given . Geno
rs10853015	cg02782634	7.3E-05	VMP1	0.038	1.2E-10	0.02	1.8E-08	0
rs8078424	cg16936953	8.8E-05	VMP1	0.0007	2.2E-13	0.0006	4.7E-10	0
rs10853015	cg18942579	9.4E-05	VMP1	1	3.3E-15	0.001	4.8E-10	1
rs8078424	cg12054453	0.0001	VMP1	1	6.7E-16	0.001	6.2E-10	1
rs8078424	cg02782634	0.0002	VMP1	1	1.2E-10	0.02	1.8E-08	1
rs10853015	cg27023597	0.0002	MIR21	1	4.0E-11	0.02	9.6E-09	1
rs10853015	cg01409343	0.0002	VMP1	1	1.5E-11	0.01	1.5E-09	1

Table 22 – Causal inference test (CIT) for VMP1/miR21 locus (p values are Holm corrected for 7 tests in VMP1 locus). The meQTL FDR p value denotes the FDR corrected P value for the association test between genotype and methylation. The CIT omnibus p value represents the overall p value for the test (represents the highest p value of the other 4 tests). Column explanation: P value for phenotype association with methylation probe. P value for Phenotype associates with genotype following adjustment for methylation. P value for Genotype association with methylation following adjustment for phenotype. P value for independence test that methylation is independent of phenotype following adjustment for genotype.

The known IBD risk allele identified in Jostins et al on chromosome 17 (rs1292053) is in linkage disequilibrium with the two SNPs identified as meQTLs for the VMP1/miR-21 locus (rs8078424, distance=13072bp, D' =1, r2=0.43 and rs10853015, distance=185198, D' = 0.93, r2=0.43 [Figure 38 d, SNAP version 2.2, Broad institute[375]]).

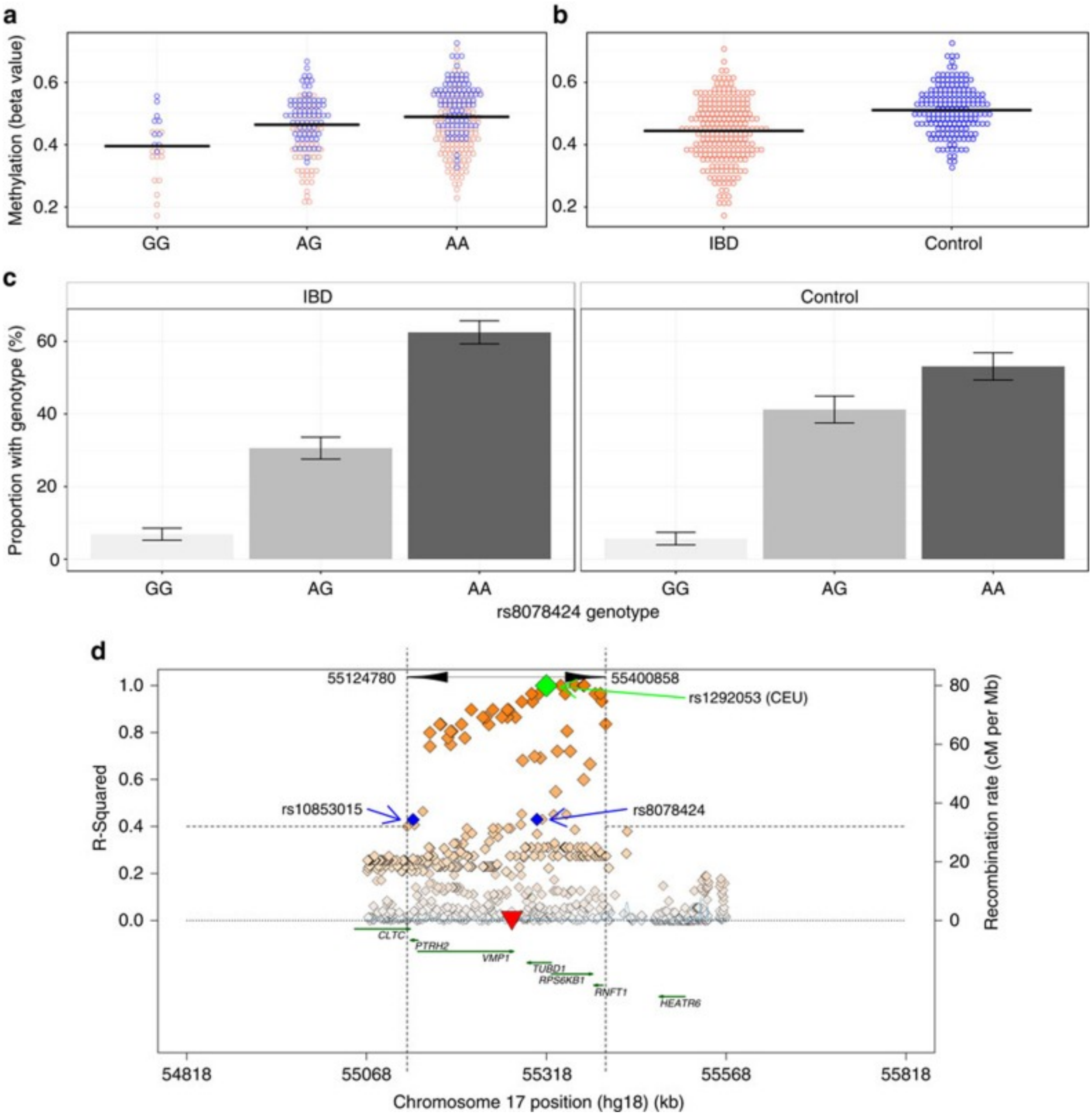


Figure 38 –VMP1 methylation associates with genotype and case status. (A- top left panel) VMP1 (cg16936953) is related to the rs8078424 genotype (FDR p=8.8 × 10-5) and case status (B-top right panel). There is some variation in the genotype between cases and controls (Cochran-Armitage test 1df χ^2=4.7 uncorrected p=0.03- C middle panel). D - Linkage disequilibrium plot in the VMP1 region between the Jostins et al IBD GWAS SNP rs1292053 (green diamond) and the SNPs associated with VMP1 methylation (rs10853015, rs8078424 blue diamonds) . The methylation probes including the VMP1 DMR are presented as red dots.(plot created using SNAP)[375]

Integrin β2

Three of the ITGB2 methylation probes were associated with three SNPs (rs2070946, rs9306118, rs2838738). Although there were associations between methylation and phenotype and methylation and genotype and genotype and disease, there was no evidence of independent effect or overall CIT (both p=1) (Table 23). The ITGB2 SNPs rs9306118 and rs2838739 are in LD (distance= 6247, r^2= 0.562, D'=0.775). The SNP rs2070946 was not included in this LD block using SNAP. The IBD GWAS SNP (rs7282490) is approx. 725kb upstream of the ITGB2 SNPs and is not in LD.

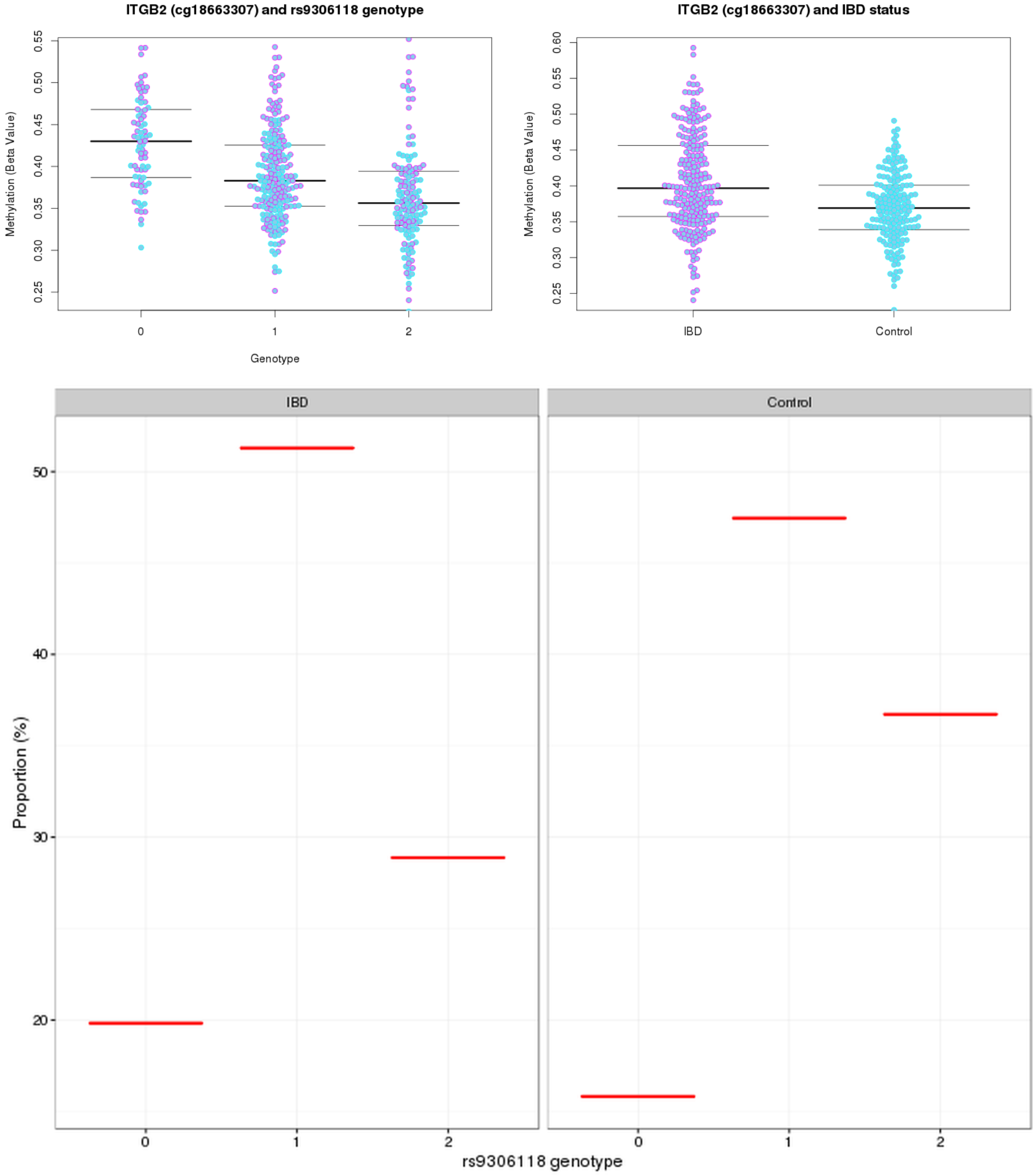


Figure 39 - ITGB2 (cg18663307) is related to the rs9306118 genotype and case status . There is some variation in the genotype between cases and controls (bottom panel). Top left panel demonstrates the association between genotype at 9306118 (X-axis) and ITGB2 cg18663307 methylation (beta-values, y-axis). Top right panel demonstrates the association between methylation at cg118663307 (beta values, y-axis) and IBD case status. Bottom panel illustrates the proportion of IBD cases (left) and controls (right) with each rs9306118 genotype (y axis=proportion of all cases).

SNP	Methylation probe	meQTL FDR p value	Meth Probe	CIT omnibus P.val	Pheno. assoc.w. Meth	Pheno. assoc.w. Geno. Given .Meth	Geno. assoc. w. Meth. given. Pheno	Meth. indep.of. Pheno .given. Geno
rs2070946	cg18663307	1.2E-15	ITGB2	1	4.9E-09	0.004	9.5E-20	1
rs9306118	cg18663307	4.6E-09	ITGB2	1	4.9E-09	0.2	4.4E-12	1
rs2070946	cg24707889	1.9E-08	ITGB2	1	3.2E-09	0.04	1.2E-12	1
rs9306118	cg24707889	8.8E-08	ITGB2	1	3.2E-09	0.2	4.3E-11	1
rs2838738	cg24707889	2.1E-06	ITGB2	1	3.2E-09	0.2	3.7E-10	1
rs2838738	cg18663307	6.9E-06	ITGB2	1	4.9E-09	0.2	1.6E-09	1
rs9306118	cg04321224	4.5E-05	ITGB2	1	1.4E-08	0.3	1.5E-08	1

Table 23 - Causal inference test for ITGB2 locus (p values are Holm corrected for 7 tests in ITGB2 locus). The meQTL FDR p value denotes the FDR corrected P value for the association test between genotype and methylation. The CIT omnibus p value represents the overall p value for the test (represents the highest p value of the other 4 tests). Column explanation: P value for phenotype association with methylation probe. P value for Phenotype associates with genotype following adjustment for methylation. P value for Genotype association with methylation following adjustment for phenotype. P value for independence test that methylation is independent of phenotype following adjustment for genotype.

Discussion

In this Chapter the disease-specific genetic basis for alteration in DNA methylation and gene expression was pursued. Systematic genome-wide data was available for all three analyses (genetics, DNA methylation and gene expression) and was integrated to identify genetic variants that impact upon DNA methylation and gene expression.

When analysing DNA methylation data it is difficult to determine whether DNA methylation is a cause or consequence of disease (i.e. chronic inflammation). A method of inferring cause (Causal Inference Test, CIT) has been developed by Millstein[373] and previously used on epigenetic data by Liu[204] and Yuan[279] in the context of complex immune diseases. Given that many of the hitherto describe genetic variants do not exist in amino acid altering positions, DNA methylation may be an important intermediary between genetics and disease. Two of five DMRs described in Chapter 3 have significant genetic associations. The two SNPs that associate with DNA methylation in the *VMP1*/miR-21 locus are in linkage disequilibrium with a known IBD-susceptibility allele.[86,88] This finding offers the tantalising possibility that the known IBD-susceptibility SNP mediates its effect on disease via DNA methylation. The second DMR with a significant genetic association is integrin beta-2 (*ITGB2*). Three SNPs associated with *ITGB2* methylation are relatively close, but not in linkage disequilibrium with another previously described IBD-susceptibility allele (rs7282490). The main limitation in this work (compared with the previously published Rheumatoid arthritis study) is the overall lack of power to determine genetic associations between disease and control. The lack of power with regard to genetic study makes the overall causal inference test unlikely to be significant as all tenants of the test (including genotype association with disease) need to be significant to prove an overall effect. Whilst this technique is useful in understanding genotype-DNA methylation relationships, observational studies such as the present work are unable to definitively prove cause and effect.

In the adult whole blood cohort, there were 220 IBD-associated meQTLs and 109 IBD-associated eQTLs, however no SNPs overlapped between the two and there was no overlap with known IBD-associated SNPs[88]. This demonstrates that genetic variation has a significant impact on site-specific methylation, however the same SNPs do not have the same impact upon gene expression. This reflects the wider scientific literature where meQTLs and eQTLs largely exist independent of each other and may contribute separately to variance in disease suseptibility.[368,377] The absence to detect shared SNPs between eQTLs and meQTLs may in part be related to reduced statistical power with regard to the gene expression experiments: only 14 controls with whole blood RNA were available for this experiment.

When the DMPs identified in Chapter 3 were investigated for a genetic association, there was only one significant meQTL in *cis* (*CLINK*) and no significant *trans* meQTLs following correction for multiple testing. This suggests that the DMPs identified in Chapter 3 are likely to exist independently of germline variation. This work has focussed on the local impact of genetic variation on DNA methylation (and gene expression) by concentrating on *cis* effects. Whilst some *trans* effects are likely to exist, there is some debate in the literature on their relative importance as *trans* or distal QTLs do not appear to be reproducible.[378] Furthermore, the large number of independent association tests warrants stringent correction for multiple testing.

Independent of disease status, we have also demonstrated an abundance of quantitative trait loci for both DNA methylation and gene expression in whole blood. meQTLs are common throughout the genome and a study in adipose tissue demonstrated that 28.5% of CpG sites are associated with SNPs.[366] Many of the disease-dependent and -independent methylation and expression QTLs were found in the major histocompatibility complex (MHC)/human leukocyte antigen (HLA) region of chromosome 6. The overrepresentation of MHC eQTLs in whole blood has previously been described in healthy individuals.[379] It is noted in the present dataset that several expression probes (and methylated probes) are associated with the same SNP. The MHC region is known to harbour many genetic variants and has complex and extended linkage disequilibrium structures.[379] Methylation and expression QTLs have previously investigated in whole blood of healthy volunteers using the Illumina 27K platform.[376] When attempting to validate the non-disease specific QTLs, a larger proportion of meQTLs validated (25%) compared to eQTLs (5%). The lack of eQTL validation may relate to sample preparation methods, in particular the method of whole blood collection, RNA extraction and globin clearance step.

In this chapter a limited GWAS study was performed in an attempt to identify or corroborate known IBD genetic associations. Following correction for multiple testing, no significant associations were identified. Clearly this small dataset is underpowered to detect such differences.

Conclusion

In this chapter compelling associations have been made between germ line variation and DNA methylation and gene expression. Quantitative trait loci may be a mechanism by which genetics contributes to disease variance. Whilst IBD-associated meQTLs and eQTLs have been described, these appear to exist independently of each other.

Chapter 6. Integrative genome-wide analysis of gene expression and DNA methylation

Abstract

Introduction

Relating epigenetic and expression data has been a major challenge in the field of epigenetics. Whilst there is a known inverse association between methylation in gene promotor regions and gene expression, the relationship is complex and unlikely to be binary. The three aims in this chapter were: (i) to define the genome wide expression profile in samples with DNA methylation data, (ii) to perform targeted expression profiling of DMRs identified in chapter 3, and (iii) to attempt to integrate DNA methylation and gene expression data.

Methods

Treatment naïve IBD patients and controls who had previously undergone DNA methylation profiling described in Chapter 3 had gene expression profiling using the Illumina HT12 expression microarray using RNA extracted from whole blood (PAXgene) and separated cells. Globin mRNA transcripts were removed from whole blood samples using GlobinClear. Differentially expressed genes were identified using linear models. Targeted gene expression profiling was performed using qPCR in an established disease cohort using RNA extracted from separated PBMCs and Granulocytes. Functional epigenetic modules were used to integrate DNA methylation and gene expression data.

Results

There were 47 differentially expressed genes in IBD cases compared with controls in whole blood following correction for multiple testing. There were no differentially expressed genes in any of the separated cell types following correction for multiple testing. Specific expression of the top DMP and DMRs was investigated. Expression of RPS6KA2 was increased at one probe and decreased at another in IBD cases. There was decreased expression of TXK (log fold change = -0.4, uncorrected $p=7.2 \times 10^{-5}$) and WRAP73 (log fold change = -0.1, uncorrected p=0.005) in IBD cases which is compatible with the hypermethylation of these genes described in Chapter 3. There was no difference in VMP1 or ITGB2 expression. Using qPCR for targeted expression of the most significant DMP/Rs, there was no significant difference in the expression of pre-miR21, RPS6KA2 and ITGB2 in PBMCs or granulocytes between cases and

controls. Functional epigenetic modules within gene networks of biological relevance were identified in whole blood for IBD as well as CD and UC separately.

Discussion

The relationship between DNA methylation and gene expression is complex and is likely to be cell specific. The location of DNA methylation change is critical when associating methylation with altered gene expression. Cell-specific changes in gene expression were seen in the top DMRs identified in chapter 3. For TXK where hypermethylation occurs within the TSS/promotor region, a reduced gene expression in whole blood and CD8+ cells was accompanied by a statistically significant negative correlation with DNA methylation in matched samples. Whilst similar convincing differences were not seen for the other DMRs/Ps, this may in part be related to type II statistical error and reflects similar experience in the wider field of epigenetics.

6.1 Introduction

The 5'-methylcytosine modification to DNA has been known for over half a century.[380] However real interest was ignited in the 1970s when pioneers such as Riggs,[381] Bird[382,383] and Cedars and Razin[384] demonstrated that DNA methylation was associated with altered gene expression. In humans, the globin gene was found to be unmethylated in tissues expressing globin, but heavily methylated in non-expressing tissues.[385]

There are several theories as to the underlying mechanism by which methylation effects gene transcription. DNA-binding transcription factors may be unable to bind to DNA recognition sites when CpGs are methylated.[383] Alternatively, protein complexes (e.g. MeCP2, methyl-CpG-binding protein 2) may preferentially bind to methylated CpGs and act as repressors to transcription by compacting chromatin structure and recruiting co-repressors.[383] Mutations in the *MeCP2* gene on Chromosome X are associated with Rett syndrome, a rare and severe neurological disorder.[386]

The relationship between epigenetic modifications and overall gene expression is complex, and unlikely to exist in a simple binary inverse relationship. Most CpG islands within transcription start sites (TSS) are unmethylated. It is now relatively well established that CpG island methylation occurring within promotor regions and transcription start sites (TSS) inhibit binding of transcription factors related to RNA polymerase and thereby inhibit gene transcription and expression.[387,388] Methylated promotor CpG islands are associated with long-term repression of genes, for example X chromosome inactivation in females.[282,389]A further controversy is the order of events; DNA methylation may occur after a gene has been silenced, thereby serving as a mechanism to reinforce or 'lock' the silent status of the gene.[282] The murine *Hprt* gene becomes methylated after inactivation of the X chromosome.[390]

The field is controversial; methylation in promotors does not always associate with gene silencing and approximately half of promotors do not contain CpG islands.[388] Many in the scientific community feel that methylation is a consequence of low transcription activity rather than a cause. It is not known whether methylation occurring out-with CpG islands can affect gene expression.[282] The relationship between expression and gene body methylation is even harder to unravel, and gene body methylation has been correlated with both increased and decreased gene expression.[391] Gene bodies generally have low CpG densities and are generally methylated and gene body methylation is a feature of transcribed genes. CpG islands do rarely occur within gene bodies, and are generally unmethylated, but can also be methylated in a tissue specific manner.[282]

With the expansion in the number of DNA methylation studies with complementary gene expression datasets, there has been a focus on developing methods to integrate the two data types.[392] The FEM package (Functional epigenetic modules) has been developed as a supervised method to identify epigenetically regulated gene networks/pathways based on an existing protein-protein interaction map.[392,393] This method has been used to integrate 27k DNA methylation data and PPI networks in the context of endometrial cancer[394] and stem cell differentiation.[395]

6.1.2 Aims

1. Define the whole genome gene expression profile in whole blood and separated cells in the same patients with DNA methylation data in previous chapters
    a. Determine the effect of globin mRNA on whole blood gene expression profiling
2. Targeted gene expression profile of top differentially methylated regions demonstrated in previous chapter
3. Integrate DNA methylation and gene expression data

6.2 Methods

6.2.1 Patient selection

Recruitment, sampling and cell separation of samples from IBD patients and controls is described in Chapter 2. For Illumina gene expression arrays a subset of the same cohort as used for DNA methylation array experiments (Chapter 3) was used. For selection of this subset of patients, priority was given to study subjects with cell separation data. An independent cohort of patients with established IBD with separated PBMC and granulocyte RNA was also used for gene expression analysis using qPCR described in this chapter. There is some overlap of patients used in the qPCR and microarray experiments.

6.2.2 Whole Blood PAXgene RNA clean-up and concentration

RNA extraction from PAXgene blood tubes is described in methods chapter 2.4.6. Given that the extracted RNA from PAXgene blood tubes was eluted in Qiagen buffer BR5 (Qiagen, Dusseldorf) it was not possible to concentrate RNA samples using a vacuum centrifuge (Speed vac). As a result, the Qiagen MinElute RNA cleanup kit was used to both clean up and concentrate the whole blood RNA samples extracted from PAXgene tubes. The MinElute kit uses a silica membrane and a column based technique to purify and concentrate RNA. The column enriches the sample for RNA with nucleotides >200nt, and therefore excluded microRNAs. Up to 10 µg of RNA was concentrated for use in downstream reactions (maximum capacity 45 µg). In all cases, the appropriate volume of sample was made up to 100 µL with RNAse free water. To the sample, 350µL of buffer RLT and 250µL of 100% ethanol was added, and mixed by pipetting. The sample was transferred to the RNeasy MinElute column and centrifuged at 10,000 × g for 15 seconds, and the flow through discarded. To the spin column, 500µL of buffer RPE was added and centrifuged at 10,000 × g for 15 seconds, and the flow through discarded. The same step was repeated using 500 µL of 80% ethanol. The membrane was dried by centrifuging the empty column for 5 minutes at maximum speed with the lid open. The RNA sample was eluted in 14 µL of buffer BR5 from the PAXgene blood RNA kit (centrifuged for 1 minute at 10,000 × g). The sample was denatured by heating on a heat block at 65 °C for 5 minutes.

Determine the effect of globin mRNA on whole blood gene expression profiling

6.2.3 Removal of globin mRNA from PAXgene whole blood RNA samples

Globin mRNA was removed from whole blood PAXgene RNA using the GlobinClear kit (Ambion, Life Technologies USA). Up to 10 µg of total RNA derived from PAXgene was used per GlobinClear preparation. PAXgene RNA samples were extracted as described in section 2.4.6 and concentrated into up to 14 µL according to section 7.2.2. Prior to starting the bead resuspension mix was prepared by combining per reaction; RNA binding beads (10 µL); RNA bead buffer (4 µL); and 100% isopranolol (6 µL). The Steptavidin binding beads were prepared by aliquoting 30 µL of newly resuspended beads into a 1.5mL RNAse free tube. The Steptavidin beads were placed on a magnetic capture stand for 3-5 minutes, and the supernatant was aspirated without disturbing the bead pellet and the same volume of

Steptavidin bead buffer was added. The bead mixture was mixed and incubated at 50 °C for 15 minutes in a hybridisation oven.

After preparation of bind beads, 1 µL of Capture Oligo Mix was added to the total RNA sample (in 14 µL) to make up a final volume of 15 µL. To the sample, 15 µL of 2X Hybridisation buffer was added and the sample was mixed. The sample was incubated at 50 °C (in hybridisation oven) for 15 minutes to hybridise globin mRNA and globin capture oligonucleotides. The globin mRNA was removed by adding the pre-prepared and pre-heated streptavidin beads to each sample (30 µL per sample), mixed and incubated at 50 °C for 30 minutes. Steptavidin beads bound to globin mRNA were magnetically captured by placing on magnetic capture stand, with the aspirated supernatant containing the globin depleted RNA sample, which was transferred to a new tube.

The sample underwent clean up in a second stage by binding the enriched RNA sample to a second set of magnetic beads (Bead resuspension mix, prepared earlier), and washed twice. To each sample, 100 µL of RNA binding buffer and 20 µL of bead resuspension mix (prepared earlier) were added and mixed. The sample was mixed and following magnetic capture, the supernatant was discarded. The beads bound to the globin depleted RNA were washed by adding 200 µL of RNA wash solution, before magnetic capture and the supernatant discarded. The beads were air dried for 5 minutes by leaving the tubes with caps off on the magnetic stand at room temperature. The globin mRNA depleted RNA sample was eluted from the beads by adding 30 µL of pre-warmed (58 °C) elution buffer and incubated at 58°C for 5 minutes. The beads were finally magnetically captured and the supernatant containing globin mRNA depleted RNA was aspirated and transferred to a new RNase free tube. The GlobinClear RNA sample was quantified using the nanodrop.

6.2.4 RNA amplification and biotin labelling

Total RNA extracted from separated cells (section 2.4.2), whole blood (section 2.4.1), or GlobinClear PAXgene tube (Sections 2.4.6, 7.2.2 & 7.2.3) was amplified using the Illumina TotalPrep RNA Amplification Kit (Ambion, Life Technology). Prior to amplification RNA quality was assessed using the Agilent BioAnalyzer (section 2.5.2). An input quantity of 250ng of RNA was used for all amplifications/expression experiments based on the nanodrop spectrophotometry result. At each stage the appropriate 'master mix' was made up on ice according to the number of samples, in a nuclease free 1.5mL Eppendorf tube by adding the following in order. The volumes of each component specified below were multiplied according to the number of samples to be amplified (+ 1.05*$n_{samples}$ overage for pipetting error). The master mix was mixed using a vortex and briefly centrifuged to collect at the bottom of the tube, and kept on ice until required.

i) Reverse Transcription to Synthesise First strand of cDNA

The volume required to obtain 250ng of total RNA was calculated (maximum of 11 µL) and inserted into an RNAse-free 0.2mL strip cap PCR tube (Ambion). Nuclease-free water was added to make the total volume up to 11 µL. A reverse transcription master mix was made up according to the appropriate number of samples. The master mix contained T7Oligo(dT) Primer(1 µL/sample), 10X First Strand Buffer (2 µL), dNTP Mix (4 µL), RNase Inhibitor (1 µL) and ArrayScript(1 µL). To each sample, 9 µL of reverse transcription master mix was added, mixed by pipetting up and down 3 times and flicking the side of the tube 3 times, and centrifuged briefly to collect at the bottom of the tube. The samples were placed in a preheated thermal cycler for 2 hours at 42 °C(lid temp 50 °C).

ii) Second Strand synthesis

Following incubation samples were removed from the thermocycler, centrifuged briefly and placed on ice. The Second strand master mix was made up by adding the following in order: Nuclease free water (63 µL/sample); 10X Second Strand Buffer (10 µL); dNTP Mix (4 µL); DNA polymerase (2 µL); and RNase H (1 µL). To each sample, 80 µL of mastermix was added, mixed by pipetting up and down 3 times and flicking the side of the tube 3 times, and centrifuged briefly to collect at the bottom of the tube. The samples were placed in a preheated thermal cycler for 2 hours at 16 °C (lid heat disabled between 16 °C and room temperature).

iii) cDNA Purification

Prior to completion of the incubation, 20 µL of nuclease-free water (per sample) was preheated to 55 °C using a heat block. Following incubation samples were removed from the thermocycler and transferred to a 0.6mL nuclease-free Eppendorf tube. To each sample, 250 µL of cDNA binding buffer was added, mixed by pipetting up and down 3 times and flicking the side of the tube 3 times, and centrifuged briefly to collect at the bottom of the tube. The sample was immediately transferred to a cDNA filter cartridge within a 2 mL tube, and centrifuged at 10,000 × g for 1 minute. The flow through was discarded and 500 µL of wash buffer was added to the cartridge, and centrifuged at 10,000 × g for 1 minute. The flow through was discarded and the empty cartridge and tube centrifuged at 10,000 × g for 1 minute to dry the column. The cDNA was eluted using 20 µL of preheated nuclease free water (55 °C) added to the cartridge filter, incubated at room temperature for 2 minutes, and centrifuged at 10,000 × g for 1 minute.

iv) in-Vitro Transcription to synthesize cRNA

The eluted cDNA sample from the previous step (~17.5-20 µL) was transferred to a labelled 0.2mL strip cap PCR tube. An IVT master mix was made up, containing: T7 10X reaction buffer (2.5 µL/sample); T7 Enzyme mix (2.5 µL); and Biotin NTP Mix (2.5 µL). To each sample, 7.5 µL of master mix was added, and transferred to a preheated thermocycler (3 7°C, lid 100 °C) for 14 hours for the in-vitro transcription reaction. Following the 14 hour incubation (usually overnight), the sample was held at 4°C. The reaction was stopped by adding 75 µL of nuclease free water to each sample to a total volume of 100 µL.

v) cRNA purification

Prior to completion of the incubation, 200 µL of nuclease free water (per sample) was preheated to 55°C using a heat block. Samples were transferred to a 0.6 mL nuclease free Eppendorf tube. To each sample, 350 µL of cDNA binding buffer was added, mixed before immediately proceeding to the next step. To each sample, 250 µL of 100% ethanol was added and mixed. The sample was immediately transferred to a cRNA filter cartridge within a 2 mL tube, and centrifuged at 10,000 × g for 1 minute (delay may result in precipitation of sample). The flow through was discarded and 650 µL of wash buffer was added to the cartridge, and centrifuged at 10,000 × g for 1 minute. The flow through was discarded and the empty cartridge and tube centrifuged at 10,000 × g for 1 minute to dry the column. The cRNA was eluted using 200 µL of preheated nuclease free water (55°C) added to the cartridge filter, incubated at 55°C for 10 minutes in a heat block, and centrifuged at 10,000 × g for 1 minute.

6.2.3 Quality and quantity assessment of cRNA

Quality assessment using the Agilent BioAnalyzer

cRNA samples were assessed using a nanoChip either using the total RNA or mRNA settings on the BioAnalyzer instrument as described in Chapter 2. The expected gel appearance of cRNA is a 'smear', with a distribution of cRNA size is expected between 250 to 5500 nucleotides, with most cRNA between 1000-1500 nt. Samples were also quantified using the nanodrop spectrophotometer (See Chapter 2).

6.2.4 Expression microarrays

Expression microarray profiling was performed at the Wellcome Trust Clinical Research Facility by Tamara Gilchrist and Louise Everden. The cRNA samples were prepared to a concentration of 150 ng/µL. Illumina HT12 human v4 expression microarrays), using a hybridisation time of 18 hours at 58 °C.

6.2.5 Data processing and analysis

Raw data

Data were analysed using the lumi[285] and limma[396] packages in R (R foundation for Statistical programming, Vienna). The raw data from the Illumina HT12 array were read into R using the lumiR.batch function. There were some inconsistencies in the gene annotation, therefore the probe profile txt. files were used instead of the gene profile files. The object type created was a lumiBatch object.

Quality control

The general quality of the raw data were assessed using several functions (summary, density plot, pairs plot and cumulative distribution function [CDF] plot, density plot of coefficient of variance). Outlying samples were excluded on the basis of cell type clustering principal component analysis (PCA) plots and presumed to have been mislabelled.

Background adjustment and normalisation

Background correction was performed using the bgAdjust function. As the probe profile file does not contain control probe information, control probe information was added to the control data slot. Variance stabilisation was performed to transform samples to possess a similar variance. This was required prior to downstream differential gene expression statistics. The quantile method was used to normalise the data (lumiN function). Each cell type was normalised separately.

Data Analysis

Data were analysed using the R package limma using a linear model with age and sex as covariates. Gene ontology was performed using goSeq on differentially expressed genes (FDR <0.05). Gene Ontology analysis was corrected using with a Benjamini-Hochberg (FDR p<0.05).

Targeted gene expression profile of top differentially methylated regions

6.2.7 Primer design

The NM (NCBI Reference Sequence) number for desired mRNA targets identified in previous chapters (e.g. DNA methylation) were obtained from NCBI Nucleotide search engine (http://www.ncbi.nlm.nih.gov/nuccore/). Primers were designed using the NCBI primer design tool (http://www.ncbi.nlm.nih.gov/tools/primer-blast/) using default options except to include exon-spanning junctions (limits amplification to mRNA only) and for primer pairs to include at least one intron on the corresponding genomic DNA (to distinguish between amplification of mRNA and genomic DNA as the latter is much longer due to presence of an intron). The in-silico PCR tool on the UCSC genome browser (https://genome.ucsc.edu/cgi-bin/hgPcr) was used to assess uniqueness of primer sequences. The four reference and target gene sequences are displayed in Table 24. The reference genes (GAPDH, TBP, SDHAP1, ACTB) were selected due to their previous stability of expression in leukocytes (specifically neutrophils and T-Cells).[284,397]

Table 24 - PCR Primer Sequences

Reference Genes	**Forward**	**Reverse**
GAPDH	TCATCTCTGCCCCCTCTGCT	CGACGCCTGCTTCACCACCT
TBP	TGCCCGAAACGCCGAATATA	TTTCTTGCTGCCAGTCTGGA
SDHAP1	AGGGCATCTGCTAAAGTTTCAGA	GATTCCTCCCTGTGCTGCAA
ACTB	GCCAGCTCACCATGGATGAT	AATCCTTCTGACCCATGCCC
Target Genes		
IGTB2	AGGAGGAGCTGAGAGGAACAG	CTGAGAGAGGACGCACCCG
Pri-MiR21	ATGGGCTGTCTGACATTTTGGTA	CATTGGATATGGATGGTCAGATGA
SBNO2	CAAGATGGCGCCCGAAAC	TGGAACAGCTTATCGTGGGT
RPS6KA2	GCCACCCTAAAAGTTCGGGA	GGGGTGATTCACTTCTGCCA
WRAP73	AGTCAGTTCCTGGCAGTTGG	ATGTTGTCGTTCCTTGTCGC
TXK	TTGTGAGTAGAGCACCGCAG	GGCAGCCTCCGTACTTCTTC

6.2.8 qPCR experimental work

Total RNA was extracted from separated cells as detailed in Chapter 2. RNA (500 ng input) was converted to first-strand cDNA using SuperScript VILO cDNA synthesis Kit (Invitrogen, Life Technologies). A master mix of 5XVILO reaction and 10XSuperScript Enzyme reagents was created and mixed with RNA samples. Incubations were as follows; 25 °C for 10 minutes, 42 °C for 60 minutes, and 85 °C for 5 minutes. Serial dilutions of the cDNA were performed and 1:100 cDNA was used for qPCRs. The Go Taq qPCR kit (Promega. WI, USA) was used for qPCRs in a volume of 25 µL (5 µL of 1:100 cDNA template, 2 µL primers [1 µL forward, µL reverse primer], 12.5 µL Go Taq master mix, 5.5 µL water). The MJ Research PTC-200 thermal cycler (Quebec, Canada) with Chromo4 (Bio-Rad, CA, USA) was used with the following cycles; hot-start activation 95 °C for 2 minutes, 40 cycles of 95 °C for 15 seconds, 60 °C for 60 seconds, dissociation/melt curve 60-95 °C. The proprietary dye used in the Go Taq kit has similar spectral properties to SYBR Green I (Molecular Probes Inc, Oregon, US): Excitation at 580nm and emission at 620nm.[398] Data were analysed using the Opticon monitor 3 software (Bio-Rad, CA, USA).

6.2.9 qPCR data analysis

Data were analysed in R (Version 3.2.0, R Statistical programming, Vienna Austria) using the normqpcR package.[399] The optimal reference genes were selected using the geNORM[400] and NormFinder[401] methods. Differential expression was calculated using ΔCT and $2^{-\Delta\Delta CT}$ methods.[402] A Wilcoxon Rank Sum test was used to compare ΔCT between cases and controls.

Integration of Gene Expression and DNA methylation data

6.2.10 Functional epigenetic modules (FEM)

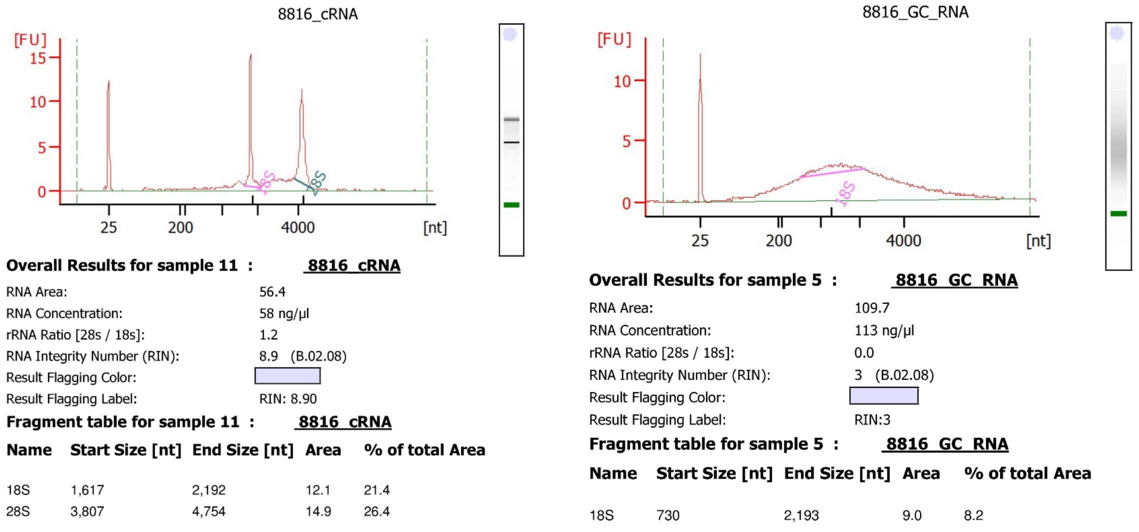
To integrate gene expression and DNA methylation DNA, the FEM R package was used (Functional Epigenetic modules).[392] This supervised algorithm aims to identify epigenetically regulated gene modules and pathways associated with disease status. The process consisted of two steps i) integration of methylation and expression data into a protein-protein interaction PPI network and ii) inference of modules based on a weighted network . The PPI network containing 8438 genes annotated to NBCI Entrez IDs was used as the adjacency matrix and has been previously used by package authors (http://sourceforge.net/projects/funepimod/).[393,395] Where several 450k probes mapped to a single gene, methylation values were summarised by taking an average of beta values of probes mapping to within 200bp of the transcription start site (TSS) or if none present, an average of betas in probes mapping to the 1st exon or within 1.5kb of the TSS were used.

Comparison of methylation (and expression) at each gene level was performed according to disease status (therefore matched samples were not necessarily compared). Gene expression and methylation data was scaled to avoid one or other data sources overly biasing the network. Genes were only included where there was a negative correlation between TSS methylation status and gene expression. Edge weights were assigned by taking the average statistic of each of the connecting gene nodes. Gene networks with higher edge weights than the rest of the network were defined as 'heavy subnetworks' or 'modules'. A spin-glass community detection algorithm was used to identify modules with the maximum edge weights. The number of seeds was set at 100 (default) indicating that the algorithm searches around the top 100 genes. The default parameter ($\gamma=0.5$) was used which typically identifies modules containing 10-100 nodes. The statistical significance was generated using Monte Carlo (MC) randomization, which performs a permutation test (set at 1000 permutations). Modules with a $p < 0.05$ were set a significant following FDR correction.

6.3 Results

6.3.1 Determine the effect of globin mRNA on whole blood gene expression profiling

The electrophrenogram profile demonstrated successful depletion of globin mRNA as indicated by the presence of a smooth distribution and smear (Figure 40 right panel) as opposed to the globin mRNA peak and band seen in uncleared samples (Figure 40 left panel). There was greater number of detected probes in those samples depleted of globin mRNA (Table 25). There was a significant reduction in the Haemoglobin Alpha 2 (p=0.001,Figure 41A), Delta ($p = 1.26 \times 10^{-8}$,Figure 41D) and Epsilon (p=0.001 Figure 41 C) expression in globin cleared samples versus non-globin cleared samples. There was no difference in the expression of haemoglobin beta (p=0.1, Figure 41 B) expression in globin cleared samples versus non-globin cleared samples.



Overall Results for sample 11 : 8816_cRNA

RNA Area:	56.4
RNA Concentration:	58 ng/µl
rRNA Ratio [28s / 18s]:	1.2
RNA Integrity Number (RIN):	8.9 (B.02.08)
Result Flagging Color:	□
Result Flagging Label:	RIN: 8.90

Fragment table for sample 11 : 8816_cRNA

Name	Start Size [nt]	End Size [nt]	Area	% of total Area
18S	1,617	2,192	12.1	21.4
28S	3,807	4,754	14.9	26.4

Overall Results for sample 5 : 8816_GC_RNA

RNA Area:	109.7
RNA Concentration:	113 ng/µl
rRNA Ratio [28s / 18s]:	0.0
RNA Integrity Number (RIN):	3 (B.02.08)
Result Flagging Color:	□
Result Flagging Label:	RIN:3

Fragment table for sample 5 : 8816_GC_RNA

Name	Start Size [nt]	End Size [nt]	Area	% of total Area
18S	730	2,193	9.0	8.2

Figure 40 – Electrophrenogram of sample 8816 before (8816_cRNA, left) and after globin mRNA depletion (8816_GC_RNA, right). The electrophenogram profiles demonstrate the characteristic peak and band caused by globin mRNA (left panel) and 'smear' appearance and smooth distribution following globin clearance (right panel)

Patient sample number	PAXgene sample without Globin Clear	Globin Clear PAXgene sample
8886	14,034	17,221
8816	12,603	17,120

Table 25 - Number of gene expression probes detected using the HT12 expression microarray in the same two samples before and after globin depletion using whole blood PAXgene mRNA samples

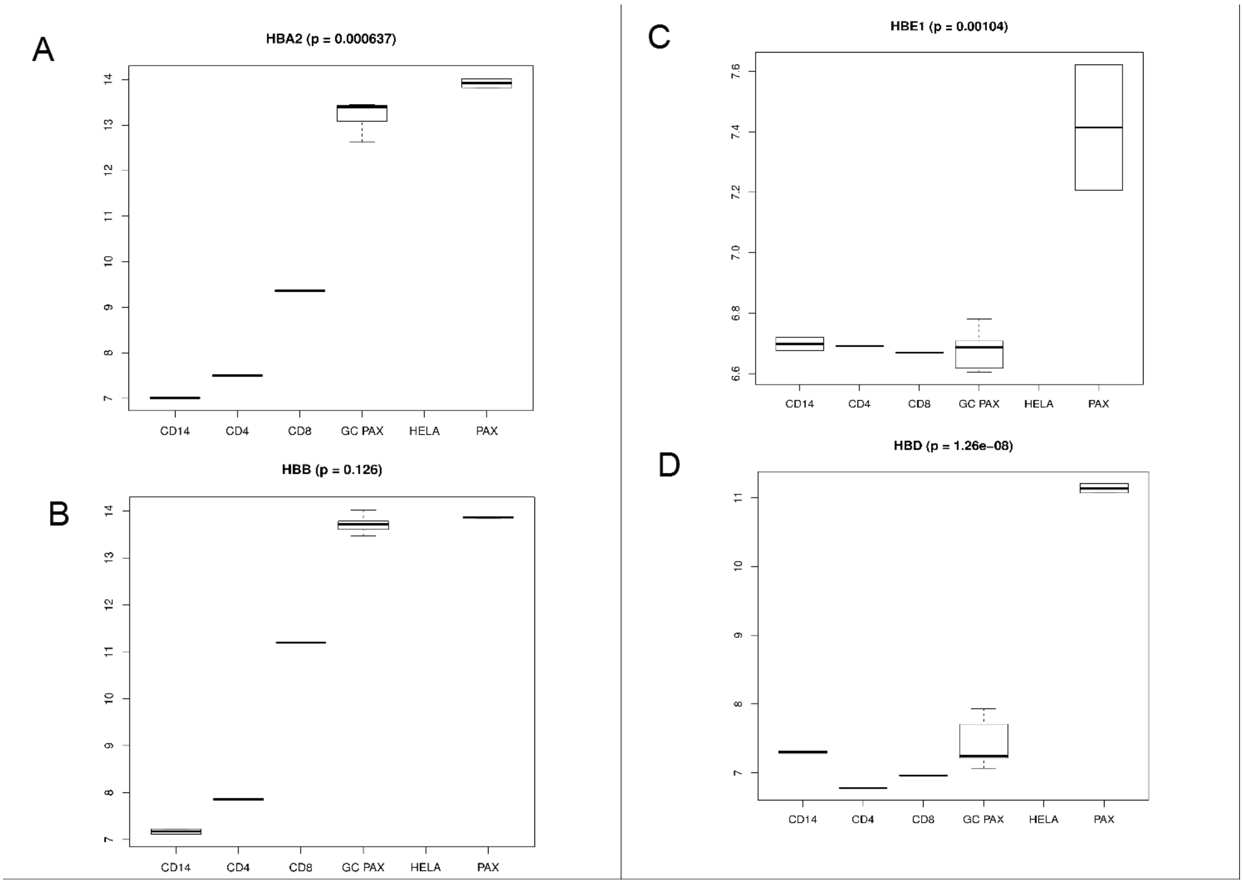


Figure 41- Relative expression of globin mRNA transcripts in separated cells, whole blood PAXgene with and without globin clearance. HBA2 = haemoglobin subunit α2, HBB = haemoglobin β, HBE = haemoglobin subunit ε, HBD = haemoglobin δ. X axis denotes sample type (CD14 = CD14+ monocytes, CD4= CD4+ T-cells, CD8=CD8+Tcells, HELA = Hela Cell line RNA provided as a control in kit, GC PAX = globin clear PAXgene Whole blood RNA, PAX = non-globin cleared PAXgene Whole blood RNA.

6.3.2 Whole genome gene expression profiling

Principal component analysis on unnormalised whole genome expression data demonstrated tight clustering according to cell type of the sample (Figure 42). Figure 43 details the RNA samples available for gene expression experiments and the overlap of separated cell samples available.

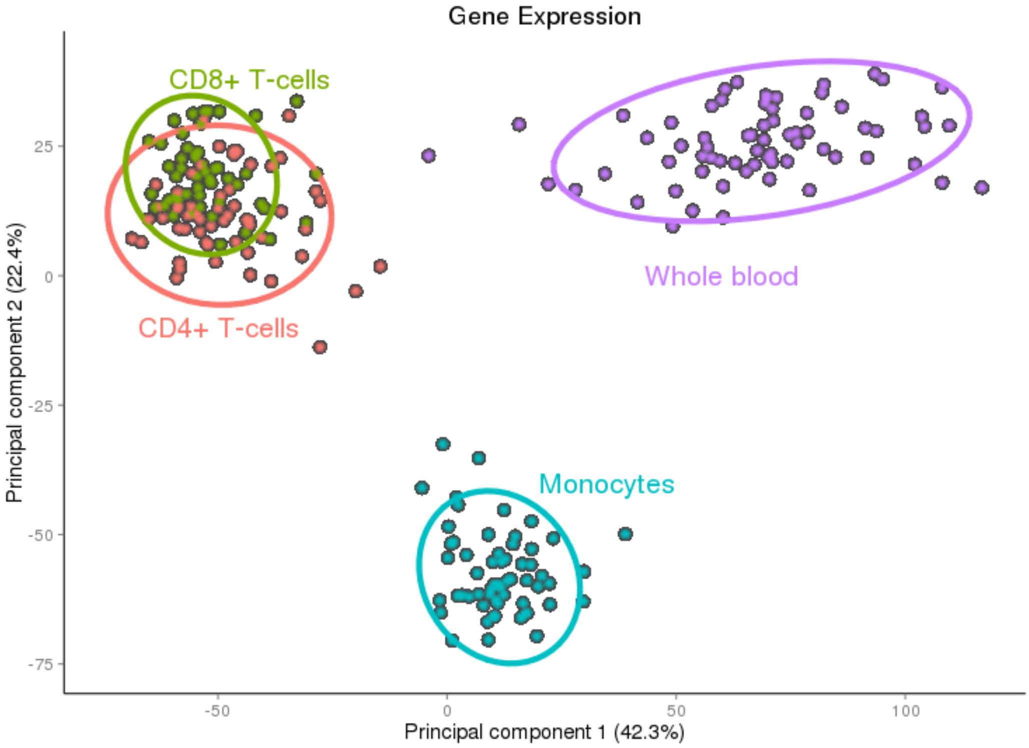


Figure 42 - Principal component analysis of whole genome expression data. Clustering was present according to cell type. Mislabelled samples were excluded based on PCA plots (data not shown).

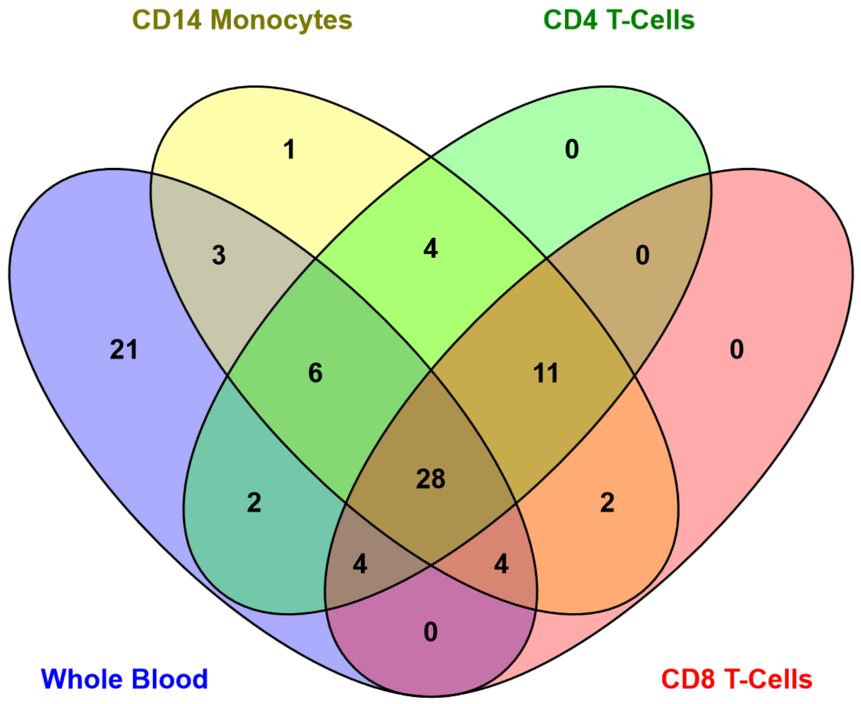


Figure 43 - Venn diagram[308] detailing samples derived from each individual participant used for whole genome expression profiling (e.g. 28 patients had all 4 cell samples available for analysis, 21 patients had whole blood cDNA alone)

6.3.2.1 Whole blood gene expression profiling

6.3.2.1 Whole blood Patient demographics

The patient demographics of the whole blood cohort for gene expression microarrays is displayed in Table 26. There was a non-significant trend towards increased numbers of current or ex-smokers in the IBD group.

	CD	**UC**	**IBD**	**Control**	**IBD versus Control**
n	22	22	44	24	
Females (%)	10 (45.5)	10 (45.5)	20 (45.5)	8 (33)	0.5
Age At Diagnosis (median, IQR)	26 (22-32)	36.5 (26-50)	28.2 (24.8-38.8)	31.5 (25.7 – 41.3)	0.7
Smoking (current or Ex)	13 (59)	12 (57)	25 (58.1)	7 (32)	0.08

Table 26 - Patient demographics of patients included in whole blood gene expression microarray (p values denote Fishers exact test for categorical variables, and Wilcoxon test for continuous variable)

6.3.2.2 Results of whole blood IBD versus control

There were 47 differentially expressed genes in whole blood in IBD cases versus control following Holm correction for multiple testing (Table 27). Gene ontology analysis revealed 116 significantly enriched terms (Table 62).The expression of the top DMP (RPS6KA2, Figure 44) and DMRs (VMP1, TXK, ITGB2 and WRAP73) identified in Chapter 3 in whole blood are displayed in Figure 45 and Figure 46.

IlluminaID	Gene symbol	logFC	AveExpr	P.Value	Holm adj.P.Val
ILMN_1802808	NA	0.70	13.03	1.11E-08	0.0005
ILMN_1663160	ZNF337	-0.41	8.91	2.24E-08	0.0011
ILMN_1723846	METTL21B	-0.26	7.83	3.72E-08	0.0017
ILMN_1764577	MFNG	-0.41	11.02	6.73E-08	0.0032
ILMN_1708323	ALAS2	1.81	10.38	7.14E-08	0.0033
ILMN_1757872	SGK223	-0.52	9.20	8.10E-08	0.0038
ILMN_1651719	MBTPS1	-0.28	9.74	8.61E-08	0.0040
ILMN_1703565	GLTSCR2	-0.53	12.98	8.95E-08	0.0042
ILMN_3242883	AGAP4	-0.58	9.47	9.61E-08	0.0045
ILMN_1789338	SORBS3	-0.30	7.88	1.17E-07	0.0055
ILMN_1654946	ZSCAN18	-0.45	9.15	1.44E-07	0.0068
ILMN_1719204	PRPF31	-0.32	9.59	1.70E-07	0.0080
ILMN_1795428	WDR59	-0.32	9.29	1.75E-07	0.0082
ILMN_3240222	SGK223	-0.56	9.87	1.87E-07	0.0087
ILMN_1755843	SLC26A8	0.99	8.74	1.99E-07	0.0093
ILMN_1683178	JAK2	0.38	8.48	2.89E-07	0.0135
ILMN_1679324	EIF1B	0.54	11.17	3.12E-07	0.0146
ILMN_1766657	STOM	0.73	11.15	3.15E-07	0.0147
ILMN_2050911	SLC22A4	0.74	9.20	3.22E-07	0.0151

Table 27 - Whole genome gene expression IBD versus control in whole blood. logFC = log fold change. AveExpr = Average expression.

