Integrative genomic analyses of bacterially-associated colorectal cancer

by

Katie S. Viljoen

Submitted in accordance with the requirements for the degree of Doctor of Philosophy

The University of Cape Town

Department of Clinical Laboratory Sciences

Division of Medical Biochemistry

November 2014

Abstract

Sporadic colorectal cancer (CRC) has been linked to various lifestyle factors, including the consumption of alcohol and red meat, smoking, and obesity. CRC is one of most extensively characterised cancers, both at a molecular and 'omic level; nevertheless, the precise mechanism driving CRC initiation remains unknown. To date, numerous studies have identified changes in the microbial profiles of CRCs compared to adjacent normal mucosa and compared to healthy controls; however, CRC-associated bacteria have not been concurrently quantified across a single cohort; nor have the relationships between CRC-associated bacteria, clinicopathological features of CRC and genomic subtypes of CRC been investigated.

The main aim of this thesis was therefore to gain insight into the potential contribution of CRC-associated bacteria in the aetiopathogenesis of CRC by leveraging both host genomic and clinicopathological data as well as to investigate patterns of tissue colonisation between different CRC-associated bacteria.

The objectives were 1) to quantify, using quantitative-PCR, CRC-associated bacteria in a cohort of 55 paired tumour and adjacent histologically normal samples collected during surgical resection as well as in an additional 18 formalin-fixed paraffin-embedded (FFPE) samples; 2) to determine their relationships to patient age, gender, ethnicity, stage of disease, site of disease and MSI status (Chapter 4); 3) to evaluate the relationship between each bacterium and host gene expression (Chapter 8) and methylation changes (Chapter 6); and 4) to determine genomic subtypes of CRC using unsupervised clustering of gene expression data in the context of patient clinicopathological features and bacterial quantitation data; and 5) to gain a deeper biological understanding of the results from the objectives 1–4 using pathway analyses of the genomic subtypes obtained (Chapter 7).

The main finding of this thesis is that a transcriptomic subtype of colorectal cancer, characterised by an increase in CpG island methylation, displays an increased frequency of colonisation by *Enterococcus faecalis* and by high levels of Fusobacterium. At the pathway-level, this subtype is enriched for pathways related to DNA and protein damage response, infection, inflammation and cellular proliferation; notably, these

findings were confirmed in a well-defined publically available CRC gene expression dataset of colorectal adenocarcinomas (N=155). These findings suggest that specific bacterial colonisation underlies a distinct genomic subtype of colorectal cancer that is characterised by inflammatory-related gene expression changes; these findings however require validation in a larger cohort. In addition, novel associations between colonisation by specific bacteria and host clinicopathological, transcriptomic and DNA methylation features were identified.

Declaration

1. I know that plagiarism is wrong. Plagiarism is to use another's work and pretend that it is one's own.

2. I have used the American Medical Association convention for citation and referencing. Each contribution to, and quotation in, this thesis from the work(s) of other people has been attributed, and has been cited and referenced.

3. This thesis is my own work

4. I have not allowed, and will not allow, anyone to copy my work with the intention of passing it off as his or her own work.

5. I acknowledge that copying some else's assignment or essay, or part of it, is wrong and declare that this is my own work.

Signature _____

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Prof Jonathan Blackburn, for his expert guidance, his positive attitude and valuable input that kept me inspired throughout my PhD, and for allowing me to develop my own research ideas.

I would like to acknowledge my family and friends for their support and belief in me, and for helping me maintain my sense of humour throughout my PhD. I would also like to thank Ryan Goosen (whose thesis was based on the same CRC cohort presented here) for the moral and technical support during the countless hours spent processing samples and deciphering hospital folders, and for the useful bioinformatics discussions; Jessica Duarte for proofreading my thesis in under a week; and Alvera Vorster for always being prepared to share her technical expertise.

Lastly, I would like to thank the Harry Crossley Foundation, the National Research Foundation (NRF) and the German Academic Exchange Service (DAAD), for their generous financial support over the past four years. My sincere thanks also goes to the Cancer Association of South Africa (CANSA) who funded this research project.

Table of Contents

List of Abbreviations 1

Chapter 1: The epidemiology and aetiopathogenesis of colorectal cancer 3

Abstract 3
Epidemiology of CRC 3
CRC risk factors 4
Southern African CRC landscape 4
Physiology of the colon 6
Intestinal barrier function 8
Specialised intestinal epithelial cells. 9
CRC aetiopathogenesis 13
CRC subtypes 13
Stem cells of the colon and CRC initiation 16
The role of inflammation in cancer 17
Epigenetic influences in cancer 21

Chapter 2: The infectious link to colorectal cancer 24

Abstract 24
The role of microbes in cancer 24
Established oncogenic pathogens 25
The CRC-associated microbiome 26
Factors influencing the composition of the colonic microbiome 26
CRC-associated bacteria 29
Escherichia coli and CRC 29
Fusobacterium spp. 32
Streptococcus gallolyticus 32
Enterococcus faecalis 33
Enterotoxigenic *Bacteroides fragilis* (ETBF) 34

Chapter 3: Study design, clinicopathological characterisation, and sample processing 35

Abstract 35
Study aims and objectives 36
Materials and methods 37
Sample collection and storage 37
A priori sample size calculation 38
Post-hoc power analysis for detection of bacterially-associated differential gene expression 39
Sample preparation 41
Microsatellite instability (MSI) testing 42
Results 44
Cohort characterisation 44
Microsatellite instability testing 46
Discussion 47

Chapter 4: Quantitative profiling of colorectal cancer-associated bacteria reveals associations between *Fusobacterium* spp., enterotoxigenic *Bacteroides fragilis* (ETBF) and clinicopathological features of CRC 48
Abstract *48*
Introduction *49*
Materials and Methods *51*
Cohort selection 51
Sample preparation and MSI testing 51
qPCR amplification conditions 54
qPCR quantification 55
qPCR quantification in FFPE samples 56
Statistical analyses 57
Results *58*
Bacterial quantification 58
ETBF and afaC-positive *E. coli* are significantly enriched in the colon compared to the rectum of CRC patients 63
Colonisation by ETBF or high-level colonisation by Fusobacterium are associated with late-stage CRC 64
Further clinical associations with high-level Fusobacterium colonisation 66
EPEC detection and characterisation 67
Discussion *71*

Chapter 5: Quality assessment and data handling methods for Affymetrix Gene 1.0 ST arrays with variable RNA integrity 74
Abstract *74*
Introduction *75*
Materials and Methods *77*
Sample preparation and quality control 77
Quantitative real-time PCR 78
Microarray analysis 79
Results *82*
Array quality 82
Transcript-dependent effects of RNA degradation on accuracy 84
Quality dependent methods of data adjustment and analysis 85
qRT-PCR validation of select genes 91
Discussion and conclusions *92*

Chapter 6: Whole-genome methylation analysis of CRC tumour and adjacent normal mucosal samples in relation to bacterial infection 96
Abstract *96*
Introduction *97*
Methylation patterns in CRC 98
Methylation analysis technology overview 101
Materials and methods *107*
Methylation analysis of *MLH1* using methylation specific PCR 107

Whole genome array-based methylation analysis 109
Results *112*
Detection of *MLH1* methylation by methylation-specific PCR 112
Characterisation of CRC methylation by multivariate analysis 112
Determining patterns of methylation alongside a large external cohort 115
Predicting CIMP status using an array-based marker panel 120
Genome-wide methylation analyses 124
Discussion *132*

Chapter 7: Specific bacterial infection, inflammation, and DNA and protein damage responses underlie a distinct genomic subtype of CRC 135
Abstract *135*
Introduction *135*
Established CRC subtypes 135
Data analysis workflow 137
Methods *140*
Establishing patient subtypes 140
Biological interpretation of CRC subtypes 141
Results *143*
Identifying tumour subtypes 143
Applying our data analysis pipeline to a large external dataset with previously defined subtypes 145
Biological features that distinguish CRC subtypes 148
Discussion *177*
Specific bacterial infection, inflammation and DNA and protein damage responses underlie a distinct genomic subtype of CRC 177
Study limitations and future recommendations 180
Conclusion *181*

Chapter 8: Gene expression analyses reveals *E. faecalis*- and *Fusobacterium*-associated genomic alterations in colorectal cancer 183
Abstract *183*
Introduction *183*
Enterococcus faecalis: an overview 184
Fusobacterium: an overview 186
Methods *186*
Sample preparation and microarray-based gene expression analysis 186
Differential gene expression analysis 187
Pathway analysis 187
Results *187*
Bacteria-associated gene expression analysis 187
Enterococcus faecalis-associated genomic alterations in CRCs 189
Genomic alterations associated with high-level colonisation by Fusobacterium in CRCs 200
Discussion and conclusions *201*

Chapter 9: Summary, conclusions and future perspectives 203
Conclusion *205*
Future perspectives *205*
References **207**
Appendix A **238**
Appendix B **252**
Appendix C **259**
Appendix D **279**

List of Abbreviations

AEEC: attaching/effacing *E. coli*

aEPEC: atypical EPEC

AIEC: adherent-invasive *E. coli*

ANOVA: analysis of variance

BMIQ: beta mixture quantile normalization

CAC: colitis-associated cancer

CD: Crohn's disease

CIMP-H: CIMP-high

CIMP-L: CIMP-low

CIMP: CpG island methylator phenotype

CIN: chromosomal instability

CRC: colorectal cancer

CV: coefficient of variation

EF: *Enterococcus faecalis*

EHEC: enterohaemorrhagic *E. coli*

EPEC: Enteropathogenic *E. coli*

ETBF: Enterotoxigenic *Bacteroides fragilis*

FAE: follicle-associated epithelium

FB: Fusobacterium

FC: fold change

FDR: false discovery rate

FFPE: formalin-fixed paraffin-embedded

GALT: gut-associated lymphoid tissue

GNUSE: global normalised, unscaled standard error

HNPCC: hereditary non-polyposis colorectal cancer

HPV: Human papilloma virus

IBD: irritable bowel disease

IPA: Ingenuity Pathway Analysis

IPLs: integrated pathway levels

LOD: limit of detection

LOH: loss of heterozygosity

MAMP: microbial-associated molecular pattern

MDS: multidimensional scaling

MMR: mismatch repair

MSI-H: MSI-high

MSI-L: MSI-low

MSI: microsatellite instability

MSP: methylation specific PCR

MSS: Microsatellite stable

N: CRC adjacent normal mucosa

NSAIDs: non-steroidal anti-inflammatory drugs

PCR: polymerase chain reaction

PRRs: pattern recognition receptors

qPCR: quantitative-PCR

qRT-PCR: quantitative reverse transcription PCR

RIN: RNA integrity number

ROS: reactive oxygen species

RPMM: recursively-partitioned mixture model

SCFAs: short chain fatty acids

SVA: Surrogate Variable Analysis

T: CRC tumour

TCGA: The Cancer Genome Atlas

UC: ulcerative colitis

Chapter 1: The epidemiology and aetiopathogenesis of colorectal cancer

Abstract

This chapter introduces various aspects of colorectal cancer (CRC) relevant to this thesis. The role of pathogenic infection and the microbiome in CRC will be discussed in Chapter 2. Chapter 1 starts with a brief overview of the epidemiology of CRC with specific reference to risk factors and the South African landscape of the disease. Then, colonic physiology will be described while emphasizing host defense mechanisms in the colon. Next, the aetiopathogenesis of CRC, with specific reference to proximal vs. distal CRCs, which are believed to have distinct aetiopathological features, is described. An overview of the role of inflammation in CRC and in irritable bowel disease (IBD) is provided, with specific emphasis on the effect of non-steroidal anti-inflammatory drugs (NSAIDs) on inflammatory-related diseases of the gastrointestinal tract. The last section is dedicated to epigenetic influences in cancer, which is closely linked to chronic inflammation.

Epidemiology of CRC

CRC is the third most commonly detected cancer in men and the second in women, with an estimate of more than 1.2 million new cases and 608,700 deaths occurring in 2008 worldwide[1]. Interestingly, CRC incidence varies at least 25 fold between countries, with a clear distinction between developed (United States of America, Canada, Japan and New Zealand) and developing (South East Asia and Africa) countries[2].

In South Africa, CRC ranks in the top five most common cancers, with a lifetime risk of 1 in 115 and 1 in 199 in men and women, respectively[3]. The 5-year survival rate for CRC patients is greater than 90% when tumours are detected at a localised, early stage. However, that rate drops to 40-65% once the cancer has spread regionally, and to 10% after distant metastases of the original tumour are detected[4]. Unfortunately, only 39% of CRCs are in fact diagnosed at an early stage—mainly due to low rates of screening[5].

Increased rates of colonoscopy-based CRC screening between 2000 and 2008 in the United States have been met with a decrease in annual CRC incidence of 2.3–4.2%, depending on tumour location, disease stage and gender[6]. These figures emphasize the need for a better understanding of the aetiology of different classes of CRC, which is crucial for early prevention and detection.

CRC risk factors

Although epidemiological studies have provided valuable clues regarding risk factors for CRC, the aetiology of sporadic CRC remains an active area of research.

CRC is significantly more prevalent in developed countries, a trend linked to lifestyle-related risk factors, such as red meat- and alcohol- consumption, obesity, physical inactivity, type-2 diabetes mellitus, and smoking[6–8]. Further risk factors include older age, male gender and a family history of CRC, polyps or IBD[6]. Conversely, CRC risk is mitigated by increased intake of fiber, NSAIDs, calcium, folate, vitamin D and hormone replacement therapy (HRT)[6,8]. Meanwhile, increasing evidence supports the role of specific pathogens[9–16], as well as dysbiosis (defined as an imbalance in the intestinal microbiome) in tumourigenesis[17–19].

The latest worldwide estimate of the proportion of infection-attributable cancers stands at 18%[20]. Currently, cancers with the largest proportion attributable to infectious agents are cervical cancers (100%), liver cancers (70%) and stomach cancers (60%)[21]. Not surprisingly, the rate of infection-attributable cancers is significantly higher in underdeveloped countries[21,22], underscoring the necessity of continued research, especially in sub-Saharan Africa where the proportion of infection-attributable cancers is estimated at 32.7%[22]. Importantly, around 30% of infection-attributable cases occur in people younger than 50 years[22].

Southern African CRC landscape

The incidence of CRC in Southern Africa is highly disparate between different racial groups—a phenomenon that appears to mirror the degree of Westernization (especially

regarding dietary composition)[23,24]; this is supported by increased rates of CRC with migration to westernized countries, as seen in African Americans[23]. Black South Africans have very low rates of CRC[3,25,26] that contrasts starkly with their white counterparts[3], which appears to be related to diet: Native Africans follow a diet of predominantly resistant starch (boiled maize meal) with a significantly lower intake of meat protein compared to Westernized populations; this dietary composition is met with high levels of microbially-derived butyrate, folate, and biotin[23,27]—molecules known for their protective effect against CRC. Interestingly, intestinal crypt cell proliferation is significantly lower in asymptomatic Native Africans compared to African Americans[28], an observation which is likely explained by the protective effect of butyrate[29]. The low levels of meat protein consumed by Native Africans likely provide further protection. Regarding species-level composition of the gut microbiome, Native Africans have significantly lower levels of certain Bacteroides spp. (*B. vulgatus* and *B. stercoris*) compared to high-risk CRC groups, and increased levels of *Lactobacillus acidophilus* 5[30].

Another intriguing observation among South African CRC patients is the disproportionately high number of *young* black CRC patients; 41–57% of black patients[25,31,32] compared to only 10% of white patients were under the age of 50[31]. CRC in young black patients do not appear to originate from colonic polyps[25] and cannot be attributed to IBD or diverticulosis[25]. Indeed, the majority of these cancers are located in the proximal colon[25,31,32] (often in the ceacum[25]), and often display mucinous histology[32,33] and a higher rate of microsatellite instability (MSI) compared to older patients[31,33]. In comparison to a developed country, in California, 10.6% of African Americans vs. 5.5% of white CRC patients presented before the age of 50 years[34], and black patients were more likely than white patients to develop proximal CRC[24,34]. Therefore, while the difference in the overall CRC incidence between African Americans and Native Africans appears to be due to differences in CRC-related lifestyle factors, the elevated risk of early-onset CRC appears to be at least partially influenced by ethnicity.

The relative increase in early-onset CRCs is also common in developing countries (where the overall CRC incidence is low) including Lebanon[35], Taiwan[36], Uganda[37], Nigeria[38], Ethiopia[39], Sri-Lanka[40], Saudi Arabia[41], India[42] and Egypt[43]. Of the aforementioned studies, rectal cancers were reported as the predominant site of disease in India, Nigeria, Uganda and Ethiopia, while an increase in proximal cancers were reported in Saudi-Arabia[41], Taiwan[36] and South Africa[33]. While most of these studies did not assess MSI status, an increase in the incidence of MSI was reported in early-onset CRCs in Taiwan[36], China[44] and South Africa[31].

Because the incidence of late-onset lifestyle-related CRC is lower in developed countries, one might expect an increase in the proportion of early-onset CRCs, which are more likely due to hereditary factors. However, in many developing countries, the rate of early-onset CRCs with a positive family history of CRC is very low (1–4%)[35,36,40,42,43]. A Taiwanese study reported no family history of CRC in 81% of early-onset MSI+ CRCs[36] , while a Chinese study reported germline mutations in mismatch repair (MMR) genes in 47% of early-onset (< 45 years) CRCs[44]. Another study composed of patients from the US and Scotland demonstrated that 58% of CRC patients ≤ 35 years exhibited MSI, compared to 12% of patients > 35 years; of the younger patients with MSI, 42% had germline mutations in an MMR gene[45]. In a Scottish cohort, 28% of CRC patients < 30 years of age had pathogenic mutations in the MMR genes *MLH1* or *MSH2*.

A substantial proportion of early-onset CRCs, especially in developing countries such as South Africa, are therefore of unknown aetiological origin, and warrant further investigation.

Physiology of the colon

Starting from the most apical layer, the colon consist of the mucosa, which includes the epithelium, the lamina propria and the muscularis mucosae; the submucosa; the muscularis propria; and the serosa[46]. Longitudinally, the regions of the colon that are proximal and distal to the splenic flexure have embryological origins in the mid- and hindgut, respectively, with distinct morphological and functional differences between these regions. Here, we consider the proximal colon to consist of the caecum, ascending

colon, hepatic flexure and transverse colon; and the distal colon consists of the descending colon, sigmoid colon, recto-sigmoid junction, splenic flexure and the rectum[6].

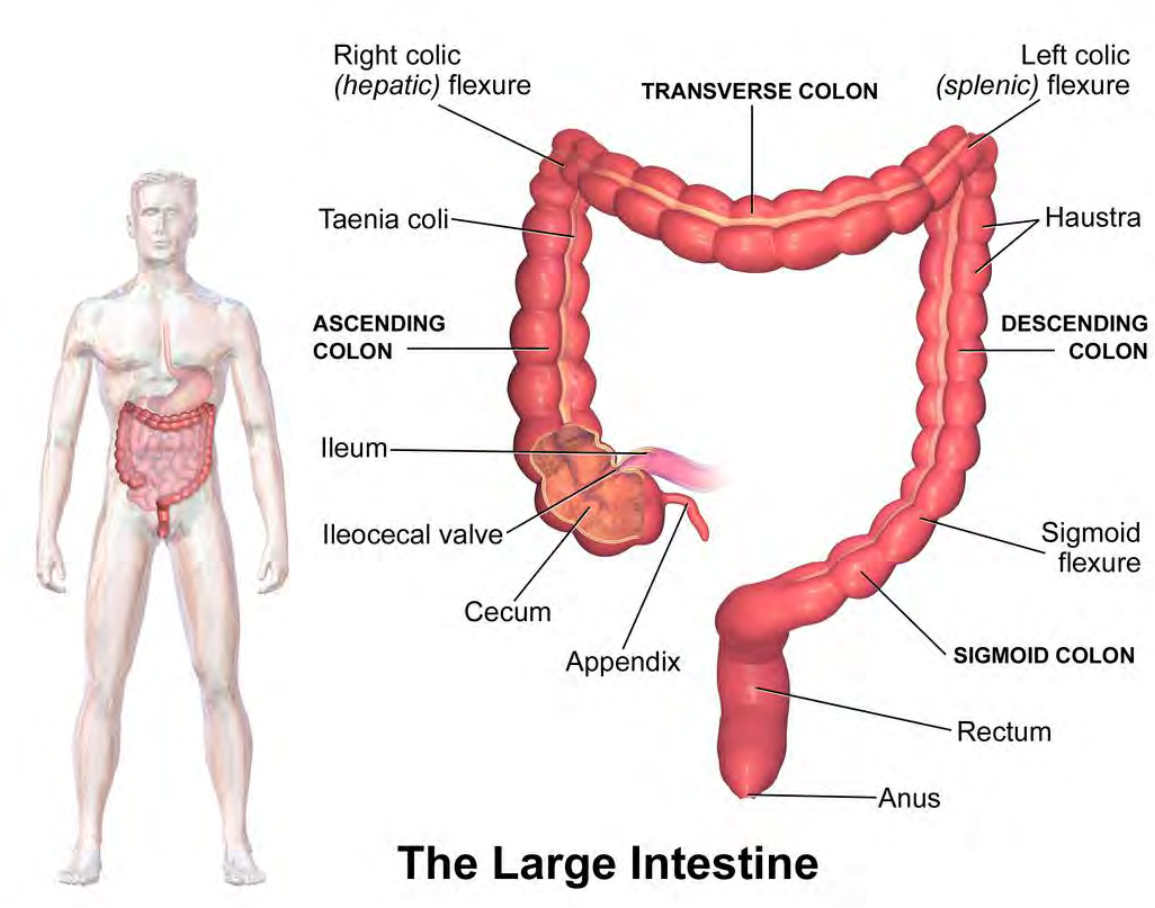


Figure 1: Illustration of different regions of the colon[47]. The proximal colon consists of the caecum, ascending colon, hepatic flexure and transverse colon. The distal colon consists of the descending colon, sigmoid colon, recto-sigmoid junction, splenic flexure and the rectum.

Although the embryological boundary, two-thirds along the transverse colon, guides proximal-distal demarcation, many features instead change gradually along the colon. LaPointe et al. mapped out gene expression in the healthy colon and found that while some genes showed significantly differentially expressed genes between the proximal and distal colon, a strict dichotomous model may be too rigid, since as a subset of genes are better described by a continuous model, where expression changes gradually along the length of the colon[48–50].

The main functions of the colon are excretion of waste, reabsorption of water and electrolytes from the luminal material, and processing resistant carbohydrates via fermentation. The latter occurs mainly in the ascending and proximal transverse colon through hundreds of saccharolytic and proteolytic bacterial species—mostly obligate anaerobes, which produce short chain fatty acids (SCFAs). Together, these SCFAs, which include butyrate (15%), propionate (25%) and acetate (60%), constitute 5–15% of the total caloric needs of an individual[51]. Once carbohydrates (the preferred nutrient source for most bacteria) are depleted, undigested proteins, originating from both the diet, colonic mucous and shed epithelial cells, are fermented in the distal colon. Proteins are converted to SCFAs, branched chain fatty acids, amines, ammonia, phenols, indoles and sulfurs. Products that are not absorbed by the host are either used by bacteria as a nitrogen source or excreted[51]. In addition to utilisable nutrients, bacteria provide the host with essential vitamins, including a variety of B vitamins and vitamin K[52].

Intestinal barrier function

The colon is home to a rich collection of flora where interaction with the host may be symbiotic, commensal or pathogenic. Accordingly, the colon is well equipped to defend itself against harmful bacteria—the first line of defense, a 150µm thick mucous layer[53] imposes a physical and chemical barrier, which consists mainly of mucin 2 (MUC2) and antimicrobial effector molecules such as antimicrobial peptides, secretory IgAs, glycoproteins and trefoil factors, which are secreted by specialised epithelial cells[54].

Structurally, the mucous layer consists of an oligomerised mesh of mucins, with MUC2 (the only gel-forming mucin expressed in the healthy colon[55]) as the major constituent. The mucous layer consists of a thin sterile inner layer, which is firmly adherent to the epithelial cells, and a partially colonised, non-adherent outer layer[56]. The importance of the mucous layer is demonstrated in mice deficient in MUC2, where bacteria are found in direct contact with the epithelium and even in colonic crypts, which harbour intestinal stem cells; these mice develop colitis[57,58] which frequently progresses to adenomas and finally adenocarcinomas of the colon and rectum[59].

Apart from the mucous layer, the intestinal epithelium poses a further physical barrier, with the negatively charged glycocalyx deterring bacteria, and tight junctions regulating intercellular flux[60]. The glycocalyx is composed of proteoglycans, glycoproteins and glycolipids, and plays an important role in cell-cell recognition, communication and intracellular adhesion.

The integrity of the epithelial barrier may be disrupted by a range of factors (many of which are known CRC risk factors), including chronic alcohol consumption, psychological stress, chronic use of NSAIDs and specific bacteria (most notably pathogenic *E. coli* strains[61]).

Specialised intestinal epithelial cells.

The harsh environmental factors that rule the intestinal lumen are met with specialised epithelial cells; these include microfold (M) cells, goblet cells, Paneth cells and endocrine cells, each with specialised functions to maintain homeostasis[54]. These specialised cells differentiate from progenitor cells called transit-amplifying cells, which in turn arise from pluripotent stem cells located in the base of intestinal crypts (Figure 2). With the exception of Paneth cells, differentiated cells rapidly migrate to the luminal border after which they are sloughed off within a few days[62,63]. Paneth cells, on the other hand, migrate to the base of crypts and have a turnover rate of 6–8 weeks[62].

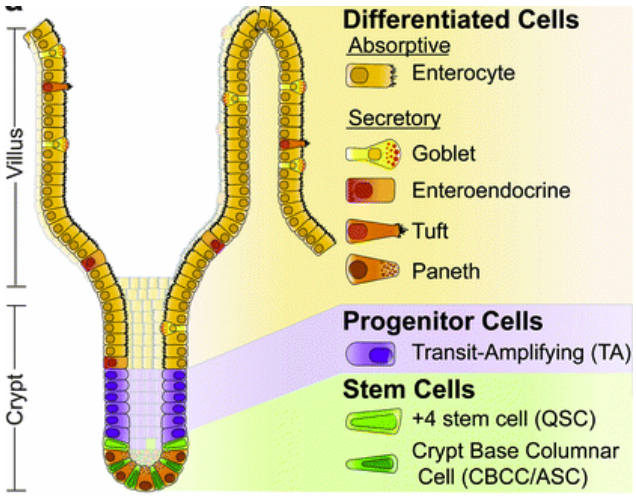


Figure 2: Organisation of specialised epithelial cells along the crypt villus axis. The intestinal epithelium is organised into crypt and villus regions, with the stem and progenitor zone localised in the crypt. Transit-amplifying (TA) progenitors arise from the stem cell compartment and differentiate into absorptive enterocytes or secretory goblet, enteroendocrine, tuft, or Paneth cells. Most of the differentiated cell populations migrate up the villi, but, uniquely, the Paneth cells move downward and reside between the stem cells [64].

Up to 70% of the body's immunocytes are found in the gut-associated lymphoid tissue (GALT), making it one of the body's largest lymphoid organs[65]. The GALT consists of isolated or aggregated lymphoid follicles, with the latter found in the caecum, colon and rectum patches. Aggregated lymphoid follicles are covered by the follicle-associated epithelium (FAE), which in turn consists of enterocytes and M cells. As demonstrated in germ-free mice, the microbiome is essential for the development of lymph node architecture, the GALT, and for antibody production[66]. This shaping of the mucosal immunity serves to establish physiological inflammation: a state of homeostasis where immune responses to commensals are down-regulated to avoid excessive and potentially self-damaging inflammatory responses, while maintaining low level inflammation to allow a rapid response to unwanted antigens[54].

Both the innate and the adaptive immune system possess feedback loops to regulate bacterial contact with the mucosal surface. The innate immune system monitors mucosa-associated bacterial density by microbial-associated molecular pattern (MAMP) concentration, to tailor the activation of epithelial antimicrobial responses[67], which include secretion of antimicrobial peptides by Paneth cells, and secretion of mucins by Goblet cells.

Meanwhile, the adaptive immune system uses dendritic cells to sample live bacteria at the mucosal surface, which are trafficked to the GALT, where B cells are induced to produce bacteria-specific IgAs[67], which serve to control the density of specific bacteria at the mucosal surface. If homeostasis is disrupted due to host and/or environmental factors, dysbiosis (a state associated with the overgrowth of a subset of harmful bacteria/single species), loss of epithelial barrier integrity and pathological inflammation may ensue (Figure 3).

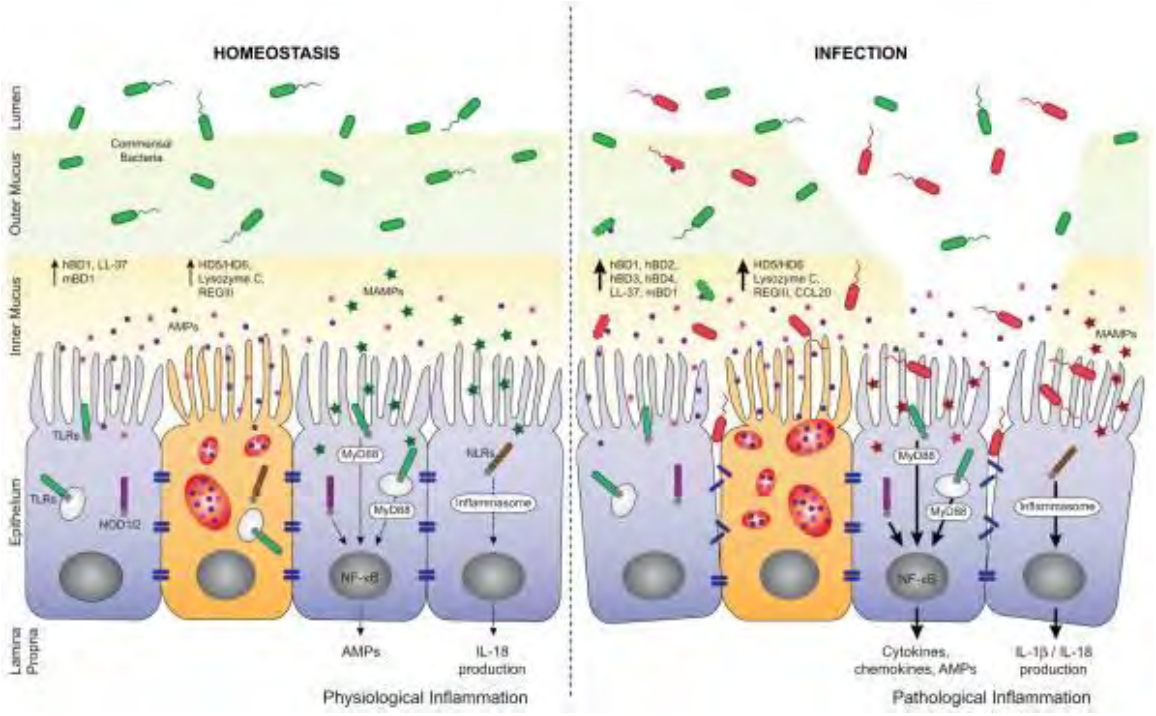


Figure 3: Colonic epithelial function during physiological vs. pathological inflammation. During physiological inflammation, microbial-associated molecular patterns, (MAMPs, green stars) from commensal bacteria are sensed by pattern recognition receptors (PRRs), which triggers basal levels of antimicrobial peptides (AMP) and interleukin-18 (IL-18) production, which contributes toward tolerance intestinal homeostasis. Pathological inflammation occurs when pathogens breach the mucus layer, contact intestinal epithelial cells and disrupt epithelial barrier integrity. Pathogens and their products (MAMPs, red stars) are also sensed by PPRs and induce expression of a pro-inflammatory cytokines/chemokines and AMPs that help limit pathogen propagation and contribute to the recruitment of immune cells. hBD, human β-defensin; mBD, mouse β-defensin; HD, human α-defensin. Figure reproduced with permission from Muniz et al[68].

Goblet cells

Goblet cells are primarily responsible for the production of mucins and increase in concentration from the duodenum (4%) to the distal colon (16%), mirroring the thickness of the mucous layer and microbial density gradient in the intestine[69]. Notably,

MUC12, *MUC17, MUC5B* and *MUC11* were found to be expressed 3.0, 1.5, 2.3 and 1.7 fold lower, respectively in the proximal colon when compared to the distal colon[48].

Mucins can be broadly classified as either membrane-bound or secretory. Of the secretory mucins, MUC2, MUC5AC, MUC5B, and MUC6 are all gel-forming mucins, with similar structures localized on chromosome 11.5.5 as a cluster[69]. Membrane bound mucins include MUC1, MUC3A, MUC3B, MUC4, MUC12 and MUC17[70]. In the healthy colon, MUC2 is the dominant mucin, with MUC3 and MUC4 expressed at much lower levels[71]. In a diseased state, including IBD, shigellosis and CRC, several mucins including MUC1, MUC5AC, MUC6 may be aberrantly expressed in the colon[71,72].

Mucin abundance is regulated by both host- and environmental factors including hormones, microbes, cytokines and reactive oxygen and nitrogen species. This regulation may target goblet cell differentiation or mucin-synthesis and -secretion[69]. In particular, the recognition of pathogen-associated molecular patterns (PAMPs) via host cell pattern recognition receptors (PRRs) play a prominent role in mucin regulation via inflammatory pathways, as is evident in germ-free mice, which have significantly fewer mucin-secreting Goblet cells[69,73]. In the case of MUC2, various extracellular stimuli are transcriptionally consolidated through pro-inflammatory Nf-**κ**B, which has binding sites in the *MUC2* promoter[69].

M cells

Unlike normal enterocytes, M cells lack microvilli, do not secrete mucous or digestive enzymes, and have a much thinner glycocalyx, allowing them to readily sample antigens from the lumen via endocytosis or phagocytosis, which are then delivered to dendritic cells and lymphocytes located in the pocket-like structure on the basolateral side of M cells. M cells therefore act as sentinels at the interface between the lumen and the lymphoid system[65]. However, having ready access to the lumen comes at the cost of exploitation by certain pathogens, which manage to escape immune detection and use M cells as an entry point for infection.

Dendritic cells themselves can also sample whole bacteria that penetrate the epithelium. These antigens are presented to B and T cells, inducing antigen-specific IgA production.

IgAs are transported and transcytosed across the epithelium where they may bind to complimentary bacteria to prevent bacterial translocation across the epithelium[53].

Paneth cells

Paneth cells are specialized secretory cells located at the base of intestinal crypts. Although Paneth cells are normally only present in the small intestine, they may arise at other sites including the colon and oesophagus in some diseased states, and their appearance is usually associated with chronic inflammation[74,75]. Paneth cells secrete a variety of antimicrobial factors including α-defensins and lysozyme C (which are active against gram-positive and -negative bacteria), as well as secretory group IIA phospholipase A2 and REG3A (a C-type lectin), which are effective against gram-positive bacteria only[75].

Defensins are 30–40 amino acids in length and can be separated into α- and ß-defensins. Whereas α-defensins (DEFA5 & DEFA6) are constitutively expressed by Paneth cells and neutrophils, ß-defensins are inducibly expressed by a variety of epithelial cells including colonocytes (with the exception of ß-defensin-1, which is constitutively expressed)[75]. ß-defensins 2, 3 and 4 are induced by pro-inflammatory and bacterial stimuli in the colon and are upregulated in colonocytes of patients with UC, which correlates with an increase in TNF-α and IL-8[76]. Similarly, REG3A is inducibly expressed by TLR signaling in Paneth cells and enterocytes.

Regarding site-specific expression, *MEP1B, DEFA5* and *REG1A* were upregulated 1.8, 2.8 and 2.7 fold respectively, in the proximal compared to the distal colon, according to LaPoint et al[48].

CRC aetiopathogenesis

CRC subtypes

As previously mentioned, the proximal and distal regions of the colon have different embryological origins, and significant morphological and functional differences exist between the two. In the early 1980s, epidemiological data highlighted gender- and age-disparities in colon cancers of the proximal vs. distal colon[77]. Subsequent molecular and cytogenetic data has supported an aetiopathogenic dichotomy in CRC[78–81]. Moreover,

some environmental risk factors seem to be site-specific, such as smoking in the case of proximal CRC and red meat consumption in the case of distal CRC[50,82,83]. Although this dichotomous division is currently the widely accepted working model, in reality many molecular features vary gradually along the length of the colon with no clear divide. In this regard, Yamauchi et al. propose a paradigm shift towards a continuum model where there is a better appreciation of subsite-specific profiles, an idea supported by Benedix et al., who demonstrated additional prognostic value by classification according to colonic subsite[84]. However, there is currently very little focus on colonic subsite in current research and as such we will adhere to the dichotomous proximal-distal model of CRC for the remainder of this discussion.

The main molecular feature that segregates the majority of proximal from distal cancers is the type of genomic instability encountered, with chromosomal instability (CIN) and MSI being more prevalent in distal and proximal cancers, respectively[85]. CIN refers to changes in chromosomal copy number and/or structure, which often result in the physical loss of genes, referred to as loss of heterozygosity (LOH). Regarding CIN by CRC subtype, chromosomal imbalances have been reported in 30% of MSI CRCs, compared to 70% of microsatellite stable (MSS) CRCs[86]. Various mechanisms of CIN, including inactivation of genes involved in the mitotic spindle checkpoint, DNA damage checkpoints, chromosome metabolism, and centrosome function[87], as well as global DNA hypomethylation[88] have been suggested. DNA hypomethylation is associated with large-scale CIN[88,89], but not point mutations and short insertions and deletions[88]. In this regard, it is interesting to note that Bond et al. recently defined two types of CIN with distinct patterns of instability in MSS cancers: BRAF mutant and wild-type cases displayed 'focal' or 'whole arm' patterns of CIN, respectively[90].

On the other hand, MSI results from a loss of MMR function, which leads to strand slippage within repetitive DNA sequence elements and consequent insertions or deletions of mono- or dinucleotide repeats (microsatellites). The resulting mutations most notably affect tumour suppressors (containing microsatellites), such as TGF-beta receptor II (*TGFBR2*) and BCL2-associated X protein (*BAX*)[91]. MSI-positive cancers

can be categorized as MSI-high (MSI-H) or MSI-low (MSI-L) according to the degree of MSI seen.

Sporadic MSI-H cancers account for 10–15%[92–95] of all CRCs, and an additional 3–10% display MSI-L[92,95,96]. Regarding the prevalence of MSI-L, one study showed by examining 377 microsatellite regions that 79% of sporadic CRC showed MSI in 1–11 of these regions, and they concluded that the Bethesda panel of microsatellite markers (the gold standard for detecting MSI-H) is not very sensitive in detecting MSI-L cases[97].

Around 3% of all CRCs display MSI-H due to a loss-of-function germline mutation[98–100] in the MMR genes *MLH1*, *MSH2* or *MSH6*, with respective frequencies of 50%, 39% and 7%[101]. CRCs originating from germline mutations in MMR genes are referred to as hereditary non-polyposis CRCs (HNPCCs)—these patients have an 80% lifetime risk of developing CRC[91] and patients may develop cancer as early as their teenage years. Although colorectal cancer is by far the most common type of cancer associated with these mutations, ovarian, gastric, small intestine, brain, urinary and biliary tract tumours may also occur[101]; it is not known why HNPCC patients are specifically predisposed to CRC. Loss of MMR function requires inactivation of both alleles, and in the case of HNPCC, this occurs via mutation or loss-of-heterozygosity (LOH), and less frequently via epigenetic silencing[91]. Regarding the mechanistic basis of MSI in sporadic CRC, Poynter et al. conducted comprehensive mutational screening of MMR genes, *MLH1* methylation testing, and immunohistochemistry of MMR proteins for 1061 population-based CRCs. They found that of the MSI-H cases with no mutational basis (i.e. sporadic), 60% could be explained by *MLH1* methylation (although some literature reports are higher), a further 28% showed loss of an MMR protein (MLH1, MSH2 or MSH6) by IHC but did not have *MLH1* methylation, while 12% were unexplained by *MLH1* methylation or loss of MLH1, MSH2 or MSH6. In MSI-L cases, the MMR proteins MLH1, MSH2 and MSH6 were present by IHC in 100% of cases, and only 3% showed promoter *MLH1* methylation, leaving 97% of MSI-L unexplained[102].

Stem cells of the colon and CRC initiation

Despite the seemingly distinct aetiopathogenic foundations of proximal vs. distal CRCs, aberrant activation of the Wnt signaling pathway plays a central role in an estimated 90% of CRCs[103,104]. Wnt signaling is an evolutionary conserved pathway involved in embryonic development and tissue homeostasis[105]. In the healthy colon, Wnt signaling is restricted to colonic crypts with Wnt ligands permanently present in a gradient along the crypt–villus axis, facilitating stem cell homeostasis, cell renewal and repair[105].

In about 60% of CRCs, Wnt signaling is activated via loss-of-function mutations in the adenomatous polyposis coli (*APC*) gene[106,107], as first discovered in patients with germline mutations in *APC*. These patients develop hundreds of colorectal adenomas following inactivation of the remaining wild type allele in a syndrome referred to as Familial Adenomatous Polyposis (FAP)[108]. Almost invariably, some of these adenomas will progress to carcinomas.

In addition to mutation in APC, Wnt signaling may also be activated by activating mutations in β-catenin, allowing translocation of β-catenin to the nucleus. Mutations in *APC* or β-catenin seem to be mutually exclusive and common in distal and proximal cancers, respectively. Further mechanisms of Wnt signaling activation include inflammatory signals through NF-κB-, PI3K-, and Akt-related pathways[109], or by promoter hypermethylation of Wnt pathway antagonists (especially in sessile serrated adenomas and proximal CRCs)[110].

Activation of the Wnt pathway allows β-catenin to translocate to the nucleus where it induces expression of an array of genes commonly implicated in CRC pathogenesis, including c-MYC, cyclin D1 and Axin 2. Using $APC^{-/-}$ mice, Oshima et al. demonstrated that the earliest consequence of loss of APC is an expansion of the proliferative zone in crypts—cells remain in a proliferative state instead of differentiating and migrating out of the crypt. This results in abnormal tissues architecture, which presents as a polyp[111,112]. Valuable insight regarding the nature of cancer initiating cells in CRC was provided by Barker et al., who demonstrated that stem-cell-specific loss of APC was required for adenoma formation in a Lgr5-cre mouse

line: these cells drove adenoma formation while remaining at the base of the crypt[105,113]. These findings demonstrate the pivotal role of Wnt signaling in CRC initiation, in a stem-cell-dependent manner.

The role of inflammation in cancer

Acute inflammation is the host's response to tissue damage as a result of environmental insults such as infections, toxins, allergic reactions and autoimmune diseases. This response is tightly regulated and usually short-lived. However, persistent activation of immune cells leads to chronic inflammation resulting in DNA damage and tissue destruction[114].

For centuries, physicians have been aware of a link between inflammation and cancer, based on shared characteristics between the two processes. In 1863, Virchow et al. observed inflammatory cells in tumour biopsies and noted that tumours often developed at sites of chronic inflammation[115,116]. Meanwhile in 1986, Dvorak et al. described tumours as "wounds that do not heal", highlighting the overlap between cancer and inflammation, including immune cell infiltration and angiogenesis[116,117].

The role of inflammation in both tumour initiation and development is now widely accepted. Cancer-related inflammation may be *intrinsic* or *extrinsic* to malignant transformation, although the two pathways are not mutually exclusive, as *extrinsic* inflammation may indirectly cause *intrinsic* inflammation following malignant transformation[118]. In the *intrinsic* pathway, an inflammatory response is activated as a consequence of malignant transformation. In the *extrinsic* pathway, chronic inflammation, which may be caused by various environmental factors, including smoking, exposure to pathogens and a Western style diet, facilitates malignant transformation[118].

During chronic inflammation, the balance between pro- and anti-inflammatory cytokines is disrupted, creating a pro-carcinogenic environment. Certain pro-inflammatory cytokines have repeatedly been implicated in tumour initiation and progression of various cancers—these include TNF-α, IL-6 and IL-1β[118–120]. Lack of

anti-inflammatory cytokines is equally detrimental, as demonstrated in IL-10$^{-/-}$ mice, which develop colitis and colitis associated cancer (CAC)[114].

Pro-inflammatory cytokines activate transcription factors such as NF-κB and STAT3, which in turn trigger genes involved in diverse cellular processes relevant to both tumour initiation and progression. In addition to activation through pro-inflammatory cytokines (including TNF-α and IL-1β), NF-κB is activated through TLRs that recognize microbial signatures[114]. NF-κB signaling has been causally implicated in pathogen-induced cancers such as HBV-related liver cancer and HPV-related cervical cancer[121]; blocking NF-κB signaling has been shown to suppress experimental colitis in mice[121–123].

In conclusion, tumour-promoting inflammation is considered one of the enabling characteristics of cancer that facilitates acquisition of the hallmarks of cancer described by Hanahan et al[124]. Inflammation can contribute to multiple cancer hallmarks including the induction of angiogenesis; invasion and metastasis; sustained proliferative signaling, resisting cell death; and evading growth suppressors[124]. Inflammation is therefore a key feature in most cancers. One key question that is particularly relevant to this thesis is the degree of inter-individual variation in extrinsic, cancer-promoting and to what extent this extrinsic inflammation is caused by host genetic factors or external stimuli such as pathogens, smoking and a Western style diet.

Inflammation, IBD and CRC

The two major forms of IBD, UC and Crohn's disease (CD), are strongly linked to chronic inflammation. While CD may affect any part of the GIT, it most commonly affects the lowest part of the small intestine, the ileum. UC on the other hand is limited to the colon and rectum. Increased severity and duration of IBD are accompanied by an increased risk of developing CRC with an average risk of 2–3 fold for all IBD patients[114]

Elevated levels of TNF-α and NF-κB signaling commonly occur in IBD and antibodies against TNF-α have proven effective in the treatment of IBD[114,121]. The aetiology of IBD involves the complex interplay between the host and its microbiome: around 100 genetic loci have been associated with IBD risk and specific bacteria and dietary factors

are also associated with the disease[114,125]. It is generally thought that IBD is the result of an inappropriate immune response to commensal bacteria; however, several studies also support the role of specific adherent pathogens, and in particular, adherent strains of *E. coli* in the pathogenesis of IBD[126,127]. The recent surge in metagenomic studies will undoubtedly increase our knowledge regarding variability in the human microbiome[128,129], which may help identify specific microbial profiles that correlate to disease susceptibility.