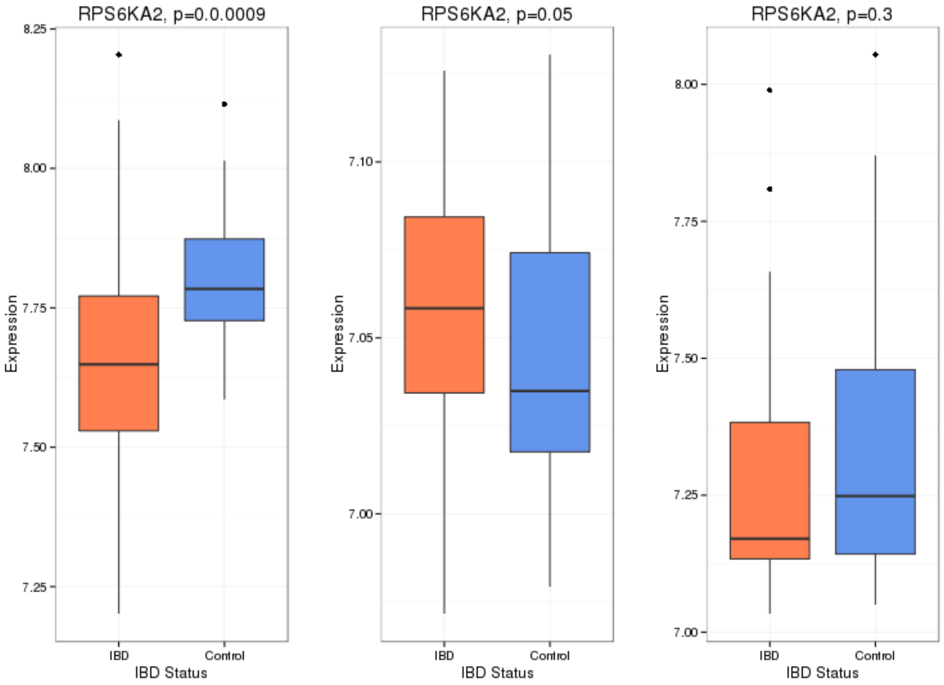


Figure 44 - Whole blood gene expression of the top DMP (RPS6KA2) in IBD cases versus controls (p values are uncorrected). There are three different gene expression probes annotated with RPS6KA2 on the Illumina HT12 microarray.

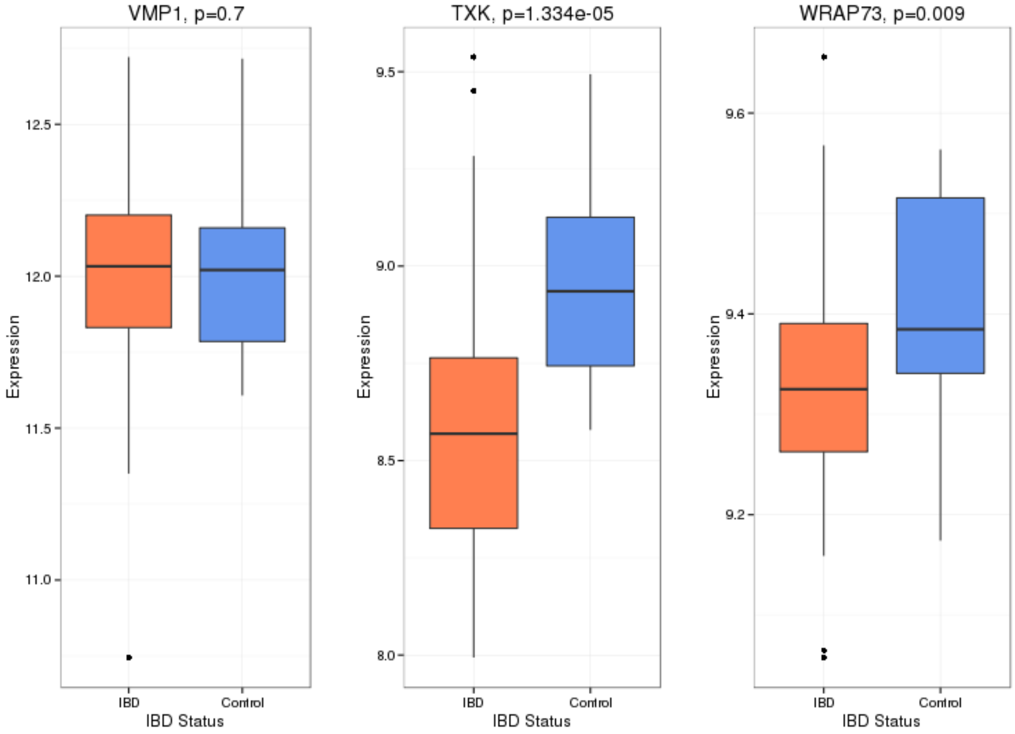


Figure 45 - Whole blood gene expression of the top DMRs (VMP1, TXK, WRAP73) in IBD cases versus controls (p values are uncorrected)

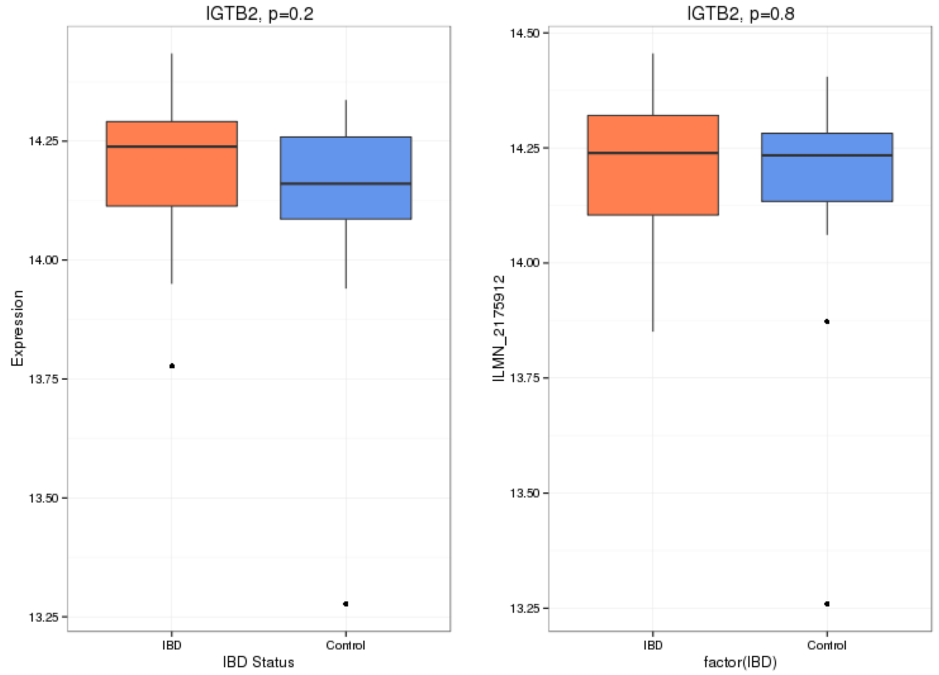


Figure 46 - Whole blood gene expression of the top DMRs (Integrin β2 subunit, ITGB2) in IBD cases versus controls (p values are uncorrected). There are two different gene expression probes annotated with ITGB2 on the Illumina HT12 microarray.

6.3.2.3 Results of separated cell gene expression in IBD versus control

The demographics of patients included in separated cell gene expression experiments are detailed in the appendix (Table 63, Table 64 and Table 65). Following correction for multiple testing there were no significant differentially expressed genes in IBD versus control in any of the separated cell types. The top lists of genes are detailed in the appendix (CD14 monocytes Table 66, CD4+ T cells Table 67 or CD8+ T-cells Table 68).

6.3.2.4 Separated cell gene expression for specific DMPs and DMRs

The expression of the top DMP (RPS6KA2, Figure 72) and DMRs (VMP1 [Figure 73], TXK, ITGB2 and WRAP73 [Figure 74]) were also investigated in separated cells.

6.3.2.5 TXK

The reduction in TXK gene expression seen in whole blood (fold change= −0.38, p=7.2 × 10^{-5}) was also seen in CD8+ (Fold change −0.41, p=0.03, Figure 47), but not other cell types. There was statistically significant negative correlation between TXK gene expression (ILMN_1741143) and all three DNA methylation probes included in the DMR in whole blood (cg02600394 Pearson's correlation = −0.48 p= 0.001, cg20981615 corr= −0.49 p=0.0007, cg17984638 corr=−0.44 p=0.003) and CD8+ cells (cg02600394 Pearson's correlation = −0.55 p= 0.0002, cg20981615 corr= −0.56 p=0.0001, cg17984638 corr=−0.7 p=2 × 10^{-7}) but not for other cell types.

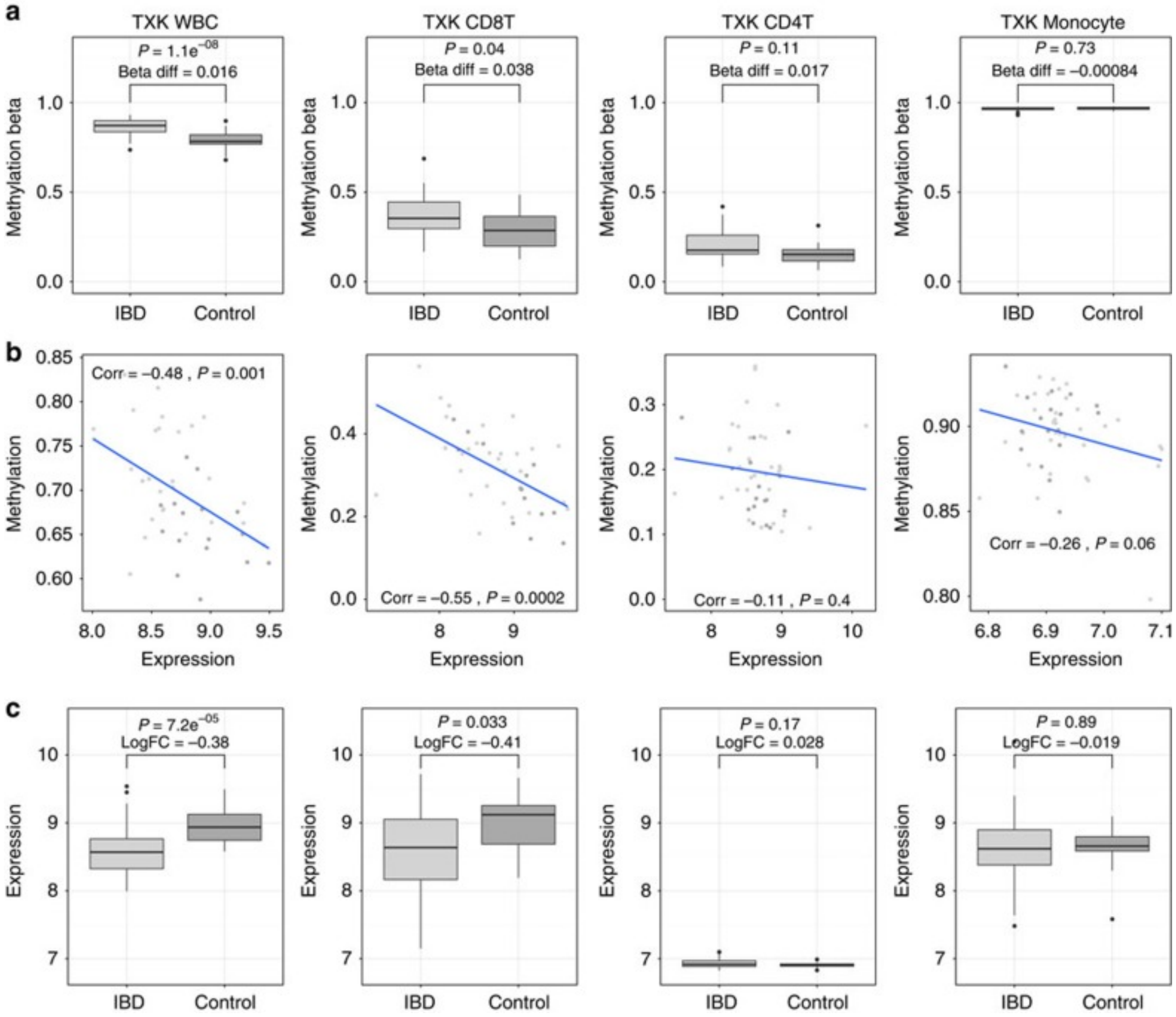


Figure 47 - TXK (Tyrosine Kinase) DNA methylation (a, top row) and gene expression (c, bottom row) in whole blood (Globin mRNA depleted), CD8+ T Cells, CD4+ T Cells and, CD14+ monocytes. Uncorrected P values derived from linear models including age and sex as covariates. Panel b demonstrate correlation between TXK gene expression and DNA methylation in matched samples. The y-axis scale has been standardised for all cell types to provide a meaningful comparison of expression/methylation levels.

6.3.2.6 RPS6KA2

For RPS6KA2, there was a slight decrease in whole blood gene expression in IBD cases (p=0.05), but the opposite gene expression pattern was seen in CD4+ cells (p= 0.3, Figure 72). In whole blood there was no significant correlation between DNA methylation (cg17501210) and gene expression (ILMN_1790801). In CD4+ cells there was a non-significant positive correlation between DNA methylation and gene expression (Pearson's corr=0.28, p=0.051).

6.3.2.7 VMP1

In whole blood, no change in VMP1 expression was seen. However in CD8+ cells, there was an increased gene expression in IBD (p=0.006, Figure 73). In matched samples, there was no significant correlation between DNA methylation and gene expression. This is notable given the whole blood 'signal' is thought to derive from CD8+ cells.

6.3.2.8 WRAP73

For WRAP73 (Figure 74) the increase in expression in whole blood (p=0.009) was also mirrored in CD8+ cells (p=0.05), but not the other cell types. In whole blood there was a non-significant negative correlation between DNA methylation and gene expression (Pearson's correlation = −0.39 p= 0.059). There was no significant correlation in CD8+ T-cells.

6.3.3 Targeted gene expression profile of top differentially methylated regions using qPCR

6.3.3.1 Patient selection for qPCR experiments

The demographics of patients and symptomatic controls with separated PBMCs and granulocytes are displayed in Table 69 and Table 70 respectively. There was a higher level of inflammation as denoted by C-reactive protein (CRP) in the symptomatic control group although this failed to reach statistical significance.

6.3.3.2 Reference gene selection

Using the geNorm and NormFinder algorithms, the best performing reference gene was TBP and SADPH for Granulocytes and SADPH for PBMCs (Table 28).

	geNorm				NormFinder			
	Gran		*PBMC*		*Gran*		*PBMC*	
	ranking	Step 1 stability M value	ranking	Step 1 stability M value	ranking	Step 1 stability Rho value	ranking	Step 1 stability Rho value
TBP	1	0.08	2	0.06	1	0.012		
SADPH	2	0.09	1	0.07	2	0.015	1	0.007
ACTB	3	0.11	5	0.08	3	0.02	3	0.01
UBC	4	0.13	3	0.07	5	0.03	2	0.007
GAPDH	5	0.14	4	0.08	4	0.02	4	0.01

Table 28 - Reference Gene selection using geNorm and NormFinder

6.3.3.2 Results of targeted qPCR of PBMC and granulocyte RNA of DMRs

The results of targeted qPCR of RNA from PBMCs and Granulocytes of DMRs identified in Chapter 3 are displayed in Table 29.

		2^dCt.IBD	IBD.sd	2^dCt.Control	Control.sd	2^-ddCt	2^ddCt.min	2^ddCt.max
RPS6KA2	**Gran**	0.62	1.57	0.45	1.52	1.36	0.46	4.04
	PBMC	0.46	0.53	0.28	0.34	1.62	NA	NA
Pri-miR21	**Gran**	4.72	3.59	2.56	2.33	1.85	0.15	22.18
	PBMC	0.46	0.54	0.76	2.2	0.60	NA	NA
ITGB2	**Gran**	0.03	4.72	0.01	3.15	4.72	0.18	124.56
	PBMC	0.36	0.01	NA	0.04	NA	NA	NA

Table 29 - $2^{-\Delta\Delta CT}$ values from targeted qPCR in PBMCs and Granulocytes

6.3.3.3 RPS6KA2

In Granulocytes there was no statistically significant difference in *RPS6KA2* expression between cases and controls (Wilcoxon rank sum test p=0.3). This was also the case in CD (p=1) and UC (p=0.07). There was no difference in the expression of *RPS6KA2* in PBMCs in IBD cases (p=0.2) compared with controls (Table 29). This was also the case in the individual diseases (CD, p=0.2, UC, p=0.3).

6.3.3.4 pri-miR21

In Granulocytes there was no statistically significant difference in pri-miR21 expression between cases and controls (Wilcoxon rank sum test p=0.8). This was also the case in CD (p=0.5) and UC (p=0.7). In PBMCs, there was no difference in pre-miR21 expression in IBD cases (p=0.1)(Table 29). There was lower pre-miR21 expression in CD (p<0.05) but not UC (p=0.7) compared with controls.

6.3.3.5 ITGB2

In Granulocytes there was no statistically significant difference in *ITGB2* expression between cases and controls (Wilcoxon rank sum test p=0.5). This was also the case in CD (p=0.9) and UC (p=0.3). In PBMCs, there was no difference in the expression of *IGTB2* in PBMC IBD cases versus control (p=0.1)(Table 29). There was also no difference in expression of *IGTB2* in CD (p=0.2) but not UC (p=0.5).

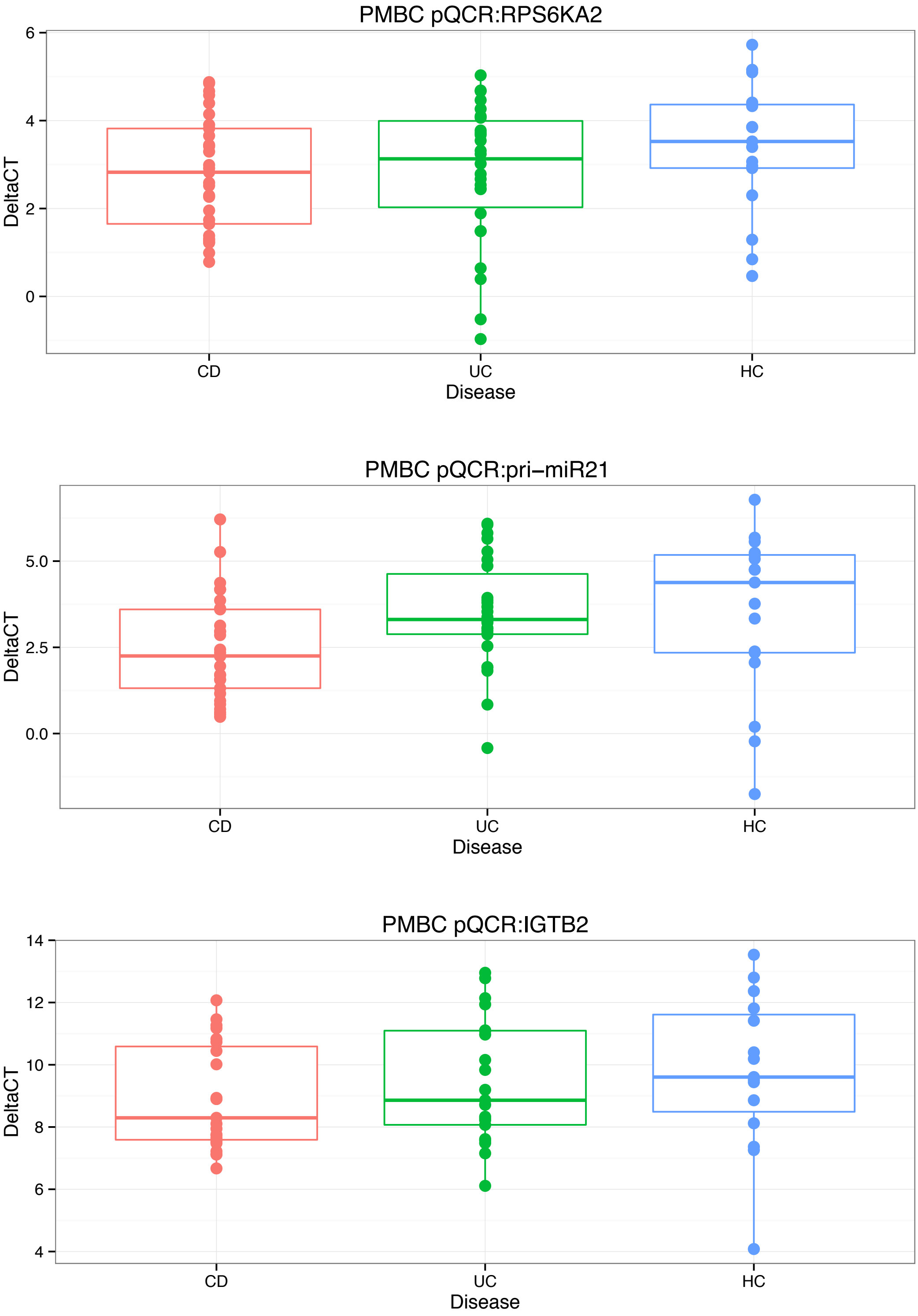


Figure 48 - PBMC qPCR results for RPS6KA2, pre-miR21 and IGTB2

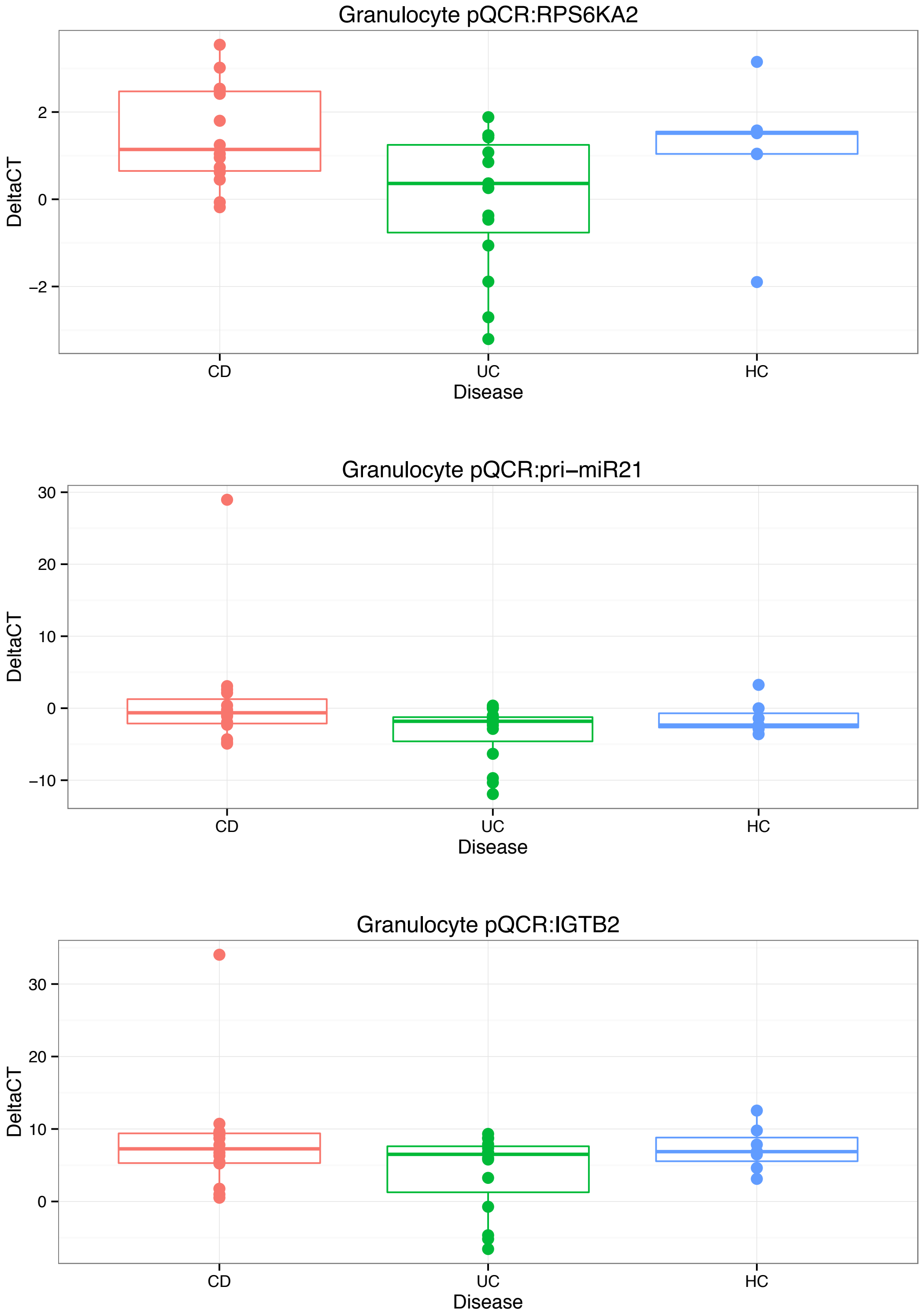


Figure 49 –Granulocyte qPCR results for RPS6KA2, pre-miR21 and IGTB2

6.3.3.6 Relationship between qPCR markers and clinical phenotype

There was no relationship between any of the measured markers and CRP or albumin. There was a weak but significant positive correlation between gene expression and the duration of the disease (i.e. time between date of diagnosis and date of blood sample, Figure, Figure 77, Figure 78).

6.3.4 Integration of Gene Expression and DNA methylation data

6.3.4.1 Functional Epigenetic Modules (FEM): Whole blood IBD versus control

To integrate gene expression and DNA methylation DNA, the Functional Epigenetic Modules (FEM) R package was used. Protein-Protein interaction (PPI) hotspots were identified displaying several linked genes demonstrating differential methylation. In whole blood when comparing IBD and controls, there were 11 significant functional epigenetic networks (Table 30). Figure 50 demonstrates the most significant network (DIABLO) containing 30 genes within the network.

	EntrezID (Seed)	Symbol (Seed)	Number of genes in network	Modularity	FDR p Value
1	56616	DIABLO	30	3.628882	0.009
2	80331	DNAJC5	38	2.701339	0.043
3	5265	SERPINA1	14	4.457891	0.006
4	1991	ELANE	59	4.583397	0
5	4353	MPO	66	3.873013	0
6	3082	HGF	12	5.652254	0.002
7	566	AZU1	13	4.021772	0.016
8	3674	ITGA2B	17	4.495567	0.004
9	1053	CEBPE	12	3.682455	0.032
10	966	CD59	46	2.791665	0.026
11	4318	MMP9	17	5.140515	0.002

Table 30- Functional epigenetic module for IBD versus control in whole blood. Modularity=average of edge weights. P Values are calculated using the Monte-Carlo procedure (a permutation test, n=1000)

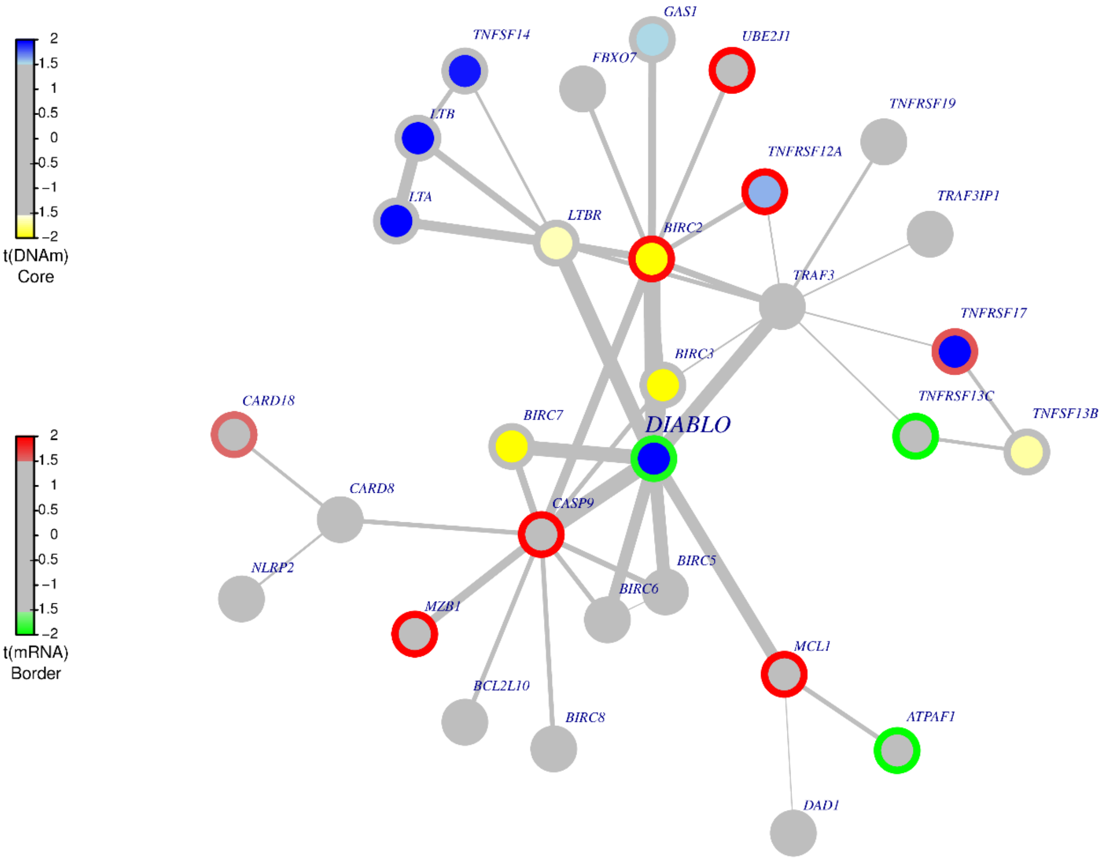


Figure 50 - DIABLO functional epigenetic module in whole blood (IBD versus control). Inner circles represent methylation (blue=hypermethylation/yellow=hypomethylation) and outer circles represent gene expression (red=increased expression, green=decreased expression)

6.3.4.2 FEM: CD and UC versus control in whole blood

FEM were also identified separately for CD (Table 71) and UC (Table 72) with a significant overlap of gene networks between the two individual diseases and with IBD.

6.3.4.3 FEM: IBD versus control in separated cells

FEM were also identified in separated cells (CD4+ [Table 74], CD8+ [Table 73], and CD14+ [Table 75]) and are detailed in the appendix.

6.4 Discussion

6.4.1 Targeted gene expression profile of top differentially methylated regions (DMRs) demonstrated in previous chapter

The location of DNA methylation change is critically important when attempting to delineate an association between methylation and altered gene expression. CpG island methylation occurring within promotor regions and transcription start sites (TSS) is known to be associated with reduced gene transcription and expression.[387,388] Therefore the most logical approach for this dataset would be to specifically investigate DMPs/DMRs occurring within the appropriate genomic context in promotor regions/TSS. The DMR TXK (*Tec* tyrosine kinase) is hypermethylated in cases (beta difference +2%) and located between the 5' UTR and 1st exon. There was a statistically significant reduction in TXK gene expression in whole blood and importantly DNA methylation and expression data in matched samples demonstrated a significant negative correlation. Given that DNA methylation and gene expression changes are likely to be subtle and cell specific, the difference seen in whole blood was only seen in CD8+ T-cells, again with a significant negative correlation between DNA methylation and gene expression. The DMR ITGB2 is also hypermethylated in cases (beta difference 4%) and located around the transcription start site/5' UTR and therefore would be another plausible candidate to alter gene expression, but there was no difference in expression of ITGB2.

The majority of DMPs/DMRs identified in Chapter 3 occur within the gene body (see volcano plot, Chapter 3), where the relationship between methylation and expression is unclear. Interestingly for VMP1 no difference in gene expression was seen in IBD cases and controls in whole blood or CD4+ cells despite a significant positive correlation between VMP1 methylation and gene expression. In CD8+ cells however, there was a reduction in VMP1 expression, but no significant correlation between DNA methylation and expression in matched samples. For the top DMP in whole blood, RPS6KA2 hypomethylation of the gene body was seen in IBD cases. For the three annotated RPS6KA2 probes included on the gene expression microarray, two demonstrated a counter-intuitive reduction in gene expression and one demonstrated increased gene expression.

To conclude from this work, at certain sites within promotor regions/TSS and within specific cell types, hypermethylation was associated with an appropriate reduction in gene expression. Where DNA methylation change was confined to the gene body the effects on gene expression were less consistent. There are several possible reasons for this potential disconnect between DNA methylation and gene expression. Firstly the absolute differences in beta values in IBD cases versus control are small. Whilst the beta differences seen in the present study are consistent with findings in other complex diseases, the absolute differences may not be enough to affect gene expression. Secondly, the overall number of subjects included in the gene expression microarray experiments was small, and there is a risk of type II error. A further limitation was the potential imbalance in baseline patient demographics in this experiment; there were no statistically significant differences between groups, however this may be related to the small numbers in each group.

In the prevailing scientific literature it has been difficult to convincingly link DNA methylation and gene expression in complex immune diseases. In the complex immune diseases EWAS literature some studies have successfully correlated DNA methylation difference with gene expression whilst others have not. In IBD, Adams et al correlated hypomethylation of the *VMP1*/miR-21 locus in peripheral blood of IBD patients with increased pri-miR-21 expression in blood and in mucosal biopsy samples, but not in matched samples.[284,403] McDermott et al demonstrated *TRAF6* hypermethylation in PBMCs in IBD cases correlated with decreased expression in a subset of the same patients.[307] In Harris et al, DNA methylation in was assessed in gut mucosal samples of children with UC and a small subset had allied gene expression data (5 UC, 5 HC). Several genes demonstrated epigenetically associated gene-expression including ITGB2, S100A9 (heterodimer with S100A8 to form calprotectin), IFITM1, SLPI and STAT3.[321] In mucosal samples, the expression of DOK2 and FUT7 correlated with gene expression.[171]

6.4.2 Integration of Gene Expression and DNA methylation data

Attempting to demonstrate an inverse correlation between DNA methylation and gene expression may be over simplistic and methods such as functional epigenetic modules[392] that consider location of DNA methylation probes within transcription start sites and/or known regulatory regions may be more appropriate. By using this method we have highlighted several gene networks of biological relevance that were significantly associated with IBD.

Functional epigenetic modules do not however take into account differential cellular proportions and are based on pre-constructed gene networks (i.e. not assumption-free). It is also worth noting that analyses are not paired, i.e. all DNA methylation and expression data are considered rather than only including paired data from the same individual.

6.4.3 Globin clearance increases the number of probes detected on downstream microarrays

Alpha and Beta globin mRNA is known to effect whole blood gene expression arrays. Globin mRNA transcripts derived from reticulocytes and red blood cells (to a lesser extent as anucleate) make up around 70% of total whole blood mRNA transcripts. These globin mRNA transcripts tend to dominate gene expression arrays, and decrease the transcript counts from cell populations of interest (i.e. leukocytes).

Preparatory work was performed to optimise gene expression microarrays from whole blood collected in PAXgene tubes. Total RNA obtained from whole blood storage media (e.g. PAXgene or Tempus tubes) contains a large proportion of globin mRNA. Globin mRNA was depleted using GlobinClear (Ambion). This pilot work demonstrated that the total number of detected probes was increased following globin mRNA depletion when analysed on the Illumina H12 expression array. Expression of haemoglobin alpha and delta, but not beta, mRNA transcripts was also significantly reduced.

The relative importance of globin clearance has been debated in the literature with some authors strongly advocating[404,405] the process in order to improve the sensitivity of microarray analysis. This methodological work was important in informing later experiments. A strength of this work was that RNA taken from the same PAXgene tube was used for both the globin cleared and non-globin cleared analyses. A limitation is that whilst certain globin mRNA transcripts were significantly reduced, the beta globin transcript was unchanged indicating that the experimental process may not have been completely effective at removing all globin mRNA. Whilst important to recognise the potential impact of globin mRNA on expression profiles, the additional steps taken to remove globin mRNA (including clean up steps either side) may degrade or deplete RNA samples if already low quality and/or quantity as well as incurring an additional consumables cost.

6.4.4 Conclusion

The relationship between DNA methylation and gene expression is complex and is likely to be cell specific. The location of DNA methylation change is critical when associating methylation with altered gene expression. Cell-specific changes in gene expression were seen in the top DMRs identified in chapter 3. For TXK where hypermethylation occurs within the TSS/promotor region, a reduced gene expression in whole blood and CD8+ cells was accompanied by a statistically significant negative correlation with DNA methylation in matched samples. Whilst similar convincing differences were not seen for the other DMRs/Ps, this may in part be related to type II statistical error and reflects similar experience in the wider field of epigenetics. Future work should be directed at those differentially expressed DNA methylation regions that occur within transcription start sites.

Chapter 7. Biomarker development from DNA methylation data

Abstract

Introduction

Existing biomarkers such as faecal calprotectin (FC) are highly sensitive at detecting patients with gut inflammation. However, FC is not specific for IBD, and the currently available biomarkers are less helpful in predicting disease prognosis. The development of biomarkers is a compelling translational application of epigenetic data. The aims of this chapter were to validate DNA methylation biomarkers previously identified in our paediatric study, and to identify new biomarkers from the present dataset capable of discriminating IBD patients from controls, as well as those IBD patients likely to experience a more severe disease course.

Methods

Paired methylation probes identified in the previous paediatric cohort were validated in the present adult cohort using linear discriminant analysis. The area under the receiver operating curve (AUC) with and without leave-out-one cross validation was calculated to assess model accuracy. The CMA package in R was used to assess different methods of new biomarker selection. Unsupervised consensus clustering was used to identify subclasses within the IBD cohort, and subclasses were assessed for need for surgery and immunomodulator therapy using Cox proportional hazards.

Results

The best performing of the previously described paired methylation probe biomarkers in the paediatric study were RPS6KA2/VMP1 probes (cg17501210/ cg12054453) and RPS6KA2/TNFSF10 probes (cg17501210/ cg01059398) which were able to accurately discriminate between disease and control in CD (AUC=0.84/0.81 respectively); IBD (AUC=0.79/0.79) and UC (AUC=0.73/0.71). Least absolute shrinkage and selection operator (lasso) modelling identified 30 methylation probes can be used to accurately discriminate IBD cases from controls (AUC = 0.898, sensitivity = 90.6%, specificity = 84.7%). Using unsupervised consensus clustering, three stable clusters were identified in the data methylation data. The three subclasses were associated with high-, moderate- and low-risk of requiring surgery (p=0.01), emergency hospital admission (p=0.0008) and

immunomodulatory therapy (p=0.02). These groups however were not independently predictive of outcome and unlike existing clinical markers.

Discussion

DNA methylation data may be used as diagnostic and prognostic biomarkers. Putative biomarkers identified in this chapter require further validation in independent cohorts.

7.1 Introduction

The existing biomarkers used to diagnose and prognosticate in IBD have been extensively described above (Section 1.9). Briefly, relatively good biomarkers are currently available to assist in the diagnosis of IBD (i.e. Faecal Calprotectin [FC]). FC is a highly sensitive marker of gut inflammation (AUROC=0.97, with a threshold of >50 μg/g demonstrating a 97% sensitivity and 0.74% specificity).[230] Whilst calprotectin is good at discriminating IBD from functional disease, calprotectin has a low specificity and can be elevated as a result of gut inflammation (e.g. infectious gastroenteritis, diverticulitis). Faecal calprotectin sampling can be challenging for both patients and laboratory staff; patients can find stool sample collection messy and distasteful, and biochemistry laboratory staff are often unable to process insufficient samples and processing delays can hinder clinical use. Therefore identifying a peripheral blood biomarker that does not rely on stool collection or mucosal biopsy sample at colonoscopy would be potentially attractive. New blood-based biomarkers could feasibly be used to complement FC. Furthermore, future diagnostic biomarker development should focus on increasing specificity of a test in identifying IBD from other inflammatory conditions, and to differentiate CD from UC.

Epigenetic markers have been proposed as putative diagnostic biomarkers for a range of conditions. Peripheral blood DNA methylation of SEPT09 is commercially available for the diagnosis of colorectal cancer.[353] There have subsequently been several studies investigating the use of SEPT09 in colorectal cancer screening.[406,407] In Adams et al,[284] two methylation probes were used to accurately discriminate children with CD and controls. Using linear discriminant analysis, the area under receiver operating characteristic curve (AUC) was as high as 0.98 which is as good as faecal calprotectin. A major limitation of much of the biomarker discovery literature is the lack of prospective validation.

Whilst FC performs well as a diagnostic biomarker, there is an unmet need for the identification of biomarkers that can accurately predict the course of disease. Such a biomarker may assist in the identification of patients who would benefit from early aggressive treatment with immunomodulators/biologic therapies or surgery ('top-down' approach), and those who could safely avoid the considerable toxicity and complications of such treatments without developing complications of IBD. As discussed in Chapter 1 (section 1.9.2), existing clinical parameters,[241] biochemical markers (CRP), [234,235] calprotectin, [243] serological markers[237–239] and genetic risk scores[254] have all been used to predict prognosis. In the field of transcriptomics which more closely aligns with epigenetic data, Lee *et al.* demonstrated that the gene expression profile of circulating CD8$^+$ T-lymphocytes is able to accurately predict a relapsing disease course from a stable one in patients with newly-diagnosed IBD.[260]

7.1.2 Aims

The aims of this chapter were:

1. To validate biomarkers identified in paediatric IBD by Adams et al
2. To identify novel biomarkers using DNA methylation data that could be used to
    a. Discriminate IBD from controls
    b. Discriminate CD from UC
    c. Predict patients with a more severe course
3. To identify sub-classes of IBD patients on the basis of DNA methylation data based on unsupervised consensus clustering.

7.2 Methods

7.2.1 Validation of LDA analysis performed in paediatric cohort published in Adams et al

Biomarkers identified in Adams et al were validated in the present adult methylation dataset. The previously published methylation probe pairings are listed in Table 76.[284] Linear discriminant analysis (LDA) was performed using beta values from the entire adult whole blood methylation dataset using the same methodology as previously described by Adams et al.[284] The MASS package in R was used for linear discriminant analyses, in which age and sex were included as covariates.[408]The Area under Receiver operating curve (AUC) was calculated for each probe pair using the ROC package.[409] As this was validation of previously described markers, the cohort was not split into testing/validation cohorts and no cross validation was performed.

7.2.2 CMA method selection

The CMA package[410] package aims to address the situation whereby the number of variables vastly outnumbers the number of samples, as is common in microarray studies. The 21 classification methods implemented in the package were compared, and the best performing method selected based on the area under the receiver operating curve (AUC). Further elaboration on the actual methods selected is provided in the results section.

7.2.3 Unsupervised Consensus Clustering

The second method employed unsupervised consensus (hierarchical) clustering[411] of median beta values of the methylation profile of IBD patients using the ConsensusClusterPlus package.[412,413] The number of stable clusters was assessed using the cumulative distribution function (CDF) [411] and the clest method.[414] Logistic regression was used to compare individuals classified according to clusters and clinical outcomes including need for surgery, emergency admission and immunomodualtor requirement/treatment escalation.

7.3 Results

7.3.1 Validation of LDA analysis performed in paediatric cohort published in Adams et al

Two of the previously described 44 CpG pairs had to be removed (those containing probe cg02292450) as the specific probes had been filtered during quality control and data processing. The AUC performance for discriminating case status ranged from between 0.84-0.52 in CD, 0.73-0.50 in UC and 0.79-0.54 in IBD. A complete summary of the data is presented in Table 77. The best performing of the paired methylation probe biomarkers described in Adams et al[284] were RPS6KA2/VMP1 probes (cg17501210/ cg12054453) and RPS6KA2/TNFSF10 probes (cg17501210/ cg01059398) which were able to accurately discriminate between disease and control in CD (AUC=0.84/0.81 respectively); IBD (AUC=0.79/0.79) and UC (AUC=0.73/0.71)(Figure 51)

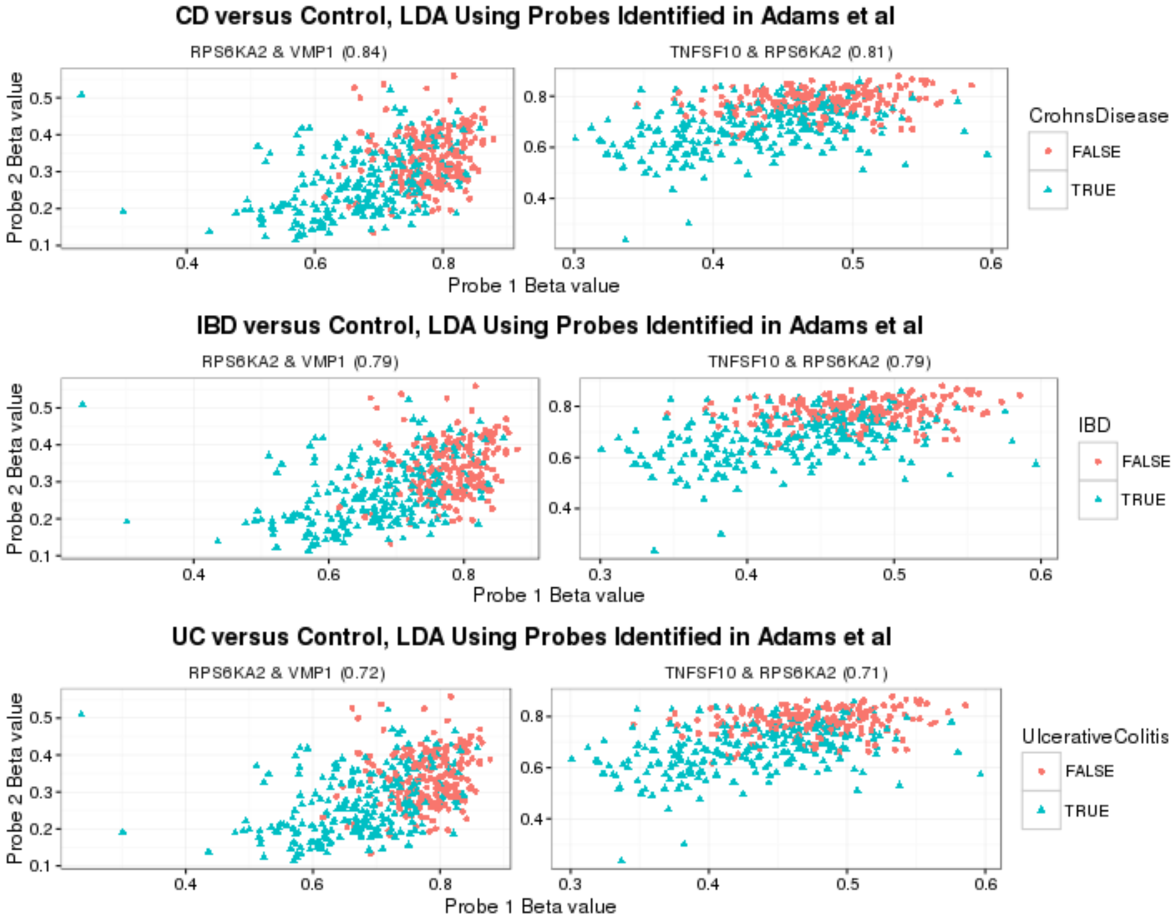


Figure 51 – Top two performing Methylation probe pairs from Adams et al at discriminating case status on Linear Discriminant Analysis (LDA). Top panel Crohn's disease versus controls, middle panel Ulcerative colitis versus controls, bottom panel ulcerative colitis versus controls. Red points = controls, Blue points = cases. Axes represent beta values of the two probes listed above the panel.

7.3.2 Novel diagnostic biomarker identification

Given that the aforementioned technique using LDA assessed only a proportion of top-ranking probes (many of which will be co-correlated) an alternative method of biomarker discovery has been employed. Using the CMA package, the available methods of variable selection were assessed (Table 78). Based on AUC, Lasso (least absolute shrinkage and selection operator[415,416]) was the best performing variable classification method. The LassoCMA function was used to perform the lasso algorithm for shrinkage and selection of CpG probes to be used as putative biomarkers. The cohort was arbitrarily split into a learning set (2/3 of the cohort = 287 individuals) and a testing set of 144 individuals. The L1 shrinkage intensity was tuned to provide the most accurate model, based on the AUC. This involved altering the shrinkage intensity (i.e. altering the number of CpG probes that algorithm could include in the model). The random seed was fixed to generate reproducible results. Both beta values and M-values were assessed.

7.3.2.1 IBD versus control

The lasso algorithm was tuned to determine the optimum shrinkage intensity (Table 31). The best performing shrinkage intensity was a normalisation fraction of 0.06, which included 30 methylation probes. This model including 30 probes was able to discriminate between IBD cases and controls with a high degree of accuracy (AUC 0.898, sensitivity 0.812, specificity 0.847, misclassification rate 0.174, Figure 52). The model included 30 probes including several of the most significant DMPs in the case control analysis presented in Chapter 3 (e.g. *RPS6KA2*, *VMP1*, and *BCL3*) but also some non-significant probes (Table 32). The number of methylation probes included in the model could be reduced to 3 probes (cg175012010 [*RPS6KA2*], cg09349128, cg25114611), however this led to a reduction in specificity for IBD and a higher misclassification rate (AUC 0.87, sensitivity 0.906, specificity 0.542 and misclassification rate 0.243).

Norm fraction	**Number of methylation probes included**	**AUC**	**Sensitivity**	**Specificity**	**Misclassification rate**
0.3	142	0.881	0.765	0.746	0.243
0.2	116	0.886	0.765	0.78	0.229
0.15	95	0.888	0.776	0.797	0.215
0.1	65	0.897	0.8	0.797	0.201
0.09	57	0.897	0.824	0.831	0.174
0.08	45	0.898	0.8	0.847	0.181
0.07	42	0.897	0.8	0.847	0.181
0.06	30	0.898	0.812	0.847	0.174
0.05	22	0.895	0.812	0.831	0.181
0.04	13	0.889	0.812	0.847	0.174
0.03	10	0.885	0.788	0.831	0.194
0.02	5	0.875	0.788	0.763	0.222
0.01	3	0.87	0.906	0.542	0.243

Table 31 – Tuning of lasso algorithm to alter the shrinkage intensity and thus the number of methylation probes included in the model. A shrinkage intensity of 0.06 was the optimum: this included 30 methylation probes and demonstrated the highest AUC/Sensitivity/Specificity and lowest misclassification rate.

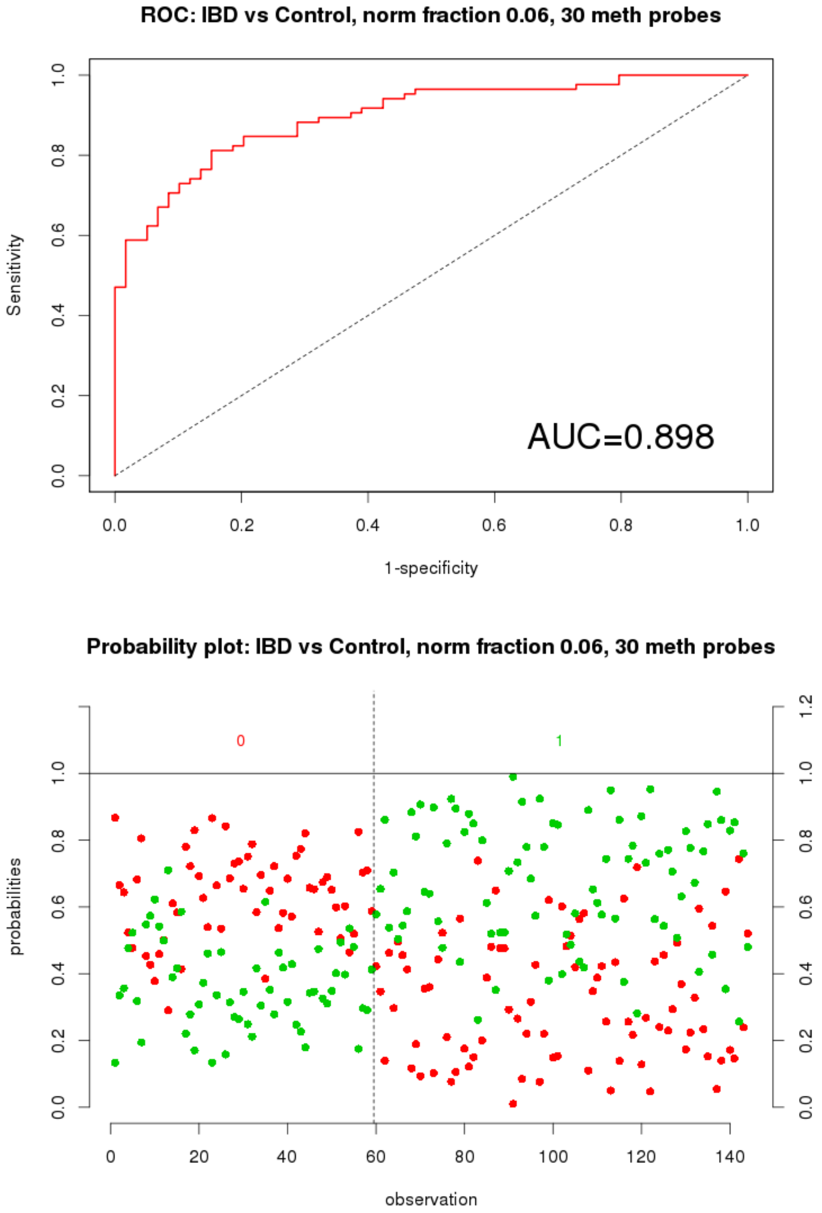


Figure 52 – Lasso modelling to discriminate IBD cases from controlsTop: Receiver operator curve for Lasso selected probes to distinguish IBD from controls using 30 methylation probes. Bottom: Probability plot. 0/red = controls, 1/green = IBD cases.

	absolute value of regression coefficient	**Probe Id**	**Chr**	**Gene Symbol**	**Δβ** IBD vs Cont	**P.Value** IBD vs Cont	**Holm adj.P.Val** IBD vs Cont
1	6.18	cg24971181	chr8	NA	-0.003	0.0001	1
2	5.48	cg22768358	chr11	ZBTB16	0.02	7.1E-12	3.2E-06
3	5.03	cg17501210	chr6	RPS6KA2	-0.08	2.7E-22	1.2E-16
4	4.65	cg25114611	chr6	NA	-0.04	1.1E-18	4.9E-13
5	4.01	cg26470501	chr19	BCL3	-0.03	5.8E-16	2.6E-10
6	2.83	ch.1.4690234F	chr1	LGALS8	-0.002	0.005	1
7	2.80	cg07012999	chr22	NA	0.01	0.005	1
8	2.79	cg23344935	chr12	PRR13	-0.005	0.019	1
9	2.37	cg16246188	chr11	ZBTB16	0.01	3.0E-09	0.001
10	2.07	cg04666911	chr11	LSP1	0.02	4.8E-08	0.02
11	1.91	cg20364632	chr6	NA	-0.02	7.6E-10	0.0003
12	1.83	cg09349128	chr22	NA	-0.04	3.1E-19	1.4E-13
13	1.60	cg17927096	chr10	NA	-0.009	0.0001	1
14	1.43	cg07242215	chr4	COX18	-0.003	0.001	1
15	1.21	cg26247646	chr1	KIF1B	-0.02	0.0003	1
16	1.05	cg24312865	chr10	PPP2R2D	-0.003	0.18	1
17	0.81	cg22881435	chr8	RAB11FIP1	0.02	1.1E-11	4.9E-06
18	0.75	cg07398517	chr3	NA	-0.04	6.1E-16	2.8E-10
19	0.53	cg09026415	chr5	TMEM161B	-0.007	0.0001	1
20	0.47	cg12582317	chr17	NA	0.05	5.7E-14	2.5E-08
21	0.43	cg01543300	chr19	TIMM13	-0.004	0.003	1
22	0.23	cg18877969	chr2	NMUR1	-0.003	0.009	1
23	0.15	cg20201143	chr6	SYNGAP1	0.01	1.9E-05	1
24	0.12	cg02716826	chr9	NA	-0.04	2.7E-15	1.2E-09
25	0.11	cg18589102	chr10	NA	0.02	0.004	1
26	0.09	cg07533100	chr16	NA	-0.01	0.12	1
27	0.09	cg08249698	chr16	RBFOX1	-0.02	1.6E-05	1
28	0.07	cg25653947	chr8	NA	0.03	2.6E-13	1.2E-07
29	0.04	cg23598089	chr1	ATP2B4	0.03	1.8E-08	0.008
30	0.01	cg12054453	chr17	VMP1	-0.07	3.9E-17	1.8E-11

Table 32 - Panel of Methylation probes selected by lasso algorithm to differentiate IBD from control. (Δβ = difference in beta values between IBD versus control, p.value and Holm adjusted p values derived from linear models IBD versus control with age, sex and estimated cell proportions as covariates)

7.3.2.2 CD and UC versus control

Similar results could be obtained for CD versus control (AUC=0.89, sensitivity=0.659, specificity=0.889, lasso norm fraction=0.13, 42 probes, Table 79, Table 80, Figure 79) and slightly inferior results for UC versus control (AUC =0.81, 12 probes, norm fraction =0.03, Table 81, Table 82, Figure 80).

7.3.2.2 CD versus UC

For CD versus UC the learning set was increased to include 3/4 of the cohort. The best performing model included 19 methylation probes was able to discriminate CD and UC with a reasonable degree of accuracy (AUC=0.719, sensitivity=1, specificity=0.111, misclassification rate=0.533, Table 83, Table 84, Figure 81).

7.3.2.3 IBD prognosis

The lasso approach was also applied to the IBD cases alone in an attempt to define models associated with specific outcomes (resectional surgery and/or colectomy, emergency hospital admission, need for immunomodulatory), however no models were identified with an AUC >0.5 .