NSAIDs and inflammatory-related diseases of the gastrointestinal tract

The ability of aspirin and other NSAIDs to decrease the risk of developing CRC, as well as other cancers, has recently received much attention, lending further support to the carcinogenic role of inflammation[130,131].

Several studies have demonstrated long-term protection against CRC (including HNPCC) with regular use of aspirin[130,132–134]. The magnitude of the effect appeared increase with treatment duration (2.5 years vs. 5 years), rather than the concentration of aspirin, with no additional reduction in the 20-year CRC incidence and mortality above 75mg/day[130]. Interestingly, the chemopreventive benefit of aspirin also varied by colonic site—with at least five years of aspirin use, the risk of developing proximal and rectal colon cancers was reduced by 70% and 50%, respectively, while no benefit was seen in distal cancers[130].

A host of putative mechanisms underlying the chemopreventive effects of aspirin have been proposed[135]. These mechanisms may be divided into COX-dependent and COX-independent methods. Aspirin inhibits both COX-1 and COX-2 through acetylation, thereby modulating arachidonic acid metabolism and exerting its analgesic and anti-inflammatory effects via COX-2. Whereas *COX-1* is constitutively expressed in most tissues, *COX-2* is inducibly expressed in response to growth factors, cytokines and tumour promoters, and is 50–100 fold less sensitive to aspirin compared to COX-1[135]. COX-2 metabolises arachadonic acid to PGH_2, but inhibition by aspirin leads to incomplete metabolism with a concomitant shift in metabolites, now favoring, among other metabolites, the production of lipoxins, which are known to inhibit cancer cell proliferation and angiogenesis[135]. Further, COX-2's infamous role in tumour promotion

and progression is largely based on the production of the pro-inflammatory prostanoid PGE_2, that stimulates proliferation, inhibits apoptosis, and promotes motility, invasion and angiogenesis[135].

Interestingly, it has been suggested that aspirin promotes apoptosis of MSI-H cells, where critically unstable cells are removed from the population, resulting in a largely MSS population, even in the presence of a dysfunctional MMR system, as evidenced from MMR deficient CRC cell lines and a mouse model of HNPCC[133,136]. Using a mouse model of HNPCC, Mcilhatton et al. recently demonstrated a reduction in MSI upon long-term administration of aspirin, with a concomitant 18–21% increase in life span. Although the difference did not reach statistical significance, untreated HNPCC mice showed MSI-L and MSI-H in 39% and 33% of cases, whereas aspirin-treated mice showed MSI-L and MSI-H in 56% and 22% of cases. This effect was however obtained in homozygous null *MSH2* mice, whereas HNPCC patients are heterozygous carriers of MMR mutations, suggesting that the effect on preventing MSI might be increased under long-term, low dose aspirin use in HNPCC carriers, compared to the 400mg/kg dose that was required in the HNPCC mouse model[133]. The finding that aspirin is more effective in preventing proximal cancers lends further credence to an MSI-based mechanism of protection.

The mechanism whereby aspirin might prevent MSI is still unclear, but it seems plausible that it is related to a reduction in inflammation and oxidative stress, since these conditions are known to promote MSI[137–139]. In relation to HNPCC carriers, Kloor et al. recently found that MSI+, MMR-deficient crypt foci (where bi-allelic MMR gene inactivation has occurred) are prevalent in the colons and small bowels of HNPCC patients: about one per 1 cm^2 or 2 cm^2, respectively[140]; this is in stark contrast to the low number of clinically manifest HNPCC cancers, a discrepancy that suggests that most MMR deficient crypt foci do not progress to cancer[136]. Indeed, despite the prevalence of MMR-deficient crypt foci in the small bowel, only ±4% of patients with HNPCC developing small bowel cancers[141]. Since the vast majority of MMR mutational carriers develop CRC specifically, it seems plausible that the colonic environment somehow potentiates tumourigenesis, either by putting extra pressure on an already MMR

deficient system thereby inducing MSI, or by an MSI-independent mechanism. In such a pro-inflammatory and -tumourigenic environment regular use of aspirin may tip the scales towards an anti-inflammatory environment, thus abrogating chronic inflammation that may be caused by chronic pathogenic infection or other environmental stimuli. However, this still does not explain why a relatively short course (~5 years) of treatment with aspirin confers long-term protection (10–20 years follow up) against CRC in HNPCC patients.

Given the preventative role of regular aspirin use in CRC, together with the association between inflammation, IBD and CRC, one would reasonably expect NSAIDs to have a similarly preventative effect on IBD-related diseases. However, the data regarding the association between NSAID use and IBD are conflicting[7,142,143], and appear to depend in part on the type of NSAID used[143]. These results may also be confounded by the accuracy of IBD diagnosis, which may be confused with drug-induced colitis, where NSAIDs cause IBD-like ulceration[144,145]. Irrespective of whether NSAIDs increase the risk of IBD, regular use appears to be at least moderately protective against CRC in IBD patients[142,146].

Epigenetic influences in cancer

According to the classical model, cancer arises from an initiating mutation, which is followed by a series of mutations in tumour suppressor genes and oncogenes, resulting in clonal selection and tumour heterogeneity. With increasing knowledge of the role of epigenetics in health and disease, it is becoming apparent that not only mutations but also epigenetic modifications play a pivotal role in various cancers.

The epigenetic landscape is sculpted by post-translational modifications to DNA and histones, which are essential to normal processes such as cellular differentiation, X-chromosome inactivation and genomic imprinting[147].

Of the epigenetic marks, changes in DNA methylation has received most attention, with a growing body of evidence supporting a causal role for various environmental factors, including alcohol consumption, caloric intake, cigarette smoking and pesticides in aberrant methylation, which in turn has been associated with various diseases including

diabetes, asthma, cancer, Alzheimer's and atherosclerosis[148,149]. Furthermore, both bacterial and viral pathogens have been linked to aberrant methylation of host DNA: *Helicobacter pylori* induces aberrant DNA methylation in gastric cells; this phenomenon is most likely triggered by an inflammatory response to *H. pylori* rather than directly by the bacteria since the immunosuppressive drug cyclosporin A blocks aberrant DNA methylation in *H. pylori*-infected mice without affecting *H. pylori* colonisation[150]. Furthermore, Epstein-Bar virus+ gastric cancers have distinct patterns of methylation[151].

Aberrant methylation is a common occurrence in CRC, with global hypomethylation, and region-specific hypermethylation associated with CIN+ and MSI+ CRCs, respectively[152–154]. About 70% of proximal cancers with MSI exhibit aberrant hypermethylation of CpG islands (referred to as the CpG island methylator phenotype (CIMP))[155] and hypermethylation of the promoter region of *MLH1* is the main cause of sporadic MSI CRCs[102].

Global hypomethylation occurs particularly in non-CpG-island regions of the genome, including repetitive sequences, and in CpG island shores[156]. While hypomethylation has been associated with activation of tumour suppressors in a few genes (e.g. *CDH3*), the most apparent consequence of global hypomethylation is CIN[88,89]. This occurs when pericentromeric regions are demethylated, which facilitates recombination and altered chromosomal replication[156].

An epigenetic basis for cancer was suggested by Feinberg et al. who suggested that cancer has a fundamentally common basis that is grounded in a polyclonal epigenetic disruption of stem/progenitor cells; these changes are mediated by 'tumour-progenitor genes' which increase the capacity for self-renewal and pluripotency[157]. Hypothetical tumour progenitor genes would have a direct effect on a) DNA with both genetic and epigenetic consequences (e.g. *APOBEC*), b) stem cell genes (e.g. *OCT4*, *FOXD3*, *NANOG*) or c) genes that affect chromatin e.g. (*EZH2*)[157].

These epigenetic changes set the stage for initiating mutations as well as genetic and epigenetic plasticity[157]. This model is supported by *in vitro* and *in vivo* studies where

tumour cells demonstrate reversibility of phenotype, indicating epigenetic control. Furthermore, global hypomethylation, and in many cases promoter hypermethylation, precedes initial mutations in cancer, with epigenetic alterations found even in benign neoplasms[157].

In conclusion, epigenetic modifications undoubtedly play a crucial role in carcinogenesis, particularly during the initiation stage. However, whole-genome epigenetic research is still in its infancy and at this stage it is unclear to what extent aberrant epigenetic changes in progenitor/stem cells drive cancer initiation and to what extent epigenetic mechanisms contribute to each of the hallmarks of cancer[124].

Chapter 2: The infectious link to colorectal cancer

Abstract

Oncogenic pathogens have been discovered in numerous cancers, including cervical, lung, liver and gastric cancer. At least 18% of all cancers were attributed to infection in a global report on infection-associated cancers in 2002, a number that is expected to grow.

From other infection-attributable cancers, we know that only a small proportion of individuals initially infected with the relevant pathogen will develop cancer, and that this progression is dependent on chronic long-term infection. Further, genetic, environmental and strain-specific factors modify the susceptibility to the oncogenic potential of these microbes.

In the case of CRC, pathogenic shifts in the microbiome (dysbiosis) and/or infection with specific pathogens may play a role in the aetiopathogenesis of CRC. Although there is ample epidemiological and/or *in vitro* evidence for an association between specific bacteria and CRC, causality has not been established thus far.

This chapter outlines established oncogenic pathogens, and their shared characteristics, followed by a discussion of factors that influence the composition of the colonic microbiome and possible links to CRC aetiopathogenesis. Finally, CRC-associated bacteria and their putative oncogenic mechanisms are described.

The role of microbes in cancer

Although several cancer-microbial associations have been documented, only a handful of pathogens have been unequivocally established as oncogenic thus far[158]. Even so, 18% of all cancers were attributed to infection in a global report on infection-associated cancers in 2002[20]. Given the steady increase of cancer-microbial associations being uncovered, this is likely an underestimate of the true fraction of infection-dependent cancers. However, establishing causality is no easy task: certain bacteria may induce tumour growth, but subsequently decline once the tumour forms; others may opportunistically infect existing tumours without contributing to disease progression[159];

while some may even selectively kill tumour cells. In fact, the observation that advanced tumours spontaneously regress in certain cancer patients following an infection prompted inquiry into the use of bacteria as a cancer therapeutic tool[160]. A multitude of bacterial strains across diverse genera have now been explored for both their standalone therapeutic potential, as well as their potential to be used as vectors for cancer therapy[160].

Clearly, one needs to sift through numerous cancer-microbe associations to uncover any true oncogenic pathogens. Further, an oncogenic pathogen, even if present at very low concentrations, may affect the system in a manner analogous to the butterfly effect, so that one cannot readily pinpoint the original trigger. As summarized in the keystone-pathogen hypothesis, certain low-abundance microbial pathogens can orchestrate inflammatory disease by remodeling a normally benign microbiota into a dysbiotic one. This is based on the observation that certain species have disproportionately large effects on their communities, given their abundance, and are thought to form the 'keystone' of the community's structure[161]. As an example, it has been proposed that *Porphyromonas gingivalis*, a minor constituent of periodontal biofilms (which is not able to induce periodontitis by itself), impairs the host's innate immune response, leading to dysbiosis-induced inflammation, which manifests as periodontitis[161]. Similar scenarios may exist in the case of IBD and CRC.

Established oncogenic pathogens

Established oncogenic viruses and the cancers they cause include Epstein–Barr virus and Burkitt's Lymphoma; Hepatitis C and B viruses and hepatocellular carcinoma; Human Herpesvirus 8 and Kaposi's sarcoma; and Human papillomavirus (HPV) and cervical cancer[158,162]. Among bacteria, only *Helicobacter pylori* has been established as oncogenic, through evidence from large-scale epidemiological studies with long-term follow up data[158,163].

In the case of HPV, which expresses oncogenic proteins that transform infected host cells[164], the pathogenic mechanisms involved are very clear. In many cases however, multiple pathogenic methods may be implicated. With *Helicobacter pylori* for instance, although chronic inflammation appears to be a major oncogenic driving force, the *H.*

pylori effector protein, CagA, which is injected directly into host cells, induces multiple cancer-related pathways in the host, including ERK-MAPK signaling pathway, which likely further contributes to oncogenesis[165].

Despite diverse pathogenic mechanisms, certain shared characteristics between oncogenic pathogens are worth noting: Firstly, only a small proportion of individuals initially infected with the relevant pathogen will develop cancer, and secondly, this progression is dependent on chronic, long-term infection. For example, 10–90% of the population is chronically infected with *H. pylori,* depending on geographical location[166], of which 10–15% will experience recurrent gastroduodenal ulceration, but only 1–2% of infected individuals will eventually develop gastric cancer[166]. These figures point to obvious genetic, environmental and strain-specific risk modifiers that govern susceptibility to the oncogenic potential of these microbes. Finally, the proportion of a given cancer type that may be ascribed to the relevant infectious agent is variable: virtually 100% in the case of HPV-related cervical cancers worldwide[5,20], and at least 60% of all gastric cancer cases (and 75% in the case of noncardia gastric cases), in the case of *H. pylori*[20,21].

The CRC-associated microbiome

The human microbiome consists of approximately tenfold as many cells as the human body and collectively weighs 1–2kg[56], with the colon containing an estimated 10^{11}–10^{12} cells/g of luminal contents[167]. Through two major initiatives—the Human Microbiome Project, and the Metagenomics of the Human Intestinal Tract (MetaHIT) consortium, the relationship between the human microbiome, the environment and host genetics is steadily being uncovered, revealing possible intricate links to various diseases, including cancer, diabetes and obesity[168].

Factors influencing the composition of the colonic microbiome

In the healthy colon, the immune system tolerates antigens derived from commensal flora[169], the mucosal barrier provides a robust barrier to colonisation by pathogens, and beneficial bacteria dominate over potentially harmful bacteria—this state is referred to as eubiosis. On the other hand, dysbiosis is described as qualitative and quantitative

changes in microbial composition, changes in microbial metabolic activity and changes in the local distribution of specific microbes[170]; dysbiosis may occur due to various environmental insults including antibiotics, long-term dietary patterns[128,171] and psychological stress[172,173]. Dysbiosis is correlated with various chronic disease states including diabetes[174], IBD[175] and CRC[16,176,177]. Based on next generation sequencing efforts, temporal associations between microbiota and developing tumours in CRC have been described[178]; similarly, a progressive decrease in SCFAs and a change in microbial profile from healthy controls to adenomas to CRCs have been noted, which suggests that dysbiosis occurs at the adenoma stage or earlier, and intensifies with tumour development[179]. The question of whether dysbiosis is a cause or a consequence of disease is at least partially answered by transplantation experiments where microbiota is transplanted from diseased to germ-free healthy animals, an event which is accompanied by transference of the donor disease phenotype. This has been demonstrated in the case of obesity, metabolic syndrome and colitis[180].

Long-term dietary patterns, particularly regarding the proportion and type of fats and polysaccharides, directly affect microbial composition due to substrate specificity[128]. While *Prevotella* spp. dominate under carbohydrate-rich diets, *Bacteroides* spp. are associated with diets high in proteins and animal fat[171]. The bacterial metabolite profile produced under different diets directly impact host health; under sulfate-rich diets for example, sulfides are produced by sulfate-reducing bacteria, leading to the production of hydrogen sulfide which damages the colonic mucosa and increases mucosal permeability[170]. Intriguingly, microbial composition might explain some of the risk modulatory effects of diet on CRC. O' Keefe et al. found that Native Africans (risk cancer incidence < 1:100 000), who consumed more resistant starch and less red meat and animal products, had significantly higher levels of butyrate (and total SCFA) when compared to African Americans (risk 65:100 000) and Caucasians (risk 50:100 000) who consumed a high animal-protein and -fat diet[23,27]. Since the presence of butyrate can be wholly attributed to bacterial activity, it can be concluded that the natural microbiota may be mediating a proportion of the risk associated with consuming a high animal-protein and -fat diet. In terms of microbial composition, *Prevotella* spp. and *Bacteroides* spp. (corresponding to different gut enterotypes)[181], predominate in Native

Africans and African Americans, respectively[27]. Together with a decrease in SCFAs, a Western-style diet rich in animal fats, red meat and alcohol (known CRC risk factors) promote microbial production of carcinogens[182] and decreases barrier function, which facilitates *E. coli* colonisation[183].

Psychological stress can also affect microbial composition[172,173]; mechanisms that may explain this effect include impairment of intestinal barrier function[184–186](which is partially explained by reduced production of mucins[187]); decreased production of immunoglobulin A (IgA); as well as the production of various catecholamines, which increase dramatically in response to stress and encourage the growth of potentially pathogenic microbes[170]—norepinephirine in particular has been linked to overgrowth and increased expression of *E. coli* virulence factors[170,188,189].

Finally, host-genetic factors influence the composition of the gut microbiome. This has been most evident in the case of IBD susceptibility genes. Determining to what extent microbial variation depends on the host genome as a whole has been more challenging. Studies comparing the microbiome of dizygotic vs. monozygotic twins, have thus far failed to identify a significant host-genotype effect, perhaps due to lack of statistical power[180]. Examples, of single gene effects include carcinoembryonic antigen-related cell adhesion molecule 6 (CEACAM6), which is often overexpressed in patients with Crohn's disease; CEACAM6 acts as a receptor for adherent-invasive *E. coli* (AIEC) that subsequently colonise the intestinal mucosa and induce inflammation[190]; in mice, NOD2-dependent dysbiosis predisposes mice to transmissible colitis and CRC by creating a pro-inflammatory environment and disrupting epithelial barrier function. The microbiota of NOD2-deficient mice, which are transferrable to wild type mice, sensitizes the colonic mucosa to injury[191]; finally fucosyltransferase 2 (FUT2), is responsible for the synthesis of an oligosaccharide moiety (H antigen) that acts as an attachment site as well as a carbon source for bacteria. Loss-of-function mutations in FUT2 increase susceptibility to Crohn's disease, through altering microbial composition (which leads to chronic inflammation)[192,193]. Other disease-predisposing genes that affect microbial composition include defensin genes, *MYD88* and *HLA* genes[180].

CRC-associated bacteria

Given the current evidence, at least three theories can explain a causative role of bacteria in CRC: 1) Dysbiosis creates a pro-carcinogenic environment due to the lack of beneficial bacteria and a relative increase in pro-carcinogenic bacteria with non-specific oncogenic effects such as inflammation, and the production of pro-carcinogenic metabolites; 2) Chronic infection with a single oncogenic pathogen causes CRC in susceptible individuals (e.g. *H. pylori* infection in gastric cancer) through directly manipulating host cellular signaling and/or non-specific effects such as chronic inflammation or 3) A combination of 1) and 2), where dysbiosis promotes a pro-carcinogenic environment, and facilitates colonisation by opportunistic pathogens (with oncogenic potential) due to disruption of mucosal barrier function.

Recent studies, either through epidemiological and/or *in vitro* results, have linked Enteropathogenic *E. coli* (EPEC)[9,10], Fusobacterium.[16,194–196], *Streptococcus gallolyticus*[12,13,197], *Enterococcus faecalis*[198–202], Enterotoxigenic *Bacteroides fragilis* (ETBF)[14,15,203], as well as viruses (polyoma viruses (JC and SV40), human papillomavirus, Epstein Barr virus, and cytomegalovirus)[204,205] to CRC. The oncogenic potential of the bacteria, as well as suspected bacterial components implicated in the aetiopathogenesis of CRC are summarised in Table 1 in Chapter 4.

Escherichia coli and CRC

While *E. coli* has a core genome of ~2000 genes, the average *E. coli* genome consists of 4721 genes, with a complete pool of around 8000 genes; a high rate of recombination underpins the adaptability of *E. coli* strains and results in the high level of strain diversity. Although up to 300 *E. coli* strains may be present in any individual, each person is commonly colonised by a single dominant strain that constitutes more than half the colonies isolated from faeces[206]. The dominant strain is usually also the resident strain, which may be present for months or years[206].

Both IBD and CRC patients are commonly colonised by *E. coli* strains with adherent and/or invasive properties[127,207,208]. Using 16S rRNA sequence analysis, Swidsinski et al. showed that while only 3% of biopsy specimens from healthy controls contained any

type of bacteria (that had successfully colonised the epithelium), ~90% of patients with adenomas or carcinomas had 10^3–10^5 bacteria in both malignant and macroscopically normal samples. *E. coli* was the dominant bacterial species in 3%, 62% and 77% of patients who were asymptomatic, had adenomas, or carcinomas, respectively; further, *E. coli* was partially intracellular in 87% of *E. coli*-positive patients[207].

These findings are supported by Martin et al. who demonstrated the presence of intramucosal *E. coli* in 33% and 14% of colon cancer tumour and matched normal biopsies respectively vs. in 9% of controls[127]. They further demonstrated that strains isolated from tumour or normal biopsies from a given patient were identical.

Only recently has further characterization identified a specific group of *E. coli* strains associated with CRC: These strains, first identified in patients with Crohn's disease, were classified as adherent-invasive *E. coli* (AIEC)—a new pathogenic group—in 1999, based solely on phenotypic traits, and the absence of genes previously related to invasion in pathogenic strains of *E. coli*[209]. With no AIEC-specific genes identified, these strains are identified by their ability to adhere to and invade intestinal epithelial cells; the ability to survive and replicate in macrophages without triggering host cell death; and the ability to trigger TNF-α release by infected macrophages[210].

AIEC

Recent studies have identified *E. coli* virulence factors commonly found in AIEC that might be relevant to CRC; these include *pks*, *afaC*, *lpfA* and *cnf1*. The polyketide synthases (*pks*) genomic island encodes the colibactin (*clb*) genotoxin, which induces DNA double-strand breaks in an ex vivo mouse intestinal model[11]. *In vitro*, epithelial cells infected with *pks*+ strains displayed increased mutation frequency and anchorage-independent colony formation, which suggests oncogenic potential[11]. Moreover, *pks*-positive *E. coli* promote CRC progression in colitis-susceptible interleukin-10–deficient ($Il10^{-/-}$) mice, through its genotoxic capabilities[211].

Clinically, these results are supported by the finding that *E. coli* strains expressing *afaC, lpfA* or *pks* (or a combination thereof) were significantly more common in CRC cases compared to healthy controls[212,213].

EPEC

EPEC causes severe watery diarrhea, particularly among infants in developing countries, and asymptomatically infects an unknown proportion of adults. EPEC is closely related to enterohaemorrhagic *E. coli* (EHEC) and *Citrobacter rodentium,* which are collectively referred to as attaching/effacing *E. coli* (AEEC) (not to be confused with AIEC), due to their ability to attach to epithelial cells and cause loss of microvilli (effacement). *C. rodentium* infects small animals and is widely used to model human EPEC and EHEC infection in mice. EPEC can be divided into typical and atypical EPEC (aEPEC); where aEPEC lack the EPEC adherence factor (EAF) plasmid[214].

The first clue that AEEC might have oncogenic potential is that *C. rodentium* causes transmissible colonic hyperplasia in immunocompetent mice[215]. Recent *in vitro* evidence, suggest that MSI-positive CRCs may be caused by EPEC infection in susceptible individuals by downregulating MMR proteins (MLH1 and MSH2) in CRC cell lines, in an attachment-dependent manner[9,10]. Following EPEC-infection of normal colonic mucosa samples, bacteria were able to enter ~10% of colonic crypts and subsequently attach to areas that contain undifferentiated proliferative epithelial cells[9], which are believed to be the cells-of-origin for CRC[113]. Maddocks et al. found AEEC in 25% (5/20) of formalin-fixed paraffin-embedded (FFPE) CRC samples and in none of the matched normal samples, as measured by PCR detection of the intimin (*eae*) gene; AEEC were present at tens to tens of thousands of bacteria/FFPE section judged by immunofluorescence staining[9].

More recently, Maddocks et al. showed that *in vitro* EPEC-induced depletion of the mismatch repair proteins occurs at the protein level and that depletion of MLH1 and MSH2 was dependent on mitochondrial targeting of the EPEC effector protein EspF, which caused depletion of mismatch repair proteins and increased mutational frequency of infected cells[10].

In a proteomic analysis of the response of intestinal epithelial cells to EPEC infection, MSH6 protein levels were increased more than 7-fold in a Caco-2 EPEC-infected cell line compared to a type III secretion-deficient mutant EPEC infection, while MLH1- and MSH2-proteins were not reported. In the same system, DNA-damage binding

protein 1 (DDB1) was downregulated two-fold in wild type EPEC-infected cells[216]. Notably, DDB1 is involved in host-virus interaction e.g. EB-virus, which has been noted as a causal factor in gastric cancer[205].

Many EPEC effectors have been studied in detail and, in terms of their potential to promote tumourigenesis, exhibit disparate effects: NIeC and NIeD inhibit NF-kB and AP-1 respectively, both oncogenes. On the other hand NIeH1 activates the oncogene Bax-inhibitor-1, while EspG and EspZ activate p21-activated kinase (PAK) and focal adhesion kinase (FAK), respectively. A recent review by Wong et al. provides a complete overview of all 27 EPEC effectors that have been identified to date, and their cellular function[217].

Fusobacterium spp.

An association between Fusobacterium spp. and CRC was discovered via metagenomic analyses; multiple independent studies have now confirmed enrichment with Fusobacterium spp. in biopsy samples from CRC patients compared to the adjacent normal mucosa and compared to healthy controls[16,194,195]. McCoy et al. also found significantly higher numbers of Fusobacterium spp. in the normal mucosa of patients with adenomas compared to healthy controls, suggesting a possible early role for Fusobacterium in CRC[195]. In the $APC^{Min/+}$ mice, *F. nucleatum* infection accelerated tumourigenesis, characterised by infiltration of specific myeloid cell subsets into tumours and an Nf-kB pro-inflammatory signature (which included COX-2, TNF-α and IL-8) that is shared in human Fusobacterium-infected CRC tissue[218]. However, *F. nucleatum* infection alone may not be oncogenic since it stimulates proliferation in CRC cell lines but not in non-neoplastic cell lines. Further, *F. nucleatum* is found at significantly higher levels in tumours compared to the adjacent normal mucosa[16,194,219], which suggests that the tumour provides a niche environment that is exploited by these bacteria.

Streptococcus gallolyticus

Streptococcus bovis, which is present in 2.5–15% of healthy individuals[159], causes 10–15% of bacterial endocarditis cases[197]. An association between *Streptococcus bovis*-

bacteraemia/endocarditis and CRC was first investigated in the 1950's[220]. Subsequently, a multitude of retrospective studies and case control studies have reported this association, but with varying incidence rates (6–67%)[13,197].

S. bovis has since been sub-classified as *S. gallolyticus* ssp. *gallolyticus* (*S. bovis* I*), S. infantarius* ssp. *infantarius* (*S. bovis* II*/*1) and *S. gallolyticus* ssp. *pasteurianus* (*S. bovis* II*/*2)[221]. *S. bovis* associated CRC was subsequently found to be much more strongly associated with *S. gallolyticus* ssp. *gallolyticus* compared to the other subtypes (pooled odds ratio = 7.26; 95% confidence interval = 3.94–13.36)[13].

A study investigating the incidence of *S. gallolyticus* ssp. *gallolyticus* infection among CRC patients with or without a history of *S. gallolyticus*/bovis bacteremia, found infection in 48.7% and 32.7% of tumours, respectively; similar levels of *S. gallolyticus* were detected in the corresponding normal tissue samples, while only 2% of control biopsies were infected[12]. Interestingly, these results were limited to tissue samples (with no significant difference between groups in feacal or mucosal infection), which indicate intimate adherence/invasion in CRC patients. *S. gallolyticus* infection was not associated with stage, grade, or location of the tumours, or with age or gender[12].

S. gallolyticus has the following pro-carcinogenic traits: *S. gallolyticus* or its wall extracted antigens (WEA) stimulates COX-2 expression in rats pretreated with the carcinogen azoxymethane, thereby triggering MAPK signaling and the progression of preneoplastic lesions[4]; this is supported by the increased expression of the pro-inflammatory cytokines (IL-1 and COX-2) in *S. gallolyticus*-positive CRC patients in both tumour and normal tissues compared to *S. gallolyticus*-negative CRC patients[12].

Enterococcus faecalis

Traditionally classified as a human commensal, *E. faecalis*, is in fact one of the most common causes of nosocomial infections[222]. *E. faecalis* has been linked to CRC due to significantly higher fecal levels of the bacterium in CRC patients compared to healthy controls[198].

Mechanistic ties to oncogenesis include the production of extracellular reactive oxygen species (ROS) and induction of COX-2 expression, which leads to inflammation and CRC in IL-10 knockout mice[199,202,223], which induces aneuploidy and tetraploidy in an *in vitro* model of infection[223].

Enterotoxigenic *Bacteroides fragilis* (ETBF)

ETBF is a pathogenic strain of the commensal *B. fragilis*, which secretes a heat-labile metalloprotease (*B. fragilis* toxin (BFT)) and causes human inflammatory diarrhea in susceptible individuals. In faecal samples, ETBF is found in ±12% of healthy controls[14,224], 27% of patients with diarrhea[224], and 38% of patients with CRC[14]. However, infection rates appear to vary largely by geographical location[225].

In addition to causing diarrhea in susceptible individuals, ETBF induces colitis and colonic tumours in Apc/+ min mice[15]; BFT cleaves the extracellular domain of E-cadherin, which triggers ß-catenin nuclear signaling, followed by c-MYC expression and cellular proliferation[203]; ETBF-induced colitis, colonic hyperplasia and tumour formation could be prevented by inhibiting the inflammatory cytokines IL-23 and IL-17.

Chapter 3: Study design, clinicopathological characterisation, and sample processing

Abstract

In this chapter, the aims and objectives of this thesis are stated and a schematic representation of the study design is supplied.

Next, the clinicopathological characteristics of our CRC cohort, sample preparation and storage methods, as well as sample size and power calculations are presented.

Our cohort consists mainly of mixed ancestry patients (70%), with around 90% of cases of sporadic origin, and a remaining 10% of cases of confirmed hereditary non-polyposis CRC (HNPCC). Regarding site of disease, 60% of CRCs were located in the rectum, 18% in the proximal colon and 22% in the distal colon. Furthermore, 81.6% of our cohort presented with stage II or III CRCs. We further recorded patient gender, age, BMI, familial history of cancer, smoking or alcohol consumption, as well as histopathological features noted in pathology reports and chemotherapy and/or radiation treatment received.

In order to minimize the effect of intra-sample heterogeneity for the fresh-frozen cohort (N=55), we simultaneously extracted DNA, RNA and protein from a single tissue sample, and bacterial detection and whole genome analysis (including transcriptomic and epigenomic analyses presented in Chapters 6–8) were conducted on these samples, when possible. In some instances, these could not be used do to lack of sample availability, poor sample quality or the nature of downstream applications (e.g. the extraction of DNA from gram-positive bacteria in tissue samples required a modified protocol). In addition to the main cohort (N=55), 18 formalin-fixed paraffin-embedded (FFPE) CRC tissues were sourced from Groote Schuur Hospital's archival materials in order to increase the number of MSI+ samples available for bacterial detection. Only Fusobacterium and EPEC (which had previously been associated with MSI in the literature) were measured in these samples.

Finally, an a priori sample size calculation was initially used as a rough estimate of the required sample size for a paired tumour vs. normal gene expression analysis. We then conducted post-hoc power analyses to evaluate the power to detect differential gene expression for the various bacteria-specific comparisons of interest (e.g. genes differentially expressed in EPEC+ vs. EPEC– groups or in groups with high or low/no colonization by Fusobacterium).

Study aims and objectives

The main aim of this thesis was to gain insight into the potential contribution of CRC-associated bacteria in the aetiopathogenesis of CRC by leveraging both host genomic and clinicopathological data as well as to investigate patterns of tissue colonisation between different CRC-associated bacteria.

As described in Chapter 2, several bacterial species have been linked to colorectal cancer, but these species have not been concurrently quantified across a single cohort, nor have any bacteria been studied in parallel with the host genomic and clinicopathological features. The aim of this study was to address these issues using a South African cohort of CRC patients in order to 1) gain insight into the potential contribution of CRC-associated bacteria in the aetiopathogenesis of CRC and 2) determine the quantitative relationship of tumour and adjacent normal tissue colonisation between the various CRC-associated bacteria studied here.

The objectives were to 1) quantify CRC-associated bacteria in a cohort of 55 paired tumour and adjacent histologically normal samples collected during surgical resection, as well as in an additional 18 FFPE CRCs (17 of which were MSI+) for the detection of EPEC and Fusobacterium; 2) to determine their relationships to patient age, gender, ethnicity, stage of disease, site of disease and MSI status (Chapter 4); 3) evaluate the relationship between each bacterium and host gene expression (Chapter 8) and methylation changes (Chapter 6); and 4) to determine genomic subtypes of CRC using unsupervised clustering of gene expression data in the context of patient clinicopathological features and bacterial quantitation data; and 5) to gain a deeper biological understanding of the results from the objectives 1–4 using pathway analyses of the genomic subtypes obtained (Chapter 7). The study design is outlined in Figure 1.

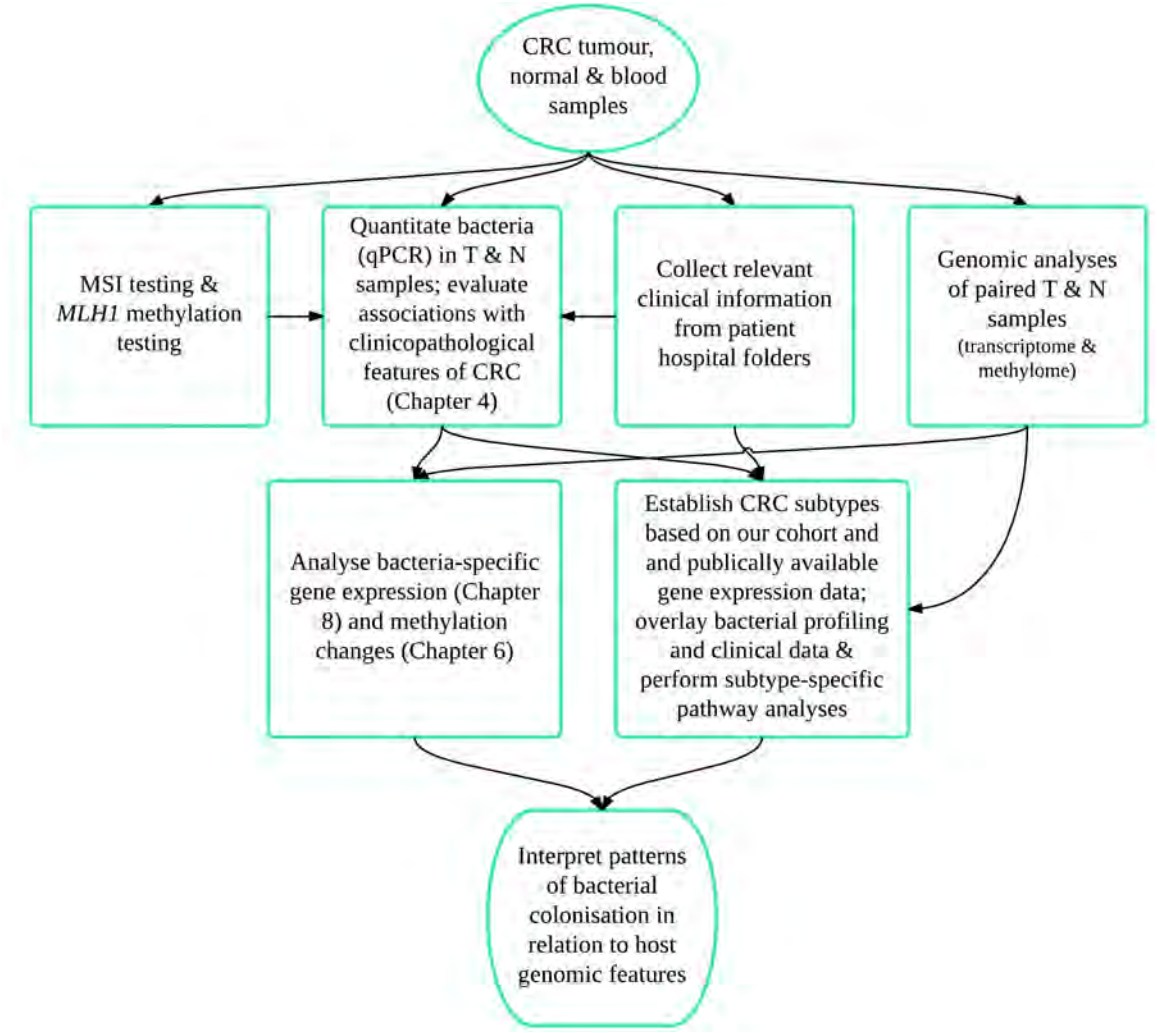


Figure 1: Thesis outline describing the use of samples from the primary cohort (fresh-frozen samples). T: tumour samples; N: normal samples. FFPE samples were used to supplement the number of MSI+ samples (N=17) for bacterial detection as described in Chapter 4.

Materials and methods

Sample collection and storage

This study consists of two cohorts: The main cohort consists of 55 paired colorectal patient samples (adenocarcinoma tissue and adjacent macroscopically normal mucosa) together with patient blood samples (for MSI testing) that were collected during surgical resection at the Groote Schuur Hospital, as part of a previous study in our laboratory by Dr. Amirtha Ganesh. The samples were frozen immediately in liquid nitrogen and

stored at -80.0°C; the second cohort was sourced in order to obtain more patients with sporadic microsatellite instability (MSI-H). For this cohort, 18 adenocarcinoma samples were selected from archival formalin-fixed paraffin embedded (FFPE) specimens that had previously been screened for MSI by immunohistochemistry of the mismatch repair genes *MLH1*, *MSH2* and *MSH6*; these patients were referred for MSI testing if CRC was diagnosed under the age of 50, or if CRC was diagnosed in two or more first or second degree relatives with an HNPCC-related tumor. For our purposes, we selected patients with MSI (absence of staining of one or more MMR proteins), who also had mutational screening data available so that we could distinguish sporadic from HNPCC cases. Of the 18 patients, two had confirmed mutations in the *MLH1* mismatch repair gene and were therefore classified as HNPCC. Although the main focus of this study is to characterize CRC-associated bacteria in sporadic CRCs, HNPCC patients were included, because they provide a valuable reference for investigating the putative role of EPEC and Fusobacterium in causing MSI (because MSI is linked to specific mutations in HNPCC).

Ethical consent was obtained for the collection of samples (UCT HREC REF 416/2005) and each patient provided written and informed consent to collect tissue and blood samples during surgical resection. Ethical consent has subsequently been renewed to collect additional patient samples and to continue analyses of the existing cohort (UCT HREC REF 366/2010).

A priori sample size calculation

Our initial estimate of the sample size for gene expression analyses was based on a standard paired tumour vs. normal comparison and performed in R using the Bioconductor package SizePower[226]; accordingly, a minimum of 17 paired patient samples were required in order to detect a minimum fold change of 2 at a false negative rate of 0.05–the parameters used are specified in Table 1. The anticipated standard deviation of the difference in log-expression between matched tumour and normal samples ($\sigma = 0.7$) was based on a conservative estimate from the literature[227].

Table 1: Parameters specified for the power analysis conducted for analysis of tumour vs. paired normal samples.

Parameter	Value
α[1]	0.001
β[2]	0.05
μ[3]	1
σ[4]	0.7

[1]The acceptable probability of a type I error for any single gene (false positive rate).

[2]The acceptable probability of a type II error for any single gene (false negative rate).

[3]Absolute mean difference in log-expression between treatment and control units as postulated under the alternative hypothesis H1 (equates to a fold change of two in this case).

[4]Anticipated standard deviation of the difference in log-expression between matched treatment and control units.

Post-hoc power analysis for detection of bacterially-associated differential gene expression

As detection of CRC-associated bacteria in tissue samples was conducted after gene expression analysis of 19 sample pairs, post-hoc power analyses were conducted to assess the power to detect bacterially-associated differential gene expression for each of the six bacterial species investigated.

In sample size and power calculations it is usually assumed that the two groups being compared are of equal size; in practice however, this is rarely the case. We therefore estimated an adjusted 'equal sample size' for each comparison of interest, using the online resource StatsToDo[228]. We show the actual and adjusted sample sizes for the comparison between high-level colonisation by a particular bacterium compared to low-level or no infection for tumours (Table 2a) and normal samples (Table 2b). The distinction between high- and low-level infections was set based on the quantitative distribution of bacteria across all samples, where samples that fell in the third quartile were graded as high-level (see Chapter 4 for further detail).

Table 2a. Description of sample groups for gene expression analysis, *tumour* samples: comparing *high-level colonisation to low/no infection*.

	High level colonisation	Low/no infection	Adjusted sample
Fusobacterium	7	12	8
afaC+ AIEC	5	14	7
ClB+ AIEC	1	18	n/a

EPEC	n/a	n/a	n/a
ETBF	3	16	5
E. faecalis	2	15	3

Table 2b. Description of sample groups for gene expression analysis, *normal* samples: comparing *high*-level colonisation to low/no infection

	High-level	Low/no infection	Adjusted sample
Fusobacterium	2	17	3
afaC+ AIEC	1	18	n/a
ClB+ AIEC	1	18	n/a
EPEC	n/a	n/a	n/a
ETBF	2	17	3
E. faecalis	1	13	n/a

Next, we show the actual and adjusted sample sizes for the comparison of samples with or without colonisation by a particular bacterium for tumours (Table 2c) and normal samples (Table 2d).

Table 2c. Description of sample groups for gene expression analysis, *tumour* samples: comparing *infection-positive vs. –negative* samples.

	Infection+	Infection–	Adjusted sample
Fusobacterium	18	1	n/a
afaC+ AIEC	11	8	9
ClB+ AIEC	6	13	8
EPEC	3	16	5
ETBF	10	9	9
E. faecalis	7	10	8

Table 2d. Description of sample groups for gene expression analysis, *normal* samples: comparing *infection-positive vs. –negative* samples.

	Infection+	Infection–	Adjusted sample
Fusobacterium	16	3	5
afaC+ AIEC	10	9	9
ClB+ AIEC	6	13	8
EPEC	2	17	3
ETBF	10	9	9
E. faecalis	4	10	5

These adjusted samples sizes were used as input for power calculations using the R package sizepower[226], where we calculated the power to detect genes differentially expressed at fold changes of two (effect size = 1) or three (effect size = 1.6), for equal-size group comparisons. We arbitrarily selected group sizes of 3, 6 or 9, in order to obtain an estimate of power at each level (Table 3). The acceptable number of false positives (α) was specified as 0.001 and the variance term (σ^2_d) was calculated as 0.66, from the median residual standard deviation across all probes when comparing two randomly chosen treatment and control samples (using the R package limma[229]), as recommended in the sizepower package.

Table 3. Post-hoc power analyses for varying effect and sample sizes.

Number of samples per group*	Effect size (μ)	Estimated power
3	1	25%
6	1	65%
9	1	89%
3	1.6	80%
6	1.6	99%
9	1.6	100%

*Due to unequal sample sizes between classes, an adjusted sample size/group was calculated and used for power calculation[228].

Sample preparation

For MSI testing, DNA was isolated from tumour and adjacent normal colonic mucosal samples and from peripheral blood lymphocytes for each patient; DNA from tissue was isolated from ~6 mm^3 of tissue using the Gentra Puregene Tissue Kit (Qiagen). Briefly, after adding cell lysis solution, the samples were incubated at 65°C for 15 minutes. In order to obtain maximum yield, Proteinase K was added to each sample followed by continued lysis at 55° C for 3 hours. If the tissue was not completely lysed after 3 hours, samples were incubated overnight at 55° C. Blood DNA was extracted from peripheral blood lymphocytes using the Gentra Puregene Blood Kit (Qiagen), according to the manufacturer's instructions.

Subsequent to the extraction of DNA for MSI testing, DNA, RNA and protein were simultaneously isolated from paired patient samples using a Dounce homogenizer and the AllPrep DNA/RNA/Protein kit (Qiagen), according to the manufacturer's

instructions. These DNA and RNA amplicons were used for genomic analysis (gene expression and methylation arrays, and for the detection of gram-negative bacteria). For the detection of gram-positive bacteria, a separate extraction had to be performed where DNA was extracted from ~6 mm^3 of tissue using the QIAamp® DNA Mini Kit (Qiagen). Briefly, each sample was incubated in 180 µl of lysozyme (20mg/ml) for 40 min at 37°C; after adding 20 µl of proteinase K, samples were incubated at 56°C until the tissue was completely lysed (at least 4 hours, or overnight if tissue was still visible after 4 hours); samples were next incubated for 30 min in Buffer AL (supplied with QIAamp® DNA Mini Kit), and thereafter DNA was isolated according to the manufacturer's instructions.

For FFPE samples, DNA was extracted using the RecoverAll Total Nucleic Acid Isolation Kit (Ambion). FFPE slides were prepared from the Groote Schuur Hospital archival FFPE wax blocks. Of the 17 patients for whom FFPE material was requested, only 6 were successfully retrieved due to lack of sample availability; FFPE slides were prepared, and tumour margins were demarcated using hematoxylin and eosin guide slides, by the department of Anatomical Pathology at UCT. Fourteen additional patients' FFPE slides (that had previously been prepared from archival FFPE wax blocks) were obtained from the Human Genetics Laboratory at UCT; unfortunately, tumour margins had not been demarcated for the majority of these slides. In cases where tumours had been demarcated, tissue inside or outside of the demarcated area was processed separately. Samples were prepared by scraping the tissue from the glass slides using a sterile scalpel blade and transferring it to a 2mL Eppendorf tube. After deparraffinization in 100% xyleen, samples were incubated in 180 µl of lysozyme (20mg/ml) for 40 min at 37°C, followed by incubation for 42 hours at 50°C in Proteinase K. For the remainder of the protocol, DNA was isolated according to the manufacturer's instructions. Of the 20 FFPE samples obtained, 18 had sufficient DNA quantity and quality for downstream analyses, and 17 were MSI+.

Microsatellite instability (MSI) testing

MSI testing was conducted on DNA extracted from paired tissue samples, as well as the corresponding blood samples for each patient, using allelic profiling of the Bethesda

panel of microsatellite markers, which includes two mononucleotide repeats (BAT25 and BAT26) and three dinucleotide repeats (D2S123, D17S225 and D5S314). Samples were classified as microsatellite stable (MSS), microsatellite instable-low (MSI-L) or microsatellite instable-high (MSI-H) if they had 0, 1 or at least 2 of the 5 markers showing instability, respectively[230], based on electrophoretic analysis of PCR products, which was conducted on an ABI PRISM® 3130xl Genetic Analyzer. Primer sequences were taken from Loukola et al.[231], and are indicated in Table 4. The five primers sets were divided into two multiplex PCRs; the optimized parameters are shown in Table 5. For each pair, the forward primer was labeled with one of the fluorescent markers 6-carboxyfluorescein (FAM), tetrachloro-6-carboxyfluorescein (TET) or hexachlorofluorescein (HEX).

Table 4. Primers used for MSI testing according to the Bethesda panel of markers.

Microsatellite marker	Primers (3'–5')	Size range (bp)
BAT25	(F) HEX-TCGCCTCCAAGAATGTAAGT	90–125
	(R) TCTGGATTTTAACTATGGCTC	
BAT26	(F) FAM-TGACTACTTTTGACTTCAGCC	80–120
	(R) AACCATTCAACATTTTTAACCC	
D2S123	(F) FAM-AAACAGGATGCCTGCCTTTA	197–227
	(R) GGACTTTCCACCTATGGGAC	
D5S346	(F) TET-ACTCACTCTAGTGATAAATCG	96–122
	(R) AGCAGATAAGACAGTATTACTAGTT	
D17S250	(F) FAM-GGAAGAATCAAATAGACAAT	140–170
	(R) GCTGGCCATATATATATTTAAACC	

For each pair, the forward primer was labeled with one of the fluorescent markers 6-carboxyfluorescein (FAM), tetrachloro-6-carboxyfluorescein (TET) or hexachlorofluorescein (HEX).

Table 5. Cycling conditions used for multiplexed PCR MSI testing.

Step	Duration	Temperature	Cycles
Denaturation	60 sec	95°C	
Denaturation	1 sec	96°C	x35
Annealing/extension*	10 sec	49.9°C or 57.5°C	
Final extension	10 sec	72°C	

*An annealing/extension temperatures were 49.9°C for the BAT25/26 multiplex and 57.5°C for the D2S123/D5S346/D17S250 multiplex.

For the FFPE cohort immunohistochemistry (IHC) had previously been performed for MSH2, MSH6 and MLH1 by the division of Anatomical Pathology at the Groote Schuur Hospital. Samples that displayed absence of staining for any of the mismatch repair proteins evaluated were considered to have MSI-H. IHC of MMR proteins has been shown to have high sensitivity (92.7%) and specificity (100%) in detecting MSI[232]. Originally, patients were referred for IHC analysis if CRC was diagnosed under the age of 50 and/or if two or more first or second-degree relatives had an HNPCC-related tumor.

Results

Cohort characterisation

We retrieved clinical information from patient hospital folders for 61 of the 68 patients initially sourced by Dr. Amirtha Ganesh. We recorded patient demographics, treatment summaries, pathology reports, and chemo- and radiotherapy treatment records in a MySQL database. We further recorded patient age, BMI, familial history of cancer, smoking or alcohol consumption, as well as histopathological features and chemotherapy and radiation treatment received.

The clinicopathological characteristics of the 55 fresh-frozen paired samples (of the 68 collected) used in this study are summarized in Table 6. The mean age of patients was 59 (SD±15.3), while gender was divided equally. MSI testing was performed for the 32 patients who had not received preoperative radio- or chemotherapy: 7 were MSI-H (of which 4 were HNPCC), 3 MSI-L, while the remaining 22 were MSS. For comparisons using MSI status, missing cases (where MSI status was not determined were excluded). The majority of cases were stage II or III cancers (81.6%), while stage I- and IV-cancers accounted for 12.2% and 6.1% of cases, respectively. The cohort consisted of 60% rectal and 40% colon cancers, with proximal cancers accounting for 45% of colon cancers. The majority of our cohort was of mixed-ancestry (70.4%), while patients of

Caucasian (14.8%), black (11.1%) and Indian ethnicities (3.7%) made up the rest of the cohort.

Table 6: Clinicopathological characteristics of the cohort of fresh-frozen tissues (N=55).

Feature (patients with missing data)	**Number of patients N=55**
Mean age (2)	59 (SD±15.3)
BMI (4)	26.8 (SD±4.7)
Gender (1)	
Female	27 (50%)
Male	27 (50%)
MSI status (23)	
MSS	22 (68.8%)
MSI-H	7 (21.9%)
MSI-L	3 (9.4%)
CRC Type	
HNPCC	6 (10.9%)
Sporadic	49 (89.1%)
Tumour stage (6)	
I	6 (12.2%)
II	18 (36.7%)
III	22 (44.9%)
IV	3 (6.1%)
Tumour site (5)	
Ceacum	4 (8%)
Ascending colon	1 (2%)
Hepatic flexure	1 (2%)
Transverse colon	3 (6%)
Splenic flexure	1 (2%)
Descending colon	3 (6%)
Sigmoid colon	3 (6%)
Rectosigmoid junction (RSJ)	4 (8%)
Rectum	30 (60%)
Radiation/Chemo received before resection (2)	
Yes	22 (41.5%)
No	31 (58.5%)
Ethnicity (1)	
Black	6 (11.1%)
Caucasian	8 (14.8%)
Indian	2 (3.7%)
Mixed-Ancestry	38 (70.4%)

In the case of age and BMI, mean values and their standard deviations (SD) are reported. The numbers in column 1 in brackets represent the number of patients with missing data in that category.

Microsatellite instability testing

Forty-eight of the 68 patients had tumour, normal, and blood samples available, and we performed MSI testing on the 32 patients who did not receive pre-operative chemo- or radiotherapy (the vast majority of these tumours were located in the colon). Of the 5 HNPCC patients, 4 were classified as MSI-H (≥ 2/5 markers show instability). Of the sporadic CRCs, 3 patients (9%) were classified as MSI-H, and an additional 3 as MSI-L, which is comparable to previous reports of MSI in 10–18% of CRCs[92,94]. An example of an electrophoretogram for 1 of the 5 markers evaluated for each sample can be seen in Figure 2; in this example microsatellite instability presents as additional peaks on the left, which indicate an insertion.

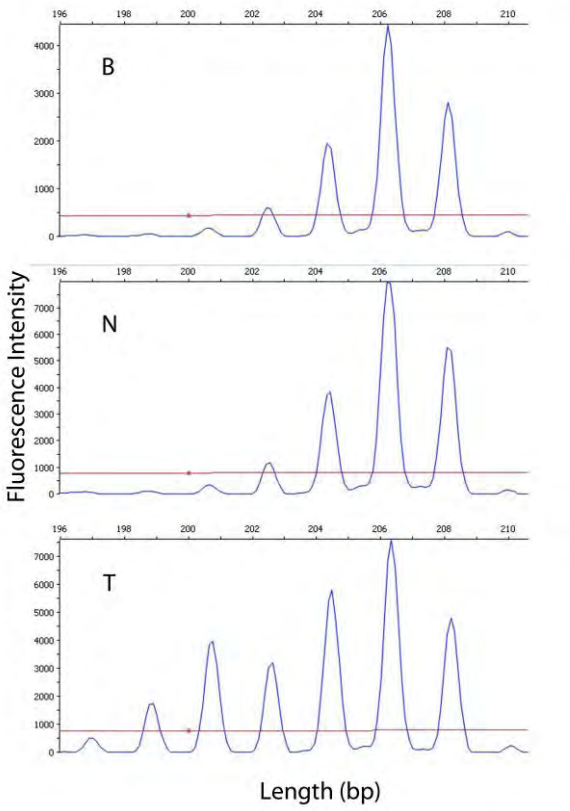


Figure 2: Capillary electrophoresis of a MSI positive cancer. Electropherograms of the D17S250 loci from normal colonic mucosa (N), blood (B) and tumour (T) tissues. The presence of additional peaks in T indicate that an increased proportion of T vs. N/B cells have an insertion in this marker which is caused by a dysfunctional mismatch repair system.

Discussion

In order to minimize the effect of intra-sample heterogeneity, we simultaneously extracted DNA, RNA and protein for each patient, and bacterial detection and whole genome analysis (including transcriptomic and epigenomic analyses presented in Chapters 6–8) were conducted on these samples, when possible.

Regarding sample size calculations, we initially estimated sample size based on a standard paired analysis between tumour and normal samples. However, given the design of our study, where we screened for multiple CRC-associated bacteria, the power to detect differences in gene expression or methylation across the genome depended on the number of positive and negative samples for a given species. Accordingly, post-hoc power analyses were conducted for the various comparisons of interest. In order to detect a fold change of 2, at least 9 samples per group are required. On the other hand, a fold change of 3 could be detected with 80% and 90% power with 3 or 6 samples per group, respectively.

Our cohort consists mainly of patients of mixed ancestry—it is important to note however that the racial distribution of our cohort largely mirrors the segment of the population who attend the Groote Schuur Hospital, rather than a representative sample of all CRCs in South Africa. The majority of our cohort has cancers of sporadic origin, but about 10% of our cohort has confirmed HNPCC. Regarding pathological characteristics, around 60% of our cohort has rectal cancers, and of the remaining 40%, proximal and distal colon cancer occurred in 45% and 55% of cases, respectively. Further, the majority of our cohort presented with stage II/III CRCs.

Chapter 4: Quantitative profiling of colorectal cancer-associated bacteria reveals associations between *Fusobacterium* spp., enterotoxigenic *Bacteroides fragilis* (ETBF) and clinicopathological features of CRC

Abstract

Various studies have presented clinical or *in vitro* evidence linking bacteria to colorectal cancer, but these bacteria have not previously been concurrently quantified by qPCR in a single cohort. Here, many of these bacteria (Fusobacterium spp., *Streptococcus gallolyticus, Enterococcus faecalis*, Enterotoxigenic *Bacteroides fragilis* (ETBF), Enteropathogenic *Escherichia coli* (EPEC), and afaC- or pks-positive *E. coli*) are quantified in paired tumour and normal tissue samples from 55 colorectal cancer patients. Further, associations between a) the presence and b) the level of colonisation by each bacterial species and site and stage of disease, age, gender, ethnicity and MSI-status are determined.

With the exception of *S. gallolyticus*, we detected all bacteria profiled here in both tumour and normal samples at varying frequencies. ETBF (FDR=0.001 and 0.002 for normal and tumour samples) and *afaC*-positive *E. coli* (FDR=0.03, normal samples) were significantly enriched in the colon compared to the rectum. ETBF (FDR=0.04 and 0.002 for normal and tumour samples, respectively) and Fusobacterium (FDR=0.03 tumour samples) levels were significantly higher in late stage (III/IV) colorectal cancers. Fusobacterium was by far the most common bacteria detected, occurring in 82% and 81% of paired tumour and normal samples. Fusobacterium was also the only bacterium that was significantly higher in tumour compared to normal samples (p=6e-5). Significant associations between high-level colonisation by Fusobacterium and MSI-H (FDR=0.05); age (FDR=0.03); or *pks*-positive *E. coli* (FDR=0.01) were also discovered. Furthermore, the EPEC detected here was exclusively identified as *atypical* EPEC, which has not been previously reported in association with colorectal cancer.

By quantifying colorectal cancer-associated bacteria across a single cohort, inter- and intra-individual patterns of colonisation not previously recognised were uncovered and

important associations with clinicopathological features were identified for *Fusobacterium* and ETBF.

Introduction

A causal link between specific pathogens and numerous cancers has now been firmly established. Clear evidence exists that the vast majority of cervical cancers are directly caused by colonisation by human papillomavirus (HPV)[233]. Similarly, *Helicobacter pylori* is a known risk factor for the development of gastric cancer and is considered a class I carcinogen by the WHO [234,235].

The possibility of oncogenic bacteria in the colon was already evident in the 1950s when a clinical association between *Streptococcus bovis* bacteraemia/endocarditis and CRC was discovered[236]. Subsequently, multiple studies have demonstrated enrichment with specific bacterial pathogens in faecal or tissue samples of CRC patients, including, Fusobacterium[194,195,237], *S. gallolyticus*[12,13,197], *E. faecalis*[198] and Enterotoxigenic *Bacteroides fragilis* (ETBF)[238].

Previously, 16S rRNA profiling of CRC paired tumour and normal biopsies has revealed that while only 3% of biopsy specimens from healthy controls contained any type of bacteria, ~90% of patients with adenomas or carcinomas had bacterial concentrations of 10^3–10^5 colony-forming units per microliter in both malignant and macroscopically normal samples[207]. This clearly demonstrates the susceptibility of these patients to pathogenic colonisation of the normally sterile colonic epithelium—not only in existing tumour tissue, but also in the surrounding macroscopically normal tissue, which may suggest a pre-existing risk to colonisation/infection.

Based on both *in vitro* and *in vivo* observations, bacterially-driven oncogenic mechanisms in CRC have been proposed to include activation of Wnt signaling (ETBF[203], EPEC[239], Fusobacterium[196]), pro-inflammatory signaling (*E. faecalis*[200,201], *S. gallolyticus*[240,241]) and genotoxicity (EPEC[10], AIEC[11,212,213]).

The oncogenic potential of these bacteria, as well as suspected bacterial components implicated in the aetiopathogenesis of CRC were discussed in Chapter 2 and are summarized in Table 1.

Table 1: Summary of the putative oncogenic mechansims and the bacterial components implicated in CRC pathogenesis for the six bacterial species quantified in this study.

Bacterial species	**Support for putative oncogenic mechanism**	**Suspected bacterial components implicated**
EPEC	Downregulates mismatch repair proteins *in vitro*[9,10]; increases mutational frequency *in vitro*[10].	*espF*[10]
Escherichia coli with adherent/and or invasive properties.	Enriched in CRC patients[207,208,242]; CRC-associated strains commonly have genes related to M-cell translocation (*lpfA*)[213]; genotoxicity (*pks*)[11,211–213], or cell cycle modulation (*cnf1*)[212].	*pks*[11,212,213], *afaC*[213], *lpfA*[213] *cnf1*[212] and *cdt*[212].
Fusobacterium	Multiple independent metagenomic studies identify Fusobacterium as overrepresented in CRC tissue compared to matched normal mucosa and healthy controls[194,195,237]. *F. nucleatum* increases tumour multiplicity in an APC Min/+ mouse model[218]; Triggers ß-catenin nuclear signaling[196].	*FadA*[196]
ETBF	Enriched in faecal samples from CRC patients[238]; Triggers ß-catenin nuclear signaling; induces c-Myc expression and cellular proliferation[203]; increases colitis and tumour in a Min/+ mice model[15].	*B. fragilis* toxin (*Bft*)[203]
Streptococcus gallolyticus	Enriched in CRC patients with[12,13,197,243] and without bacteremia[12]. *S. infantarius* or its wall extracted antigens promote progression of preneoplastic lesions in rats and promotes pro-inflammatory *COX-2* signaling[240,241].	Cell wall extracted antigens
E. faecalis	Enriched in faecal samples from CRC patients[198]; Produces extracellular superoxide[199], promotes inflammation and CRC in IL-10 knockout mice[200,201], and promotes *COX-2*-related chromosomal instability[223].	Reactive oxygen species (superoxide, hydrogen peroxide)

To date, however, the presence and levels of multiple CRC-associated bacteria have not been examined across a single cohort. Further, to our knowledge, ETBF and *E. faecalis* have only been quantified by quantitative PCR (qPCR) in feacal samples of CRC patients, and EPEC has only been quantified in a small CRC cohort with archival FFPE samples[9]. Here qPCR was used to measure the presence of six bacteria, previously reported in association with CRC, in paired adenocarcinoma and adjacent normal mucosal samples; these include Fusobacterium, *Streptococcus gallolyticus, Enterococcus faecalis*, ETBF, EPEC and *afaC*- or *pks*-positive *E. coli*. Detailed participant-level characteristics are presented in Appendix A (Tables 1 & 2).

Materials and Methods

Cohort selection

Cohort selection for bacterial quantification is described in Chapter 3. Briefly, the primary cohort consists of 55 paired colorectal patient samples (adenocarcinoma tissue and adjacent normal mucosa), whilst the second cohort was sourced in order to obtain more patients with sporadic microsatellite instability (MSI-H). For this cohort, 18 adenocarcinoma samples were selected from archival FFPE specimens that had previously been screened for MSI by immunohistochemistry of the mismatch repair proteins MLH1, MSH2 and MSH6.

Sample preparation and MSI testing

DNA isolation protocols used for the extraction of gram-negative or gram-positive bacterial DNA from host tissue samples or from FFPE sections are described in Chapter 3. MSI testing performed for tissue samples and for FFPE sections are also described in Chapter 3.

Primers and control DNA

Primers for the detection of each bacterial species were sourced from the literature or designed in-house, and their specificity was confirmed using Primer BLAST[244]. All primers, along with their limits of detection (LODs) and qPCR efficiencies, are listed in Table 2.

Table 2: Primers used for bacterial detection together with their annealing temperatures, efficiencies and LODs.

Target	**Test gene**	**F & R primers (5'–3')**	**Product (bp)**	**Annealing T (C)**	**qPCR efficiency (%)**	**qPCR error**	**LOD***	**% +ve at LOD**
EPEC	*eaeA*[9]	F-GTGACGATGGGGATCGAT R1-ACGGCTGCCTGATAATGTT	150	70–60	87	0.010	4 (20fg)	73
	eaeA (intimin epsilon, gamma, zeta, alpha, pi, rho, beta, lambda, iota, kappa, eta, delta, xi, mu, kapp, jota)	R2-GGAACTGCATTGAGTAAAGGAG	70	70-60	80	0.055	4 (20fg)	100
	eaeA (intimin theta)	R3-GAAGCTGCATTGAGTAAAGAAG	70	60	ND	ND	~ 10[#]	ND
	bfp[245]	F-GGAAGTCAAATTCATGGGGGTAT R-GGAATCAGACGCAGACTGGTAGT	299	70–64	92	0.010	20 (100fg)	100
EHEC	*stx1*[246]	F-ACATTGTCTGGTGACAGTAGC R-CGACATTAAATCCAGATAAGAAGTAGT	114	70–60	86	0.008	16 (100fg)	83
	stx2[246]	F-ATGACAACGGACAGCAGTTAT R-CTGAACTCCATTAACGCCAGATA	116	70–60	89	0.015	16 (100fg)	100
AIEC	*Clb*[211] *(pks)*	F-GCGCATCCTCAAGAGTAAATA	280	60	80	0.022	20	50

		R-GCGCTCTATGCTCATCAACC					(100fg)	
	afaC[247]	F-GCGCTATGTGGTGCAGAGTA R-AAAACCGGTATTCACCAGGA	185	65–60	86	0.026	20 (100fg)	53
E. faecalis	16S rRNA[248]	F-CCGAGTGCTTGCACTCAATTGG R-CTCTTATGCCATGCGGCATAAAC	137	60	80	0.017	3 (10 fg)	87
S. gallolyticus	*sodA*[12]	F-CAATGACAATTCACCATGA R-TTGGTGCTTTTCCTTGTG	408	60–50	83	0.016	5 (20fg)	100
ETBF	*Bft1*[249]	F-GACGGTGTATGTGATTTGTCTGAGAGA R-ATCCCTAAGATTTTATTATCCCAAGTA	294	65-55	78	0.020	4 (20fg)	70
Fusobacterium	16S rRNA[250]	F-CGGGTGAGTAACGCGTAAAG R1-GCCGTGTCTCAGTCCCCT	228	60	85	0.026	2 (5fg)	75
		R2-GCATTCGTTTCCAAATGTTGTCC	61	60	83	0.020	2 (5fg)	79
FFPE QC	*COX1*	F-TATGGCGTTTCCCCGCATAA R-GCGAGCAGGAGTAGGAGAGA	69	57	98		N/A	N/A

ND: not determined; [#]The LOD for the specific detection of intimin theta could not be accurately determined since an intimin theta+ control strain was not available at the time; *LOD: limit of detection (bacterial copies) where positive identification was made in at least 50% of replicates.

The following reagents were obtained through the NIH Biodefense and Emerging Infections Research Resources Repository, (NIAID, NIH) as part of the Human Microbiome Project: *Streptococcus gallolyticus* subsp. *gallolyticus*, Strain TX20005, HM-272D; Genomic DNA from *Bacteroides fragilis*, Strain 3_1_12, HM-20D, Genomic DNA from *Clostridium difficile*, Strain NAP07 (CDC#2007054), HM-88D; Genomic DNA from *Enterococcus faecalis*, Strain HH22, HM-200D; Genomic DNA from *Escherichia coli*, Strain B171, NR-9297; and Genomic DNA from *Fusobacterium nucleatum* subsp. *polymorphum*, Strain F0401. ETBF genomic DNA (ATCC43858) was kindly provided by Dr Annalisa Pantosti from the Istituto Superiore di Sanità, Italy. DNA from enterohemorrhagic *E. coli* (to confirm EPEC identity) was kindly supplied by Dr. Anthony Smith at the National Institute for Communicable Diseases, South Africa. DNA from AIEC (strains HM358, HM229 and HM334) was kindly provided by Dr. Barry Campbell from the University of Liverpool, UK.

qPCR amplification conditions

Experiments were performed in triplicate on a Roche LightCycler® 480 Real-Time PCR System in 96-well format, using 50 ng patient DNA per well. Separate assays were performed for each bacterial gene detected; the cycling conditions are specified in Appendix A (Table 3). EPEC (*eaeA*, *bfpA* and *stx1* and *stx2*), ETBF and *S. gallolyticus*, were each detected in 20 µl reactions using SensiFAST SYBR No-ROX Kit (Bioline); AIEC, Fusobacterium and *E. faecalis* were each detected in 25 µl reactions using Maxima SYBR green qPCR Master Mix (Thermo Scientific). In order to increase specificity, it was necessary in some cases to perform touchdown PCR, whereby the annealing temperature is lowered in a stepwise manner to discourage amplification of off-targets during the first 10 cycles of PCR[251,252]; touchdown qPCR was performed for detection of EPEC (*bfpA* and *eaeA*), *S. gallolyticus*, ETBF, EHEC (*stx1* and *stx2*) and AIEC (*afaC*).

qPCR quantification

For each qPCR assay, absolute quantification was performed using a standard curve constructed from serially diluted genomic DNA for each of the positive control strains. The concentration of bacterial DNA found was expressed in terms of genome copies by calculating the weight of one genome copy for each species as used by Dolezel et al.[253]: DNA content (pg) = genome size (bp)/(0.978 x 10^9).

For example, Fusobacterium have an estimated genome size of 2.2 Mb and since one picogram of DNA equals approximately 978 Mb, a single Fusobacterium genome weighs approximately 2.25 fg (2.2Mb/978Mb/pg) and therefore 1 ng of DNA from Fusobacterium equates to about 445k copies (1000pg/(2.2Mb/978Mb)) of the bacterium (Table 3).

Table 3: Estimates of bacterial genome copies per nanogram of bacterial DNA.

	Strain, genome size	**Estimated bacterial copies/ng bacterial DNA.**
EPEC (*eaeA/bfp*)	E2348/69, 4.97 Mb	2 x 10^5
ETBF (*bft*)	3_1_12, 5.49 Mb	1.8 x 10^5
E. faecalis (16s rRNA)	V583, 3.34 Mb	3 x 10^5
Fusobacterium (16s rRNA)	NA, 2.2 Mb	4.5 x 10^5
AIEC (*afaC*)	LF82, 4.88 Mb[210]	2 x 10^5
AIEC (*ClB*)	LF82, 4.88 Mb	2 x 10^5
S. gallolyticus (*sodA*)	UCN34, 2.35[254]	4 x 10^5
EHEC (*stx1*)	O157:H7, 5.6	1.8 x 10^5
EHEC (*stx2*)	O157:H7, 5.6	1.8 x 10^5

In the case of AIEC strains, genome size may vary substantially between strains, since these strains are classified according to phenotypic traits and not sequence similarity. We opted to use the prototypical LF82 AIEC strain, which has a genome size of 4.88 for quantification.

Positive control standards were spiked with the same amount of human genomic DNA (extracted from uninfected human cell cultures) used in the patient sample reactions (50ng DNA). The LOD was defined as the lowest concentration at which a positive result (correct meltcurve) could be obtained in at least 50% of replicates (Table 1). For all assays except *ClB* and *afaC* at least 70% of replicates were positive at the relevant LOD. In cases where results were inconsistent (1/3 replicates positive), samples were retested and taken as positive if a positive meltcurve was obtained in both runs (the results were then averaged across the two runs to obtain quantitative data). Negative controls were included in each assay.

qPCR quantification in FFPE samples

We first evaluated the quality of DNA extracted from archival FFPE slides (which had been stored between 2 and 23 years) using three primer pairs designed to amplify 100bp, 200bp and 300bp amplicons of the *GAPDH* gene[255]. For most samples we detected either a very faint or no visible band at 100bp, whilst a 200bp amplicon could only be amplified in a few samples. On testing a shorter amplicon (69bp) of the *COX1* gene (which we found to be stably expressed in our cohort and is therefore assumed to have no significant differences in copy number between samples), all samples could be amplified by qPCR; the difference in cycle threshold (Ct) between the highest and lowest quality sample was 9.3. We therefore redesigned the reverse primers for bacterial detection to shorten the resulting amplicons to 60–70bp, and used the *COX1* results to account for degradation in our bacterial quantification. In the case of *eaeA*, two reverse primers were designed, one that detects intimin subtypes epsilon, gamma, zeta, alpha, pi, rho, beta, lambda, iota, kappa, eta, delta, xi, mu, kappa and jota; while the second was designed to specifically detect intimin theta (which was found in both EPEC-positive MSI-H samples from the fresh-frozen cohort). The efficiencies for the *COX1* qPCR was calculated using 5-fold serial dilutions constructed using a high- and low quality patient sample, as 1.96 and 2, respectively. A 'fold change' value was then calculated for *COX1* in each sample, using the ΔΔCt method and the mean Ct across 6 randomly selected DNA samples from the fresh-frozen (high-quality DNA) sample cohort was used as reference (the maximum ΔCt between fresh-frozen samples was 1.8). These

sample-specific 'fold change' values for *COX1* between FFPE samples to be tested and the reference set of fresh-frozen samples were used as a correction factor to adjust for DNA sample quality. A theoretical limit of detection was also calculated for each sample by multiplying the correction factor for each sample with the LOD that had previously determined for high quality DNA. A theoretical limit of detection was also calculated for each sample by multiplying the correction factor for each sample with the LOD that had previously determined for high quality DNA. After performing absolute quantification, the result was multiplied by the correction factor for each sample. The validity of this method was assessed by comparing Fusobacterium quantitation obtained from DNA extracted from fresh frozen samples to that of matched FFPE samples (which we had available for four patients); after removing a single outlier sample, the Pearson's correlation coefficient was 0.94, and the median fold change between matched fresh-frozen and FFPE samples was 1. FFPE samples that tested negative for Fusobacterium were set to 'NA' for downstream analysis, since the negative results could be due either to sample quality or to absence of the bacterium.

Statistical analyses

In order to assess quantitative differences between paired tumour and normal samples for each bacterium, the Wilcoxon signed rank test was applied to the subset of samples that had at least one positive sample in a pair (tumour or normal).

To assess the association between each bacterium and clinicopathological features, we compared a) samples with vs. without colonisation by a particular bacterium and b) samples with high vs. low/no-colonisation by a particular bacterium. Except for Fusobacterium, all other bacterial quantitative data had a large proportion of colonization-negative cases, which lead to non-normal distributions and unequal variances between groups. To address this issue, we converted the quantitative data to categorical data where for each bacterium, samples were categorised as 'no-infection', 'low-infection' or 'high-infection'. Quantitative data (copies/50ng) were log2 transformed and samples with no-colonisation were arbitrarily set to 1 before log2 transformation; the third quartile (calculated across infection-positive cases only) was used to discriminate low- and high-colonisation cases (see Figure 2 for categories).

Associations with clinicopathological features were examined using Fisher's exact test. Meanwhile, in the case of Fusobacterium (where the data was normally distributed), we used the Kruskal-Wallis test to evaluate differences between groups stratified by the clinicopathological parameter of interest. In order to investigate which CRC stages were significantly different in terms of the level of colonisation by Fusobacterium we used Dunn's test to compare individual stages in a pairwise manner. Results with an FDR $\leq$ 0.05 after applying multiple testing correction (Benjamini-Hochberg method) over all clinicopathological comparisons made for each species, were considered significant.

Results

Bacterial quantification

CRC-associated bacteria were quantified in adenocarcinoma and matched normal mucosal samples by qPCR, using a serial dilution of genomic DNA from each bacterium as standards. With the exception of *S. gallolyticus*, all species were present in tumour and normal samples at varying frequencies. While the association between *S. gallolyticus* bacteremia or infective endocarditis and CRC is well established[13], only one study, by Abdulamir et al.[12], has measured *S. gallolyticus* in CRC patients *without* a history of bacteremia or infective endocarditis. That study found that 4% of healthy controls but 48.7% and 32.7% of CRC patients with or without bacteremia were infected with *S. gallolyticus* in the relevant colonic tissue[12]. In contrast, we did not detect *S. gallolyticus* in any of our adenocarcinoma or matched normal mucosa samples using the same primers used by Abdulamir et al.[12]who used these primers for both for conventional PCR and qPCR. It should be noted that the levels reported in that study were typically very low and that none of our cohort had any reported history of bacteremia/bacterial-endocarditis. It is important in this regard that our qPCR assay was very sensitive (LOD = 5 copies/50ng DNA) and allowed for the detection of gram-positive bacteria such as *S. gallolyticus* by the addition of lysozyme to the homogenized human tissue prior to DNA extraction; we can therefore only speculate that the discrepancy between our results and those of Abdulamir et al. could be explained by a) differences in sample preparation, b) ethnic differences in the susceptibility to colonisation by *S. gallolyticus* or c) geographical differences in *S. gallolyticus* strains

found in Southern Africa that may have precluded detection of the bacterium in our cohort. Further investigation is therefore required to clarify this discrepancy.

Of the bacteria that we detected, Fusobacterium was by far the most common, occurring in 82% and 81% of paired tumour and normal samples, respectively, with 80% concurrent colonisation in paired samples. Fusobacterium was also the only bacterium that occurred at significantly higher levels in tumour compared to normal samples ($p = 6e-5$, Wilcoxon signed rank test), which is in agreement with previous studies[194,219,237]. The qPCR results are summarized in Figure 1 and Table 4.

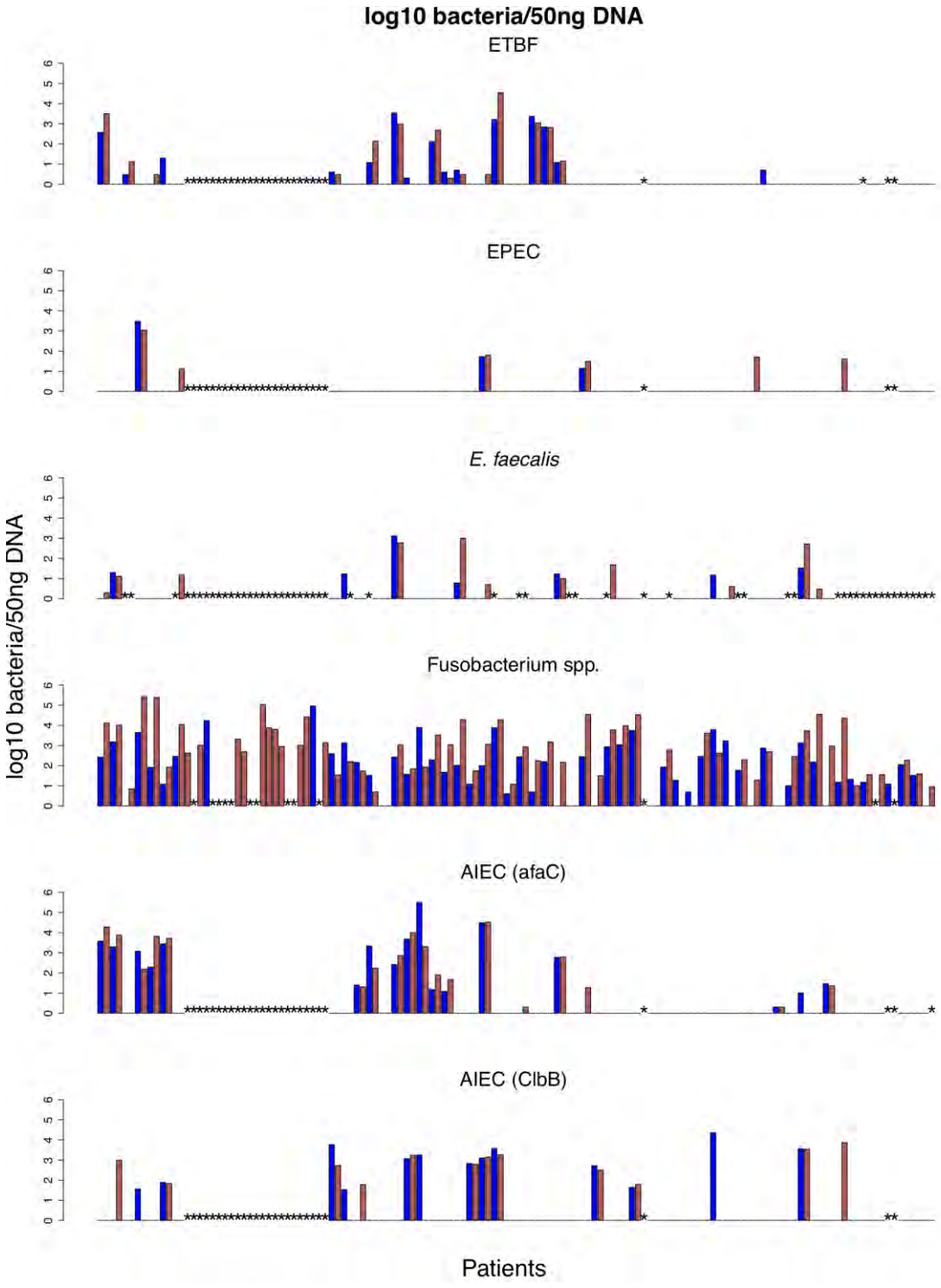


Figure 1: qPCR quantification of bacteria in paired patient samples, expressed as log10 bacteria/50ng of patient DNA. Each bar represents one samples (either tumour or normal) and the order of samples are the same for each bacterium. Red (tumour); blue (normal); *(Not determined, the majority of these, block of * on the left of figure were FFPE samples, where only Fusobacterium was successfully measured).

Table 4: Quantification of bacteria in colorectal cancer and adjacent normal tissues

Pathogen	**Colonisation rate T (%)**	**Colonisation rate N (%)**	**Concurrent colonisation in T & N (%)**
Fusobacterium spp.	58/71 (82%)	48/59 (81%)	43/54 (80%)
AIEC (afaC)	19/53 (36%)	17/54 (31%)	16/20 (80%)
AIEC (pks)	12/54 (22%)	13/55 (24%)	9/16 (56%)
E. faecalis	11/40 (28%)	7/38 (18%)	5/10 (50%)
ETBF	14/54 (26%)	15/53 (28%)	12/17 (71%)
EPEC	6/54 (11%)	3/54 (6%)	3/6 (50%)
S. gallolyticus	0/45 (0%)	0/45 (0%)	0/45 (0%)

T and N denote adenocarcinoma and adjacent normal mucosa, respectively. Rates of concurrent colonisation in T and N samples were calculated as a fraction of the number of patients who were infected in T and/or N with a particular bacterium.

In our cohort, ETBF was detected in 14/54 (26%) of colorectal adenocarcinomas and 15/53 (28%) of adjacent normal mucosa samples. This is largely consistent with previous studies on faecal samples, which have reported ETBF in ±12% of healthy controls[224,238], 27% of patients with diarrhea[224], and 38% of patients with CRC[238] with colonisation rates appearing to vary widely by geographical location[15]. Further, 71% of ETBF+ patients were infected in both adenocarcinoma and matched adjacent normal samples.

Although Balamurugan et al. demonstrated significantly higher levels of faecal *E. faecalis* in CRC patients compared to healthy controls[198], this is the first study to quantitatively measure *E. faecalis* in paired adenocarcinoma (28% *E. faecalis*-positive) and normal mucosa samples (18% *E. faecalis*-positive) with 50% of infected patients being infected in both adenocarcinoma and matched normal mucosa samples. However, no significant clinical associations with *E. faecalis* colonisation were found.

To investigate the presence of *E. coli* genes that are commonly found in AIEC in IBD and CRC patients, presence of *ClB* (part of the *pks* genomic island) and *afaC* (present in all operons of the afimbrial adhesin family) were evaluated in paired CRC samples. *pks*+ *E. coli* has previously been detected in 55–67% of CRC patients[213,256], compared to 8% of healthy controls[213]. By contrast, in our cohort, 22% of adenocarcinomas and 24% of adjacent normal mucosa samples were *pks*+, and 56% of *pks*+ patients were infected in both adenocarcinoma and matched normal mucosa samples. Meanwhile, *afaC* was detected in 36% and 31% of adenocarcinoma and normal mucosa samples, respectively, and found 80% of *afaC*+ patients were infected in both adenocarcinoma and matched normal mucosa samples. These rates are much lower than that found by Prorok-Hamon et al., who found 67% of CRC patients to be *afaC*+ compared to 17% of controls[213]. This discrepancy could be explained by our relatively high LOD for *afaC* and *pks* (Table 1). In contrast to Buc et al.[256], who found *pks* to be more common in distal compared to the proximal colon, no significant association between the presence of *pks* and site of disease was found. Lastly, it should be noted that in addition to *pks* and *afaC* many other AIEC-related genes including cyclomodulins[256] and *lpfA*[213] have been noted for their relevance to CRC.

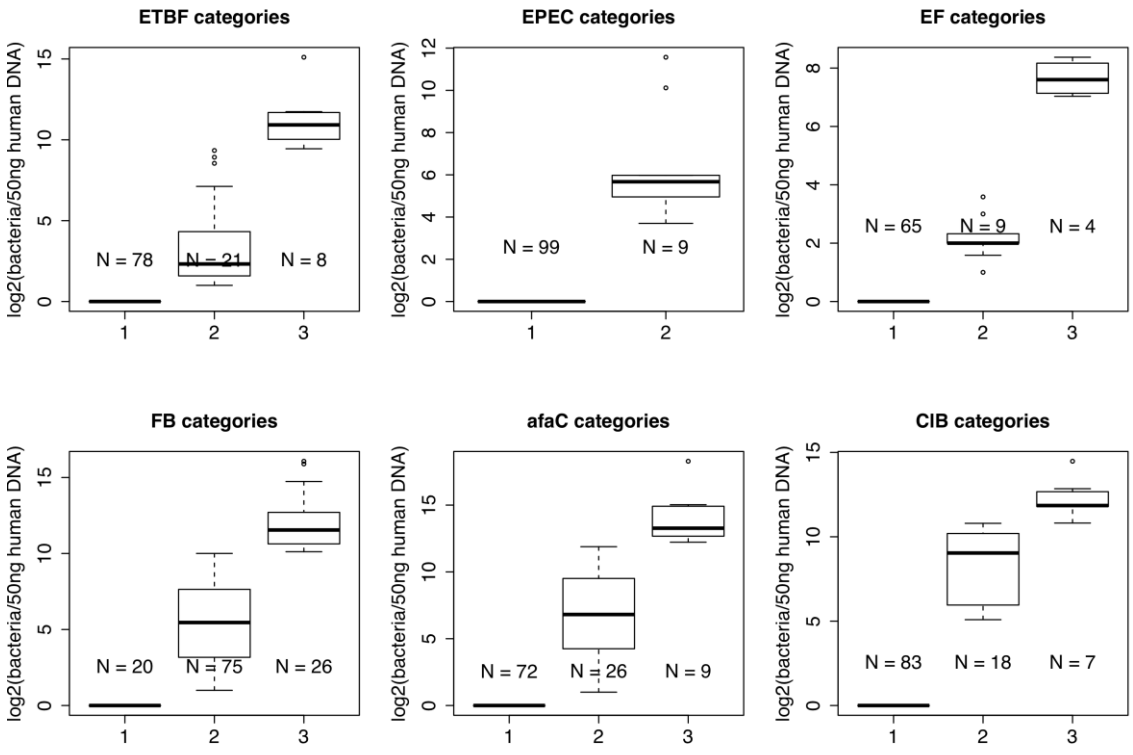


Figure 2: Colonisation levels for each bacterium/gene were categorized using the third quartile taken across infection-positive samples as a cutoff for high- or low-level infection. Categories: 1 (No infection), 2 (low infection, any positive result < third quartile), 3 (high infection, any positive results ≥ third quartile). In the case of EPEC, because there were so few EPEC-positive patients (N=6), samples were analysed as positive or negative only. EF: *E. faecalis*; FB: Fusobacterium.

ETBF and afaC-positive *E. coli* are significantly enriched in the colon compared to the rectum of CRC patients

As shown in Figure 3, the presence of ETBF and *afaC*-positive strains were significantly associated with the colon compared to the rectum in normal samples (FDR = 0.001 and 0.03, respectively), as well as in tumour samples in the case of ETBF (FDR = 0.002). No significant differences were found between the proximal and distal colon for any of the bacteria in this study.

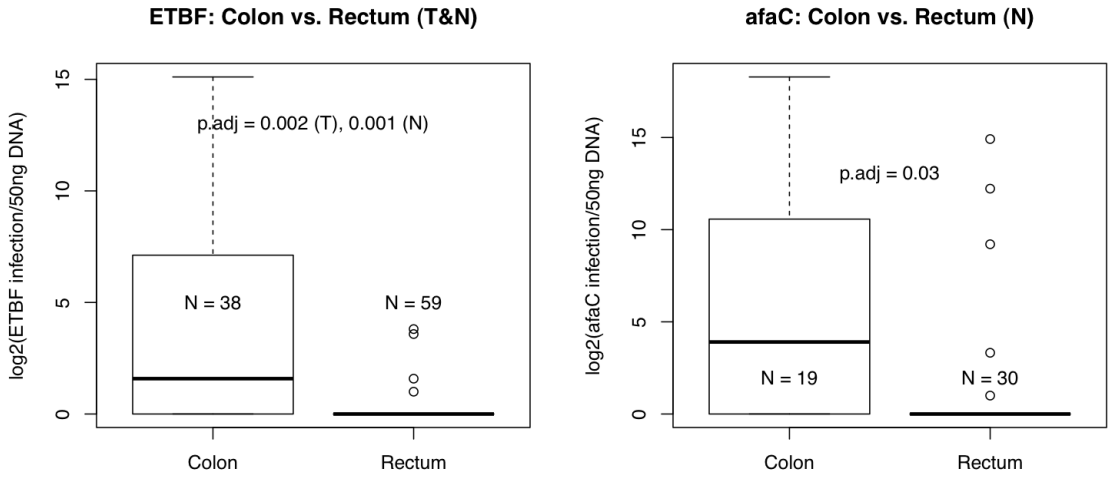


Figure 3: ETBF and afaC is significantly more prevalent in colon vs. rectal cancers. This applies to both tumour and normal tissue in the case of ETBF (FDR = 0.002, 0.001, respectively) and normal tissue only in the case of afaC (FDR = 0.03).

Colonisation by ETBF or high-level colonisation by Fusobacterium are associated with late-stage CRC

As shown in Figure 4, the presence of ETBF was significantly associated with stage of disease (Fisher's exact FDR=0.04 and 0.002 for normal and tumour samples, respectively). Similarly, in the case of Fusobacterium, late stage (III/IV) tumour samples were significantly associated with high-level colonisation by Fusobacterium (Kruskal-Wallis, FDR=0.03). In order to investigate which CRC stages were significantly different in terms of the level of colonisation by Fusobacterium we used Dunn's test to compare individual stages in a pairwise manner. Fusobacterium levels were significantly higher in stage III CRCs compared to stage I or II CRCs (p=0.002 for both comparisons). For ETBF, for which we found a difference in the presence or absence of ETBF between stages, individual stages were compared in a pairwise manner using Fisher's exact test. ETBF was found more frequently in stage III or IV CRCs compared to stage I or II CRCs (stage I vs. IV p=0.01; stage II vs. stage IV p=0.01; stage II vs. stage III p=0.003) as well as in the normal mucosa of stage IV CRCs compared to stage I CRCs (stage I vs. IV p=0.01; stage II vs. IV p=0.01).

High-level colonisation by Fusobacterium also seems to correlate with chronic inflammation in CRC. For example, McCoy et al. found a significant positive correlation between Fusobacterium abundance and local inflammation in adenoma cases[195] whilst we found that there is a trend towards high-level colonisation by Fusobacterium in patients with noted inflammation in normal tissue (Kruskal-Wallis, p = 0.01, FDR = 0.07, Figure 5), and tumour tissue (Kruskal-Wallis, p = 0.18, FDR = 0.2). We also found a positive association between high levels of colonisation by Fusobacterium and *pks*-positve *E. coli* in normal tissue (Fisher's exact, FDR = 0.007) or EPEC in tumour tissue (Fisher's exact, p = 0.08, FDR = 0.2). These data suggest that certain individuals may be more susceptible to infection, irrespective of the bacterial species. The finding that invasive strains of Fusobacterium are more prevalent in inflamed tissue in IBD, therefore suggests a possible inflammation-driven role of Fusobacterium in CRC development that warrants further investigation.

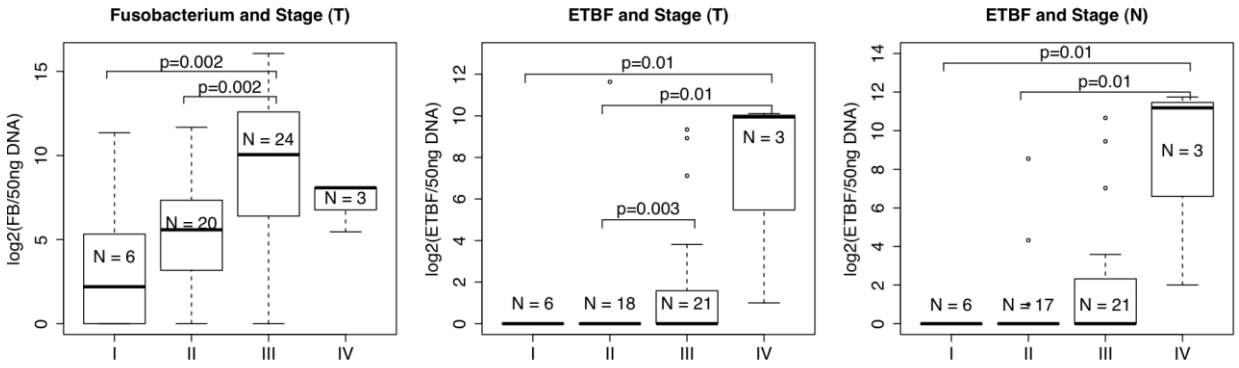


Figure 4: **ETBF and Fusobacterium are found at significantly higher levels in late stage (III/IV) cancers**. For Fusobacterium, individual stages were compared in a pairwise manner using Dunn's test. For ETBF, individual stages were compared in a pairwise manner using Fisher's exact test. Fusobacterium is found at significantly higher levels in stage III CRCs compared to stage I or II CRCs. ETBF is found more frequently in stage III or IV CRCs compared to stage I or II CRCs; and in the corresponding normal mucosa of stage IV CRCs compared to stage I CRCs.