Using DNA methylation data to predict prognosis in IBD

7.3.3 DNA methylation data and disease location and behaviour

Following correction of multiple testing, there were no DNA methylation markers that predicted Montreal disease location or behaviour in CD or Paris disease extent in UC (mm = ~1 + Extent + CD8 + CD4 + NK + BCell + Mono + Gran + EverSmoked + Age + Sex). Using multi-dimensional scaling, there was no obvious clustering according to disease location (CD, Figure 82), behaviour (CD, Figure 83) or extent (UC, Figure 84).

7.3.4 Unsupervised consensus clustering

An attempt to identify subclasses of IBD patients based on DNA methylation data unsupervised consensus clustering was performed on the top 5000 DMPs identified in the primary analysis (IBD versus controls, whole blood). Kmeans clustering based on the Pearson correlation coefficient was used as the final clustering method, although similar results were obtained by using other methods (kmeans clustering based on Spearman's correlation, hierarchical clustering, PAM clustering). Three stable clusters (Figure 53 A) were used for downstream analysis, chosen on the basis of the cumulative distribution function (CDF, Figure 53 B) [411] and the clest method.[414] Survival analysis demonstrated that the three subclasses of IBD based on the consensus clustering formed high-, intermediate- and low- risk groups for requiring surgery (resectional surgery [CD] and/or colectomy) and immunomodulatory requirement. The demographics of each group are presented in Table 33. The results of Cox proportional hazards regression performed to determine factors (including consensus subtypes of IBD) independently associated with surgery and need for immunomodulator and escalation of therapy as defined by Lee et al.[260]

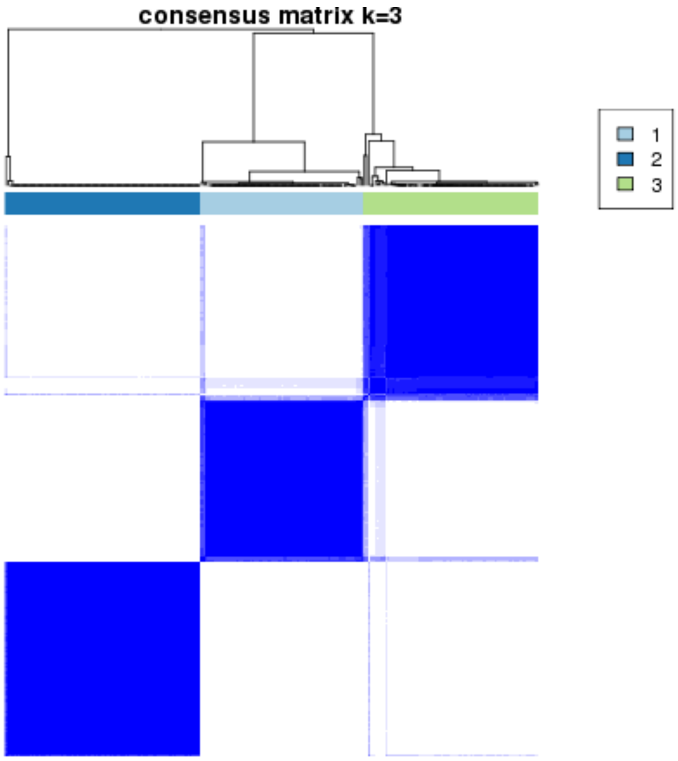


Figure 53- Unsupervised Consensus Clustering to identify IBD subclasses based on DNA methylation data. Three stable clusters were formed.

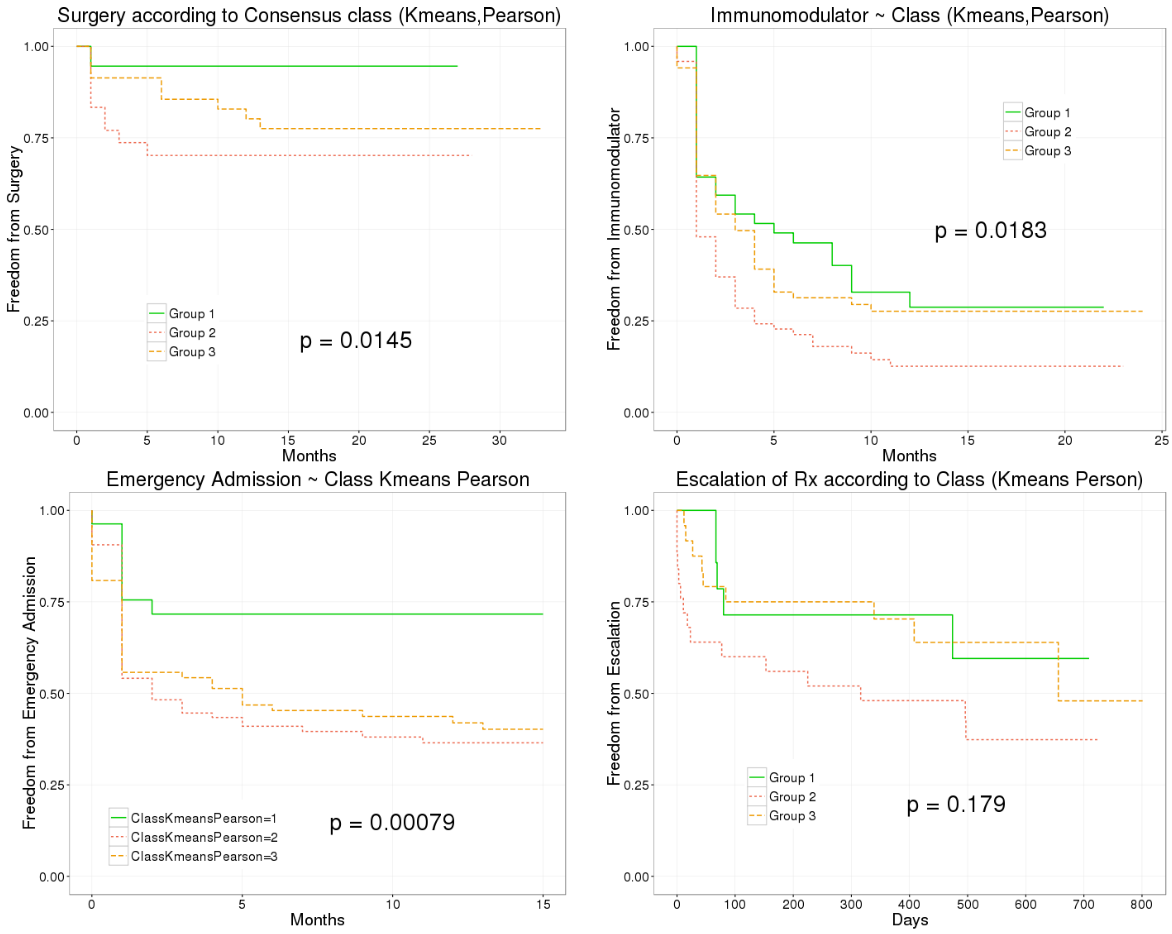


Figure 54 - Survival analysis according to DNA methylation consensus class. Top left panel – Time to surgery. Bottom Left – Time until emergency admission. Top right panel – Time until requirement for Immunomodulator (oral or IV steroid, anti-TNFalpha drug, ciclosporin, Methotrexate, thiopurine). Bottom right panel – Criteria of treatment escalation defined by Lee et al[260] (surgery, step up to 2 or more immunomodulators). P values denote ChiSquared test for difference between survival curves with 2 degrees of freedom

	Group 1 (low risk) (n=62)	**Group 2** (High risk) (n=91)	Wilcox p value (high risk versus low risk)	**Group 3** (intermediate risk) (n=87)	Kruskall-Wallis p value (all three groups)
Age	37.6 (27-50.3)	30.8 (24.6-47.7)	0.1	33.4 (25-47.4)	0.3
Female (%)	31 (50)	36 (39.6)	0.3	42 (48.3)	0.4*
Follow up length	18 (13-28)	19 (7-25)	0.6	20 (10-29)	0.4
CRP	3.5 (1.25-5.75)	14.5 (4-37.5)	0.002	11 (2.25-26)	0.004
Albumin	40 (37.25-41)	34 (29-36.75)	1.7E-5	37 (33-39)	1.89E-5
Haemoglobin	141 (127-149.5)	132.5 (115.5 – 142)	0.02	130 (122-144)	0.06

Table 33 - Demographics of three IBD subgroups generated by Unsupervised consensus clustering (data presented as medians (interquartile range), * = chi squared test)

	Hazard ratio	95% CI lower	95% CI upper	P value
Surgery (n=106, events= 19)				
Consensus Class	1.5	0.8	3	0.2
CRP	1	0.99	1.01	0.9
Albumin	0.9	0.86	1.04	0.3
Age	1	0.97	1.03	0.9
Male sex	2.3	0.74	6.93	0.2
Immunomodulator (n=107, n=83)				
Consensus Class	0.8	0.62	1.12	0.2
CRP	1	0.99	1.01	0.2
Albumin	0.9	0.86	0.95	5.2E-5***
Age	1	0.98	1.01	0.7
Male sex	0.9	0.56	1.40	0.6
Escalation of therapy (n=60, events=28)				
Consensus Class	0.8	0.41	1.34	0.3
CRP	1	0.99	1.01	0.7
Albumin	0.9	0.78	0.92	0.0001**
Age	1	0.97	1.04	0.7
Male sex	1.5	0.60	3.58	0.4
Emergency admission (n=100, events=68)				
Consensus Class	1.2	0.86	1.7	0.3
CRP	1	0.998	1.0072	0.6
Albumin	0.9	0.87	0.96	0.0007
Age	0.99	0.98	1.02	0.7
Male sex	0.80	0.5	1.3	0.4

Table 34 - Cox proportional hazards model for factors associated with poor outcome in IBD. Risk of surgery (resection surgery in CD +/- colectomy), need for emergency hospital admission, Time until requirement for Immunomodulator (oral or IV steroid, anti-TNFalpha drug, ciclosporin, Methotrexate, thiopurine). need for immunomodulatory and escalation of therapy as defined by Lee et al.

In an effort to understand contributors to the three subclasses of IBD, the residuals of a linear model of cell proportion (estimated using Houseman method) and smoking were subsequently used in a further consensus clustering analysis. Figure 55 demonstrates the effect of including cell proportion (Figure 55 A), and both cell proportion and smoking (Figure 55 B) on the strength of the clustering.

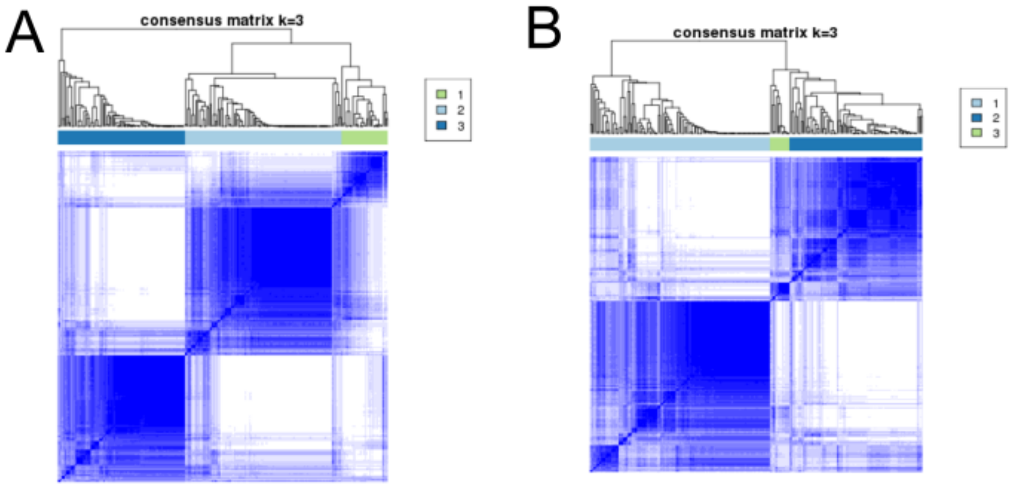


Figure 55 - Consensus Clustering using residuals of a linear model based on cell proportions estimated by Houseman method and Smoking status

7.4 Discussion

This study provides prospective validation of the paired two probe methylation biomarkers described by paediatric onset CD.[284] Of the 44 probes pairings described in Adams et al tested in this dataset, the best performing pairings were RSP6KA2/VMP1 and RPS6KA2/TNFSF10. The strongest discriminatory value was seen in CD, and this may reflect the nature of the testing cohort (CD-only) in the previous study. However there was also good discriminatory value in both IBD and UC. In the previous study, the AUC in the early-onset group was as high as 0.98, however in this study the maximum AUC was 0.84 in CD. As has previously been noted, the IBD/CD-specific methylation pattern weakens with age and may result from the accumulation of confounding epigenetic marks caused by aging and environmental exposures in development from childhood to adulthood.[284,417]

Whilst linear discriminant analysis performed relatively well in discriminating IBD cases from controls, there are other machine learning methods better suited to such large datasets where the number of variables vastly exceeds the number of samples (so-called "$p \gg n$" setting). The

CMA package was designed to bring together the multiple methods available for class prediction in microarray-type datasets, and allow comparison and selection of the best performing method.[410] Following assessment of classification methods, the lasso method was selected.[416] Lasso (least absolute shrinkage and operator selection) an established machine learning method developed by Tibshirani[416] in the 1990's and has been used widely for class prediction the biomolecular setting. Briefly, the model shrinks the absolute value of regression coefficients to less than a constant, and sets non-relevant variables to zero, thereby producing a simpler model without these variables.[416] Using this method a model including 30 probes was able to discriminate IBD from controls with a high degree of accuracy (AUC=0.898) and outperformed the LDA techniques. The final 30-probe model is easily scalable into a high-throughput pyrosequencing panel and such a non-invasive peripheral blood biomarker could be used to stratify patients to further intrusive investigations such as colonoscopy. Existing clinically available biomarkers such as faecal calprotectin[230] already provide similar utility but are unable to distinguish the two forms of IBD. A different 19-probe methylation-based panel may confer an additional benefit in discriminating CD and UC (AUC=0.719), which can be critical for decision-making in terms of medical and surgical management. In this chapter the cohort was split into a testing and validation set rather than using cross validation. Whilst arbitrarily splitting the cohort is robust method, the main disadvantage to this is that the power to detect variables is reduced in a smaller dataset. Expression and genetic data derived in later chapters was also incorporated into biomarker discovery models, however the high level of statistical significance achieved by the methylation data in discriminating cases and controls meant that the genetic and gene expression variables was not selected in the best performing models.

Lee et al used unsupervised consensus clustering on CD8+ transcriptomic data to subclassify IBD patients, and demonstrated certain expression signatures were associated with worse clinical prognoses (need for surgery, treatment escalation).[260] When the same method has been applied to this whole blood DNA methylation data, three distinct subclasses of IBD were identified, corresponding to a high-, intermediate- and low-risk groups of a more severe disease prognosis. Using survival analysis, there was a significant difference in the risk of resectional surgery or colectomy (p=0.02), need for emergency hospital admission (p=0.008) and for the time to immunomodulatory requirement (p=0.02) in the three groups. There was no significant difference when the same criteria for escalation of treatment used by Lee et al[260] (p=0.2), however the number of patients with this level of clinical phenotype data was small. Whilst this method subclassifying IBD patients was initially hopeful, on further examination there were several important limitations. Firstly, when using Cox proportional hazards regression analysis, the consensus clustering group was not independently predictive of need for surgery or immunomodulatory use, whereas an existing clinically available biomarker, albumin, was. Secondly, a major strength of the Lee et al paper, was the lack of differences in baseline characteristics, disease extent and behaviour and key biochemical markers between the two IBD subclasses.[260] This was an exciting finding suggesting the gene expression signature was not the result of baseline differences in IBD subclasses that are already known or can be readily assayed in clinic. In the present analysis of DNA methylation data however, the three groups all demonstrated significant baseline differences, with the "high-risk" group containing more males, younger patients, with higher baseline CRP and lower baseline albumin and haemoglobin (all known to be associated with worse prognosis). A further limitation of this analysis is the known effect of cell proportion and other factors (e.g. smoking) on DNA methylation data. When differences in cell proportion and smoking status were accounted for, the strength of clustering decreased significantly. Lastly, since the publication of Lee et al,[260] there have been criticisms of unsupervised consensus clustering in the literature, in particular the fact that stable clusters can be derived from randomly generated data.[418] Taken together, this model would not be suitable as a biomarker, given the expense of generating the methylation data, and that presently available clinical parameters contribute to the clustering effect seen and may be superior at predicting patients at risk of a worse prognosis.

Conclusions

These data demonstrate the considerable translational potential of DNA methylation data as diagnostic and prognostic biomarkers. The underlying biomarker signals may be driven by differences in cell proportions between cases and controls. Care must be taken that in developing biomarkers from sophisticated (and expensive) '-omic' technologies such as DNA methylation that the putative biomarkers are not merely a surrogate for easily obtainable simple clinical parameters such as white cell count.

Chapter 8. MicroRNA in IBD

Abstract

Introduction

MicroRNAs (miRNA) are small non-coding nucleic acids that have the capacity to modulate gene expression through direct inhibition of translation or by inducing cleavage of mRNA. miRNAs have been increasingly implicated in many of the important IBD pathogenic pathways including autophagy, intestinal epithelial barrier integrity and the Th17 pathway. In common with all epigenetic mechanisms, miRNA expression is dynamic and cell-specific. The aim of this study was to characterise miRNA expression in separated peripheral blood immune cells in CD compared with controls.

Methods

Small RNA sequencing (RNA-seq) was performed on RNA extracted from CD14+, CD4+ and CD8+ cells isolated from 8 newly diagnosed cases of ileal or ileocolonic CD and 8 age and sex matched controls. A small RNA-seq analysis pipeline was optimised. TargetScan and DIANA/miRPath were used to predict downstream mRNA targets for differentially expressed miRNAs.

Results

There was a median of 2.4 million reads per sample (range 132,800-12.8 million reads per sample). Several normalisation methods were tested, and filtered quantile scaled normalisation was selected on the basis of multidimensional scaling plots. One microRNA was differentially expressed in CD compared with controls (hsa-miR-503-5p log fold change = 0.7, FDR adjusted p = 9.1×10^{-5}) in CD4+ lymphocytes. Using different normalisation methods miR-503-5p was no longer statistically significantly differently expressed between cases and controls. There were no other differentially expressed miRNAs in the other cell types. TargetScan demonstrated 101 gene targets for miR-503 and DIANA/miRPath highlighted 61 related pathways.

Discussion

The small number of cases used in this experiment raises the possibility of both type I and II error. Larger number of participants are required to detect a true difference or otherwise

between cases and controls. The methodology used to generate these data is useful in informing future small RNA-seq experiments in the context of complex disease.

8.1 Introduction

MicroRNAs (miRNAs) are non-coding RNAs between 18-23 nucleotides in length that can act as post-transcriptional modifiers of gene expression.[314] miRNAs were initially described in the nematode *C.elegans*[419] in 1993 and subsequently Fire and Mello won the Nobel prize in 2006 for their elucidation of the mechanisms by which miRNAs affect gene expression.[420] miRNAs have been shown to regulate and influence many aspects of plant and animal health and disease.[421] The number of miRNA-related publications is rapidly expanding (Figure 56).

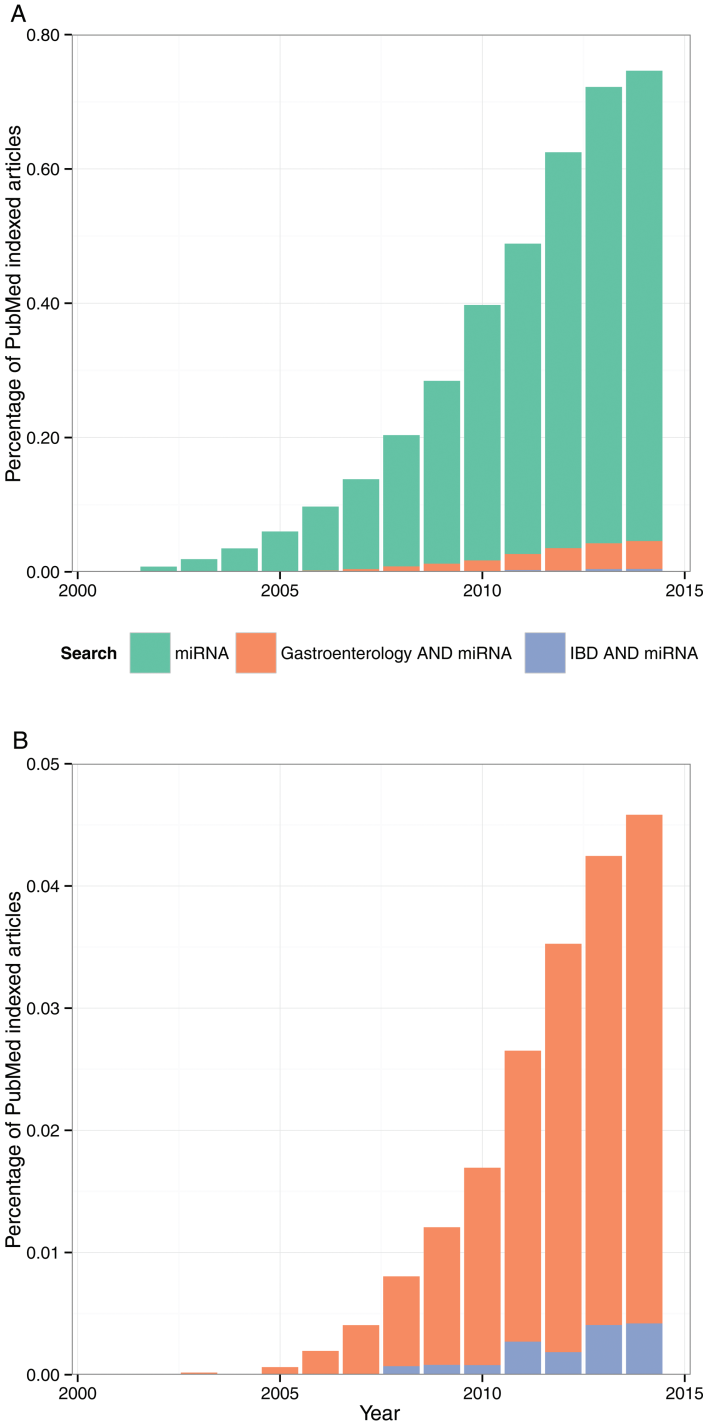


Figure 56 –Number of miRNA related citations on Pubmed.org. A – Total miRNA publications (green) with proportion related to "Gastroenterology". B – miRNA AND Gastroenterology and IBD (purple) publications. Normalised for total number of PubMed citations. Taken with permission from Kalla, Ventham et al with permission (License number 3755840927686)[314]

MiRNAs have been increasingly implicated in many of the important IBD pathogenic pathways including autophagy,[422–424] intestinal epithelial barrier integrity[425] and the Th17 pathway.[422,426] Many of the assumption-free screens of miRNAs in IBD to date have used microarray-based platforms that work by hybridisation of miRNAs to complementary probes immobilised on a chip or bead.[427] More recently next generation sequencing (NGS), specifically small RNA sequencing (RNA-seq) has been used to characterise miRNA profiles. RNA-seq utilizes massive parallel sequencing, generating millions of small RNA reads per sample.[427] Small RNA-seq has a number of potential advantages over microarray platforms. RNA-seq provides much more comprehensive view of the RNA sample, and profiles all species of RNA (not confined to microRNA). RNA-seq outputs data as actual reads rather than probe intensities used in microarrays (which often require normalisation and can lead to batch effects). RNAseq can detect microRNAs at very low copy numbers (rather than probe intensity used in microarrays), however this is dependent on the depth of coverage. Microarrays rely on pre-designed probes based on previously described microRNAs (also leading to problematic probe redundancy and annotation), whilst RNA-seq is 'assumption-free' and is not dependent on prior sequence knowledge. RNA-seq also allows discovery of novel microRNAs. A study comparing both methods demonstrated RNA-seq to perform better in identifying low abundance transcripts, genetic variants and differentiating biologically different isoforms.[428] Microarray use still predominates due to its current relative cheap price and well defined data analysis pipelines.

Despite the advantages of the RNA-seq technique described above, only one study has used this technique in the context of IBD.[429] Lin and colleges used the Illumina platform to sequence small RNAs obtained from fresh frozen resection specimens from active and inactive IBD patients (CD=9, UC=10). The comparator group used was diverticular disease (n=18). Of 44 differentially expressed microRNAs, nine were aberrantly expressed in IBD patients compared to diverticular disease controls. Four of the nine differentially expressed microRNAs were successfully replicated using qPCR (miR-31,206, 424, 146a) and were differentially expressed in formalin-embedded samples in IBD compared to diverticular disease, ischaemic colitis and infectious colitis.

The aim of this study was to identify differentially expressed miRNAs in circulating leucocytes of newly diagnosed patients with Crohn's disease compared with controls.

8.2 Methods

8.2.1 Patient samples

Patients were recruited as described in the methodology section (2.1.1 Patient selection). Eight treatment naïve patients with ileal (L1) or ileocolonic (L3) CD and eight age and sex matched healthy controls were selected for this study. Lymphocyte subsets (CD4, CD8 and CD14) were isolated from peripheral blood samples using immunomagnetic separation as described in Chapter 2 (2.2 Cell separation). RNA and microRNA was extracted and quality assessed using methods described in Chapter 2 (2.5 Nucleic acid quantification and quality assessment).

8.2.2 Vacuum concentrating RNA samples

An input volume between 1-6 µL was required for creation of small RNA libraries. As the RNA samples derived from lymphocytes were dilute, it was necessary to concentrate RNA samples to ascertain approx. 1000 ng in 5 µL. The samples were aliquoted in the required volume to a 96 well plate for a total RNA quantity of 2 µg, therefore half of this volume was taken to obtain a total RNA quantity of 1 µg. Five mircolitres of 0.1M EDTA was added to each sample to inactivate RNases. The samples were placed in eppendorf tubes with lids off in a vacuum concentrator for a variable amount of time to concentrate the samples depending on the initial volume. Following concentration, dry samples were resuspended with 6 µL of 0.1M EDTA.

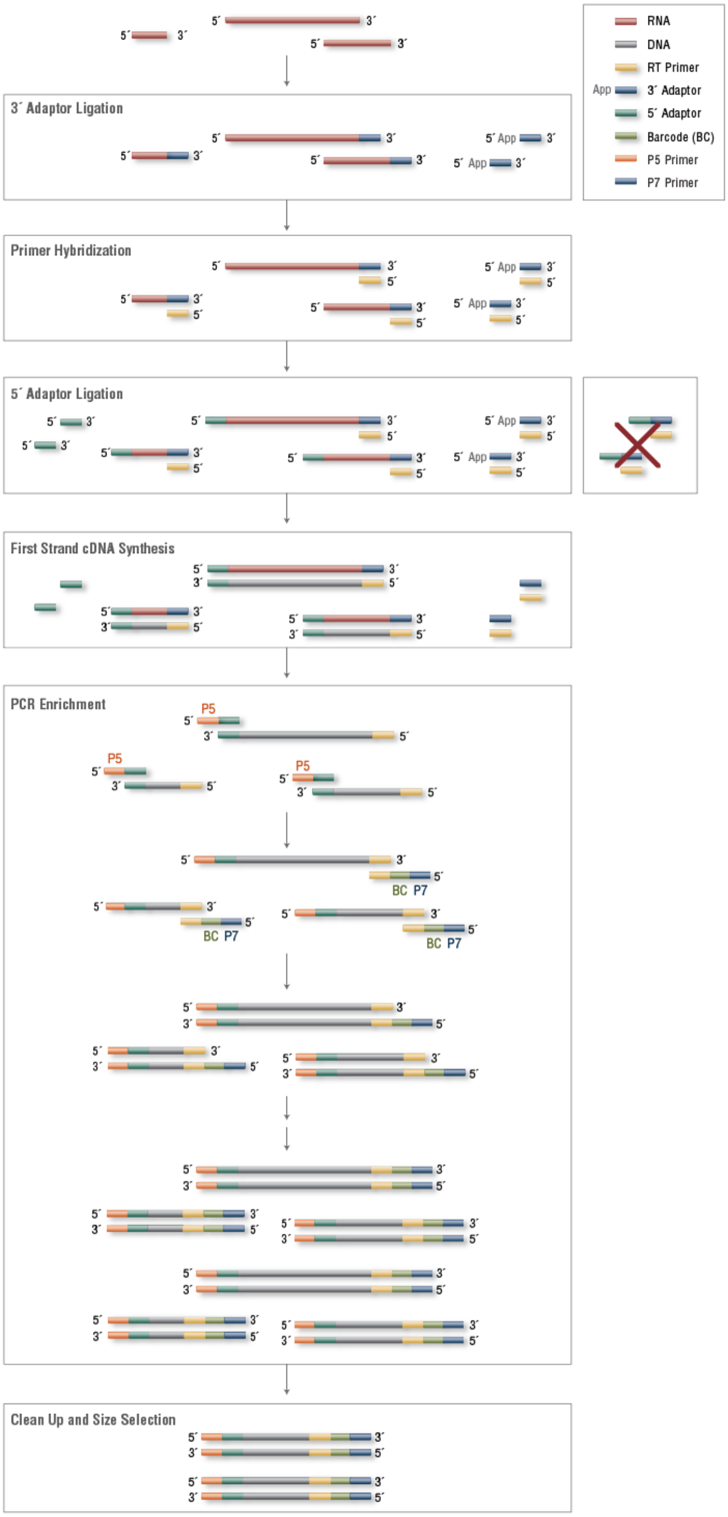


Figure 57 – Overview of NEBNext multiplex kit Small RNA Library Prep Set for Illumina. Taken from https://www.neb.com/products/e7300-nebnext-multiplex-small-rna-library-prep-set-for-illumina-set-1

8.2.3 Small RNA library creation

The following experiments were conducted with the help and supervision of Juan Quintana and Dr Amy Buck in the Centre for Immunity, Infection and Evolution, Ashworth Laboratories, Kings Buildings, Edinburgh. Small libraries were created using RNA from separated lymphocytes using the NEBNext multiplex kit Small RNA Library Prep Set for Illumina (New England Biolabs).

8.2.3.1 Ligate 3' SR Adaptor to RNA samples

Initially the 3'SR adaptor was diluted 1:2 in nuclease free water. In a nuclease free tube, 1 µL of the diluted adaptor, input RNA (variable according to concentration of RNA samples) and a variable amount of nuclease free water were combined to add up to 6 µL. Samples were incubated in a pre-heated thermal cycler for two minutes at 70 °C. The 3'Ligation buffer (half volume used, therefor 5 µL per samples), and 3'Ligation enzyme mix (half vol, 1.5 µL) were added to the sample and incubated for 1 hour in the thermal cycler.

8.2.3.2 Reverse Transcription primer to remove excess unbound 3' Adaptor.

Half volumes of SR RT reverse transcription primer (0.5 µL per sample, not diluted) were combined with nuclease free water (2.25 µL per sample) and added to the 3'SR ligated sample and placed on the heat cycler for: 5minutes @ 75 °C; 15 minutes at 37 °C, and 15minutes at 25 °C. This step is done to prevent adaptor dimer formation; by allowing the SR RT Primer to bind excess 3' SR Adaptor (single stranded DNA) into double stranded DNA. As dsDNA is not a substrate for T4 RNA Ligase 1 used in the subsequent step, the 5' SR adaptor is not added to these dsDNA products.

8.2.3.3 Ligate the 5'SR Adaptor to the RNA sample

When used for the first time, the 5' SR Adaptor was re-suspended from powder form to solution in 120 µL of nuclease free water. From this an aliquot required for the library was taken, in this case 1.6 µL of 5'Adaptor, and diluted with 1.6 µL of water, providing 1 µL of 1:2 diluted adaptor for each RNA sample. Prior to adding to the sample, the adaptor was denatured at 70 °C for 2 minutes and immediately placed on ice, and used within 30minutes. The following components were added to the RNA sample, 1 µL of denatured and diluted 5' SR Adaptor, 5' ligation buffer mix (half volume, 0.5 µL), and 5' Ligation enzyme mix (1.25 µL, half volume). The sample was incubated in the thermal cycler for 1 hour.

8.2.3.4 Reverse transcription of the sample

To the adaptor ligated RNA sample, 4 µL of first strand synthesis reaction buffer, 0.5 µL of murine RNase inhibitor, and 0.5 µL of Protoscript II reverse transcriptase was added to the sample and incubated for 1 hour at 60 °C.

8.2.3.5 PCR Amplification

To the reverse transcribed reaction mix, the following reagents were added: 25 µL of LongAmp *Taq*X2 Master Mix, 1.25 µL SR Primer, 2.5 µL nuclease free water, and 1.25 µL Index Primer (Primers 1-12, different number/barcoded primer added to each sample). Two different PCR cycles lengths were attempted: 15 and 22 cycle PCR. Gel assessment demonstrated that 22 cycles provided a better result (Figure 58). The final protocol used is summarised in Table 35.

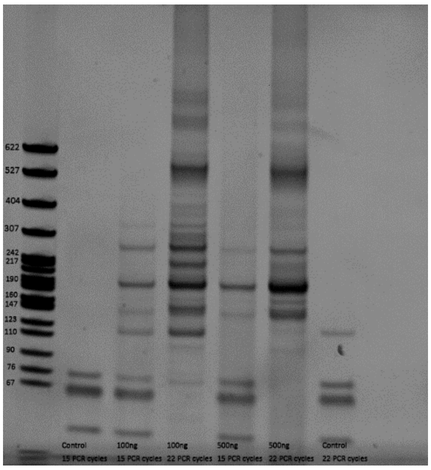


Figure 58 – TBE polyacrylamide gel demonstrating optimisation of PCR cycle length and input RNA amount for small RNA library creation

Table 35 – Summary of thermocycler settings for library creation (Taken from NEBNext multiplex Small RNA Library Prep Set for Illumina (Set 1) Protocol)

Cycle Step	Temp	Time	Cycles
Initial Denaturation	94°C	30 seconds	1
Denaturation	94°C	15 sec	22
Annealing	62°C	30sec	
Extension	70°C	15sec	
Final Extension	70°C	5minutes	1

8.2.3.6 Gel assessment of PCR purity and microRNA band

The purified PCR product was run on a 6% TBE acrylamide gel (Invitrogen) using TBE running buffer (100mls of 5X TBE, diluted with 400mls of purified water). The PCR product was mixed with 5 μL of gel dye. The first well was loaded with Quick-Load pBR322 DNA-Mspl Digest (5 μL). The following wells were loaded with 3.5uL of appropriate sample. Electrophoresis was performed for 40 minutes at 180V, or until the blue dye reached the end of the gel. The gel was carefully removed and put into the TBE buffer, and stained with ethidium bromide. The gel was carefully placed in the syngene UV transilluminator and image capture performed.

8.2.3.7 Pooling of samples for sequencing

Samples of cDNA were pooled prior to sequencing according to PCR product band intensity and barcode primer (Appendix 6 - Chapter 8 microRNAs in IBD
Table 85). Two pools were created of high PCR band intensity, and two for low PCR band intensity. Groups were compiled to ensure samples with the same barcode were not pooled together. (Also group so that in group comparisons could be performed if there was a problem with sequencing e.g. CD14 in pool 1, CD4 in Pool 5, and CD8 in Pool 6). For high PCR band intensity samples, the relative volume contribution to the pool was reduced (5 μL) and for medium PCR band intensity (Between 7-10 μL). Low PCR band intensity all contributed 10 μL to the pool from each sample.

8.2.3.8 Size selection of the purified amplified cDNA library on 6% polyacrylamide gel

The purified PCR product was run on a 6% TBE polyacrylamide gel (Invitrogen) using TBE running buffer (100 mL of 5X TBE, diluted with 400 mL of purified water) to size select microRNA. The pooled PCR products were mixed with blue gel loading dye (5 μL). The first well was loaded with Quick-Load pBR322 DNA-Mspl Digest (5 μL). The volume of pooled sample was loaded according the size of pool (e.g. Pool 1 total volume 96 μL, split between 9 wells=11 μL in each well, plus one ladder). Electrophoresis was performed for 1 hour at 180V, or until the blue dye reached the end of the gel. The gel was stained with ethidium bromide stain for 2-3 minutes. The exposure to UV light and ethidium bromide was kept to a minimum to prevent cDNA damage in the purified samples.[430] Before cutting out the size selected band on the gel, gel was imaged on a UV transilluminator to observe an expected microRNA band at ~140bp (21 nucleotide microRNA) and 150bp (30 nucleotide RNA) corresponding to the adapter ligated constructs. The gel was taken to the dark room where the desired gel band is cut out of the gel using a scalpel blade with the gel on the UV transilluminator (performed by Juan Quintana, Figure 59). The post-cut gel was again imaged to ensure that the correct band had been removed.

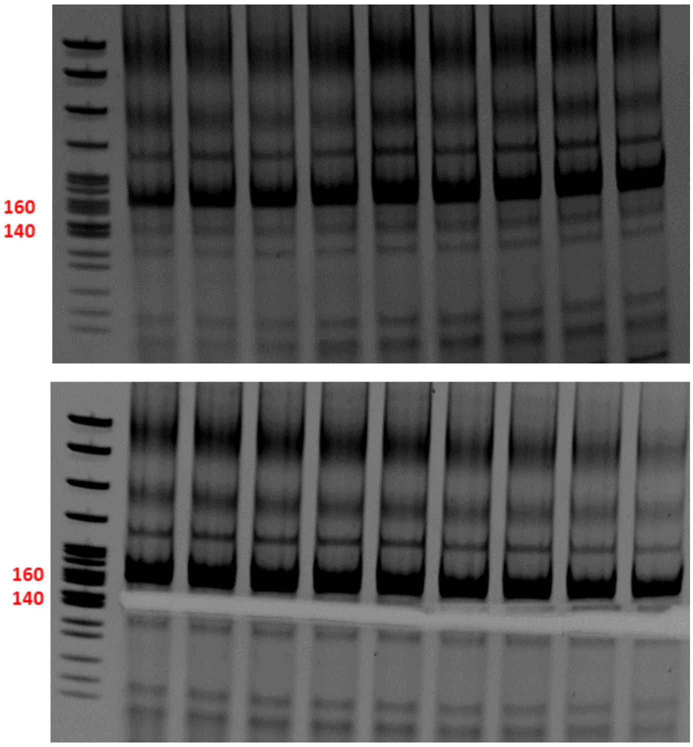


Figure 59 – Size selection of purified amplified cDNA library on 6% polyacrylamide gel. Expected microRNA exists between 140 and 150 bp before and after cut. The image shown is Pool 6 (medium PCR band intensity).

The gel was crushed by centrifugation through a small hole made in a microfuge tube placed inside a larger Eppendorf (7 minutes at 16000RCM). The crushed gel was incubated in 300 µL of water overnight at 4 °C on a rotating incubator to elute cDNA. Samples were then quality assessed using the Agilent DNA chip (Table 36).

	Pool concentrations (ng/ µL)	Total cDNA in 15 uL	Vol DNA added (µL)	EB buffer (µL)	Concentration in sample for GenePool (ng/ µL)
Pool 1	5.17	77.55	7.25	7.75	2.499
Pool 2	6.96	104.4	5.39	9.61	2.501
Pool 5	4.9	73.5	7.64	7.36	2.496
Pool 6	3.17	47.55	11.83	3.17	2.500

Table 36 – Pooled samples for sequencing

8.2.4 Sequencing

Sequencing was performed using the Illumina HiSeq 2000/2500 platform (performed at Edinburgh Genomics). The Illumina HiSeq uses the polymerase-based, sequence by synthesis technique. This system uses a polymerase to amplify (bridge amplification) the DNA into DNA colonies or 'clusters'. Four types of florescent reversible-terminator base (RT base) are added, corresponding to the four bases for sequencing, and the non-incorporated bases are washed away. A camera takes an image of the fluorescently dyes nucleotides, and converts the analog image to digital output (analogue to digital conversion). The marker nucleotides are chemically removed from the DNA, together with the terminal 3' blocker, allowing the next base to be sequenced. Therefore DNA chains are extended one nucleotide at a time (sequence by synthesis technique). The process is carried out in parallel (massively parallel signature sequencing, next generation sequencing [MPSS]) and cameras with a faster analogue to digital conversion rate have allowed increasingly rapid sequencing capabilities. One rapid flow cell lane (V1 chemistry, Yielding approx. 100million reads per lane) was used per sample pool with single-ended reads. Rapid mode provides a faster processing time, but less depth of sequencing compared to high output mode. Individual samples within the pool were indexed using unique barcode sequences and the 5' adaptor sequences were removed. Data was outputted in the form of fastq files.

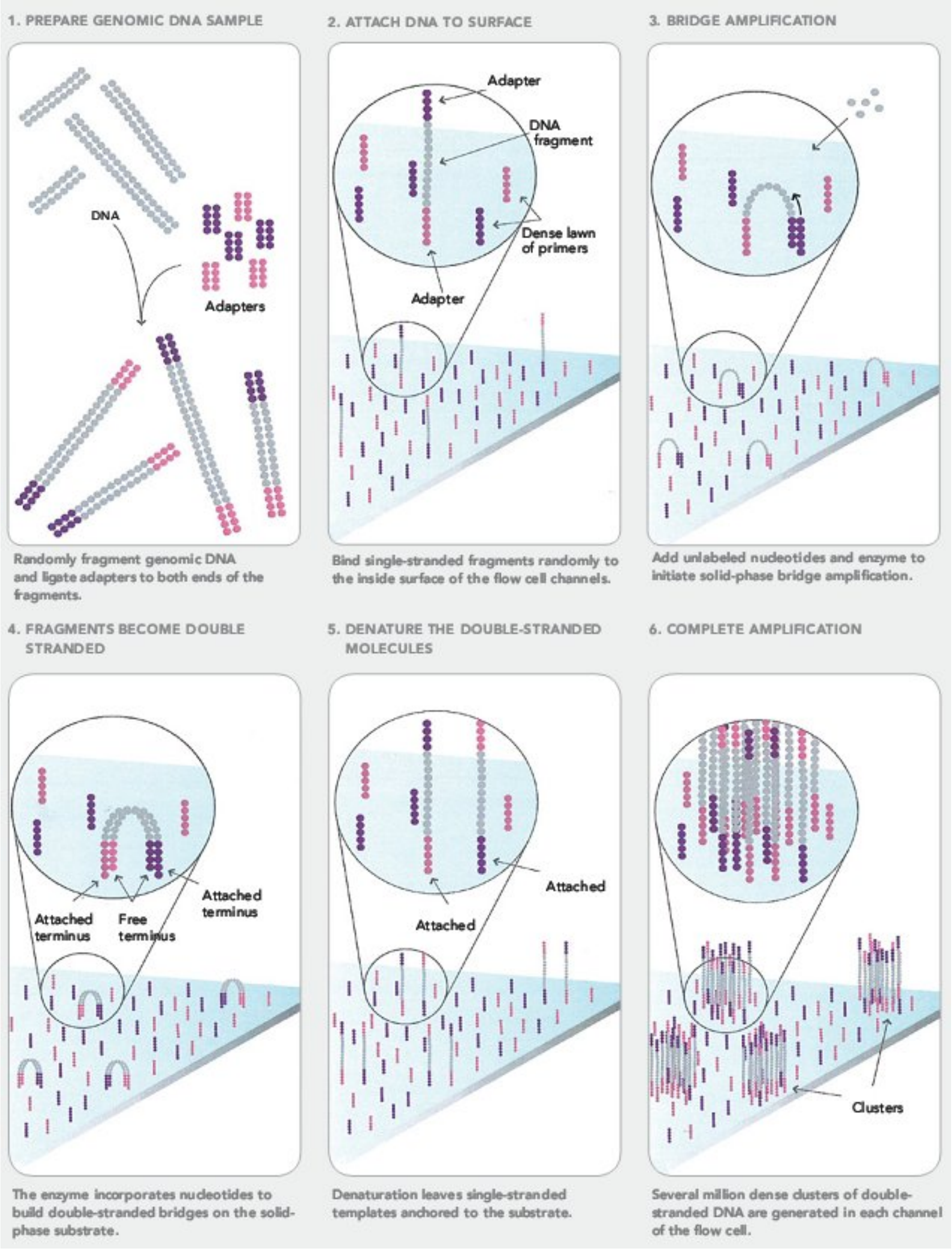

1. PREPARE GENOMIC DNA SAMPLE
DNA
Adapters
Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.
2. ATTACH DNA TO SURFACE
Adapter
DNA fragment
Dense lawn of primers
Adapter
Bind single-stranded fragments randomly to the inside surface of the flow cell channels.
3. BRIDGE AMPLIFICATION
Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.
4. FRAGMENTS BECOME DOUBLE STRANDED
Attached terminus
Free terminus
Attached terminus
The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.
5. DENATURE THE DOUBLE-STRANDED MOLECULES
Attached
Attached
Denaturation leaves single-stranded templates anchored to the substrate.
6. COMPLETE AMPLIFICATION
Clusters
Several million dense clusters of double-stranded DNA are generated in each channel of the flow cell.

Figure 60 – Summary of Illumina bridging reaction during next generation sequencing (http://seqanswers.com/forums/showthread.php?t=21)

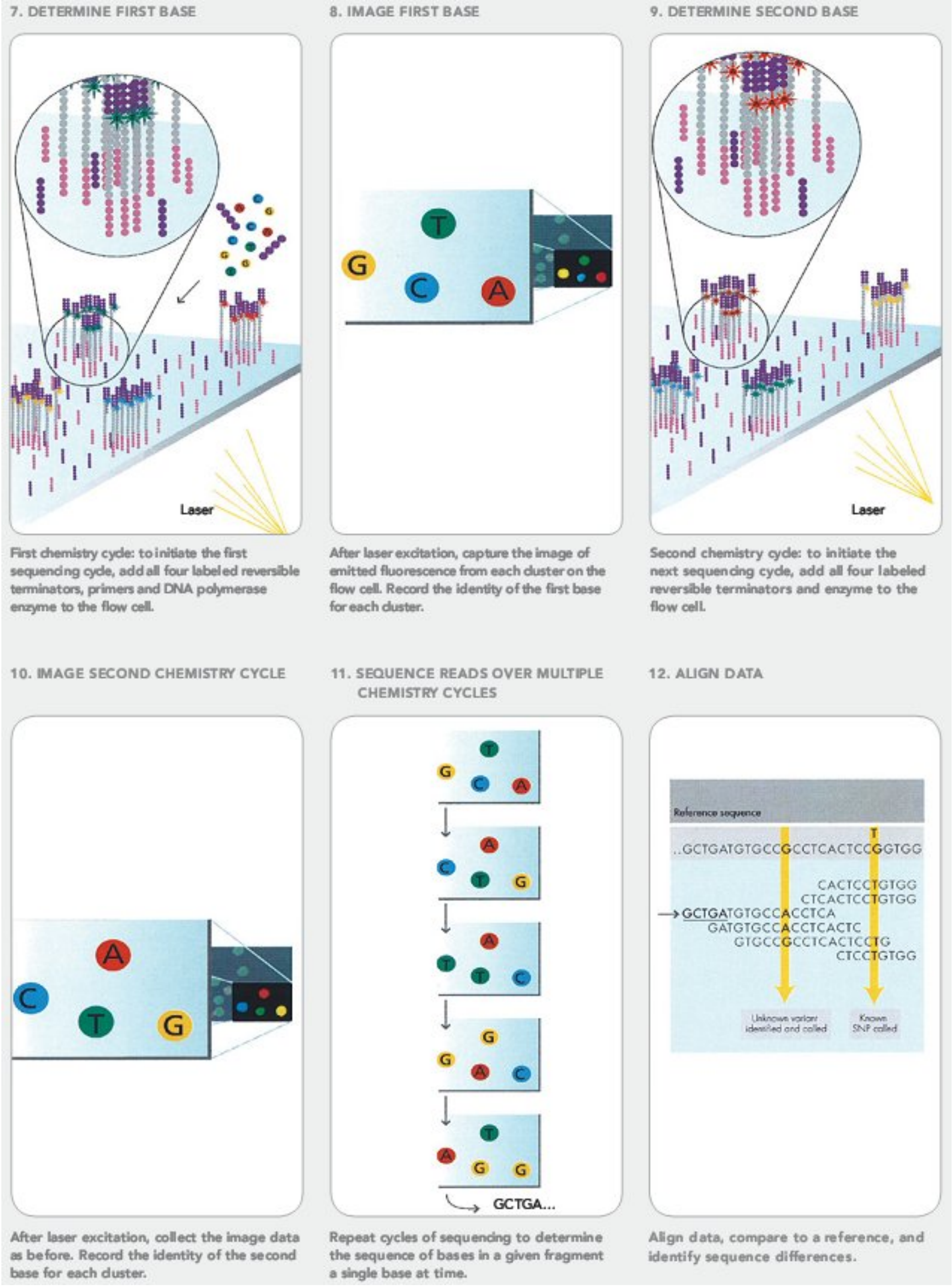

7. DETERMINE FIRST BASE
8. IMAGE FIRST BASE
9. DETERMINE SECOND BASE
Laser
G
T
C
A
Laser
First chemistry cycle: to initiate the first sequencing cycle, add all four labeled reversible terminators, primers and DNA polymerase enzyme to the flow cell.
After laser excitation, capture the image of emitted fluorescence from each cluster on the flow cell. Record the identity of the first base for each cluster.
Second chemistry cycle: to initiate the next sequencing cycle, add all four labeled reversible terminators and enzyme to the flow cell.
10. IMAGE SECOND CHEMISTRY CYCLE
11. SEQUENCE READS OVER MULTIPLE CHEMISTRY CYCLES
12. ALIGN DATA
A
C
T
G
GCTGA...
Reference sequence
T
..GCTGATGTGCCGCCTCACTCCGGTGG
CACTCCTGTGG
CTCACTCCTGTGG
GCTGATGTGCCACCTCA
GATGTGCCACCTCACTC
GTGCCGCCTCACTCCTG
CTCCTGTGG
Unknown variant identified and called
Known SNP called
After laser excitation, collect the image data as before. Record the identity of the second base for each cluster.
Repeat cycles of sequencing to determine the sequence of bases in a given fragment a single base at time.
Align data, compare to a reference, and identify sequence differences.

Figure 61 Summary of Illumina next generation sequencing (http://seqanswers.com/forums/showthread.php?t=21)

8.2.5 Data processing

Data were processed in conjunction with Dr Nick Kennedy with additional input from Dr Al Ivens.

Fastq files consist four lines that display the following information:

1) *Sample ID* (usually starts with @, then instrument name(), flowcell lane()title number within flowcell lane(), x co-ordinate within cluster, y co-ordinate within cluster, index number for barcode of multiplex sample(), and number within pair(not applicable in this case, only when paired end reads).

2) *Raw sequence reads*

3) '+' (can contain another sample identifier)

4) *Quality value* based on the ASCII scale of increasing quality form left to right (corresponding to 33 to 126)

```
!"#$%&'()*+,./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrst
uvwxyz{|}~
```

This can be converted into a phred score (0-60) a measure of the probability of a correct base call (Phred + 33). The phred score is calculated from the standard Sanger variant (Q_{sanger} = -10* $\log_{10}p$, where p is the probability of a base call being correct).

8.2.5.1 Quality assessment

The sequencing data was quality assessed using the Fastqc program. Low quality reads were filtered using CutAdapt in the following step.

8.2.5.2 Trim adaptors/adapters

During post-sequencing processing the 5' adaptor was removed, and the raw reads are of 50 base length. Given microRNAs are between 18-24 nucleotides in length, this invariably means that the 3' adaptor needs to be trimmed. Various packages can be used for this process, and indeed this step is incorporated into many of the 'complete' microRNA processing pipelines detailed below. To remove adaptors the program Cutadapt was used.[431] A permitted error rate was tested (0.1, 0.15, 0.2, 0.3), and the 0.1 (=10%) permitted error rate was optimal (this is the default for the package). This package also allows removal of sequences longer or shorter than a specified length, which in the case of attempting to identify mature microRNAs, was set as 17-33nt. CutAdapt also allows trimming on the basis of quality. Low quality ends are trimmed prior to adapter removal.

The adaptor sequence is:

AGATCGGAAGAGCACACGTCTGAACTCCAGTCTACACTCTTTCCCTACACGACGCTCTTCCGAT

(Reference https://www.neb.com/~/media/Catalog/All-Products/33E45F5CD69B497E92C1073F5D551DC7/Datacards%20or%20Manuals/manualE7335.pdf; see http://seqanswers.com/forums/showthread.php?t=40289)

8.2.5.3 Align to human genome and mapping to miRBAse

The sRNA bench program (formerly miRAnalyzer) was used to align adaptor trimmed reads to the human genome (1000 genomes, version 37) and map reads to miRbase.[432] A bowtie index of the human genome was firstly created and placed in a newly created seqOBJ folder. The current miRBase list of miRNA sequences was downloaded from miRBase and placed in the libs folder of sRNAbench directory. Genome mapping mode of sRNA bench was used, which additionally provides files not included when the library mode is used (maps only to reference sequences [i.e. miRBase], not genome). These include reads mapped to the genome, but not assigned to a reference sequence, the read lengths, mapped antisense read. The genome mapping mode additionally provides the chromosomal location of mapped mature miRNAs (may have >1 chromosomal location). The mature_sense.grouped.txt files were then exported into R for further analysis (using read.delim).

```
sRNAbench microRNA=hsa species=human_g1k_v37 input=$inFile
output=$outSubdir > '$logSRB'/$inBn.log
```

The following loop was used to process all samples (script written by Dr Nick Kennedy)

```
./doSRNAbench.sh

#!/bin/bash

inDir=./trimmedCA

logSRB=./logSRB

outDir=./outSRB

if [ ! -d $logSRB ]; then mkdir $logSRB; fi

if [ ! -d $outDir ]; then mkdir $outDir; fi

ls $inDir/*.fastq | xargs -i --max-procs 8 bash -c
'inFile={};inBn=`basename $inFile`;inPath=`dirname $inFile`;echo
Processing $inFile with sRNAbench;'\

'outSubdir='$outDir'/`echo $inBn | sed s/\\.fastq$//`;'\

'if [ ! -d $outSubdir ]; then mkdir $outSubdir; fi;'\

'sRNAbench microRNA=hsa species=human_g1k_v37 input=$inFile
output=$outSubdir > '$logSRB'/$inBn.log'
```

Several software packages were tested to mapping and alignment of miRNA sequences: miRExpress,[433] miRDeep*,[434] and The UEA Small RNA workbench.[435]

8.2.5.4 Normalisation of data

It is not possible to compare the raw counts of microRNA sequences (mapped to mirBase) due to the variance of total number of reads obtained from each sample. Raw counts from all samples were normalized to a common scale to allow comparison. There are several normalization techniques, and there is a lack of consensus to the best type of normalization.[436]

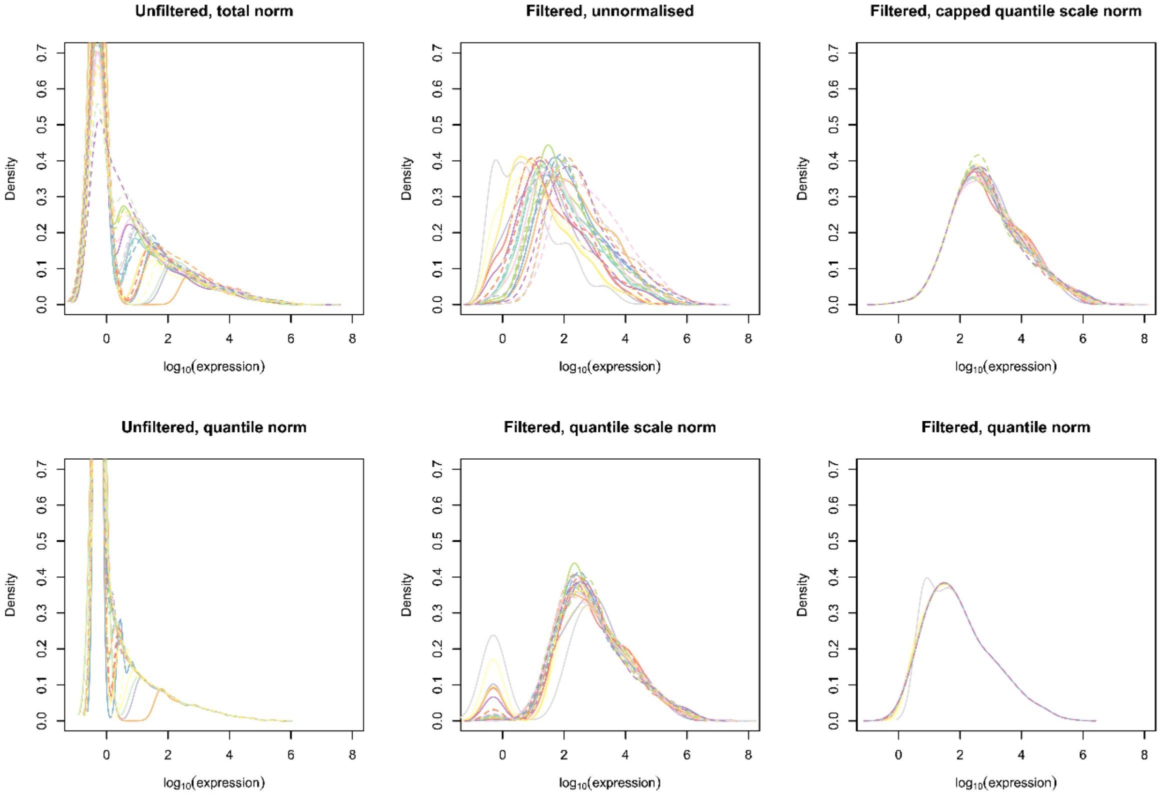
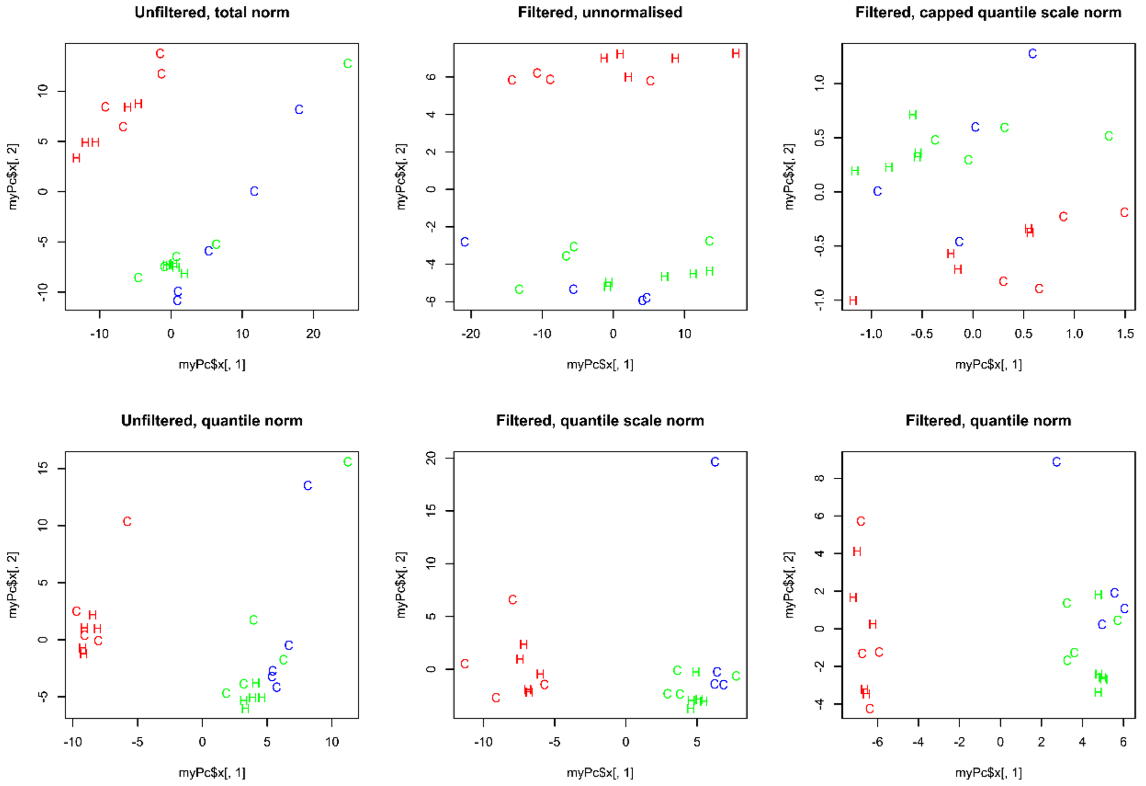


Figure 62 – Density plots of log10 expression of small RNAs using different normalisation techniques


Unfiltered, total norm

Filtered, unnormalised

Filtered, capped quantile scale norm

Unfiltered, quantile norm

Filtered, quantile scale norm

Filtered, quantile norm

Figure 63 – Multidimensional scaling plots using different normalisation techniques. H= healthy control, C = Crohn's disease. Red= Monocytes (CD14+), Green = CD4+ lymphocytes, Blue = CD8+ lymphocytes. See text for explanations for each normalisation method.

8.2.5.4.1 Normalize to total mapped reads

It is possible to normalize data to the total number reads mapping to miRBase. This is a form of simple scaling where RNA expression accounts are divided by the total number of mapped reads and multiplied by the mean total account across all samples.[437](Figure 63 Figure unfiltered, total norm)

8.2.5.4.2 Quantile normalization

Quantile normalization is performed by ordering the expression value in each sample, and then taking an average (usually mean) across all microRNAs (or probes etc.). The highest expression level becomes a mean of all the highest expression values; the second highest value becomes the mean of all the second highest values etc. The new values are then substituted back in for each sample according to the rank within that sample. The new normalized samples therefore have the same distribution and are more easily compared.[290]

Quantile normalization is used extensively when analysing micro-array datasets. This method is employed when expected changes are likely to be due to technical rather than biological variation. Quantile normalization may not be entirely appropriate for the reason that it is based on the strong assumption that all samples must have identical read count distributions,[437] and may be over harsh(intrusive) on the data.

The following two methods of normalisation adapt this method have been suggested by Rahmann et al.[436] Both methods discard microRNAs which do not reach expression levels of 5 counts in at least half of the samples. The authors state that below this level of expression, no meaningful statistical interpretation is possible and therefore these microRNAs are discarded from further analysis. Furthermore, samples with low overall read counts are excluded (<500,000). This constitutes the 'Filtered' samples in Figure 63.

8.2.5.4.3 Quantile based scaled normalization

This method uses one experiment/sample as the reference sample. The sample should be one where there are high total read counts. The authors of the method suggest that the sample with the highest 0.75th/third quartile should be used. Other samples are then scaled to the reference sample, using the median of the ratios between expression levels between the two samples. (Figure 63, Filter, quantile scale norm)

8.2.5.4.4 Capped quantile normalization

This method employs standard quantile normalization, but does not include 'extreme' values in this normalization. 'The 'extreme' values are normalized using a scaling factor of the median of the ratios of the highly expressed (but not extreme) quantiles. (Figure 63, Filtered, capped quantile scale norm)

8.2.5.5 Differential expression

Differential expression was calculated using linear modelling using R package limma. P values are corrected for multiple testing using a false discovery rate (FDR) according to Benjamini-Hochberg and Holm.[301]

8.2.5.6 Identifying downstream miRNA targets and pathways

Putative messenger RNA (mRNA) targets of differentially expressed miRNAs were identified using TargetScan (v7.0; targetscan.org).[438] Pathway analysis based on differentially expressed miRNAs was performed using DIANA-miRPath (DIANA-microT-4.0 beta version).[439]

8.3 Results

8.3.1 Patient demographics

Patient demographics are outlined in Table 37.

Table 37 –Patient demographics for patients in small RNA sequencing experiment

	Crohn's disease (n=8)	**Healthy controls (n=8)**	p value
Age (median, IQR)	27 (20-29.5)	30 (28.75-30.75)	0.15 (Wilcox.test)
Sex M:F (%males)	5:3 (62.5)	5:3 (62.5)	1 (fishers exact)
Smoking status Current:Ex:Never	4:2:2	0:2:6	Fishers exact for current smoking p=0.08
CRP (Median, IQR)	72 (55-93.5)	0 (0-1)	P=0.0007
Fecal Calprotectin (Median, IQR)	990 (480-1190)*	Available on 5/8 patients and no controls	
Disease location L1:L3	2:6		
Disease behavior B1:B2:B3	4:1:3		
Required azathioprine therapy Y:N	5:3		
Required Biologic/anti-TNFα therapy Y:N	3:5		
Required surgery Y:N	2:6		

Individual sample details, including RNA concentration, quantity and integrity are displayed in Table 86. In total four out of 48 samples were excluded from sequencing for the following reasons: two failed electrophrenogram quality control (RNA integrity number not calculable);

one sample looked abnormal on PAGE gel assessment of PCR product purity; and one sample was lost.

8.3.2 Quality control

Sample fastqc plots are displayed. Pools 5 and 6 (low PCR band intensity) demonstrated uniformly higher quality than the high PCR band intensity pools (Pools 1 and 2, Figure 64).

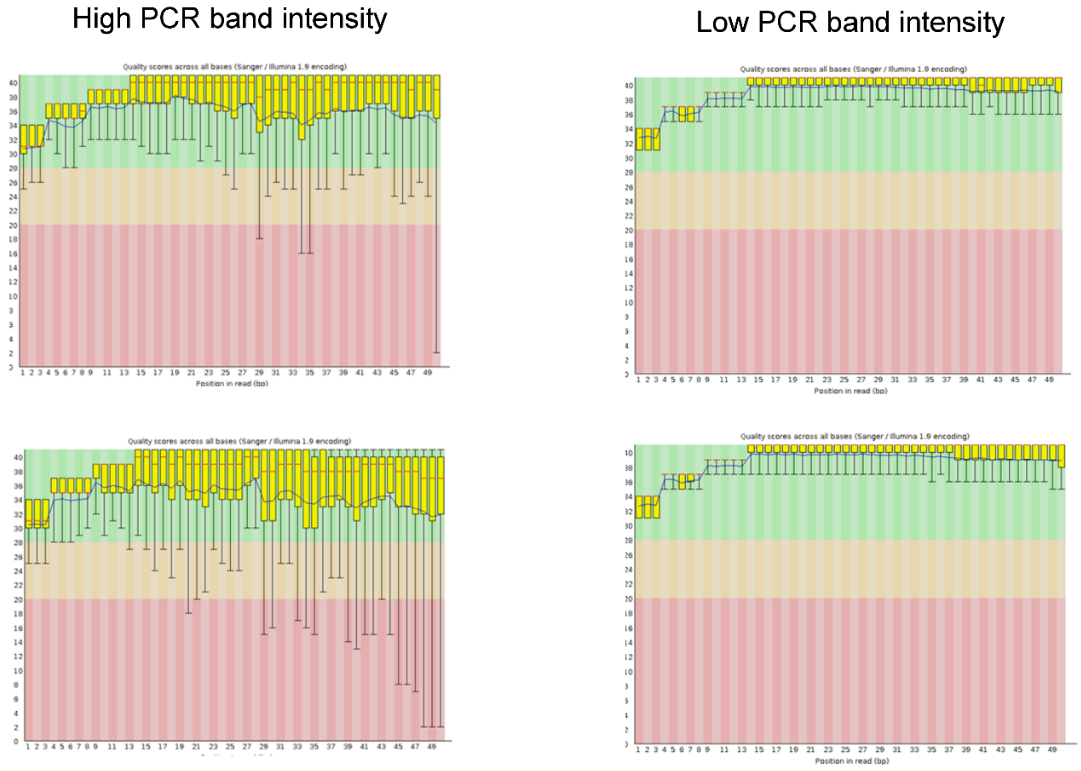


Figure 64 - Fastqc images of High PCR and Low PCR band intensity

There was a median of 2.4 million reads per sample (range 132,800-12.8 million reads per sample). The number of reads per cell-type and by case/control status is outlined in Table 38. Low quality reads were trimmed during the adaptor trimming stage of data processing. Samples with less than 100,000 reads were discarded (1 sample) and miRNAs with less than 5 reads per sample were filtered. There was a total of 389 miRNAs included in the final analyses. The proportions of each miRNA in cell type are displayed in Figure 65. A large proportion of the reads were made up of miR-21-5p reads, most notably for CD14 monocytes. Based on density (Figure 62) and MDS (Figure 63) plots, **Filtered quantile normalisation** was used as the normalisation method.

Table 38 - Total median read counts (range) according to case status and cell type

Median read count(range)	CD14	CD4	CD8
CD	2,394,136 (575,600-5,620,000)	771,617 (200,400-7,132,000)	2,351,457 (132,800-4,145,000)
Control	3,364,827 (2,222,000-12,790,000)	3,552,862 (1,128,000-11,340,000)	1,426,372 (206,800-4,649,000)

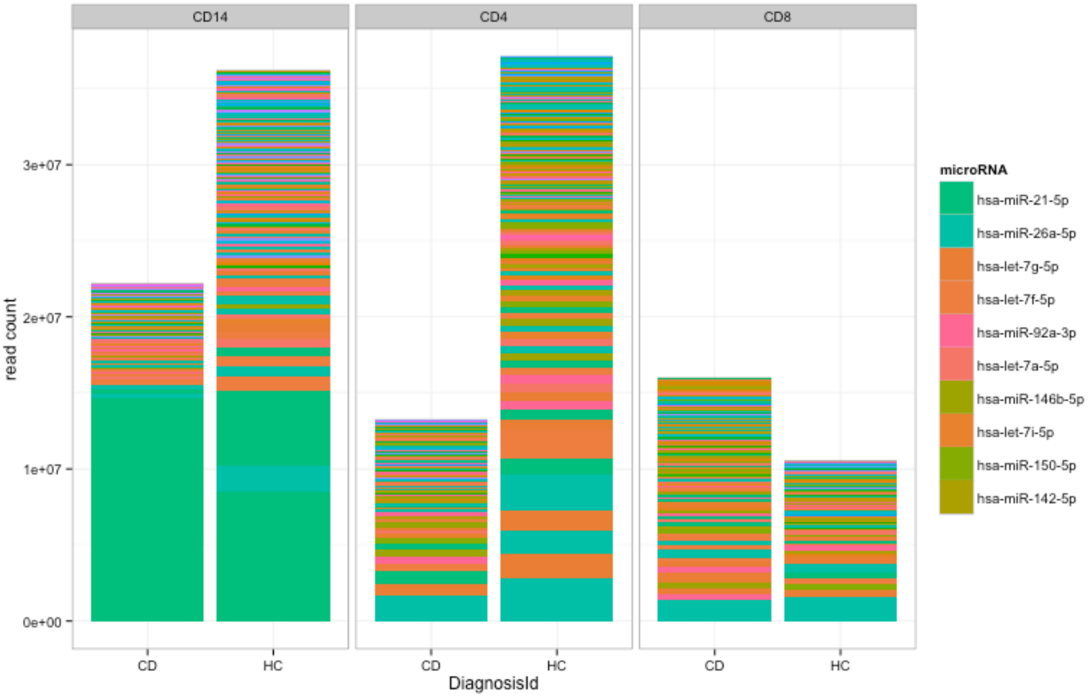


Figure 65 - Proportions of miRNAs (uncorrected) according to case status and cell type

8.3.3 Differentially expressed microRNAs in different lymphocyte in CD versus controls

8.3.3.1 CD4 lymphocytes

Following correction for multiple testing (FDR) there was one miRNA demonstrating increased expression in CD compared with control, miR-503-5p (log fold change =0.7, FDR adjusted p = 9.1 × 10^{-5}). This effect disappeared when other normalisation methods were applied to the data (Table 40). Another miR, miR-223-5p was also almost statistically significant (FDR p=0.06), and was statistically significant when different normalisation methods were used (Table 40). These miRNAs also demonstrated a trend towards increased expression in CD8 cells but did not reach statistical significance (

Figure 66). The top 10 differentially expressed miRNAs in CD4+ lymphocytes are detailed in Table 39.

Table 39 - Top list of differentially expressed miRNAs in CD4 lymphocytes in CD cases and controls. Filtered scaled quantile normalised data shown.

	logFC	AveExpr	t	P.Value	FDR adj.P.Val	B
hsa-miR-503-5p	0.69	1.97	5.23	9.13E-05	0.04	1.56
hsa-miR-223-5p	0.50	3.63	4.63	0.0003	0.06	0.51
hsa-miR-542-3p	0.48	2.09	4.05	0.001	0.11	-0.55
hsa-miR-574-3p	0.63	2.09	3.99	0.001	0.11	-0.66
hsa-miR-3614-5p	0.59	2.07	3.61	0.002	0.19	-1.37
hsa-miR-182-5p	1.13	2.12	3.42	0.004	0.20	-1.72
hsa-miR-455-5p	-0.95	1.97	-3.42	0.004	0.20	-1.73
hsa-miR-3909	0.36	2.77	3.27	0.005	0.24	-2.02
hsa-miR-26b-5p	-0.28	5.19	-3.18	0.006	0.26	-2.18
hsa-miR-223-3p	0.57	4.11	3.07	0.008	0.27	-2.38

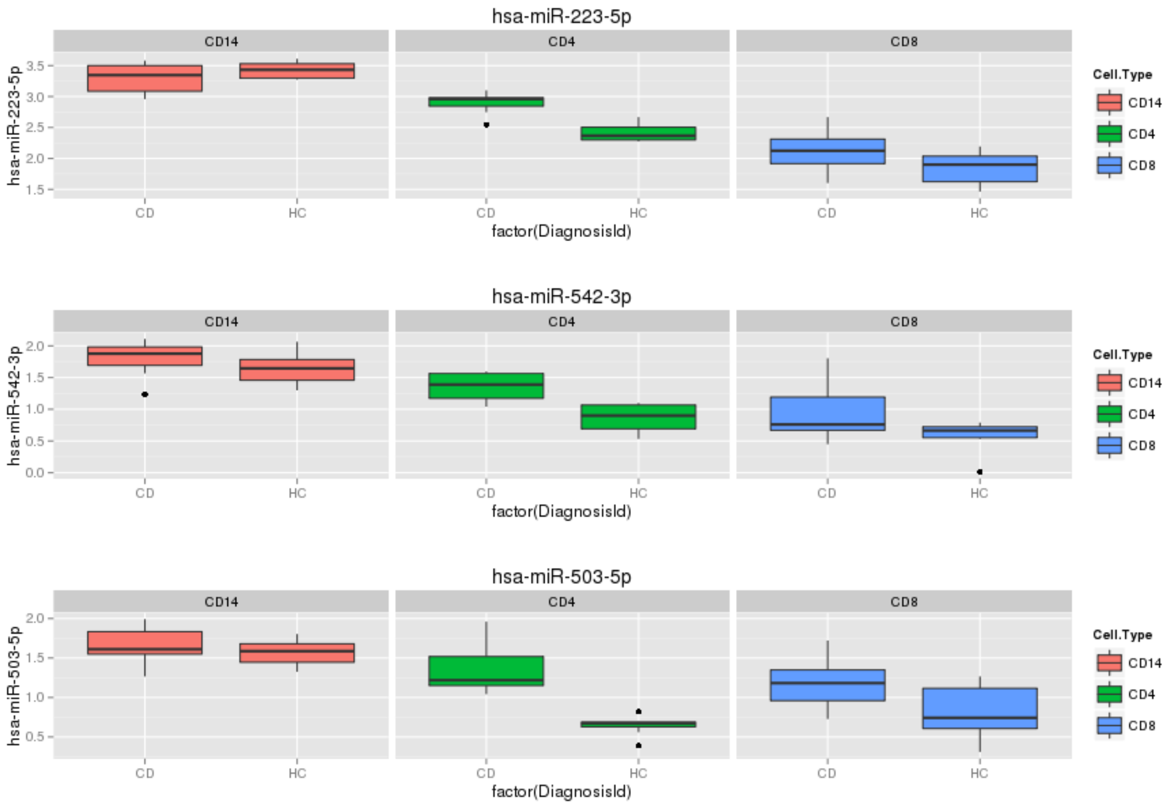


Figure 66 - Expression of miR-223, miR-542 and miR-503 according to case status and cell type. (Filtered scaled quantile normalisation)

		Filtered, scaled quantile	**Filtered, quantile**	**Filtered, scaled capped quantile**	**Filtered normalised to total**	**Unfiltered normalised to total**	**Filtered unnormalised**
miR-503-5p	*Log FC*	0.69	0.16	0.03	*0.5*	*0.52*	*0.28*
	FDR adj p value	0.04	0.002	*1*	*0.8*	*0.7*	*0.5*
miR-223-5p	*Log FC*	0.50	0.06	-0.07	*0.49*	*0.49*	*0.08*
	FDR adj p value	0.06	0.04	*1*	*0.8*	*0.7*	*0.8*

Table 40 – Differences in log fold change in expression and FDR adjusted p values for top two differentially expressed miRs in CD4+ using different normalisation methods.

8.3.3.2 CD8 Lymphocytes

There were no differentially expressed miRNAs in CD8 lymphocytes following correction for multiple testing (Table 41).

	logFC	**AveExpr**	**t**	**P.Value**	**FDR adj.P.Val**	**B**
hsa-miR-1237-3p	1.29	0.92	3.51	0.003	0.7	-3.95
hsa-miR-30e-5p	-0.34	4.80	-3.00	0.009	0.7	-4.08
hsa-miR-6726-3p	1.05	1.68	2.85	0.013	0.7	-4.12
hsa-miR-425-3p	-0.39	3.18	-2.84	0.013	0.7	-4.12
hsa-miR-26a-5p	-0.32	6.40	-2.80	0.014	0.7	-4.13
hsa-miR-26b-5p	-0.34	5.26	-2.80	0.014	0.7	-4.13
hsa-miR-374a-5p	-0.34	2.99	-2.63	0.020	0.7	-4.18
hsa-miR-186-5p	-0.38	4.56	-2.57	0.022	0.7	-4.20
hsa-miR-505-3p	-0.32	2.78	-2.55	0.023	0.7	-4.20
hsa-miR-103a-3p	-0.28	4.96	-2.52	0.024	0.7	-4.21

Table 41 - Top list of differentially expressed miRNAs in CD8+ lymphocytes in CD cases and controls. Filtered scaled quantile normalised data shown.

8.3.3.3 CD14 Monocytes

There were no differentially expressed miRNAs in CD14 monocytes following correction for multiple testing (Table 42).

	logFC	AveExpr	t	P.Value	FDR adj.P.Val	B
hsa-miR-3913-5p	-1.06	1.27	-3.71	0.002	0.67	-3.32
hsa-miR-3176	0.39	1.73	3.05	0.01	0.79	-3.65
hsa-miR-509-3-5p	-0.71	0.59	-2.90	0.01	0.79	-3.73
hsa-miR-21-3p	0.34	3.77	2.78	0.01	0.79	-3.79
hsa-miR-424-3p	0.32	3.53	2.78	0.01	0.79	-3.79
hsa-miR-2277-5p	0.42	2.00	2.71	0.01	0.79	-3.83
hsa-miR-100-5p	-0.45	3.64	-2.68	0.02	0.79	-3.85
hsa-miR-339-3p	-0.22	3.14	-2.60	0.02	0.79	-3.89
hsa-miR-6513-3p	-0.73	1.40	-2.47	0.02	0.79	-3.96
hsa-miR-152-3p	-0.30	3.21	-2.39	0.03	0.79	-4.00

Table 42 - Top list of differentially expressed miRNAs in CD14+ monocytes in CD cases and controls. Filtered scaled quantile normalised data shown.

8.3.4 Unsupervised Hierarchical Clustering

Hierarchical clustering was performed on the whole dataset and is displayed in Figure 67 and Figure 68. Figure 67 demonstrates one clear outlier that was excluded from further analysis. The monocytes appeared to cluster together, however there was less defined clustering between CD4 and CD8 lymphocytes. Figure 68 demonstrates that samples did not tend to cluster according to the individual from which the cells were obtained.

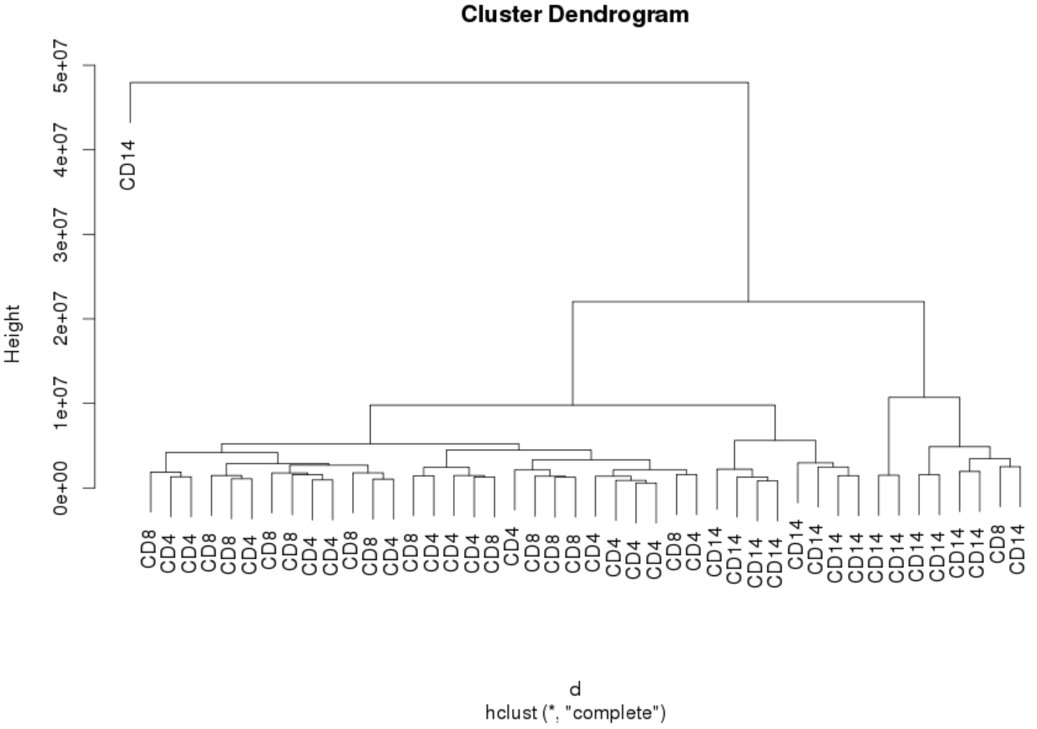


Figure 67 - Cluster dendrogram of small RNA sequencing samples according to cell type

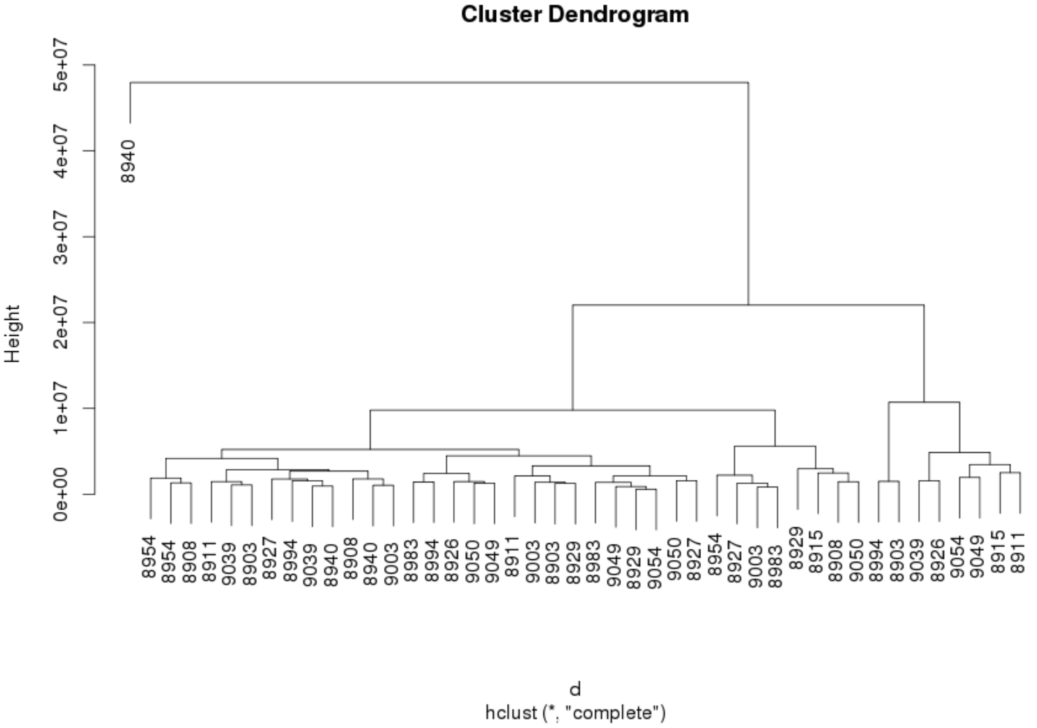


Figure 68 - Cluster dendrogram of small RNA sequencing samples according to sample number

8.3.5 Principal component analysis

Principal component analysis demonstrated accurate clustering of samples according to cell type (Figure 69).

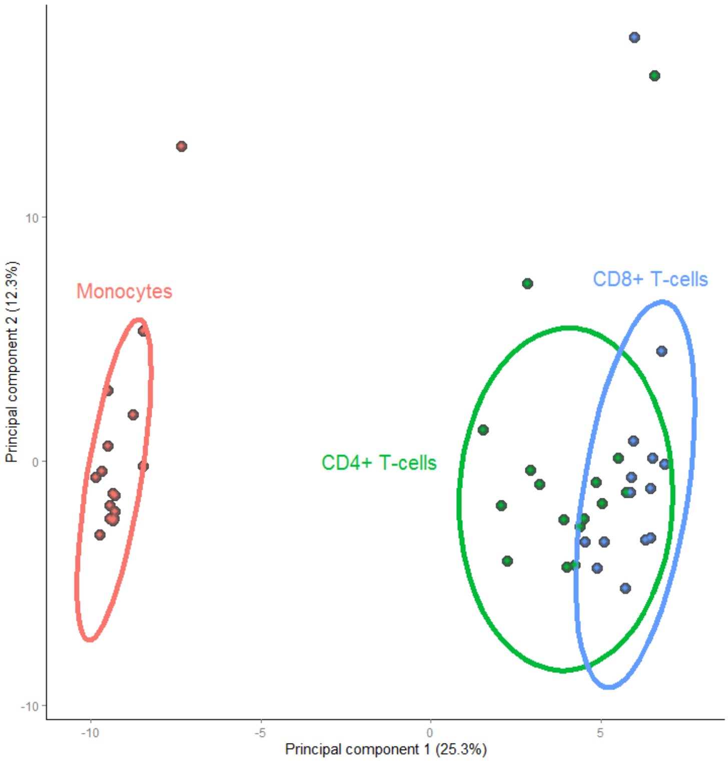


Figure 69 - Principal component analysis based on all filtered, scaled quantile normalised data according to cell type

8.3.6 Target scan of top differentially expressed miR in Crohn's disease

Putative messenger RNA (mRNA) targets of miR-503-5p were identified using TargetScan (v7.0; targetscan.org, Table 43).[438] Pathway analysis based on mRNA targets of miR-503-5p was performed using DIANA-miRPath (DIANA-microT-4.0 beta version)[439]. There were 61 pathways identified and are presented in Table 87.

Target gene	Gene name	3P-seq tags + 5	Cumulative weighted context++ score	Total context++ score	Aggregate PCT
ARL2	ADP-ribosylation factor-like 2	7073	-1.03	-1.03	0.88
CNTNAP1	contactin associated protein 1	821	-0.89	-0.89	0.7
LCE6A	late cornified envelope 6A	5	-0.88	-0.88	< 0.1
MAPK8IP2	mitogen-activated protein kinase 8 interacting protein 2	88	-0.85	-1.1	0.16
TMEM74B	transmembrane protein 74B	33	-0.75	-0.91	0.81
FAM122A	family with sequence similarity 122A	1277	-0.74	-0.78	0.18
CCNE1	cyclin E1	954	-0.72	-0.72	0.97
RP11-144F15.1	Uncharacterized protein	5	-0.72	-0.72	< 0.1
CYB561	cytochrome b561	10	-0.7	-0.7	< 0.1
LURAP1L	leucine rich adaptor protein 1-like	266	-0.69	-0.73	< 0.1
CTD-2207O23.12	Uncharacterized protein	32	-0.68	-0.85	ORF
INSR	insulin receptor	577	-0.68	-0.73	< 0.1
CCND2	cyclin D2	66	-0.67	-0.81	0.98
DGCR2	DiGeorge syndrome critical region gene 2	1754	-0.64	-0.65	< 0.1
APLN	apelin	48	-0.63	-0.63	0.84
NDUFA4	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 4, 9kDa	143	-0.62	-0.78	< 0.1
AL117190.3	Oesophagus cancer-related gene-2 interaction susceptibility protein; Uncharacterized protein	30	-0.62	-0.63	0.19
CMC4	C-x(9)-C motif containing 4 homolog (S. cerevisiae)	87	-0.6	-0.6	0.36
SLC46A1	solute carrier family 46 (folate transporter), member 1	580	-0.6	-0.68	< 0.1
CCND1	cyclin D1	4687	-0.59	-0.63	0.82
TNFSF13B	tumour necrosis factor (ligand) superfamily, member 13b	5	-0.58	-0.58	0.87

Table 43 – messenger RNA targets of miR-503-5p as identified by TargetScan[438]

8.4 Discussion

This chapter provides a detailed characterisation of miRNAs within circulating leucocytes in patients newly diagnosed with Crohn's disease. As has been demonstrated using DNA methylation and gene expression data elsewhere in this thesis, based on miRNA sequencing data alone, samples cluster accurately according to cell type. When comparing CD cases and controls, miR-503-5p was the only differentially expressed miR in CD4+ lymphocytes following correction for multiple testing. Interestingly, the third most differentially expressed miRNA in CD4+ cells was miR-542, and although not statistically significant (uncorrected p=0.001, FDR corrected p=0.1) arises from the same cluster (<10kb) as miR-503. MiR-503 downregulation has been described in several cancers and may act as a tumour suppressor.[440–442] MiR-503 has not previously been investigated in the context of IBD.[314] Overexpression of the second ranking miR-223 (FDR corrected p=0.06) has previously been demonstrated in colonic biopsies in active UC and inactive CD[196] and in biopsies in active small bowel CD.[197] In Rheumatoid arthritis, miR-223 was found to be overexpressed in CD4+ T-cells compared with controls.[443] Notably miR-223 is known to be associated with granulocyte differentiation and activation.[444,445] A potential explanation for increased expression of miR-223 may be disproportional contamination of CD4+ samples with granulocytes in CD cases compared with controls.

Most studies to date have used a microarray based platforms in the context of IBD miRNA research.[144,314] The advantages of next generation sequencing have been discussed above. Whilst an increasing number of papers utilising small RNA sequencing have been published, there is not yet a consensus on the optimal pipeline for data normalisation and analysis.[436] Previous studies in human disease have mostly utilised the approach of normalising to total mapped reads which is likely to lead to bias as a result of changes in highly expressed miRNAs significantly affecting lesser expressed miRNAs. Various normalisation methods were explored in this study prior on deciding the best method for processing data. On the basis of multidimensional scaling and density plots, filtered, scaled quantile normalisation was performed in this study and has previously been advocated by other authors.[436] MiR-503 was a differentially expressed using filtered, scaled quantile normalised data and the same finding was also seen when data was normalised using filtered quantile normalisation, but not other techniques (Table 40). The dependence on the normalisation method to identify this differentially expressed miR reduces confidence in this finding raises the possibility of a type I error. Similarly the small sample size may lead to a type II error. As such extensive downstream validation and functional work is not warranted based on this finding.

Whilst reservations on the validity of the differentially expressed miRNA (503-5p) have been discussed above, putative downstream targets of this miRNA were investigated in order to explore the process. TargetScan was used to identify messenger RNA targets of the only differentially expressed miRNA (miR-503-5p). An alternative method (DIANA-miRPath) based on gene network/pathway analysis was also performed for miR-503. Most of the top ranked KEGG pathways were cancer related (Prostate cancer, Melanoma, Glioma, Colorectal cancer) but other pertinent pathways were included (mTOR signalling, focal adhesion, TGF β signalling). Whilst such methods has been widely used, there are inherent biases that will lead to over-identification of related biological processes.[446] Much of the current literature relating to miRNA gene targets has been conducted in the context of cancer, and therefore results are likely to be biased toward cancer-associated pathways, as was experienced in the present study.[446] Therefore findings of associated target genes or related pathways are likely to be non-specific and should be treated with caution.

The strength of this study is the use of a purified cell type to demonstrate differences in miRNA between cases and controls. Many of the studies to date comparing miRNA expression on a case-control basis have used whole tissue, leading to significant uncertainty on the cell type of origin of differentially expressed miRNAs.[144,314] Another positive aspect of this study design is the use of age and sex matched cases and controls and the inclusion of CD patients with ileal or ileal-colonic disease. Given the significant heterogeneity in the disease itself, this leads to increased uncertainly on the reliability of small datasets such as this. The main limitations of the study are discussed above; namely the small overall sample size and lack of consensus on data normalisation method.

This work has provided a useful introduction into NGS for microRNA research, however is limited by a small sample size. The wet-lab and bioinformatic techniques employed will inform further study in this area. In the future it may be possible to sequence small RNAs from larger numbers of separated cells form newly diagnosed IBD patients to more definitely address the question.

Chapter 9. Conclusions, implications and future research

9.1 Conclusions

9.1.1 Results of study

This study has demonstrated site-specific methylation changes in IBD compared with controls that were strongly significant following stringent correction for multiple testing. In lieu of a consensus on an accepted significance threshold for EWAS, this study has used a correction method traditionally used in GWAS. Using this conservative threshold 439 significant DMPs and 5 DMRs have been identified. Whereas many early EWAS results have not been replicated, the highly replicable nature of DMPs and DMRs in independent cohorts in this study increases the confidence in these findings. A comprehensive approach was employed to study genome-wide DNA methylation, allied with genomic and transcriptomic data in matched individuals allowing truly integrative analysis.

9.1.2 Literature in IBD and other complex immune diseases

The rationale for epigenetic research in IBD is compelling: IBD has a significant genetic contribution and well known environmental risk factors. The inclusion of newly diagnosed patients minimises the potentially confounding effect of potent immunomodulation drugs, whilst still allowing exploration of the relationship between inflammation and DNA methylation. This is the largest study of DNA methylation in IBD to date. Several other smaller studies have investigated DNA methylation in IBD including the study into paediatric CD performed in Edinburgh. McDermott et al [307] have assayed DNA methylation in PBMCs of IBD patients with established disease with varying degrees of disease activity and immunosuppressive duration. The present data corroborate that patients with IBD do not have systematic changes in methylation unlike the global hypomethylation seen in systemic lupus erythematosus.[447]

9..1.3 Cell types & Gene expression

The impact of cellular heterogeneity on DNA methylation data is a commonly cited limitation of EWAS studies conducted using whole tissue such as blood.[346,347] Statistical algorithms that provide estimated cell proportions and allow adjustment for cellular heterogeneity are now widely performed throughout the EWAS literature.[205,302] There are comparatively fewer epigenetic studies with separated cell data, particularly disease-relevant cells and this is a significant strength of the present study.[348,447–449] Detailed characterisation of matched genetic, DNA methylation and expression data in separated leukocytes has highlighted several cell-specific findings. For example, the top DMP in whole blood, RPS6KA2, was differentially methylated in CD14+ monocytes but not CD4+ or CD8+ lymphocytes, potentially unmasking the cell type of origin of methylation signals seen in whole tissue. Cell-specificity may become even more relevant when analysing the relationship between methylation and gene expression. A major finding of this study is IBD-associated hypermethylation of the *TXK* TSS occurring specifically within CD8+ cells, with an appropriate negative correlation with decreased gene expression in CD8+ T-cells of the same individual. Expression of TXK, a member of the Tec family of non-receptor tyrosine kinases, in Th1 T-cells is obligatory for the production of interferon gamma.[450] CD8+ T-Cells have an established role in IBD pathogenesis[260,451–453] with recent data suggesting that CD8+ T-Cell exhaustion may be a critical prognostic factor in immune-mediated diseases.[454]

9.1.4 Origin of epigenetic signals

Whilst the present study design does not allow functional interrogation of the origin of the DNA methylation profile seen here in IBD, it is interesting to speculate on the origin of such signals. The strong correlation between clinical inflammatory markers and the top DMP/Rs perhaps indicate that the observed methylation changes are a consequence of inflammation. It is notable that these signals endure in the replication cohort (Chapter 4) that consists of patients with established disease with varying levels of inflammation following treatment. Contrary to this hypothesis is the strong association between germ line variation and the variance of methylation of two of the five DMRs (VMP1/miR-21 and ITGB2). The VMP1/miR-21 locus was the major finding of the paediatric CD study, with miR-21, a pro-inflammatory microRNA previously implicated in colitis being an obvious candidate for further investigation. The novel finding is the association of methylation in the VMP1/mcroRNA-21 region with two nearby SNPs acting as methylation quantitative trait loci (meQTLs). The two polymorphisms (rs10853015, rs8078424) acting as meQTLs are in linkage disequilibrium with a known IBD-susceptibility GWAS locus (rs1292053). DNA methylation may be a mechanism by which genetic variants outside of protein coding regions may contribute to the disease phenotype. In Rheumatoid arthritis, Liu et al used mediation analyses to demonstrate that methylation was the causal mechanism by which genotype conferred disease risk.[204] Most associations occurred in the major histocompatibility complex region, known to harbour many genetic variants with complex and extended linkage disequilibrium structures.[379]The present analysis established several of the tenants of causal inference but not independence of genotype and phenotype following adjustment for methylation, a finding reflected in the complex disease literature.[279] The strong association between top ranking IBD-associated DMP/R and nearby genetic variants nevertheless represents a major finding and goes some way to explain these significant site-specific methylation differences in IBD cases and controls. Further work in larger cohorts is required to disentangle this complex relationship between genetics, DNA methylation, inflammation and other environmental factors.

9.1.5 Translational potential

DNA methylation data offer immediate translational potential as biomarkers. The SEPT09 blood-based DNA methylation biomarker has been used in diagnosis and screening for colorectal cancer.[353,406,407] Adams et al previously demonstrated two methylation probes can accurately differentiate CD and controls and these have been prospectively validated using the present adult dataset.[284] The previous work has been expanded upon by using lasso,[416] an established machine learning technique that can help avoid over-fitting in large datasets where the number of variables vastly exceed the number of samples. The final 30-probe model is easily scalable into a high-throughput pyrosequencing panel and such a non-invasive peripheral blood biomarker could be used to stratify patients to further intrusive investigations such as colonoscopy. Existing clinically available biomarkers such as faecal calprotectin[230] already provide similar utility but are unable to distinguish the two forms of IBD. A different 19-probe methylation-based panel may confer an additional benefit in discriminating CD and UC, which can be critical for decision-making in terms of medical and surgical management. Currently there are no reliable prognostic biomarkers that can identify patients requiring early aggressive treatment from those who would experience a quiescent disease course who could safely avoid the potentially toxic side effects of potent immunosuppresants. There has been some anticipation that emerging '–omic' data may provide such a biomarker.[260] In the present work a DNA methylation signature that associates with high-, intermediate- and low-risk of specific deleterious outcomes has been described. Consensus clustering has recognised limitations[418] and it is noteworthy that the methylation subclasses are not independently predictive of outcome and are likely to be driven by underlying differences in cell count or other clinical parameters.

9.1.6 Conclusions and implications

This is the most detailed characterisation of the circulating IBD methylome to date. Highly statistically significant and replicable site-specific differences in DNA methylation have been demonstrated at sites pertinent to disease pathogenesis. DNA methylation may be a factor of underlying germ line variation and may represent a mechanism by which genetic polymorphisms contribute to disease variance. Cell sorting of disease-relevant immune cells has highlighted subtle cell-specific relationships between DNA methylation and gene expression. The immediate agenda for epigenetic research in IBD is discussed below.

9.2 Future research

9.2.1 DNA methylation

Whilst this is the largest DNA methylation dataset in whole blood in IBD to date, several questions remain unaddressed. Firstly, although the presented adult dataset strongly corroborates findings from the previous paediatric CD data,[284] all of the whole blood data in IBD to date has been generated in Edinburgh using Scottish samples. The presented epigenetic profile may be driven by regional genetic or environmental factors.[144] As such, it would be important to repeat these experiments in large cohorts from other geographical regions and other ancestries. Ongoing work as part of the IBD-CHARACTER project (www.ibd-character.eu) will profile whole blood DNA methylation in 400 newly diagnosed IBD patients from Northern Europe (Edinburgh, Oslo, Linköping, Orebro, Zaragoza). Furthermore, collaborations as part of the IBD-BIOM project (www.ibd-biom.eu) may allow replication of the present findings in patients from Italy and USA (including a large Ashkenazi Jewish population).

Another important question to address would be the disease specificity of the DNA methylation signals demonstrated in the present study in IBD. A commonly cited criticism for this kind of work is that the observed changes may be a consequence of inflammation and that findings are not-disease specific. A future study design should include samples from subjects with other inflammatory diseases such as rheumatoid arthritis. This is particularly pertinent when developing disease-specific biomarkers.

In order to further increase to power and sample size for epigenomic discovery, it would be possible to combine and meta-analyse multiple 450k methylation datasets. Meta-analysis has been a major success in the context of genomic study[88,97] however there are several differences with DNA methylation data that require further consideration. Firstly there are significant and well-recognised batch-effects associated with the 450K Illumina methylation microarray. Particular care must be taken when combining such datasets such that technical variation is adequately examined and adjusted for (if possible). Secondly the tissue from which the data have been generated must be considered when combining datasets. Whilst it may be possible to combine datasets generated using whole blood DNA, it would be difficult to combine data generated in PBMCs[168,307] or other cell types.[171,321] Lastly, any meta-analysis should be designed to ensure that the included patients possess similar clinical characteristics to avoid the introduction of further heterogeneity into the study. This is the main reason for not combining the adult and paediatric 450k datasets, despite both using whole blood DNA.

An interesting aspect of this thesis is the link between genetics and DNA methylation. In this study there was an association between methylation and genotype for some of the most differentially methylated regions. It was not possible however to demonstrate that DNA methylation mediates the genetic effect on phenotype (using the CIT test[373], see Chapter 5). The major factor limiting this work was the lack of power for the genome-wide association tests. Another method that it may be possible to use is Mendelian randmosiation[455] to link genetic and DNA methylation data. The major advantage of this technique is that paired data are not required. Therefore it may be possible to exploit the wealth of available IBD genetic data to address this question.

The main theme of the work presented in this thesis is a case control analysis of IBD patients and controls. DNA methylation data may also be used to address other important questions in IBD research. It would be interesting to attempt to identify an epigenetic signature associated with response/non-response to particular therapies, or as an expansion to the successful pharmacogenetics[122] studies aiming to identify polymorphisms associated with drug toxicity. Such studies may be possible given the large cohorts of IBD patients on anti-TNF alpha monoclonal antibody therapy being assembled (PANTS [https://www.pantsdb.co.uk/], biocycle [http://cordis.europa.eu/project/rcn/193180_en.html]). Another homogenous group that it may be interesting to study would be CD patients following ileocaecal resection. The disease is essentially reset at this point, and it may be possible to detect a methylation signature associated with disease recurrence. The large number of patients collected locally as part of the TOPPIC trial (randomised trial of thiopurine to prevent disease recurrence following ileocaecal resection) may provide such a cohort. The patients with severe refractory CD undergoing autologous stem cell transplant[456] would be another interesting group to study, however the very low white blood cell counts resulting from treatment may make this difficult to study using whole blood samples.

Lastly, it would be interesting to investigate to what extent the DNA methylation profile changes during an individual's disease and treatment course. During the IBD-BIOM project, serial samples were collected from the same patient allowing longitudinal study of the methylation profile. A small study of mucosal DNA methylation in children with UC (n=2) suggests that the methylome reverted back towards that of healthy controls following treatment.[321]

9.2.1.1 Hydroymethyl DNA methylation

During the course of this research, collaboration has been developed with Cambridge Epigenetix (http://www.cambridge-epigenetix.com/) for the investigation of hydroxymethyl-DNA methylation. The Illumina 450K platform used for throughout this thesis reports DNA methylation, but is unable to differentiate 5-methylcytosine (5mC) and 5-hydroxymethylcytosine (5hmC). The Cambridge Epigenetix group have developed (and patented) use of oxidative bisulphite sequencing (ox-BS seq) that allows differentiation between 5mC and 5hmC.[457] The technique employs a subtractive approach in which standard bisulphite converted DNA is assayed initially using the platform of choice (to detect 5mC and 5hmC, i.e. Illumina 450K/WGBS), and then the same sample is repeated following oxidative bisulphite conversion (to detect 5hmC). This workflow is associated with high costs as each sample must be converted and assayed twice. It remains unclear the overall value of perusing this line of research as 5hmC will make up a very small proportion of overall DNA methylation in IBD index tissues (blood, gut).[458] Most of the focus of 5hmC research has been focused on brain tissue where 5hmC marks are relatively more abundant.[457,458] 5hmC is being studied in other non-neurological diseases including cancer research.[459–462] Other modifications of cytosine include 5-formylcytosine and 5-carboxycytosine.[463] Cambridge Epigenetix are developing other methylation profiling platforms that may straddle the current coverage gap between the Illumina 450K array and WGBS. The new Illumina EPIC methylation array will similarly provide much greater coverage (850K) and is likely to supersede the 450K array in the near future.

9.2.2 Whole genome bisulphite sequencing

As part of the IBD-CHARACTER collaboration links have been made with the Centro Nacional de Análisis Genómico (http://www.cnag.cat/) for the development of whole genome bisulphite sequencing (WGBS). Whilst the Illumina 450K array used throughout this project provides good genome-wide coverage of relevant CpG sites, it is subject to the usual biases of microarray-based platforms and only provides information on pre-selected probes with non-customisable content. In particular probes are biased towards cancer and inflammatory pathways. The arrays are also associated with higher levels of technical/non-biological variation compared with other microarrays (genotyping, gene expression). Whole genome bisulphite sequencing (WGBS) allows assaying of CpGs throughout the genome. WGBS is becoming an established technology however at the time of writing remains expensive and data analysis is computationally demanding. Sequencing must be performed in high coverage/depth (>30x) in order to be confident about methylation differences, which results in the generation of large amounts of data. It is unclear at present the niche in EWAS research that WGBS will occupy at the present. It is too expensive to perform large comparative studies and use has largely been confined to smaller separated and single-cell studies. The Barcelona group have published several high impact papers using separated cells collected as part of the IHEC collaboration (http://ihec-epigenomes.org/) particularly in the field of haematological cancers.[464–466] An interesting experiment using this technology would be to characterise the differentially methylated positions and regions identified in this study in detail in a small number of patients. The WGBS data also provides the underlying sequence so the relationship between genotype and methylation could be more thoroughly explored.

9.2.3 DNA methylation in other tissues

There is clear rationale for study of DNA methylation in circulating leukocytes in IBD. IBD is an immune mediated disease with well-recognised extra-intestinal manifestations.[15] Much of the current armamentarium of IBD therapy targets the peripheral immune system[123] and there has been recent interest in autologous stem cell transplant as a treatment for severe,[350,456] refractory CD. Lastly, CD is known to recur following intestinal transplant indicating IBD is not exclusively a gut-based pathology.[351] Whilst blood does appear to be the disease-relevant tissue to study in IBD, and is appealing when developing biomarkers, it would nevertheless be important to study DNA methylation in the gut and its relationship with the blood DNA methylation profile. Several small studies have studied mucosal DNA methylation in a candidate gene fashion, and a small number using a microarray approach.[170,171] There have been no outstanding findings that have been replicable in more than one study.[144] The IBD-CHARACTER project will be using Illumina 450K microarrays to assay DNA methylation in gut tissue of newly-diagnosed IBD patients, and will be in a position to perform integrative analyses with matched whole blood DNA methylation, genetic, blood and mucosal transcriptome, microbiome and proteomic data. One of the critical issues when looking to study DNA methylation in gut biopsies is the lack of information on cell proportions/tissue composition in the healthy and disease state.[321,346] In blood, there are several reference datasets consisting of cell-sorted methylation data from most of the major blood cell populations.[303] This has facilitated the development of statistical algorithms to estimate cell proportions based on DNA methylation data and now adjustment for cell heterogeneity is accepted (and expected) in the EWAS literature.[205,302,361] A study conducted in Cambridge attempted to perform cell separation in gut tissue.[348] The main difficulty with cell separation in gut tissue is that epithelial cells are highly abundant and easy to isolate, but are more likely to be the bystander rather than the cell-type of interest in IBD. Intra-epithelial lymphocytes by contrast make up a very small proportion of cells and are very difficult to isolate using either flow cytometry or immunomagnetic separation. Whilst working with Miltenyi-Biotech on blood based cell separation equipment and reagents there is a possibility that a gentleMACs cell dissociator (order no. 130-093-235) can be loaned to the department. By using this equipment in conjunction with a collagenase enzyme, solid tissue (e.g. gut biopsy[467]) can be homogenised into a single-cell suspension solution that can then be used for downstream analysis (flow cytometry/immunomagnetic cell separation). The first step in helping to inform whole tissue studies would be to generate a reference dataset in healthy gut tissue in order to provide an estimate of the cell composition. This is likely to be different at different sites within the colon and small bowel, and infiltrating leukocytes in areas of inflammation will alter the cell make-up yet further. In summary, DNA methylation profiling in the gut requires a greater understanding of the constituent cell types. Generating reference data sets from separated intestinal cells will be difficult and likely to require significant funding. A large international consortium (perhaps following on from IHEC) may be best placed to carry out this type of work.

9.2.4 Other Future work (genetics, gene expression, MicroRNA and biomarkers)

In addition to the potential future work described above relating to DNA methylation, further work would be valuable in the other 'omics' disciplines described in this thesis.

9.2.4.1 Genetics

The genetic component of this study was underpowered to detect differences in germ line variation between IBD cases and controls. IBD genetic research now has progressed beyond single-centre research to International collaboratives (http://www.ibdgenetics.org/). As such it would be valuable to contribute the genetic data generated here for inclusion into work (i.e. meta-analyses) being performed internationally. The investigation into methylation quantitative trait loci has been the most value use of the genetic data to the present work. The aforementioned method of Mendelian Randomisation[455] may be the next step to take this work.

9.2.4.2 Gene expression

The gene expression part of this study was performed only in a subset of patients included in the 450k analysis and as such was potentially underpowered to detect differences in gene expression between cases and controls. This was in part limited by the fact that not all subjects included in the DNA methylation study had matched whole blood RNA samples. In the future it would be beneficial to perform a larger study including matched DNA methylation and gene expression data, in order to delineate the relationship between methylation and gene expression. As discussed in Chapter 6, in my opinion rather than perusing the 'top hits' identified during genome-wide screens, it would be better to target genes where differential methylation occurs in gene promotor/transcription start sites. Furthermore it would be interesting to collaborate with the Cambridge group (Prof Ken Smith, Dr James Lee et al) in an attempt to replicate their work on identifying a transcriptomic signature in CD8+ cells that can predict disease outcome.[260] Lastly, as part of the IBD-CHARACTER project transcriptomic study will be performed using the Ion Torrent Ampliseq platform (Life Technologies). The Ion Torrent platform has been used in a number of next-generation sequencing settings and works by detection of a PH change when the appropriate base is incorporated into a single strand of DNA (with subsequent release of a hydrogen ion). The major benefits of this technique are a small starting RNA requirement, rapid processing time and wider coverage that provide by microarrays (e.g. HT12 array used in this study) as well as the other benefits of mRNA sequencing (see introduction to Chapter 8).

9.2.4.3 microRNA

As was discussed in the conclusions to chapter 8, the microRNA aspect of this project was an exploratory analysis and underpowered to detect differences between cases and controls. This work is subject to current grant applications to perform small RNA sequencing on the remaining separated cell samples and also on small RNAs derived from serum and whole blood (PAXgene).

9.2.5 Functional work

Like the large genetic studies in IBD, whilst there has been tremendous success in identifying variants associated with disease, very little is known regarding the functional implications of the polymorphisms. There is a similar danger in the sphere of EWAS that large numbers of differentially methylated loci are identified with a lack of understanding into the underlying function of these modifications. Whilst this has not be a focus of this thesis there are several important functional experiments that would be possible to further investigate the role of the key genes (VMP1/miR21/RPS6KA2/TXK) identified in this project. The Gastrointestinal lab in Edinburgh specialises in functional genomics and several techniques are already optimised that could be used to explore these genes. Established gut (e.g SW80, HCT-116) and non-gut (HEK293, THP-1) cell lines could be used to investigate these genes. It would be important initially to characterise the normal methylation state and gene expression levels of the target genes in the planned cell lines, as well as characterising the genotype of key SNPs may form meQTLs with the genes of interest. The genes could knocked down (for example using small interfering RNA [siRNA]) or overexpressed (e.g. using viral vectors). There are several DNA methyltransferase inhibitors such as 5-azacytidine and decitabine could potentially be used experimentally to prevent/reduce levels of DNA methylation, although these are pan-DMT inhibitors and would not target specific methylation sites. An extremely exciting new technology is the use of caspase9/clustered regularly-interspaced short palindromic regions (CRISPR) to edit specific portions of the genome and may be used to target specific methylation sites. Recently a group have used engineered a fusion of transcription-like effectors (TALEs) and the Ten eleven translocation (TET1, enzymes that remove DNA methylation by oxidising 5mC to 5hmC) enzymes to modify specific methylated promotor regions to alter gene expression.[468] Similar experiments could be performed in primary cultured human blood cells (i.e. immunomagnetically separated as described in Chapter 2) or in Epstein - Barr virus transformed human blood cells.