Further clinical associations with high-level Fusobacterium colonisation

A significant relationship between high-level colonisation by Fusobacterium and MSI-H, compared to samples that were MSS or MSI-L (Kruskal-Wallis, FDR=0.05) was found, Figure 5. Furthermore, a significant increase in Fusobacterium levels in CRCs of younger patients (< 60 years), (Kruskal-Wallis, FDR=0.03) was noted, with 31% vs. 11% of patients under or over the age of 60 falling into the Fusobacterium-high group of colonisation (Figure 5). In normal samples, there was a trend towards high-level colonisation in males compared to females (Kruskal-Wallis, p=0.09, Figure 5).

In order to objectively assess the levels of colonisation by Fusobacterium between different ethnic groups, age-matched subsets of the data were used to account for the significant difference in patient age by ethnicity (ANOVA p=7.2e-6); across all patients, the mean age of black patients was 36, that of mixed ancestry patients was 58, and that of white patients was 77. Two age- and gender-matched comparisons were therefore performed: a) black patients (mean age=35, N=6) vs. mixed ancestry patients under the age of 50 (mean age=42, N=10) and b) caucasian patients (mean age=77, N=8) vs. mixed ancestry patients over the age of 70 (mean age=72, N=19). Fusobacterium was found at significantly higher levels in black patients compared to their age-matched mixed ancestry counterparts in adjacent normal samples, (Kruskal-Wallis, p=0.03, Figure 5), but not in tumour samples (Kruskal-Wallis, p=0.6). No significant differences were found between age-matched caucasian and mixed ancestry patients in terms of Fusobacterium colonisation levels.

Finally, the Fusobacterium-high group was also significantly associated with the presence of *pks*-positive *E. coli* in normal samples (Fisher's exact, FDR=0.01) and two of the three EPEC+ tumours were also infected with Fusobacterium-high (Fisher's exact, p=0.08, FDR=0.2), Figure 5.

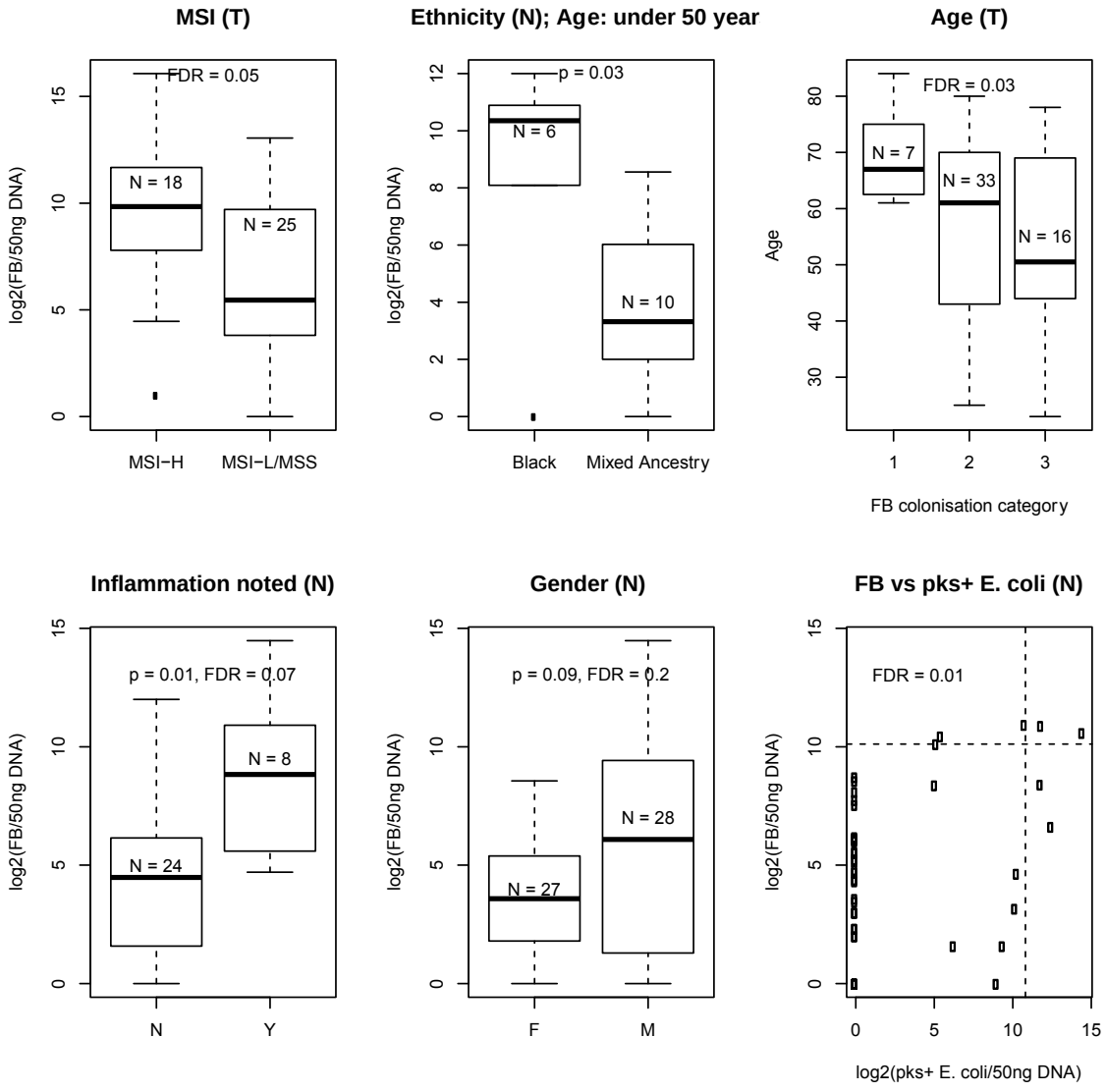


Figure 5: **Fusobacterium clinicopathological associations.** High-level colonisation by Fusobacterium is significantly more prevalent in younger patients, males and patients of black ethnicity. Due to the disproportionately high number of young, black patients seen in our cohort the relationship between ethnicity and levels of colonisation by Fusobacterium was assessed using the subset of patients ≤ 50 years. A borderline significant relationship was seen between high-level colonisation by Fusobacterium and MSI-H compared to MSS/MSI-L (In our cohort three MSI-L cases were included with the MSS cohort). The vertical and horizontal dotted lines in the bottom right figure represent the cutoff for high-level colonisation by *pks+ E. coli* and Fusobacterium, respectively (see methods for further detail). FB: Fusobacterium; T: CRC tumour tissue; N: adjacent normal mucosa; F: Female; M: Male; |nflammation noted: yes (Y), no (N).

EPEC detection and characterisation

In the fresh-frozen cohort, *eaeA* was detected in 11% and 6% of tumour and normal samples, respectively, with 50% concurrent colonisation in paired samples.

Colonisation levels varied from ±13–3037 bacteria/50ng of DNA extracted. The *eaeA* gene is found in EPEC and in enterohemorrhagic *E. coli* (EHEC). In our cohort, all *eaeA*-positive cases (N = 6) lacked the *bfpA* gene, which is present in typical EPEC, but not atypical EPEC (aEPEC) or EHEC. All *eaeA*+ samples were then identified as aEPEC and not EHEC due to the absence of *stx1* (present in EHEC, but not aEPEC). aEPEC has not been previously reported in association with CRC, although the EPEC detected in FFPE CRC samples by Maddocks et al.[9] could be aEPEC since they only profiled *eaeA*, and not *bfpA* or *stx1*.

No significant clinical associations were found for EPEC—this is not surprising given the small number of EPEC positive patients (N=6). However, of the six patients infected with EPEC, 67% (2/3) of sporadic MSI-H cases (fresh-frozen cohort), and only 9% (2/22) of MSS were EPEC-positive; the remaining two EPEC-infected patients were of unknown MSI status. Therefore, similar to Fusobacterium, there seems to be a trend towards EPEC- colonisation in sporadic MSI-H patients. Furthermore, in light of the effect of intimin subtype (of which there are currently 27 known variants[257]) on tissue tropism[258–260] we sequenced the 150bp amplicon amplified during intimin detection, which is located in the variable region of intimin and identified intimin theta exclusively in the two EPEC-positive MSI-H cases and in one case with unknown MSI-status. In the remaining EPEC-positive samples, intimin subtype could not be conclusively identified based on the 150bp product, but produced equal BLAST scores for the intimin subtypes: zeta 2&3, alpha 2, pi, iota 1, delta, beta 2, epsilon 2&8, jota and lambda in all of the remaining samples. In samples with concurrent colonisation in paired samples, the 150bp product sequences were identical. Intimin sequencing results are summarized in Table 5.

Table 5: Summary of BLAST search query to identify intimin subtypes. Samples highlighted in bold were used to determine the effect of FFPE fixation on the ability to detect EPEC by qPCR.

Sample	Highest scoring BLAST hits	% identity	Patient MSI status (T)	*MLH1* hypermethylated	Paired T&N sequences identical
44T	theta	100	MSI-H	Y	Y

44N	theta	99	MSI-H	N	Y
63T	theta	100	MSI-H	Y	NA
34N	zeta 2&3, alpha 2, pi, iota 1, delta, beta 2, epsilon 2&8, jota, lambda	100	MSS	N	Y
34T	zeta 2&3, alpha 2, pi, iota 1, delta, beta 2, epsilon 2&8, jota, lambda	100	MSS	N	Y
22T	theta, gamma	98	ND	ND	NA
45T	zeta 2&3, alpha 2, pi, iota 1, delta, beta 2, epsilon 2& 8, jota, lambda	97	ND	ND	NA
29N	zeta, alpha 2, pi, iota 1, delta, beta 2, epsilon, jota, lambda	100	MSS	ND	Y
29T	zeta 2&3, alpha 2, pi, iota 1, delta, beta 2, epsilon2& 8, jota, lambda	100	MSS	ND	Y

ND: not determined; NA: not applicable.

Our finding that intimin theta was exclusively identified in MSI-H EPEC positive cases (both located in the caecum), and in one case of unknown MSI status (located in the rectum) is interesting. Moreover, the two MSI-H patients infected with intimin-theta aEPEC were also the only two patients (with available MSI data) where *MLH1* was hypermethylated, as determined by methylation-specific qPCR. Both these patients were also infected with high levels of Fusobacterium (2730 and 68700 copies/50ng in tumour samples). Our data support the finding by Maddocks et al. that EPEC may decrease MMR functionality, but the implied mechanisms thereof appear to contrast: Maddocks et al. demonstrated *in vitro* EPEC-induced depletion of the mismatch repair proteins occurring at the protein level, despite an apparent increase in *MLH1* and *MSH2* mRNA following infection of HT29 cells with EPEC (strain E2348/69)[10]. Maddocks et al. concluded that EPEC-induced depletion of MLH1 and MSH2 proteins was dependent on mitochondrial targeting of the EPEC effector protein EspF and that this

depletion significantly increased the mutational frequency of infected cells[10]. On the other hand, we identified decreased MMR functionality, which is based on epigenetic silencing of *mlh1* in intimin theta+ aEPEC+ samples. Further work therefore seems needed to reconcile the apparently differing molecular origins of MLH1 depletion suggested by the cell-line-based studies of Maddocks et al. and our studies on clinical samples.

Lastly, although we did not sequence the entire intimin gene, the 150bp amplified sequences were consistently identical within but not between patients, suggesting that strains isolated from tumour or normal biopsies from a given patient are identical, in agreement with the findings by Martin et al. concerning *E. coli* strains in paired CRC samples[242].

Next, given the reported relationship between EPEC and MSI *in vitro*[9,10] as well as the relationship between intimin theta+ aEPEC and MSI seen here, we sourced 18 additional MSI-H samples from archival FFPE samples. However, none of which tested positive for EPEC, but because the median level of EPEC colonisation across EPEC-positive samples from the fresh-frozen cohort was relatively low (51 copies/50ng DNA), we investigated whether the level of degradation in the FFPE samples precluded detection in these samples. To this extent we compared the qPCR results from fresh-frozen (150bp amplicon) and matched archival FFPE samples (70bp amplicon) for three EPEC-positive patients (5 EPEC-positive T or N samples). EPEC could only be detected in one of the five matched FFPE samples—the sample that displayed the highest level of colonisation (3037 copies/50ng) in the fresh frozen tissue. Further, the median estimated LOD for the FFPE samples was 191 copies/50ng DNA (see Methods for further details), which is higher than the median level detected in fresh frozen samples (51 copies/50ng DNA). We therefore conclude that if EPEC were present in the MSI-H FFPE samples at levels similar to that seen in fresh-frozen samples, the level of degradation in the FFPE samples would have precluded detection of EPEC, even when attempting to amplify a 70bp amplicon.

Discussion

By quantifying multiple CRC-associated bacteria in one cohort, we uncovered inter- and intra-individual patterns of colonisation not previously recognised. We further identified significant associations with clinicopathological features including MSI-H (Fusobacterium), stage of disease (ETBF and Fusobacterium), tumour location (ETBF and *afaC*-positive *E. coli*), age (Fusobacterium), as well as a positive association between Fusobacterium and *pks*-positive strains.

Notably, the finding that late stage (III) tumour samples were significantly associated with high-level colonisation by Fusobacterium is consistent with previous studies demonstrating a positive association between high-level colonisation by Fusobacterium and regional lymph node metastases[194,219]. Bonnet et al. found a similar trend between cyclomodulin-positive *E. coli*, and stage III/IV colon cancers, which we however did not observe here[261]. Tumour tissue provides a nutrient-rich surface that is not protected by an intact mucosal layer, and the tumour-homing activity of certain bacteria is well documented[159]; but this does not necessarily imply oncogenic potential. However, in addition to the enrichment of Fusobacterium in tumour vs. normal tissues and in late stage CRCs, Fusobacterium spp. are also enriched in irritable bowel disease (IBD) patients (who have a 2–3 fold increased risk of developing CRC)[114] compared to healthy controls. Interestingly, Fusobacterium spp. isolated from inflamed tissue in IBD patients were significantly more invasive in a subsequent *in vitro* assay compared to non-inflamed tissue from IBD patients or healthy controls[262], possibly suggesting an active role for Fusobacterium in gastrointestinal diseases.

Tahara et al. have previously observed an association between high-level colonisation by Fusobacterium and MSI-H, *MLH1* methylation as well as the CpG island methylator phenotype (CIMP)[219] suggesting that Fusobacterium might promote MSI by inducing *MLH1* hypermethylation. Importantly, the association between MSI and Fusobacterium observed in our study was independent of the origin of MSI in our cohort, with 4/8 HNPCC adenocarcinoma samples falling into the Fusobacterium-high group of infection. HNPCC requires inactivation of both alleles of the affected mismatch-repair gene and it is tempting to speculate that Fusobacterium precipitates loss of the wild-type

allele through methylation. However, the role of aberrant methylation in the aetiopathogenesis of HNPCC remains questionable: Kaz et al. found promoter methylation of *MLH1* in 53% of HNPCC adenomas[263], compared to only 4% of sporadic adenomas, whilst Speake et al. found 40% and 25% of hyperplastic polyps of sporadic or HNPCC origin to be CIMP-H[264]. However, LOH or gene conversion are the most frequent mechanism of inactivation of the wild type *MLH1* allele in HNPCC tumours[265–269]. Further, HNPCC and sporadic MSI-H cancers have distinct histological and molecular features: While, both cancer types display lymphocytic infiltration, mucin secretion and poor differentiation[270], HNPCCs tends to originate from classical adenomas compared to sessile-serrated adenomas in the case of MSI-H CRCs[271] whilst on a molecular level, HNPCCs are strongly associated with mutations in *APC* or ß-catenin and/or *KRAS*[270,271], while MSI-H sporadic CRCs instead exhibit *BRAF* mutations, which are present in CRC precursor lesions[271]. Therefore, while it is tempting to speculate that Fusobacterium might cause MSI (and thereby CRC), it seems more likely that Fusobacterium preferentially flourishes in MSI-H compared to MSS cancers, perhaps due to the altered glycosylation profile in MSI-H cancers[272], that could facilitate adherence of certain bacteria[273]. Additionally, *F. nucleatum* infection has been shown to stimulate proliferation in CRC- but not in non-neoplastic-cell lines[196] and Fusobacterium stimulates cellular proliferation following an initial oncogenic hit (affecting a component of the WNT signaling pathway) in mice[274] and in CRC cell lines[196]. Taken together, it therefore seems most likely that Fusobacterium is not oncogenic itself, but may contribute to tumourigenesis by promoting inflammation and cancer cell proliferation.

It has long been appreciated that certain individuals are more susceptible to aberrant pathogenic colonisation of the gut epithelium, which may be accompanied by chronic inflammation, for example in patients with IBD. However, our finding that colonisation by certain bacteria are significantly associated with clinicopathological features in CRC–including the stage and site of disease–is new and might be linked to differential susceptibilities in relation to clinical features, such as age and ethnicity; these associations do not necessarily imply oncogenicity since many of the CRC-associated bacteria investigated in this study are asymptomatically present in a significant

percentage of the population [224,238]. One might therefore expect a pathogenic trend similar to that of *H. pylori* where genetic, environmental and strain-specific risk modifiers govern susceptibility to bacterially-mediated oncogenesis in the colon and where only a small fraction of individuals infected with the bacterium will eventually develop cancer. Evaluating the distribution of bacteria in relation to ethnicity, lifestyle- and clinicopathological factors is the first step in evaluating host-susceptibility to infection and putative bacterially-mediated oncogenic mechanisms. Furthermore, bacterial abundance is not the only factor that may be correlated with clinicopathological features since low-abundant bacteria may exert a significant effect on the host through the secretion of toxins at high levels. For example, Dutilh *et al*. showed that Enterobacterial toxins were among the most highly expressed in metatranscriptomic sequencing data from CRC paired tumour and normal tissues [275], including toxins from *E. coli, Salmonella enterica* and *Shigella flexneri* [275]. Evaluating the presence of bacterial toxins with oncogenic potential at the transcriptional or proteomic level will thus provide an additional layer of information to unravel complex host-pathogen interactions with relevance to CRC in the future. Future studies should also be aimed at validating our findings in a larger cohort (particularly in MSI-H CRCs in the case of EPEC); and at profiling Fusobacterium at the species level, as well as other AIEC toxins implicated in CRC not examined here, such as *lpfA*.

Establishing causality for any of the bacteria examined here remains a challenge and would require rigorous investigation in animal models as well as large scale epidemiological data, as was used in establishing causality in the case of *H. pylori* and gastric cancer. However, by evaluating the distribution of bacteria in relation to ethnicity, lifestyle- and clinicopathological factors in a South African cohort, we have taken a first step towards this goal and we expect that our data will facilitates the development of targeted research questions for future studies.

Chapter 5: Quality assessment and data handling methods for Affymetrix Gene 1.0 ST arrays with variable RNA integrity

Abstract

One of the primary objectives of this thesis was to conduct gene expression analysis on paired CRC samples to investigate host-signaling pathways in the context of specific bacterial colonisation. However, clinical samples often have variable RNA integrity due to a range of factors from host factors (type of tissue/sample, disease state etc.) to sample collection and storage, which have important implications for downstream gene expression analyses. RNA and microarray quality assessment form an integral part of gene expression analysis and, although methods such as the RNA integrity number (RIN) algorithm reliably asses RNA integrity, the relevance of RNA integrity in gene expression analysis as well as analysis methods to accommodate the possible effects of degradation requires further investigation.

We investigated the relationship between RNA integrity and array quality on the commonly used Affymetrix Gene 1.0 ST array platform using reliable within-array and between-array quality assessment measures. The possibility of a transcript specific bias in the apparent effect of RNA degradation on the measured gene expression signal was evaluated after either excluding quality-flagged arrays or compensation for RNA degradation at different steps in the analysis.

Using probe-level and inter-array quality metrics to assess 34 Gene 1.0 ST arrays derived from historical, paired tumour and normal primary colorectal cancer samples, 7 arrays (20.6%), with a mean sample RIN of 3.2 (SD = 0.42), were flagged during array quality assessment while 10 arrays from samples with RINs < 7 passed quality assessment, including one sample with a RIN < 3. We detected a transcript length bias in RNA degradation in only 5.8% of annotated transcript clusters (p-value 0.05, FC ≥ |2|), with longer and shorter than average transcripts under- and overrepresented in quality-flagged samples respectively. Applying compensatory measures for RNA degradation performed at least as well as excluding quality-flagged arrays, as judged by hierarchical clustering, gene expression analysis and Ingenuity Pathway Analysis;

importantly, use of these compensatory measures had the significant benefit of enabling lower quality array data from irreplaceable clinical samples to be retained in downstream analyses.

Here, we demonstrate an effective array-quality assessment strategy, which will allow the user to recognize lower quality arrays that can be included in the analysis if appropriate measures are applied to account for known or unknown sources of variation, such as array quality- and batch- effects, by implementing ComBat or Surrogate Variable Analysis. This approach of quality control and analysis will be especially useful for clinical samples with variable and low RNA qualities, with RIN scores ≥ 2.

The results from this Chapter were published in BMC genomics in 2013[276] (Appendix D).

Introduction

RNA degradation is a common concern in gene expression analysis, especially for clinical samples where RNA degradation may occur before sample collection[277]. A wealth of archival material, either snap frozen or formalin fixed and paraffin embedded (FFPE), could potentially be used for gene expression analysis, given an appropriate method to evaluate and account for the effect of RNA degradation on the quality of downstream gene expression data. Methods such as the RNA integrity number (RIN) algorithm reliably assess RNA integrity by extracting features from the RNA electropherogram. The RIN algorithm was developed using learning tools to identify regions (features) indicative of RNA integrity in the electropherogram, which are then used to compile the RNA integrity number on a scale of 1 to 10. However, the relevance of RNA integrity in gene expression analysis, especially when there is large variability between samples, requires further investigation and validation on a platform specific basis. The impact of RNA integrity on gene expression analysis has been investigated on both qRT-PCR and certain microarray platforms[278–283]. Opitz et al. investigated the impact of RNA degradation on Agilent 44 k gene expression profiling by subjecting RNA from clinical biopsies to temperature-induced RNA degradation and comparing gene expression to the original, intact samples. Notably, less than 1% of genes were affected, even after substantial RNA degradation, where control and test samples had

RINs of 9 and 5 respectively. The affected transcripts were relatively shorter, had lower GC content, or had probes relatively closer to the 5' region of the gene compared to more robust genes[282]. Although the process of RNA degradation is not fully understood, both exonuclease and endonuclease activity is likely to play an important role[282]. Classical oligo-dT based cDNA synthesis, which starts at the poly-A tail, will most certainly be compromised by exonuclease activity. In contrast random priming does not rely on full length mRNA and therefore is in theory at least partially relieved from the affects of RNA degradation[282–285].

When using semi-degraded RNA for gene expression studies, reliable measures of array quality provide valuable information that can be used to guide downstream analysis. Microarray data quality may be defined in terms of accuracy (systematic bias between the true and measured value), precision (the uncertainty in replicated measures), specificity (the selective power of the measurement to respond only to the specific targets) and sensitivity (the expression range potentially covered by the measurement)[286]. Any attempt to utilise array quality results to guide downstream analysis should ideally take into account the possible effects of RNA degradation on sensitivity, specificity and accuracy. In previous work, Binder et al. proposed a single-array preprocessing method that allows correction for systematic biases such as RNA degradation by utilising information on the 3'/5'-amplification bias and the sample-specific calling rate[286]. Lassmann et al. proposed using a data adjustment method to allow comparative analysis of microarray datasets derived from fresh frozen vs. FFPE samples by centering the log intensities of each probe set independently to a mean of zero in both groups[284]. Chow et al. evaluated the suitability of different quality control and preprocessing strategies for use with partially degraded RNA samples on the Illumina DASL-based gene expression assay using mean inter-array correlation and multivariate distance matrix regression (MDMR) as a measure of success[287]. Unfortunately none of these studies are directly applicable to one of the most commonly used human transcriptomic microarray platforms, namely Affymetrix Gene 1.0 ST arrays, either because they do not use a random priming approach or because the design of the microarray platform differs substantially from Gene 1.0 ST arrays. We therefore identified two alternative approaches that might be used as compensatory methods:

Firstly, Johnson et al. developed an empirical Bayes algorithm, ComBat, to directly adjust for non-biological experimental variation. As the name implies, this method is most often used to adjust for batch effects i.e. when microarrays are processed on different dates[288]. Secondly, Leek et al. developed a method called Surrogate Variable Analysis (SVA), which examines the contribution of sources of signal due to unknown (surrogate) variables in high-dimensional data sets, which may confound the biological signal of interest[289]. The surrogate variables are constructed directly from the gene expression data where groups of genes that are affected by each source of variation are identified, factors are then estimated for each array which can be included in a linear model to adjust for unknown sources of noise e.g. RNA- or array-quality.

Here, we investigate the relationship between RNA integrity and array quality on Affymetrix Gene 1.0 ST arrays for 34 paired colorectal tumour and adjacent normal biopsies of highly variable RNA integrity. We assume that at a certain point on the RIN scale, RNA will be degraded to the extent where fragments are too small to analyse reliably and for the purpose of this analysis we arbitrarily select a RIN cutoff of 2. We describe the within- and between-array quality control measures and analysis methods that we found most relevant for gene expression analysis of samples with highly variable RINs on Affymetrix Gene 1.0 ST arrays. We then investigate the possibility of a transcript-length dependency in RNA degradation. Finally, we apply array quality information to either exclude quality-flagged arrays, to directly adjust the data using the ComBat algorithm, or to account for unknown sources of variation (such as RNA integrity or array quality) in the model fitting process using SVA. The data discussed, have been submitted to ArrayExpress, with accession number E-MEXP-3715.

Materials and Methods

Sample preparation and quality control

Frozen samples were transitioned to RNA®later -ICE (Ambion), an RNA stabilisation solution, using dry ice to prevent thawing of the tissue at any stage. RNA was extracted using a Dounce homogenizer and the AllPrep DNA/RNA/Protein kit (Qiagen) including DNAse treatment. RNAseZap (Ambion) was used to eliminate RNAse from the work surface, pipettes and glassware. RNA integrity assessment was conducted on an Agilent

Bioanalyser 2100.

Quantitative real-time PCR

From a biological perspective, we used the stability of expression of housekeeping genes to investigate the effect of RNA integrity on array- and qRT-PCR performance. Gene candidates were selected from those previously been specifically identified as good reference genes for colorectal cancer[290–294]. Expression stabilities were ranked using the Normfinder algorithm[295] and three genes were selected for use as reference genes. All primers except those for *B2M*[296] were designed using Primer-BLAST; sequences are shown in Table 4.

Table 4: Primers used for qRT-PCR analysis.

Test genes	F primers (5' - 3')	R primers (5' - 3')	Product (bp)
dpep1	GACAACTGGCTGGTGGACA	ACCACACGCTGCCCAAA	74
cldn1	GCTGTCATTGGGGGTGCGAT	GGCAACTAAAATAGCCAGACCTGC	54
Reference genes			
ubc	GGTCGCAGTTCTTGTTTGTGG	CACGAAGATCTGCATTGTCAAG	59
b2m	TGCTGTCTCCATGTTTGATGTATCT	TCTCTGCTCCCCACCTCTAAGT	86
atp5e	CTGGACTCAGCTACATCCGA	GCATCTCTCACTGCTTTTGCAC	55

Experiments were performed in triplicate on a Roche LightCycler® 480 Real-Time PCR System in 96-well format. Efficiency was determined for each primer pair using a two-fold dilution series across five points for five patient samples of varying RNA integrity. For each patient, tumour vs. normal fold change was determined based on the method of Antonov et al. whereby the Ct of the test gene is normalised by the geometric mean of multiple control genes[285]. Since our efficiencies were quite low in some cases, we adapted the Antonov et al. method to include primer efficiency as shown in the expression below:

$$\frac{e^{\Delta Ct(t)}}{\sqrt[n+1]{e_i^{\Delta Ct(i)} \times e_{i+1}^{\Delta Ct(i+1)} \times \ldots \times e_{i+n}^{\Delta Ct(i+n)}}}$$

where t represents the test gene, e represents efficiency and i represents the control gene(s).

Microarray analysis

Affymetrix HuGene 1.0 ST expression arrays

Thirty-four samples with A260/230 ratios of at least 1.6, RINs of at least 2 and no sign of genomic DNA contamination, were selected for microarray analysis. The samples were amplified from 200ng of total RNA in accordance with the Ambion® WT Expression assay kit and fragmented and end labeled in accordance with the Affymetrix® GeneChip® WT Terminal Labeling protocol. The prepared targets were hybridized overnight to Affymetrix Human Gene 1.0 ST arrays. Following hybridization, the arrays were washed and stained using the GeneChip Fluidics Station 450 and scanned using the GeneChip® Scanner 3000 7G. Arrays were processed in two batches - batch one had 10 arrays, and batch two 24. Individual patient pairs were not split across batches.

Microarray quality assessment and data analysis

Standard Affymetrix quality control was conducted using Expression Console® Software: The quality of cDNA preparation and array hybridisation was assessed using appropriate spike-in controls at each stage.

Raw array quality was investigated at the probe level by 1) the difference between the mean of the perfect match probes and the mean of the background probes for each array as well as 2) the coefficient of variation (CV) across all probes for each array. A threshold for the CV across probes was set as two standard deviations from the mean CV, where the mean was calculated from arrays with RINs > 6. The data were preprocessed in R using the Bioconductor packages frma[297], oligo[298], and the ComBat algorithm for batch correction[288]. Preprocessed data quality was assessed using the global normalised, unscaled standard error (GNUSE)[299]. The SE estimates are normalized such that for each probe set, the median standard error across all arrays is equal to 1. Since most genes are not expected to be differentially expressed, boxplots for each array should be centered around 1. Samples with a median GNUSE of greater than 1.25 were flagged for downstream analysis. This threshold is fairly arbitrary and has not been validated for the Gene 1.0 ST platform but roughly equates to having a precision that is on average 25% worse than the average Gene 1.0 ST array[299].

Five comparative methods for analysis of differential expression were individually applied to the preprocessed data: 1) The arrayWeights function in the Bioconductor package limma[229] was used to estimate array quality weights which were then included in the linear model fit; 2) Arrays that were flagged in array quality assessment were excluded from the analysis; 3) The ComBat algorithm for batch correction was applied to directly adjust the data according to quality, where arrays were divided into two categories according to the array quality assessment; 4) "Quality" and "batch" were included as a factors in the linear model together with disease status; 5) Surrogate variable analysis was applied to frma-processed data without any direct adjustment, the output from SVA being incorporated into the linear model fit[289].

To rank genes by evidence for differential expression, the eBayes function in limma was applied to compute moderated t-statistics, moderated F-statistic, and log-odds of differential expression by empirical Bayes shrink- age of the standard errors towards a common value[229]. Next, using the topTable function in limma, p-values were adjusted for multiple hypothesis testing using the Benjamini and Hochberg method[300]. Transcript clusters were annotated in R using the Bioconductor package hugene10sttranscriptcluster.db (Affymetrix Human Gene 1.0-ST Array Transcriptcluster Revision 8 annotation data, assembled using data from public repositories).

The subset of genes differentially affected by RNA quality was similarly obtained, now using array quality for grouping, instead of disease status. Genes with adjusted p-values ≤ 0.05 and FCs $\geq |2|$ were included in the analysis. Transcript length was obtained for all annotated transcript clusters using the Bioconductor package goseq[301]. Hierarchical clustering with average linkage and Euclidian distance as distance measure was performed in R using the hclust function.

For Ingenuity Pathway Analysis, genes that were found to be significantly differentially expressed for each method (adjusted p-value ≤ 0.01), were used as input for IPAs "Core Analysis". Here, statistically significant over-representation of our listed genes in a given process such as "colorectal tumour" or "infection of embryonic cell lines" is scored by p-value, using the right-tailed Fisher's Exact Test. In the case of upstream regulators, the predicted activation state and activation z-score is based on the direction

of fold change values for genes in the input dataset for which an experimentally observed causal relationship has been established. Performance was assessed using the top 10 functions in terms of p-values for each method while taking into account the relevance of the function to colorectal cancer.

The data analysis pipeline which can be divided into three phases: quality control; data adjustment and/or analyses; and performance evaluation, is summarized in Figure 1.

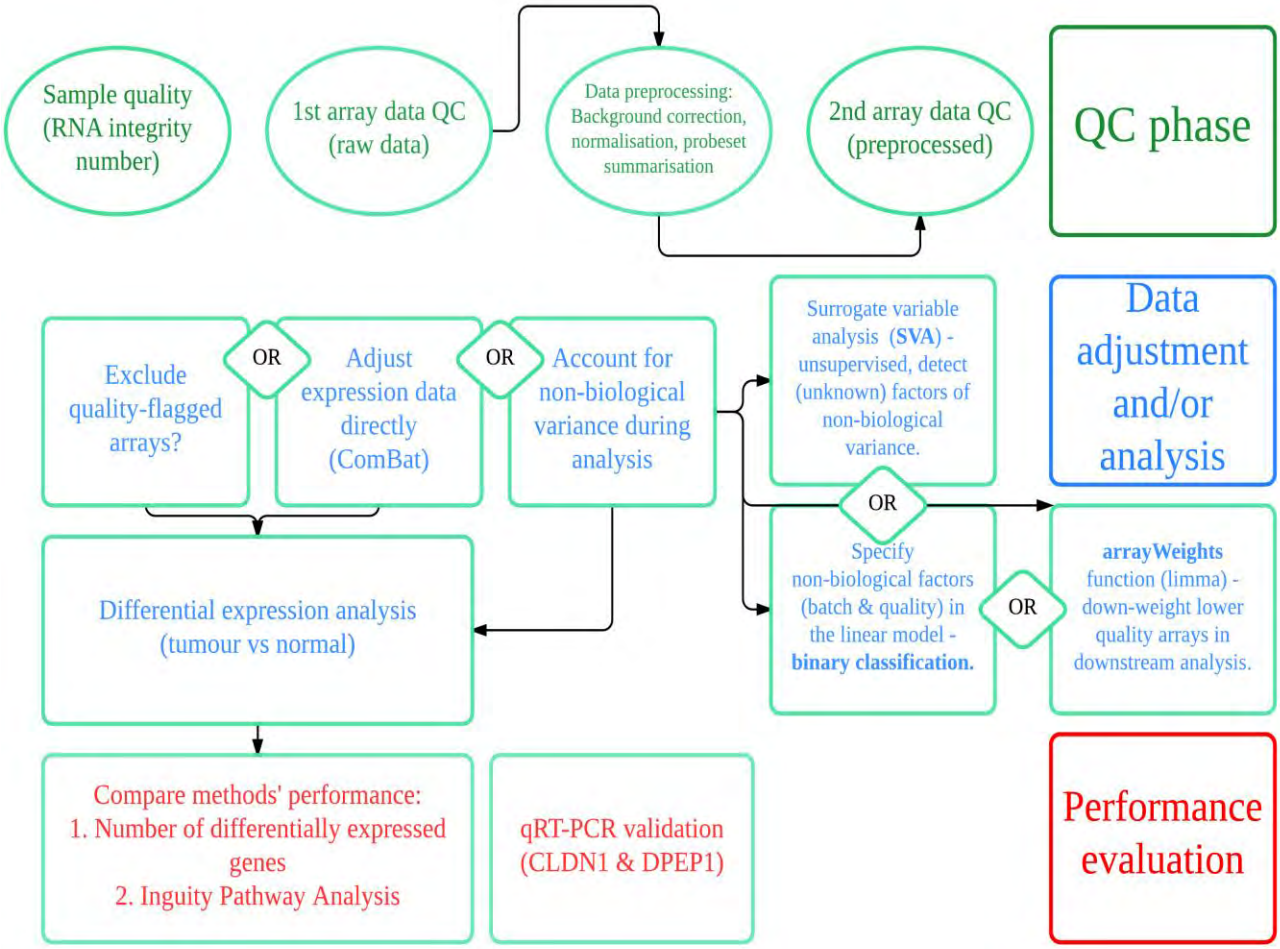


Figure 1: Summary of the full data analysis pipeline which consists of three phases: 1) Quality control (QC), 2) data adjustment and/or analysis and 3) performance evaluation.

Results

Array quality

We assessed array quality using within- and between-array measures—the former to assess raw data quality (Figures 2a & 2b), and the latter to assess the quality of an array relative to a large publically available collection of high quality Gene 1.0 ST arrays (Figure 2c). Raw array quality was investigated at the probe level by calculating the difference between the means of perfect match- and background-probes for each array as well as the coefficient of variation (CV) across all probes for each array. Preprocessed data quality was assessed using the global normalised, unscaled standard error (GNUSE)[299]. See Methods section for details.

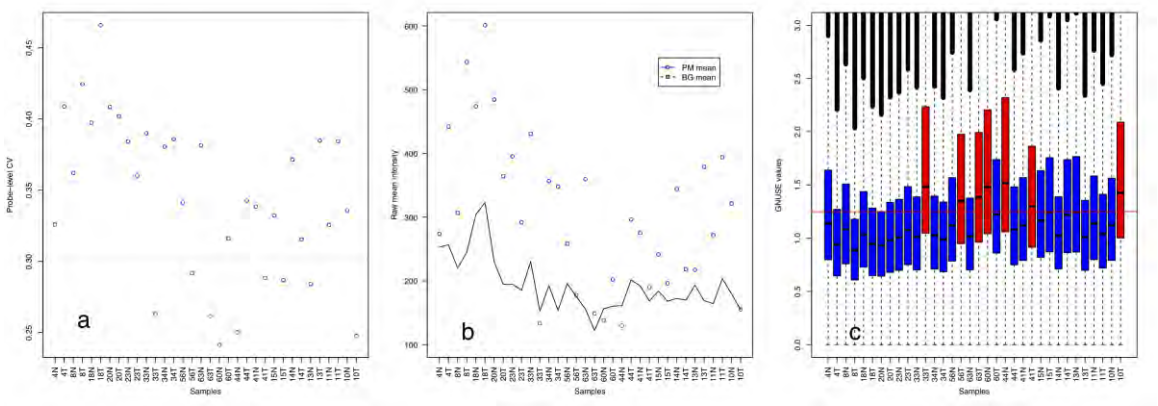


Figure 2: Array quality metrics. a) Raw coefficient of variation across all probes by sample, the red line represents our chosen threshold which is calculated as 2SD from the mean of CVs for arrays with RINs > 6. Samples labeled in red denote samples that were eventually flagged for downstream analysis because they failed ≥ two of the three quality measures b) Raw perfect match mean - background mean c) Global normalised unscaled errors (GNUSE) across probes for each array. Samples with a median GNUSE greater than 1.25 were flagged for downstream analysis.

The 34 RNA samples used in this study had mean RIN of 6.3 and a standard deviation of 2.0. Samples that failed all three measures of quality had RINs between 2 and 3.3 as summarised in Table 1. Samples were ranked by GNUSE median and we found a good concordance in terms of ranking between the different quality control metrics. Samples that failed at least two out of the three quality measures were flagged for downstream analysis, resulting in 7 out of 34 samples being flagged (mean RIN = 3.2; SD = 0.42, samples represented in red in Figure 2). Interestingly, for one sample with a RIN of 2.6, array quality was not compromised, judged by our quality measures. The possibility of

a RIN-independent RNA quality factor, such as chemical purity, was investigated by performing a two-tailed Student's t-test, comparing A260/230 ratios between quality-flagged and quality-passed sample groups but no significant association was found (p-value = 0.14).

Table 1: Array quality assessment summary.

Sample ID	RIN	RNA 260/23(	GNUSE	probe-level (	PM-BG	Array weight
44N	3	2.41	**fail (1)**	**fail (3)**	**fail (1)**	**0.22 (1)**
33T	2.8	2.08	**fail (2)**	**fail (5)**	**fail (2)**	**0.28 (2)**
60N	3.2	2.03	**fail (3)**	**fail (1)**	**fail (3)**	**0.42 (3)**
63T	3	2.2	**fail (4)**	**fail (4)**	**pass**	**0.59 (6)**
10T	3.2	2.18	**fail (5)**	**fail (2)**	**fail (4)**	**0.60 (7)**
56T	3.3	1.87	**fail (6)**	**fail (10)**	**fail (5)**	**0.42 (4)**
41T	4.2	2.21	**fail (7)**	**fail (9)**	**pass**	**0.78 (8)**
13N	4.6	2.24	pass	fail (7)	pass	0.82 (9)
15T	4.8	2.15	pass	fail (8)	pass	1.07 (15)
4N	2.6	1.62	pass	pass	pass	0.44 (5)
18N	7.1	1.66	pass	pass	pass	0.83 (10)
8T	8.5	2.16	pass	pass	pass	0.85 (11)
56N	6.5	1.94	pass	pass	pass	0.95 (12)
20T	7.4	1.6	pass	pass	pass	1.02 (13)
44T	6.9	1.72	pass	pass	pass	1.03 (14)
11T	8.6	2.16	pass	pass	pass	1.07 (16)
60T	6.4	1.64	pass	pass	pass	1.09 (17)
14T	6.4	1.76	pass	pass	pass	1.09 (18)
13T	8.3	2	pass	pass	pass	1.11 (19)
23T	7	2.17	pass	pass	pass	1.18 (20)
8N	7.1	2.22	pass	pass	pass	1.25 (21)
18T	7.4	1.85	pass	pass	pass	1.26 (22)
33N	8.1	1.82	pass	pass	pass	1.45 (23)
34T	8	2.25	pass	pass	pass	1.49 (24)
11N	6.8	1.94	pass	pass	pass	1.50 (25)
20N	7.3	2.11	pass	pass	pass	1.50 (26)
63N	7.5	2.13	pass	pass	pass	1.61 (27)
23N	8.4	2.02	pass	pass	pass	1.61 (28)
34N	8.3	2.21	pass	pass	pass	1.61 (29)
14N	8.1	2.36	pass	pass	pass	1.74 (30)
41N	5.4	2.07	pass	pass	pass	1.76 (31)
10N	7.3	1.78	pass	pass	pass	1.78 (32)
15N	6.9	2.16	pass	pass	pass	1.90 (33)
4T	8.4	2.25	pass	pass	pass	2.14 (34

Array performance is ranked for each measure with 1 considered the worst quality. Samples highlighted in bold were flagged for downstream analysis.

Transcript-dependent effects of RNA degradation on accuracy

To investigate a possible probe-positional intensity bias related to RNA integrity, we plotted the mean probe intensity from the 5'- to 3' end of the sequence using 4644/32321 (14.4%) of transcript clusters for Gene 1.0 ST arrays and 54130/54675 (99%) of probesets for HGU133-plus2 arrays. The number of probes per set varies for GeneST arrays, so we selected the largest group (N = 4664), which had exactly 25 probes/set. Interestingly, from the 4644 transcript clusters displayed in Figure 3, Gene ST 1.0 arrays, do not display the same probe-positional intensity bias typically seen in oligo-dT based arrays such as the HGU133-plus2 arrays.

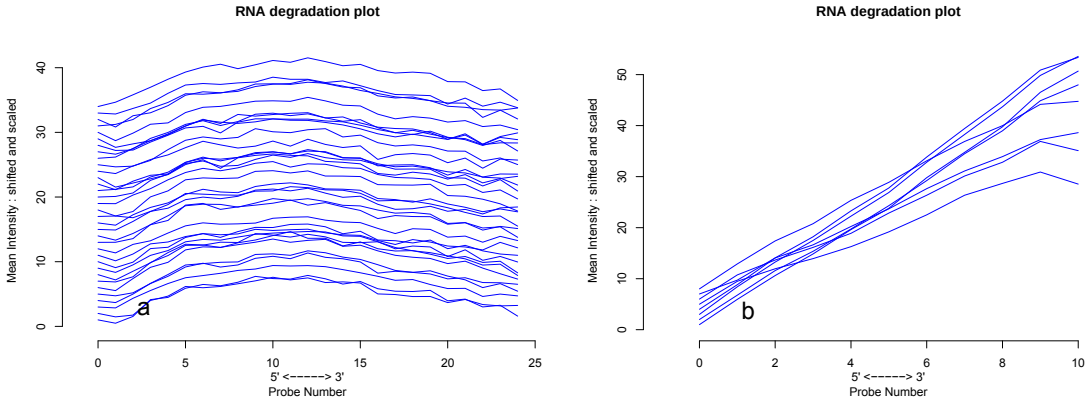


Figure 3: Mean probe intensity by probe position, where each line represents an array for a) Gene 1.0 ST arrays: transcript clusters with exactly 25 probes (N = 4644) and b) HGU133-plus2 arrays previously analysed with a subset of the cohort: probesets with exactly 11 probes per probeset.

We next investigated which genes were most affected in our quality-flagged category and identified 1994 out of 21943 annotated transcript clusters (with 1172 uniquely identified genes) that were significantly different (fold change $\geq |2|$, adjusted p-value ≤ 0.05) between the two quality categories previously discussed. Of the 1172 uniquely identified genes, 1032 and 140 showed decreased or increased intensity in the quality-flagged category respectively (Figure 4a). To investigate transcript characteristics in the genes most affected, we compared transcript lengths (taken as the median cDNA length for each gene) between the different groups. Compared to the unaffected genes, median cDNA lengths of genes that showed increased intensity were significantly shorter (p-value $< 2.2e-16$) while those with decreased intensity significantly longer (p-value $= 2.9e-9$) with regards to quality, judged using the Mann Whitney test (Figure 4b).