Chapter 10. References

1 Molodecky NA, Soon IS, Rabi DM, Ghali WA, Ferris M, Chernoff G, *et al.* Increasing incidence and prevalence of the inflammatory bowel diseases with time, based on systematic review. *Gastroenterology* 2012;**142**:46–54.e42; quiz e30.

2 Henderson P, Hansen R, Cameron FL, Gerasimidis K, Rogers P, Bisset M, *et al.* Rising incidence of pediatric inflammatory bowel disease in Scotland. *Inflamm Bowel Dis* 2012;**18**:999–1005.

3 Benchimol E, Guttmann A, Griffiths AM, Rabeneck L, Mack DR, Brill H, *et al.* Increasing incidence of paediatric inflammatory bowel disease in Ontario, Canada: evidence from health administrative data. *Gut* 2009;**58**:1490–7.

4 Ng SC, Bernstein CN, Vatn MH, Lakatos PL, Loftus E V, Tysk C, *et al.* Geographical variability and environmental risk factors in inflammatory bowel disease. *Gut* 2013;:1–17.

5 Gunesh S, Thomas GAO, Williams GT, Roberts A, Hawthorne AB. The incidence of Crohn's disease in Cardiff over the last 75 years: an update for 1996-2005. *Aliment Pharmacol Ther* 2008;**27**:211–9.

6 IBD Standards Group UK. Quality Care Service standards for the healthcare of people who have Inflammatory Bowel Disease (IBD). 2009.

7 van der Valk ME, Mangen M-JJ, Leenders M, Dijkstra G, van Bodegraven A a, Fidder HH, *et al.* Healthcare costs of inflammatory bowel disease have shifted from hospitalisation and surgery towards anti-TNFα therapy: results from the COIN study. *Gut* 2012;:1–8.

8 Bewtra M, Kaiser LM, TenHave T, Lewis JD. Crohn's disease and ulcerative colitis are associated with elevated standardized mortality ratios: a meta-analysis. *Inflamm Bowel Dis* 2013;**19**:599–613.

9 Kaplan GG, McCarthy EP, Ayanian JZ, Korzenik J, Hodin R, Sands BE. Impact of hospital volume on postoperative morbidity and mortality following a colectomy for ulcerative colitis. *Gastroenterology* 2008;**134**:680–7.

10 Roberts SE, Williams JG, Yeates D, Goldacre MJ. Mortality in patients with and without colectomy admitted to hospital for ulcerative colitis and Crohn's disease: record linkage studies. *BMJ* 2007;**335**:1033.

11 Nicholls RJ, Clark DN, Kelso L, Crowe AM, Knight AD, Hodgkins P, *et al.* Nationwide linkage analysis in Scotland implicates age as the critical overall determinant of mortality in ulcerative colitis. *Aliment Pharmacol Ther* 2010;**31**:1310–21.

12 Kennedy NA, Clark DN, Bauer J, Crowe AM, Knight AD, Nicholls RJ, *et al.* Nationwide linkage analysis in Scotland to assess mortality following hospital admission for Crohn's disease: 1998-2000. *Aliment Pharmacol Ther* 2012;**35**:142–53.

13 Ventham NT, Kennedy NA, Duffy A, Clark DN, Crowe AM, Knight AD, *et al.* Nationwide linkage analysis in Scotland-Has mortality following hospital admission for Crohn's disease changed in the early 21st century? J. Crohn's Colitis. 2014.

14 Ventham NT, Kennedy NA, Duffy A, Clark DN, Crowe AM, Knight AD, *et al.* Comparison of mortality following hospitalisation for ulcerative colitis in Scotland between 1998-2000 and 2007-2009. *Aliment Pharmacol Ther* 2014;**39**:1387–97.

15 Kalla R, Ventham NT, Satsangi J, Arnott IDR. Crohn's disease. *Br Med J* 2014;**349**:g6670–g6670.

16 Pimentel M, Chang M, Chow EJ, Tabibzadeh S, Kirit-Kiriak V, Targan SR, *et al.* Identification of a prodromal period in Crohn's disease but not ulcerative colitis. *Am J Gastroenterol* 2000;**95**:3458–62.

17 Fine KD, Schiller LR. AGA technical review on the evaluation and management of chronic diarrhea. *Gastroenterology* 1999;**116**:1464–86.

18 Truelove SC, Willoughby CP, Lee EG, Kettlewell MG. Further experience in the treatment of severe attacks of ulcerative colitis. *Lancet* 1978;**2**:1086–8.

19 Dinesen LC, Walsh AJ, Protic MN, Heap G, Cummings F, Warren BF, *et al.* The pattern and outcome of acute severe colitis. *J Crohns Colitis* 2010;**4**:431–7.

20 Silverberg MS, Satsangi J, Ahmad T, Arnott ID, Bernstein CN, Brant SR, *et al.* Toward an integrated clinical, molecular and serological classification of inflammatory bowel disease: Report of a Working Party of the 2005 Montreal World Congress of Gastroenterology. *Can J Gastroenterol* 2005;**19 Suppl A**:5–36.

21 Levine A, Griffiths A, Markowitz J, Wilson DC, Turner D, Russell RK, *et al.* Pediatric modification of the Montreal classification for inflammatory bowel disease. *Inflamm Bowel Dis* 2011;**17**:1314–21.

22 Van Assche G, Dignass A, Reinisch W, van der Woude CJ, Sturm A, De Vos M, *et al.* The second European evidence-based Consensus on the diagnosis and management of Crohn's disease: Special situations. *J Crohns Colitis* 2010;**4**:63–101.

23 Mowat C, Cole A, Windsor A, Ahmad T, Arnott I, Driscoll R, *et al.* Guidelines for the management of inflammatory bowel disease in adults. *Gut* 2011;**60**:571–607.

24 Henriksen M, Jahnsen J, Lygren I, Sauar J, Schulz T, Stray N, *et al.* Change of diagnosis during the first five years after onset of inflammatory bowel disease: results of a prospective follow-up study (the IBSEN Study). *Scand J Gastroenterol* 2006;**41**:1037–43.

25 Sidhu R, Sanders DS, Morris AJ, McAlindon ME. Guidelines on small bowel enteroscopy and capsule endoscopy in adults. *Gut* 2008;**57**:125–36.

26 Johnson GJ, Cosnes J, Mansfield JC. Review article: smoking cessation as primary therapy to modify the course of Crohn's disease. *Aliment Pharmacol Ther* 2005;**21**:921–31.

27 Dam AN, Berg AM, Farraye FA. Environmental influences on the onset and clinical course of Crohn's disease-part 1: an overview of external risk factors. *Gastroenterol Hepatol (N Y)* 2013;**9**:711–7.

28 Mahid SS, Minor KS, Soto RE, Hornung CA, Galandiuk S. Smoking and inflammatory bowel disease: a meta-analysis. *Mayo Clin Proc* 2006;**81**:1462–71.

29 Seksik P, Nion-Larmurier I, Sokol H, Beaugerie L, Cosnes J. Effects of light smoking consumption on the clinical course of Crohn's disease. *Inflamm Bowel Dis* 2009;**15**:734–41.

30 Takeuchi K, Smale S, Premchand P, Maiden L, Sherwood R, Thjodleifsson B, *et al.* Prevalence and mechanism of nonsteroidal anti-inflammatory drug-induced clinical relapse in patients with inflammatory bowel disease. *Clin Gastroenterol Hepatol* 2006;**4**:196–202.

31 Meyer AM, Ramzan NN, Heigh RI, Leighton JA. Relapse of inflammatory bowel disease associated with use of nonsteroidal anti-inflammatory drugs. *Dig Dis Sci* 2006;**51**:168–72.

32 Evans JM, McMahon AD, Murray FE, McDevitt DG, MacDonald TM. Non-steroidal anti-inflammatory drugs are associated with emergency admission to hospital for colitis due to inflammatory bowel disease. *Gut* 1997;**40**:619–22.

33 Beaugerie L, Sokol H. Clinical, serological and genetic predictors of inflammatory bowel disease course. *World J Gastroenterol* 2012;**18**:3806–13.

34 Dignass A, Van Assche G, Lindsay JO, Lémann M, Söderholm J, Colombel JF, *et al.* The second European evidence-based Consensus on the diagnosis and management of Crohn's disease: Current management. *J Crohns Colitis* 2010;**4**:28–62.

35 Zachos M, Tondeur M, Griffiths AM. Enteral nutritional therapy for induction of remission

in Crohn's disease. *Cochrane database Syst Rev* 2007;:CD000542.

36 Ford AC, Bernstein CN, Khan KJ, Abreu MT, Marshall JK, Talley NJ, *et al.* Glucocorticosteroid therapy in inflammatory bowel disease: systematic review and meta-analysis. *Am J Gastroenterol* 2011;**106**:590–599; quiz 600.

37 Lichtenstein GR, Feagan BG, Cohen RD, Salzberg BA, Diamond RH, Price S, *et al.* Serious infection and mortality in patients with Crohn's disease: more than 5 years of follow-up in the TREAT^{TM} registry. *Am J Gastroenterol* 2012;**107**:1409–22.

38 Summers RW, Switz DM, Sessions JT, Becktel JM, Best WR, Kern F, *et al.* National Cooperative Crohn's Disease Study: results of drug treatment. *Gastroenterology* 1979;**77**:847–69.

39 Benchimol EI, Seow CH, Otley AR, Steinhart AH. Budesonide for maintenance of remission in Crohn's disease. *Cochrane database Syst Rev* 2009;:CD002913.

40 Marshall JK, Thabane M, Steinhart AH, Newman JR, Anand A, Irvine EJ. Rectal 5-aminosalicylic acid for induction of remission in ulcerative colitis. *Cochrane database Syst Rev* 2010;:CD004115.

41 Ford AC, Khan KJ, Achkar J-P, Moayyedi P. Efficacy of oral vs. topical, or combined oral and topical 5-aminosalicylates, in Ulcerative Colitis: systematic review and meta-analysis. *Am J Gastroenterol* 2012;**107**:167–76; author reply 177.

42 Ford AC, Achkar J-P, Khan KJ, Kane S V, Talley NJ, Marshall JK, *et al.* Efficacy of 5-aminosalicylates in ulcerative colitis: systematic review and meta-analysis. *Am J Gastroenterol* 2011;**106**:601–16.

43 Laharie D, Bourreille A, Branche J, Allez M, Bouhnik Y, Filippi J, *et al.* Ciclosporin versus infliximab in patients with severe ulcerative colitis refractory to intravenous steroids: a parallel, open-label randomised controlled trial. *Lancet* 2012;**380**:1909–15.

44 Lichtiger S, Present DH, Kornbluth A, Gelernt I, Bauer J, Galler G, *et al.* Cyclosporine in severe ulcerative colitis refractory to steroid therapy. *N Engl J Med* 1994;**330**:1841–5.

45 Panaccione R, Ghosh S, Middleton S, Marquez JR, Scott BB, Flint L, *et al.* Combination therapy with infliximab and azathioprine is superior to monotherapy with either agent in ulcerative colitis. *Gastroenterology* 2014;**146**.

46 Schreiber S, Reinisch W, Colombel JF, Sandborn WJ, Hommes DW, Robinson AM, *et al.* Subgroup analysis of the placebo-controlled CHARM trial: increased remission rates through 3 years for adalimumab-treated patients with early Crohn's disease. *J Crohns Colitis* 2013;**7**:213–21.

47 D'Haens G, Baert F, van Assche G, Caenepeel P, Vergauwe P, Tuynman H, *et al.* Early combined immunosuppression or conventional management in patients with newly diagnosed Crohn's disease: an open randomised trial. *Lancet* 2008;**371**:660–7.

48 Schreiber S, Colombel J-F, Bloomfield R, Nikolaus S, Schölmerich J, Panés J, *et al.* Increased response and remission rates in short-duration Crohn's disease with subcutaneous certolizumab pegol: an analysis of PRECiSE 2 randomized maintenance trial data. *Am J Gastroenterol* 2010;**105**:1574–82.

49 Allen PB, Peyrin-Biroulet L. Moving towards disease modification in inflammatory bowel disease therapy. *Curr Opin Gastroenterol* 2013;**29**:397–404.

50 Colombel JF, Sandborn WJ, Reinisch W, Mantzaris GJ, Kornbluth A, Rachmilewitz D, *et al.* Infliximab, azathioprine, or combination therapy for Crohn's disease. *N Engl J Med* 2010;**362**:1383–95.

51 Bressler B, Siegel CA. Beware of the swinging pendulum: anti-tumor necrosis factor monotherapy vs combination therapy for inflammatory bowel disease. *Gastroenterology* 2014;**146**:884–7.

52 Louis E, Mary J-Y, Vernier-Massouille G, Grimaud J-C, Bouhnik Y, Laharie D, *et al.* Maintenance of remission among patients with Crohn's disease on antimetabolite therapy after infliximab therapy is stopped. *Gastroenterology* 2012;**142**:63–70.e5; quiz e31.

53 Pittet V, Froehlich F, Maillard MH, Mottet C, Gonvers J-J, Felley C, *et al.* When do we dare to stop biological or immunomodulatory therapy for Crohn's disease? Results of a multidisciplinary European expert panel. *J Crohns Colitis* 2013;**7**:820–6.

54 Hutas G. Golimumab, a fully human monoclonal antibody against TNFalpha. *Curr Opin Mol Ther* 2008;**10**:393–406.

55 Sandborn WJ, Feagan BG, Marano C, Zhang H, Strauss R, Johanns J, *et al.* Subcutaneous golimumab induces clinical response and remission in patients with moderate-to-severe ulcerative colitis. *Gastroenterology* 2014;**146**:85–95.

56 Prefontaine E, Sutherland LR, Macdonald JK, Cepoiu M. Azathioprine or 6-mercaptopurine for maintenance of remission in Crohn's disease. *Cochrane database Syst Rev* 2009;:CD000067.

57 Khan KJ, Dubinsky MC, Ford AC, Ullman T a, Talley NJ, Moayyedi P. Efficacy of immunosuppressive therapy for inflammatory bowel disease: a systematic review and meta-analysis. *Am J Gastroenterol* 2011;**106**:630–42.

58 Patel V, Wang Y, MacDonald JK, McDonald JWD, Chande N. Methotrexate for maintenance of remission in Crohn's disease. *Cochrane database Syst Rev* 2014;**8**:CD006884.

59 Fraser AG. Methotrexate: first-line or second-line immunomodulator? *Eur J Gastroenterol Hepatol* 2003;**15**:225–31.

60 Kennedy NA, Kalla R, Warner B, Gambles CJ, Musy R, Reynolds S, *et al.* Thiopurine withdrawal during sustained clinical remission in inflammatory bowel disease: relapse and recapture rates, with predictive factors in 237 patients. *Aliment Pharmacol Ther* 2014;**40**:1313–23.

61 Ford AC, Khan KJ, Sandborn WJ, Hanauer SB, Moayyedi P. Efficacy of topical 5-aminosalicylates in preventing relapse of quiescent ulcerative colitis: a meta-analysis. *Clin Gastroenterol Hepatol* 2012;**10**:513–9.

62 Ramadas A V, Gunesh S, Thomas GAO, Williams GT, Hawthorne AB. Natural history of Crohn's disease in a population-based cohort from Cardiff (1986-2003): a study of changes in medical treatment and surgical resection rates. *Gut* 2010;**59**:1200–6.

63 Nguyen GC, Nugent Z, Shaw S, Bernstein CN. Outcomes of patients with Crohn's disease improved from 1988 to 2008 and were associated with increased specialist care. *Gastroenterology* 2011;**141**:90–7.

64 Nordgren SR, Fasth SB, Oresland TO, Hultén LA. Long-term follow-up in Crohn's disease. Mortality, morbidity, and functional status. *Scand J Gastroenterol* 1994;**29**:1122–8.

65 Kim NK, Senagore AJ, Luchtefeld MA, MacKeigan JM, Mazier WP, Belknap K, *et al.* Long-term outcome after ileocecal resection for Crohn's disease. *Am Surg* 1997;**63**:627–33.

66 Tilney HS, Constantinides VA, Heriot AG, Nicolaou M, Athanasiou T, Ziprin P, *et al.* Comparison of laparoscopic and open ileocecal resection for Crohn's disease: a metaanalysis. *Surg Endosc* 2006;**20**:1036–44.

67 He X, Chen Z, Huang J, Lian L, Rouniyar S, Wu X, *et al.* Stapled side-to-side anastomosis might be better than handsewn end-to-end anastomosis in ileocolic resection for Crohn's disease: a meta-analysis. *Dig Dis Sci* 2014;**59**:1544–51.

68 Yamamoto T, Watanabe T. Surgery for luminal Crohn's disease. *World J Gastroenterol* 2014;**20**:78–90.

69 Solberg IC, Lygren I, Jahnsen J, Aadland E, Høie O, Cvancarova M, *et al.* Clinical course during the first 10 years of ulcerative colitis: results from a population-based inception

cohort (IBSEN Study). *Scand J Gastroenterol* 2009;**44**:431–40.

70 Turner D, Walsh CM, Steinhart AH, Griffiths AM. Response to corticosteroids in severe ulcerative colitis: a systematic review of the literature and a meta-regression. *Clin Gastroenterol Hepatol* 2007;**5**:103–10.

71 Annese V, Vecchi M. Use of biosimilars in inflammatory bowel disease: Statements of the Italian Group for Inflammatory Bowel Disease. *Dig Liver Dis* 2014;**46**:963–8.

72 Park W, Hrycaj P, Jeka S, Kovalenko V, Lysenko G, Miranda P, *et al.* A randomised, double-blind, multicentre, parallel-group, prospective study comparing the pharmacokinetics, safety, and efficacy of CT-P13 and innovator infliximab in patients with ankylosing spondylitis: the PLANETAS study. *Ann Rheum Dis* 2013;**72**:1605–12.

73 Gecse KB, Lovász BD, Farkas K, Banai J, Bene L, Gasztonyi B, *et al.* Efficacy And Safety Of The Biosimilar Infliximab CT-P13 Treatment In Inflammatory Bowel Diseases: A Prospective, Multicentre, Nationwide Cohort. *J Crohn's Colitis* 2015;:jjv220.

74 Wyant T, Leach T, Sankoh S, Wang Y, Paolino J, Pasetti MF, *et al.* Vedolizumab affects antibody responses to immunisation selectively in the gastrointestinal tract: randomised controlled trial results. *Gut* 2015;**64**:77–83.

75 Langer-Gould A, Atlas SW, Green AJ, Bollen AW, Pelletier D. Progressive multifocal leukoencephalopathy in a patient treated with natalizumab. 2005.

76 Feagan BG, Rutgeerts P, Sands BE, Hanauer S, Colombel J-F, Sandborn WJ, *et al.* Vedolizumab as induction and maintenance therapy for ulcerative colitis. *N Engl J Med* 2013;**369**:699–710.

77 Sandborn WJ, Feagan BG, Rutgeerts P, Hanauer S, Colombel J-F, Sands BE, *et al.* Vedolizumab as induction and maintenance therapy for Crohn's disease. *N Engl J Med* 2013;**369**:711–21.

78 Sands BE, Feagan BG, Rutgeerts P, Colombel J-F, Sandborn WJ, Sy R, *et al.* Effects of Vedolizumab Induction Therapy for Patients With Crohn's Disease in Whom Tumor Necrosis Factor Antagonist Treatment Had Failed. *Gastroenterology* 2014;**147**:618–627.e3.

79 Vermeire S, O'Byrne S, Keir M, Williams M, Lu TT, Mansfield JC, *et al.* Etrolizumab as induction therapy for ulcerative colitis: a randomised, controlled, phase 2 trial. *Lancet* 2014;**384**:309–18.

80 Krueger GG, Langley RG, Leonardi C, Yeilding N, Guzzo C, Wang Y, *et al.* A human interleukin-12/23 monoclonal antibody for the treatment of psoriasis. 2007.

81 Sandborn WJ, Feagan BG, Fedorak RN, Scherl E, Fleisher MR, Katz S, *et al.* A Randomized Trial of Ustekinumab, a Human Interleukin-12/23 Monoclonal Antibody, in Patients With Moderate-to-Severe Crohn's Disease. *Gastroenterology* 2008;**135**:1130–41.

82 Sandborn WJ, Gasink C, Gao L-L, Blank M a, Johanns J, Guzzo C, *et al.* Ustekinumab induction and maintenance therapy in refractory Crohn's disease. *N Engl J Med* 2012;**367**:1519–28.

83 Gottlieb AB, Kalb RE, Langley RG, Krueger GG, de Jong EMGJ, Guenther L, *et al.* Safety observations in 12095 patients with psoriasis enrolled in an international registry (PSOLAR): experience with infliximab and other systemic and biologic therapies. *J Drugs Dermatol* 2014;**13**:1441–8.

84 Monteleone G, Neurath MF, Ardizzone S, Di Sabatino A, Fantini MC, Castiglione F, *et al.* Mongersen, an Oral SMAD7 Antisense Oligonucleotide, and Crohn's Disease. *N Engl J Med* 2015;**372**:1104–13.

85 Xavier RJ, Podolsky DK. Unravelling the pathogenesis of inflammatory bowel disease. *Nature* 2007;**448**:427–34.

86 Liu JZ, van Sommeren S, Huang H, Ng SC, Alberts R, Takahashi A, *et al.* Association analyses

identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat Genet* 2015;**47**:979–89.

87 Van Limbergen J, Stevens C, Nimmo ER, Wilson DC, Satsangi J. Autophagy: from basic science to clinical application. *Mucosal Immunol* 2009;**2**:315–30.

88 Jostins L, Ripke S, Weersma RK, Duerr RH, McGovern DP, Hui KY, *et al.* Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* 2012;**490**:119–24.

89 Khor B, Gardet A, Xavier RJ. Genetics and pathogenesis of inflammatory bowel disease. *Nature* 2011;**474**:307–17.

90 Ahmad T, Armuzzi A, Bunce M, Mulcahy-Hawes K, Marshall SE, Orchard TR, *et al.* The molecular classification of the clinical manifestations of Crohn's disease. *Gastroenterology* 2002;**122**:854–66.

91 Futami S, Aoyama N, Honsako Y, Tamura T, Morimoto S, Nakashima T, *et al.* HLA-DRB1*1502 allele, subtype of DR15, is associated with susceptibility to ulcerative colitis and its progression. *Dig Dis Sci* 1995;**40**:814–8.

92 Lesage S, Zouali H, Cézard J-P, Colombel J-F, Belaiche J, Almer S, *et al.* CARD15/NOD2 mutational analysis and genotype-phenotype correlation in 612 patients with inflammatory bowel disease. *Am J Hum Genet* 2002;**70**:845–57.

93 Cleynen I, Boucher G, Jostins L, Schumm LP, Zeissig S, Ahmad T, *et al.* Inherited determinants of Crohn ' s disease and ulcerative colitis phenotypes : a genetic association study. *Lancet* 2015;**6736**:1–12.

94 Anderson CA, Boucher G, Lees CW, Franke A, D'Amato M, Taylor KD, *et al.* Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. *Nat Genet* 2011;**43**:246–52.

95 Brand S. Crohn's disease: Th1, Th17 or both? The change of a paradigm: new immunological and genetic insights implicate Th17 cells in the pathogenesis of Crohn's disease. *Gut* 2009;**58**:1152–67.

96 Franke A, Balschun T, Karlsen TH, Sventoraityte J, Nikolaus S, Mayr G, *et al.* Sequence variants in IL10, ARPC2 and multiple other loci contribute to ulcerative colitis susceptibility. *Nat Genet* 2008;**40**:1319–23.

97 Franke A, Mcgovern DPB, Barrett JC, Wang K, Graham L, Ahmad T, *et al.* Meta-Analysis Increases to 71 the Tally of Confirmed Crohn's Disease Susceptibility Loci. *Nat Genet* 2010;**42**:1118–25.

98 Glocker E-O, Kotlarz D, Boztug K, Gertz EM, Schäffer A a, Noyan F, *et al.* Inflammatory bowel disease and mutations affecting the interleukin-10 receptor. *N Engl J Med* 2009;**361**:2033–45.

99 Glocker E-O, Frede N, Perro M, Sebire N, Elawad M, Shah N, *et al.* Infant colitis--it's in the genes. *Lancet* 2010;**376**:1272.

100 Anderson CA, Boucher G, Lees CW, Franke A, D'mato M, Taylor K, *et al.* Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. *Nat Genet* 2011;**43**:246–52.

101 Barrett JC, Hansoul S, Nicolae DL, Cho JH, Duerr RH, Rioux JD, *et al.* Genome-wide association defines more than thirty distinct susceptibility loci for Crohn's disease. *Nat Genet* 2009;**40**:955–62.

102 Cooney R, Baker J, Brain O, Danis B, Pichulik T, Allan P, *et al.* NOD2 stimulation induces autophagy in dendritic cells influencing bacterial handling and antigen presentation. *Nat Med* 2010;**16**:90–7.

103 Travassos LH, Carneiro LAM, Ramjeet M, Hussey S, Kim Y-G, Magalhães JG, *et al.* Nod1 and

Nod2 direct autophagy by recruiting ATG16L1 to the plasma membrane at the site of bacterial entry. *Nat Immunol* 2010;**11**:55–62.

104 Aldhous MC, Soo K, Stark L a, Ulanicka A a, Easterbrook JE, Dunlop MG, *et al.* Cigarette smoke extract (CSE) delays NOD2 expression and affects NOD2/RIPK2 interactions in intestinal epithelial cells. *PLoS One* 2011;**6**:e24715.

105 Cho JH, Brant SR. Recent insights into the genetics of inflammatory bowel disease. *Gastroenterology* 2011;**140**:1704–12.

106 Intemann CD, Thye T, Niemann S, Browne ENL, Amanua Chinbuah M, Enimil A, *et al.* Autophagy gene variant IRGM -261T contributes to protection from tuberculosis caused by Mycobacterium tuberculosis but not by M. africanum strains. *PLoS Pathog* 2009;**5**:e1000577.

107 Barrett JC, Lee JC, Lees CW, Prescott NJ, Anderson CA, Phillips A, *et al.* Genome-wide association study of ulcerative colitis identifies three new susceptibility loci, including the HNF4A region. *Nat Genet* 2009;**41**:1330–4.

108 Yamagata K, Furuta H, Oda N, Kaisaki PJ, Menzel S, Cox NJ, *et al.* Mutations in the hepatocyte nuclear factor-4gene in maturity-onset diabetes of the young (MODY1). *Nature* 1996;**384**:458–60.

109 Cargill M, Schrodi SJ, Chang M, Garcia VE, Brandon R, Callis KP, *et al.* A Large-Scale Genetic Association Study Confirms IL12B and Leads to the Identification of IL23R as Psoriasis-Risk Genes. *Am J Hum Genet* 2007;**80**:273–90.

110 Burton P, Clayton D, Cardon L, Craddock N, Deloukas P, Duncanson A, *et al.* Association scan of 14,500 nsSNPs in four common diseases identifies variants involved in autoimmunity. *Nat Genet* 2009;**39**:1329–37.

111 Cho JH, Gregersen PK. Genomics and the multifactorial nature of human autoimmune disease. *N Engl J Med* 2011;**365**:1612–23.

112 Wang K, Baldassano R, Zhang H, Qu H-Q, Imielinski M, Kugathasan S, *et al.* Comparative genetic analysis of inflammatory bowel disease and type 1 diabetes implicates multiple loci with opposite effects. *Hum Mol Genet* 2010;**19**:2059–67.

113 Muriboberg K, Melum E, Folseraas T, Schrumpf E. Three ulcerative colitis susceptibility loci are associated with primary sclerosing cholangitis and indicate a role for IL2, REL and CARD9. *Hepatology* 2012;**53**:1977–85.

114 Lees CW, Barrett JC, Parkes M, Satsangi J. New IBD genetics: common pathways with other diseases. *Gut* 2011;**60**:1739–53.

115 Worthey EA, Mayer AN, Syverson GD, Helbling D, Bonacci BB, Decker B, *et al.* Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. *Genet Med* 2011;**13**:255–62.

116 Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012;**489**:57–74.

117 Brant SR. Promises, Delivery, and Challenges of Inflammatory Bowel Disease Risk Gene Discovery. *Clin Gastroenterol Hepatol* 2012;**11**:22–6.

118 Dassopoulos T, Nguyen GC, Talor MV, Datta LW, Isaacs KL, Lewis JD, *et al.* NOD2 mutations and anti-Saccharomyces cerevisiae antibodies are risk factors for Crohn's disease in African Americans. *Am J Gastroenterol* 2010;**105**:378–86.

119 Inoue N. Lack of common NOD2 variants in Japanese patients with Crohn's disease. *Gastroenterology* 2002;**123**:86–91.

120 US Food and Drug Administration. Imuran (azathioprine) product information. Drugs@FDA. 2011;:1–9.

121 Roberts RL, Barclay ML. Current relevance of pharmacogenetics in immunomodulation treatment for Crohn's disease. *J Gastroenterol Hepatol* 2012;**27**:1546–54.

122 Heap GA, Weedon MN, Bewshea CM, Singh A, Chen M, Satchwell JB, *et al.* HLA-DQA1-HLA-DRB1 variants confer susceptibility to pancreatitis induced by thiopurine immunosuppressants. *Nat Genet* 2014;**46**:1131–4.

123 Danese S. New therapies for inflammatory bowel disease: from the bench to the bedside. *Gut* 2012;**61**:918–32.

124 Petronis A. Epigenetics as a unifying principle in the aetiology of complex traits and diseases. *Nature* 2010;**465**:721–7.

125 Calkins BM. A meta-analysis of the role of smoking in inflammatory bowel disease. *Dig Dis Sci* 1989;**34**:1841–54.

126 Andersson RE, Olaison G, Tysk C, Ekbom A. Appendectomy and protection against ulcerative colitis. *N Engl J Med* 2001;**344**:808–14.

127 Chan SSM, Luben R, Bergmann MM, Boeing H, Olsen A, Tjonneland A, *et al.* Aspirin in the aetiology of Crohn's disease and ulcerative colitis: a European prospective cohort study. *Aliment Pharmacol Ther* 2011;**34**:649–55.

128 Cornish JA, Tan E, Simillis C, Clark SK, Teare J, Tekkis PP. The risk of oral contraceptives in the etiology of inflammatory bowel disease: a meta-analysis. *Am J Gastroenterol* 2008;**103**:2394–400.

129 Hviid A, Svanström H, Frisch M. Antibiotic use and inflammatory bowel diseases in childhood. *Gut* 2011;**60**:49–54.

130 Ananthakrishnan AN, Khalili H, Higuchi LM, Bao Y, Korzenik JR, Giovannucci EL, *et al.* Higher predicted vitamin D status is associated with reduced risk of Crohn's disease. *Gastroenterology* 2012;**142**:482–9.

131 Tjonneland A, Overvad K, Bergmann MM, Nagel G, Linseisen J, Hallmans G, *et al.* Linoleic acid, a dietary n-6 polyunsaturated fatty acid, and the aetiology of ulcerative colitis: a nested case-control study within a European prospective cohort study. *Gut* 2009;**58**:1606–11.

132 de Silva PS a, Olsen A, Christensen J, Schmidt EB, Overvaad K, Tjonneland A, *et al.* An association between dietary arachidonic acid, measured in adipose tissue, and ulcerative colitis. *Gastroenterology* 2010;**139**:1912–7.

133 Ananthakrishnan AN, Khalili H, Konijeti GG, Higuchi LM, de Silva P, Korzenik JR, *et al.* A prospective study of long-term intake of dietary fiber and risk of Crohn's disease and ulcerative colitis. *Gastroenterology* 2013;**145**:970–7.

134 Clarke SF, Murphy EF, O'Sullivan O, Lucey AJ, Humphreys M, Hogan A, *et al.* Exercise and associated dietary extremes impact on gut microbial diversity. *Gut* 2014;**63**:1913–20.

135 Hold GL. The gut microbiota, dietary extremes and exercise. *Gut* 2014;**11**:gutjnl-2014-307305.

136 Morgan HD, Sutherland HGE, Martin DIK, Whitelaw E. Epigenetic inheritance at the agouti locus in the mouse. *Nat Genet* 2006;**23**:314–8.

137 Heijmans BT, Tobi EW, Stein AD, Putter H, Blauw GJ, Susser ES, *et al.* Persistent epigenetic differences associated with prenatal exposure to famine in humans. *Proc Natl Acad Sci U S A* 2008;**105**:17046–9.

138 Schaible TD, Harris RA, Dowd SE, Smith CW, Kellermayer R. Maternal methyl-donor supplementation induces prolonged murine offspring colitis susceptibility in association with mucosal epigenetic and microbiomic changes. *Hum Mol Genet* 2011;**20**:1687–96.

139 Waterland RA, Jirtle RL. Transposable Elements : Targets for Early Nutritional Effects on

Epigenetic Gene Regulation. *Mol Cell Biol* 2003;**23**:5293–300.

140 Tobi EW, Lumey LH, Talens RP, Kremer D, Putter H, Stein AD, *et al.* DNA methylation differences after exposure to prenatal famine are common and timing- and sex-specific. *Hum Mol Genet* 2009;**18**:4046–53.

141 Heijmans BT, Tobi EW, Lumey LH, Slagboom PE. Archive of the prenatal environment. *Epigenetics* 2009;**4**:526–31.

142 Horvath S. Erratum to: DNA methylation age of human tissues and cell types. *Genome Biol* 2013;**14**:96.

143 Relton CL, Davey Smith G. Epigenetic Epidemiology of Common Complex Disease: Prospects for Prediction, Prevention, and Treatment. *PLoS Med* 2010;**7**:e1000356.

144 Ventham NT, Kennedy NA, Nimmo ER, Satsangi J. Beyond gene discovery in inflammatory bowel disease: the emerging role of epigenetics. *Gastroenterology* 2013;**145**:293–308.

145 Brain O, Allan P, Pichulik T, Khatamzas E, Simpson P, Jewell D, *et al.* NOD2 regulation of micrornas. *Gut* 2011;**60**:A37–A37.

146 Janson PCJ, Winerdal ME, Winqvist O. At the crossroads of T helper lineage commitment-Epigenetics points the way. *Biochim Biophys Acta* 2009;**1790**:906–19.

147 Faulk C, Dolinoy DC. Timing is everything: The when and how of environmentally induced changes in the epigenome of animals. *Epigenetics* 2011;**6**:791–7.

148 De Santis M, Selmi C. The therapeutic potential of epigenetics in autoimmune diseases. *Clin Rev Allergy Immunol* 2012;**42**:92–101.

149 Ushijima T, Watanabe N, Okochi E, Kaneda A, Sugimura T, Miyamoto K. Fidelity of the methylation pattern and its variation in the genome. *Genome Res* 2003;**13**:868–74.

150 Petronis A. Epigenetics as a unifying principle in the aetiology of complex traits and diseases. *Nature* 2010;**465**:721–7.

151 Breton C V, Byun H-M, Wenten M, Pan F, Yang A, Gilliland FD. Prenatal tobacco smoke exposure affects global and gene-specific DNA methylation. *Am J Respir Crit Care Med* 2009;**180**:462–7.

152 Tsaprouni LG, Yang TP, Bell J, Dick KJ, Kanoni S, Nisbet J, *et al.* Cigarette smoking reduces DNA methylation levels at multiple genomic loci but the effect is partially reversible upon cessation. *Epigenetics* 2014;**9**:1382–96.

153 Takahashi K, Sugi Y, Hosono A, Kaminogawa S. Epigenetic regulation of TLR4 gene expression in intestinal epithelial cells for the maintenance of intestinal homeostasis. *J Immunol* 2009;**183**:6522–9.

154 Leung C-H, Lam W, Ma D-L, Gullen E a, Cheng Y-C. Butyrate mediates nucleotide-binding and oligomerisation domain (NOD) 2-dependent mucosal immune responses against peptidoglycan. *Eur J Immunol* 2009;**39**:3529–37.

155 Smallwood S a, Kelsey G. De novo DNA methylation: a germ cell perspective. *Trends Genet* 2012;**28**:33–42.

156 Blewitt ME, Vickaryous NK, Paldi A, Koseki H, Whitelaw E. Dynamic reprogramming of DNA methylation at an epigenetically sensitive allele in mice. *PLoS Genet* 2006;**2**:e49.

157 Grossniklaus U, Kelly B, Ferguson-Smith AC, Pembrey M, Lindquist S. Transgenerational epigenetic inheritance: how important is it? *Nat Rev Genet* 2013;**14**:228–35.

158 Lane N, Dean W, Erhardt S, Hajkova P, Surani A, Walter J, *et al.* Resistance of IAPs to methylation reprogramming may provide a mechanism for epigenetic inheritance in the mouse. *Genesis* 2003;**35**:88–93.

159 Cubas P, Vincent C, Coen E. An epigenetic mutation responsible for natural variation in

floral symmetry. *Nature* 1999;**401**:157–61.

160 Youngson N a, Whitelaw E. Transgenerational epigenetic effects. *Annu Rev Genomics Hum Genet* 2008;**9**:233–57.

161 Portela A, Esteller M. Epigenetic modifications and human disease. *Nat Biotechnol* 2010;**28**:1057–68.

162 Doi A, Park I, Wen B, Murakami P, Aryee MJ, Herb B, *et al.* Differential methylation of tissue- and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts. *Nat Genet* 2010;**41**:1350–3.

163 Hughes T, Webb R, Fei Y, Wren J, Sawalha A. DNA methylome in human CD4+ T cells identifies transcriptionally repressive and non-repressive methylation peaks. *Genes Immun* 2010;**11**:554–60.

164 Kuroda A, Rauch T a, Todorov I, Ku HT, Al-Abdullah IH, Kandeel F, *et al.* Insulin gene expression is regulated by DNA methylation. *PLoS One* 2009;**4**:e6953.

165 Clark SJ, Melki J. DNA methylation and gene silencing in cancer: which is the guilty party? *Oncogene* 2002;**21**:5380–7.

166 Issa JP, Ahuja N, Toyota M, Bronner MP, Brentnall TA. Accelerated age-related CpG island methylation in ulcerative colitis. *Cancer Res* 2001;**61**:3573–7.

167 Nimmo ER, Prendergast JG, Aldhous MC, Kennedy NA, Henderson P, Drummond HE, *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* 2012;**18**:889–99.

168 Harris RA, Nagy-Szakal D, Pedersen N, Opekun A, Bronsky J, Munkholm P, *et al.* Genome-wide peripheral blood leukocyte DNA methylation microarrays identified a single association with inflammatory bowel diseases. *Inflamm Bowel Dis* 2012;**2399**:1–8.

169 Lin Z, Hegarty JP, Yu W, Cappel J a, Chen X, Faber PW, *et al.* Identification of disease-associated DNA methylation in B cells from Crohn's disease and ulcerative colitis patients. *Dig Dis Sci* 2012;**57**:3145–53.

170 Häsler R, Feng Z, Bäckdahl L, Spehlmann ME, Franke A, Teschendorff A, *et al.* A functional methylome map of ulcerative colitis. *Genome Res* 2012;**22**:2130–7.

171 Cooke J, Zhang H, Greger L, Silva A-L, Massey D, Dawson C, *et al.* Mucosal genome-wide methylation changes in inflammatory bowel disease. *Inflamm Bowel Dis* 2012;**18**:2128–37.

172 Kouzarides T. Chromatin modifications and their function. *Cell* 2007;**128**:693–705.

173 Wang Z, Zang C, Rosenfeld JA, Schones DE, Barski A, Cuddapah S, *et al.* Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat Genet* 2008;**40**:897–903.

174 Shahbazian MD, Grunstein M. Functions of site-specific histone acetylation and deacetylation. *Annu Rev Biochem* 2007;**76**:75–100.

175 Kouzarides T. Chromatin modifications and their function. *Cell* 2007;**128**:693–705.

176 Tsaprouni LG, Ito K, Powell JJ, Adcock IM, Punchard N. Differential patterns of histone acetylation in inflammatory bowel diseases. *J Inflamm (Lond)* 2011;**8**:1.

177 Latham T, Mackay L, Sproul D, Karim M, Culley J, Harrison DJ, *et al.* Lactate, a product of glycolytic metabolism, inhibits histone deacetylase activity and promotes changes in gene expression. *Nucleic Acids Res* 2012;**40**:4794–803.

178 Halili M a, Andrews MR, Sweet MJ, Fairlie DP. Histone deacetylase inhibitors in inflammatory disease. *Curr Top Med Chem* 2009;**9**:309–19.

179 Glauben R, Siegmund B. Inhibition of histone deacetylases in inflammatory bowel diseases.

Mol Med 2011;**17**:426–33.

180 Park J-S, Lee E-J, Lee J-C, Kim W-K, Kim H-S. Anti-inflammatory effects of short chain fatty acids in IFN-gamma-stimulated RAW 264.7 murine macrophage cells: involvement of NF-kappaB and ERK signaling pathways. *Int Immunopharmacol* 2007;**7**:70–7.

181 Xu WS, Parmigiani RB, Marks P a. Histone deacetylase inhibitors: molecular mechanisms of action. *Oncogene* 2007;**26**:5541–52.

182 Butzner JD, Parmar R, Bell CJ, Dalal V. Butyrate enema therapy stimulates mucosal repair in experimental colitis in the rat. *Gut* 1996;**38**:568–73.

183 Glauben R, Batra A, Fedke I, Zeitz M, Lehr HA, Leoni F, *et al.* Histone Hyperacetylation Is Associated with Amelioration of Experimental Colitis in Mice. *J Immunol* 2006;**176**:5015–22.

184 Plöger S, Stumpff F, Penner GB, Schulzke J-D, Gäbel G, Martens H, *et al.* Microbial butyrate and its role for barrier function in the gastrointestinal tract. *Ann N Y Acad Sci* 2012;**1258**:52–9.

185 Lührs H, Gerke T, Müller JG, Melcher R, Schauber J, Boxberge F, *et al.* Butyrate inhibits NF-kappaB activation in lamina propria macrophages of patients with ulcerative colitis. *Scand J Gastroenterol* 2002;**37**:458–66.

186 Glauben R, Batra a, Stroh T, Erben U, Fedke I, Lehr H a, *et al.* Histone deacetylases: novel targets for prevention of colitis-associated cancer in mice. *Gut* 2008;**57**:613–22.

187 O'Connell RM, Rao DS, Chaudhuri A a, Baltimore D. Physiological and pathological roles for microRNAs in the immune system. *Nat Rev Immunol* 2010;**10**:111–22.

188 Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 2004;**116**:281–97.

189 O'Connell RM, Rao DS, Chaudhuri A a, Baltimore D. Physiological and pathological roles for microRNAs in the immune system. *Nat Rev Immunol* 2010;**10**:111–22.

190 Dalal SR, Kwon JH. The Role of MicroRNA in Inflammatory Bowel Disease. *Gastroenterol Hepatol* 2010;**6**:714–22.

191 Baek D, Villén J, Shin C, Camargo FD, Steven P, Bartel DP. The impact of microRNAs on protein output. *Nature* 2009;**455**:64–71.

192 McKenna LB, Schug J, Vourekas A, McKenna JB, Bramswig NC, Friedman JR, *et al.* MicroRNAs control intestinal epithelial differentiation, architecture, and barrier function. *Gastroenterology* 2010;**139**:1654–64, 1664.e1.

193 Wu F, Zikusoka M, Trindade A, Dassopoulos T, Harris ML, Bayless TM, *et al.* MicroRNAs are differentially expressed in ulcerative colitis and alter expression of macrophage inflammatory peptide-2 alpha. *Gastroenterology* 2008;**135**:1624–1635.e24.

194 Bian Z, Li L, Cui J, Zhang H, Liu Y, Zhang C-Y, *et al.* Role of miR-150-targeting c-Myb in colonic epithelial disruption during dextran sulphate sodium-induced murine experimental colitis and human ulcerative colitis. *J Pathol* 2011;**225**:544–53.

195 Takagi T, Naito Y, Mizushima K, Hirata I, Yagi N, Tomatsuri N, *et al.* Increased expression of microRNA in the inflamed colonic mucosa of patients with active ulcerative colitis. *J Gastroenterol Hepatol* 2010;**25 Suppl 1**:S129-33.

196 Fasseu M, Tréton X, Guichard C, Pedruzzi E, Cazals-Hatem D, Richard C, *et al.* Identification of restricted subsets of mature microRNA abnormally expressed in inactive colonic mucosa of patients with inflammatory bowel disease. *PLoS One* 2010;**5**:e13160.

197 Wu F, Zhang S, Dassopoulos T, Harris ML. Identification of MicroRNAs Associated with Ileal and Colonic Crohn's Disease. *Inflamm Bowel Dis* 2010;**16**:1729–38.

198 Brest P, Lapaquette P, Souidi M, Lebrigand K, Cesaro A, Vouret-Craviari V, *et al.* A

synonymous variant in IRGM alters a binding site for miR-196 and causes deregulation of IRGM-dependent xenophagy in Crohn's disease. *Nat Genet* 2011;**43**:242–5.

199 Wu F, Guo NJ, Tian H, Kwon JH. Peripheral blood microRNAs distinguish active ulcerative colitis and Crohn's disease. *Inflamm Bowel Dis* 2012;**17**:241–50.

200 Zahm AM, Thayu M, Hand NJ, Horner A, Leonard MB, Friedman JR. Circulating MicroRNA is a biomarker of pediatric Crohn disease. *J Pediatr Gastroenterol Nutr* 2011;**53**:26–33.

201 Archanioti P, Gazouli M, Theodoropoulos G, Vaiopoulou A, Nikiteas N. Micro-RNAs as regulators and possible diagnostic bio-markers in inflammatory bowel disease. *J Crohns Colitis* 2011;**5**:520–4.

202 Rakyan VK, Down T a, Balding DJ, Beck S. Epigenome-wide association studies for common human diseases. *Nat Rev Genet* 2011;**12**:529–41.

203 Drong a W, Lindgren CM, McCarthy MI. The Genetic and Epigenetic Basis of Type 2 Diabetes and Obesity. *Clin Pharmacol Ther* 2012;:1–9.

204 Liu Y, Aryee MJ, Padyukov L, Fallin MD, Hesselberg E, Runarsson A, *et al.* Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nat Biotechnol* 2013;**31**:142–7.

205 Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, *et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* 2012;**13**:86.

206 Liu Y, Aryee MJ, Padyukov L, Fallin MD, Hesselberg E, Runarsson A, *et al.* Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nat Biotechnol* 2013;**31**.

207 Bell CG, Finer S, Lindgren CM, Wilson GA, Rakyan VK, Teschendorff AE, *et al.* Integrated Genetic and Epigenetic Analysis Identifies Haplotype-Specific Methylation in the FTO Type 2 Diabetes and Obesity Susceptibility Locus. *PLoS One* 2010;**5**:e14040.

208 Toperoff G, Aran D, Kark JD, Rosenberg M, Dubnikov T, Nissan B, *et al.* Genome-wide survey reveals predisposing diabetes type 2-related DNA methylation variations in human peripheral blood. *Hum Mol Genet* 2012;**21**:371–83.

209 Tycko B. Allele-specific DNA methylation: beyond imprinting. *Hum Mol Genet* 2010;**19**:R210-20.

210 Byun H-M, Siegmund KD, Pan F, Weisenberger DJ, Kanel G, Laird PW, *et al.* Epigenetic profiling of somatic tissues from human autopsy specimens identifies tissue- and individual-specific DNA methylation patterns. *Hum Mol Genet* 2009;**18**:4808–17.

211 Korman B, Kastner DL, Gregersen PK, Remmers EF. STAT4: Genetics, Mechanisms, and Implications for Autoimmunity Review for Current Allergy and Asthma Reports. *Curr Allergy Asthma Rep* 2009;**8**:398–403.

212 Remmers EF, Plenge RM, Lee AT, Graham RR, Hom G, Behrens TW, *et al.* STAT4 and the risk of rheumatoid arthritis and systemic lupus erythematosus. *N Engl J Med* 2007;**357**:977–86.

213 Martínez a, Varadé J, Márquez a, Cénit MC, Espino L, Perdigones N, *et al.* Association of the STAT4 gene with increased susceptibility for some immune-mediated diseases. *Arthritis Rheum* 2008;**58**:2598–602.

214 Diaz-Gallo LM, Palomino-Morales RJ, Gómez-García M, Cardeña C, Rodrigo L, Nieto A, *et al.* STAT4 gene influences genetic predisposition to ulcerative colitis but not Crohn's disease in the Spanish population: a replication study. *Hum Immunol* 2010;**71**:515–9.

215 Glas J, Seiderer J, Nagy M, Fries C, Beigel F, Weidinger M, *et al.* Evidence for STAT4 as a common autoimmune gene: rs7574865 is associated with colonic Crohn's disease and early disease onset. *PLoS One* 2010;**5**:e10373.

216 Abelson a-K, Delgado-Vega a M, Kozyrev S V, Sánchez E, Velázquez-Cruz R, Eriksson N, *et al.* STAT4 associates with systemic lupus erythematosus through two independent effects that correlate with gene expression and act additively with IRF5 to increase risk. *Ann Rheum Dis* 2009;**68**:1746–53.

217 Kim SW, Kim ES, Moon CM, Kim T Il, Kim WH, Cheon JH. Abnormal genetic and epigenetic changes in signal transducer and activator of transcription 4 in the pathogenesis of inflammatory bowel diseases. *Dig Dis Sci* 2012;**57**:2600–7.

218 Shin H-J, Park H-Y, Jeong S-J, Park H-W, Kim Y-K, Cho S-H, *et al.* STAT4 expression in human T cells is regulated by DNA methylation but not by promoter polymorphism. *J Immunol* 2005;**175**:7143–50.

219 Zwiers A, Kraal L, van de Pouw Kraan TCTM, Wurdinger T, Bouma G, Kraal G. Cutting edge: a variant of the IL-23R gene associated with inflammatory bowel disease induces loss of microRNA regulation and enhanced protein production. *J Immunol* 2012;**188**:1573–7.

220 Brest P, Lapaquette P, Souidi M, Lebrigand K, Cesaro A, Vouret-Craviari V, *et al.* A synonymous variant in IRGM alters a binding site for miR-196 and causes deregulation of IRGM-dependent xenophagy in Crohn's disease. *Nat Genet* 2011;**43**:242–5.

221 Parkes M, Barrett JC, Prescott N, Tremelling M, Carl A, Fisher SA, *et al.* Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn disease susceptibility. *Nat Genet* 2009;**39**:830–2.

222 Okubo M, Tahara T, Shibata T, Yamashita H, Nakamura M, Yoshioka D, *et al.* Association study of common genetic variants in pre-microRNAs in patients with ulcerative colitis. *J Clin Immunol* 2011;**31**:69–73.

223 Atkinson A.J. J, Colburn W a., DeGruttola VG, DeMets DL, Downing GJ, Hoth DF, *et al.* Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework. *Clin Pharmacol Ther* 2001;**69**:89–95.

224 Rogler G, Biedermann L. Clinical Utility of Biomarkers in IBD. *Curr Gastroenterol Rep* 2015;**17**:26.

225 Sands BE. Biomarkers of Inflammation in Inflammatory Bowel Disease. *Gastroenterology* 2015;**149**:1275–1285.e2.

226 Saibeni S, Folli C, de Franchis R, Borsi G, Vecchi M. Diagnostic role and clinical correlates of anti-Saccharomyces cerevisiae antibodies (ASCA) and anti-neutrophil cytoplasmic antibodies (p-ANCA) in Italian patients with inflammatory bowel diseases. *Dig Liver Dis* 2003;**35**:862–8.

227 Gerich ME, McGovern DPB. Towards personalized care in IBD. *Nat Rev Gastroenterol Hepatol* 2014;**11**:287–99.

228 Panaccione R, Sandborn WJ. Is antibody testing for inflammatory bowel disease clinically useful? *Gastroenterology* 1999;**116**:1001-2-3.

229 Quinton JF, Sendid B, Reumaux D, Duthilleul P, Cortot A, Grandbastien B, *et al.* Anti-Saccharomyces cerevisiae mannan antibodies combined with antineutrophil cytoplasmic autoantibodies in inflammatory bowel disease: prevalence and diagnostic role. *Gut* 1998;**42**:788–91.

230 Kennedy NA, Clark A, Walkden A, Chang JCW, Fascí-Spurio F, Muscat M, *et al.* Clinical utility and diagnostic accuracy of faecal calprotectin for IBD at first presentation to gastroenterology services in adults aged 16-50 years. *J Crohns Colitis* 2015;**9**:41–9.

231 D'Haens G, Ferrante M, Vermeire S, Baert F, Noman M, Moortgat L, *et al.* Fecal calprotectin is a surrogate marker for endoscopic lesions in inflammatory bowel disease. *Inflamm Bowel Dis* 2012;**18**:2218–24.

232 van Rheenen PF, Van de Vijver E, Fidler V. Faecal calprotectin for screening of patients with

suspected inflammatory bowel disease: diagnostic meta-analysis. *BMJ* 2010;**341**:c3369.

233 Yamamoto T, Shiraki M, Bamba T, Umegae S, Matsumoto K. Faecal calprotectin and lactoferrin as markers for monitoring disease activity and predicting clinical recurrence in patients with Crohn's disease after ileocolonic resection: A prospective pilot study. *United Eur Gastroenterol J* 2013;**1**:368–74.

234 Henriksen M, Jahnsen J, Lygren I, Stray N, Sauar J, Vatn MH, *et al.* C-reactive protein: a predictive factor and marker of inflammation in inflammatory bowel disease. Results from a prospective population-based study. *Gut* 2008;**57**:1518–23.

235 Niewiadomski O, Studd C, Hair C, Wilson J, Ding NS, Heerasing N, *et al.* Prospective population-based cohort of inflammatory bowel disease in the biologics era: Disease course and predictors of severity. *J Gastroenterol Hepatol* 2015;**30**:1346–53.

236 Amre DK, Lu S-E, Costea F, Seidman EG. Utility of serological markers in predicting the early occurrence of complications and surgery in pediatric Crohn's disease patients. *Am J Gastroenterol* 2006;**101**:645–52.

237 Dubinsky MC, Lin Y-C, Dutridge D, Picornell Y, Landers CJ, Farrior S, *et al.* Serum immune responses predict rapid disease progression among children with Crohn's disease: immune responses predict disease progression. *Am J Gastroenterol* 2006;**101**:360–7.

238 Vasiliauskas EA, Kam LY, Karp LC, Gaiennie J, Yang H, Targan SR. Marker antibody expression stratifies Crohn's disease into immunologically homogeneous subgroups with distinct clinical characteristics. *Gut* 2000;**47**:487–96.

239 Mow WS, Vasiliauskas EA, Lin Y-C, Fleshner PR, Papadakis KA, Taylor KD, *et al.* Association of antibody responses to microbial antigens and complications of small bowel Crohn's disease. *Gastroenterology* 2004;**126**:414–24.

240 Ferrante M, Henckaerts L, Joossens M, Pierik M, Joossens S, Dotan N, *et al.* New serological markers in inflammatory bowel disease are associated with complicated disease behaviour. *Gut* 2007;**56**:1394–403.

241 Travis SPL, Farrant JM, Ricketts C, Nolan DJ, Mortensen NM, Kettlewell MGW, *et al.* Predicting outcome in severe ulcerative colitis. *Gut* 1996;**38**:905–10.

242 Ho GT, Mowat C, Goddard CJR, Fennell JM, Shah NB, Prescott RJ, *et al.* Predicting the outcome of severe ulcerative colitis: development of a novel risk score to aid early selection of patients for second-line medical therapy or surgery. *Aliment Pharmacol Ther* 2004;**19**:1079–87.

243 Molander P, Färkkilä M, Ristimäki A, Salminen K, Kemppainen H, Blomster T, *et al.* Does fecal calprotectin predict short-term relapse after stopping TNFα-blocking agents in inflammatory bowel disease patients in deep remission? *J Crohns Colitis* 2015;**9**:33–40.

244 Mooiweer E, Severs M, Schipper MEI, Fidder HH, Siersema PD, Laheij RJF, *et al.* Low fecal calprotectin predicts sustained clinical remission in inflammatory bowel disease patients: a plea for deep remission. *J Crohns Colitis* 2015;**9**:50–5.

245 Osterman MT, Aberra FN, Cross R, Liakos S, McCabe R, Shafran I, *et al.* Mesalamine dose escalation reduces fecal calprotectin in patients with quiescent ulcerative colitis. *Clin Gastroenterol Hepatol* 2014;**12**:1887–93.e3.

246 Lasson A, Strid H, Ohman L, Isaksson S, Olsson M, Rydström B, *et al.* Fecal calprotectin one year after ileocaecal resection for Crohn's disease--a comparison with findings at ileocolonoscopy. *J Crohns Colitis* 2014;**8**:789–95.

247 Ho GT, Lee HM, Brydon G, Ting T, Hare N, Drummond H, *et al.* Fecal calprotectin predicts the clinical course of acute severe ulcerative colitis. *Am J Gastroenterol* 2009;**104**:673–8.

248 Kennedy N, Van Ross JE, Hare NC, Ho G-T, Drummond HE, Shand AG, *et al.* Acute severe ulcerative colitis: the last 12 years in Edinburgh. *Gut* 2012;**61**:A235–6.

249 Annese V, Lombardi G, Perri F, D'Incà R, Ardizzone S, Riegler G, *et al.* Variants of CARD15 are associated with an aggressive clinical course of Crohn's disease--an IG-IBD study. *Am J Gastroenterol* 2005;**100**:84–92.

250 Brant SR, Picco MF, Achkar J-P, Bayless TM, Kane S V, Brzezinski A, *et al.* Defining complex contributions of NOD2/CARD15 gene mutations, age at onset, and tobacco use on Crohn's disease phenotypes. *Inflamm Bowel Dis* 2003;**9**:281–9.

251 Economou M, Trikalinos T a, Loizou KT, Tsianos E V, Ioannidis JPA. Differential effects of NOD2 variants on Crohn's disease risk and phenotype in diverse populations: a metaanalysis. *Am J Gastroenterol* 2004;**99**:2393–404.

252 Adler J, Rangwalla SC, Dwamena BA, Higgins PDR. The prognostic power of the NOD2 genotype for complicated Crohn's disease: a meta-analysis. *Am J Gastroenterol* 2011;**106**:699–712.

253 Weersma RK, Stokkers PCF, van Bodegraven AA, van Hogezand RA, Verspaget HW, de Jong DJ, *et al.* Molecular prediction of disease risk and severity in a large Dutch Crohn's disease cohort. *Gut* 2009;**58**:388–95.

254 Cleynen I, González JR, Figueroa C, Franke A, McGovern D, Bortlík M, *et al.* Genetic factors conferring an increased susceptibility to develop Crohn's disease also influence disease phenotype: results from the IBDchip European Project. *Gut* 2013;**62**:1556–65.

255 Satsangi J, Welsh KI, Bunce M, Julier C, Farrant JM, Bell JI, *et al.* Contribution of genes of the major histocompatibility complex to susceptibility and disease phenotype in inflammatory bowel disease. *Lancet* 1996;**347**:1212–7.

256 Roussomoustakaki M, Satsangi J, Welsh K, Louis E, Fanning G, Targan S, *et al.* Genetic markers may predict disease behavior in patients with ulcerative colitis. *Gastroenterology* 1997;**112**:1845–53.

257 Ahmad T, Armuzzi A, Neville M, Bunce M, Ling K-L, Welsh KI, *et al.* The contribution of human leucocyte antigen complex genes to disease phenotype in ulcerative colitis. *Tissue Antigens* 2003;**62**:527–35.

258 Ho G-T, Soranzo N, Nimmo ER, Tenesa A, Goldstein DB, Satsangi J. ABCB1/MDR1 gene determines susceptibility and phenotype in ulcerative colitis: discrimination of critical variants using a gene-wide haplotype tagging approach. *Hum Mol Genet* 2006;**15**:797–805.

259 Haritunians T, Taylor KD, Targan SR, Dubinsky M, Ippoliti A, Kwon S, *et al.* Genetic predictors of medically refractory ulcerative colitis. *Inflamm Bowel Dis* 2010;**16**:1830–40.

260 Lee JC, Lyons PA, McKinney EF, Sowerby JM, Carr EJ, Bredin F, *et al.* Gene expression profiling of CD8+ T cells predicts prognosis in patients with Crohn disease and ulcerative colitis. *J Clin Invest* 2011;**121**:4170–9.

261 Kabakchiev B, Turner D, Hyams J, Mack D, Leleiko N, Crandall W, *et al.* Gene expression changes associated with resistance to intravenous corticosteroid therapy in children with severe ulcerative colitis. *PLoS One* 2010;**5**:1–8.

262 Gevers D, Kugathasan S, Denson LA, Vázquez-Baeza Y, Van Treuren W, Ren B, *et al.* The treatment-naive microbiome in new-onset Crohn's disease. *Cell Host Microbe* 2014;**15**:382–92.

263 Michail S, Durbin M, Turner D, Griffiths AM, Mack DR, Hyams J, *et al.* Alterations in the gut microbiome of children with severe ulcerative colitis. *Inflamm Bowel Dis* 2012;**18**:1799–808.

264 Ooi M, Nishiumi S, Yoshie T, Shiomi Y, Kohashi M, Fukunaga K, *et al.* GC/MS-based profiling of amino acids and TCA cycle-related molecules in ulcerative colitis. *Inflamm Res* 2011;**60**:831–40.

265 Schicho R, Shaykhutdinov R, Ngo J, Nazyrova A, Schneider C, Panaccione R, *et al.*

Quantitative metabolomic profiling of serum, plasma, and urine by (1)H NMR spectroscopy discriminates between patients with inflammatory bowel disease and healthy individuals. *J Proteome Res* 2012;**11**:3344–57.

266 Stephens NS, Siffledeen J, Su X, Murdoch TB, Fedorak RN, Slupsky CM. Urinary NMR metabolomic profiles discriminate inflammatory bowel disease from healthy. *J Crohns Colitis* 2013;**7**:e42-8.

267 Miyahara K, Nouso K, Saito S, Hiraoka S, Harada K, Takahashi S, *et al.* Serum glycan markers for evaluation of disease activity and prediction of clinical course in patients with ulcerative colitis. *PLoS One* 2013;**8**:e74861.

268 Theodoratou E, Campbell H, Ventham NT, Kolarich D, Pučić-Baković M, Zoldoš V, *et al.* The role of glycosylation in IBD. *Nat Rev Gastroenterol Hepatol* 2014;**11**:588–600.

269 Lichtenstein GR, Targan SR, Dubinsky MC, Rotter JI, Barken DM, Princen F, *et al.* Combination of genetic and quantitative serological immune markers are associated with complicated Crohn's disease behavior. *Inflamm Bowel Dis* 2011;**17**:2488–96.

270 Dubinsky MC, Kugathasan S, Kwon S, Haritunians T, Wrobel I, Wahbeh G, *et al.* Multidimensional prognostic risk assessment identifies association between IL12B variation and surgery in Crohn's disease. *Inflamm Bowel Dis* 2013;**19**:1662–70.

271 Lennard-Jones JE. Classification of inflammatory bowel disease. *Scand J Gastroenterol Suppl* 1989;**170**:2-6-9.

272 Sharratt CL, Gilbert CJ, Cornes MC, Ford C, Gama R. EDTA sample contamination is common and often undetected, putting patients at unnecessary risk of harm. *Int J Clin Pract* 2009;**63**:1259–62.

273 Duale N, Brunborg G, Rønningen KS, Briese T, Aarem J, Aas KK, *et al.* Human blood RNA stabilization in samples collected and transported for a large biobank. *BMC Res Notes* 2012;**5**:510.

274 Greiner Bio-One. Data Sheet: Isolation of mononuclear cells from human peripheral blood by density gradient centrifugation. 2008.

275 Böyum A. Isolation of mononuclear cells and granulocytes from human blood. Isolation of monuclear cells by one centrifugation, and of granulocytes by combining centrifugation and sedimentation at 1 g. *Scand J Clin Lab Invest Suppl* 1968;**97**:77–89.

276 Böyum A. Isolation of leucocytes from human blood. Further observations. Methylcellulose, dextran, and ficoll as erythrocyteaggregating agents. *Scand J Clin Lab Invest Suppl* 1968;**97**:31–50.

277 Bøyum A. Isolation of lymphocytes, granulocytes and macrophages. *Scand J Immunol* 1976;**Suppl 5**:9–15.

278 Hansen R, Russell RK, Reiff C, Louis P, McIntosh F, Berry SH, *et al.* Microbiota of de-novo pediatric IBD: increased Faecalibacterium prausnitzii and reduced bacterial diversity in Crohn's but not in ulcerative colitis. *Am J Gastroenterol* 2012;**107**:1913–22.

279 Yuan W, Xia Y, Bell CG, Yet I, Ferreira T, Ward KJ, *et al.* An integrated epigenomic analysis for type 2 diabetes susceptibility loci in monozygotic twins. *Nat Commun* 2014;**5**:5719.

280 Dick KJ, Nelson CP, Tsaprouni L, Sandling JK, Aïssi D, Wahl S, *et al.* DNA methylation and body-mass index: a genome-wide analysis. *Lancet* 2014;**383**:1990–8.

281 Callaway E. Epigenomics starts to make its mark. *Nature* 2014;**508**:22.

282 Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet* 2012;**13**:484–92.

283 Jaffe AE, Irizarry RA. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biol* 2014;**15**:R31.

284 Adams AT, Kennedy NA, Hansen R, Ventham NT, O'Leary KR, Drummond HE, *et al.* Two-stage genome-wide methylation profiling in childhood-onset Crohn's Disease implicates epigenetic alterations at the VMP1/MIR21 and HLA loci. *Inflamm Bowel Dis* 2014;**20**:1784–93.

285 Du P, Kibbe WA, Lin SM. lumi: a pipeline for processing Illumina microarray. *Bioinformatics* 2008;**24**:1547–8.

286 Du P, Huang S, Kibbe WA, Lin S. Analyze Illumina Infinium methylation microarray data Major classes of Illumina methylation microarray data. 2014.

287 Hansen K, Aryee M, Iriizarry R. minfi and shinyMethyl: a winning pair of R-packages for the analysis of methylation data. *Tutor BioC* 2013.

288 Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* 2012;**491**:56–65.

289 Marabita F, Almgren M, Lindholm M, Ruhrmann S, Fagerström-Billai F, Jagodic M, *et al.* An evaluation of analysis pipelines for DNA methylation profiling using the Illumina HumanMethylation450 BeadChip platform. *Epigenetics* 2013;**8**.

290 Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 2003;**19**:185–93.

291 Teschendorff AE, Marabita F, Lechner M, Bartlett T, Tegner J, Gomez-Cabrero D, *et al.* A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics* 2013;**29**:189–96.

292 Schalkwyk ALC, Pidsley R, Wong CCY, Touleimat N, Defrance M, Teschendorff A, *et al.* Package ' wateRmelon '. 2014.

293 Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 2007;**8**:118–27.

294 Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* 2012;**28**:882–3.

295 Gower J. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* 1966;**53**:325–328.

296 Gentleman R, Carey V, Dudoit S, Irizarry R, Huber W. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer, New York 2005.

297 Triche T. IlluminaHumanMethylation450k.db: Illumina Human Methylation 450k annotation data. R Packag. version 2.0.7. 2012.

298 Du P, Zhang X, Huang C-C, Jafari N, Kibbe WA, Hou L, *et al.* Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics* 2010;**11**:587.

299 Holm S. A simple sequentially rejective multiple test procedure. *Scand J Stat* 1979;**6**:65–70.

300 Morris TJ, Butcher LM, Feber A, Teschendorff AE, Chakravarthy AR, Wojdacz TK, *et al.* ChAMP: 450k Chip Analysis Methylation Pipeline. *Bioinformatics* 2014;**30**:428–30.

301 Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B* 1995;**57**:289–300.

302 Houseman EA, Molitor J, Marsit CJ. Reference-free cell mixture adjustments in analysis of DNA methylation data. *Bioinformatics* 2014;**30**:1431–9.

303 Reinius LE, Acevedo N, Joerink M, Pershagen G, Dahlén S-E, Greco D, *et al.* Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. *PLoS One* 2012;**7**:e41361.

304 Hansen KD, Ayree M, Irizarry RA, Jaffe AE, Maksimovic J, Houseman EA. Package ' minfi '. 2014.

305 Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, *et al.* Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* 2014;**30**:1363–9.

306 Young MD, Wakefield MJ SG and OA. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol* 2010;**11**:14.

307 McDermott E, Ryan EJ, Tosetto M, Gibson D, Burrage J, Keegan D, *et al.* DNA methylation profiling in inflammatory bowel disease provides new insights into disease pathogenesis. *J Crohns Colitis* 2015;**10**:77–86.

308 Oliveros J. Venny. An interactive tool for comparing lists with Venn's diagrams. ;:http://bioinfogp.cnb.csic.es/tools/venny/index.htm.

309 Calvo-Garrido J, Carilla-Latorre S, Escalante R. Vacuole membrane protein 1, autophagy and much more. *Autophagy* 2008;**4**:835–7.

310 Molejon MI, Ropolo A, Re A Lo, Boggio V, Vaccaro MI. The VMP1-Beclin 1 interaction regulates autophagy induction. *Sci Rep* 2013;**3**:1055.

311 Kang R, Zeh HJ, Lotze MT, Tang D. The Beclin 1 network regulates autophagy and apoptosis. *Cell Death Differ* 2011;**18**:571–80.

312 Ribas J, Ni X, Castanares M, Liu MM, Esopi D, Yegnasubramanian S, *et al.* A novel source for miR-21 expression through the alternative polyadenylation of VMP1 gene transcripts. *Nucleic Acids Res* 2012;**40**:6821–33.

313 Kanaan Z, Rai SN, Eichenberger MR, Roberts H, Keskey B, Pan J, *et al.* Plasma miR-21: a potential diagnostic marker of colorectal cancer. *Ann Surg* 2012;**256**:544–51.

314 Kalla R, Ventham NT, Kennedy N a., Quintana JF, Nimmo ER, Buck a. H, *et al.* MicroRNAs: new players in IBD. *Gut* 2015;**64**:504–17.

315 Shi C, Liang Y, Yang J, Xia Y, Chen H, Han H, *et al.* MicroRNA-21 knockout improve the survival rate in DSS induced fatal colitis through protecting against inflammation and tissue injury. *PLoS One* 2013;**8**:e66814.

316 Wu F, Dong F, Arendovich N, Zhang J, Huang Y, Kwon JH. Divergent influence of microRNA-21 deletion on murine colitis phenotypes. *Inflamm Bowel Dis* 2014;**20**:1972–85.

317 Koshizuka Y, Ikegawa S, Sano M, Nakamura K, Nakamura Y. Isolation, characterization, and mapping of the mouse and human WDR8 genes, members of a novel WD-repeat gene family. *Genomics* 2001;**72**:252–9.

318 Tan S. The leucocyte β2 (CD18) integrins: the structure, functional regulation and signalling properties. Biosci. Rep. 2012;**32**:241–69.

319 Kishimoto TK, Hollander N, Roberts TM, Anderson DC, Springer TA. Heterogeneous mutations in the beta subunit common to the LFA-1, Mac-1, and p150,95 glycoproteins cause leukocyte adhesion deficiency. *Cell* 1987;**50**:193–202.

320 van de Vijver E, van den Berg TK, Kuijpers TW. Leukocyte Adhesion Deficiencies. Hematol. Oncol. Clin. North Am. 2013;**27**:101–16.

321 Harris RA, Nagy-Szakal D, Mir SA V, Frank E, Szigeti R, Kaplan JL, *et al.* DNA methylation-associated colonic mucosal immune and defense responses in treatment-naïve pediatric ulcerative colitis. *Epigenetics* 2014;**9**:1131–7.

322 Hutterer E, Asslaber D, Caldana C, Krenn PW, Zucchetto A, Gattei V, *et al.* CD18 (ITGB2) expression in chronic lymphocytic leukaemia is regulated by DNA methylation-dependent and -independent mechanisms. *Br J Haematol* 2015;**169**:286–9.

323 Nardone S, Sams DS, Reuveni E, Getselter D, Oron O, Karpuj M, *et al.* DNA methylation

analysis of the autistic brain reveals multiple dysregulated biological pathways. *Transl Psychiatry* 2014;**4**:e433.

324 Haire RN, Ohta Y, Lewis JE, Fu SM, Kroisel P, Litman GW. TXK, a novel human tyrosine kinase expressed in T cells shares sequence identity with Tec family kinases and maps to 4p12. *Hum Mol Genet* 1994;**3**:897–901.

325 Gomez-Rodriguez J, Kraus ZJ, Schwartzberg PL. Tec family kinases Itk and Rlk/Txk in T lymphocytes: Cross-regulation of cytokine production and T-cell fates. FEBS J. 2011;**278**:1980–9.

326 Nagafuchi H, Takeno M, Yoshikawa H, Kurokawa MS, Nara K, Takada E, *et al.* Excessive expression of Txk, a member of the Tec family of tyrosine kinases, contributes to excessive Th1 cytokine production by T lymphocytes in patients with Behcet's disease. *Clin Exp Immunol* 2005;**139**:363–70.

327 Suzuki N, Nara K, Suzuki T. Skewed Th1 responses caused by excessive expression of Txk, a member of the Tec family of tyrosine kinases, in patients with Behcet's disease. Clin. Med. Res. 2006;**4**:147–51.

328 Rönn T, Volkov P, Davegårdh C, Dayeh T, Hall E, Olsson AH, *et al.* A Six Months Exercise Intervention Influences the Genome-wide DNA Methylation Pattern in Human Adipose Tissue. *PLoS Genet* 2013;**9**.

329 Benton MC, Johnstone A, Eccles D, Harmon B, Hayes MT, Lea RA, *et al.* An analysis of DNA methylation in human adipose tissue reveals differential modification of obesity genes before and after gastric bypass and weight loss. *Genome Biol* 2015;**16**:8.

330 Maller JL, Foulkes JG, Erikson E, Baltimore D. Phosphorylation of ribosomal protein S6 on serine after microinjection of the Abelson murine leukemia virus tyrosine-specific protein kinase into Xenopus oocytes. *Proc Natl Acad Sci U S A* 1985;**82**:272–6.

331 Anjum R, Blenis J. The RSK family of kinases: emerging roles in cellular signalling. *Nat Rev Mol Cell Biol* 2008;**9**:747–58.

332 Pancholi S, Lykkesfeldt AE, Hilmi C, Banerjee S, Leary A, Drury S, *et al.* ERBB2 influences the subcellular localization of the estrogen receptor in tamoxifen-resistant MCF-7 cells leading to the activation of AKT and RPS6KA2. *Endocr Relat Cancer* 2008;**15**:985–1002.

333 Bignone PA, Lee KY, Liu Y, Emilion G, Finch J, Soosay AER, *et al.* RPS6KA2, a putative tumour suppressor gene at 6q27 in sporadic epithelial ovarian cancer. *Oncogene* 2007;**26**:683–700.

334 El Kasmi KC, Smith AM, Williams L, Neale G, Panopolous A, Watowich SS, *et al.* Cutting Edge: A Transcriptional Repressor and Corepressor Induced by the STAT3-Regulated Anti-Inflammatory Signaling Pathway. *J Immunol* 2007;**179**:7215–9.

335 Wiley SR, Schooley K, Smolak PJ, Din WS, Huang CP, Nicholl JK, *et al.* Identification and characterization of a new member of the TNF family that induces apoptosis. *Immunity* 1995;**3**:673–82.

336 Almasan A, Ashkenazi A. Apo2L/TRAIL: Apoptosis signaling, biology, and potential for cancer therapy. Cytokine Growth Factor Rev. 2003;**14**:337–48.

337 Johnstone RW, Frew AJ, Smyth MJ. The TRAIL apoptotic pathway in cancer onset, progression and therapy. *Nat Rev Cancer* 2008;**8**:782–98.

338 Begue B, Wajant H, Bambou JC, Dubuquoy L, Siegmund D, Beaulieu J, *et al.* Implication of TNF-Related Apoptosis-Inducing Ligand in Inflammatory Intestinal Epithelial Lesions. *Gastroenterology* 2006;**130**:1962–74.

339 Brost S, Koschny R, Sykora J, Stremmel W, Lasitschka F, Walczak H, *et al.* Differential expression of the TRAIL/TRAIL-receptor system in patients with inflammatory bowel disease. *Pathol Res Pract* 2010;**206**:43–50.

340 McKeithan TW, Ohno H, Dickstein J, Hume E. Genomic structure of the candidate proto-oncogene BCL3. *Genomics* 1994;**24**:120–6.

341 Román J, Planell N, Lozano JJ, Aceituno M, Esteller M, Pontes C, *et al.* Evaluation of responsive gene expression as a sensitive and specific biomarker in patients with ulcerative colitis. *Inflamm Bowel Dis* 2013;**19**:221–9.

342 Abdelbaqi M, Chidlow JH, Matthews KM, Pavlick KP, Barlow SC, Linscott AJ, *et al.* Regulation of dextran sodium sulfate induced colitis by leukocyte beta 2 integrins. *Lab Invest* 2006;**86**:380–90.

343 Palmen MJ, Dijkstra CD, van der Ende MB, Peña AS, van Rees EP. Anti-CD11b/CD18 antibodies reduce inflammation in acute colitis in rats. *Clin Exp Immunol* 1995;**101**:351–6.

344 Uzel G, Kleiner DE, Kuhns DB, Holland SM. Dysfunctional LAD-1 neutrophils and colitis. *Gastroenterology* 2001;**121**:958–64.

345 Granlund A van B, Flatberg A, Østvik AE, Drozdov I, Gustafsson BI, Kidd M, *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* 2013;**8**:e56818.

346 Kellermayer R. Hurdles for epigenetic disease associations from peripheral blood leukocytes. *Inflamm Bowel Dis* 2013;**19**:E66-7.

347 Murphy TM, Mill J. Epigenetics in health and disease: heralding the EWAS era. *Lancet* 2014;**383**:1952–4.

348 Jenke AC, Postberg J, Raine T, Nayak KM, Molitor M, Wirth S, *et al.* DNA methylation analysis in the intestinal epithelium-effect of cell separation on gene expression and methylation profile. *PLoS One* 2013;**8**:e55636.

349 Hawkey C. OC-007 Haemopoetic Stem Cell Transplantation For Severe Resistant Crohn's Disease: Preliminary Evidence For Durable Benefit. *Gut* 2014;**63 Suppl 1**:A4.

350 Hawkey CJ. Stem cells as treatment in inflammatory bowel disease. *Dig Dis* 2012;**30 Suppl 3**:134–9.

351 Kaila B, Grant D, Pettigrew N, Greenberg H, Bernstein CN. Crohn's Disease Recurrence in a Small Bowel Transplant. *Am J Gastroenterol* 2004;**99**:158–62.

352 Church TR, Wandell M, Lofton-Day C, Mongin SJ, Burger M, Payne SR, *et al.* Prospective evaluation of methylated SEPT9 in plasma for detection of asymptomatic colorectal cancer. *Gut* 2013;**304149**:1–9.

353 Grützmann R, Molnar B, Pilarsky C, Habermann JK, Schlag PM, Saeger HD, *et al.* Sensitive detection of colorectal cancer in peripheral blood by septin 9 DNA methylation assay. *PLoS One* 2008;**3**:e3759.

354 Chen C, Grennan K, Badner J, Zhang D, Gershon E, Jin L, *et al.* Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. *PLoS One* 2011;**6**:e17238.

355 Horvath S, Erhart W, Brosch M, Ammerpohl O, von Schönfels W, Ahrens M, *et al.* Obesity accelerates epigenetic aging of human liver. *Proc Natl Acad Sci U S A* 2014;**111**:15538–43.

356 Horvath S, Garagnani P, Bacalini MG, Pirazzini C, Salvioli S, Gentilini D, *et al.* Accelerated epigenetic aging in Down syndrome. *Aging Cell* 2015;**14**:491–5.

357 Horvath S, Levine AJ. HIV-1 Infection Accelerates Age According to the Epigenetic Clock. *J Infect Dis* 2015;**212**:1563–73.

358 Horvath S. Erratum to: DNA methylation age of human tissues and cell types. *Genome Biol* 2015;**16**:96.

359 Horvath S, Mah V, Lu AT, Woo JS, Choi O-W, Jasinska AJ, *et al.* The cerebellum ages slowly

according to the epigenetic clock. *Aging (Albany NY)* 2015;**7**:294–306.

360 Kraft P, Zeggini E, Ioannidis JPA. Replication in genome-wide association studies. *Stat Sci* 2009;**24**:561–73.

361 Wilhelm-Benartzi CS, Koestler DC, Karagas MR, Flanagan JM, Christensen BC, Kelsey KT, *et al.* Review of processing and analysis methods for DNA methylation array data. *Br J Cancer* 2013;**109**:1394–402.

362 Roessler J, Ammerpohl O, Gutwein J, Hasemeier B, Anwar S, Kreipe H, *et al.* Quantitative cross-validation and content analysis of the 450k DNA methylation array from Illumina, Inc. BMC Res. Notes. 2012;**5**:210.

363 Burisch J, Pedersen N, Čuković-Čavka S, Brinar M, Kaimakliotis I, Duricova D, *et al.* East-West gradient in the incidence of inflammatory bowel disease in Europe: the ECCO-EpiCom inception cohort. *Gut* 2014;**63**:588–97.

364 Dixon AL, Liang L, Moffatt MF, Chen W, Heath S, Wong KCC, *et al.* A genome-wide association study of global gene expression. *Nat Genet* 2007;**39**:1202–7.

365 Kerkel K, Spadola A, Yuan E, Kosek J, Jiang L, Hod E, *et al.* Genomic surveys by methylation-sensitive SNP analysis identify sequence-dependent allele-specific DNA methylation. *Nat Genet* 2008;**40**:904–8.

366 Grundberg E, Meduri E, Sandling JK, Hedman AK, Keildson S, Buil A, *et al.* Global analysis of DNA methylation variation in adipose tissue from twins reveals links to disease-associated variants in distal regulatory elements. *Am J Hum Genet* 2013;**93**:876–90.

367 Heyn H, Sayols S, Moutinho C, Vidal E, Sanchez-Mut J V, Stefansson OA, *et al.* Linkage of DNA methylation quantitative trait loci to human cancer risk. *Cell Rep* 2014;**7**:331–8.

368 Gamazon ER, Badner JA, Cheng L, Zhang C, Zhang D, Cox NJ, *et al.* Enrichment of cis-regulatory gene expression SNPs and methylation quantitative trait loci among bipolar disorder susceptibility variants. *Mol Psychiatry* 2013;**18**:340–6.

369 Rushton MD, Reynard LN, Young DA, Aubourg G, Gee F, Darlay R, *et al.* Methylation quantitative trait locus (meQTL) analysis of osteoarthritis links epigenetics with genetic risk. 2015;**24**:7432–44.

370 Shah TS, Liu JZ, Floyd JAB, Morris JA, Wirth N, Barrett JC, *et al.* OptiCall: A robust genotype-calling algorithm for rare, low-frequency and common variants. *Bioinformatics* 2012;**28**:1598–603.

371 Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007;**81**:559–75.

372 Shabalin AA. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* 2012;**28**:1353–8.

373 Millstein J, Zhang B, Zhu J, Schadt EE. Disentangling molecular relationships with a causal inference test. *BMC Genet* 2009;**10**:23.

374 Hong X, Hao K, Ladd-Acosta C, Hansen KD, Tsai H-J, Liu X, *et al.* Genome-wide association study identifies peanut allergy-specific loci and evidence of epigenetic mediation in US children. *Nat Commun* 2015;**6**:6304.

375 Johnson AD, Handsaker RE, Pulit SL, Nizzari MM, O'Donnell CJ, de Bakker PIW. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics* 2008;**24**:2938–9.

376 van Eijk KR, de Jong S, Boks MP, Langeveld T, Colas F, Veldink JH, *et al.* Genetic analysis of DNA methylation and gene expression levels in whole blood of healthy human subjects. *BMC Genomics* 2012;**13**:636.

377 Gibbs JR, van der Brug MP, Hernandez DG, Traynor BJ, Nalls MA, Lai S-L, *et al.* Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. *PLoS Genet* 2010;**6**:e1000952.

378 Peirce JL, Li H, Wang J, Manly KF, Hitzemann RJ, Belknap JK, *et al.* How replicable are mRNA expression QTL? *Mamm Genome* 2006;**17**:643–56.

379 de Jong S, van Eijk KR, Zeegers DWLH, Strengman E, Janson E, Veldink JH, *et al.* Expression QTL analysis of top loci from GWAS meta-analysis highlights additional schizophrenia candidate genes. *Eur J Hum Genet* 2012;**20**:1004–8.

380 Razin A, Cedar H. DNA methylation and gene expression. *Microbiol Rev* 1991;**55**:451–8.

381 Riggs AD. X inactivation, differentiation, and DNA methylation. *Cytogenet Cell Genet* 1975;**14**:9–25.

382 Kriaucionis S, Bird A. DNA methylation and Rett syndrome. *Hum Mol Genet* 2003;**12 Spec No**:R221–7.

383 Bird A. DNA methylation patterns and epigenetic memory DNA methylation patterns and epigenetic memory. *Genes Dev* 2002;:6–21.

384 Razin a, Cedar H. Distribution of 5-methylcytosine in chromatin. *Proc Natl Acad Sci U S A* 1977;**74**:2725–8.

385 van der Ploeg LH, Flavell RA. DNA methylation in the human gamma delta beta-globin locus in erythroid and nonerythroid tissues. *Cell* 1980;**19**:947–58.

386 Lyst MJ, Bird A. Rett syndrome: a complex disorder with simple roots. *Nat Rev Genet* 2015;**16**:261–75.

387 Farthing CR, Ficz G, Ng RK, Chan C-F, Andrews S, Dean W, *et al.* Global mapping of DNA methylation in mouse promoters reveals epigenetic reprogramming of pluripotency genes. *PLoS Genet* 2008;**4**:e1000116.

388 Han H, Cortez CC, Yang X, Nichols PW, Jones PA, Liang G. DNA methylation directly silences genes with non-CpG island promoters and establishes a nucleosome occupied promoter. *Hum Mol Genet* 2011;**20**:4299–310.

389 Venolia L, Gartler SM. Comparison of transformation efficiency of human active and inactive X-chromosomal DNA. *Nature* 1983;**302**:82–3.

390 Lock LF, Takagi N, Martin GR. Methylation of the Hprt gene on the inactive X occurs after chromosome inactivation. *Cell* 1987;**48**:39–46.

391 Maunakea AK, Nagarajan RP, Bilenky M, Ballinger TJ, D'Souza C, Fouse SD, *et al.* Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature* 2010;**466**:253–7.

392 Jiao Y, Widschwendter M, Teschendorff AE. A systems-level integrative framework for genome-wide DNA methylation and gene expression data identifies differential gene expression modules under epigenetic control. *Bioinformatics* 2014;**30**:2360–6.

393 Cerami EG, Gross BE, Demir E, Rodchenkov I, Babur Ö, Anwar N, *et al.* Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res* 2011;**39**.

394 Jones A, Teschendorff AE, Li Q, Hayward JD, Kannan A, Mould T, *et al.* Role of DNA Methylation and Epigenetic Silencing of HAND2 in Endometrial Cancer Development. *PLoS Med* 2013;**10**.

395 West J, Beck S, Wang X, Teschendorff AE. An integrative network algorithm identifies age-associated differential methylation interactome hotspots targeting stem-cell differentiation pathways. *Sci Rep* 2013;**3**:1630.

396 Smyth GK. Limma: linear models for microarray data. In: Gentleman R, Carey V, Dudoit S, *et al.*, eds. *Bioinformatics and Computational Biology Solutions Using {R} and Bioconductor*.

New York: : Springer 2005. 397–420.

397 Ledderose C, Heyn J, Limbeck E, Kreth S. Selection of reliable reference genes for quantitative real-time PCR in human T cells and neutrophils. *BMC Res Notes* 2011;**4**:427.

398 Promega. GoTaq pCR Master Mix Technical Manual (TM318). 2014;:1–5.

399 Perkins JR, Dawes JM, McMahon SB, Bennett DL, Orengo C, Kohl M. ReadqPCR and NormqPCR: R packages for the reading, quality checking and normalisation of RT-qPCR quantification cycle (Cq) data. *BMC Genomics* 2012;**13**:296.

400 Vandesompele J, De Preter K, Pattyn F, Poppe B, Van Roy N, De Paepe A, *et al.* Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol* 2002;**3**:RESEARCH0034.

401 Andersen CL, Jensen JL, Ørntoft TF. Normalization of real-time quantitative reverse transcription-PCR data: A model-based variance estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets. *Cancer Res* 2004;**64**:5245–50.

402 Livak KJ, Schmittgen TD. Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the 2−ΔΔCT Method. *Methods* 2001;**25**:402–8.

403 Noble CL, Abbas AR, Cornelius J, Lees CW, Ho G-T, Toy K, *et al.* Regional variation in gene expression in the healthy colon is dysregulated in ulcerative colitis. *Gut* 2008;**57**:1398–405.

404 Bayatti N, Cooper-Knock J, Bury JJ, Wyles M, Heath PR, Kirby J, *et al.* Comparison of blood RNA extraction methods used for gene expression profiling in amyotrophic lateral sclerosis. *PLoS One* 2014;**9**:e87508.

405 Vartanian K, Slottke R, Johnstone T, Casale A, Planck SR, Choi D, *et al.* Gene expression profiling of whole blood: comparison of target preparation methods for accurate and reproducible microarray analysis. *BMC Genomics* 2009;**10**:2.

406 Jin P, Kang Q, Wang X, Yang L, Yu Y, Li N, *et al.* Performance of a second-generation methylated SEPT9 test in detecting colorectal neoplasm. *J Gastroenterol Hepatol* 2015;**30**:830–3.

407 Ladabaum U, Alvarez-Osorio L, Rösch T, Brueggenjuergen B. Cost-effectiveness of colorectal cancer screening in Germany: current endoscopic and fecal testing strategies versus plasma methylated Septin 9 DNA. *Endosc Int open* 2014;**2**:E96–104.

408 Venables WN, Ripley BD. *Modern Applied Statistics with S*. 4th ed. Springer 2002.

409 Carey V, Redestig H. ROC: utilities for ROC, with uarray focus. R Packag. version 1.44.0.

410 Slawski M, Boulesteix A, Bernau C. CMA: Synthesis of microarray-based classification. R Packag. version 1.26.0. 2009.

411 Monti S, Tamayo P, Mesirov J, Golub T. Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data. *Mach Learn* 2003;**52**:91–118.

412 Wilkerson M, Waltman P. ConsensusClusterPlus: ConsensusClusterPlus. R Packag. version 1.22.0. 2013.

413 Wilkerson MD, Hayes DN. ConsensusClusterPlus: A class discovery tool with confidence assessments and item tracking. *Bioinformatics* 2010;**26**:1572–3.

414 Dudoit S, Fridlyand J. A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biol* 2002;**3**:RESEARCH0036.

415 Park MY, Hastie T. L 1 -regularization path algorithm for generalized linear models. *J R Stat Soc Ser B (Statistical Methodol* 2007;**69**:659–77.

416 Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Ser B*

1996;**58**:267–88.

417 Bell JT, Tsai P-C, Yang T-P, Pidsley R, Nisbet J, Glass D, *et al.* Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population. *PLoS Genet* 2012;**8**:e1002629.

418 Şenbabaoğlu Y, Michailidis G, Li JZ. Critical limitations of consensus clustering in class discovery. *Sci Rep* 2014;**4**:6207.

419 Lee RC, Feinbaum RL, Ambros V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* 1993;**75**:843–54.

420 Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE, Mello CC. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature* 1998;**391**:806–11.

421 Esteller M. Non-coding RNAs in human disease. Nat. Rev. Genet. 2011;**12**:861–74.

422 Brain O, Owens BMJ, Pichulik T, Allan P, Khatamzas E, Leslie A, *et al.* The intracellular sensor NOD2 induces microrna-29 expression in human dendritic cells to limit IL-23 release. *Immunity* 2013;**39**:521–36.

423 Nguyen HTT, Dalmasso G, Müller S, Carrière J, Seibold F, Darfeuille-Michaud A. Crohn's disease-associated adherent invasive escherichia coli modulate levels of microRNAs in intestinal epithelial cells to reduce autophagy. *Gastroenterology* 2014;**146**:508–19.

424 Lu C, Chen J, Xu H-GG, Zhou X, He Q, Li Y-LL, *et al.* MIR106B and MIR93 prevent removal of bacteria from epithelial cells by disrupting ATG16L1-mediated autophagy. *Gastroenterology* 2014;**146**:188–99.

425 Yang Y, Ma Y, Shi C, Chen H, Zhang H, Chen N, *et al.* Overexpression of miR-21 in patients with ulcerative colitis impairs intestinal epithelial barrier function through targeting the Rho GTPase RhoB. *Biochem Biophys Res Commun* 2013;**434**:746–52.

426 Xue X, Feng T, Yao S, Wolf KJ, Liu C-G, Liu X, *et al.* Microbiota downregulates dendritic cell expression of miR-10a, which targets IL-12/IL-23p40. *J Immunol* 2011;**187**:5879–86.

427 van Rooij E. The art of microRNA research. *Circ Res* 2011;**108**:219–34.

428 Zhao S, Fung-Leung W-P, Bittner A, Ngo K, Liu X. Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells. *PLoS One* 2014;**9**:e78644.

429 Lin J, Welker NC, Zhao Z, Li Y, Zhang J, Reuss S a, *et al.* Novel specific microRNA biomarkers in idiopathic inflammatory bowel disease unrelated to disease activity. *Mod Pathol* 2013;:1–7.

430 Cariello NF, Keohavong P, Sanderson BJS, Thilly WG. DNA damage produced by ethidium bromide staining and exposure to ultraviolet light. *Nucleic Acids Res* 1988;**16**:4157.

431 Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMEnet J* 2011;**17**:10–2.

432 Barturen G, Rueda A, Hamberg M, Alganza A, Lebron R, Kotsyfakis M, *et al.* sRNAbench: profiling of small RNAs and its sequence variants in single or multi-species high-throughput experiments. *Methods Next Gener Seq* 2014;**1**:21–31.

433 Wang W-C, Lin F-M, Chang W-C, Lin K-Y, Huang H-D, Lin N-S. miRExpress: analyzing high-throughput sequencing data for profiling microRNA expression. *BMC Bioinformatics* 2009;**10**:328.

434 An J, Lai J, Lehman ML, Nelson CC. miRDeep*: an integrated application tool for miRNA identification from RNA sequencing data. *Nucleic Acids Res* 2013;**41**:727–37.

435 Stocks MB, Moxon S, Mapleson D, Woolfenden HC, Mohorianu I, Folkes L, *et al.* The UEA sRNA workbench: a suite of tools for analysing and visualizing next generation sequencing microRNA and small RNA datasets. *Bioinformatics* 2012;**28**:2059–61.

436 Rahmann S, Martin M, Schulte JH, Köster J, Marschall T, Schramm A. Identifying transcriptional miRNA biomarkers by integrating high-throughput sequencing and real-time PCR data. *Methods* 2013;**59**:154–63.

437 Dillies M-A, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, *et al.* A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief Bioinform* 2013;**14**:671–83.

438 Agarwal V, Bell GW, Nam J-W, Bartel DP. Predicting effective microRNA target sites in mammalian mRNAs. *Elife* 2015;**4**:e05005.

439 Papadopoulos GL, Alexiou P, Maragkakis M, Reczko M, Hatzigeorgiou AG. DIANA-mirPath: Integrating human and mouse microRNAs in pathways. *Bioinformatics* 2009;**25**:1991–3.

440 Ide S, Toiyama Y, Shimura T, Kawamura M, Yasuda H, Saigusa S, *et al.* MicroRNA-503 promotes tumor progression and acts as a novel biomarker for prognosis in oesophageal cancer. *Anticancer Res* 2015;**35**:1447–51.

441 Long J, Ou C, Xia H, Zhu Y, Liu D. MiR-503 inhibited cell proliferation of human breast cancer cells by suppressing CCND1 expression. *Tumour Biol* 2015;**36**:8697–702.

442 Chong Y, Zhang J, Guo X, Li G, Zhang S, Li C, *et al.* MicroRNA-503 acts as a tumor suppressor in osteosarcoma by targeting L1CAM. *PLoS One* 2014;**9**:e114585.

443 Fulci V, Scappucci G, Sebastiani GD, Giannitti C, Franceschini D, Meloni F, *et al.* miR-223 is overexpressed in T-lymphocytes of patients affected by rheumatoid arthritis. *Hum Immunol* 2010;**71**:206–11.

444 Sun W, Shen W, Yang S, Hu F, Li H, Zhu T-H. miR-223 and miR-142 attenuate hematopoietic cell proliferation, and miR-223 positively regulates miR-142 through LMO2 isoforms and CEBP-β. *Cell Res* 2010;**20**:1158–69.

445 Johnnidis JB, Harris MH, Wheeler RT, Stehling-Sun S, Lam MH, Kirak O, *et al.* Regulation of progenitor cell proliferation and granulocyte function by microRNA-223. *Nature* 2008;**451**:1125–9.

446 Godard P, van Eyll J. Pathway analysis from lists of microRNAs: common pitfalls and alternative strategy. *Nucleic Acids Res* 2015;**43**:3490–7.

447 Absher DM, Li X, Waite LL, Gibson A, Roberts K, Edberg J, *et al.* Genome-wide DNA methylation analysis of systemic lupus erythematosus reveals persistent hypomethylation of interferon genes and compositional changes to CD4+ T-cell populations. *PLoS Genet* 2013;**9**:e1003678.

448 Maltby VE, Graves MC, Lea RA, Benton MC, Sanders KA, Tajouri L, *et al.* Genome-wide DNA methylation profiling of CD8+ T cells shows a distinct epigenetic signature to CD4+ T cells in multiple sclerosis patients. *Clin Epigenetics* 2015;**7**:118.

449 Bos SD, Page CM, Andreassen BK, Elboudwarej E, Gustavsen MW, Briggs F, *et al.* Genome-wide DNA methylation profiles indicate CD8+ T cell hypermethylation in multiple sclerosis. *PLoS One* 2015;**10**:e0117403.

450 Kashiwakura J, Suzuki N, Nagafuchi H, Takeno M, Takeba Y, Shimoyama Y, *et al.* Txk, a nonreceptor tyrosine kinase of the Tec family, is expressed in T helper type 1 cells and regulates interferon gamma production in human T lymphocytes. *J Exp Med* 1999;**190**:1147–54.

451 Punit S, Dubé PE, Liu CY, Girish N, Washington MK, Polk DB. Tumor Necrosis Factor Receptor 2 Restricts the Pathogenicity of CD8(+) T Cells in Mice With Colitis. *Gastroenterology* 2015;**149**:993–1005.e2.

452 Funderburg NT, Stubblefield Park SR, Sung HC, Hardy G, Clagett B, Ignatz-Hoover J, *et al.* Circulating CD4(+) and CD8(+) T cells are activated in inflammatory bowel disease and are associated with plasma markers of inflammation. *Immunology* 2013;**140**:87–97.

453 Nancey S, Holvöet S, Graber I, Joubert G, Philippe D, Martin S, *et al.* CD8+ cytotoxic T cells induce relapsing colitis in normal mice. *Gastroenterology* 2006;**131**:485–96.

454 McKinney EF, Lee JC, Jayne DRW, Lyons PA, Smith KGC. T-cell exhaustion, co-stimulation and clinical outcome in autoimmunity and infection. *Nature* 2015;**523**:612–6.

455 Davey Smith G, Hemani G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum Mol Genet* 2014;**23**:R89–98.

456 Hawkey CJ, Allez M, Clark MM, Labopin M, Lindsay JO, Ricart E, *et al.* Autologous Hematopoetic Stem Cell Transplantation for Refractory Crohn Disease: A Randomized Clinical Trial. *JAMA* 2015;**314**:2524–34.

457 Plongthongkum N, Diep DH, Zhang K. Advances in the profiling of DNA modifications: cytosine methylation and beyond. *Nat Rev Genet* 2014;**15**:647–61.

458 Gackowski D, Zarakowska E, Starczak M, Modrzejewska M, Olinski R. Tissue-Specific Differences in DNA Modifications (5-Hydroxymethylcytosine, 5-Formylcytosine, 5-Carboxylcytosine and 5-Hydroxymethyluracil) and Their Interrelationships. *PLoS One* 2015;**10**:e0144859.

459 Liu Y, Liu P, Yang C, Cowley AW, Liang M. Base-Resolution Maps of 5-Methylcytosine and 5-Hydroxymethylcytosine in Dahl S Rats: Effect of Salt and Genomic Sequence. *Hypertension* 2014;**63**:827–38.

460 Strand SH, Hoyer S, Lynnerup A-S, Haldrup C, Storebjerg TM, Borre M, *et al.* High levels of 5-hydroxymethylcytosine (5hmC) is an adverse predictor of biochemical recurrence after prostatectomy in ERG-negative prostate cancer. *Clin Epigenetics* 2015;**7**:111.

461 Tsai K-W, Li G-C, Chen C-H, Yeh M-H, Huang J-S, Tseng H-H, *et al.* Reduction of global 5-hydroxymethylcytosine is a poor prognostic factor in breast cancer patients, especially for an ER/PR-negative subtype. *Breast Cancer Res Treat* 2015;**153**:219–34.

462 Uribe-Lewis S, Stark R, Carroll T, Dunning MJ, Bachman M, Ito Y, *et al.* 5-hydroxymethylcytosine marks promoters in colon that resist DNA hypermethylation in cancer. *Genome Biol* 2015;**16**:69.

463 Bachman M, Uribe-Lewis S, Yang X, Burgess HE, Iurlaro M, Reik W, *et al.* 5-Formylcytosine can be a stable DNA modification in mammals. *Nat Chem Biol* 2015;**11**:555–7.

464 Lee S-T, Muench MO, Fomin ME, Xiao J, Zhou M, de Smith A, *et al.* Epigenetic remodeling in B-cell acute lymphoblastic leukemia occurs in two tracks and employs embryonic stem cell-like signatures. *Nucleic Acids Res* 2015;**43**:2590–602.

465 Kulis M, Heath S, Bibikova M, Queirós AC, Navarro A, Clot G, *et al.* Epigenomic analysis detects widespread gene-body DNA hypomethylation in chronic lymphocytic leukemia. *Nat Genet* 2012;**44**:1236–42.

466 Kulis M, Merkel A, Heath S, Queirós AC, Schuyler RP, Castellano G, *et al.* Whole-genome fingerprint of the DNA methylome during human B cell differentiation. *Nat Genet* 2015;**47**:746–56.

467 Geremia A, Arancibia-Cárcamo C V, Fleming MPP, Rust N, Singh B, Mortensen NJ, *et al.* IL-23-responsive innate lymphoid cells are increased in inflammatory bowel disease. *J Exp Med* 2011;**208**:1127–33.

468 Maeder ML, Angstman JF, Richardson ME, Linder SJ, Cascio VM, Tsai SQ, *et al.* Targeted DNA demethylation and activation of endogenous genes using programmable TALE-TET1 fusion proteins. *Nat Biotechnol* 2013;**31**:1137–42.

469 McLachlan G. *Discriminant Analysis and Statistical Pattern Recognition*. John Wiley & Sons 2004.

470 Ripley BD. *Pattern Recognition and Neural Networks*. Cambridge University Press 1996.

471 Zhu J, Hastie T. Classification of gene microarrays by penalized logistic regression. *Biostatistics* 2004;**5**:427–443.

472 Boulesteix A-L, Strimmer K. Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Brief Bioinform* 2006;**8**:32–44.

473 Breiman L. Random Forests. *Mach Learn* 2001;**45**:5–32.

474 Bühlmann P, Yu B. Boosting With the L 2 Loss. *J Am Stat Assoc* 2003;**98**:324–39.

475 Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Ser B (Statistical Methodol* 2005;**67**:301–20.

Appendix 1-Chapter 3 Genome wide DNA methylation analysis

Gene	Position	CHR	Δβ	P Value	Adj P value
RPS6KA2	166970252	chr6	-0.09	1.26E-29	5.63E-24
SBNO2	1130866	chr19	-0.09	3.61E-26	1.62E-20
VMP1	57915665	chr17	-0.11	3.04E-23	1.36E-17
SBNO2	1130965	chr19	-0.05	1.15E-22	5.16E-17
VMP1	57915717	chr17	-0.10	3.52E-22	1.58E-16
NA	50327986	chr22	-0.05	4.81E-21	2.15E-15
NA	101901234	chr3	-0.06	2.73E-18	1.22E-12
SOCS3	76354621	chr17	-0.07	3.63E-18	1.63E-12
NA	12890029	chr19	-0.06	1.92E-17	8.62E-12
FKBP5	35654363	chr6	-0.07	3.42E-17	1.53E-11
NA	35696870	chr6	-0.04	7.67E-17	3.44E-11
VMP1	57915773	chr17	-0.06	1.09E-16	4.90E-11

Table 44 - Top table of differentially methylated positions (DMPs) between Crohn's disease (CD) cases and controls in whole blood

Gene	Position	CHR	Δβ	P Value	Adj P value
NA	35696870	chr6	-0.04	1.12E-16	5.02E-11
NA	101901234	chr3	-0.06	1.35E-16	6.07E-11
TNFSF10	172235808	chr3	-0.05	2.85E-16	1.28E-10
SBNO2	1130866	chr19	-0.06	1.42E-15	6.36E-10
NA	12890029	chr19	-0.06	1.57E-15	7.03E-10
AIM2	159047163	chr1	-0.06	6.96E-15	3.12E-09
ICA1	8201134	chr7	0.04	1.45E-14	6.48E-09
RPS6KA2	166970252	chr6	-0.07	2.22E-14	9.95E-09
VMP1	57915665	chr17	-0.09	2.48E-14	1.11E-08
ZEB2	145172035	chr2	-0.08	5.65E-14	2.53E-08
NA	50327986	chr22	-0.04	2.11E-13	9.45E-08
FRMD4A	13913931	chr10	0.05	2.45E-13	1.10E-07

Table 45 - Top table of differentially methylated positions (DMPs) between Ulcerative colitis (UC) cases and controls in whole blood

Gene	Probe	Position	CHR	Δβ	P Value	Holm Adj P value
MNDA	cg05304729	158800024	chr1	-0.06	3.44E-07	0.15
NA	cg19683494	74908142	chr5	-0.06	5.94E-07	0.27
ZEB2	cg10502206	145182344	chr2	0.02	1.08E-06	0.49
NA	cg02573091	74908125	chr5	-0.07	1.97E-06	0.88
NLRC5	cg07839457	57023022	chr16	-0.06	2.81E-06	1
TK1	cg25069807	76171191	chr17	0.03	4.75E-06	1
NA	cg25730685	2375010	chr1	0.01	7.07E-06	1
SPG7	cg04879696	89574810	chr16	0.01	7.66E-06	1
TK1	cg06098276	76171208	chr17	0.04	1.30E-05	1
CENPV	cg05238069	16257135	chr17	0.01	1.56E-05	1
CAPN5	cg08103551	76777993	chr11	0.00	1.67E-05	1
ADK	cg23198334	76179907	chr10	0.01	2.10E-05	1

Table 46 - Top table of differentially methylated positions (DMPs) between Crohn's disease (CD) and Ulcerative colitis in whole blood

Gene	Probe	Position	CHR	Δβ	P Value	Adj P value
ROCK1	cg09449490	18690843	chr18	0.01	1.17E-06	0.522723
PFKFB3	cg27545615	6249748	chr10	0.03	1.55E-06	0.693497
PIK3R6	cg00409104	8762014	chr17	0.02	1.55E-06	0.696945
CELF2	cg11832281	11211022	chr10	0.01	1.83E-06	0.818404
NA	cg01686975	138816336	chr7	0.02	2.19E-06	0.981941
NA	cg24448340	179921042	chr1	0.03	3.00E-06	1
TXNDC11	cg03382501	11794641	chr16	0.02	3.06E-06	1
EFHD2	cg25978218	15738732	chr1	0.02	3.08E-06	1
CPD	cg23344321	28707212	chr17	0.02	4.57E-06	1
XPO6	cg26730763	28205389	chr16	0.01	5.56E-06	1
KIAA1033	cg13622209	105501127	chr12	0.01	5.75E-06	1
NA	cg08900384	72546168	chr11	0.02	6.03E-06	1

Table 47 - Top table of differentially methylated positions (DMPs) between Symptomatic controls and healthy volunteers in whole blood

Illumina 450k Probe id	Chr	Gene symbol	Δβ	P.Value	FDR adj.P.Val
cg17501210	chr6	RPS6KA2	-0.07	2.09E-21	9.38E-16
cg18608055	chr19	SBNO2	-0.07	2.20E-20	9.84E-15
cg16936953	chr17	VMP1	-0.08	1.20E-18	5.36E-13
cg09349128	chr22	NA	-0.04	6.81E-18	3.05E-12
cg12170787	chr19	SBNO2	-0.04	6.96E-18	3.12E-12
cg25114611	chr6	NA	-0.04	1.19E-17	5.35E-12
cg12992827	chr3	NA	-0.05	3.58E-17	1.60E-11
cg19821297	chr19	NA	-0.06	1.79E-16	8.04E-11
cg12054453	chr17	VMP1	-0.07	4.29E-16	1.92E-10
cg01059398	chr3	TNFSF10	-0.04	1.78E-15	7.98E-10
cg26804423	chr7	ICA1	0.04	4.04E-15	1.81E-09
cg02716826	chr9	NA	-0.03	8.06E-15	3.61E-09
cg13619623	chr7	BBS9	0.04	1.00E-14	4.49E-09
cg18942579	chr17	VMP1	-0.05	1.05E-14	4.70E-09
cg03546163	chr6	FKBP5	-0.06	1.72E-14	7.70E-09

Table 48 - Top table of differentially methylated positions (DMPs) between inflammatory bowel disease (IBD) cases and controls in whole blood with smoking included as a covariate along with age, sex and the estimated blood cell proportions

Illumina 450k Probe id	**Chr**	**Gene Symbol**	**Δβ**	**P.Value**	**FDR. adj.P.Val**
cg11870956	chr6	NA	0.05	3.54E-08	0.009
cg07215298	chr2	HDAC4	0.08	4.30E-08	0.009
cg17501210	chr6	RPS6KA2	-0.12	5.75E-08	0.009
cg02508743	chr8	LYN	0.08	1.09E-07	0.010
cg17901584	chr1	DHCR24	-0.09	1.13E-07	0.010
cg22610434	chr1	CD1C	0.04	1.85E-07	0.013
cg18931633	chr8	CLU	-0.08	2.17E-07	0.013
cg17192381	chr18	BCL2	0.04	2.25E-07	0.013
cg01799015	chr19	PALM	-0.07	2.94E-07	0.015
cg19458697	chr2	NA	0.03	3.48E-07	0.016
cg19590591	chr12	GOLGA3	0.12	3.83E-07	0.016
cg24843346	chr8	NA	-0.06	4.18E-07	0.016
cg05832823	chr13	TBC1D4	0.07	5.47E-07	0.019
cg12488187	chr12	MSRB3	-0.07	7.05E-07	0.022
cg06996599	chr6	C6orf136	0.06	7.28E-07	0.022
cg26663590	chr16	NA	0.07	7.82E-07	0.022
cg27293155	chr2	NA	0.08	8.88E-07	0.023

Table 49- Top Table of differentially methylated positions (DMPs) in inflammatory bowel disease (IBD) cases and controls in CD14+ monocytes.

Illumina 450k Probe id	Chr	Gene Symbol	Δβ	P.Value	FDR. adj.P.Val
cg02478369	chr17	NA	0.01	7.97E-08	0.036
cg04485603	chr4	CLOCK	0.01	1.86E-06	0.164
cg08616681	chr16	RHOT2	-0.02	1.98E-06	0.164
cg00977895	chr3	NA	0.02	2.09E-06	0.164
cg22617703	chr20	DNAJC5	0.02	2.69E-06	0.164
cg18360149	chr9	ZNF79	-0.01	3.07E-06	0.164
cg08540929	chr5	NA	0.02	3.72E-06	0.164
cg17445101	chr4	GAK	0.01	3.81E-06	0.164
cg04587829	chr17	FN3K	0.05	4.47E-06	0.164
cg03850936	chr6	NA	0.07	4.49E-06	0.164
cg08463024	chr6	DDX39B	-0.03	4.57E-06	0.164
cg02976575	chr7	LMTK2	0.01	5.20E-06	0.164
cg26511507	chr10	DIP2C	0.02	5.27E-06	0.164
cg19807612	chr21	NA	-0.01	5.33E-06	0.164
cg14153061	chr9	STX17	-0.06	5.71E-06	0.164
cg24011441	chr19	NA	-0.01	5.83E-06	0.164
cg09977980	chr17	EVPLL	0.02	6.95E-06	0.164

Table 50- Top Table of differentially methylated positions (DMPs) in inflammatory bowel disease (IBD) cases and controls in CD4+ lymphocytes.