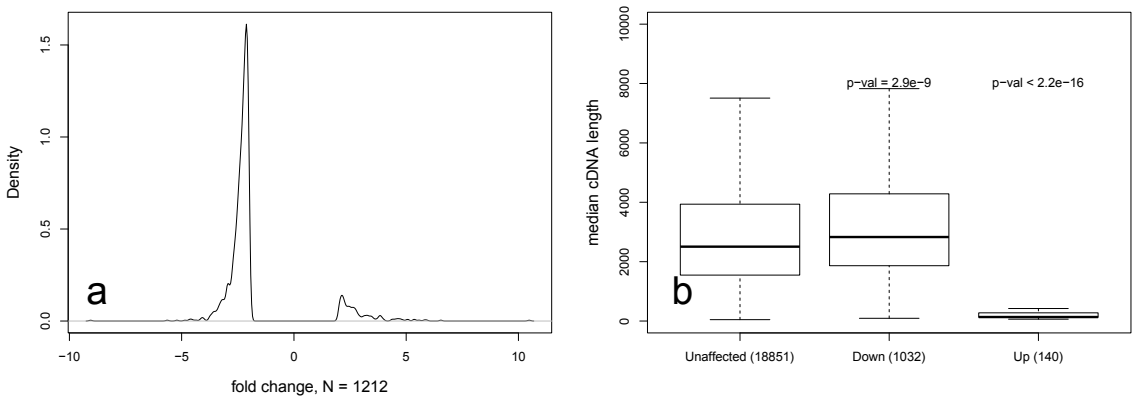


Figure 4: Characteristics of genes most affected by RNA degradation (adjusted p-value ≤ 0.01, |fold change| > 2) when comparing samples that either passed or were flagged during QC. a) Fold change distribution of annotated transcript clusters comparing samples that were flagged vs. samples that passed QC b) Gene lengths of uniquely identified genes. Expression signal significantly increased (Up) or decreased (Down) with respect to the 'Unaffected' group, judged using a Mann-Whitney test.

Quality dependent methods of data adjustment and analysis

After assigning samples to two categories according to array quality measures, we next assessed the performance of the five preprocessing and analysis methods. Broadly speaking, the data was either directly adjusted for quality effects using ComBat, or quality-flagged samples were excluded from the analysis, or possible quality effects were addressed by including known or unknown sources of non- biological variance in the linear model fit to assess differential expression.

The five methods of data preprocessing and analysis, further detailed in the Methods section, were: 1) Estimating array quality weights which were then included in the linear model fit; 2) Excluding quality-flagged arrays from the analysis; 3) Applying a batch correction algorithm, ComBat[288], to directly adjust the data according to quality, where arrays were divided into two categories according to the array quality assessment; 4) "Quality" and "batch" were included as a factors in the linear model together with disease status; 5) Possible unknown sources of non-biological variance, such as quality, was estimated by SVA, with the output incorporated into the linear model fit[289].

To assess the effect of using ComBat for direct data adjustment, hierarchical clustering using Euclidian distance was performed before and after direct adjustment (Figure 5).

We chose to use Euclidian distance based on research by Gibbons et al. who demonstrated that, for log-transformed expression data, using Euclidian distance is more appropriate than Pearson's correlation coefficients[302].

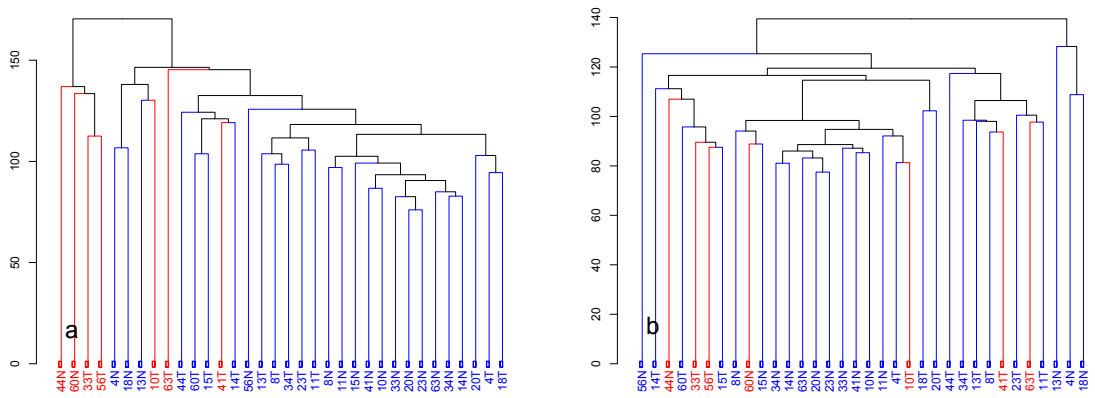


Figure 5: Samples' expression profiles were clustered using average linkage hierarchical clustering. The dissimilarity measure (height) used was 1- Pearson correlation of the log2-transformed expression values. a) Sample clustering after preprocessing. b) Sample clustering after preprocessing and correction for batch and quality using ComBat. Samples that were flagged during quality assessment are highlighted in red.

Before adjustment, samples that were flagged during quality assessment cluster closely together, irrespective of the disease status of the samples. After adjustment, the maximum distance between samples is greatly reduced, and quality-flagged samples no longer cluster together. Also, samples segregate more clearly by disease status after adjustment.

Furthermore, applying ComBat clearly has a stabilising effect on the transcript clusters most affected by RNA quality (Figures 6b & 6c).

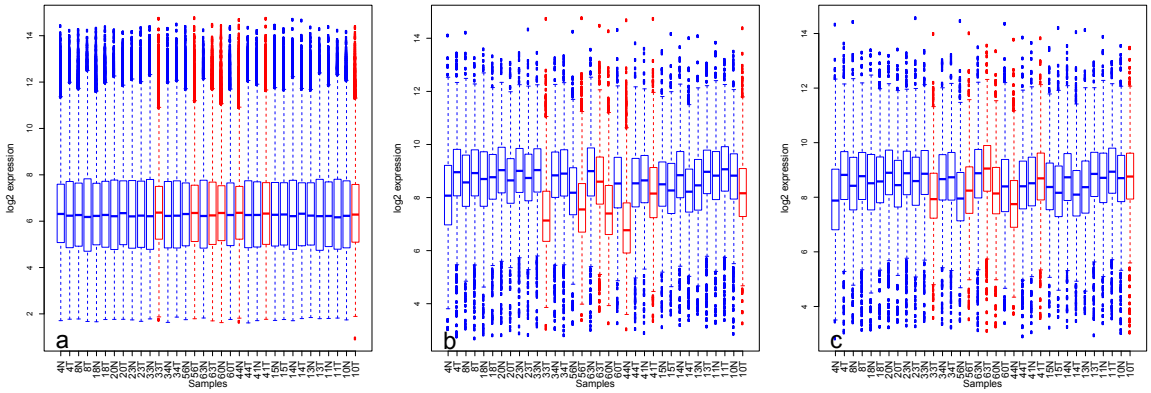


Figure 6: Boxplots of frma-normalised expression values. a) All transcript clusters. b) Genes most affected by quality (adjusted p-value ≤ 0.01, |fold change| > 2. c) Genes most affected by quality after adjustment for batch- and quality-effects using ComBat. Samples that were flagged during quality assessment are highlighted in red.

SVA identified two surrogate variables that were subsequently used in downstream analysis. Plotting the estimates of these surrogate variables for each sample revealed a pattern whereby samples were clearly grouped by batch and quality (Figure 7). Importantly, SVA identified these two variables without supervision.

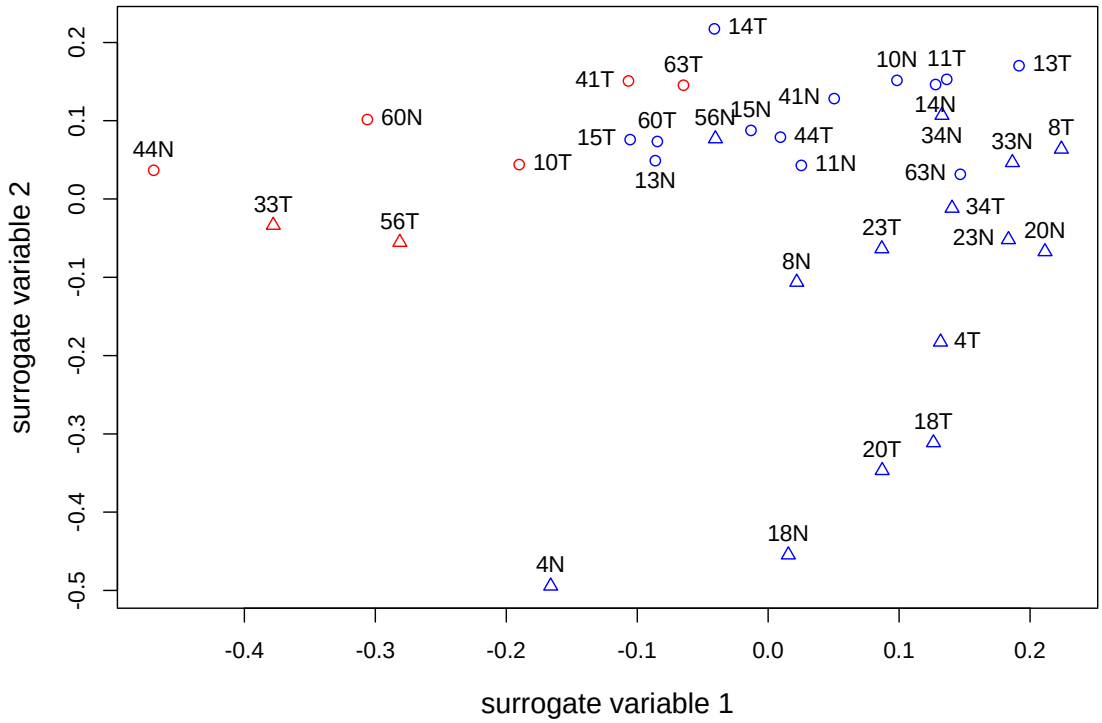


Figure 7: Surrogate variable analysis results. Two latent variables were identified by SVA. Samples highlighted in triangles or circles were from two different batches. Samples that were flagged during quality assessment are highlighted in red.

To evaluate the performance of each method, we first compared the number of differentially expressed genes detected between tumour and normal samples at a stringent p-value of 0.01. For our analysis, we did not use a fold change cutoff since we feel that artificial fold change cutoffs, which exclude subtle changes in the expression of many genes, may result in the loss of valuable biological information, or worse, affect the interpretation of the data—this is particularly true for applications such as network/pathway analysis[303].

SVA and ComBat detected 2137 and 1945 genes (p-value $\leq$ 0.01), respectively. The top four methods had 1117 differentially expressed genes in common (Figure 8). At the commonly used p-value- and fold change-cutoffs of 0.05 and 2 respectively—SVA, Combat, ArrayWeights and excluding arrays, produced 447, 475, 461 and 521 differentially expressed genes – suggesting similar performance under these criteria. We next assessed the relevance of these differentially expressed genes, in colorectal cancer, using Ingenuity Pathway Analysis where, statistically significant over-representation of our listed genes in a given process such as *colorectal tumour* or *infection of embryonic cell lines* is scored by p-value.

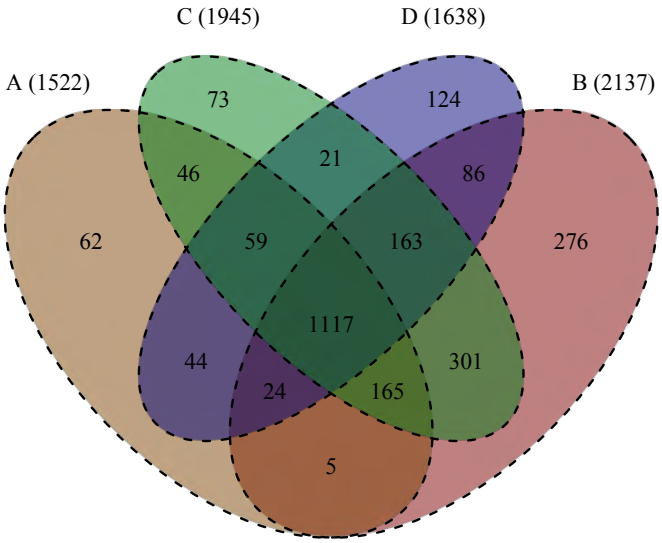


Figure 8: Venn diagram of unique differentially expressed genes (tumour vs. normal) with adjusted p-values ≤ 0.01 for the four best-performing methods. A) removing quality-flagged arrays before analysis. B) applying SVA to batch corrected data. C) ComBat used to correct for batch and quality. D) Array weights included in the linear model.

We considered the top 10 functions for each method (Table 2) from which it was clear that the 615 and 423 additional genes identified as differentially expressed by SVA and ComBat, compared to that obtained when excluding quality-flagged arrays, were certainly relevant to colorectal cancer.

Table 2: P-values for evidence for overrepresentation in the functions listed for each method.

Functions	A	B	C	D	E
Cancer	7.72E-29	NA	8.15E-24	NA	3.38E-23
cancer	NA	2.58E-25	NA	2.70E-26	NA
carcinoma	8.64E-37	2.52E-33	1.87E-34	2.87E-32	5.56E-30
colon cancer	1.30E-26	1.19E-36	1.10E-26	1.99E-21	3.29E-21
colon tumor	1.10E-26	4.31E-37	3.65E-27	1.80E-21	7.28E-22
colorectal cancer	2.27E-26	4.74E-29	2.43E-26	1.11E-21	1.98E-23
colorectal tumor	2.28E-26	6.80E-29	2.97E-26	4.67E-22	1.72E-23
digestive organ tumor	2.68E-31	6.82E-32	1.24E-28	7.27E-27	2.72E-29
epithelial tumor	2.16E-38	NA	2.27E-35	NA	1.11E-30
gastrointestinal tract cancer	2.35E-25	2.42E-28	4.00E-24	3.19E-21	5.31E-22
intestinal cancer	2.02E-26	5.77E-29	2.58E-26	1.03E-21	1.55E-23
neoplasia	NA	1.63E-24	NA	1.10E-25	NA
solid tumor	3.31E-35	8.07E-32	6.80E-33	4.65E-31	3.88E-29
tumorigenesis	NA	1.55E-26	NA	3.31E-28	NA
uterine serous papillary cancer	3.46E-21	1.71E-20	8.61E-25	1.26E-22	1.14E-15

A) excluding quality-flagged arrays from the analysis; B) applying SVA to batch corrected data; C) ComBat used to correct for batch and quality; D) Array weights included in the linear model; E) including batch and quality as factors in the linear model.

Using IPA, we considered the top 10 upstream regulators (highest absolute activation z-scores) when comparing tumour vs. normal samples, to further investigate the utility of SVA or ComBat as suitable analysis methods when including low-RIN samples (Table 3). We found considerable overlap in the identity and direction of activation of these upstream regulators between the methods compared.

Table 3: Top 10 IPA-derived upstream regulators, by absolute activation z-score. A) excluding quality-flagged arrays from the analysis; B) applying SVA to batch corrected data; C) ComBat used to correct for batch and quality.

A

Upstream Regulator	Log Ratio	Molecule Type	Predicted Activation State	Activation z-score	p-value of overlap
TP53 (includes EG:22059)		transcription regulator	Inhibited	-4.88	1.05E-16
CDKN1A	-0.469	kinase	Inhibited	-3.274	4.20E-10
TRAF2		enzyme	Activated	2.804	3.06E-06
CCNK		other	Activated	2.905	3.83E-04
TNF		cytokine	Activated	2.935	7.69E-04
IL1B		cytokine	Activated	2.952	1.76E-01
TP63		transcription regulator	Activated	3.181	8.37E-10
TREM1		other	Activated	3.352	3.69E-05
FOXM1	1.37	transcription regulator	Activated	4.28	3.71E-17
Mek		group	Activated	4.336	2.38E-07

B

Upstream Regulator	Log Ratio	Molecule Type	Predicted Activation State	Activation z-score	p-value of overlap
TP53 (includes EG:22059)	0.622	transcription regulator	Inhibited	-5.749	6.48E-12
TGM2 (includes EG:21817)		enzyme	Inhibited	-4.243	3.64E-02
CDKN1A	-0.485	kinase	Inhibited	-3.548	1.85E-10
KDM5B		transcription regulator	Inhibited	-3.126	3.31E-08
NFkB (complex)		complex	Activated	3.034	3.59E-03
TREM1		other	Activated	3.073	2.18E-05
TP63		transcription regulator	Activated	3.63	6.25E-06
IL1B		cytokine	Activated	3.686	4.13E-01
FOXM1	1.29	transcription regulator	Activated	3.925	5.82E-11
Mek		group	Activated	4.771	7.08E-08

C

Upstream Regulator	Log Ratio	Molecule Type	Predicted Activation State	Activation z-score	p-value of overlap
TP53 (includes EG:22059)		transcription regulator	Inhibited	-5.126	1.30E-13
CDKN1A	-0.496	kinase	Inhibited	-3.534	5.99E-10
TGM2 (includes EG:21817)		enzyme	Inhibited	-3.402	4.25E-02
miR-483-3p (miRNAs w/seed CACUCCU)		mature microRNA	Inhibited	-3.153	6.49E-03
EGFR		kinase	Activated	3.104	4.43E-03
IL1B		cytokine	Activated	3.281	1.73E-01
TP63		transcription regulator	Activated	3.524	1.48E-09
TREM1		other	Activated	3.845	5.74E-06
FOXM1	1.398	transcription regulator	Activated	4.386	4.18E-16
Mek		group	Activated	4.654	9.72E-08

qRT-PCR validation of select genes

In order to ascertain whether or not data obtained by microarray analysis with low-RIN samples were comparable to the results obtained using the method designed by Antonov et al. for qPCR analysis of low-RIN samples, we selected two genes, dipeptidase 1 (*DPEP1*) and claudin 1 (*CLDN1*), for qRT-PCR validation. Given that our microarray data analysis suggests ~95% of genes are unaffected by RNA integrity, we wished to

compare microarray and qPCR data for genes that were apparently unaffected by RNA integrity; DPEP1and CLDN1 were found to be significantly differentially expressed in our microarray data by all of the five methods used and, in addition, there is strong literature evidence for their differential expression between tumour and normal samples. From reference genes previously cited as suitable for colorectal cancer studies, we selected those most stably expressed in our cohort using the Normfinder algorithm (UBC, B2M, ATP5E)[290–294]. We found good correlations, for both CLDN1 (Adjusted $R^2 = 0.81$) and DPEP1 (Adjusted $R^2 = 0.83$), between qRT-PCR- and microarray-based fold change values (Figure 9), irrespective of RIN score.

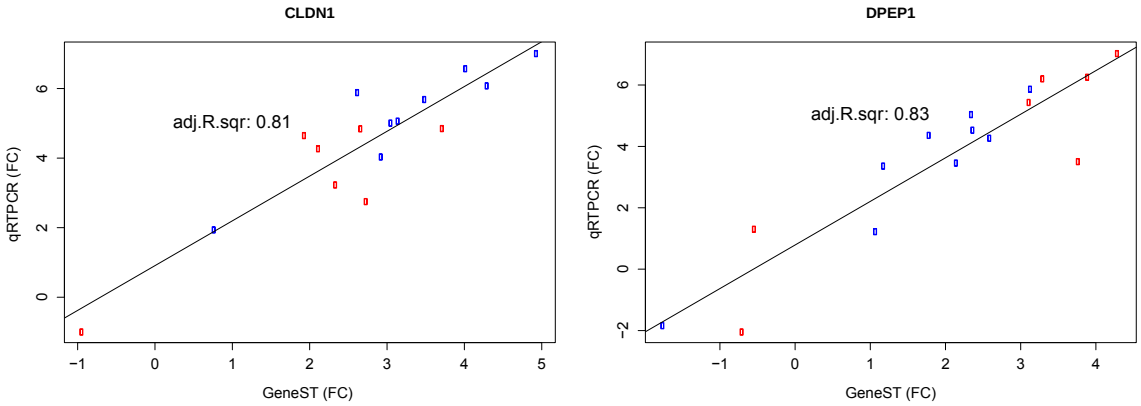


Figure 9: DPEP1 and CLDN1 tumour vs. normal fold change (FC) results for qRT-PCR and microarray results. Samples that were flagged during quality assessment are highlighted in red.

Discussion and conclusions

RNA is extremely vulnerable to degradation and as such has the potential to introduce a systematic bias in gene expression measures. Reliable measures of sample and data quality are therefore essential to evaluate the effects of RNA integrity on accuracy, sensitivity and specificity of gene expression results. From previous studies as well as our own, it is now clear that the level of acceptable RNA degradation within an experiment depends largely on the experimental design, platform and application. Multiple studies have demonstrated an improvement in microarray and qRT-PCR performance by using random priming when RNA integrity is in doubt. Here we observed a direct association between RINs and array quality in the majority of cases. To gauge the consequences of using these arrays in downstream analysis, we compared quality-flagged to quality-passed arrays and found a relatively small subset of genes,

1172/20019, to be significantly affected (p-value 0.05, FC ≥ |2|) in our samples on the Gene 1.0 ST platform. It is of course possible that the exact identity and proportion of the affected genes may differ between studies on Gene 1.0 ST arrays but, based on our data, we suggest that the overall proportion of affected genes is unlikely to be significantly different to that observed here. Depending on the application, this may or may not have an effect on the study outcome. However, the most common microarray applications such as finding differentially expressed genes between two conditions, pathway analysis, and clustering do not rely on interrogating specific genes and appear to be largely robust to the effects of RNA degradation on this platform (Table 2).

Using within- and between-array quality measures, we investigated the relationship between RNA integrity and array quality on Affymetrix Gene 1.0 ST arrays. We found a combination of within- and between-array quality measures useful to rank samples by array quality. However, the single most useful array quality measure appears to be GNUSE, since it provides a more general measure of array quality relative to a large set of publically available arrays. We found that 86% of samples with RINs ≤ 3.3 were flagged by at least two of our quality control measures. One sample with RIN score < 3 passed all three quality measures, although it did have relatively low array quality weight. Furthermore, 10 out of 17 samples with RIN scores ≤ 7 passed at least 2 out of 3 quality measures, suggesting that the widely used RIN cutoff of 7 is too stringent for Gene 1.0 ST arrays.

We then examined the genes most affected by RNA degradation and demonstrated a relationship between accuracy and length of the original transcript, with both longer than average, and very short transcripts being under- and overrepresented in quality-flagged samples respectively. This is in contrast to the findings by Opitz et al. who found that short transcripts were more vulnerable to the perceived effects of degradation, whereas long transcripts were more stable relative to the average length transcript[282]. Interestingly, of the genes that were overrepresented in quality-flagged samples, 70% were small non-protein coding RNAs, including 94 small nucleolar RNAs, and 4 microRNAs, consistent with reports that microRNAs are more robust to RNA degradation compared to mRNA[304], perhaps because they are more thermodynamically stable than mRNAs.

Without excluding any genes, we then compared the orthogonal approaches of either excluding quality-flagged arrays or compensating for RNA degradation at different steps in the analysis. Sample clustering showed that when using ComBat adjustment, quality-flagged samples no longer clustered together. Furthermore, samples tend to segregate more clearly by disease status following adjustment, which suggests that the algorithm is not introducing artifacts. It is worth noting that patients 13, 4 and 18 were diagnosed with a hereditary form of CRC (HNPCC)—it is therefore not surprising that the 'normal' samples from these patients form a separate cluster.

Irrespective of sample/array quality, applying compensatory measures for RNA degradation performed at least as well as excluding arrays that were flagged during quality assessment, as judged by gene expression analysis and IPA. At a p-value of 0.01, SVA and Combat detected the highest number of differentially expressed genes between tumour and normal samples and the top four methods applied here had 1117 differentially expressed genes in common. To evaluate the biological plausibility of the genes deemed significantly differentially expressed between tumour and normal samples, we harnessed the results from IPA to show that, in terms of the top scoring biological functions and upstream regulators, there is considerable overlap in the identity and direction of biological activation when comparing analysis methods that either excluded or included quality-flagged arrays. These results suggest that our analysis strategies are biologically sound and not biased by non-biological variance.

The relevance of each method will depend on the downstream application and the proportion of quality-flagged arrays: If a small percentage of arrays are flagged, there might not be much benefit in including them for downstream analysis. However, if a large proportion of the arrays are affected by RNA quality—which is likely to often be the case where the RNA is derived from irreplaceable historical clinical samples—the ability to retain all arrays and to account for these effects in the analysis will be valuable. Here, ComBat may be useful if direct data adjustment is required, e.g. for sample/gene clustering. On the other hand, for analysis of differential expression, especially when the source of non-biological variance is not immediately apparent, SVA may be most useful since it does not require supervision; notably, in our hands SVA was able to identify two surrogate variables which closely corresponded to "batch" and "quality"

factors, judged by the grouping of samples. To establish whether the measures used here to compensate for quality-effects are superior to excluding these arrays from the analysis will require a controlled study with known true- and false-positives where the discriminatory power of each method can be objectively investigated. However, the significant overlap observed between the differentially expressed genes identified by the different approaches used here, combined with the considerable overlaps in both biological function and upstream regulators identified by pathway analysis of the resultant data, argues against a simple expansion of false positives when lower quality array data is included in the analyses. The quality assessment and data analysis methods discussed here should in principle be as useful for Affymetrix Exon ST array analysis as well.

In conclusion, array quality measures can be used to set quality thresholds, to provide valuable information that can be used to improve the linear model of differential expression, or to correct expression signal prior to assessing differential expression. We suggest that accounting for known or unknown sources of variation, such as variable RNA integrity and batch, by implementing ComBat or Surrogate Variable Analysis for analysis of differential gene expression enables robust analysis of microarray datasets derived from variable and low quality RNA, thereby extending the range of clinical samples that are suitable for microarray analysis.

Chapter 6: Whole-genome methylation analysis of CRC tumour and adjacent normal mucosal samples in relation to bacterial infection

Abstract

Whole-genome methylation analysis was conducted on 24 pairs of colorectal adenocarcinoma and matched normal mucosal samples (19 of these 24 pairs were also analysed by whole-genome gene expression analysis, Chapters 7 and 8).

In this Chapter a brief overview of DNA methylation in general and in the context of CRC is provided. Next, whole-genome methylation technologies are reviewed, with a detailed description of the Illumina HumanMethylation450 BeadChip technology used thereafter.

Methylation patterns were assessed using a model-based approach to unsupervised clustering in our cohort both before and after merging our cohort with a large publically available dataset of colorectal adenocarcinomas (N=361), with confirmed CpG island methylator phenotype (CIMP) annotation, available from The Cancer Genome Atlas (TCGA).

In the merged dataset (N=385), four clusters with varying levels of CpG island methylation were obtained: two clusters were dominated by CIMP-low (L) samples, another by CIMP-high (H) samples, while the fourth cluster was almost entirely composed of CIMP-stable samples. Clinically relevant features included the enrichment of MSI-H samples in the CIMP-H cluster and the predominance of proximal cancers and stage I/II cancers in the two clusters displaying the highest level of CpG island methylation.

CIMP-status in our cohort was predicted using unsupervised clustering of CpG island probes mapping to a published five-gene array-based marker panel; 9 of 24 samples were classified as CIMP-L, 1 as CIMP-H and 14 as CIMP-stable.

Analysis of variance by CIMP-status revealed 2172 probes associated with CIMP-status ($p \leq 0.05$, $|\Delta \text{ beta}| \geq 0.2$), 93% of which mapped to CpG islands and 99.8% of which showed increased methylation in CIMP+ vs. CIMP– samples; moreover, of the 94 genes significantly associated with CIMP status (FDR ≤ 0.25), at least 23% (22/94) had previously been found to show increased CpG island methylation in CIMP+ vs. CIMP– tumours[305]; of these, *BDNF*[306] and *KCNK13*[305] have been used in published CIMP-marker panels.

In addition to array-based methylation analyses, *MLH1* methylation (the main cause of MSI-H in sporadic CRCs) was validated by methylation-specific PCR. Two samples tested positive for *MLH1* methylation, both of which were sporadic MSI-H CRCs that were predicted to be CIMP+.

Lastly, any patterns of methylation associated with CRC associated pathogens (including Fusobacterium, *Enterococcus faecalis*, Enterotoxigenic *Bacteroides fragilis* (ETBF), Enteropathogenic *Escherichia coli* (EPEC), ClB+ *E. coli* and afaC+ *E. coli*) were determined. EPEC was associated with altered patterns of host methylation in tumour samples, but the majority of CpG sites were not significantly associated with EPEC- colonisation after multiple testing correction. This is not surprising since there were only three EPEC+ samples in this cohort, and multiple testing correction imposed a substantial penalty on the p-values, given that ~300 000 probes were analysed concurrently. Nevertheless, two of the three EPEC+ samples had confirmed *MLH1* promoter methylation by methylation-specific PCR, and all three EPEC+ samples were predicted to be CIMP+. Taken together, these results point towards a possible link between EPEC and aberrant methylation in the colon, which warrants validation in a larger cohort.

Introduction

DNA methylation–the covalent addition of a methyl group to cytosine–is catalyzed by DNA methyltransferases (DNMTs), a family that includes members with de novo- and maintenance-methylation functions.

CpG dinucleotides (a cytosine followed by a guanosine) are the most frequent site of methylation, with ~80% methylated in healthy cells. Because methylated cytosines tend to be deaminated to thymine over evolutionary time, CpG dinucleotides occur at a frequency of 20–25% less than expected by chance[307]. However, more than 60% of genes contain CpG-rich regions, referred to as CpG islands; these regions are 300–3000bp in length and have a GC percentage of at least 50%, and an observed-to-expected CpG ratio of at least 0.6, where the observed-to-expected CpG ratio is calculated by the formula ((Number of CpG/(Number of C × Number of G)) × Total number of nucleotides in the sequence)[308]. The cutoffs for this definition are however quite arbitrary and may vary between studies[309].. In contrast to CpG poor regions, CpG islands are typically unmethylated. DNA methylation across the genome therefore follows a bimodal distribution, with the majority of CpG sites methylated at very high levels (>85%), and CpG islands largely unmethylated (<15%)[310].

Although DNA methylation is commonly associated with gene silencing, its function is more accurately region-specific. While DNA methylation serves to immobilise transposable elements genome-wide[311], CpG-island-based methylation may activate or repress transcription of specific genes. Methylation of promoter-based CpG islands is strongly associated with transcriptional repression due to the inhibition of binding of methylation-sensitive transcriptional activators, or to affects on the binding of proteins that orchestrate changes in chromatin conformation[312]. On the other hand, developmentally programmed methylation of CpG islands in 3' regions have been associated with tissue- and cell-type-specific transcriptional activation[312]. Further, non-CpG island methylation in gene bodies is even thought to affect gene splicing[313]. It is therefore becoming apparent that DNA methylation is not simply an on-off switch for transcription, but has various, more subtle roles across different regions of the genome.

Methylation patterns in CRC

Aberrant methylation is a common occurrence in CRC, with global hypomethylation, and region-specific hypermethylation associated with chromosomal instability (CIN) and microsatellite instability (MSI), respectively[152–154]. About 70% of proximal MSI+

cancers exhibit frequent hypermethylation of CpG islands[155], and hypermethylation of the promoter region of *MLH1* is the main cause of MSI-H in sporadic CRCs[102].

Prior to the development of technologies for evaluating methylation genome-wide, Toyota et al. defined 33 markers that were differentially methylated between Caco-2 cells and normal mucosal samples[314]. These markers were subsequently divided into cancer-specific and age-related markers[315]. The subset of CRCs that displayed methylation of cancer-specific markers was referred to as CIMP[315]. Five genes were subsequently selected as surrogate markers of CIMP *(CDKN2A, MLH1, MINT1, MINT2* and *MINT31*) that could be used to classify samples as CIMP+ or CIMP-stable. This panel has since been updated based on analysis of a larger number (195) of CpG loci. The new panel consists of *CACNA1G, IGF2, NEUROG1, RUNX3* and *SOCS1*[155,306], and is commonly used for PCR-based detection of CIMP.

The classification of samples as CIMP+ has since been subdivided into CRCs with high- (CIMP-H) or intermediate/low (CIMP-L) levels of CpG island methylation. Importantly, CIMP-H and CIMP-L CRCs appear to have heterogeneous molecular and clinicopathological features: CIMP-H CRCs are associated with the proximal colon, older age, female sex, frequent *BRAF* mutation, *MLH1* methylation, MSI and rarely *KRAS* and *TP53* mutations; meanwhile CIMP-L CRCs are associated with *KRAS* mutation, but not MSI, or mutations in *BRAF* or *TP53*[316]. The mechanisms underlying these associations remain unclear. Regarding the mechanistic basis of CIMP, the DNA methyltransferase DNMT3B as well as SNPs in folate metabolizing enzymes have been implicated[316] but additional clarification is required. Further, numerous studies have investigated CRC prognosis according to MSI, CIN, CIMP and *BRAF*/*KRAS* mutation status, but the relationship between these factors appears complex.

Unsupervised clustering of whole-genome methylation data supports the association between the degree of CpG island methylation encountered and specific clinicopathological and molecular features[317]. Using a model-based approach to unsupervised clustering, Hinoue et al. identified four classes of methylation (determined across 125 CRCs): a relatively small subset of CRCs displayed a particularly high-level and -frequency of CpG island methylation, along with frequent mutations in *BRAF,*

MLH1 promoter methylation, MSI-H and proximal colonic location—these cases were classified as CIMP-H; the remaining three clusters were defined as: CIMP-L (characterised by hypermethylation of a subset of CIMP-H markers), cluster 3 (which displayed less CpG island methylation compared to the CIMP-L clusters), and cluster 4 (which displayed even less CpG island methylation compared to cluster 3). Hinoue et al. developed two diagnostic DNA methylation gene marker panels, each consisting of five genes, based on Illumina Infinium DNA methylation data of 125 CRCs; one panel is used to identify CIMP (CIMP-H or CIMP-L), and the other is used to segregate CIMP-H tumors from CIMP-L tumors[317].

Meanwhile, global hypomethylation occurs particularly in non-CpG-island regions of the genome, including repetitive sequences and in CpG island shores[156]. While hypomethylation has been associated with activation of a handful of tumour supressors (e.g. *CDH3*), the most apparent consequence of global hypomethylation is chromosomal instability[88,89,153]. This occurs when pericentromeric regions are demethylated, which facilitates recombination and altered chromosomal replication[156].

While certain disease-related methylation patterns such as global DNA hypomethylation are quite obvious, whole-genome methylation analysis is still in its infancy, and it is essential to first establish a baseline tissue-specific methylation profile, as well as the extent of inter-individual and temporal variations in methylation profiles, before more subtle disease-related methylation patterns can be appreciated[147]. For example, age-related methylation occurs in the majority of tissues, where a subset of genes undergo hyper- and hypomethylation in CpG-rich and -poor regions, respectively[147]. In cancer, cigarette smoking and body mass index have been associated with increased methylation of CpGs, which was attenuated by hormone replacement therapy or aspirin use[318].

Finally, numerous pathogens can induce aberrant methylation in host cells, including *H. pylori*[150,319,320] and Epstein Barr in gastric cancer[151], as well as Influenza virus, *Pseudomonas syringae*, *Wolbachia pipientis*[321] and *Campylobacter rectus*[322]. Whether or not microbially-induced methylation plays a role in CRC remains unknown.

Methylation analysis technology overview

Although DNA methylation testing has been conducted on a per-gene basis for over 20 years, microarray and sequencing technologies have only recently been implemented allowing cost-effective, whole-genome methylation analysis.

Methylation patterns can be captured in a high-throughput manner following one of two strategies to record methylation sites: 1) bisulfite conversion of DNA, which specifically mutates unmethylated cytosines (through deamination) to uracil, allowing genotype-based analysis of methylation (using microarrays or sequencing) at single nucleotide level, or 2) enrichment-based methods that employ DNA-methylation-specific antibodies, methyl-binding domain proteins, or restriction enzymes to enrich for a fraction of highly methylated (or unmethylated) DNA fragments. Enrichment of specific fragments are then quantified by next-generation sequencing[323].

An overwhelming variety of platforms are currently available for DNA methylation testing, selection from which should be guided by the nature of the sample to be analysed (specifically regarding the amount and integrity of DNA available), as well as the tradeoff between cost and genomic coverage.

Because enrichment-based methods require relatively large amounts of intact (in the case of restriction enzyme methods) DNA, bisulfite-based methods are most suitable for clinical samples, which often have variable quality (e.g. FFPE-derived DNA) and quantity (e.g. scarce clinical samples). Bisulfite-based methods are also more accurate and reproducible, as long as bisulfite conversion and PCR efficiencies are consistent. Further, bisulfite-based methods allow profiling of CpG-poor regions that are largely excluded by CpG-enrichment based methods[324].

One disadvantage of bisulfite-based methods is the inability to distinguish between bisulfite-induced deamination of unmethylated cytosines to uracil, and single nucleotide polymorphisms (SNPs), which are particularly prevalent in CpG dinucleotides, since

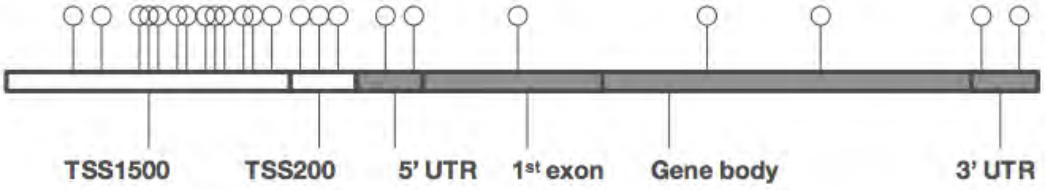
methylated cytosine are naturally deaminated to thymine over evolutionary time. One solution is to exclude known SNPs from the downstream analysis. In the case of bisulfite-sequencing, both the forward and reverse strand can be sequenced to allow discrimination between SNPs and unmethylated CpGs, since the opposing strand would have been propagated as an A as opposed to a G in the case of a true SNP[324].

Bisulfite-based microarrays in particular are commercially available only for human samples, and although they are more cost-effective and require less technical expertise than sequencing, this comes at the cost of genomic coverage.

Lastly, enrichment-based methods efficiently assess genome-wide methylation on a broad scale, but they do not yield information on individual CpG dinucleotides, and quantification may be biased by copy number variations (particularly in cancer samples)[324]. In the case of enzyme-based methods, this may be remedied by measuring the ratio between methylated and unmethylated versions of a sequence, as opposed to either on their own[324]. The main advantage of enrichment-based methods is that they provide cost-effective assessment of genome-wide DNA methylation, albeit at relatively lower resolution and with higher susceptibility to experimental bias[323].

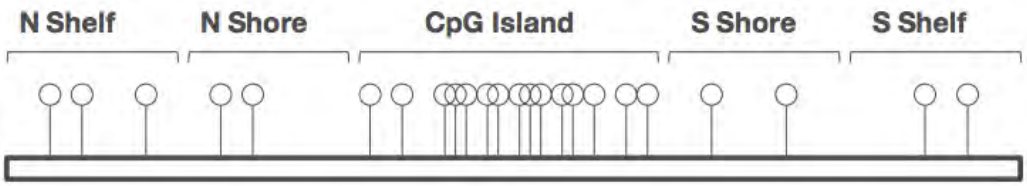
Illumina HumanMethylation450 BeadChips

The Illumina HumanMethylation450 BeadChip was selected for this study. This is a bisulfite-based array, which covers selected areas of 99% of Refseq genes, with an average of 17 CpG sites per gene region, where each gene region is divided into six feature types (Figure 1). Further, 94% of CpG islands are covered, with similar coverage for adjacent regions (Figure 2).



Feature Type	Genes Mapped	Percent Genes Covered	Number of Loci on Array
NM_TSS200	14895	0.79	2.56
NM_TS1500	17820	0.94	3.41
NM_5'UTR	13865	0.78	3.34
NM_1stExon	15127	0.80	1.62
NM_3'UTR	13042	0.72	1.02
NM_GeneBody	17071	0.97	8.97
NR_TSS200	1967	0.65	1.84
NR_TSS1500	2672	0.88	2.92
NR_GeneBody	2345	0.77	5.34

Figure 1: Coverage of NM and NR transcripts from UCSC database. Each transcript is divided into function regions: TSS200 is the region from the transcription start site (TSS) to –200 nt upstream of the TSS; TSS1500 covers –200 nt to –1500 nt upstream of TSS; 5' untranslated region (5'UTR), 1st exon, gene body and 3'untranslated region (3'UTR) are covered separately. Figure reproduced with permission from Illumina[325].



Feature Type	Islands Mapped	Percent Islands Covered	Average Number of Loci on Array
Island	26153	0.94	5.08
N_Shore	25770	0.93	2.74
S_Shore	25614	0.92	2.66
N_Shelf	23896	0.86	1.97
S_Shelf	23968	0.86	1.94

Figure 2: Coverage of regions in relation to CpG islands. The 2 kb regions immediately upstream and downstream of the CpG island boundaries are referred to as CpG island shores, and the 2 kb regions upstream and downstream of the CpG island shores are referred to as CpG island shelves. N: north; S: south. Figure reproduced with permission[325].

Regions covered outside CpG islands include non-CpG methylated sites, sites differentially methylated in various tumour versus normal samples and across several tissue types, CpG islands outside coding regions, and miRNA promoter regions.

The assay includes whole-genome amplification of bisulfite-converted DNA, followed by fragmentation and hybridization to the array. For each locus, the percentage of methylation is determined using a combination of sequence-specific hybridization capture and allele-specific single-base primer extension. One of two array chemistries are used to achieve this (Figure 3): Infinium I beadtypes consist of an unmethylated and a methylated bead for each locus examined; if the locus happens to be methylated, it can only bind to a methylated bead type (and vice versa), upon which a single, fluorescently labeled nucleotide (A or T (red), and C or G (green)) is incorporated immediately after the CpG locus. Infinium II beadtypes (which make up 72% of probes on the array) have one bead for both methylated and unmethylated loci; the nucleotide that binds reveals the methylation status since it completes the CpG locus; if an A (red) or G (green) is incorporated, the locus was unmethylated or methylated, respectively. Importantly, type I probes map to more CpG islands (57%) compared with type II probes (21%)[326].

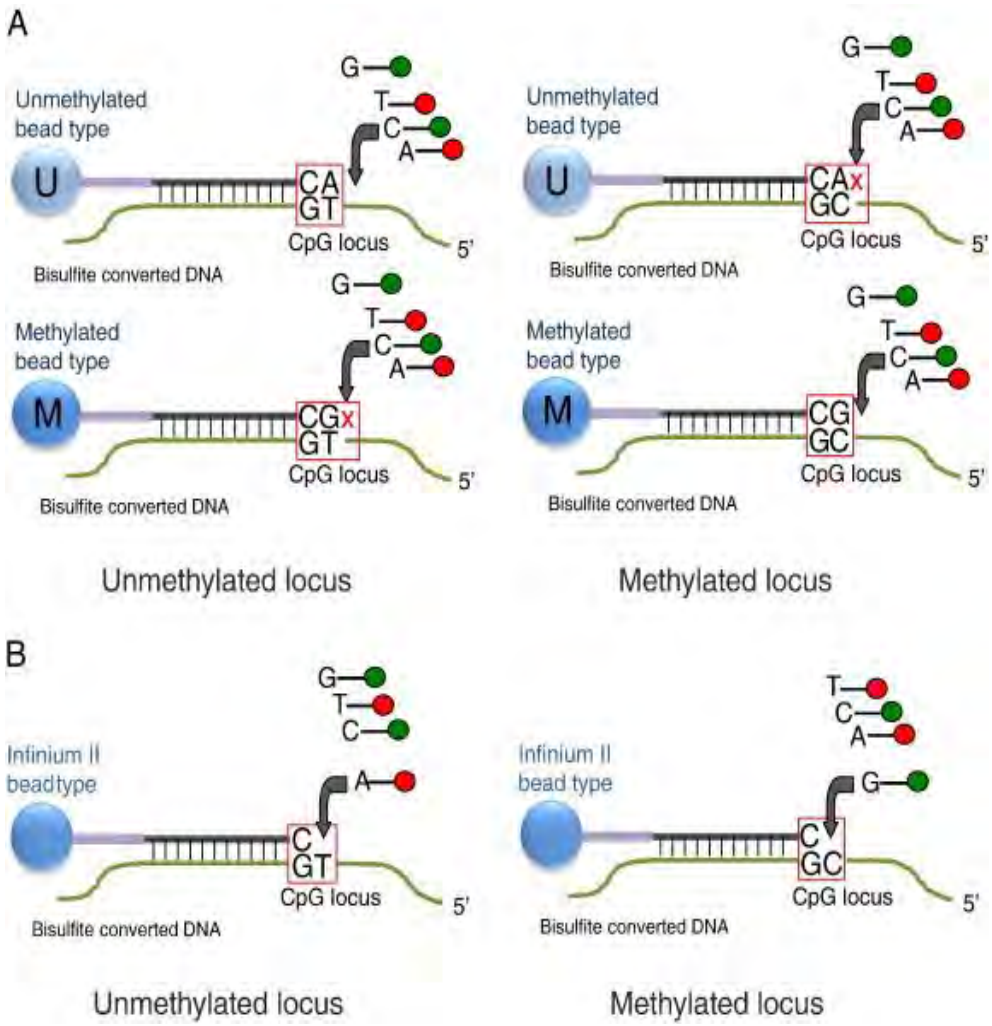


Figure 3: Illumina bead chemistries used to detect methylation. A: Infinium I beadtypes consist of an unmethylated and a methylated bead for each locus examined; if the locus happens to be methylated, it can only bind to a methylated bead type (and vice versa), upon which a single, fluorescently labeled nucleotide (A or T (red), and C or G (green)) is incorporated immediately after the CpG locus. B: Infinium II beadtypes have one bead for both methylated and unmethylated loci; the nucleotide that binds reveals the methylation status since it completes the CpG locus; if an A (red) or G (green) is incorporated, the locus was unmethylated or methylated, respectively. Figure reproduced with permission[325].

The level of methylation, on a scale of 0–1 at a particular CpG site, is summarized as a beta-value, where the ith interrogated CpG site is defined as,

$$ß = \frac{\max(y_{i,methy}, 0)}{\max(y_{i,methy}, 0) + \max(y_{i,unmethy}) + \alpha}$$

where $y_{i,methy}$ and $y_{i,unmethy}$ are the intensities measured by the ith methylated and unmethylated probes respectively, and any negative values after background adjustment

are set to 0. A constant offset α (by default, $\alpha = 100$) is added to the denominator to regularise beta when both methylated and unmethylated probe intensities are low[327]. The beta-statistic ranges between 0 (100% unmethylated) and 1 (100% methylated).

Performance-wise, precision and accuracy on the Illumina HumanMethylation450 BeadChip is very good, with $R > 0.98$ between technical replicates; the maximum sensitivity measured is a beta value of 0.2 (<1% false positive rate).

Quality may be assessed by a total of 850 control bead types on the array, about 70% of which are 'negative' controls used by Illumina's GenomeStudio software for background correction[328]. The control probes include non-specific binding controls, staining controls, extension controls, target removal controls, and hybridization controls.