Illumina 450k Probe id	Chr	Gene Symbol	Δβ	P.Value	FDR. adj.P.Val
cg02985240	chr1	ARID4B	0.07	2.95E-08	0.009
cg04405547	chr3	NA	-0.07	3.81E-08	0.009
cg16339434	chr11	NA	-0.07	1.21E-07	0.013
cg08179431	chr6	HFE	0.11	1.55E-07	0.013
cg07632771	chr7	NA	-0.06	1.60E-07	0.013
cg08831348	chr19	EML2	0.01	2.23E-07	0.013
cg23665802	chr13	NA	-0.08	2.47E-07	0.013
ch.3.38006391R	chr3	NA	-0.01	2.58E-07	0.013
cg01995548	chr16	RPL13	-0.05	2.76E-07	0.013
cg09373727	chr1	PTP4A2	-0.09	2.81E-07	0.013
cg14277403	chr9	ANP32B	-0.10	3.20E-07	0.013
cg10909790	chr10	ALOX5	-0.07	3.84E-07	0.014
cg27141915	chr19	IRGQ	0.05	4.40E-07	0.015
cg18150958	chr17	NA	-0.04	4.75E-07	0.015
cg22801913	chr11	C11orf49	0.03	5.00E-07	0.015
cg13709639	chr12	TUBA1B	-0.09	5.67E-07	0.016
cg02679745	chr9	NA	-0.06	5.89E-07	0.016

51- Top Table of differentially methylated positions (DMPs) in inflammatory bowel disease (IBD) cases and controls in CD8+ lymphocytes.

Illumina 450k Probe id	Chr	Gene Symbol	Δβ	P.Value	FDR. adj.P.Val
cg24773560	12	IL23A	0.06	6.66E-09	0.003
cg07215298	2	HDAC4	0.08	1.57E-08	0.004
cg22017309	11	RAB3IL1	-0.04	2.40E-08	0.004
cg18931633	8	CLU	-0.08	3.67E-08	0.004
cg26701810	2	NA	-0.03	6.00E-08	0.005
cg02368508	16	TNFRSF17	0.09	9.03E-08	0.006
cg07457727	8	NA	0.08	1.25E-07	0.006
cg00719771	6	NA	0.11	1.29E-07	0.006
cg00962903	3	MECOM	-0.07	1.39E-07	0.006
cg23216724	6	GPR31	0.13	1.48E-07	0.006
cg25921544	2	HDAC4	0.11	1.50E-07	0.006
cg15020801	17	PNPO	0.07	1.58E-07	0.006
cg05941027	17	LIMD2	0.06	2.02E-07	0.006
cg12992827	3	NA	-0.11	2.26E-07	0.006
cg02508743	8	LYN	0.06	2.26E-07	0.006
cg00779858	3	WDR49	0.07	2.45E-07	0.006
cg17579089	3	KIF9	0.07	2.57E-07	0.006

Table 52 - Top table of differentially methylated positions (DMPs) between Crohn's disease (CD) cases and controls in CD14+ monocytes

Illumina 450k Probe id	Chr	Gene Symbol	Δβ	P.Value	FDR. adj.P.Val
cg08936645	4	TBC1D1	0.02	2.16E-06	0.4
cg04587829	17	FN3K	0.04	5.51E-06	0.4
cg24743237	2	D2HGDH	0.03	6.87E-06	0.4
cg00345704	17	KRTAP2-3	0.02	7.33E-06	0.4
cg07652774	6	NA	0.02	7.89E-06	0.4
cg25647784	17	WNK4	0.09	9.15E-06	0.4
cg25904183	4	NA	-0.07	1.02E-05	0.4
cg06521149	1	NA	-0.03	1.04E-05	0.4
cg22460123	12	KRT7	-0.03	1.11E-05	0.4
cg07287949	17	NA	0.04	1.13E-05	0.4
cg16261737	2	FN1	-0.05	1.38E-05	0.4
cg27013382	2	NA	0.03	1.46E-05	0.4
cg17454857	20	NA	0.04	1.61E-05	0.4
cg05063856	9	NA	-0.04	1.67E-05	0.4
cg07649744	20	HELZ2	-0.01	2.09E-05	0.4
cg10827159	12	TMEM132B	-0.02	2.12E-05	0.4
cg20989942	12	NA	0.03	2.16E-05	0.4

Table 53- Top table of differentially methylated positions (DMPs) between Crohn's disease (CD) cases and controls in CD4+ lymphocytes

Illumina 450k Probe id	Chr	Gene Symbol	Δβ	P.Value	FDR. adj.P.Val
cg20239639	1	LCK	0.07	1.08E-06	0.5
cg02985240	1	ARID4B	0.08	2.44E-06	0.5
cg16575998	2	C2orf61	0.07	3.45E-06	0.5
cg06713373	3	SLC12A8	0.063	4.70E-06	0.5
cg17181543	1	AK5	0.03	7.47E-06	0.6
cg18338046	5	TCF7	0.10	8.33E-06	0.6
cg08113002	19	ASPDH	-0.10	9.76E-06	0.6
cg09789252	19	ASPDH	-0.11	1.13E-05	0.6
cg00619505	13	TMCO3	0.08	1.30E-05	0.6
cg07702548	16	ZC3H18	0.02	1.50E-05	0.6
cg19145607	3	NA	0.15	1.72E-05	0.6
cg01219426	11	MAML2	0.02	1.78E-05	0.6
cg09053247	3	NA	0.06	1.84E-05	0.6
cg01631226	22	NA	0.06	1.86E-05	0.6
cg04587829	17	FN3K	0.04	2.19E-05	0.6
cg13298528	11	CXCR5	0.07	2.29E-05	0.6
cg13992008	9	FAM102A	0.10	2.53E-05	0.7

Table 54 - Top table of differentially methylated positions (DMPs) between Crohn's disease (CD) cases and controls in CD8+ lymphocytes

Illumina 450k Probe id	Chr	Gene Symbol	Δβ	P.Value	FDR. adj.P.Val
cg26508200	12	SSH1	-0.04	7.03E-08	0.03
cg08272368	3	NA	-0.06	3.47E-07	0.08
cg24843346	8	NA	-0.07	8.19E-07	0.1
cg21120249	9	NA	-0.07	1.61E-06	0.2
cg04999691	7	NA	-0.03	1.83E-06	0.2
cg04152675	2	CASP10	-0.04	2.18E-06	0.2
cg14847009	1	KIAA0040	-0.10	2.76E-06	0.2
cg03270340	6	TRIM27	-0.02	3.36E-06	0.2
cg17501210	6	RPS6KA2	-0.15	3.49E-06	0.2
cg19515398	8	NA	0.07	4.23E-06	0.2
cg00575066	6	NA	-0.03	5.08E-06	0.2
cg14703482	4	FGF2	-0.02	6.53E-06	0.2
cg05380304	4	AFAP1-AS1	0.04	7.66E-06	0.3
cg11906021	17	NA	0.06	9.18E-06	0.3
cg22920418	19	RPS5	-0.03	1.02E-05	0.3
cg15766101	19	CCDC61	-0.02	1.09E-05	0.3
cg23866916	19	SBNO2	-0.08	1.31E-05	0.3

Table 55 - Top table of differentially methylated positions (DMPs) between Ulcerative colitis (UC) cases and controls in CD14+ Monocytes

Illumina 450k Probe id	Chr	Gene Symbol	Δβ	P.Value	FDR. adj.P.Val
cg18074189	12	TMTC1	-0.06	2.34E-07	0.04
cg07869023	20	PCSK2	-0.06	3.10E-07	0.04
cg21214232	15	SNORD115-7	-0.05	3.20E-07	0.04
cg05483184	6	TNXB	-0.01	3.96E-07	0.04
cg06641503	3	ARIH2	-0.05	7.66E-07	0.07
cg19652483	6	FILIP1	-0.06	9.81E-07	0.07
cg09077443	11	NA	-0.03	1.29E-06	0.08
cg13594903	9	STOML2	-0.03	1.84E-06	0.09
cg08594554	17	TTYH2	0.06	1.93E-06	0.09
cg03638432	16	NA	-0.02	2.11E-06	0.09
cg20302533	7	POU6F2	0.14	2.40E-06	0.09
cg09435170	4	NA	-0.07	2.53E-06	0.09
cg18803306	20	NA	0.07	2.59E-06	0.09
cg08835956	7	POU6F2	0.08	2.81E-06	0.09
cg03139435	20	NA	-0.07	2.84E-06	0.09
cg18850127	7	POU6F2	0.16	3.10E-06	0.09

Table 56 - Top table of differentially methylated positions (DMPs) between Ulcerative colitis (UC) cases and controls in CD4+ Lymphocytes

Illumina 450k	Chr	Gene Symbol	Δβ	P.Value	FDR.

Probe id					adj.P.Val
cg22108567	7	NA	-0.06	2.90E-08	0.01
cg09033472	13	RASA3	-0.08	1.28E-07	0.03
cg23045404	17	NA	-0.02	2.95E-07	0.04
cg04071118	20	FAM83D	-0.07	3.94E-07	0.04
ch.12.2487434F	12	NA	-0.07	3.97E-07	0.04
cg16339434	11	NA	-0.07	9.98E-07	0.08
cg07091220	4	ZNF827	-0.08	1.39E-06	0.09
cg12785694	3	NA	-0.10	1.87E-06	0.09
cg26819611	16	KIFC3	-0.05	1.94E-06	0.09
cg14277403	9	ANP32B	-0.10	2.14E-06	0.09
cg01966091	16	HAS3	0.05	2.66E-06	0.09
cg16290996	1	NA	-0.07	3.29E-06	0.09
cg02835421	1	IPO13	-0.05	3.38E-06	0.09
cg11733958	17	MPRIP	0.08	3.58E-06	0.09
cg03364486	8	NA	0.09	3.61E-06	0.09
cg01756756	19	C19orf12	-0.05	3.67E-06	0.09
cg09649266	2	NA	0.05	3.75E-06	0.09

Table 57 – Top table of differentially methylated positions (DMPs) between Ulcerative colitis (UC) cases and controls in CD8+ Lymphocytes

Appendix 2 - Chapter 4 Targeted replication of Epigenome-wide DNA Methylaiton findings

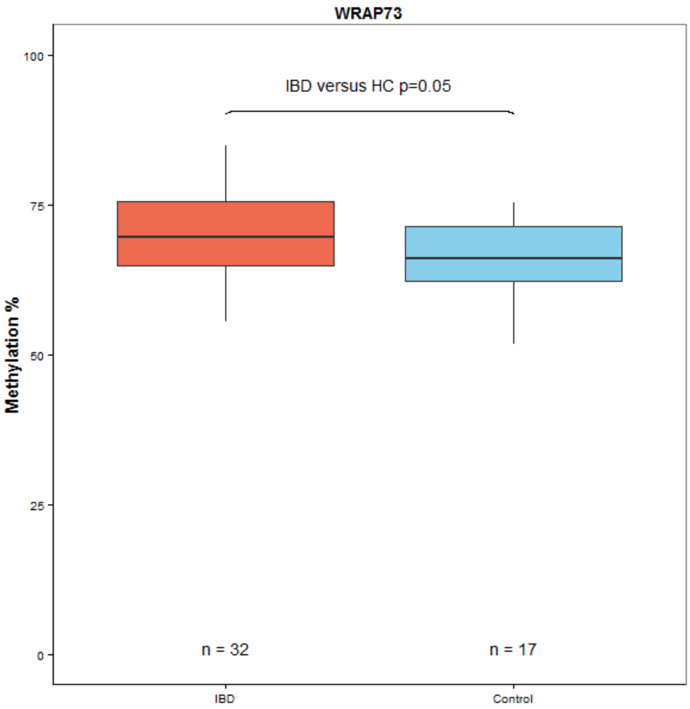


Figure 70 - WRAP73 independent cohort pyrosequencing

Appendix 3 - Chapter 5 Integrative analysis of genetic and DNA methylation data

Table 58 – Samples that failed quality assurance testing either by failing genotyping or by failing sex check.

Failed Genotyping	Failed Sex	Hetrozygosity
0474CD	0315HC	0.12
P008464-08112011-TB-S001	0037HC	0.09
0157HC		
P009007-07112013-TB-02		
P009111-28042014-TB-01		
P007843-30052012-TB-S001		
0013HC		
P009048-08012014-TB-01		
0740UC		
P009036-16122013-TB-01		
P009140-23062014-TB-01		
0281HC		
P008646-26062012-TB-S001		
0026H		
0131HC		
P008775-18012013-TB-01		
0529CD		
P007860-30052012-TB-S001		
P009021-02122013-TB-01		

Table 59 - Top list of DMPs with genetic association (meQTL). Only the top SNP association is shown for each methylation probe. The variables used to search for meQTLs were as follows: MAF>10%, cis distance 1 megabase, covariates age, sex, cell proportions, no disease covariates. The table is ordered according to the significant of the DMP in the IBD vs. Control methylation comparison (DMP rank, Holm Adjusted P.Vlaue correspond to results of linear modelling carried out in Chapter 3) rather than the significance test of the association between SNP and methylation probe (meQTL rank). The results of the significance test of the association between SNP and methylation probe are presented in columns marked meQTL P value and FDR corrected as meQTL FDR P Value

ProbeID	Ch	Meth symbol	Holm adj P.Val	DMP rank	Top SNP	meQTL P Value	meQTL FDR P Value	meQTL rank
cg16936953	17	**VMP1**	6.0E-14	3	rs8078424	2.9E-07	8.8E-05	265
cg12054453	17	**VMP1**	1.8E-11	9	rs8078424	4.4E-07	1.2E-04	284
cg18942579	17	**VMP1**	5.2E-10	14	rs10853015	3.1E-07	9.4E-05	267
cg02448796	1	KCNAB2	6.9E-09	18	rs546526	2.2E-13	2.5E-10	71
cg12582317	17	NA	2.5E-08	20	rs886926	7.4E-35	1.0E-30	6
cg16724148	1	AGL	5.4E-08	22	rs2640911	3.4E-24	1.4E-20	20
cg01409343	17	**VMP1**	1.5E-06	45	rs10853015	9.1E-07	2.3E-04	322
cg16755922	17	FOXK2	6.8E-06	61	rs11658011	1.3E-08	5.9E-06	176
cg27023597	17	**MIR21**	7.4E-06	62	rs10853015	9.0E-07	2.3E-04	320
cg02508743	8	LYN	1.9E-05	80	rs2719236	2.3E-08	1.0E-05	184
cg24469729	7	HOXA3	2.4E-05	82	rs2465276	7.1E-16	1.3E-12	45
cg14722693	8	CSGALNACT1	3.0E-05	86	rs10107533	1.3E-07	4.5E-05	232
cg24707889	21	ITGB2	3.3E-05	88	rs2070946	2.6E-11	2.0E-08	108
cg08423142	15	MYO1E	3.4E-05	89	rs17236536	2.1E-07	6.7E-05	252
cg12807764	5	NA	4.9E-05	95	rs17106769	3.4E-11	2.5E-08	110
cg02719954	8	NA	5.5E-05	97	rs1438455	7.8E-11	5.6E-08	114
cg02782634	17	**VMP1**	5.8E-05	99	rs10853015	2.3E-07	7.3E-05	258
cg21106695	14	NA	7.6E-05	108	rs10873477	1.2E-12	1.3E-09	79
cg13696490	3	AADACP1	1.2E-04	120	rs16846748	3.0E-15	4.5E-12	54
cg21066748	19	NA	3.0E-04	136	rs758761	5.8E-09	2.9E-06	164
cg12053291	12	SCARB1	3.5E-04	141	rs3782287	1.5E-07	5.1E-05	239
cg07168939	8	PSCA	3.9E-04	143	rs6471588	3.8E-09	2.0E-06	156
cg25368647	5	MXD3	5.4E-04	157	rs9885210	4.4E-07	1.3E-04	285
cg26663590	16	NA	7.8E-04	173	rs11150675	2.1E-07	6.8E-05	253
cg12535090	11	NAV2	9.6E-04	178	rs4757026	5.8E-07	1.6E-04	298
cg10241823	17	NXN	1.1E-03	180	rs12939584	2.3E-09	1.3E-06	147
cg05554192	11	USH1C	1.1E-03	181	rs1055574	1.5E-12	1.5E-09	81
cg08791347	10	FRMD4A	1.3E-03	188	rs10796127	9.1E-09	4.3E-06	170
cg20692268	1	NA	1.4E-03	196	rs311426	7.7E-10	4.6E-07	137
cg12229367	17	NA	1.5E-03	197	rs886926	9.2E-15	1.3E-11	58
cg14849578	12	SCARB1	1.6E-03	201	rs3782287	2.3E-07	7.2E-05	255
cg23298114	8	EXT1	1.7E-03	204	rs11781245	3.3E-08	1.4E-05	191
cg18663307	21	ITGB2	2.0E-03	207	rs2070946	4.5E-19	1.2E-15	30
cg23669118	16	PTX4	2.1E-03	208	rs2667675	5.1E-20	1.7E-16	25

meQTL.rank	Chr	SNP	Meth Probe	Meth Probe Symbol	FDR P Val	CIT omnibus Pval Holm	Pheno. assoc.w. Meth Holm	Pheno. assoc. w. Geno. given. Meth Holm	Geno. assoc. w. Meth. given. Pheno Holm	Meth. indep. of. Pheno. given. Geno Holm
1	8	rs678839	cg26126879	NA	6.2E-53	0.01	1.3E-05	0.01	1.7E-52	0
2	6	rs9369640	cg03951877	PHACTR1	5.6E-38	1	1.9E-06	0.02	9.5E-41	1
3	6	exm-rs1332844	cg03951877	PHACTR1	4.3E-36	0.003	1.9E-06	0.003	1.9E-38	0
4	21	rs8127895	cg13951069	SLC37A1	6.6E-36	0.1	1.2E-05	0.1	1.4E-38	0
5	21	rs8134499	cg13951069	SLC37A1	2.4E-32	1	1.2E-05	0.009	1.3E-35	1
6	17	rs886926	cg12582317	NA	1.0E-30	1	1.8E-09	1	1.4E-31	1
7	8	rs2436856	cg26126879	NA	1.3E-29	1	1.3E-05	1	1.1E-29	1
8	3	rs4687267	cg01526748	FGF12	1.0E-27	1	1.7E-06	0.05	6.8E-31	1
9	3	rs12494587	cg01526748	FGF12	1.9E-27	1	1.7E-06	0.03	1.6E-30	1
10	11	rs7103299	cg04074945	PHF21A	3.5E-25	1	1.3E-05	1	1.7E-28	1
11	11	rs3929339	cg04074945	PHF21A	4.8E-25	1	1.3E-05	1	1.4E-27	1
12	1	rs2777840	cg11847933	ADCK3	4.2E-24	1	5.1E-06	1	2.3E-27	1
13	8	rs1055376	cg26126879	NA	1.2E-23	1	1.3E-05	1	1.8E-25	1
14	8	rs571241	cg26126879	NA	1.2E-23	1	1.3E-05	1	8.8E-25	1
15	5	rs4958715	cg14968553	GALNT10	1.2E-21	1	1.3E-05	0.2	8.6E-24	1
16	11	rs2959103	cg04074945	PHF21A	2.6E-21	1	1.3E-05	1	2.2E-24	0
17	11	exm-rs16938437	cg04074945	PHF21A	2.6E-21	1	1.3E-05	1	1.5E-24	1
18	6	rs394522	cg15706657	GPR31	1.1E-20	1	1.2E-05	1	5.1E-23	1
19	6	rs13219256	cg03951877	PHACTR1	1.4E-20	1	1.9E-06	1	8.7E-24	1
20	1	rs2640911	cg16724148	AGL	1.4E-20	1	1.8E-09	0.5	1.8E-23	1

Table 60 - List of top DMPs with genetic association (DMP/meQTLs). The meQTL FDR p value denotes the FDR corrected P value for the association test between genotype and methylation.

The table is ordered by meQTL ranking (i.e. the most significant methylation-genotype association). The CIT omnibus p value represents the overall p value for the test (represents the highest p value of the other 4 tests). In Yellow are the results from the CIT test to assess whether genotype mediates the effect of methylation on disease susceptibility (Methylation CIT). Column explanation: P value for phenotype association with methylation probe. P value for Phenotype associates with genotype following adjustment for methylation. P value for Genotype association with methylation following adjustment for phenotype. P value for independence test that methylation is independent of phenotype following adjustment for genotype.

meQTL. rank	Chr	SNP	Meth Probe	Meth Probe Symbol	FDR P Val	CIT omnibus P val Holm	Pheno.assoc. w. Geno Holm	Pheno. Assoc.w. Meth. given. Geno Holm	Meth. assoc.w. Geno. given. Pheno Holm	Geno. indep.of. Pheno. given. Meth Holm
1	8	rs678839	cg26126879	NA	6.2E-53	1	1	4.6E-08	1.7E-52	1
2	6	rs9369640	cg03951877	PHACTR1	5.6E-38	1	1	7.9E-10	1.2E-40	1
3	6	exm-rs1332844	cg03951877	PHACTR1	4.3E-36	1	1	1.3E-10	1.9E-38	1
4	21	rs8127895	cg13951069	SLC37A1	6.6E-36	1	1	7.3E-08	1.4E-38	1
5	21	rs8134499	cg13951069	SLC37A1	2.4E-32	1	1	5.7E-09	1.0E-34	1
6	17	rs886926	cg12582317	NA	1.0E-30	1	1	2.2E-10	2.7E-32	1
7	8	rs2436856	cg26126879	NA	1.3E-29	1	1	1.7E-05	5.3E-30	1
8	3	rs4687267	cg01526748	FGF12	1.0E-27	1	1	1.5E-09	5.8E-30	1
9	3	rs12494587	cg01526748	FGF12	1.9E-27	1	1	9.1E-10	1.7E-29	1
10	11	rs7103299	cg04074945	PHF21A	3.5E-25	1	1	1.9E-05	8.3E-29	1
11	11	rs3929339	cg04074945	PHF21A	4.8E-25	1	1	6.7E-06	2.2E-28	1
12	1	rs2777840	cg11847933	ADCK3	4.2E-24	1	1	2.3E-06	3.7E-26	1
13	8	rs1055376	cg26126879	NA	1.2E-23	1	1	2.1E-06	2.1E-26	1
14	8	rs571241	cg26126879	NA	1.2E-23	1	1	1.4E-05	8.8E-25	0
15	5	rs4958715	cg14968553	GALNT10	1.2E-21	1	1	1.6E-07	8.3E-24	1
16	11	rs2959103	cg04074945	PHF21A	2.6E-21	1	1	1.9E-05	2.2E-24	1
17	11	exm-rs16938437	cg04074945	PHF21A	2.6E-21	1	1	1.9E-05	3.9E-23	1
18	6	rs394522	cg15706657	GPR31	1.1E-20	1	1	1.1E-05	3.9E-22	1
19	6	rs13219256	cg03951877	PHACTR1	1.4E-20	1	1	4.8E-08	3.3E-22	1
20	1	rs2640911	cg16724148	AGL	1.4E-20	1	1	4.35E-11	2.1E-23	1

Table 61 - List of top DMPs with genetic association (DMP/meQTLs). The meQTL FDR p value denotes the FDR corrected P value for the association test between genotype and methylation. The table is ordered by meQTL ranking. The CIT omnibus p value represents the overall p value for the test (highest p value of the other 4 tests). In Green are the results from the CIT test to assess whether methylation mediates genetic risk of disease susceptibility (Genetics CIT). Column explanation: P value for phenotype association with methylation probe. P value for Phenotype associates with genotype following adjustment for methylation. P value for Genotype association with methylation following adjustment for phenotype. P value for

independence test that methylation is independent of phenotype following adjustment for genotype.

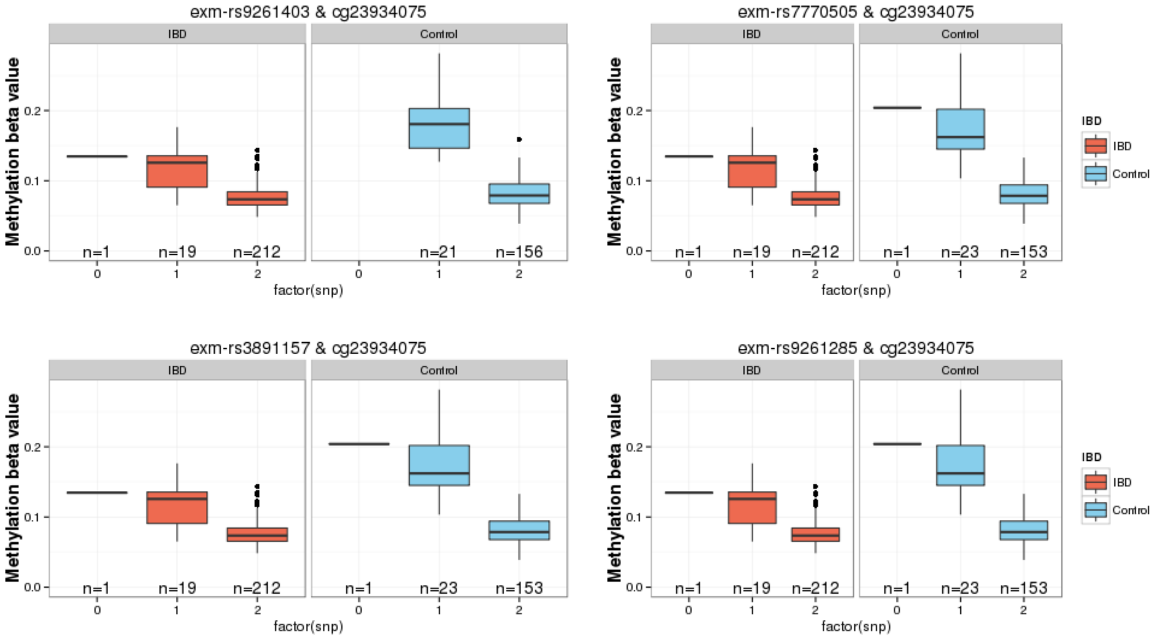


Figure 71 - cg23934075 meQTL on chromosome 6 in IBD cases versus controls in cis (panel 1 of 2). The title of each panel denotes SNP (first) followed by methylation probe.

Appendix 4 - Chapter 6 Integrative analysis of genome-wide gene expression and DNA methylation data

category	No. genes differentially expressed in category	Total No of genes in category	GO term	P Val	FDR Adj P.Val
GO:0044822	171	1156	poly(A) RNA binding	2.02E-13	4.09E-09
GO:0003723	207	1530	RNA binding	3.46E-12	3.50E-08
GO:0044445	46	204	cytosolic part	1.29E-11	8.71E-08
GO:0003676	430	3756	nucleic acid binding	2.04E-11	1.03E-07
GO:1901363	596	5579	heterocyclic compound binding	3.04E-10	1.23E-06
GO:0022626	27	102	cytosolic ribosome	4.49E-10	1.34E-06
GO:0006614	28	108	SRP-dependent cotranslational protein targeting to membrane	5.00E-10	1.34E-06
GO:0097159	601	5655	organic cyclic compound binding	5.31E-10	1.34E-06
GO:0045047	29	114	protein targeting to ER	6.37E-10	1.43E-06
GO:0044237	948	9698	cellular metabolic process	9.37E-10	1.73E-06
GO:0006613	28	110	cotranslational protein targeting to membrane	9.44E-10	1.73E-06
GO:0044424	1207	12846	intracellular part	1.76E-09	2.96E-06
GO:0072599	29	118	establishment of protein localization to endoplasmic reticulum	2.01E-09	3.12E-06
GO:0008152	1071	11186	metabolic process	2.34E-09	3.38E-06
GO:0005622	1230	13167	intracellular	4.73E-09	6.37E-06
GO:0000184	29	120	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	5.48E-09	6.92E-06
GO:0010467	529	5002	gene expression	6.43E-09	7.46E-06
GO:0019058	66	392	viral life cycle	6.65E-09	7.46E-06
GO:0090304	506	4745	nucleic acid metabolic process	8.80E-09	9.29E-06
GO:0043043	87	576	peptide biosynthetic process	9.20E-09	9.29E-06

Table 62 - Gene ontology analysis using FDR corrected differentially expressed probes in whole blood IBD versus control

	CD	**UC**	**IBD**	**Control**
n	24	18	42	17
Females (%)	9 (37.5)	6 (33.3)	15 (35.7)	6 (32.3)
Age At Diagnosis (median, IQR)	27.5 (23-37.3)	32.5 (26-41)	30 (24.3-38)	33 (30-41)
Smoking (current or Ex)	15 (62.5)	10 (58.8)	25 (61)	5 (33.3)

Table 63 - CD14 whole genome patient demographics

	CD	**UC**	**IBD**	**Control**
n	20	20	40	16
Females (%)	9 (45)	6 (30)	15 (37.5)	7 (43.8)
Age At Diagnosis (median, IQR)	29 (22.5-38.5)	33.5 (26-50)	32.5 (23.8-40.5)	35.5 (30-42.3)
Smoking (current or Ex)	11 (55)	10 (52.6)	21 (53.9)	3 (21)

Table 64 - CD4 whole genome patient demographics

	CD	**UC**	**IBD**	**Control**
n	18	18	36	14
Females (%)	6 (33.3)	6 (33.3)	12 (33.3)	4 (28.6)
Age At Diagnosis (median, IQR)	27.5 (23.5-37)	32.5 (26-48)	30.5 (24.8-44)	35 (30.3 -41.8)
Smoking (current or Ex)	11 (61.1)	10 (58.9)	21 (60)	4 (28.6)

Table 65 – CD8 whole genome patient demographics

IlluminaID	symbol	logFC	AveExpr	P.Value	Holm. Adj. P.Val
ILMN_1906187	CAMK1D	-0.42	8.64	4.79E-06	0.22
ILMN_1759743	SLC38A10	0.30	8.13	5.03E-06	0.24
ILMN_1695036	TPPP3	-0.09	6.96	5.64E-06	0.26
ILMN_2361603	NDRG2	-0.29	7.54	6.58E-06	0.31
ILMN_1738539	OPLAH	0.15	7.18	6.95E-06	0.33
ILMN_3276676	NA	-0.44	7.50	1.28E-05	0.60
ILMN_1733956	IARS	-0.30	8.09	1.30E-05	0.61
ILMN_1671891	PID1	-0.74	9.73	1.47E-05	0.69
ILMN_2119224	KIFAP3	-0.21	7.46	1.47E-05	0.69
ILMN_2393994	CSPP1	-0.16	7.63	1.82E-05	0.85
ILMN_1679580	KCNIP4	-0.09	6.92	2.31E-05	1.00
ILMN_1809013	MYL6	0.23	13.90	2.36E-05	1.00
ILMN_1725352	NA	-0.07	6.96	2.48E-05	1.00
ILMN_1660166	NA	-0.07	6.96	2.82E-05	1.00
ILMN_2337336	PVRL2	0.26	7.37	2.87E-05	1.00
ILMN_1753249	DDX10	-0.24	8.28	4.47E-05	1.00
ILMN_1761113	GNL2	-0.25	10.18	4.69E-05	1.00
ILMN_1797425	DDX55	-0.34	9.01	4.85E-05	1.00
ILMN_1808395	ACAP1	0.40	8.53	6.35E-05	1.00
ILMN_3242377	NACC2	0.26	7.60	7.07E-05	1.00

Table 66 - Whole genome expression in CD14 monocytes IBD versus controls

IlluminaID	symbol	logFC	AveExpr	P.Value	Holm.Adj.P.Val
ILMN_1688404	ZMYM4	-0.24	7.80	1.84E-05	0.9
ILMN_1864422	POU2F1	-0.21	7.23	2.01E-05	0.9
ILMN_1664065	NA	-0.12	6.78	2.90E-05	1
ILMN_2219131	RPS15	0.27	7.17	4.56E-05	1
ILMN_1737205	MCM4	0.50	7.99	5.63E-05	1
ILMN_1707493	RCC1	0.41	8.25	5.89E-05	1
ILMN_1696837	NA	0.08	6.72	9.34E-05	1
ILMN_1732296	ID3	-0.77	8.52	9.47E-05	1
ILMN_1673177	ISM2	-0.09	6.71	0.000116	1
ILMN_2411236	NRCAM	-0.15	6.81	0.000117	1
ILMN_1737184	CDCA7	0.69	8.27	0.000118	1
ILMN_1724493	LYSMD2	-0.37	10.20	0.000129	1
ILMN_1679324	EIF1B	-0.40	9.79	0.000132	1
ILMN_1795852	CCNE1	0.20	7.09	0.000145	1
ILMN_3265439	SPANXA2-OT1	0.21	6.86	0.000154	1
ILMN_1723253	ZNF222	-0.16	6.94	0.000164	1
ILMN_1804907	OR1F2P	-0.08	6.75	0.000165	1
ILMN_3226961	OAZ1	-0.13	6.83	0.000165	1
ILMN_1859942	NA	-0.15	7.08	0.000179	1
ILMN_2334693	NARF	0.47	9.52	0.000184	1

Table 67 - Whole genome expression in CD4 T-cells IBD versus controls

IlluminaID	symbol	logFC	AveExpr	P.Value	Holm.Adj.P.Val
ILMN_1813314	HIST1H2BK	0.57	8.22	3.76E-07	0.02
ILMN_1796179	HIST1H2BK	0.56	10.06	3.02E-06	0.14
ILMN_2331062	CBFA2T2	-0.25	7.69	6.89E-06	0.32
ILMN_1674763	NA	-0.11	6.46	1.64E-05	0.77
ILMN_1777233	E2F2	1.35	8.62	1.67E-05	0.78
ILMN_1790537	RMI2	0.52	7.61	4.43E-05	1.00
ILMN_2363621	RBBP8	0.78	7.47	4.52E-05	1.00
ILMN_1746403	GTF2IRD2P1	-0.32	7.55	4.53E-05	1.00
ILMN_1673112	ZNF32	-0.12	6.53	4.59E-05	1.00
ILMN_1895853	NA	0.13	6.57	4.67E-05	1.00
ILMN_2176251	MGME1	0.41	8.62	4.92E-05	1.00
ILMN_1658607	DLEU2L	0.18	6.67	5.17E-05	1.00
ILMN_1760410	NA	-0.11	6.44	5.46E-05	1.00
ILMN_1687273	BMS1P6	-0.13	6.74	6.00E-05	1.00
ILMN_2390974	DNAJB2	-0.35	9.82	6.51E-05	1.00
ILMN_1656840	VPS13D	-0.22	7.28	6.51E-05	1.00
ILMN_1672664	NA	0.18	6.85	6.58E-05	1.00
ILMN_2409298	NUSAP1	0.57	7.13	6.65E-05	1.00
ILMN_1813175	ADGRL1	-0.48	7.79	6.73E-05	1.00
ILMN_1671651	NA	-0.12	6.52	6.90E-05	1.00

Table 68 - Whole genome expression in CD8 T-cells IBD versus controls

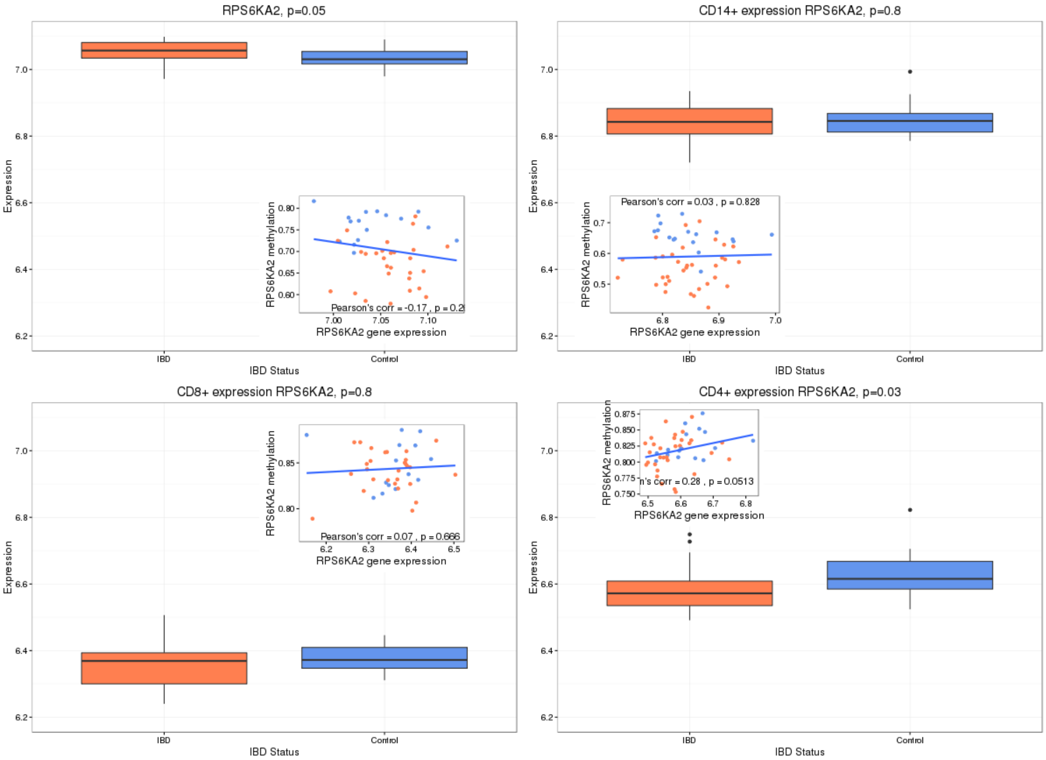


Figure 72 - Cell specific RPS6KA2 gene expression according to cell type in IBD cases and controls. Insert demonstrates correlation between methylation (beta) and gene expression.

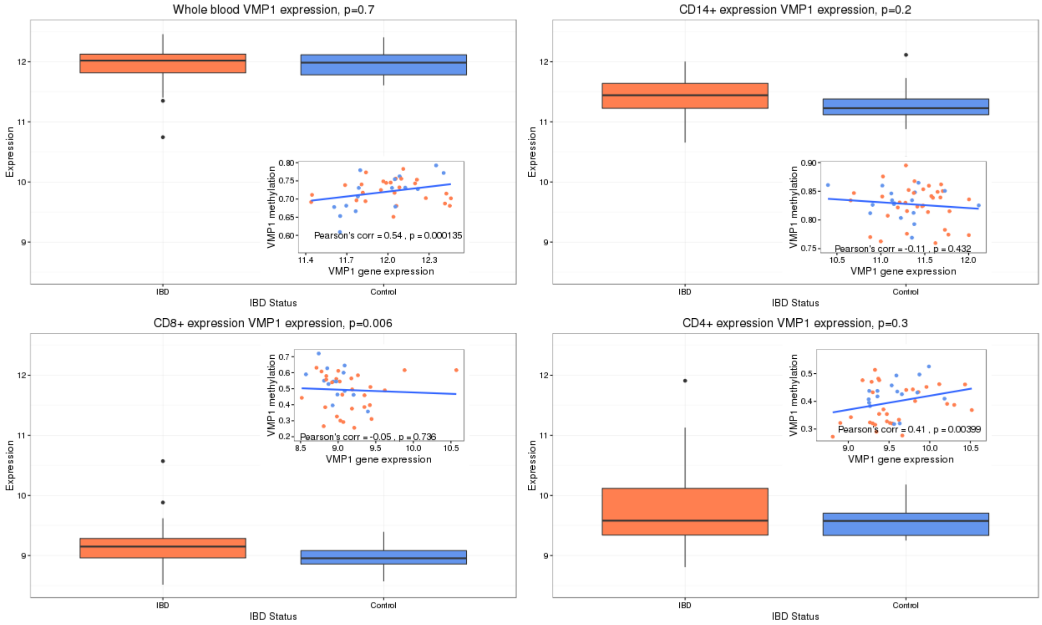


Figure 73 - Cell specific VMP1 gene expression according to cell type in IBD cases and controls. Insert demonstrates correlation between methylation (beta) and gene expression

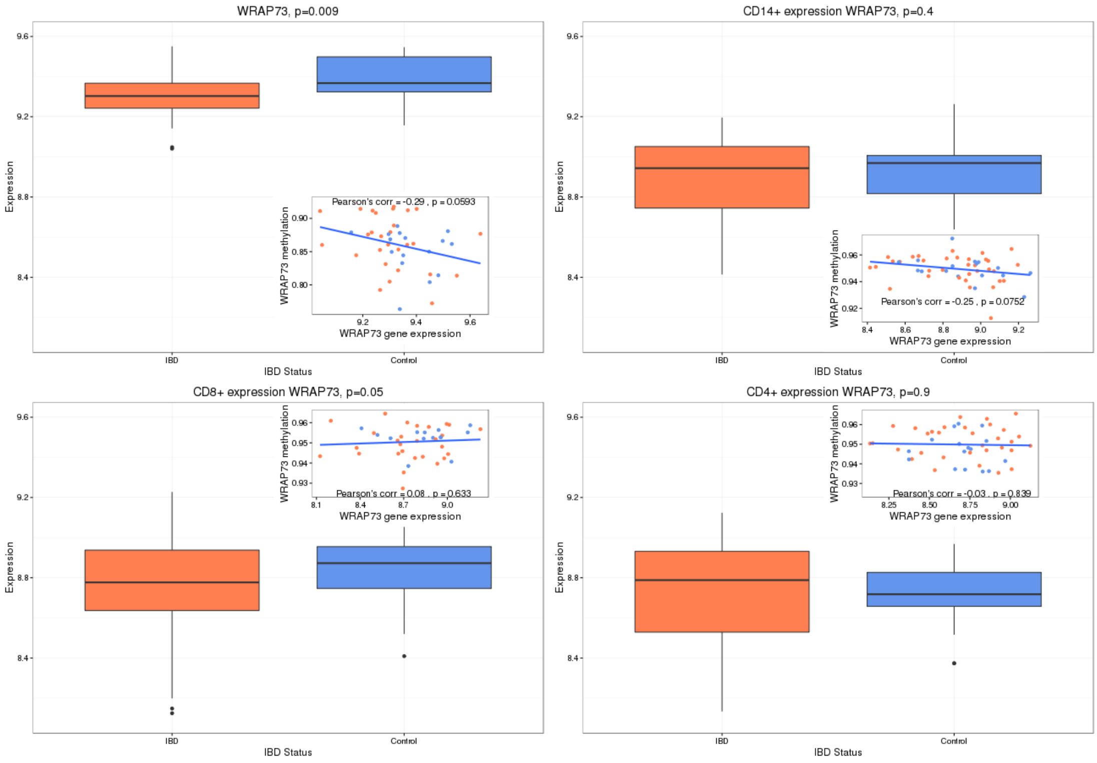


Figure 74 - Cell specific WRAP73 expression according to cell type in IBD cases and controls. Insert demonstrates correlation between methylation (beta) and gene expression

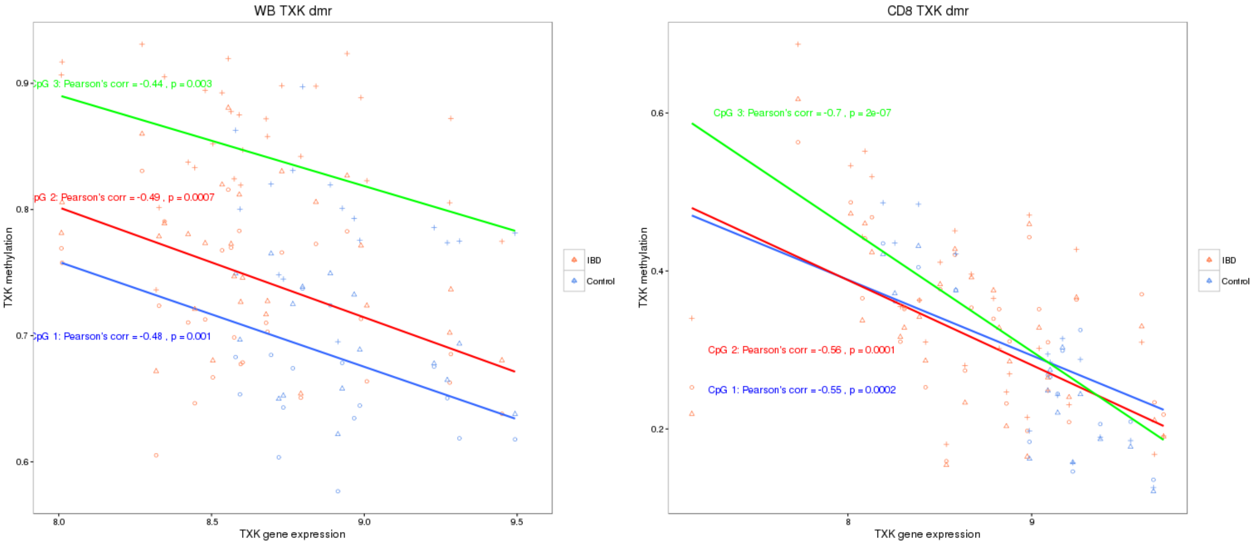


Figure 75 - Correlation between TXK gene expression and methylationCorrelation between TXK gene expression for the one probe represented on the HT12 expression array and Methylation at the three CpGs included in the DMR on the 450k array (circles/blue line = "cg02600394", triangles/red line = "cg20981615", crosses/green line = "cg17984638")(red = IBD, blue = control)

PBMCs	**IBD**	**Symptomatic control**	P Value (IBD versus Control)
n	64	17	
Age (median, IQR)	32.9 (26.2-40.3)	32.2 (27.2-40.3)	1
Females (%)	35(54.7%)	12 (70.6%)	0.2
Smoker (Current/Ex, %)	33 (57%) (missing=7)	10 (63%) (missing=1)	0.08
CRP (median, IQR)	5.5 (1-17) (missing=34)	11 (3.5-70.8) (missing=14)	0.4
Hb (median, IQR)	129 (118.5-141.8) (missing=30)	144.5 (138-149.2) (missing=12)	0.02
Alb (median, IQR) (*missing*)	36 (33.5-39.5) (missing=33)	39.5 (38.3-40) (missing=14)	0.2
Disease Duration (median, IQR, months)	3 (1-53.2)		

Table 69 - Separated PBMC Patient demographics for targeted qPCR of DMRs

Granulocytes	**IBD**	**Symptomatic control**	P Value (IBD versus Control)
n	31	7	
Age (median, IQR)	30.2 (24.9-39.8)	29.5 (22.8-42.7)	0.9
Females (%)	21(67.7%)	4 (57.1%)	0.9
Smoker (Current/Ex, %)	16 (59%) (missing=4)	2 (20%) (missing=2)	0.7
CRP (median, IQR)	7 (0-43) (missing=22)	52 (26.6-78.3) (missing=5)	0.9
Hb (median, IQR)	111(103.5-135) (missing=20)	120.5 (115-125) (missing=5)	0.8
Alb (median, IQR) (*missing*)	32.5 (26.75-39) (missing=23)	36.5 (33.5-39.75) (missing=5)	0.6
Disease Duration (median, IQR, months)	16 (2-154)		

Table 70 - Separated granulocyte Patient demographics for targeted qPCR of DMRs

	EntrezID(Seed)	Symbol(Seed)	Number of genes in network	Modularity	FDR p Value
1	1991	ELANE	74	3.983128	0
2	56616	DIABLO	7	4.780265	0.017
3	246778	IL27	12	3.555689	0.032
4	4353	MPO	41	4.722692	0
5	566	AZU1	18	4.01307	0.01
6	5265	SERPINA1	14	4.26562	0.008
7	3082	HGF	13	5.205599	0.002
8	7188	TRAF5	7	4.296099	0.016
9	4680	CEACAM6	5	4.96616	0.013
10	306	ANXA3	5	4.470587	0.038
11	2161	F12	9	4.43636	0.013

Table 71 - Functional epigenetic module Crohn's disease versus Control in Whole blood. Modularity=average of edge weights. P Values are calculated using the Monte-Carlo procedure (a permutation test, n=1000)

	EntrezID(Seed)	Symbol(Seed)	Number of genes in network	Modularity	FDR p Value
1	80331	DNAJC5	45	2.350306	0.01
2	306	ANXA3	5	4.043512	0.02
3	1991	ELANE	87	3.404352	0
4	56616	DIABLO	13	3.296102	0.008
5	4353	MPO	38	3.457653	0
6	863	CBFA2T3	9	3.658851	0.016
7	5265	SERPINA1	14	3.224906	0.019
8	2161	F12	7	4.677451	0.011
9	3674	ITGA2B	7	5.934862	0
10	3082	HGF	12	4.786923	0
11	4210	MEFV	7	5.7651	0.002
12	1378	CR1	5	4.205504	0.011
13	7188	TRAF5	7	3.479061	0.021
14	4680	CEACAM6	5	3.621719	0.021
15	8915	BCL10	9	4.437442	0.008
16	634	CEACAM1	5	3.621719	0.037
17	4318	MMP9	21	4.114022	0.001
18	54210	TREM1	16	3.982474	0.003
19	3687	ITGAX	5	4.044724	0.011
20	7292	TNFSF4	5	3.560199	0.039

Table 72 - Functional epigenetic module for UC versus control in whole blood. Modularity=average of edge weights. P Values are calculated using the Monte-Carlo procedure (a permutation test, n=1000).

	EntrezID(Seed)	Symbol(Seed)	Number of genes in network	Modularity	FDR p Value
1	930	CD19	9	1.841151	0.014
2	23495	TNFRSF13B	8	2.160652	0.001
3	3552	IL1A	13	1.834739	0.002
4	22893	BAHD1	8	1.864792	0.043
5	1277	COL1A1	49	1.20666	0.044
6	29760	BLNK	19	1.650736	0.014
7	5592	PRKG1	11	2.46726	0.004
8	608	TNFRSF17	8	2.160652	0.002
9	22936	ELL2	66	1.153672	0.022
10	7349	UCN	16	1.450829	0.031
11	2815	GP9	5	2.057541	0.018

Table 73 - Functional epigenetic module IBD versus Control in CD8+ T Cells. Modularity=average of edge weights. P Values are calculated using the Monte-Carlo procedure (a permutation test, n=1000)

	EntrezID(Seed)	Symbol(Seed)	Number of genes in network	Modularity	FDR p Value
1	1230	CCR1	30	1.349744	0.04
2	6809	STX3	12	1.764312	0.008
3	5824	PEX19	17	1.506995	0.047
4	1280	COL2A1	27	1.720637	0.015
5	2767	GNA11	12	1.680574	0.029
6	3933	LCN1	8	2.392497	0
7	567	B2M	28	1.39469	0.033
8	10333	TLR6	15	1.616574	0.026
9	6457	SH3GL3	15	1.802954	0.023
10	1432	MAPK14	32	1.879617	0.04

Table 74 - Functional epigenetic module IBD versus Control in CD4+ T Cells. Modularity=average of edge weights. P Values are calculated using the Monte-Carlo procedure (a permutation test, n=1000)

	EntrezID(Seed)	Symbol(Seed)	Number of genes in network	Modularity	FDR p Value
1	1604	CD55	7	2.375207	0.004
2	7124	TNF	31	1.739603	0.001
3	911	CD1C	6	1.918549	0.024
4	4634	MYL3	55	1.403718	0.01
5	134	ADORA1	8	1.865922	0.024
6	10666	CD226	7	1.836013	0.029
7	634	CEACAM1	5	2.561193	0.005
8	1824	DSC2	6	1.872248	0.034
9	7412	VCAM1	10	2.500476	0
10	5819	PVRL2	7	1.836013	0.04
11	7881	KCNAB1	7	2.160383	0.005
12	55729	ATF7IP	12	2.1151	0.008
13	7158	TP53BP1	15	1.585348	0.047
14	4004	LMO1	6	1.94922	0.043

Table 75 - Functional epigenetic module IBD versus Control in CD14+ Monocytes. Modularity=average of edge weights. P Values are calculated using the Monte-Carlo procedure (a permutation test, n=1000)

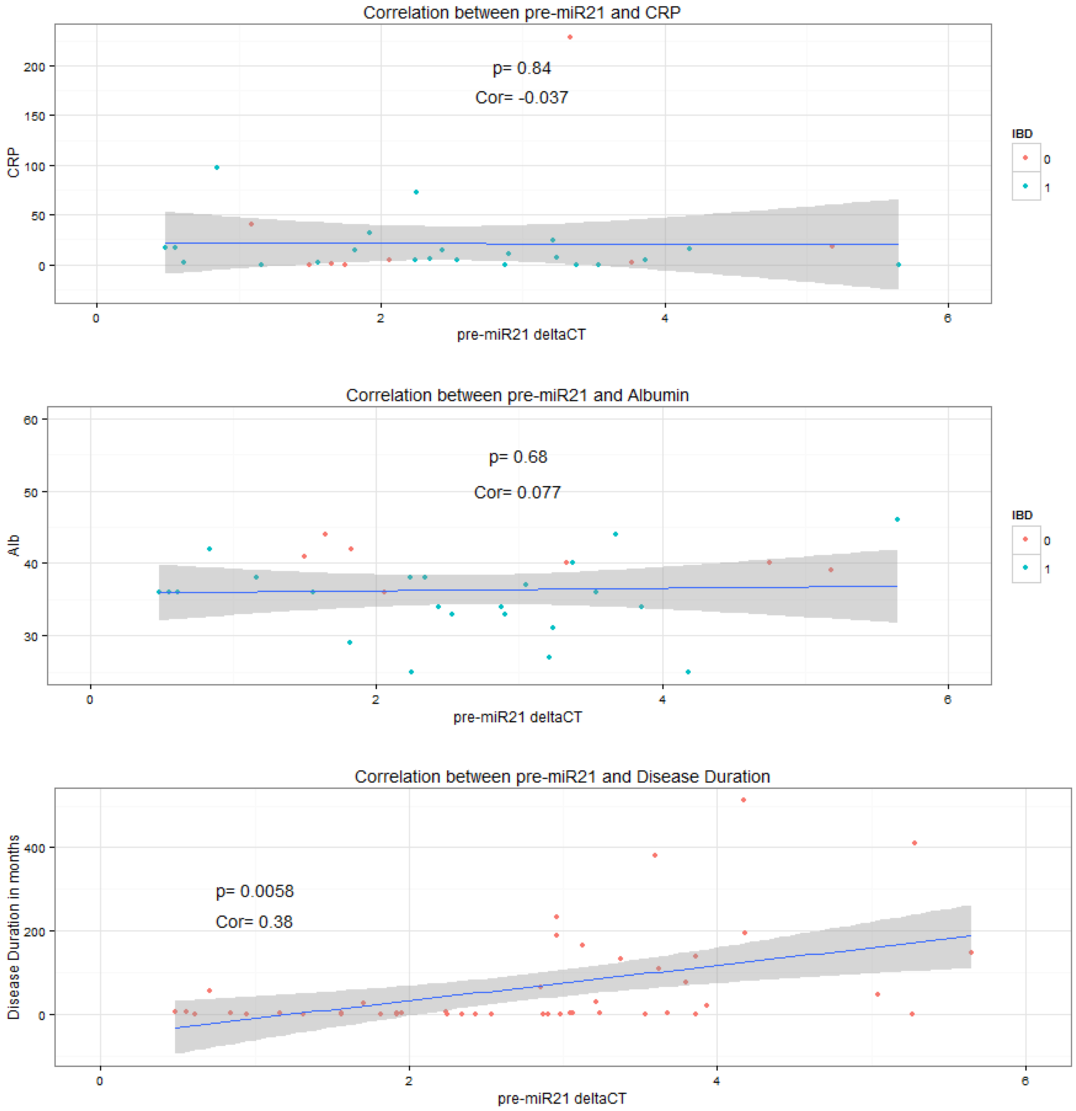


Figure 76 - Correlation between PBMC pri-miR21 expression and clinical parameters

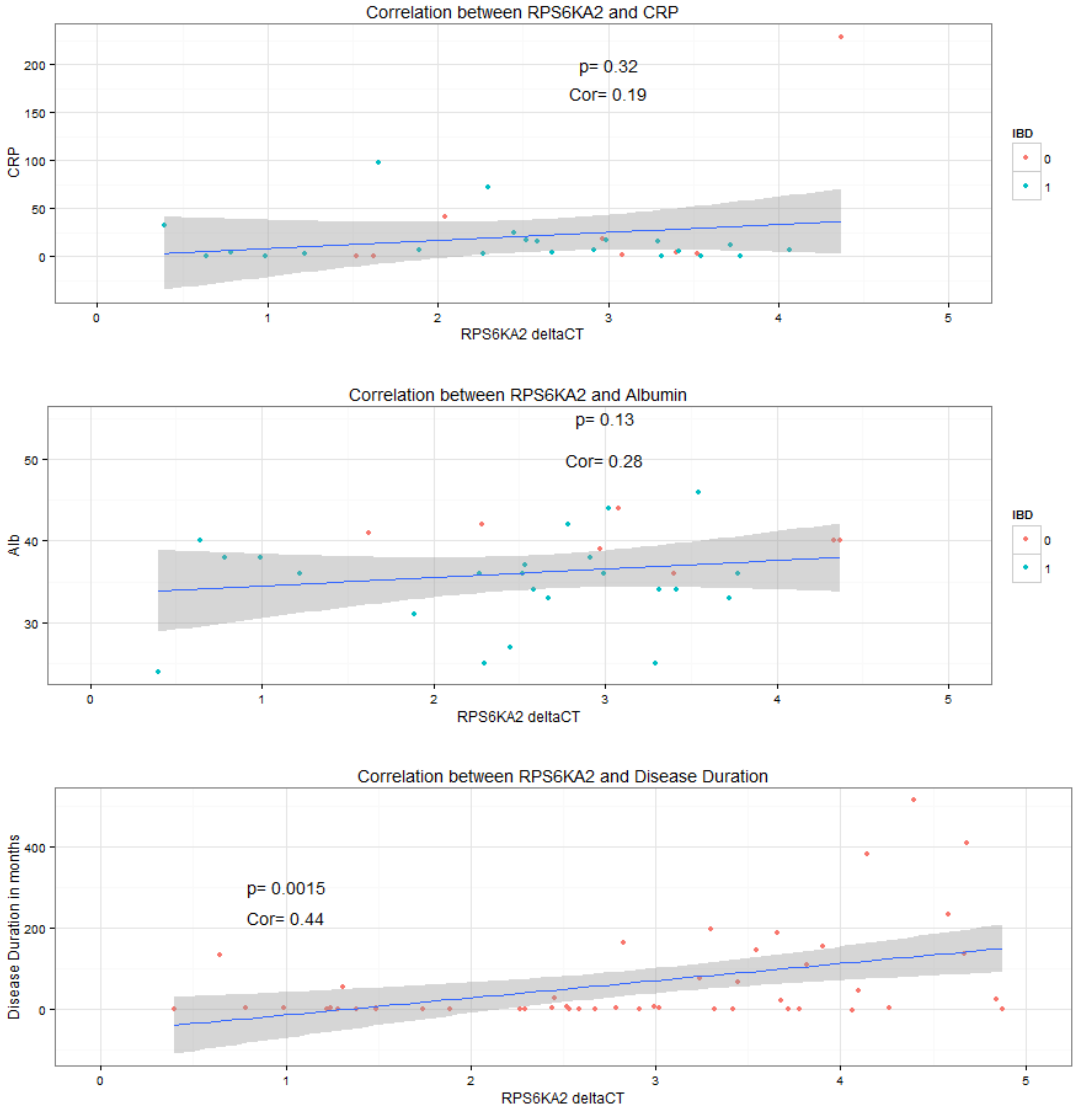


Figure 77 - Correlation between PBMC RPS6KA2 expression and clinical parameters

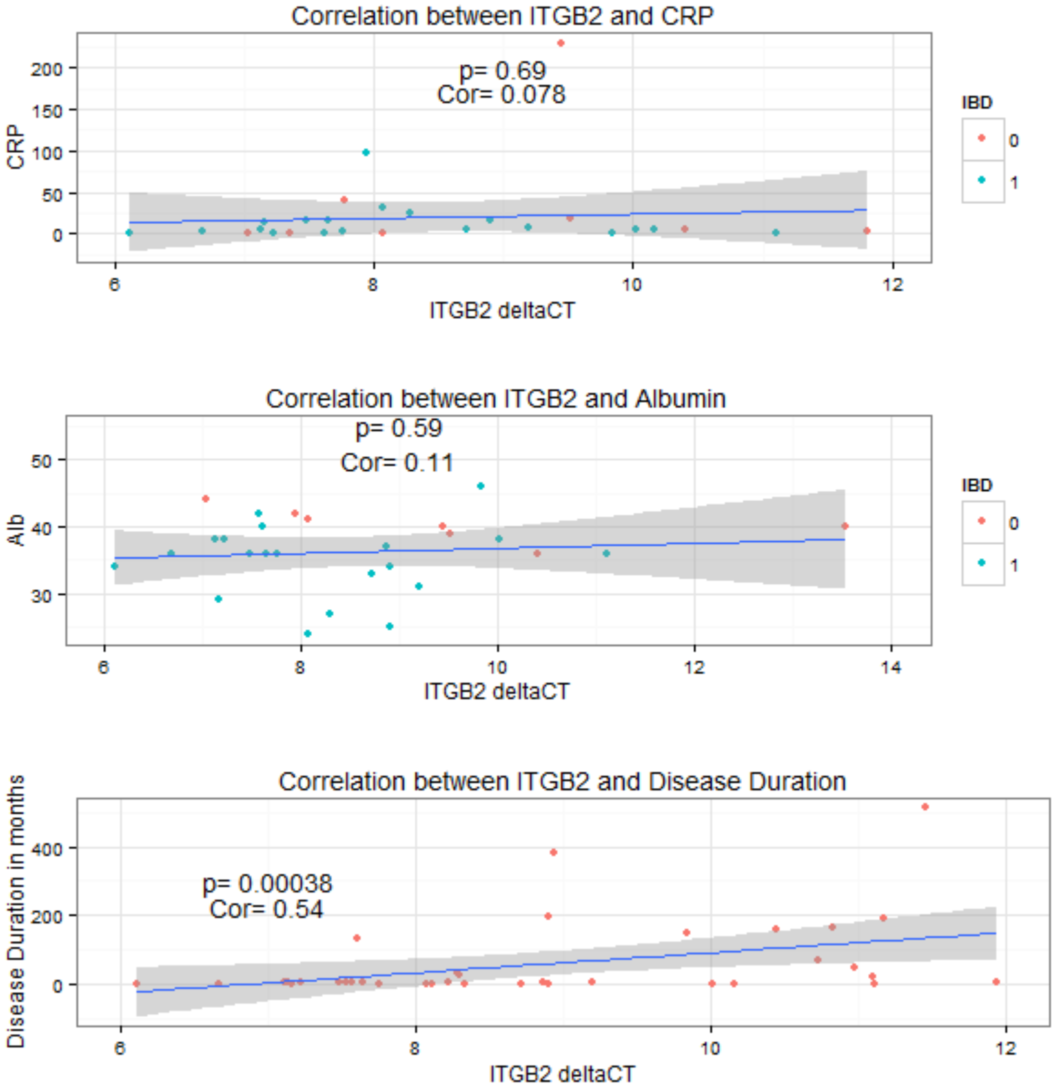


Figure 78 - Correlation between PBMC IGTB2 expression and clinical parameters

Appendix 5 – Chapter 7 Biomarkers

gene1	probe1	gene2	probe2
ARHGEF3	cg04389058	NA	cg09304397
ARHGEF3	cg04389058	TOLLIP	cg26599989
ARHGEF3	cg04389058	YWHAE	cg06219337
CDK6	cg14100946	NA	cg05740793
CSMD3	cg02292450	MIR1973	cg22914762
CSMD3	cg20323509	MIR1973	cg22914762
CSMD3	cg02292450	NA	cg27534567
CSMD3	cg20323509	NA	cg05740793
GNAS	cg10011623	NA	cg27534567
GNAS	cg01748573	NA	cg27534567
GNAS	cg26767990	NA	cg05740793
GPRIN3	cg02734358	D2HGDH	cg24743237
GPRIN3	cg02734358	NA	cg05740793
GPRIN3	cg02734358	NDUFS4	cg12351310
MIR1973	cg22914762	GNAS	cg01748573
MIR1973	cg22914762	TNS1	cg12338137
MIR21	cg27023597	NA	cg21653586
PWWP2B	cg07733247	NA	cg05740793
RPS6KA2	cg17501210	Sep-09	cg01749539
RPS6KA2	cg17501210	ANKRD11	cg16525838
RPS6KA2	cg17501210	ATP9A	cg07339236
RPS6KA2	cg17501210	HEATR2	cg10472711
RPS6KA2	cg17501210	HK2	cg27049094
RPS6KA2	cg17501210	NA	cg09304397
RPS6KA2	cg17501210	NA	cg09349128
RPS6KA2	cg17501210	NA	cg12992827
RPS6KA2	cg17501210	NMUR1	cg01616956
RPS6KA2	cg17501210	TOLLIP	cg26599989
RPS6KA2	cg17501210	YWHAE	cg06219337
RPS6KA2	cg17501210	ZBTB16	cg22768358
SOCS3	cg18181703	D2HGDH	cg13613174
TNFSF10	cg01059398	NA	cg21653586
TOLLIP	cg26599989	MYO1E	cg08423142
TOLLIP	cg26599989	NA	cg09304397
VMP1	cg12054453	Sep-09	cg01749539
VMP1	cg12054453	ITGB2	cg13315706
VMP1	cg12054453	MYO1E	cg08423142
VMP1	cg16936953	MYO1E	cg08423142
VMP1	cg12054453	NA	cg09304397
VMP1	cg16936953	NA	cg09304397
VMP1	cg12054453	TOLLIP	cg26599989
VMP1	cg16936953	TOLLIP	cg26599989
VMP1	cg12054453	YWHAE	cg06219337
VMP1	cg16936953	YWHAE	cg06219337

Table 76 - Methylation Probe Pairings described in Adams et al

Probe 1	Probe 2	AUC_CD	Rank CD	AUC_UC	Rank UC	AUC_IBD	Rank IBD
cg18181703	cg13613174	0.836	1	0.725	1	0.787	1
cg22914762	cg12338137	0.809	2	0.713	2	0.786	2
cg20323509	cg22914762	0.752	3	0.699	3	0.629	25
cg20323509	cg05740793	0.728	4	0.673	4	0.627	26
cg02734358	cg12351310	0.728	5	0.671	5	0.601	27
cg17501210	cg01616956	0.715	6	0.670	6	0.650	16
cg01059398	cg21653586	0.711	7	0.669	7	0.655	14
cg12054453	cg26599989	0.707	8	0.668	8	0.696	3
cg26767990	cg05740793	0.694	9	0.666	9	0.642	18
cg16936953	cg26599989	0.679	10	0.666	10	0.633	24
cg27023597	cg21653586	0.678	11	0.666	11	0.694	4
cg17501210	cg16525838	0.672	12	0.666	12	0.661	9
cg17501210	cg26599989	0.664	13	0.665	13	0.635	21
cg04389058	cg26599989	0.663	14	0.663	14	0.633	23
cg26599989	cg09304397	0.656	15	0.658	15	0.601	28
cg17501210	cg27049094	0.649	16	0.656	16	0.642	19
cg26599989	cg08423142	0.648	17	0.653	17	0.659	10
cg12054453	cg13315706	0.648	18	0.645	18	0.657	11
cg22914762	cg01748573	0.641	19	0.631	19	0.688	5
cg17501210	cg12992827	0.638	20	0.627	20	0.662	8
cg10011623	cg27534567	0.636	21	0.617	21	0.634	22
cg04389058	cg06219337	0.615	22	0.614	22	0.547	36
cg07733247	cg05740793	0.615	23	0.608	23	0.543	37
cg01748573	cg27534567	0.615	24	0.606	24	0.641	20
cg04389058	cg09304397	0.613	25	0.591	25	0.558	32
cg14100946	cg05740793	0.613	26	0.585	26	0.558	33
cg17501210	cg22768358	0.603	27	0.583	27	0.662	7
cg12054453	cg06219337	0.599	28	0.571	28	0.654	15
cg16936953	cg08423142	0.597	29	0.569	29	0.647	17
cg12054453	cg01749539	0.596	30	0.557	30	0.655	13
cg12054453	cg08423142	0.592	31	0.546	31	0.684	6
cg02734358	cg05740793	0.591	32	0.540	32	0.553	34
cg12054453	cg09304397	0.588	33	0.540	33	0.657	12
cg17501210	cg07339236	0.573	34	0.533	34	0.517	41
cg16936953	cg06219337	0.562	35	0.531	35	0.571	30
cg16936953	cg09304397	0.560	36	0.520	36	0.571	29
cg17501210	cg09349128	0.549	37	0.520	37	0.563	31
cg17501210	cg09304397	0.546	38	0.519	38	0.538	39
cg17501210	cg06219337	0.546	39	0.519	39	0.549	35

cg02734358	cg24743237	0.544	40	0.515	40	0.512	42
cg17501210	cg01749539	0.536	41	0.515	41	0.531	40
cg17501210	cg10472711	0.515	42	0.504	42	0.543	38

Table 77 - LDA comparing cases and controls using paired methylation markers described in Adams et al (AUC=Area under Receiver Operator Curve) for Crohn's disease (CD), ulcerative colitis (UC) and inflammatory bowel disease (IBD). Rank denotes the ranking of the biomarker probe sets for each disease group (ordered on rank in CD).

classifier	AUC	Sensitivity	Specificity	Misclassification rate	Brier score	Average probability	Method of ranking variables
Lasso1	0.865	0.783	0.733	0.239	0.317	0.642	Lasso
LDA	0.859	0.729	0.786	0.246	0.310	0.707	T-test
FDA	0.859	0.700	0.833	0.241	0.498	0.501	T-test
Lasso5	0.854	0.775	0.759	0.232	0.317	0.658	Elastic
Lasso2	0.853	0.775	0.754	0.234	0.317	0.668	Boost
DLDA5	0.847	0.688	0.838	0.246	0.434	0.756	Elastic
Lasso4	0.845	0.821	0.686	0.239	0.366	0.592	Forest
plr4	0.844	0.767	0.754	0.239	0.320	0.684	Forest
DLDA1	0.844	0.721	0.827	0.232	0.394	0.754	Lasso
LDA4	0.844	0.754	0.806	0.223	0.322	0.689	Forest
FDA4	0.844	0.708	0.853	0.227	0.498	0.501	Forest
Lasso	0.843	0.767	0.738	0.246	0.319	0.662	T-test
pls_rf4	0.838	0.792	0.728	0.237	0.330	0.687	Forest
plr	0.836	0.763	0.743	0.246	0.328	0.678	T-test
LDA2	0.834	0.738	0.770	0.248	0.336	0.698	Boost
FDA2	0.834	0.713	0.806	0.246	0.498	0.501	Boost
DLDA2	0.826	0.704	0.828	0.241	0.433	0.746	Boost
LDA5	0.817	0.721	0.733	0.274	0.369	0.692	Elastic
FDA5	0.817	0.683	0.765	0.281	0.498	0.501	Elastic
DLDA4	0.817	0.700	0.770	0.269	0.479	0.733	Forest
QDA2	0.808	0.738	0.759	0.253	0.385	0.711	Boost
QDA	0.805	0.738	0.701	0.278	0.430	0.708	T-test

Table 78 - CMA package classifier comparison methods ordered according to area under receiving operating curve (AUC). (Classification methods = lasso = least absolute shrinkage and selection operator[415,416], LDA = linear discriminant analysis[469], FDA = Fisher's discriminant analysis[470], DLDA = diagonal discriminant analysis[469], plr = penalised logistic regression[471], pls_rf = Partial least squares random forest[472], QDA = quadratic discriminant analysis) (Method of ranking variables = T-test according to T-statistic, Forest= random forest[473], Boost= component-wise boosting[474], elastic = elastic net[475], lasso = lasso[415,416]).

Norm fraction	Number of methylation probes included	AUC	Sensitivity	Specificity	Misclassification rate
0.3	87	0.858	0.634	0.873	0.221
0.2	61	0.877	0.634	0.873	0.221
0.16	48	0.887	0.659	0.873	0.212
0.15	47	0.889	0.659	0.873	0.212
0.14	47	0.889	0.659	0.873	0.212
0.13	42	0.89	0.659	0.889	0.202
0.12	38	0.888	0.659	0.889	0.202
0.11	31	0.885	0.659	0.889	0.202
0.1	26	0.883	0.683	0.889	0.192
0.09	22	0.879	0.659	0.905	0.192
0.08	19	0.878	0.659	0.905	0.192
0.07	12	0.875	0.634	0.905	0.202
0.06	10	0.874	0.683	0.937	0.163
0.05	8	0.875	0.683	0.968	0.144
0.04	7	0.87	0.585	0.952	0.192
0.03	4	0.863	0.512	0.968	0.212
0.02	2	0.853	0.36	0.984	0.26
0.01	1	0.853	0.049	1	0.375

Table 79 - Lasso CD versus HC. Tuning of lasso algorithm to altering the shrinkage intensity and thus the number of methylation probes included in the model.

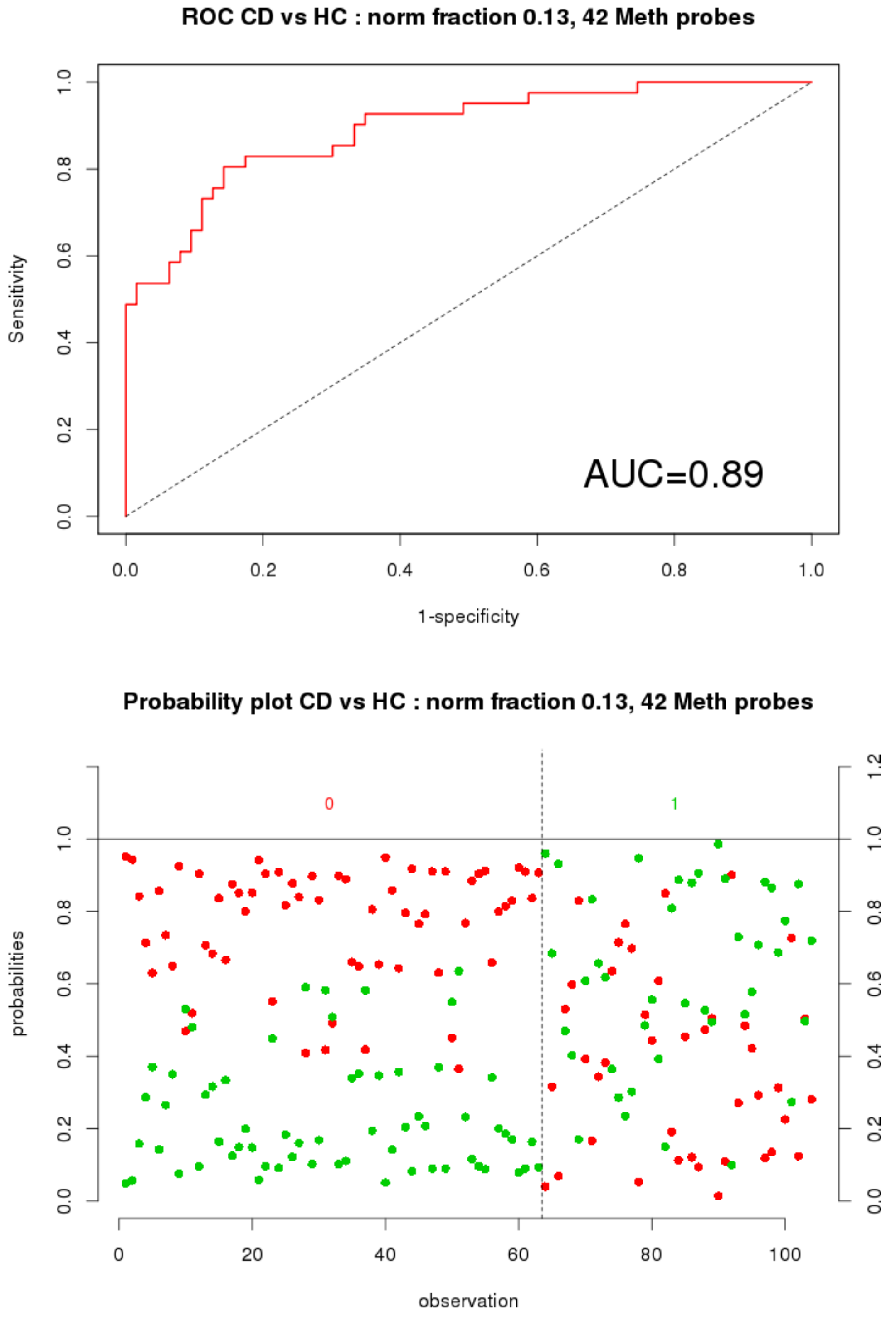


Figure 79 – Lasso modelling to discriminate Crohn's disease from controls. Receiver operator curve for CD versus control. 42 methylation probes selected using lasso penalized regression. Top: Receiver operator curve for Lasso selected probes to distinguish CD from controls using 30 methylation probes. Bottom: Probability plot. 0/red = controls, 1/green = CD cases.

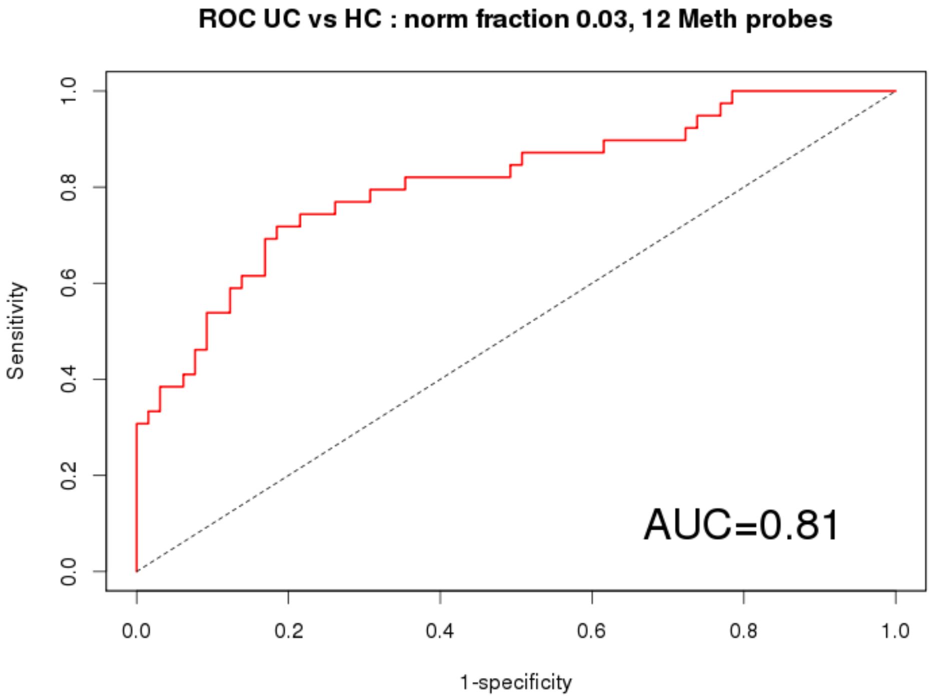
ProbeId	absolute	Chr	GeneSymbol	logFC	P.Value	adj.P.Val

	value of regression coefficient					
cg17501210	9.501	chr6	RPS6KA2	-0.08	2.71E-22	1.22E-16
cg05487424	6.148	chr2	RPIA	0.00	6.44E-05	1
cg03956353	5.496	chr1	EPS8L3	0.01	0.002766	1
cg03538833	5.211	chr3	LOC152225	0.01	0.001368	1
cg22881435	4.375	chr8	RAB11FIP1	0.02	1.09E-11	4.89E-06
cg24319178	4.156	chr6	NA	0.00	0.098107	1
cg25105536	3.863	chr6	KLHL32	0.00	0.00234	1
cg03546163	3.376	chr6	FKBP5	-0.06	3.93E-15	1.76E-09
cg07641807	2.992	chr13	NA	-0.02	8.87E-07	0.397195
cg17147182	2.875	chr2	NA	0.00	0.0024	1
cg04691264	2.737	chr10	PTCHD3P1	0.02	3.76E-06	1
cg14153654	2.496	chr1	TNFRSF9	0.01	0.065064	1
cg15591678	2.331	chr10	ZNF365	0.01	0.000651	1
cg12054453	2.194	chr17	VMP1	-0.07	3.98E-17	1.78E-11
cg17078686	2.078	chr2	POU3F3	0.01	0.000924	1
cg12357606	1.572	chr4	CORIN	0.01	0.00067	1
cg19628456	1.481	chr6	HSPA1L	0.00	0.052486	1
cg25114611	1.334	chr6	NA	-0.04	1.10E-18	4.93E-13
cg02297838	1.330	chr13	NA	-0.02	3.04E-11	1.36E-05
cg21383151	1.293	chr10	TBC1D12	0.01	0.000129	1
cg04138502	1.036	chr3	ADCY5	0.01	8.90E-07	0.398548
cg16166062	0.883	chr10	NA	-0.01	0.006794	1
cg07392460	0.802	chr2	NA	0.00	0.001946	1
cg27389562	0.778	chr19	CEACAM8	-0.01	0.008141	1
cg04666911	0.716	chr11	LSP1	0.02	4.84E-08	0.02169
cg03254161	0.698	chr17	DNAH17	0.01	0.16172	1
cg15877906	0.662	chr12	ATF1	0.00	0.712024	1
cg01487195	0.604	chr10	BMPR1A	0.00	0.061407	1
cg15025240	0.544	chr2	M1AP	-0.03	5.59E-07	0.250392
cg25591573	0.543	chr19	ZNF442	0.01	0.000588	1
cg14118946	0.519	chr9	NA	0.01	0.009499	1
cg24430034	0.474	chr13	NA	0.02	7.09E-13	3.18E-07
cg00382138	0.437	chr4	CFI	-0.03	2.45E-11	1.10E-05
cg13807509	0.271	chr11	FOSL1	0.00	0.015506	1
cg01395047	0.238	chr3	TLR9	0.01	0.000136	1
cg09479241	0.182	chr17	TLCD1	-0.02	0.000251	1
cg03526905	0.126	chr17	NA	0.00	0.134838	1
cg25676074	0.068	chr8	NCOA2	0.01	0.000218	1
cg15118665	0.068	chr8	NA	0.01	0.00675	1
cg03524147	0.065	chr10	NA	-0.01	0.04408	1
cg15168577	0.055	chr5	NA	-0.01	0.292116	1
cg20072241	0.010	chr2	RAB3GAP1	0.00	0.208996	1

Table 80 - Panel of Methylation probes selected by lasso algorithm to differentiate CD from control. (Δβ = difference in beta values between IBD versus control, p.value and Holm adjusted p values derived from linear models IBD versus control with age, sex and estimated cell proportions as covariates)

Norm fraction	Number of methylation probes included	AUC	Sensitivity	Specificity	Misclassification rate
0.2	84	0.722	0.462	0.862	0.288
0.15	66	0.733	0.462	0.862	0.288
0.1	50	0.772	0.333	0.877	0.327
0.06	27	0.794	0.385	0.923	0.279
0.05	23	0.803	0.385	0.938	0.269
0.04	20	0.808	0.41	0.91	0.269
0.03	12	0.81	0.385	0.938	0.269
0.02	3	0.806	0.359	0.969	0.26
0.01	1	0.798	0.179	1	0.308

Table 81 – Lasso UC versus HC. Tuning of lasso algorithm to altering the shrinkage intensity and thus the number of methylation probes included in the model.



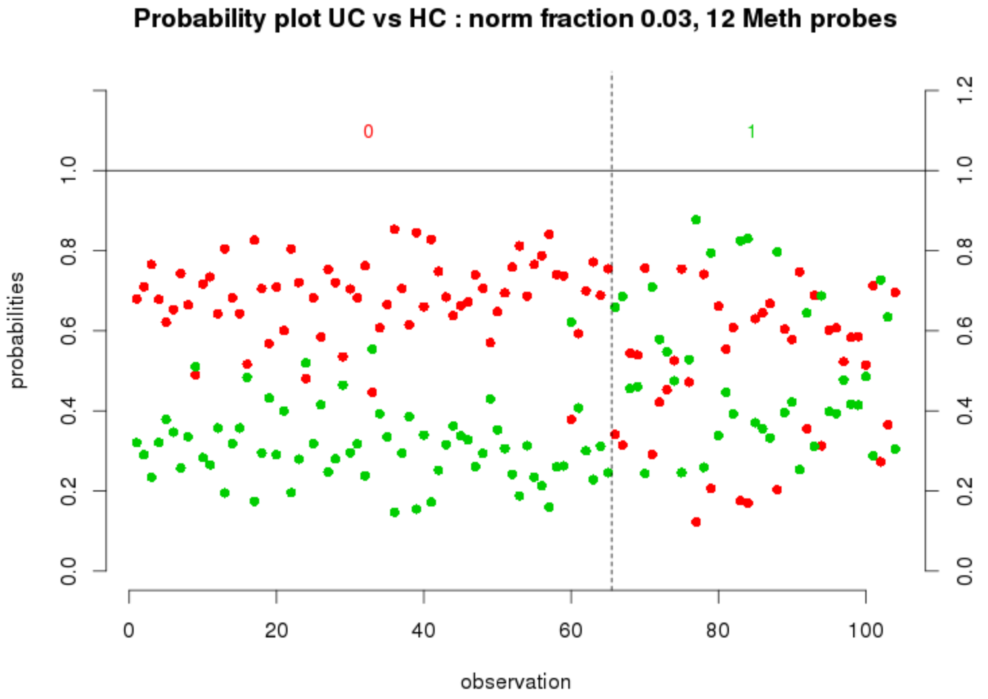


Figure 80 – Lasso modelling to discmrinate patients with ulcerative colitis compared with controls. Top: Receiver operator curve for Lasso selected probes to distinguish UC from controls using 30 methylation probes. Bottom: Probability plot. 0/red = controls, 1/green = UC cases.

absolute value of regression coefficient	ProbeID	Chromosome	GeneSymbol	logFC	P.Value	adj.P.Val
10.08	cg25114611	chr6	NA	-0.04	1.10E-18	4.93E-13
1.43	cg08423142	chr15	MYO1E	-0.02	7.69E-11	3.45E-05
0.85	cg07398517	chr3	NA	-0.04	6.14E-16	2.75E-10
0.79	cg22881435	chr8	RAB11FIP1	0.02	1.09E-11	4.89E-06
0.73	cg19821297	chr19	NA	-0.06	3.66E-17	1.64E-11
0.70	cg03013636	chr16	NA	0.01	0.070511	1
0.45	cg17501210	chr6	RPS6KA2	-0.08	2.71E-22	1.22E-16
0.20	cg04304450	chr22	BIK	0.03	1.88E-08	0.008412
0.05	cg17515347	chr1	AIM2	-0.05	1.10E-11	4.92E-06
0.04	cg07035454	chr8	OXR1	0.01	0.00809	1
0.02	cg13772414	chr2	EPHA4	0.02	5.14E-07	0.230221
0.00	cg00254470	chr8	PVT1	0.03	2.85E-06	1

Table 82 - Panel of Methylation probes selected by lasso algorithm to differentiate UC from control. (Δβ = difference in beta values between IBD versus control, p.value and Holm adjusted p values derived from linear models IBD versus control with age, sex and estimated cell proportions as covariates)

Norm fraction	Number of methylation probes included	AUC	Sensitivity	Specificity	Misclassification rate
0.2	95	0.649	0.667	0.583	0.383
0.15	80	0.659	0.667	0.5	0.433
0.1	60	0.667	0.667	0.5	0.433
0.05	37	0.712	0.792	0.5	0.383
0.03	25	0.716	0.875	0.389	0.417
0.02	21	0.714	0.917	0.194	0.517
0.01	19	0.719	1	0.111	0.533
0.005	15	0.718	1	0	0.6

Table 83 - Lasso UC versus CD. Tuning of lasso algorithm to altering the shrinkage intensity and thus the number of methylation probes included in the model

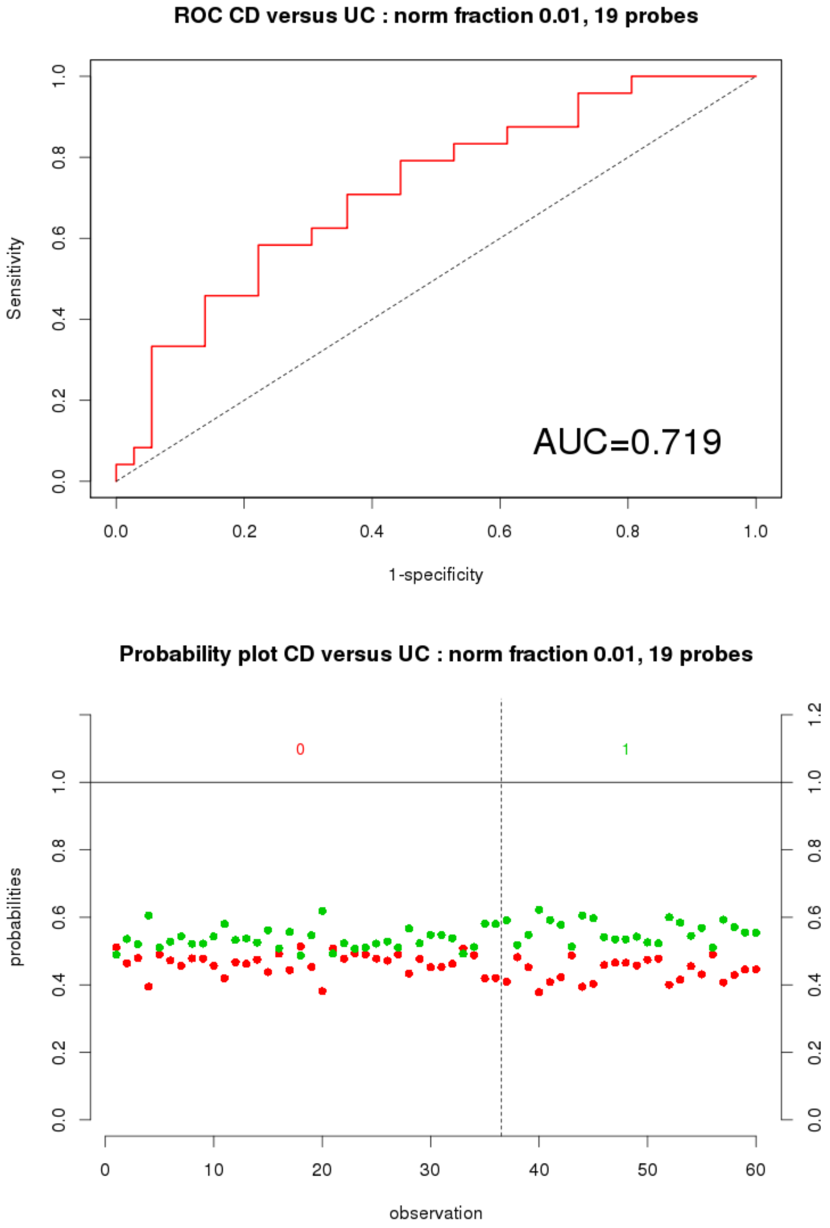


Figure 81 – Lasso modelling to differentiate Crohn's disease (CD) from ulcerative colitis (UC).Top: Receiver operator curve for Lasso selected probes to distinguish UC from CD using 19 methylation probes. Bottom: Probability plot. 0/red = controls, 1/green = UC cases.

absolute value of regression coefficient	ProbeID	Chromosome	GeneSymbol	logFC	P.Value	adj.P.Val
3.94	cg04879696	chr16	SPG7	0.006	7.66E-06	1
2.07	cg07390924	chr1	PLEKHN1	0.007	4.36E-04	1
1.71	cg13492133	chr1	SMG7	-0.010	2.50E-04	1
1.49	cg10502206	chr2	ZEB2	0.023	1.08E-06	0.49
1.46	cg03776464	chr2	EPHA4	-0.004	2.07E-04	1
0.78	cg17964532	chr6	CDKN1A	-0.008	9.16E-05	1
0.52	cg21911019	chr4	GPR78	0.030	8.37E-05	1
0.35	cg19683494	chr5	NA	-0.065	5.94E-07	0.27
0.35	cg07080946	chr16	LUC7L	-0.008	2.30E-03	1
0.30	cg10904715	chr13	NA	0.004	3.56E-05	1
0.16	cg05304729	chr1	MNDA	-0.064	3.44E-07	0.15
0.16	cg20569772	chr2	NA	-0.011	1.63E-03	1
0.06	cg15409712	chr4	NFKB1	-0.017	5.72E-03	1
0.06	cg12547959	chr5	TRIO	-0.026	1.32E-03	1
0.01	cg18181703	chr17	SOCS3	-0.042	2.59E-05	1
0.00	cg13847437	chr9	NA	0.005	4.87E-03	1
0.00	cg13708869	chr17	CDC6	-0.006	2.48E-03	1
0.00	cg04792065	chr15	TTBK2	-0.006	4.85E-04	1
0.00	cg22322277	chr6	NA	0.035	2.18E-04	1

Table 84 - Panel of Methylation probes selected by lasso algorithm to differentiate UC from CD. (Δβ = difference in beta values between IBD versus control, p.value and Holm adjusted p values derived from linear models CD versus UC with age, sex and estimated cell proportions as covariates)

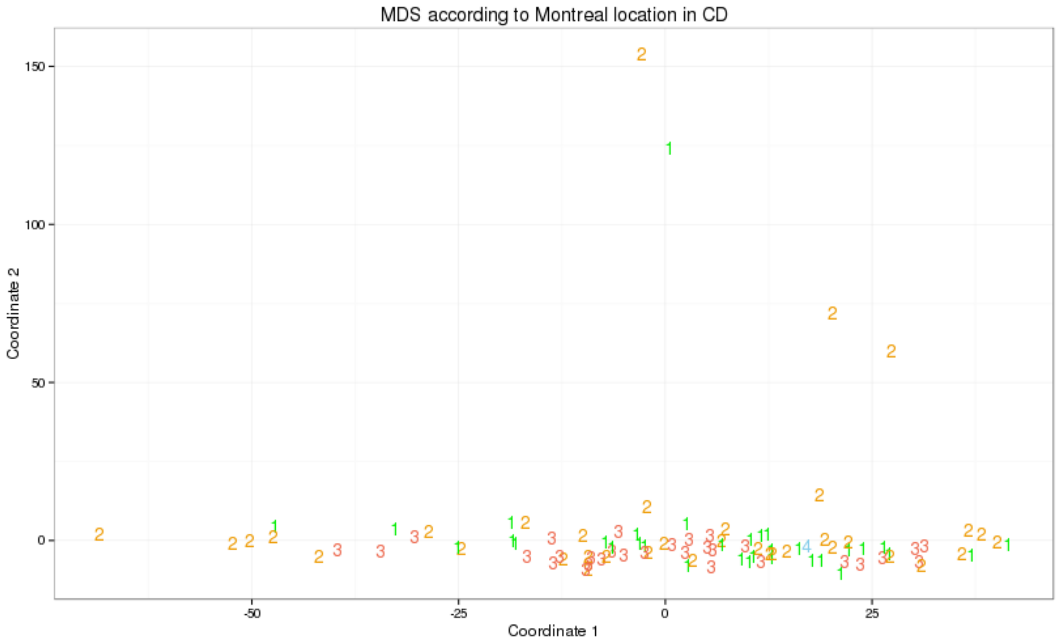


Figure 82 - multidimensional scaling plot (MDS) of all methylation data according to Montreal location in CD

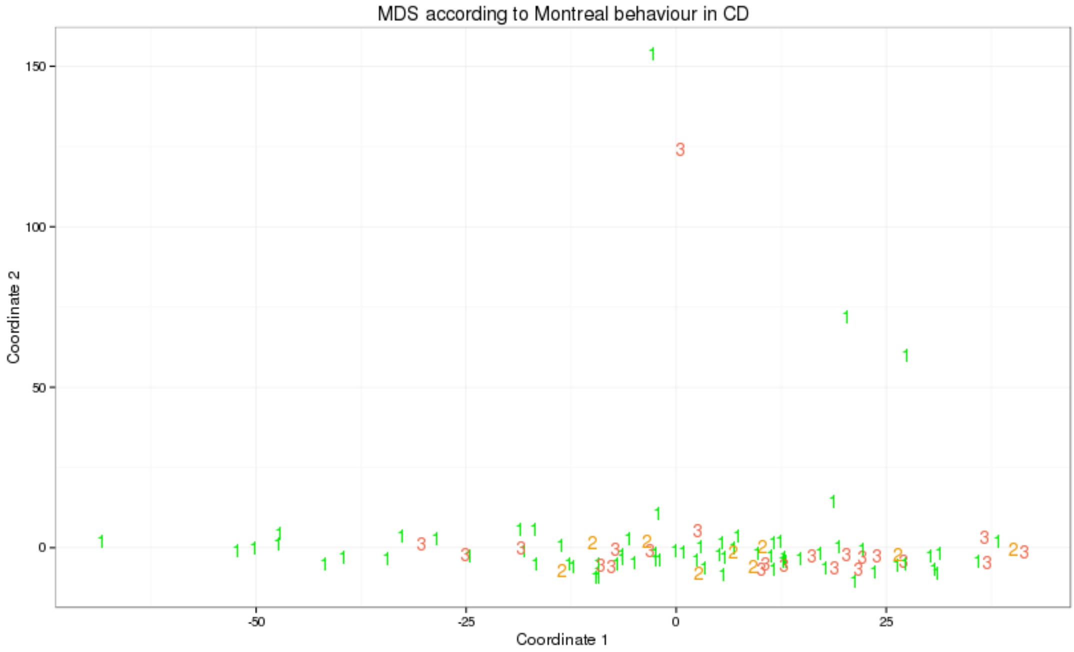


Figure 83 - multidimensional scaling plot (MDS) of all methylation data according to Montreal behaviour in CD

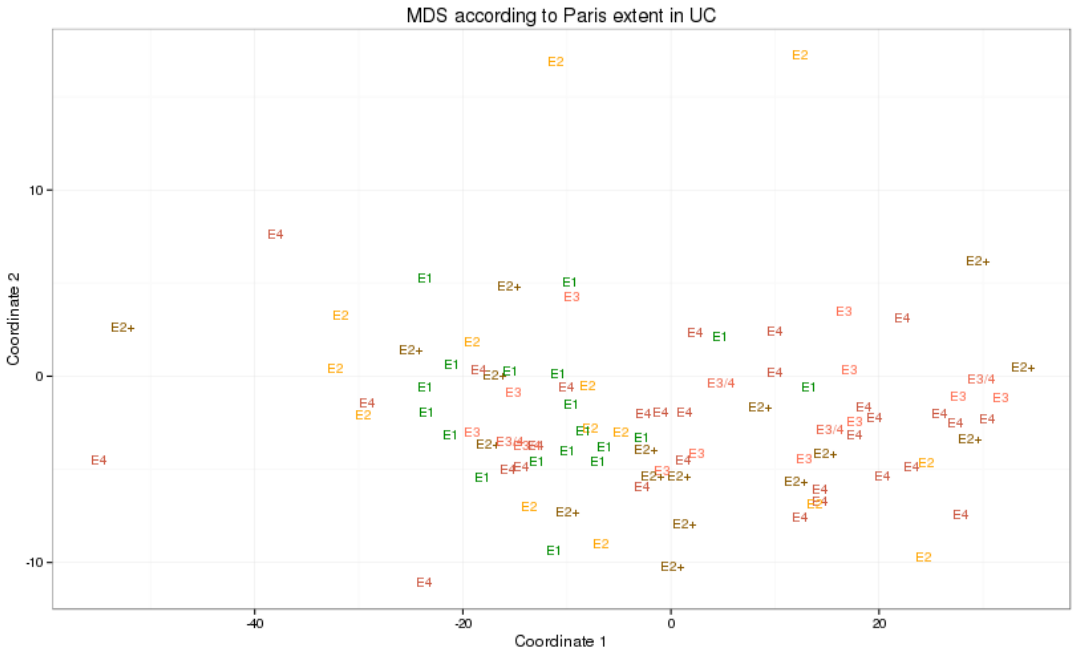


Figure 84 - multidimensional scaling plot (MDS) of all methylation data according to Montreal extent in UC

Appendix 6 - Chapter 8 microRNAs in IBD

Table 85 – Summary of sample pooling according to PCR band intensity and barcode

	Pool 1:High(CD14: CDx4,HCx4)				**Volume (µL)**
1	A5	2	CD14	HC	10
2	C6	3	CD14	CD	10
3	A3	4	CD8	CD	5
4	C10	7	CD14	CD	10
5	B5	8	CD14	CD	5
6	B6	9	CD4	CD	5
7	A12	10	CD14	CD	7
8	D5	14	CD14	HC	10
9	B9	18	CD14	HC	10
10	B10	19	CD14	HC	7
11	D6	20	CD8	CD	10
12	C1	22	CD8	CD	7
	Pool 2:High				**Volume (µL)**
1	A2	3	CD8	HC	10
2	A6	5	CD4	HC	7
3	A8	6	CD14	CD	7
4	A10	8	CD8	HC	10
5	D3	11	CD8	HC	7
6	D4	13	CD14	CD	10
7	B7	14	CD14	CD	10
8	B8	16	CD4	CD	7
9	B11	20	CD14	HC	5
10	B12	21	CD14	CD	5
11	C2	23	CD14	HC	5

Pool 5: Low (CD4: CDx4,HCx5)			
C5	2	CD4	HC
C7	4	CD4	CD
C9	6	CD4	HC
C12	8	CD4	CD
A11	9	CD8	CD
D2	10	CD4	CD
B1	11	CD4	HC
B2	13	CD4	HC
A1	14	CD8	CD
A4	16	CD14	HC
A7	18	CD4	CD
C3	25	CD4	HC

Pool 6: Low (CD8: CDx3,HCx3)			
C8	5	CD8	HC
A9	7	CD8	CD
D1	9	CD14	HC
B3	14	CD8	CD
B4	16	CD8	HC
C11	19	CD4	HC
D7	21	CD8	HC
D8	22	CD4	CD
D10	23	CD8	CD
D11	25	CD4	HC
C4	27	CD4	CD

Table 86 – Summary of total RNA sample concentration and quality assessment

PatientId	SampleTypeId	Concentration ng/ul	Total RNA ng	Total RNA in ug	RIN number
8903	MicroRNA (CD4)	50.9	3054	3.054	8.1
8903	MicroRNA (CD8)	24	1440	1.44	8
8903	MicroRNA (monocy	266	15960	15.96	NA
8908	MicroRNA (CD4)	23.5	1410	1.41	8.8
8908	MicroRNA (CD8)	27.8	1668	1.668	9.1
8908	MicroRNA (monocy	203.6	12216	12.21	8
8911	MicroRNA (CD4)	49.5	2970	2.97	9.1
8911	MicroRNA (CD8)	15.1	906	0.906	9.5
8911	MicroRNA (monocy	90.5	5430	5.43	8.7
8915	MicroRNA (CD4)	62.7	3762	3.762	8.2
8915	MicroRNA (CD8)	62.8	3768	3.768	9.2
8915	MicroRNA (monocy	135.2	8112	8.112	9.2
8926	MicroRNA (CD4)	60.1	3606	3.606	8.2
8926	MicroRNA (CD8)	95.8	5748	5.748	
8926	MicroRNA (monocy	90.8	5448	5.448	8.9
8927	MicroRNA (CD4)	37.7	2262	2.262	NA
8927	MicroRNA (CD8)	38.8	2328	2.328	8.7
8927	MicroRNA (monocy	37.9	2274	2.274	8.8
8929	MicroRNA (CD4)	45.5	2730	2.73	8.1
8929	MicroRNA (CD8)	11.4	684	0.684	8.8
8929	MicroRNA (monocy	87.8	5268	5.268	8.9
8940	MicroRNA (CD4)	39.7	2382	2.382	9.4
8940	MicroRNA (CD8)	29.8	1788	1.788	9.2
8940	MicroRNA (monocy	82.3	4938	4.938	8.7
8954	MicroRNA (CD4)	33.6	2016	2.016	9.4
8954	MicroRNA (CD8)	50	3000	3	8.5
8954	MicroRNA (monocy	258.5	15510	15.51	9
8983	MicroRNA (CD4)	33.8	2028	2.028	8.7
8983	MicroRNA (CD8)	66.3	3978	3.978	9.4
8983	MicroRNA (monocy	142.7	8562	8.562	8.6
8994	MicroRNA (CD4)	64.7	3882	3.882	8.3
8994	MicroRNA (CD8)	53.4	3204	3.204	8.6

8994	MicroRNA (monocy	217.8	13068	13.06	9.3
9003	MicroRNA (CD4)	141.2	8472	8.472	8.5
9003	MicroRNA (CD8)	28.2	1692	1.692	9.3
9003	MicroRNA (monocy	232.6	13956	13.95	9.4
9039	MicroRNA (CD4)	35.2	2112	2.112	8.9
9039	MicroRNA (CD8)	25	1500	1.5	9.7
9039	MicroRNA (monocy	78.1	4686	4.686	8.6
9049	MicroRNA (CD4)	47.5	2850	2.85	8.9
9049	MicroRNA (CD8)	43.2	2592	2.592	9.1
9049	MicroRNA (monocy	112.5	6750	6.75	8.9
9050	MicroRNA (CD4)	82.4	4944	4.944	8.6
9050	MicroRNA (CD8)	76.5	4590	4.59	9
9050	MicroRNA (monocy	94.7	5682	5.682	9.5
9054	MicroRNA (CD4)	81.3	4878	4.878	7.9
9054	MicroRNA (CD8)	58.1	3486	3.486	8.6
9054	MicroRNA (monocy	101.2	6072	6.072	8.4

KEGG Pathway	Gene Name	Found Genes	-ln(p-value)	KEGG Pathway ID
Prostate cancer	BCL2, CCNE1, E2F3, IGF1R, MAP2K1, PIK3R1, CREB5, AKT3, CCND1	9	22.13	hsa05215
Melanoma	E2F3, IGF1R, MAP2K1, PIK3R1, FGF7, FGF2, AKT3, CCND1	8	21.65	hsa05218
Glioma	E2F3, IGF1R, MAP2K1, PIK3R1, AKT3, CCND1	6	12.82	hsa05214
Colorectal cancer	BCL2, IGF1R, MAP2K1, PIK3R1, DCC, AKT3, CCND1	7	12.78	hsa05210
mTOR signaling pathway	PIK3R1, VEGFA, ENSG00000164327, AKT3, EIF4E	5	11.98	hsa04150
p53 signaling pathway	CCNE1, CHEK1, PPM1D, CCND2, CCND3, CCND1	6	11.85	hsa04115
Pancreatic cancer	E2F3, MAP2K1, PIK3R1, VEGFA, AKT3, CCND1	6	10.8	hsa05212
Non-small cell lung cancer	E2F3, MAP2K1, PIK3R1, AKT3, CCND1	5	10.09	hsa05223
Cell cycle	CCNE1, E2F3, WEE1, CHEK1, CCND2, CCND3, CCND1	7	8.81	hsa04110
Small cell lung cancer	BCL2, CCNE1, E2F3, PIK3R1, AKT3, CCND1	6	8.78	hsa05222
Bladder cancer	E2F3, MAP2K1, VEGFA, CCND1	4	8.09	hsa05219
Focal adhesion	BCL2, IGF1R, MAP2K1, PIK3R1, VEGFA, CCND2, CCND3, AKT3, CCND1	9	7.58	hsa04510
VEGF signaling pathway	MAP2K1, PIK3R1, VEGFA, AKT3, PPP3CB	5	7.32	hsa04370
Chronic myeloid leukemia	E2F3, MAP2K1, PIK3R1, AKT3, CCND1	5	6.53	hsa05220
Endometrial cancer	MAP2K1, PIK3R1, AKT3, CCND1	4	6.16	hsa05213
Wnt signaling pathway	WNT3A, FOSL1, BTRC, CCND2, CCND3, CCND1, PPP3CB	7	6.02	hsa04310
Apoptosis	BCL2, IRAK2, PIK3R1, AKT3, PPP3CB	5	5.85	hsa04210
Acute myeloid leukemia	MAP2K1, PIK3R1, AKT3, CCND1	4	5.72	hsa05221
Jak-STAT signaling pathway	PIK3R1, GHR, CCND2, SPRED1, CCND3, AKT3, CCND1	7	5.62	hsa04630
C5-Branched dibasic acid metabolism	SUCLA2	1	4.98	hsa00660
Ubiquitin mediated proteolysis	SMURF1, BTRC, FBXW7, TRIM37, UBE4B, SMURF2	6	4.77	hsa04120
Renal cell carcinoma	MAP2K1, PIK3R1, VEGFA, AKT3	4	4.24	hsa05211
Axon guidance	EFNB2, SEMA6D, EPHA7, DCC, PPP3CB	5	3.11	hsa04360

TGF-beta signaling pathway	SMURF1, ACVR2B, SMURF2, BMPR1A	4	2.83	hsa04350
Hedgehog signaling pathway	WNT3A, BTRC, IHH	3	2.78	hsa04340
B cell receptor signaling pathway	PIK3R1, AKT3, PPP3CB	3	2.43	hsa04662
Thyroid cancer	MAP2K1, CCND1	2	2.32	hsa05216
Fc epsilon RI signaling pathway	MAP2K1, PIK3R1, AKT3	3	1.65	hsa04664
Neurodegenerative Diseases	BCL2, FBXW7	2	1.5	hsa01510
Thiamine metabolism	ENSG00000107902	1	1.28	hsa00730
ErbB signaling pathway	MAP2K1, PIK3R1, AKT3	3	1.22	hsa04012
Insulin signaling pathway	MAP2K1, PIK3R1, AKT3, EIF4E	4	1.2	hsa04910
T cell receptor signaling pathway	PIK3R1, AKT3, PPP3CB	3	1.16	hsa04660
Taurine and hypotaurine metabolism	BAAT	1	1.12	hsa00430
Valine, leucine and isoleucine biosynthesis	PDHA1	1	0.99	hsa00290
Protein export	SRPR	1	0.99	hsa03060
Toll-like receptor signaling pathway	MAP2K1, PIK3R1, AKT3	3	0.88	hsa04620
Reductive carboxylate cycle (CO2 fixation)	SUCLA2	1	0.88	hsa00720
Prion disease	BCL2	1	0.7	hsa05060
Long-term potentiation	MAP2K1, PPP3CB	2	0.57	hsa04720
Natural killer cell mediated cytotoxicity	MAP2K1, PIK3R1, PPP3CB	3	0.51	hsa04650
Riboflavin metabolism	ENSG00000107902	1	0.5	hsa00740
Glycolysis / Gluconeogenesis	PDHA1	1	0.43	hsa00010
PPAR signaling pathway	ACSL4	1	0.43	hsa03320
MAPK signaling pathway	MAP2K1, FGF7, FGF2, AKT3, PPP3CB	5	0.41	hsa04010
Adipocytokine signaling pathway	ACSL4, AKT3	2	0.41	hsa04920
Glycan structures - biosynthesis 2	PIGA	1	0.4	hsa01031
Purine metabolism	AK3L1,AK3L2	1	0.39	hsa00230

Phosphatidylinositol signaling system	OCRL, PIK3R1	2	0.39	hsa04070
Adherens junction	IGF1R	1	0.37	hsa04520
Heparan sulfate biosynthesis	HS6ST2	1	0.36	hsa00534
Glycosylphosphatidylinositol(GPI)-anchor	PIGA	1	0.36	hsa00563
Amyotrophic lateral sclerosis (ALS)	BCL2	1	0.36	hsa05030
Tight junction	ASH1L, PARD6B, AKT3	3	0.34	hsa04530
Arachidonic acid metabolism	MAML1	1	0.33	hsa00590
Long-term depression	IGF1R, MAP2K1	2	0.33	hsa04730
Basal cell carcinoma	WNT3A	1	0.32	hsa05217
Polyunsaturated fatty acid biosynthesis	BAAT	1	0.28	hsa01040
Regulation of actin cytoskeleton	MAP2K1, PIK3R1, FGF7, FGF2	4	0.23	hsa04810
Cell adhesion molecules (CAMs)	CNTNAP1	1	0.23	hsa04514
Inositol phosphate metabolism	OCRL	1	0.23	hsa00562
ECM-receptor interaction	HSPG2	1	0.23	hsa04512
Fatty acid metabolism	ACSL4	1	0.2	hsa00071
Pyrimidine metabolism	CMPK	1	0.19	hsa00240
Butanoate metabolism	PDHA1	1	0.18	hsa00650
Glycan structures - biosynthesis 1	HS6ST2, MGAT4A	2	0.18	hsa01030
Oxidative phosphorylation	ENSG00000107902	1	0.16	hsa00190
Pyruvate metabolism	PDHA1	1	0.15	hsa00620
N-Glycan biosynthesis	MGAT4A	1	0.13	hsa00510
Fructose and mannose metabolism	ENSG00000107902	1	0.13	hsa00051
Dorso-ventral axis formation	MAP2K1	1	0.13	hsa04320
Notch signaling pathway	NUMB	1	0.12	hsa04330
Neuroactive ligand-receptor interaction	GPR63, PTGFR, GHR, HTR4	4	0.11	hsa04080

Cytokine-cytokine receptor interaction	VEGFA, GHR, ACVR2B, BMPR1A	4	0.11	hsa04060
Sphingolipid metabolism	FVT1	1	0.1	hsa00600
Type II diabetes mellitus	PIK3R1	1	0.1	hsa04930
Citrate cycle (TCA cycle)	SUCLA2	1	0.09	hsa00020
Aminosugars metabolism	ENSG00000107902	1	0.09	hsa00530
Leukocyte transendothelial migration	PIK3R1	1	0.09	hsa04670
Gap junction	MAP2K1	1	0.06	hsa04540
GnRH signaling pathway	MAP2K1	1	0.05	hsa04912
SNARE interactions in vesicular transport	BNIP1	1	0.05	hsa04130
Bile acid biosynthesis	BAAT	1	0.05	hsa00120
Calcium signaling pathway	PTGFR, HTR4, PPP3CB	3	0.03	hsa04020
Propanoate metabolism	SUCLA2	1	0.02	hsa00640
Melanogenesis	WNT3A, MAP2K1	2	0.01	hsa04916
Alanine and aspartate metabolism	PDHA1	1	0	hsa00252

Table 87 -DIANA miRPAth pathways of miR-503-5p