Differential methylation analysis considerations

Unlike gene expression analysis, averaging methylation values across a given gene has little biological meaning since each gene consists of multiple functional regions, each of which may be differentially affected by methylation. Differential analysis can therefore either be conducted for individual CpG sites (which have been annotated as CpG island or non-CpG island), or by region of interest (e.g. TSS200, 1st exon, gene-body) after calculating region-specific methylation values.

The beta distribution of methylation, together with the natural heterogeneity in overall intensity between samples, presents a further challenge to analysis—the widely used statistical framework for preprocessing and analysing gene expression data (which is based on the assumption of normality and similar average intensities between samples) may not be suitable. Statistical methods for differential methylation analysis is therefore an active area of investigation, which will no doubt see much change over the next few years. Wilhelm-Benartzi et al. have reviewed current processing and analysis methods for DNA methylation array data[329].

Materials and methods

Methylation analysis of *MLH1* using methylation specific PCR

MLH1 promoter methylation was determined by methylation specific PCR (MSP) in DNA extracted from tumour, normal and blood samples for 30 of the 32 patients for whom MSI testing was conducted. DNA was first bisulfite converted using the EZ DNA Methylation-Gold Kit™ (Zymo Research), where DNA was denatured for 10 minutes at 98ºC (since bisulfite only reacts with single stranded DNA), followed by incubation for 4 hours at 64ºC with bisulfite. The resulting bisulfite converted DNA was cleaned and desulphonated using spin columns.

The *MYOD* gene is considered to be constitutively methylated and was therefore selected as a positive control to assess DNA integrity, bisulphite conversion efficiency, and to normalize DNA input. Serial dilutions were constructed for *MYOD* and *MLH1* plasmids to include 10^6, 10^5, 10^4, 10^3, 10^2, 50 and 25 copies of each gene for quantification purposes. Samples in which at least 1000 copies of the *MYOD* template were detected passed quality control—this equates to a minimum analytical sensitivity of 2.5% of all alleles. Samples that had a negative result for both *MYOD* and *MLH1* were repeated. Appropriate controls were included throughout the protocol, which are listed in Table 1. The primers used for MSP of *MLH1*, and the control gene *MYOD* are listed in Table 2.

Table 1. Positive and negative controls included in the methylation detection protocol.

Control type	**Stage where used**
No template control	Bisulfite conversion
Human methylated DNA standard (Zymo Research)	Bisulfite conversion
No template control	Methylation specific PCR
Human Bisulfite converted methylation DNA standard (Zymo Research)	Methylation specific PCR

Table 2. Primers used for *MLH1* and *MYOD* MSP.

Gene	F & R primers (5'–3')
MYOD	F-CCAACTCCAAATCCCCTCTCTAT
	R-TGATTAATTTAGATTGGGTTTAGAGAAGG
MLH1	F-CGTTAAGTATTTTTTTCGTTTTGCG
	R-TAAATCTCTTCGTCCCTCCCTAAAACG

PCRs were conducted in 20 µl reactions in 96-well plates, using 1µl forward and 1µl reverse primer (from 6mM stock), 10µl BIORAD IQ supermix (from 2x stock), 2.5µl bisulfite converted DNA or 2µl plasmid standard and 5.5µl water. The cycling conditions are specified in Tables 3a (*MYOD*) and 3b (*MLH1*), and qPCR was conducted on a Bio-Rad C1000 Thermocycler.

Table 3a: Cycling protocol for *MYOD* MSP.

Step	Duration	Temperature	Cycles
Denaturation	60 seconds	95 °C	1
Denaturation	30 seconds	95 °C	
Annealing	30 seconds	59 °C	37
Extension	30 seconds	72 °C (plate read)	
Melt curve	5 seconds	65°C to 95°C (increment 0.5 °C)	1

Table 3b: Cycling protocol for *MLH1* MSP.

Step	Duration	Temperature	Cycles
Denaturation	60 seconds	95 °C	1
Denaturation	30 seconds	95 °C	
Annealing	30 seconds	61.5 °C	39
Extension	30 seconds	72 °C	
Extension	20 seconds	74 °C (plate read)	

Melt curve	5 seconds	65°C to 95°C (increment 0.5 °C)	1

Whole genome array-based methylation analysis

For whole-genome array-based methylation analsyis, samples were bisulfite converted using the EZ DNA Methylation-Gold Kit™, as described for MSP-based detection of *MLH1* methylation; again bisulfite conversion efficiency was assessed using *MYOD* as a positive control. Bisulfite-converted DNA was whole genome amplified and enzymatically fragmented prior to hybridization to Illumina HumanMethylation 450k BeadChip arrays, according to the manufacturer's instructions[330]. Arrays were scanned on an Illumina HiSeq 2000 using Illumina's iScan technology.

Data preprocessing

Raw data was extracted and overall quality was assessed using Illumina's GenomeStudio Data Analysis software. Further quality checks were performed using the R package minfi's 'qcreport' function[331]. All samples passed quality control assessment. Further exploration using multivariate analysis showed a clear distinction between tumour and normal samples by hierarchical clustering, even when using raw ß-values.

Data was normalized using beta mixture quantile normalisation (BMIQ)[332] implemented in the R package wateRmelon[333]. BMIQ addresses the difference in distributions seen between type I and II Infinium probes by using quantiles to normalize the type II probes to a distribution comparable to type I probes, using a ß-mixture model fit to the type I and type II probes separately[332].

Next, sites were filtered using the IMA[334] filter function IMA.methy450PP, where probes that had ≥10% of samples with detection P-value > 0.05 were excluded, as well as probes on the X and Y chromosomes, and probes containing known SNP sites. The results were as follows: 11847 sites with missing values were removed, 90401 sites contained SNPs and were removed, 10417 sites on the X and Y chromosomes were removed; 372912 sites were retained from the original 485577 sites.

Differential methylation analysis

Differential methylation analyses were conducted in CIMP+ vs. CIMP– tumours and by specific bacterial colonisation in tumour and normal samples, using the R package CpGassoc[335], which implements an analysis of variance (ANOVA) model. The Benjamini-Hochberg method was used for multiple testing corrections[300]. The magnitude of change between groups was estimated by calculating the difference in median beta values between the two groups. These analyses were conducted on the subset of CpG probes which had previously been shown to provide the most accurate results for the Illumina HumanMethylation450 BeadChips by comparison to bisulfite sequencing data[336]. Using this subset (294840 of the 372912 filtered probes) reduces the risk of false discovery, while maximizing the power to detect differential methylation.

Multivariate analysis

We used unsupervised clustering to explore patterns of methylation using recursively-partitioned mixture model (RPMM) clustering[337]. RPMM is a hierarchical, model-based clustering method that determines cluster membership by recursively comparing the model goodness-of-fit between n-class and n+1-class mixture models (starting with n=1). If, for example, the 2-class model fits the data better than the 1-class model, these classes are further split into 2 new classes and compared to the previous split in terms of model goodness-of-fit. Recursion continues until the algorithm arrives at the most parsimonious representation of the data. This procedure results in an estimate of the number of clusters, as well as the posterior probabilities of class membership[338].

RPMM-based analysis was conducted in R, and based on a script from Hinoue et al.[339]. The top 1% (3707) most variable probes by median absolute deviation was used as input for RPMM. Following assignment of class membership, samples within each cluster were ordered for heatmap-display using the R package seriation[340]. Heatmaps were generated using the function 'aheatmap' from the R package NMF[341].

In order to provide a visual representation of the underlying pattern (regarding similarities/dissimilarities) among samples, multidimensional scaling (MDS) was conducted using the mdsPlot function from the R package minfi[331] to visually explore

the relationship between samples and to investigate the effect of varying the number of CpG probes used for clustering.

Cohort integration and analysis

In order to evaluate methylation patterns in a broader context, our cohort was merged with a large publically available dataset of whole-genome methylation data (also run on HumanMethylation450 BeadChips), which was generated from 361 colorectal adenocarcinomas.

Raw data, together with clinical sample annotations, were downloaded from The Cancer Genome Atlas (TCGA). The merged dataset was normalized and filtered and quality control (QC) was conducted using the 'qcreport' function in minfi, and all samples passed QC. Samples were normalized using BMIQ normalization, as previously described. A filtering step was applied to the BMIQ-normalised data based on the detection p-value of probes using the R package wateRmelon, with default settings: sites where more than 5% of samples had a beadcount < 3; sites with a detection p-value > 0.05; samples where > 1% of sites had a detection p-value > 0.05; and sites where > 1% of samples had a detection p-value > 0.05 were removed.

The TCGA dataset consists of several different batches (samples processed separately). Unfortunately, batch identifiers are not available for these samples. Notably though, given that 85% of the TCGA cohort is Caucasian, while the majority of our cohort is mixed ancestry, batch correction could erase biologically relevant differences between our cohort and the TCGA cohort. However, there is a possibility that batch-specific variation remains in the dataset. Nevertheless, hierarchical clustering (Euclidian distance, complete linkage) showed a clear distinction between tumour and normal samples, irrespective of cohort, except for four tumour samples from our cohort, which clustered closely with the normal samples (10T, 15T, 1T, 4T).

Ingenuity Pathway Analysis (IPA)

IPA was used to conduct pathway-level analysis of probes associated with the phenotype of interest at a p-value ≤ 0.05. IPA is described in more detail in Chapters 7 and 8.

Results

Detection of *MLH1* methylation by methylation-specific PCR

Methylation-specific PCR of the promoter region of the *MLH1* gene was conducted on DNA extracted from tumour, normal and blood samples of 30 patients, including all of patients for whom array-based DNA methylation analysis was performed.

Two patients, 44 and 63, had *MLH1* promoter methylation in tumour samples, while the remaining 28 patients did not have *MLH1* promoter methylation in tumour, normal or blood samples, as judged by quantification against the *MLH1* standard curve. Three of the 84 samples (tumour, normal and blood for the 28 negative patients) failed QC.

Characterisation of CRC methylation by multivariate analysis

Whole-genome profiling was conducted for 24 pairs of tumour and matched normal samples using Illumina HumanMethylation450k BeadChips. Nineteen of these pairs were also run on Affymetrix Gene 1.0 ST arrays for transcriptomic profiling (See Chapters 5, 7 and 8).

Multidimensional scaling of the 1000, 10000 and 100000 most variable probes across tumour and normal samples were conducted to visually examine the relationship between samples, and to gauge the role of the number of probes used on the clustering outcome (Figure 4).

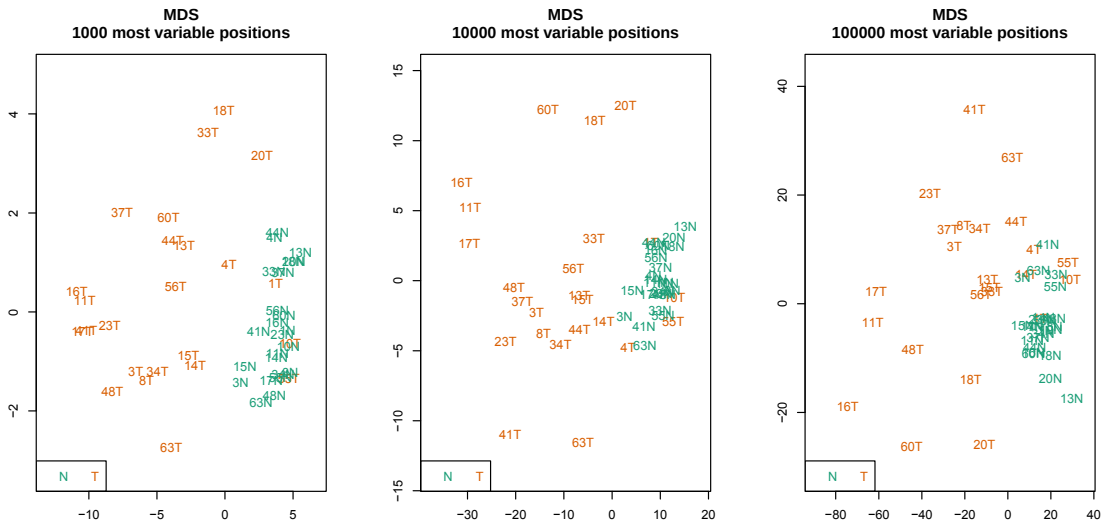


Figure 4. MDS of the 1000 (left), 10000 (middle) and 100000 (right) most variable probes across tumour and normal samples, using BMIQ normalized data. T: Tumour samples (orange); N: adjacent normal samples (green).

The most striking feature is the high degree of similarity between normal samples, which excludes the majority of tumour samples. There was no appreciable improvement in clustering when increasing the number of probes used, so the top 1% (N=3077) most variable probes (by median absolute deviation) was therefore arbitrarily selected for RPMM-based clustering. Due to the size of the overall dataset, the smaller subset was favoured to limit the computational power required to execute downstream analyses.

RPMM-based clustering of the top 1% most variable probes revealed four methylation clusters across tumour samples, with varying levels of CpG island methylation (Figure 5).

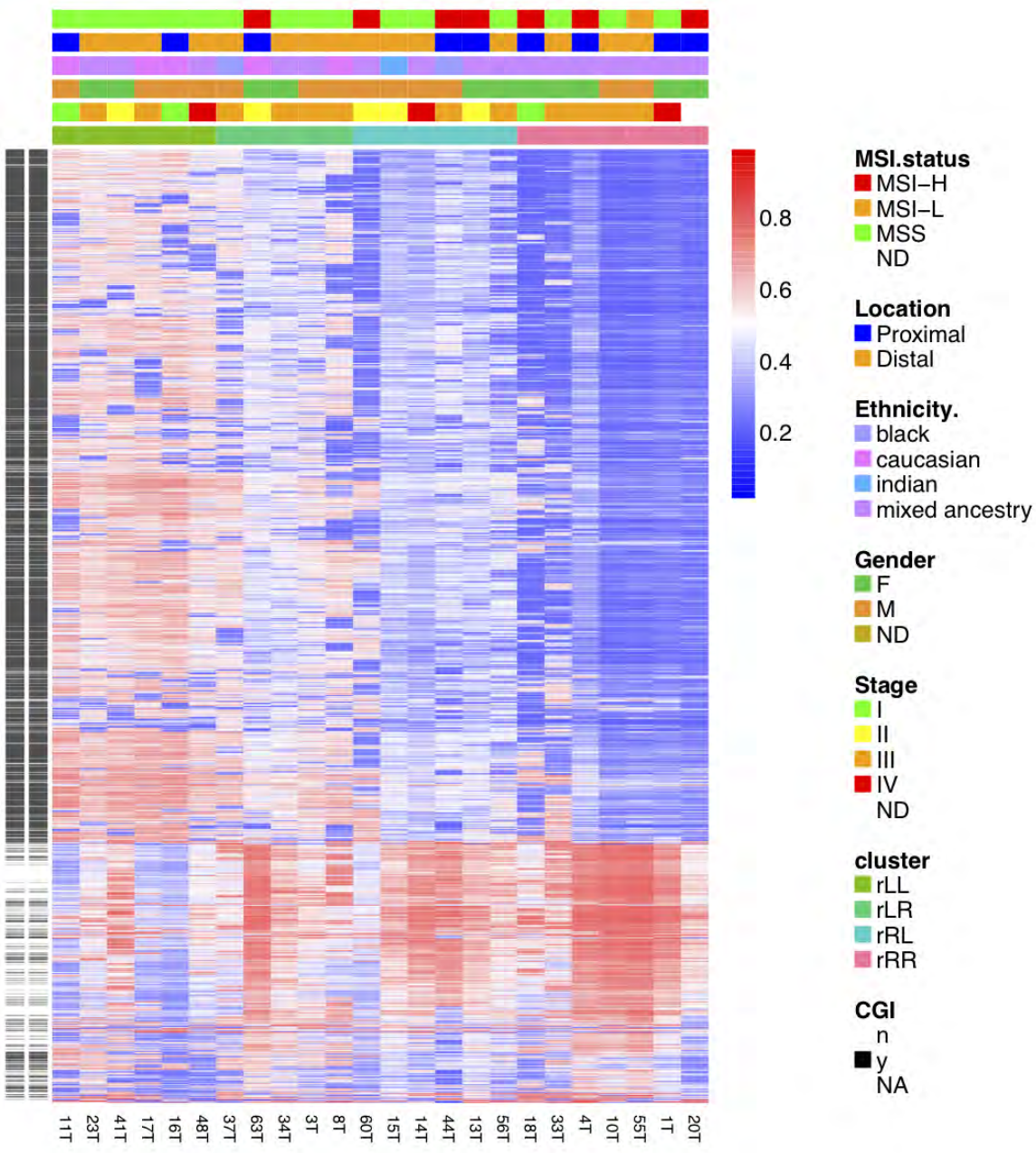


Figure 5: RPMM based clustering of the top 1% (N=3077) most variable methylation probes for tumour samples in our cohort. The row annotation on the left of the heatmap indicates whether the probe interrogates a CpG island (CGI), black=CpG island, white=non-CpG island. The scale on the right of the heatmap indicates beta values (0–1).

The relationship between individual tumour samples for the top 1% most variable probes was also explored by multidimensional scaling (Figure 6), which confirmed the RPMM-derived clusters.

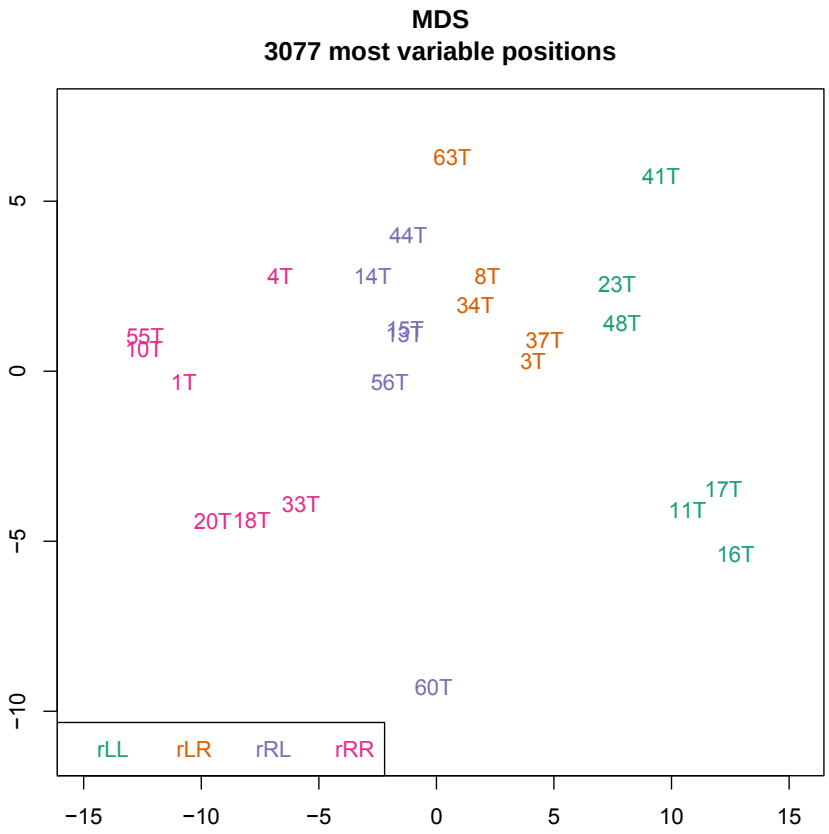


Figure 6: MDS of the top 1% (3707) most variable probes by MAD. Samples have been coloured according to the RPMM clusters to which they belong.

Most of the MSI-H samples fell into the two RPMM clusters with the least amount of CpG island methylation; however four of the five MSI-H samples in these two clusters are of hereditary origin (confirmed HNPCC)—these are 13T, 18T, 20T, 4T. Although the role of CpG island methylation across the genome is unclear in HNPCC samples, the majority of studies indicate that *MLH1* methylation does not play an appreciable role in the pathogenesis of these cancers[265,267–269,342]. Meanwhile, there were several samples that displayed a relative increase in CpG island methylation but no MSI in the rLL and rLR clusters.

Determining patterns of methylation alongside a large external cohort

In order to address the limitations of our cohort (i.e. small sample size and lack of CIMP status annotation), tumour samples from our cohort were merged with the TCGA cohort (N=361) for downstream analysis. This allowed us to a) increase sample size and

to b) to make inferences regarding CIMP status for our cohort based on the clustering alongside the TCGA samples, which had available molecularly determined CIMP classifications.

Of the 353 adenocarcinoma samples that had site of disease information, 25.5% were rectal, 46.7% proximal, and 27.7% distal. Regarding patient ethnicity (of whom 324 had available information), 85.5% were Caucasian, while 10.8% were black or African American, and 3% were Asian. Gender was roughly equal, with 45% female and 55% male. Regarding MSI status (where 347 cases had available information), 69% were MSS, 15.3% were MSI-H, and 15.6% were MSI-L. Meanwhile, 343 cases had available CIMP classifications (by MSP): 9% were classified as CIMP-H, 48.8% as CIMP-L and 31.2% as CIMP–stable. This cohort was therefore enriched for Caucasian patients, cancers of the proximal colon, and CIMP+ cancers.

RPMM-based clustering of the top 1% (N=4614) most variable probes across the 361 (TCGA) plus 24 (our cohort) adenocarcinoma samples resulted in four clusters (rLL, rLR, rRL and rRR), Figure 7.

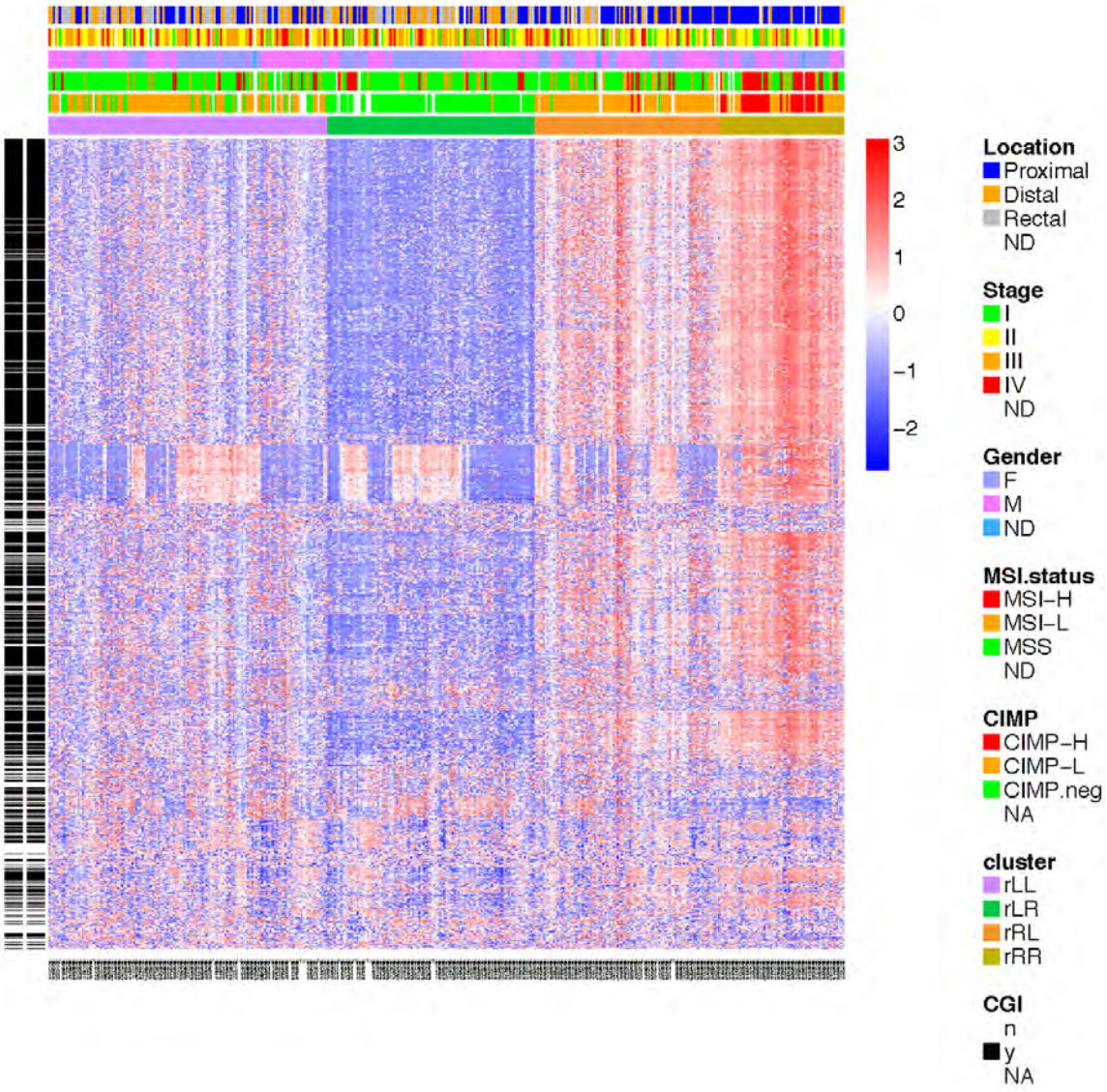


Figure 7: RPMM based clustering of the top 1% (N=4614) most variable methylation probes for tumour samples in the merged cohort (N=385). RPMM-based clustering of the top 1% (4614 probes). The row annotation on the left of the heatmap indicates whether the probe interrogates a CpG island (CGI), black=CpG island, white=non-CpG island. The scale on the right represents row-scaled beta values (scaling was done to assist visualization across a large cohort of samples). The legend categories on the right are in the same order as the row annotations at the top of the graph.

The rRR cluster displayed the highest level of CpG island methylation, followed by progressively less frequent and less intense methylation of CpG islands in the rRL, rLL and rLR clusters. The rRR, rRL, rLL and rLR clusters most likely correspond to the CIMP-H, CIMP-L, cluster 3 and cluster 4 of Hinoue et al. However, both the rRL and the rLL clusters are considered to be CIMP-L here, based on the enrichment of CIMP-L

samples in both these clusters; on the other hand Hinoue et al. referred to one of these as CIMP-L and the other as CIMP-stable.

A summary of descriptive characteristics by cluster is presented in Table 4.

Table 4: Summary of descriptive characteristics by RPMM cluster

	rLL (N=136)	**rLR (N=100)**	**rRL (N=90)**	**rRR (N=60)**
Gender (ND=4)				
Female	41%	49%	38%	57%
Male	59%	51%	62%	43%
Ethnicity (ND=38)				
Caucasian	82.6%	81.5%	74%	86%
Black	11.6%	7.6%	14.3%	8.6%
Mixed ancestry	4.1%	8.7%	3.9%	0%
Indian	0.8%	0%	0%	0%
Asian	0.8%	2.2%	7.8%	5.2%
BMI (ND=105)				
< 25	38.5%	24.3%	36%	36%
≥ 25	61.5%	75.7%	64%	64%
Stage (ND=19)				
I/II	42.5%	42.7%	61.6%	69%
III/IV	57.5%	57.3%	38.4%	31%
Site of disease (9=ND)				
Proximal colon	34.3%	24.5%	62.5%	86%
Distal colon	32.8%	43.9%	21.6%	12.3%
Rectum	32.8%	31.6%	15.9%	1.8%
MSI status (ND=15)				
MSI-H	9.2%	12.2%	10.6%	47.4%
MSI-L	12.2%	13.3%	22.4%	12.3%

MSS	78.6%	74.5%	67%	40.4%
CIMP status (ND=44)				
CIMP-H	0%	0%	2.5%	57.9%
CIMP-L	72.3%	1.2%	95%	42.1%
CIMP-stable	37.7%	98.8%	2.5%	0%

Percentages were calculated based on the cases with available information in that category; percentages may not add up due to rounding. ND: indicates the number of cases for whom that particular information was not available.

There was little variation in gender, ethnicity and BMI across the four clusters. As anticipated, the distribution of CIMP-H and CIMP-L samples—defined by MSP of a CIMP marker panel—varied greatly by cluster: the rRR cluster was heavily enriched for CIMP-H samples, which occurred at a frequency of 57.9% and 94% of all CIMP-H samples were present in this group; CIMP-L samples occurred at frequencies of 72.3% and 95% in the rLL and rRL clusters, while 98.8% of the rLR cluster consisted of CIMP-stable samples. As expected from the literature, CIMP+ rRL and rRR clusters were enriched for MSI+ cancers of the proximal colon, with notable enrichment of MSI-H samples and female gender in the rRR cluster. Regarding CIMP status according to CRC stage, early stage CRCs (I/II) were significantly more frequent in the CIMP-enriched rRL and rRR clusters compared to the CIMP-poor rLL and rLR clusters (Fisher's exact p-value=0.0001). By comparison, in the Hinoue et al. cohort, 41% of CIMP-L (corresponding to our rRL cluster) and 66% of CIMP-H (corresponding to our rRR cluster) CRCs compared to 41% of their cluster 3 (our rLR cluster) and 52% of CIMP-stable (our rLL cluster) CRCs were stage I/II[317]. The reason for this discrepancy is not clear. To our knowledge, a difference in CIMP status between early and late stage CRCs has not previously been reported in the literature, and is perhaps uncovered here due the use of whole-genome methylation profiles as opposed to five-gene CIMP panels. The basis for the difference seen here is not known–one hypothesis is that aberrant CpG island methylation is an early event in a subgroup of CRCs and that these regions are again demethylated with increasing tumour progression.

Regarding our cohort, 12 samples clustered with the predominantly CIMP-L rLL cluster, 8 with the CIMP-stable rLR cluster, and 4 with the CIMP-L rRL cluster. The CIMP-

stable cluster included three of five HNPCC patients, while two of three sporadic MSI-H patients clustered with the CIMP-L clusters.

Predicting CIMP status using an array-based marker panel

RPMM-based clustering of our samples alongside a larger CIMP-defined cohort provided a good indication of CIMP status for our cohort according to cluster membership. However, the Hinoue et al. array-based CIMP panel was utilized in an attempt to further refine the classification of samples.

As previously mentioned, Hinoue et al. published a CIMP-defining marker panel *(B3GAT2, FOXL2, KCNK13, RAB31,* and *SLIT1*) that identifies CIMP-H or CIMP-L tumours with 100% sensitivity and 95.5% specificity, with 2.4% misclassification using the condition of DNA methylation of three or more markers with a ß-value threshold of ≥ 0.1. They also defined a CIMP-H-specific marker panel (*FAM78A, FSTL1, KCNC1, MYOCD,* and *SLC6A4*) that identified CIMP-H tumours with 100% sensitivity and specificity, using the condition of DNA methylation of three or more markers with a ß-value of ≥ 0.1. However, because the majority of probes mapping to the CIMP-marker panel had ß-values of ≥ 0.1 in the present study a ß-value threshold of 0.1 could not be used as a meaningful threshold to define CIMP status. This could be due to differences in array chemistries between the HumanMethylation27 BeadChip used by Hinoue et al. (who used a ß-value threshold of 0.1) and the HumanMethylation450k BeadChip used in our study. CIMP status was therefore evaluated by RPMM-based subtyping of probes mapping to CpG islands of the Hinoue et al. CIMP+ marker panel of genes (Figure 8).

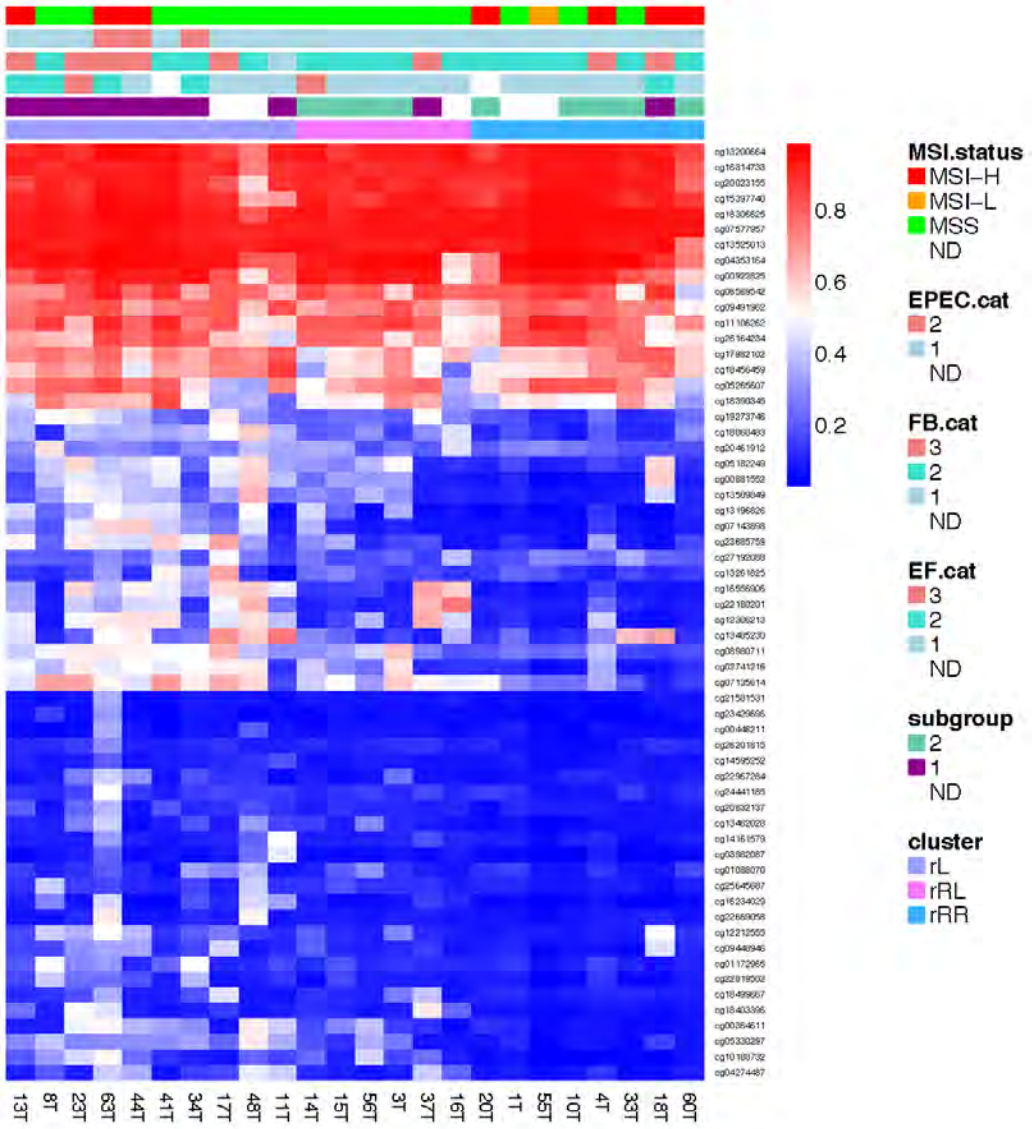


Figure 8: Predicting CIMP-status using an array-based marker panel. RPMM-based clustering of probes mapping to CpG islands in the Hinoue CIMP marker panel (*B3GAT2, FOXL2, KCNK13, RAB31*, and *SLIT1*). Samples in the rL cluster are considered to be CIMP+. The legend categories on the right are in the same order as the row annotations at the top of the graph. The scale on the right of the heatmap indicates beta values (0–1). EF: *E. faecalis*; FB: Fusobacterium.

Three clusters were obtained, two of which (rRL and rRR) more closely resembled each other compared to the rL cluster. An increased frequency of CpG island methylation was seen in the rL cluster–the ten samples in this cluster were therefore considered to be CIMP+. These ten samples (along with 37T, which was classified as CIMP-stable according to Figure 8) were the only samples where at least three of the five genes in the panel had gene-level methylation ß values of ≥ 0.3 (gene-level methylation ß values were obtained by calculating the median ß value of probes mapping to a particular gene).

The predictive ability of this panel is demonstrated in Figure 9, where RPMM clustering of the CIMP marker panel probes was performed on the merged dataset, which included 343 samples with MSP-confirmed CIMP status (CIMP-H, CIMP-L or CIMP-stable).

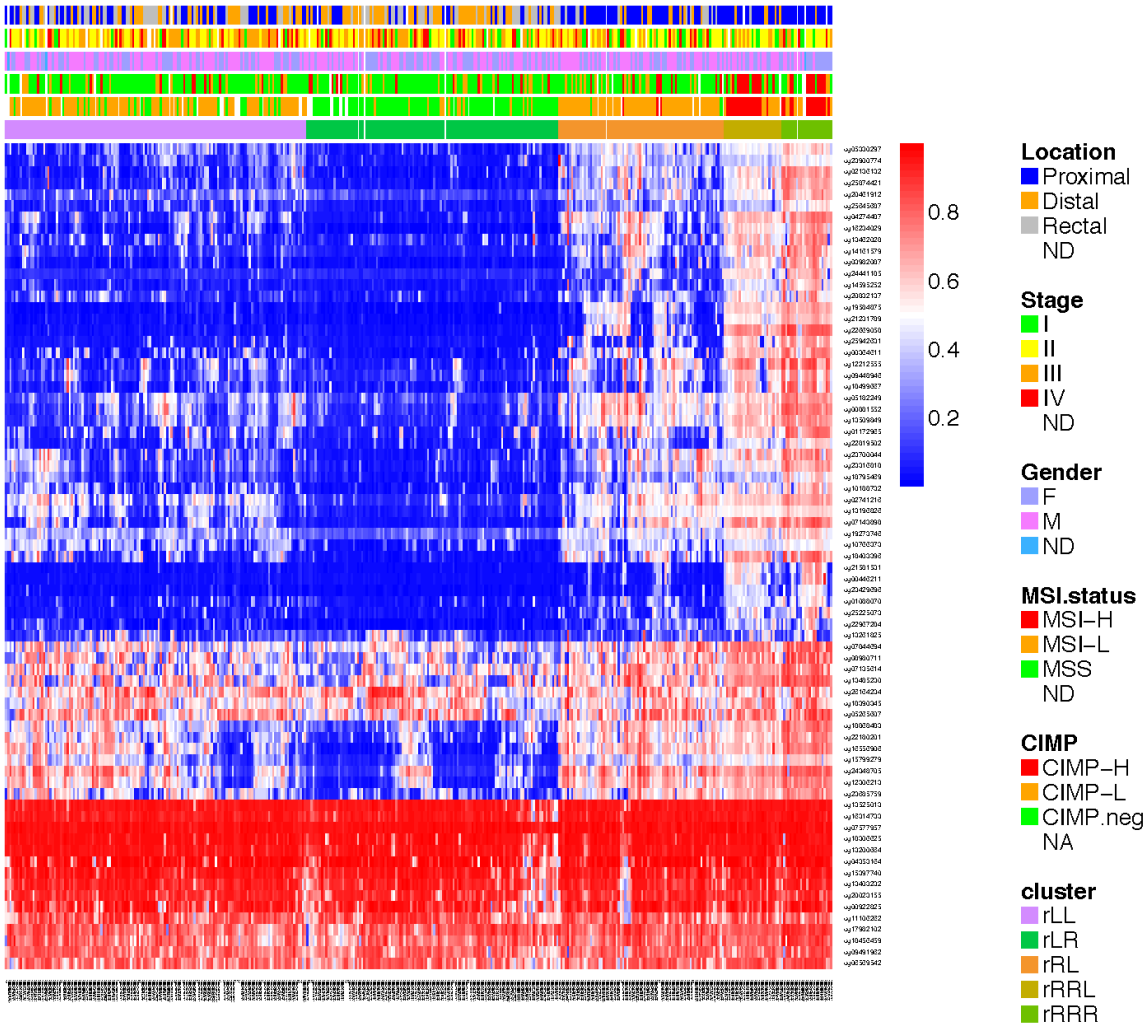


Figure 9: Predicting CIMP-status using an array-based marker panel after merging our cohort (N=24) and the TCGA cohort (N=361). RPMM-based clustering of probes matching to CpG islands in the Hinoue CIMP marker panel (B3GAT2, FOXL2, KCNK13, RAB31, and SLIT1) using a large cohort of CRC samples with confirmed CIMP status. The legend categories on the right are in the same order as the row annotations at the top of the graph. The scale on the right of the heatmap indicates beta values (0–1). F: female; M: male; CIMP.neg: CIMP-stable.

Five RPMM clusters were obtained: the rLL cluster was composed of 78% CIMP-L and 22% CIMP-stable samples; the rLR cluster was composed of 87.5% CIMP-stable and

12.5% CIMP-L samples; the rRL cluster was composed of 92.9% CIMP-L, 4.3% CIMP-H and 2.9% CIMP-stable samples; and the rRRL and rRRR clusters combined were composed of 68% CIMP-H and 32% CIMP-L samples. These results were comparable to the cluster membership obtained when using the top 1% most variable probes, with regards to the distribution of CIMP-H, CIMP-L and CIMP-stable samples (Figure 7).

We compared CIMP classifications when using a) the CIMP marker panel of genes (Figure 8) to b) the top 1% most variable probes of the merged cohort (N=385) (Figure 7). The two classifications were largely in agreement, except for six samples (15T, 14T, 37T, 16T, 3T and 4T) that were classified as CIMP-stable using the CIMP marker panel, but CIMP-L when using the top 1% most variable probes across the merged cohort. These six samples were classified as CIMP-stable on the basis that (with the exception of 37T) they all had less than three genes with ß values of ≥ 0.3.

When applying the Hinoue et al. CIMP-H panel (which specifically identifies CIMP-H tumours) to our cohort, only one sample, 63T, was identified CIMP-H (Appendix B, Figure 1); Sample 63T is MSI-H and displays hypermethylation of the *MLH1* promoter, as confirmed by PCR-based analysis, and it is therefore not surprising that this sample displayed a relative increase in CpG island methylation compared to the CIMP-L samples.

In summary, ten samples, for which methylation data was generated here, were classified as CIMP+ using the Hinoue et al. CIMP panel (13T, 8T, 23T, 63T, 44T, 41T, 34T, 17T, 48T, 11T), one of which (63T) is likely CIMP-H. The remaining 14 samples (14T, 15T, 56T, 3T, 37T, 16T, 20T, 1T, 55T, 10T, 4T, 33T, 18T, 60T) were classified as CIMP-stable. All samples classified as CIMP+ had gene-level methylation ß values of ≥ 0.3 in at least three of the five genes in the CIMP+ marker panel. The sample classification is largely supported by evaluation against cluster membership obtained by RPMM-clustering of the top 1% most variable probes, with the exception of six samples that display very low levels of CpG methylation, which were therefore classified as CIMP-stable.

Genome-wide methylation analyses

The CpGassoc package was used to detect probes associated with a) CIMP-status and b) specific bacterial infection.

Evaluation of CIMP-associated methylation

Analysis of variance by CIMP-status revealed 2172 probes (mapping to 1027 unique genes) that were associated with CIMP status in our cohort at a p-value cutoff of 0.05, and an absolute difference in median beta values between CIMP+ and CIMP– samples of 0.2 (which according to Illumina, can be detected with a FDR of < 1%). However, after multiple testing correction, only 5 probes were significantly associated with CIMP-status at an FDR cutoff of 0.05 and a Δ beta cutoff of 0.2; even at a more relaxed FDR of 0.25, only 124 (mapping to 94 unique genes) probes were associated with CIMP status. Nevertheless, an increase in CpG island methylation is evident in CIMP+ cancers, since 93% of the 2172 probes associated with CIMP-status mapped to CpG islands, and 99.8% showed increased methylation in CIMP+ vs. CIMP– samples. Moreover, of the 94 genes significantly associated with CIMP status at an FDR ≤ 0.25 and Δ beta ≥ 0.2, at least 23% (22/94) had previously been found to show increased CpG island methylation in CIMP+ vs. CIMP– tumours[305] (Table 3); of these, *BDNF*[306] and *KCNK13*[305] have been used in published CIMP-marker panels.

Table 3: Overlap in CpG probes (representing 22 genes) significantly associated with CIMP-status (FDR ≤ 0.25, Δ beta ≥ 0.2) between our cohort and Hinoue et al.[305].

Probe ID	Gene symbol	Gene Name	P value	FDR	Δ beta CIMP+ vs. CIMP–
cg22949149	AK5	Adenylate kinase 5	1.9E-04	0.23	0.21
cg12644885	ATP8B2	ATPase, Aminophospholipid Transporter, Class I, Type 8B, Member 2	6.8E-05	0.16	0.40
cg06816235	BDNF	Bone-derived neurotrophic factor	1.5E-04	0.20	0.23
cg14021073	BRSK2	BR serine/threonine kinase 2	1.1E-04	0.18	0.38
cg19202058	CBLN2	Cerebellin-2	1.2E-04	0.19	0.24
cg12208258	CDH2	Neural cadherin	2.4E-04	0.24	0.30

cg05878337	CTNND2	Catenin (cadherin-associated protein), delta 2	1.4E-04	0.20	0.22
cg03406394	FOXB1	Forkhead Box B1	4.3E-06	0.06	0.32
cg03848675	FOXF2	Forkhead Box F2	4.8E-06	0.06	0.23
cg21661027	GALR1	Galanin receptor 1	1.5E-04	0.20	0.23
cg23170850	GLB1L3	Galactosidase, Beta 1-Like 3	3.7E-05	0.13	0.29
cg05621343	GLB1L3		6.4E-05	0.15	0.35
cg05330297	KCNK13	Potassium channel, subfamily K, member 13	1.1E-04	0.18	0.23
cg00687686	NDRG4	NDRG family member 4	2.1E-05	0.11	0.39
cg05469759	NDRG4		2.1E-05	0.11	0.28
cg04190807	NDRG4		4.2E-05	0.14	0.27
cg04942472	NDRG4		7.7E-05	0.17	0.43
cg25924217	NPTX1	Neuronal pentraxin I	3.6E-05	0.13	0.29
cg20357628	PHACTR3	Phosphatase and actin regulator 3	1.0E-04	0.18	0.20
cg24530250	PHOX2A	Paired-like homeobox 2a	1.8E-04	0.22	0.32
cg22410750	PTPRZ1	Protein Tyrosine Phosphatase, Receptor-Type, Z Polypeptide 1	1.9E-05	0.11	0.28
cg18581445	SHD	Src homology 2 domain containing transforming protein D	3.8E-05	0.13	0.28
cg15173134	SLC18A2	Solute carrier family 18 (vesicular monoamine transporter), member 2	1.0E-04	0.18	0.38
cg21692194	SLC47A1	Solute carrier family 47 (multidrug and toxin extrusion), member 1	4.9E-05	0.14	0.28
cg09775582	TBX21	T-box 21	2.3E-05	0.11	0.42
cg16172408	ZNF804B	zinc finger protein 804B	2.0E-05	0.11	0.25
cg23219720	ZNF804B		2.4E-04	0.25	0.23

Probes mapping to CpG islands are noted as '1'; Δ beta = median beta difference between CIMP+ vs. CIMP– tumours; All except cg22410750 map to CpG island regions of the genome.

Evaluation of bacterially-associated methylation

For each bacterial species quantified in our cohort, differential host methylation analysis was conducted for a) high-level vs. low/no colonisation and b) positive- vs. negative-colonisation status in both tumour and normal samples, on the condition that each group had at least three samples. High- and low-level colonisation categories were determined in Chapter 4. The results for all comparisons made for tumour and normal samples are presented in Tables 4a and 4b respectively.

Table 4a: Summary of differential methylation analyses conducted for *tumour* samples showing the number of probes (that mapped to annotated genes) differentially methylated for each comparison made ($|\Delta \text{ beta}| \geq 0.2$).

Model specified	Sample number	$p \leq 0.05$	FDR ≤ 0.25	FDR ≤ 0.05
EF+ vs. EF-	7 vs. 15	315	0	0
FB-H vs. FB-L/N	7 vs. 17	403	7	0
afaC-H vs. afaC-L/N	6 vs. 18	62	0	0
afaC+ vs. afaC-	14 vs. 10	85	0	0
ClB+ vs. ClB-	7 vs. 17	331	0	0
ETBF-H vs. ETBF L/N	4 vs. 20	488	17	0
ETBF+ vs. ETBF-	14 vs. 10	215	0	0
EPEC+ vs. EPEC-	3 vs. 21	1019	141	62

EF: *E. faecalis*; FB: Fusobacterium.

Table 4b: Summary of differential methylation analyses conducted for *normal* samples showing the number of probes (that mapped to annotated genes) differentially methylated for each comparison made ($|\Delta \text{ beta}| \geq 0.2$).

Model specified	Sample number	$p \leq 0.05$	FDR ≤ 0.25	FDR ≤ 0.05
ETBF+ vs. ETBF-	14 vs. 10	2	0	0
ETBF-H vs. ETBF-L/N	4 vs. 20	4	0	0
ClB+ vs. ClB-	8 vs. 16	1	0	0
afaC+ vs. afaC-	13 vs. 11	3	0	0
afaC-H vs. afaC-L/N	3 vs. 21	9	0	0
EF+ vs. EF-	4 vs. 15	9	0	0
FB-H vs. FB-L/N	3 vs. 21	22	1	1

EF: *E. faecalis*; FB: Fusobacterium.

Only the comparison between EPEC+ and EPEC– CRCs yielded significant results at an FDR ≤ 0.05; for this comparison, 141 or 62 CpG sites were significantly methylated in association with EPEC colonisation at FDR cutoffs of 0.25 or 0.05, respectively. This is considerable given that the CIMP+ vs. CIMP– comparison produced only 124 CpG probes at an FDR cutoff of 0.25. EPEC-associated CpG sites with an FDR ≤ 0.05 are presented in Table 5. These results should however be interpreted with caution, given the unbalanced comparison (3 EPEC+ vs. 21 EPEC– samples) that is more susceptible to unequal variances between groups, as well as outliers in the smaller EPEC+ group.

Table 5: CpG sites significantly associated with EPEC colonisation in tumours (FDR ≤ 0.05, |Δ beta| ≥ 0.2) in alphabetical order by gene symbol. For the CpG island column 1=CpG island and 0=non-CpG island.

GeneID	Gene Symbol	CpG island	P value	FDR	EPEC+ vs. EPEC-
cg19778944	A4GALT	1	3.9E-06	1.4E-02	0.29
cg03940620	B3GAT1	1	8.3E-06	2.3E-02	0.22
cg19104475	CABP7	1	2.6E-08	1.0E-03	0.34
cg23614229	CACNA1G	1	3.5E-07	3.8E-03	0.26
cg16111791	CACNB4	1	1.6E-05	3.4E-02	0.21
cg07131655	CAMKV	1	9.0E-10	2.2E-04	0.23
cg14469019	CAPSL	0	2.6E-05	4.4E-02	0.25
cg00879447	CCDC85A	1	3.0E-05	4.8E-02	0.24
cg02299940	CHP2	1	9.5E-06	2.3E-02	0.21
cg12120327	COL2A1	1	8.8E-06	2.3E-02	0.28
cg07295964	CPLX2	1	2.2E-06	1.1E-02	0.40
cg04986675	CPQ	1	4.6E-06	1.6E-02	0.20
cg06531007	CR2	1	1.2E-05	2.8E-02	0.22
cg14557714	CTSL1	1	2.3E-05	4.0E-02	0.24
cg26267854	CXCL12	1	5.6E-08	1.3E-03	0.30
cg17267805	CXCL12	1	1.7E-06	9.0E-03	0.40
cg06048524	CXCL12	1	2.1E-06	1.0E-02	0.40
cg14948279	DCLK2	1	2.1E-05	3.7E-02	0.27
cg02934930	EPHB1	1	6.9E-06	2.0E-02	0.40
cg24183261	EYA1	1	7.2E-06	2.1E-02	0.24
cg02102020	FAM5B	1	8.4E-08	1.5E-03	0.26
cg20509780	FBLN1	1	2.8E-05	4.6E-02	0.32
cg14778074	GABRA2	1	1.6E-05	3.4E-02	0.20
cg10451078	GATA4	1	8.7E-07	5.8E-03	0.33
cg05412333	GATA4	1	1.7E-05	3.5E-02	0.23
cg17816394	GNG4	1	5.2E-06	1.8E-02	0.24
cg10154926	HAP1	1	1.5E-05	3.2E-02	0.22
cg05660795	IGFBP1	1	7.2E-07	5.4E-03	0.27
cg14093886	INHBA-	1	9.9E-07	6.0E-03	0.24
cg22063259	ISLR2	1	6.3E-07	5.4E-03	0.27
cg24766308	JAKMIP2	0	8.1E-06	2.3E-02	0.23
cg15811719	KCNJ6	1	1.7E-05	3.5E-02	0.20
cg27147871	LOC400891	1	3.0E-06	1.2E-02	0.27
cg10004574	LOC400891	1	9.7E-06	2.3E-02	0.22
cg05328197	MAPK4	1	9.0E-06	2.3E-02	0.34
cg25803927	MN1	1	3.4E-06	1.3E-02	0.38
cg23791592	MSI1	1	2.0E-05	3.7E-02	0.35
cg01425670	NEGR1	1	2.8E-05	4.6E-02	0.28
cg18834338	NEUROG3	1	2.0E-07	2.7E-03	0.25

cg05886671	NEUROG3	1	2.1E-05	3.7E-02	0.25
cg14005139	NKX2-1	1	2.0E-05	3.7E-02	0.24
cg03666741	NRIP3	1	5.1E-06	1.7E-02	0.37
cg07140751	P2RX5	1	4.0E-06	1.4E-02	0.27
cg26244952	PRDM6	1	2.3E-06	1.1E-02	0.31
cg15193171	PRDM6	1	1.9E-05	3.6E-02	0.28
cg20416031	PRKG1	1	3.0E-05	4.8E-02	0.23
cg23123909	RAB3C	1	1.6E-05	3.4E-02	0.32
cg08621203	RAET1G	1	7.3E-07	5.4E-03	0.31
cg05680085	RHBDD1	1	1.9E-05	3.6E-02	0.23
cg13364903	SCUBE3	1	1.1E-05	2.6E-02	0.24
cg19671026	SELV	0	3.9E-08	1.2E-03	0.29
cg18586919	SELV	0	1.1E-07	1.8E-03	0.30
cg03703707	SHC4	1	4.5E-09	3.6E-04	0.27
cg23637124	SHC4	1	6.0E-08	1.3E-03	0.25
cg22967284	SLIT1	1	3.0E-05	4.8E-02	0.22
cg21110939	SV2B	1	8.2E-06	2.3E-02	0.25
cg05003791	SYNGR1	1	2.2E-06	1.1E-02	0.26
cg00786658	TSHZ3	1	3.2E-05	4.9E-02	0.21
cg26226202	WNT9B	1	4.3E-09	3.6E-04	0.26
cg26217504	XKR5	1	2.1E-05	3.7E-02	0.51
cg14156581	ZFPM2	1	9.7E-06	2.3E-02	0.27
cg05629186	ZNF391	1	6.0E-07	5.4E-03	0.22

Of the 62 probes significantly associated with EPEC infection, 95% mapped to CpG islands and 100% displayed increased methylation in EPEC+ CRCs. Many of these genes have previously been found to be methylated in various cancers: *WNT9B* and *GNG4* are both members of the Wnt signaling pathway, and *GNG4* is specifically methylated in *BRAF* mutant CRCs[343]; *GATA4* is often methylated in CRCs and is proposed to have diagnostic potential[344]; epigenetic silencing of *CXCL12* has been shown to promote metastasis in breast[345], non-small cell lung[346], and colon cancer[347]; *FBLN1* is methylated in various cancers[348–350]; *CTSL1* is methylated in HPV+ vs. HPV- oral squamous cell carcinoma[351,352]; and *CACNA1G* is a member of the CIMP panel proposed by Weisenberger et al.[155,306].

Next, the degree of correspondence between EPEC-associated changes in gene methylation and downstream gene expression (using the results of differential gene expression analysis for EPEC+ vs. EPEC– samples, described in Chapter 8) was

assessed. The 1019 probes associated with EPEC- colonisation status ($p \leq 0.05$, $|\Delta \text{ beta}| \geq 0.2$) mapped to 701 unique genes, of which 35 also showed differential gene expression in an EPEC-specific manner ($p \leq 0.05$; Chapter 8), as shown in Table 6. The relatively low level of correspondence between genes differentially methylated and differentially expressed (35/701, 5%) is similar to the findings of Hinoue et al., who showed that 7.3% of genes that showed hypermethylation ($|ß| > 0.20$) in CIMP-H tumours also showed at least a two-fold reduction in gene expression[317].

Table 6: Genes with altered methylation and gene expression in association with EPEC colonisation ($P \leq 0.05$; $|\Delta \text{ beta}| \geq 0.2$) in alphabetical order. Consistent, yes: direction of change in gene expression consistent with the expected downstream effect of increased or decreased methylation.

Gene symbol	Gene name	P (gene expression)	EPEC + vs. EPEC – (FC)	P (methylation)	EPEC + vs. EPEC – (Δ beta)	consistent
C12orf68	chromosome 12 open reading	2.2E-02	-1.49	1.6E-02	0.23	yes
C6orf223	chromosome 6 open reading	4.5E-03	1.87	4.0E-02	-0.27	yes
C7orf50	chromosome 7 open reading	1.7E-02	-1.45	1.9E-02	0.33	yes
CDK6	cyclin-dependent	2.7E-02	1.95	4.9E-03	-0.25	yes
CLDN10	claudin 10	1.5E-02	1.71	1.2E-02	-0.22	yes
CPPED1	calcineurin-like phosphoesterase	2.2E-02	-1.59	3.5E-03	0.25	yes
CRMP1	collapsin response	8.6E-03	-1.40	3.6E-03	0.21	yes
EFNB2	ephrin-B2	1.3E-02	-1.89	4.4E-02	0.23	yes
EPDR1	ependymin related protein 1	2.8E-02	-1.44	1.2E-02	0.46	yes
EPHB1	EPH receptor B1	2.3E-02	-1.47	2.6E-04	0.31	yes
FAM13A	family with sequence	1.4E-02	-1.93	7.1E-04	-0.21	no
HACE1	HECT domain and ankyrin	2.7E-02	1.57	2.2E-02	0.27	no
HDAC9	histone deacetylase 9	2.7E-02	-1.49	3.4E-02	0.32	yes
LRP11	low density lipoprotein	2.5E-02	-1.31	4.8E-02	0.75	yes
MCMDC2	minichromosome maintenance	2.6E-02	-1.55	1.9E-02	0.28	yes
MN1	meningioma (disrupted in	1.1E-02	-1.37	2.2E-02	0.32	yes

MYEF2	myelin expression	3.6E-03	-2.40	2.1E-02	0.26	yes
NAV1	neuron navigator 1	8.9E-03	-1.49	2.1E-02	0.24	yes
NFASC	neurofascin	1.5E-02	-1.47	4.4E-03	0.25	yes
NOG	noggin	2.3E-02	-1.27	6.1E-03	0.27	yes
OSBPL5	oxysterol binding protein-	7.4E-04	-1.54	3.1E-02	0.22	yes
QPCT	glutaminyl-peptide	2.3E-02	-2.18	3.5E-03	0.26	yes
RBFOX1	RNA binding protein, fox-1	2.5E-02	-1.47	1.2E-02	0.29	yes
RPS6KA2	ribosomal protein S6	1.3E-02	-1.47	1.5E-02	0.37	yes
SCUBE3	signal peptide, CUB domain,	3.7E-03	1.71	6.4E-03	0.22	no
SOX1	SRY (sex determining	9.2E-04	-1.76	2.4E-02	0.31	yes
SPAG16	sperm associated	1.2E-02	-1.33	2.6E-02	0.23	yes
SPATS1	spermatogenesis associated,	3.9E-03	-1.42	2.0E-02	0.23	yes
TLX2	T-cell leukemia homeobox 2	1.9E-02	-1.36	2.0E-02	0.40	yes
WSCD1	WSC domain containing 1	7.6E-03	-1.60	2.2E-04	0.29	yes
ZNF287	zinc finger protein 287	1.8E-03	-1.55	3.2E-02	0.24	yes
ZNF419	zinc finger protein 419	4.7E-03	-1.43	9.0E-03	0.35	yes
ZNF546	zinc finger protein 546	3.6E-04	-1.78	3.5E-02	0.32	yes
ZNF606	zinc finger protein 606	7.9E-03	-1.33	1.6E-03	0.33	yes
ZNF83	zinc finger protein 83	4.5E-03	-1.82	3.7E-05	0.32	yes

The 35 genes presented in Table 6 encompassed 51 probes on the methylation array, which primarily mapped to CpG islands (36/51, 71%). Regarding gene regions, 35 probes (69%) mapped to TSS1500, TSS200, 5'UTR or 1st Exon gene regions, while 16 probes (31%) mapped to gene body or 3'UTR regions. Detailed methylation analysis and annotation results for the 35 genes with concomitant changes in gene expression are listed in Appendix B, Table 1. In cases where more than one probe per gene showed differential methylation, median p- and Δ beta-values were calculated across probes. Given the enrichment of probes mapping to CpG islands in the promoter region (including TSS1500, TSS200, 5'UTR and 1st Exon), the resulting gene-level

methylation values are considered to be reasonable predictors of promoter methylation and hence gene silencing.

Thirty-two of 35 genes (91%) displayed gene expression changes that were consistent with the underlying methylation change (Table 5), and 31 genes (89%) had increased methylation in EPEC+ samples.

Pathway analysis was conducted using the 1019 probes associated with EPEC colonisation in CRCs ($p \leq 0.05$) and, since fold-change values are not applicable to methylation data, only p-values were used as input. The top 10 most significant canonical pathways associated with EPEC colonisation are listed in Table 7.

Table 7: Top 10 IPA canonical pathways associated with EPEC+ CRCs. The Ratio indicates the proportion of genes significantly associated with EPEC compared to the full list of genes in each pathway.

Ingenuity Canonical Pathways	-log(p-value)	Ratio
Axonal Guidance Signaling	7.93	9.03E-02
Ephrin Receptor Signaling	5.34	1.09E-01
Glutamate Receptor Signaling	4.85	1.75E-01
Neuropathic Pain Signaling In Dorsal Horn Neurons	3.36	1.10E-01
Ephrin B Signaling	3.21	1.23E-01
CREB Signaling in Neurons	2.81	8.19E-02
Hepatic Fibrosis / Hepatic Stellate Cell Activation	2.66	7.61E-02
BMP signaling pathway	2.47	1.05E-01
Embryonic Stem Cell Differentiation into Cardiac Lineages	2.45	3.00E-01
G Protein Signaling Mediated by Tubby	2.38	1.52E-01

Notably, the pathways *Ephrin Receptor Signaling*, *Ephrin B Signaling* and *BMP signaling pathway* are all major players in intestinal stem cell maintenance and differentiation[353]. This is relevant to CRC since crypt stem cells are thought to be the cells-of-origin of intestinal cancer[113]. Moreover, EPEC has been shown to enter colonic crypts in normal colonic mucosa co-cultured with EPEC[9], which supports to the possibility of EPEC-dependent epigenetic manipulation of intestinal stem cells.

The top ten upstream regulators most significantly associated with EPEC+ CRCs are presented in Table 8.

Table 8: Top 10 IPA upstream regulators associated with EPEC+ CRCs.

Upstream Regulator	Molecule Type	p-value of overlap
REST	transcription regulator	4.08E-10
Calmodulin	group	9.04E-09
TWIST1	transcription regulator	3.49E-07
POU4F1	transcription regulator	8.99E-07
ESR2	ligand-dependent nuclear	1.11E-06
CTNNB1	transcription regulator	3.61E-06
FGF8	growth factor	3.85E-06
HTT	transcription regulator	5.52E-06
decitabine	chemical drug	6.13E-06
NANOG	transcription regulator	1.02E-05

Amongst these, the transcription factors β-catenin (CTNNB1) and NANOG both have critical roles in stem cell homeostasis, which is in agreement with the canonical pathways results. Interestingly, the DNA methyltransferase inhibitor Decitabine, which is used in epigenetic cancer therapy[354], is also listed as an upstream regulator in EPEC+ CRCs, which further alludes to a role of EPEC in epigenetic manipulation of host cells.

Discussion

Whole-genome methylation patterns in CRC and adjacent normal mucosa samples were evaluated using unsupervised RPMM clustering. By merging our smaller cohort with a large, well-annotated, publically available dataset, we were able to draw inferences across a large pool of samples, which facilitated methylation subtyping of our samples with greater statistical confidence. Four clusters with varying levels of CpG island methylation were obtained from the merged dataset; two clusters were dominated by CIMP-L samples, another by CIMP-H samples, while the fourth cluster was almost entirely composed of CIMP-stable samples. Clinically relevant features included the enrichment of MSI-H samples in the CIMP-H cluster, the predominance of proximal cancers, and stage I/II cancers in the CIMP-H and one of the CIMP-L clusters.

Samples were classified as CIMP+ or CIMP– using a published array-based marker panel of CIMP-status, and all samples classified as CIMP+ had gene-level methylation ß-values of ≥ 0.3 in at least three of the five genes in the CIMP+ marker panel. The marker panel-based classification of CIMP is largely supported by comparison against cluster membership obtained by RPMM-clustering of the top 1% most variable probes for the merged dataset.

Analysis of variance by CIMP-status revealed 2172 probes associated with CIMP-status ($p \leq 0.05$, $|\Delta \text{ beta}| \geq 0.2$), 93% of which mapped to CpG islands and 99.8% of which showed increased methylation in CIMP+ vs. CIMP– samples; moreover, of the 94 genes associated with CIMP status at an FDR ≤ 0.25, at least 23% (22/94) had previously been found to show increased CpG island methylation in CIMP+ vs. CIMP– tumours[305]; of these, *BDNF*[306] and *KCNK13*[305] have been used in published CIMP-marker panels.

Cancer-specific CpG island methylation of specific genes has been associated with a distinct clinicopathological and molecular features of CRC. However, with the exception of a few key genes, including DNA repair genes (*MLH1*), Wnt antagonists (SFRPs) and tumour suppressors (*CDKN2A*)[355], the mechanistic relevance of CpG island methylation of these genes remains largely obscure. One obvious hurdle in delineating the role of gene methylation in CRC is the often poor correlation between promoter methylation and transcriptional silencing. For instance, we demonstrate in Chapter 7 that a specific transcriptomic subtype of CRC is associated with increased promoter-methylation (and therefore presumably silencing) of several WNT pathway inhibitor genes; however, these effects do not appear to translate to decreased transcription of these genes.

Apart from the etiological contribution of DNA methylation in CRC, another clinically relevant question, which is perhaps easier to address is: to what extent does the epigenetic landscape change from tumour initiation to progression and metastasis, and what are the pathological, diagnostic and prognostic implications thereof? In the present study, a significant difference in the proportion of early (I/II) vs. late (III/IV) stage CRCs was identified between methylation clusters with higher (rRL, rRR) or lower (rLR, rLL) levels of CpG island methylation (p=0.0001). This difference has not

previously been reported and requires further study to uncover the basis and putative clinical implications thereof. There is currently great interest in identifying DNA methylation biomarkers for CRC risk, diagnosis, treatment and prognosis[355,356]. DNA methylation biomarkers are potentially more robust compared to gene expression biomarkers due to the cancer-specific methylation patterns found in certain genes; indeed and a handful of biomarkers including *SEPT9* (ColoVantage®) and *VIM* (ColoSure™) are already commercially available[356].

This is the first study to investigate the putative association between specific CRC-associated pathogens and genome-wide DNA methylation in paired CRC tissue samples. Although no specific methylation changes were found for the majority of bacteria examined here, EPEC-infected tumours displayed significantly altered patterns of host methylation in 62 CpG sites (FDR≤0.05). Further, of the 35 genes that were differentially methylated and differentially expressed in EPEC+ samples (p≤0.05), 91% displayed gene expression changes that were consistent with the underlying methylation change and 89% had increased methylation in EPEC+ samples. Moreover, two of the three EPEC+ samples had confirmed *MLH1* promoter methylation by MSP, and all three EPEC+ samples were predicted to be CIMP+. Although certainly intriguing, these results should be interpreted with caution, since we only had three EPEC+ samples available in this study. Maddocks et al. previously demonstrated that EPEC infection can cause decreased expression of *MLH1 in vitro*; however, this observed reduction apparently occurred at the protein level, not at the mRNA level[10]; this apparent discrepancy between our clinical data showing EPEC-associated epigenetic inactivation of *MLH1*, and the *in vitro* data of Maddocks et al. showing protein-level depletion of MLH1, requires further clarification.

At the pathway level, EPEC+ CRCs show methylation of genes belonging to pathways related to stem cell homeostasis and, although we have not proven a causal link between EPEC and aberrant methylation of host DNA here, the fact that EPEC can colonise stem-cell rich intestinal crypts[9] increases the probability of these effects being mediated by EPEC. Future studies should validate our findings in a larger cohort of EPEC+ CRCs and in adjacent normal samples as well as in *in vitro* infection models.

Chapter 7: Specific bacterial infection, inflammation, and DNA and protein damage responses underlie a distinct genomic subtype of CRC

Abstract

In this chapter, unsupervised clustering of gene expression data together with integrative network analysis of whole-genome gene expression and methylation data was used to investigate a) possible CRC subtypes in our cohort (which consisted of 19 adenocarcinomas) and b) the biological basis for these subgroups. In addition, the same workflow was applied to a well-defined publically available CRC gene expression dataset (GSE13294) of 155 colorectal adenocarcinomas to evaluate the relevance of the findings in our cohort to a larger CRC cohort.

A subtype of CRC that is associated with a relative increase in the frequency of colonisation by *E. faecalis* and high-level colonisation with Fusobacterium as well as CpG island methylation and microsatellite instability (MSI) was identified. Comparison of our classification of the larger cohort (GSE13294) to previous classifications of this cohort revealed a substantial overlap with the de Sousa et al.[357] CCS2/3 subtypes and the inflammatory subtype of Sadanandam et al.[358]. Pathway-level analyses of genes differentially expressed between the two CRC subtypes identified here revealed an increased response to DNA and protein damage, infection, inflammation and proliferation in one of these subtypes. Various genes and pathways linked to increased metastasis and CRC progression were also highlighted in this subtype.

Introduction

Established CRC subtypes

Colorectal cancers (CRCs) are frequently classified by the type of genomic instability encountered, with chromosomal instability (CIN) and microsatellite instability (MSI) being more frequent in the distal and proximal colon, respectively. Recent studies have refined CRC subtype classification using unsupervised clustering of whole-genome

transcriptome, methylome and copy number variation data to describe clinically and biologically relevant CRC subtypes[104,357,358].

In a comprehensive molecular characterisation of CRCs (encompassing whole-genome copy number variation, methylation, miRNA, mRNA expression and mutation data) The Cancer Genome Atlas (TCGA) divided CRCs into hypermutated (around 16% of CRCs) and non-hypermutated, which could be further subdivided into two classes by mRNA expression data and other molecular data[104]. Hypermutated CRCs had near-diploid genomes and were highly enriched for *MLH1* hypermethylation, CpG island methylator phenotype (CIMP) and *BRAF* (V600E) mutations. Interestingly, in the non-hypermutated tumours copy-number variation, gene expression and methylation data were very similar between different sites (proximal, distal, rectal). Furthermore, only the hypermutated class showed a high agreement between mRNA expression- and methylation-derived clusters, while none of the non-hypermutated classes showed good agreement between copy number variation methylation or mRNA-derived clusters. Four methylation subgroups were identified: CIMP-H, CIMP-L and two CIMP stable clusters that were predominantly non-hypermutated and showed some difference by tumour location and *KRAS* mutation status[104].

De Sousa et al. identified three CRC subtypes based on unsupervised clustering of gene expression data: one is characterised by a high frequency of proximal, MSI- and CIMP-positive CRCs (CSS2), another by distal CIN-positive CRCs (CCS1) and a third, novel category, that is heterogeneous with regards to MSI- and CIMP-status, as well as site of disease (CSS3)[357].

Sadanandam et al. applied a similar method to de Sousa to classify 445 CRCs, but instead defined five CRC subtypes, each of which show similarities to different cell types along the colonic crypt, with concomitant degrees of 'stemness' and Wnt signaling[358]. The five transcriptional CRC subtypes were named as goblet-like, enterocyte, transit-amplifying, inflammatory or stem-like, according to the degree of correspondence to a published gene signature that discriminates terminally differentiated cells (at the normal colon crypt top) from crypt stem cells (at the normal crypt base). The enterocyte- and goblet-like subtypes were significantly associated with

the top of the crypt, and the stem-like type was significantly associated with signatures of the base of the crypt. Meanwhile, the inflammatory subtype was not associated with the position in the crypt, and samples in the transit-amplifying type could be divided between the bottom and the top of the crypt. In relation to MSI status, 94% of the inflammatory subtype samples were MSI+, whereas 14% and 33% of the transit-amplifying and stem-like subtypes were MSI+, respectively[358].

Interestingly, both de Sousa et al. and Sadanandam et al. demonstrated that their CRC subtypes (4/5 in the case of Sadananadam et al.) could be matched to existing CRC cell lines and that these cell lines maintained their original classification in mouse xenografts, which may imply a cell-type-specific etiological basis[357,358], propagated via epigenetic modifications.

Given the discrepancy in classification between the de Sousa et al. and the Sadananadam et al. studies (five vs. three subtypes), these groups compared and reconciled their results, stating that although specifying k=3 provided the most robust clustering solution, k=5 had similar stability and was therefore also valid[359]. By swapping their datasets to compare the two different subgroup classification schemes, they found that the Sadanandam et al. enterocyte and transit amplifying types could be collapsed into the CIN+ CSS1 type of de Sousa et al.; the inflammatory and goblet type could be collapsed into the MSI+/CIMP+ CSS2 type of de Sousa et al.; and the stem-like type matched the CSS3 type of de Sousa et al.[359].

Numerous additional studies have described similar CRC subtype classification systems based on gene expression and/or other whole-genome data, which will not be discussed here[305,360–362].

Data analysis workflow

Unsupervised clustering of gene expression data together with integrative network analysis of whole-genome gene expression and methylation data was used to investigate a) possible CRC subtypes in our cohort (which consisted of 19 adenocarcinomas) and b) the biological basis for these subgroups. In addition, the same workflow was applied to a well-defined publically available CRC gene expression dataset (GSE13294) of 155

colorectal adenocarcinomas to evaluate the relevance of our results in a larger cohort. A summary of this workflow is presented in Figure 1.

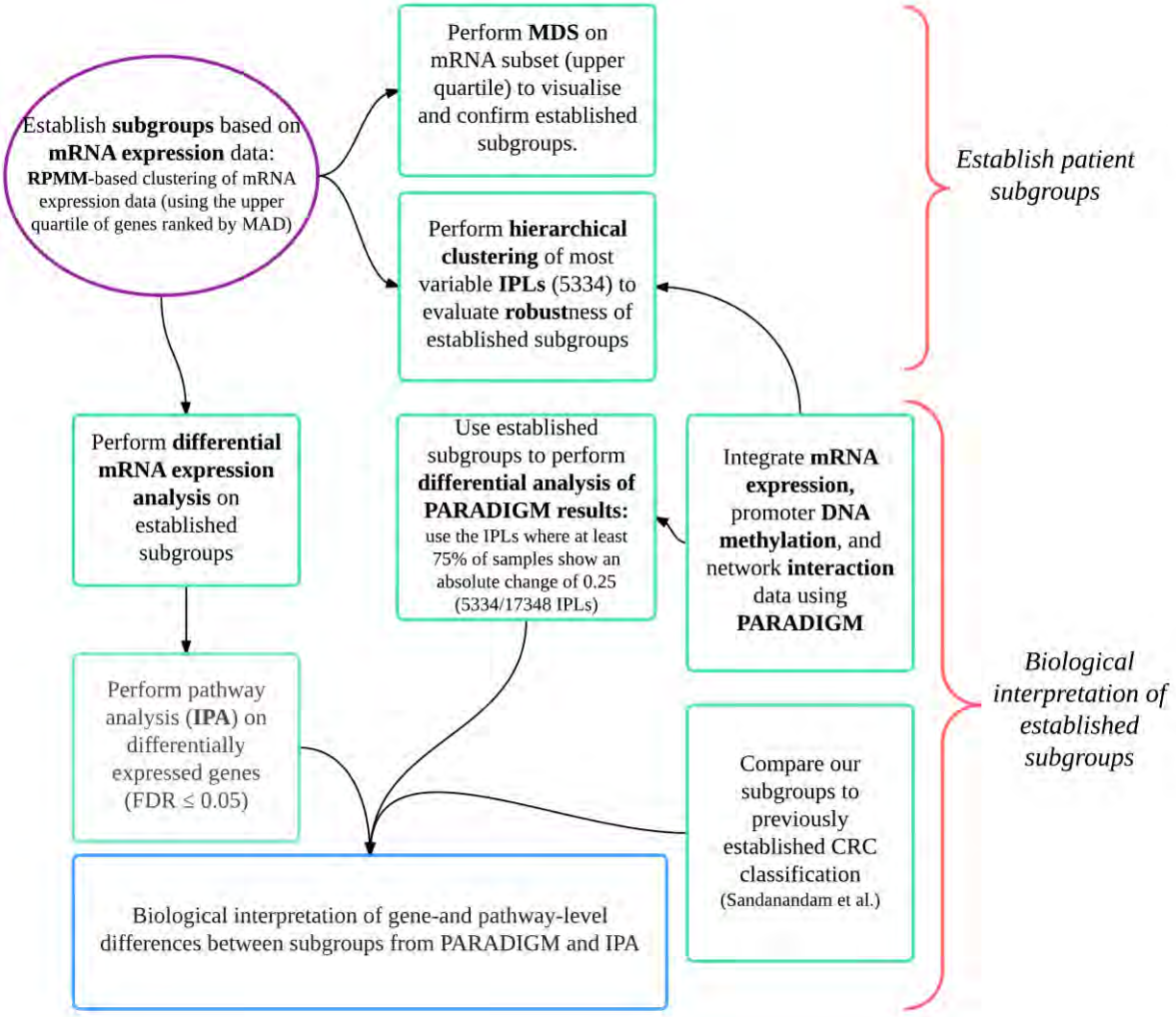


Figure 1: CRC subgroup classification and biological interpretation workflow. This method flowchart was applied to our cohort as well as to an external cohort of 155 CRC samples (for which mRNA expression profiles (GSE13294) and MSI status was available). Numbers indicated on the figure relate to our cohort. Median absolute deviation (MAD); integrated pathway level (IPL); colorectal cancer (CRC); Ingenuity Pathway Analysis (IPA).

Briefly, a mixture-model based approach called recursively-partitioned mixture model (RPMM)[337] was used for unsupervised clustering. One advantage of RPMM compared to other methods is that it does not require specification of the number of expected

subgroups. The data was next assessed using multidimensional scaling (MDS) in order to visualize the relationship between all samples and to confirm the RPMM-based subgroups.

Established subgroups were next evaluated from a network-level perspective using PARADIGM[363]—a pathway-activity inference approach which infers patient-specific alterations in genes, complexes and abstract processes by modeling each gene by a factor graph as a set of interconnected variables encoding the expression and known activity of a gene and its products, and allowing the incorporation of many types of omics data as evidence[363]. PARADIGM has been suggested as an alternative method of identifying cancer subtypes[363], having the advantage of producing pathway-level results on a per-patient basis. However, for the sake of comparing our subtypes to previously established subtypes derived from gene-expression data, subtypes were defined here solely on the basis of gene expression data.

Next, in order to investigate the underlying biology of these subgroups, differential gene expression analysis was performed to compare the two RPMM-derived groups using the R package limma[364]. Ingenuity pathway analysis (IPA) was then applied to the subset of significantly differentially expressed genes. These results were compared to those obtained through differential analysis of PARADIGM integrated pathway levels (IPLs) between subtypes.

The CRCassigner-786[358]–a gene signature subtype classifier that separates the five subtypes proposed by Sadananamdam et al.–was also used to evaluate our subgroups in the context of an external classification system. Conversely, RPMM-based classification was applied to the GSE13294 dataset (which had previously been classified by Sadanandam et al. and de Sousa et al.), and the three sets of classifications were compared. As with our cohort, IPA and PARADIGM were used to investigate biological features of the resultant RPMM-based subgroups, and the results were compared to those obtained through subgroup analyses of our cohort.

Lastly, the contribution of CpG island methylation in our CRC subgroups was evaluated by determining the frequency of CIMP+ in each subgroup (Chapter 6), as well as CpG

island methylation of Wnt pathway antagonists—a known mechanism of Wnt pathway activation in CIMP+ CRCs. Analysis of variance was used to find CpG sites associated with CRC subtypes.

Methods

Establishing patient subtypes

All data analyses described here, except PARADIGM, were conducted in the R statistical framework.

Recursively-partitioned mixture model (RPMM) clustering implemented in the R package RPMM[337] was used to identify tumour subtypes based on mRNA expression data (RPMM is explained in broader detail in Chapter 6). RPMM was applied to the third quartile most variable gene expression data by median absolute deviation (MAD), which left 8325/33297 transcript clusters in our cohort (Affymetrix Human Gene 1.0 ST arrays) and 13669/54675 probesets in the Jorissen cohort (GSE13294, Affymetrix HGU133-plus2 arrays). A Gaussian distribution was specified to suit the distribution of gene expression data. The gene expression data used for RPMM was ComBat-corrected[365] to adjust for batch and quality factors, while specifying the disease status (tumour vs. normal) as the phenotype of interest.

Multidimensional scaling (from the R package minfi[331]) was applied to the 8325 transcript clusters used as input for RPMM to visually explore the underlying pattern (regarding similarities/dissimilarities) among samples, using Euclidian distance as a measure of similarity.

Established subgroups were next evaluated from a network level in PARADIGM. Data were prepared for input to PARADIGM by median-centering the $\log_2$ gene expression data across samples for each gene, as recommended by the authors of PARADIGM[363]. Preprocessing and quality control of gene expression data is described in Chapter 5. For the methylation data, beta values in the 1500 bp region upstream from the transcription start-site were summarized from probe-level beta values, for each gene, using the IMA[334] package function indexregionfunc. These values were then inverted, since

PARADIGM relates high values to increased gene expression, and low values to decreased gene expression (and promoter methylation is generally associated with decreased gene expression); the inverted values were median-centered across samples and used as input to PARADIGM alongside gene expression data. The online version of PARADIGM, which is available on request at https://dna.five3genomics.com was used. Default analysis settings were used, except for decreasing the log-likelihood percent threshold from 0.05 to 0.01% as recommended by the authors of PARADIGM for smaller cohorts (personal communication).

To evaluate patient subgroups, hierarchical clustering (Euclidian distance, complete linkage) was performed on the subset of IPLs where at least 75% of samples had absolute activation scores of ≥ 0.25[363], which left 5334/17348 IPLs.

Biological interpretation of CRC subtypes

Once subgroups were established, all downstream analyses were conducted on gene expression data that had been corrected for batch and quality factors using ComBat[365], while specifying the RPMM-subgroups as the phenotype of interest (as opposed to disease status). This allows conservation of biologically meaningful subgroup-specific variation, while adjusting the data for known sources of technical variation.

Bacterial subgroup associations were evaluated using Fisher's exact test applied to categorical data (which was derived in Chapter 4). For Fusobacterium the level of colonisation by Fusobacterium rather than the absence or presence of Fusobacterium was compared since the majority of samples are colonised by Fusobacterium; for the rest of the bacteria positive vs. negative- colonisation status was evaluated between subgroups.

Differential gene expression analyses of RPMM-based subgroups were conducted using the R package limma[364], and the R package hugene10sttranscriptcluster.db[366] was used for annotation. Genes with an FDR ≤ 0.05 and an absolute fold change ≥ 1.25 were used to investigate pathway-level alterations that define CRC subtypes using IPA. For each cohort genes significantly altered between subtypes were used to investigate the IPA categories: canonical pathways, upstream regulators and diseases and functions.

IPA's canonical pathways, and diseases and functions analyses compute a right-tailed Fisher's exact test, considering a) the number of differentially expressed genes that participate in a given pathway/function, and b) the total number of genes known to participate in this pathway/function in the selected reference set (i.e. the full set of genes for the array platform used).

IPA's upstream regulator analysis predicts which genes, chemical drugs or toxicants act as upstream regulators, based on downstream gene expression data and prior knowledge of expected effects between upstream regulators and their target genes stored in the Ingenuity® Knowledge Base. Two statistical measures are produced: the overlap p-value is used to infer whether there is a statistically significant overlap between the dataset genes and the genes that are regulated by a transcriptional regulator (using Fisher's exact test), and the activation z-score is used to infer likely activation states of upstream regulators based on comparison with a model that assigns random regulation directions.

To assess the difference between established subgroups based on PARADIGM results, the R package limma was used to identify IPLs that showed differential activity between RPMM-based subgroups; IPLs with an FDR ≤ 0.05 and an absolute difference in median activity score between groups of at least 0.25 were deemed significant.

To apply the CRCassigner-786 to our cohort, each of the 786 genes in the panel were assigned to the Sadanandam subtype that had the maximum Prediction Analysis of Microarrays (PAM) score (specified by Sadanandam et al.[358]) for that gene. Hierarchical clustering (Euclidian distance, complete linkage) was applied to the gene expression data for each subset, to evaluate which of our samples most closely resembled a given subtype.

For methylation analyses, the CpGassoc[335] package was used to test for associations between CpG sites and CRC subgroups. To evaluate the effect of CpG island methylation on the activity of the Wnt signaling pathway, Wnt antagonists that were reported to be methylated in CRCs, including *SFRP1, SFRP2, SFRP5*[103], *DKK-1*[367], *DKK2, DKK4*[368], *WIF1*[369], *WNT5A*[370], *SOX17*[371], *APC1A*[372] and *APC2*[372], were selected

and RPMM-based clustering was performed on the methylation and gene expression data for these genes.

Results

Identifying tumour subtypes

RPMM-based clustering of the top quartile most variable transcript clusters (8325 transcript clusters) was applied whereby two main RPMM groups were obtained, one of which had two subgroups (Figure 2). The related clusters “rLL” (Left-Left) and “rLR” (Left-Right) were designated as group A, while the cluster “rR” (Right) was designated as group B.

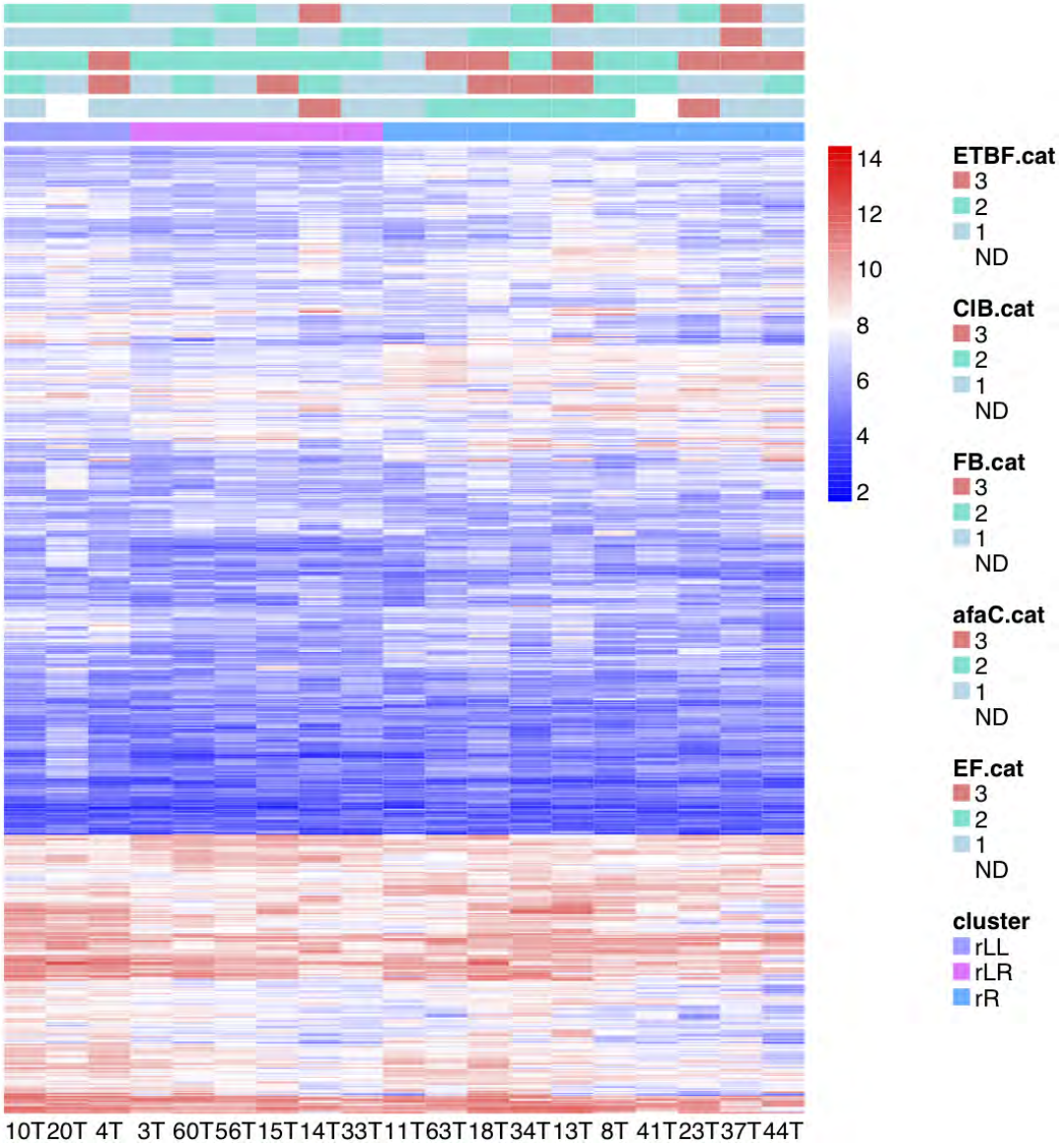


Figure 2: RPMM-based clustering of the top quartile (N=8325) most variable transcript clusters. Levels of bacterial colonisation (as determined in Chapter 4) are indicated on the figure legend, where 3: high-level infection, 2: low-level infection, 1: no infection. ETBF: Enterotoxigenic *Bacteroides fragilis*; ClB: *ClB*/*pks*+ *E. coli*, FB: Fusobacterium., afaC: *afaC*+ *E. coli*; EF: *Enterococcus faecalis.* The legend categories on the right are presented in the same order as the row annotations at the top of the graph. The scale on the right represents log2 expression values.

Multidimensional scaling (MDS) was applied to the same 8325 probes used for RPMM, which showed good agreement with the RPMM subgroups (Figure 3, left). MDS of the adjacent normal samples (Figure 3, right) demonstrated a moderate degree of correspondence with the tumour-derived subgroups, but there were no distinct clusters.

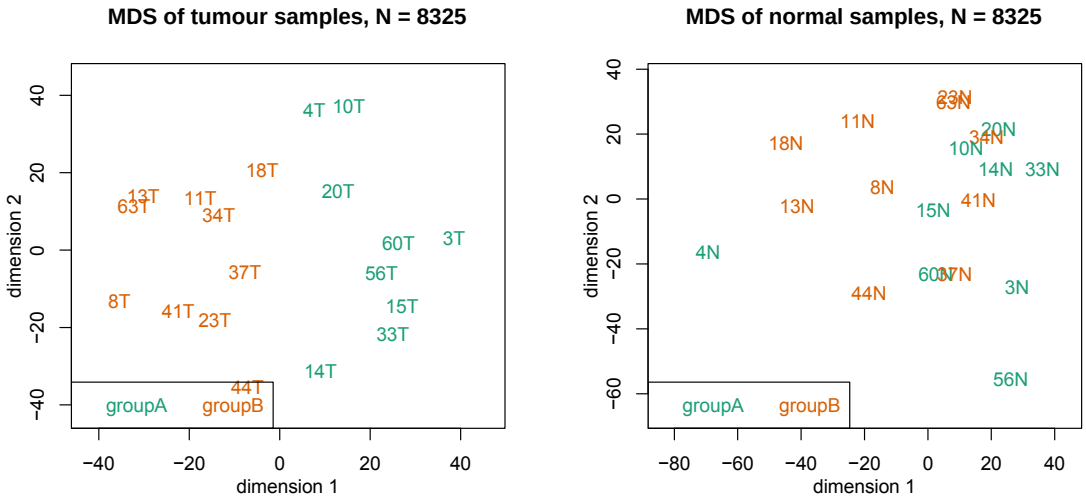


Figure 3: Multidimensional scaling of the top quartile (N=8325) most variable transcript clusters in tumour (left figure) and normal (right figure) samples. The RPMM-derived groups A and B are highlighted in green and orange, respectively. The tumour-derived subgroups are also highlighted in the adjacent normal samples (right figure) to evaluate the level of agreement between clustering of normal samples with the tumour-derived subgroups.

The subgroups obtained from RPMM-based gene expression clustering was compared to those obtained through hierarchical clustering of PARADIGM IPLs (using the top 5334/17348 IPLs), as shown in Appendix C. Two distinct clusters were obtained, and with the exception of one sample (18T), the clusters obtained for PARADIGM were identical to the two main RPMM-based gene expression clusters.

Applying our data analysis pipeline to a large external dataset with previously defined subtypes

The same RPMM-based subtyping method was applied to a publically available CRC expression dataset (GSE13294), which will be referred to as the Jorissen cohort[361]. This cohort had already been classified into subgroups by de Sousa et al.[357] and Sadanandam et al.[358]. Three RPMM clusters were obtained here, two of which were more closely related (rRL and rRR) and were therefore combined for downstream analysis—this group will be referred to as group B (Figure 4).

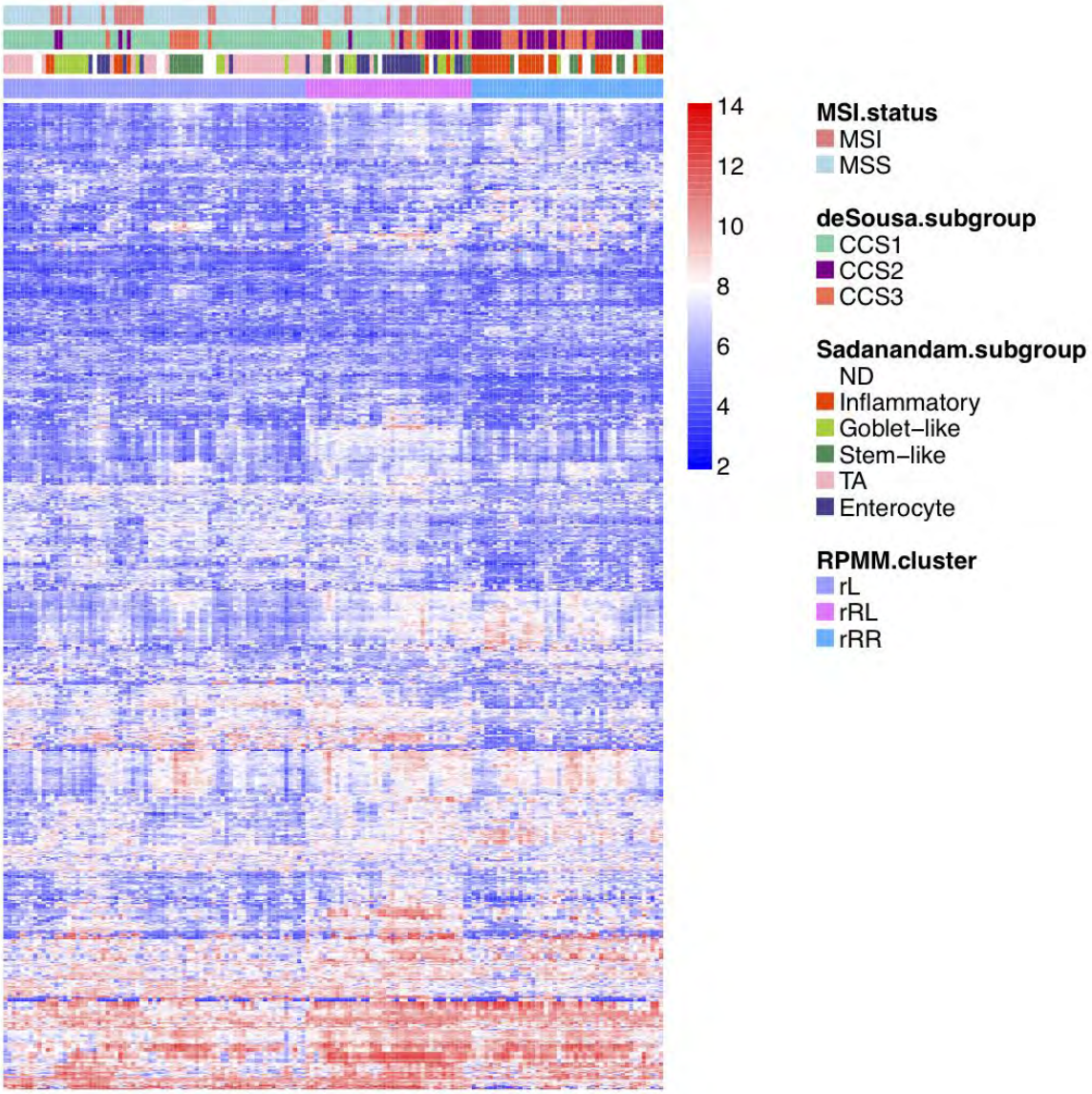


Figure 4: RPMM-based clustering of the most variable quartile of probesets (N=13669) for the Jorissen cohort (N=155). RPMM clusters are displayed alongside the previously established de Sousa and Sadanandam subtypes and MSI-status for each sample. Here, only the top 1000 most variable probes are displayed, although clustering was conducted on the top 25% most variable probes. The legend categories on the right are presented in the same order as the row annotations at the top of the graph. The scale on the right represents log2 expression values.

Multidimensional scaling again confirmed the RPMM-based clusters, which provided good separation of the data (Figure 5).

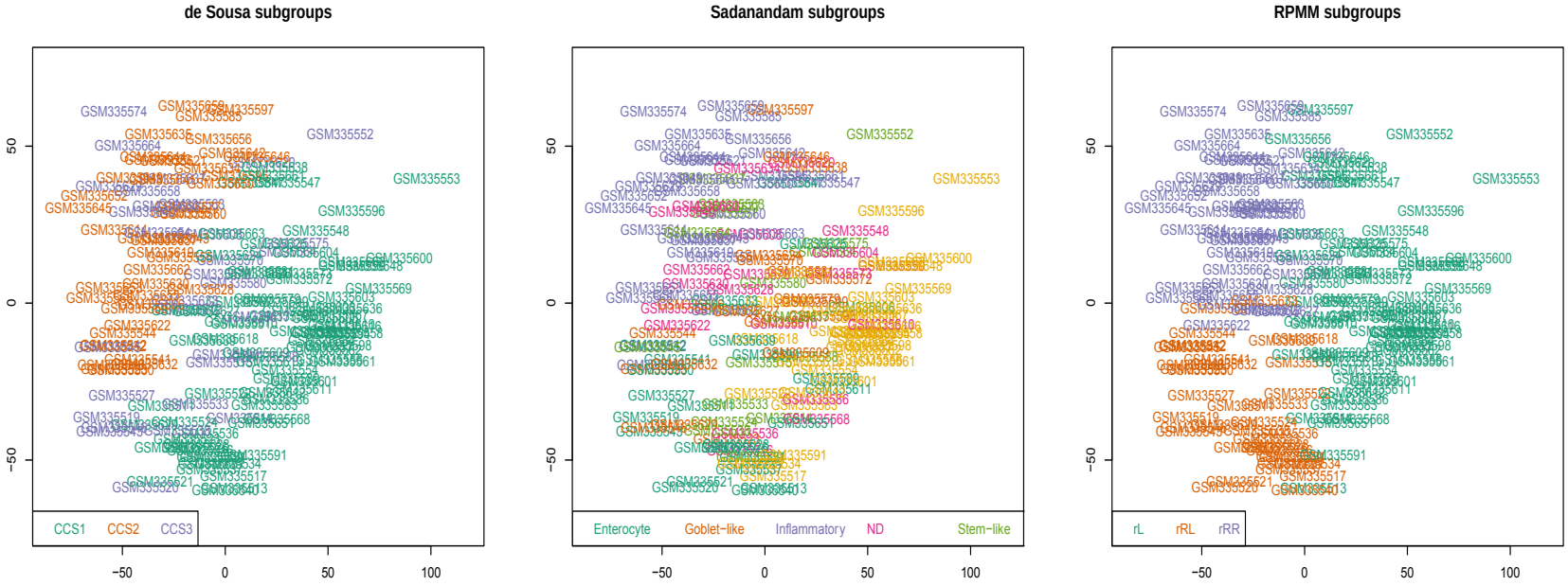


Figure 5: Multidimensional scaling of the top quartile (N=13669) most variable probesets used for RPMM clustering of the Jorissen cohort. The three figures are identical apart from the difference in annotation, where samples have been coloured by the de Sousa (left), Sadanandam (middle) or RPMM (right) subgroups.

Group B contained 84/155 samples and was heavily enriched for the de Sousa et al. CCS2- and CCS3-type samples, with 82% (32/39) and 72% (23/32) of CCS2 and CCS3 samples present in this group; meanwhile 72.5% (58/80) of the CCS1 group fell in group A. Regarding MSI-status, 79% of MSI+ patients fell in the B group.

Most of the inflammatory-like samples (85% (29/34)) and the stem-like samples (60% (12/20)), which have been reported to agree with the CCS3 subtype[359], fell into group B. However, goblet-like samples, which together with inflammatory-like samples have been reported to agree with the CCS2 subtype[359], were fairly equally distributed between groups A and B, with 43% (9/21) of goblet-like samples in group B. Further, 71% (17/24) of the Enterocyte-like samples and 14% (5/36) of the transit-amplifying-like type (which together corresponded to the CCS1 subtype[359]) fell into the B group.

To summarise, 40% of group B was composed of inflammatory-like samples (94% of which fell into the rRR subgroup of group B), 17% were stem-like samples, 12.5% were goblet-like, 24% were enterocyte-like and 7% were transit amplifying-like. Group B thus has an overrepresentation of: inflammatory-, stem- and enterocyte-like samples; CCS2 and CCS3 subtype samples; and MSI+ samples.

Biological features that distinguish CRC subtypes

No significant differences in clinical characteristics between group A and group B CRCs in our cohort were found; however, there was a trend for increased cancers of the proximal colon, and for patients of white or black ethnicity (as opposed to mixed ancestry) in group B patients (Table 1).

Table 1: Descriptive characteristics by CRC subtype.

	Group A (N = 9)	Group B (N = 10)	P (Fisher's exact)
MSI-H	3	4	1.00
HNPCC+	2	2	1.00
Stage*			
I/II	2	5	0.37
III/IV	6	5	0.37
Site			
Proximal colon	2	5	0.35
Distal colon	5	2	0.35

Rectum	2	3	1.00
Age*			
≤ 60	2	4	0.63
> 60	6	6	1
Gender			
Male	5	4	0.66
Female	4	6	0.66
Ethnicity			
Mixed ancestry	8	5	0.14
Black	0	2	0.47
White	0	3	0.2
Indian	1	0	0.47

*One case had missing information

Visual inspection of annotated heatmaps (Figure 2), suggested an overrepresentation of certain bacteria in group B. Using Fisher's exact test a) the number of Fusobacterium-high vs. Fusobacterium-low/negative cases between group A and group B, and b) positive vs. negative- colonisation status for the rest of the bacteria were compared. The frequency of high-level colonisation by any bacterium between the two groups was also compared. *E. faecalis,* Fusobacterium-high and EPEC had the strongest association with group B (p-values: 0.05, 0.06, 0.2), but this association was not significant after Benjamini-Hochberg multiple testing correction. There was also an increased frequency of high-level colonisation by any bacterium in group B (p=0.1, FDR=0.3) (Table 2).

Table 2: Comparison of bacterial colonisation between group A and group B samples using Fisher's exact test.

Comparison	Group A (N=9)	Group B (N=10)	P (Fisher's exact test)	FDR
FB-H	1	6	0.06	0.2
EF+ *	1	6	0.05	0.2
ETBF+	5	5	1	1
ClB+	3	3	1	1
afaC+	5	6	1	1
EPEC+	0	3	0.2	0.4
Colonisation-H (any)	5	13	0.1	0.3

*Two samples did not have data available. FB-H: Fusobacterium-high; EF: *E. faecalis*; ETBF: Enterotoxigenic *B. fragilis*; ClB+: colibactin+ E. coli; EPEC: Enteropathogenic *E. coli*; Colonisation-H: samples that had one or more high-level colonisations by any of the species tested.

CpG island methylation by subgroup

In Chapter 6 CIMP status was predicted using an established array-based marker panel of CIMP[305]. Interestingly, although the RPMM clustering did not take methylation data into consideration, the majority of CIMP+ samples fall in group B and in group B 80% of cases were CIMP+.

Next, given the central role of Wnt pathway activation in CRC, methylation of CpG islands in Wnt pathway antagonist genes that had previously been reported as methylated in CRC were evaluated. These included *SFRP1, SFRP2, SFRP5*[103]*, DKK-1*[367]*, DKK2, DKK4*[368] *,WIF1*[369]*, WNT5A*[370]*, SOX17*[371]*, APC1A*[372] and *APC2*[372]. Since CpG islands are frequently located at the 5' promoter of genes, CpG island methylation was used as a proxy for gene silencing. Wnt antagonists showed increased CpG island methylation in 7/10 group B cancers (Figure 6). However, these changes did not appear to translate to decreased expression of these genes (Appendix C).

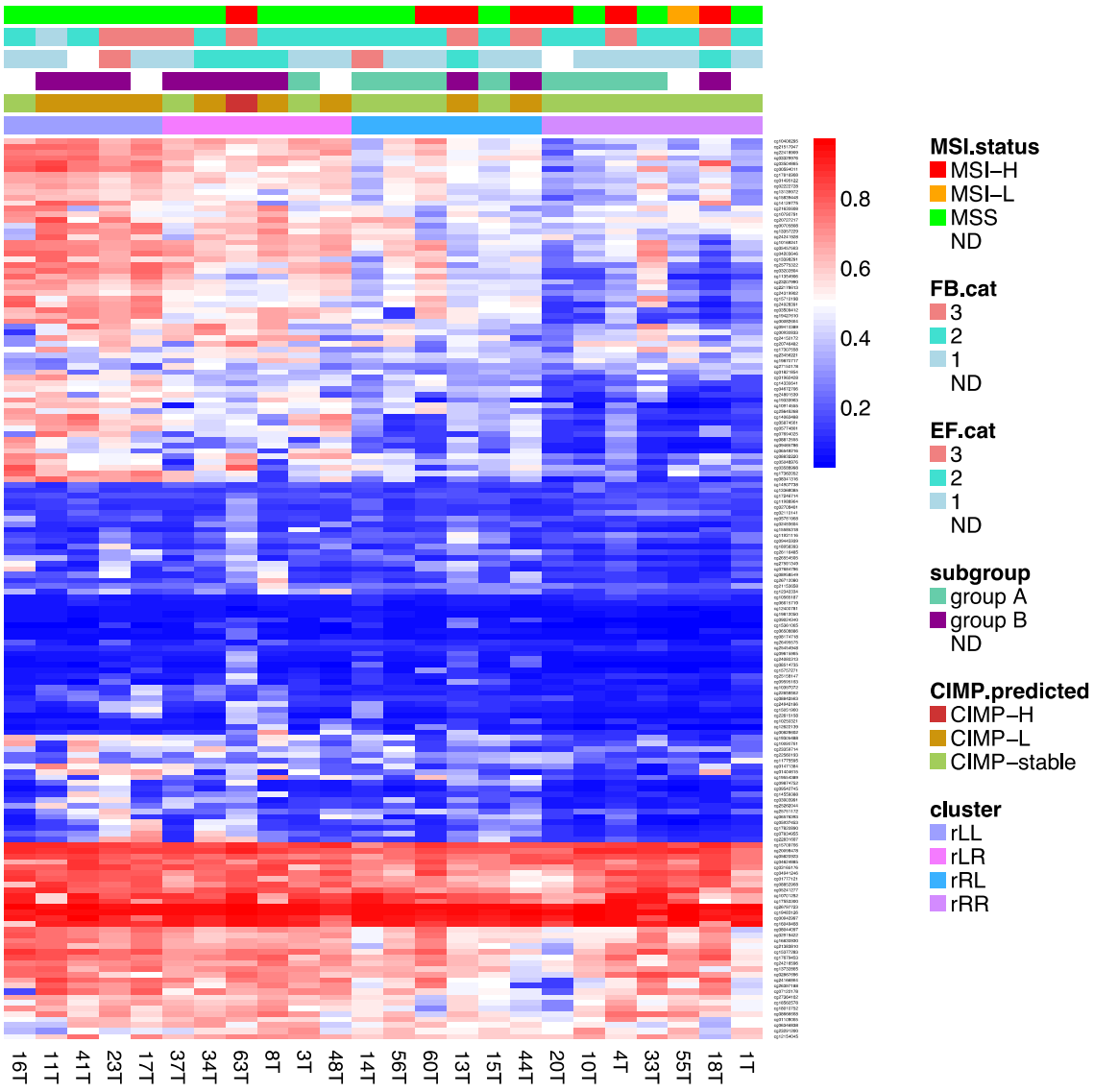


Figure 6: RPMM-based clustering of CpG islands in Wnt pathway antagonists known to be methylated in CRC. The legend categories on the right are presented in the same order as the row annotations at the top of the graph. The scale on the right represents methylation beta values (0–1). EF.cat: *E. faeclis* colonisation category (1=negative; 2=low-level; 3=high-level); FB.cat: Fusobacterium colonisation category (1=negative; 2=low-level; 3=high-level); ND: not determined.

Analysis of variance of genome-wide methylation between group A and group B cancers produced only 37 differentially methylated CpG sites at a relaxed FDR cutoff of 0.25 (Table 3). Even when analysing CIMP+ vs. CIMP- tumours (Chapter 6), the list of sites significantly associated with CIMP status was relatively small (124 sites with FDR $\leq$ 0.25). This is largely due to the drastic effect of multiple testing correction on this platform, the degree of which is proportional to the number of probes evaluated (in this case ~300 000). Nevertheless all 37 probes indicated increased methylation in group B

cancers, and 33 of 37 mapped to CpG islands, which supports the predicted CIMP+ status found in the majority of group B samples.

Table 3: Analysis of variance of CpG sites between group A and group B samples.

Probe ID	Gene symbol	Chromosome	CpG island	P value	FDR	Δ beta (group B vs. group A)
cg0505368	S1PR1	1	0	1.57E-07	0.04	0.27
cg0231594	CLSTN2	3	1	5.37E-07	0.06	0.39
cg0675867	GRIN2A	16	1	1.05E-06	0.06	0.25
cg1358910	FAM5B	1	0	1.44E-06	0.06	0.30
cg2615668	PAX2	10	1	1.58E-06	0.06	0.30
cg2710651	RGS20	8	1	1.76E-06	0.06	0.26
cg0327953	GALNT14	2	1	2.24E-06	0.07	0.46
cg2362480	ZSCAN30	18	0	2.57E-06	0.07	0.20
cg0678120	FADS2	11	1	3.99E-06	0.10	0.32
cg1224345	RAB11FIP	17	1	4.44E-06	0.10	0.21
cg1277752	LMX1B	9	1	7.02E-06	0.13	0.31
cg1391674	ZNF582	19	1	8.24E-06	0.14	0.39
cg1884438	EFS	14	1	1.10E-05	0.16	0.32
cg2710002	RAB39A	11	0	1.04E-05	0.16	0.28
cg0663447	KCNK10	14	1	2.20E-05	0.20	0.26
cg1030463	NRN1	6	1	1.77E-05	0.20	0.31
cg1811919	GATA4	8	1	2.32E-05	0.20	0.20
cg1838715	NRG1	8	1	2.66E-05	0.20	0.20
cg1839175	GRIN2A	16	1	2.33E-05	0.20	0.17
cg1949244	CELF4	18	1	1.79E-05	0.20	0.15
cg2101386	EFS	14	1	1.66E-05	0.20	0.29
cg2309282	PODN	1	1	2.21E-05	0.20	0.20
cg0973696	ECEL1	2	1	3.01E-05	0.21	0.26
cg0031954	ASTN1	1	1	3.40E-05	0.22	0.23
cg0736919	PCP4L1	1	1	3.82E-05	0.22	0.35
cg0966985	TLL1	4	1	3.81E-05	0.22	0.22
cg1185552	MPPED2	11	1	3.67E-05	0.22	0.41
cg1658075	XKR5	8	1	4.02E-05	0.23	0.22
cg0751450	HS6ST3	13	1	4.32E-05	0.23	0.26
cg1707868	POU3F3	2	1	4.21E-05	0.23	0.34
cg1725070	PTPN5	11	1	4.32E-05	0.23	0.24
cg1181542	ONECUT1	15	1	4.82E-05	0.25	0.30
cg1334707	UNC80	2	1	4.96E-05	0.25	0.16
cg1771676	APBA2	15	1	5.07E-05	0.25	0.22
cg1864217	WNT3A	1	1	5.01E-05	0.25	0.26
cg0995588	ITIH5	10	1	5.50E-05	0.25	0.14

cg2615818	NTRK3	15	1	5.35E-05	0.25	0.22

The Δ beta value was calculated for each CpG site by subtracting the median beta value across group A samples from that of group B samples.

CRCassigner-786 subtypes

In order to assess our CRC subtypes in the context of the Sadanandam et al. subgroups, and to gain biological insight into their defining characteristics, the CRCassigner-786 classifier was applied to our dataset as follows: each of the five subtypes was allocated the subset of the 786 genes that had maximum Prediction Analysis of Microarrays (PAM) scores (published by Sadanandam et al.) for that subtype. For each of the five resulting sub-classifiers, hierarchical clustering was performed in order to visually assess the degree of correspondence to each of the five subtypes; samples with increased gene expression relative to the rest of the cohort suggests increased correspondence to that subtype. For the transit amplifying-like panel of genes there was very little discernable difference between samples (Appendix C). For the remaining four subtypes (Figure 7), two main subgroups were clearly visible from the clustering dendrograms.

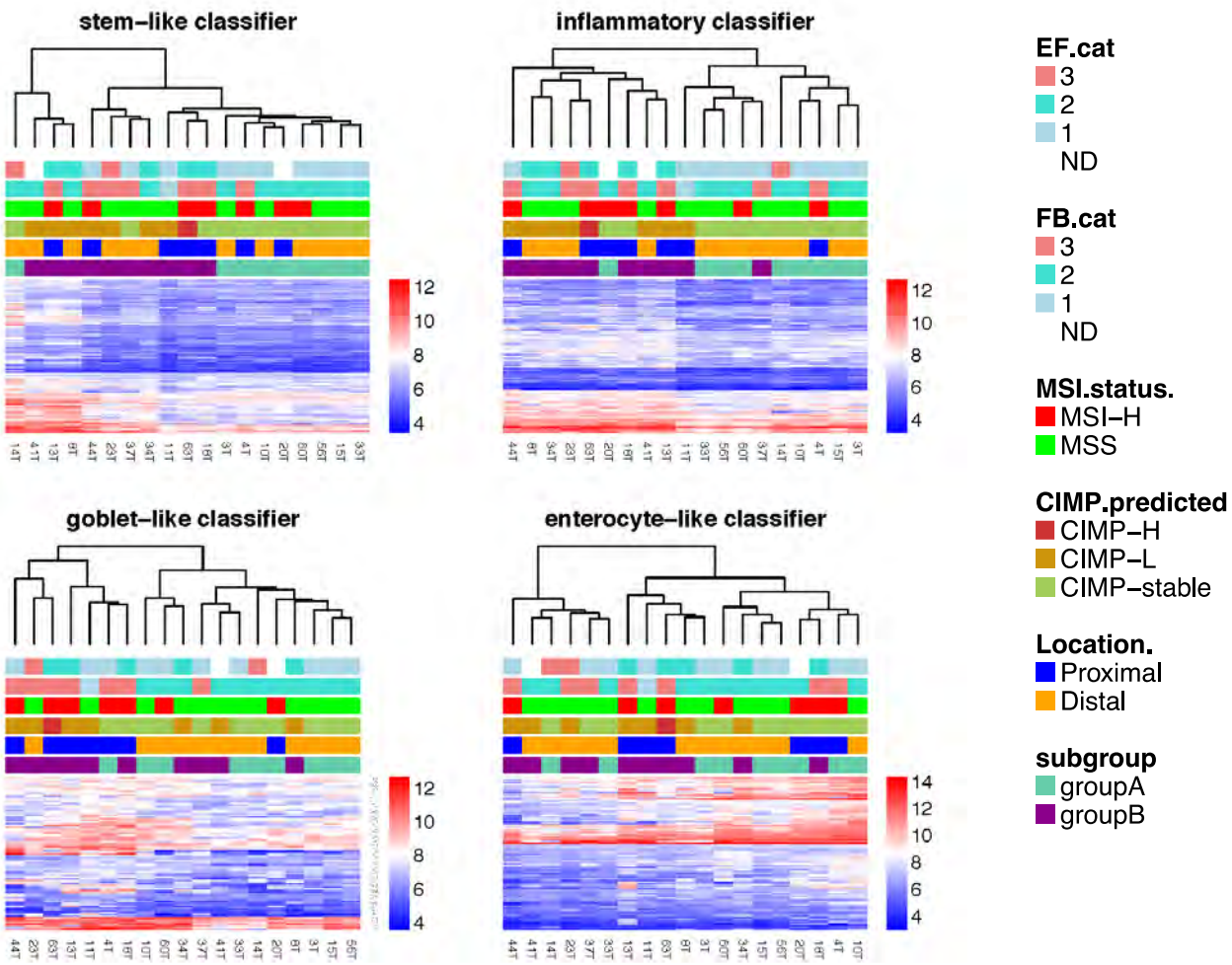


Figure 7: CRC classification according to the CRCassigner-786 classifier of Sadanandam et al. EF.cat: *E. faecalis* colonisation level category (1=negative; 2=low-level; 3=high-level); FB: Fusobacterium colonisation level category (1=negative; 2=low-level; 3=high-level). The legend categories on the right are presented in the same order as the row annotations at the top of the graph. The scale on the right represents log2 expression values.

Four samples showed a relative increase in stem-like expression (14T, 41T, 13T, 8T), three of which had been classified as group B samples; in accordance with the literature, most (3 of 4) of these samples were of distal origin[358]. Seven samples (44T, 23T, 63T, 13T, 11T, 4T, 18T) had a relative increase in goblet-like expression, and most MSI+ samples (5 of 7) belonged to this cluster—these included both HNPCC and sporadic MSI+ samples. Five of seven goblet-like samples were also predicted to be CIMP+, and 6 of 7 were located in the proximal colon. Meanwhile, the inflammatory classifier separated 9 of 10 group B samples from the rest of the cohort and included 44T, 8T, 34T, 23T, 63T, 20T, 18T, 41T and 13T—this group included 8 of 10 CIMP+ samples

and 5 of 7 MSI+ samples. With the enterocyte-like classifier it was more difficult to define subgroups: the six samples showing the lowest expression (44T, 41T, 14T, 23T, 37T, 33T) mostly belonged to group B, whereas samples that displayed the highest expression (60T, 34T, 15T, 56T, 20T, 18T, 4T, 10T) mostly belonged to group A.

Both the inflammatory samples and the goblet-like samples were enriched for high-level colonisation by Fusobacterium, with 6 of 7 of Fusobacterium-high infected samples present in the goblet-like cluster. Further, 6 of 7 *E. faecalis*+ samples fell in the inflammatory cluster.

The inflammatory/goblet characteristics of group B samples are supported by tumour biopsy pathology reports since 5 of 9 patients (9 of 10 group B samples had available records) displayed signs of inflammation and/or had a visible mucinous component. Two more patients in the inflammatory-subtype presented with diverticular disease and mucinous metaplasia of the appendix, respectively. In group A, only one patient (8 of 9 group A samples had available records) had a reported mucinous component (4T), while another had diverticulae (10T).

Our results thus agree with the merging of the goblet and inflammatory subtypes into one subtype (CCS2), as proposed by Sadanandam et al.[359], and demonstrate that many of these samples are MSI-H (5/7) or CIMP+. It can be concluded that our infection-dominated subtype (group B) most closely resembles the inflammatory subtype of Sadanandam et al., with certain samples presenting with features of the stem-like or goblet type. Group A samples are more likely of the enterocyte-type.

Gene- and pathway-level CRC subtype comparison

Patients belonging to group B of our cohort and of the Jorissen cohort were suspected to have similar biological features, since both were enriched for inflammatory-like samples, and likely also for CIMP+ samples, given the enrichment with CCS2 and CCS3 samples in group B of the Jorissen cohort. Gene expression and pathway-level differences between the two RPMM-derived subgroups were therefore compared between the cohorts. For each cohort a) differential gene expression analysis between subgroups, followed by b) IPA of the significantly differentially expressed genes, and c) differential analysis of IPLs obtained through PARADIGM were performed. Finally,

the overlap between the two cohorts regarding subgroup-specific gene- and pathway-level alterations was established.

Gene expression analysis between group A and group B of our cohort produced 4671 genes at an absolute fold change cutoff of 1.25, and 296 genes at an absolute fold change cutoff of 2 (FDR ≤ 0.05). For the Jorissen cohort, 5771 genes were differentially expressed at an absolute fold change cutoff of 1.25, and 546 genes at a fold change cutoff of 2 (FDR ≤ 0.05). Of the 4671 genes differentially expressed in our cohort, 1619 overlapped with Jorissen subgroup comparison results, 78% (1266/1619) of which were consistently up- or down-regulated in *both* cohorts. Meanwhile, 19 genes were differentially altered in both cohorts at an absolute fold change ≥ 2 (FDR ≤ 0.05), 18 of which were consistently up- or down-regulated in group B of both cohorts (Table 4).

Table 4: Genes differentially expressed at an |FC| ≥ 2 and FDR ≤ 0.05 between subtypes in both cohorts.

Gene Symbol	GeneST (FDR)	GeneST (FC)	Jorissen (FDR)	Jorissen (FC)	consistent
C10orf99	3.3E-02	-2.5	4.3E-03	-2.4	yes
COL12A1	1.4E-03	3.6	4.8E-06	2.1	yes
CXCL10	4.9E-02	2.8	2.3E-06	2.8	yes
FCGR2A	1.0E-02	2.5	7.1E-14	2.4	yes
HSPA4L	1.6E-02	2.8	9.1E-11	2.6	yes
IL1B	7.0E-03	3.2	8.9E-09	2.7	yes
IL8	5.4E-03	4.1	6.1E-06	2.9	yes
MMP1	4.6E-02	3.0	2.7E-05	2.6	yes
MMP12	1.3E-03	4.4	5.0E-08	3.0	yes
NR4A2	1.5E-02	2.2	4.8E-06	2.1	yes
PKIB	4.1E-03	-2.2	7.7E-05	2.0	no
PLA2G4A	3.2E-02	2.4	1.9E-04	2.3	yes
PLK2	5.8E-04	2.0	3.8E-11	2.8	yes
POSTN	1.8E-02	3.7	2.1E-07	3.0	yes
PTGS2	2.6E-03	3.9	1.2E-08	3.0	yes
REG1A	3.3E-02	6.2	3.2E-02	2.5	yes
TDO2	4.1E-02	3.2	7.5E-06	2.0	yes
TNFAIP6	3.7E-03	3.0	5.1E-09	2.3	yes
VCAN	2.1E-02	2.4	7.1E-08	2.3	yes

The biological roles of many of the 18 genes supported the relevance of inflammation in the pathogenesis of group B cancers: the pro-inflammatory chemokine CXCL10 and cytokines IL-8 and IL-1B, along with the infamous prostaglandin-endoperoxide synthase 2 (PTGS2 or COX-2), which is induced by IL-1B[373], were all upregulated in group B cancers. Two downstream targets of these inflammatory genes were also upregulated in group B cancers: regenerating islet-derived 1 alpha (REG1A) is upregulated by IL-8[374], and nuclear receptor 4A2 (NR4A2) is upregulated by COX-2 and rescues cells from undergoing programmed cell death[375]. Further, polo-like kinase 2 (PLK2), also upregulated in group B cancers, is upregulated in response to ROS-induced oxidative stress, where it facilitates cell survival by promoting the NRF2 antioxidant pathway[376].

IL-8 expression is of particular importance, since it is a central element of the innate immune response that is regulated by a number of different stimuli, including inflammatory signals (TNF-α, IL-1β), chemical and environmental stress (such as hypoxia), steroid hormones[377], infection with certain viruses[378,379] and bacteria. Bacteria known to induce IL-8 include Fusobacterium[196,380,381] and *H. pylori*—which induces IL-8-dependent REG1A expression[374]. Intriguingly, both REG1A and REG3A were potently upregulated in cancers with high levels of Fusobacterium infection, and to a lesser extent IL-8 (Chapter 8).

Given their role in inflammation, it is not surprising that many of the genes in Table 4 are implicated in the pathogenesis of irritable bowel disease (IBD) and/or are linked to host response to infection. For instance, *FCGR2A* polymorphisms modify susceptibility to IBD[382], and *IL-8*[383]*, COX-2, REG3A, REG1A*[384]*, IL-1B*[385]*, MMP1*[386]*, MMP-12*[387,388] and *POSTN*[389] are all overexpressed in patients with IBD.

Extracellular matrix proteoglycans versican (*VCAN*) and lumican (*LUM*) are both significantly upregulated in group B cancers (lumican with fold changes of 4 and 1.7 in our cohort and the Jorissen cohort, respectively). In addition to their structural role in the extracellular matrix, proteoglycans act as signaling molecules by virtue of the their complex structures, which facilitate interaction with ligands and receptors[390] of endogenous or exogenous origin. In their soluble form, proteoglycans can act as danger

signals, acting as damage-associated molecular patterns (DAMPs), which induce non-pathogen mediated inflammation through TLR signaling upon tissue stress[391]. Accordingly, proteoglycans have established roles in various cancers[390], including CRC[392]. Versican in particular regulates cell adhesion and survival, cell proliferation, cell migration and extracellular matrix assembly[393], and activates TLR2 and TLR6 in a CD14-dependent manner, inducing the expression of *TNFA*, *IL6* and several growth factors[391]. Further, glycosaminoglycans linked to proteoglycans are an important element of the host-pathogen interface that can mediate pathogenic infection with viruses, bacteria and parasites[394]. Lumican binds and presents bacterial lipopolysaccharide (LPS) to CD14, which initiates an immune response leading to increased phagocytosis[391].

Many of the genes upregulated in group B cancers also have roles in tumour progression and metastasis; these include matrix metalloproteases (MMPs) 1 and 12, which have opposing roles in promoting and inhibiting tumour progression, respectively[395,396]; CXCL10 also elicits seemingly contradictory effects on tumour progression since it is antiangiogenic and antiproliferative in CRC[397], yet is associated with advanced cancer and may promote invasion through the induction of cell migration[398]; Periostin (POSTN) promotes metastatic CRC growth by increased cell survival[399]; IL-8 promotes angiogenesis[400] and lastly; *REG1A* expression in CRC has been linked to poor prognosis and correlates with recurrence and/or distant metastasis in MSI+ patients[401].

Lastly, tryptophan 2,3-dioxygenase (TDO) is an important mediator of immune tolerance, and its overexpression may contribute to tumoral immune resistance to tumour-specific antigens in group B cancers[402].

Pathway-level CRC subgroup comparison

Ingenuity Pathway Analyses

IPA was used to investigate pathway-level alterations that define CRC subtypes. For each cohort genes significantly altered between subtypes (FDR ≤ 0.05, FC ≥ 1.25) were used to investigate a) canonical pathways b) upstream regulators and c) diseases and functions that defined CRC subtypes.

Canonical pathway analyses

Fifty-four and 96 canonical pathways showed significant overrepresentation ($p \leq 0.05$) in group B cancers in our cohort and in the Jorissen cohort, respectively. In our cohort, *EIF2 signaling* was the top scoring canonical pathway, while numerous pathways related to DNA and protein damage response were altered between subtypes (Table 5a). In the Jorissen cohort, top-scoring pathways included *Antigen Presentation Pathway*, *IGF-1 Signaling* and *Protein Ubiquitination Pathway* (Table 5b).

Table 5a: Top 20 most significant IPA canonical pathways in group B vs. group A samples of our cohort.

Ingenuity Canonical Pathways	-log(p-value)	Downregulated	Upregulated
EIF2 Signaling	13.4	8/169 (5%)	109/169 (64%)
Role of BRCA1 in DNA Damage Response	10.6	0/60 (0%)	46/60 (77%)
Protein Ubiquitination Pathway	9.27	20/249 (8%)	126/249 (51%)
Cell Cycle Control of Chromosomal Replication	9.25	1/27 (4%)	25/27 (93%)
Cell Cycle: G2/M DNA Damage Checkpoint Regulation	8.48	1/48 (2%)	32/48 (67%)
Mitotic Roles of Polo-Like Kinase	8.32	4/63 (6%)	39/63 (62%)
Hereditary Breast Cancer Signaling	7.29	7/111 (6%)	61/111 (55%)
Role of CHK Proteins in Cell Cycle Checkpoint Control	6.52	4/55 (7%)	36/55 (65%)
tRNA Charging	5.72	0/38 (0%)	26/38 (68%)
RAN Signaling	5.61	0/16 (0%)	15/16 (94%)
Mismatch Repair in Eukaryotes	5.61	0/16 (0%)	13/16 (81%)
Cell Cycle: G1/S Checkpoint Regulation	4.86	7/63 (11%)	32/63 (51%)
Regulation of eIF4 and p70S6K Signaling	4.54	10/141 (7%)	72/141 (51%)
Cyclins and Cell Cycle Regulation	4.46	10/77 (13%)	36/77 (47%)
ATM Signaling	4.09	2/59 (3%)	31/59 (53%)
Telomere Extension by Telomerase	4.09	0/15 (0%)	14/15 (93%)
Purine Nucleotides De Novo Biosynthesis II	4.03	1/11 (9%)	8/11 (73%)
DNA Double-Strand Break Repair by Homologous Recombination	3.60	0/14 (0%)	10/14 (71%)
DNA Double-Strand Break Repair by Non-Homologous End Joining	3.60	0/14 (0%)	11/14 (79%)

DNA damage-induced 14-3-3“€ Signaling	3.47	3/19 (16%)	11/19 (58%)
Estrogen-mediated S-phase Entry	3.45	3/24 (13%)	16/24 (67%)

Table 5b: Top 20 most significant IPA canonical pathways in group B vs. group A samples of the Jorissen cohort.

Ingenuity Canonical Pathways	-log(p-value)	Downregulated	Upregulated
Antigen Presentation Pathway	9.28	1/34 (3%)	28/34 (82%)
IGF-1 Signaling	8.40	14/96 (15%)	46/96 (48%)
Glucocorticoid Receptor Signaling	8.38	24/256 (9%)	106/256 (41%)
Role of NFAT in Regulation of the Immune Response	8.31	14/158 (9%)	74/158 (47%)
CD28 Signaling in T Helper Cells	7.83	9/107 (8%)	55/107 (51%)
OX40 Signaling Pathway	7.71	2/46 (4%)	32/46 (70%)
Role of Tissue Factor in Cancer	7.35	11/107 (10%)	52/107 (49%)
Cdc42 Signaling	7.21	8/121 (7%)	61/121 (50%)
B Cell Receptor Signaling	7.17	16/167 (10%)	73/167 (44%)
Leukocyte Extravasation Signaling	6.76	13/191 (7%)	85/191 (45%)
Integrin Signaling	6.68	18/194 (9%)	81/194 (42%)
Protein Ubiquitination Pathway	6.54	21/251 (8%)	101/251 (40%)
Virus Entry via Endocytic Pathways	6.51	10/89 (11%)	43/89 (48%)
Caveolar-mediated Endocytosis Signaling	6.19	6/71 (8%)	38/71 (54%)
HGF Signaling	6.16	12/104 (12%)	47/104 (45%)
PI3K/AKT Signaling	6.14	13/120 (11%)	53/120 (44%)
IL-17 Signaling	5.94	4/72 (6%)	40/72 (56%)
Type I Diabetes Mellitus Signaling	5.63	3/100 (3%)	53/100 (53%)
Prostate Cancer Signaling	5.61	7/80 (9%)	40/80 (50%)
PKCëü Signaling in T Lymphocytes	5.60	7/107 (7%)	52/107 (49%)

Next, the overlap between subtype analyses for the two cohorts was established, which resulted in 15 significantly overrepresented canonical pathways, all of which were

predicted to be activated in group B samples. Included were pathways related to DNA and protein damage response, as well as cell cycle regulation (Table 5c).

Table 5c: Canonical pathways predicted to be significantly altered in both cohorts in group B vs. group A samples. The first three data columns refer to our cohort, and the last three to the Jorissen cohort.

Ingenuity Canonical Pathways	log(p-value)	Downregulated	Upregulated	log(p-value) (Jorissen)	Downregulated (Jorissen)	Upregulated (Jorissen
Role of BRCA1 in DNA Damage Response	10.6	0/60 (0%)	46/60 (77%)	1.85	12/61 (20%)	17/61 (28%)
Protein Ubiquitination Pathway	9.27	20/249 (8%)	126/249 (51%)	6.54	21/251 (8%)	101/251 (40%)
Cell Cycle: G2/M DNA Damage Checkpoint Regulation	8.48	1/48 (2%)	32/48 (67%)	1.79	7/49 (14%)	17/49 (35%)
Hereditary Breast Cancer Signaling	7.29	7/111 (6%)	61/111 (55%)	1.68	16/112 (14%)	32/112 (29%)
Cell Cycle: G1/S Checkpoint Regulation	4.86	7/63 (11%)	32/63 (51%)	1.38	12/63 (19%)	16/63 (25%)
Regulation of eIF4 and p70S6K Signaling	4.54	10/141 (7%)	72/141 (51%)	1.42	13/140 (9%)	44/140 (31%)
Cyclins and Cell Cycle Regulation	4.46	10/77 (13%)	36/77 (47%)	1.31	11/77 (14%)	22/77 (29%)
DNA damage-induced 14-3-3“€ Signaling	3.47	3/19 (16%)	11/19 (58%)	1.63	2/19 (11%)	9/19 (47%)
Gluconeogenesis I	2.83	2/24 (8%)	12/24 (50%)	1.33	2/23 (9%)	10/23 (43%)
mTOR Signaling	2.19	16/178 (9%)	73/178 (41%)	1.42	21/181 (12%)	51/181 (28%)
Polyamine Regulation in Colon Cancer	2.12	1/22 (5%)	12/22 (55%)	1.68	0/21 (0%)	12/21 (57%)
dTMP De Novo Biosynthesis	1.86	0/5 (0%)	4/5 (80%)	1.35	0/5 (0%)	4/5 (80%)
Androgen Signaling	1.44	11/109 (10%)	42/109 (39%)	1.84	13/110 (12%)	35/110 (32%)
Calcium Transport I	1.37	3/9 (33%)	2/9 (22%)	1.38	2/9 (22%)	4/9 (44%)

Endoplasmic Reticulum Stress Pathway	1.34	0/21 (0%)	12/21 (57%)	1.68	2/21 (10%)	10/21 (48%)

Upstream regulator analyses

As described in the Materials and Methods section, IPA's upstream regulator analysis predicts which genes (including transcriptional regulators) act as upstream regulators based on downstream gene expression data and prior knowledge of expected effects between transcriptional regulators and their target genes. Two statistical measures are produced: the p-value of overlap, and the activation z-score (where negative and positive scores indicate inhibition or activation by the upstream regulator).

There were 150 and 894 IPA upstream regulator entries in our CRC cohort and the Jorissen cohort, respectively, that showed significant overrepresentation between subgroups (p-value of overlap ≤ 0.05). The top 20 scoring upstream regulators in our cohort included multiple transcription factors that were predicted to be activated (E2F1, E2F2, MYC, MYCN, XBP1, NFE2L2, TBX2 and CCND1) or deactivated (RB1, CDKNA2, KDM5B and NUPR1) in group B cancers (Table 6a). Meanwhile, in the Jorissen cohort, top-scoring subtype-specific upstream regulators were dominated by various cytokines, including TNF, IFNG and IFNA2 and growth factors including TGFB1 and HGF (Table 6b).

Table 6a: Top 20 most significant upstream regulators predicted to be significantly altered in group B vs. group A samples in our cohort, with absolute activation z-scores ≥ 2. Chemical upstream regulators were excluded.

Upstream Regulator	Molecule Type	Activation z-score	p-value of overlap
CDKN1A	Other	-3.688	3.68E-17
E2F1	transcription regulator	5	1.91E-15
MYC	transcription regulator	7.745	1.84E-14
XBP1	transcription regulator	7.864	5.63E-14
let-7	microRNA	-7.67	2.62E-13
MYCN	transcription regulator	5.696	1.88E-12
RB1	transcription regulator	-4.904	1.91E-12

CDKN2A	transcription regulator	-5.151	3.04E-12
NFE2L2	transcription regulator	6.705	3.33E-10
E2f	Group	3.208	8.82E-10
TBX2	transcription regulator	5.857	2.86E-09
PTGER2	G-protein coupled receptor	5.092	4.28E-09
EP400	Other	4.029	7.37E-09
CCND1	transcription regulator	4.56	8.15E-09
NUPR1	transcription regulator	-9.207	3.79E-08
miR-1 (and other miRNAs w/seed GGAAUGU)	mature microRNA	-7.472	1.92E-07
E2F2	transcription regulator	2.5	2.75E-07
RICTOR	Other	-7.97	4.30E-07
Rb	Group	-4.619	5.80E-07
KDM5B	transcription regulator	-5.879	3.53E-06

Table 6b: Top 20 most significant upstream regulators predicted to be significantly altered in group B vs. group A samples in the Jorissen cohort, with absolute activation z-scores ≥ 2. Chemical upstream regulators were excluded.

Upstream Regulator	Molecule Type	Activation z-score	p-value of overlap
TGFB1	growth factor	7.62	4.62E-43
IFNG	cytokine	11.73	9.13E-42
TP53	transcription regulator	4.115	1.14E-41
TNF	cytokine	11.051	7.45E-33
IL2	cytokine	6.836	3.86E-26
STAT3	transcription regulator	6.512	1.81E-23
IFNA2	cytokine	8.837	7.99E-23
OSM	cytokine	7.171	1.63E-22
CD40LG	cytokine	7.058	9.83E-22
IL4	cytokine	5.298	4.38E-21
CD3	complex	-4.992	5.04E-21
IL1B	cytokine	10.491	1.23E-20
HRAS	enzyme	2.703	1.50E-20
IL6	cytokine	8.821	8.31E-20
HGF	growth factor	5.829	1.13E-19
APP	other	4.657	1.50E-19
Interferon alpha	group	8.861	1.51E-19
NFKBIA	transcription regulator	4.071	2.99E-19

Vegf	group	7.696	2.77E-18
CSF2	cytokine	8	3.87E-18

Of the 68 upstream regulators that overlapped between the two datasets, 67 had consistent direction of predicted activation state. The upstream regulators (excluding molecules of chemical origin such as chemical toxicants and chemical drugs) with consistent direction of activation between the two cohorts are listed in Table 6c.

Table 6c: Upstream regulators predicted to be significantly altered in our cohort and the Jorissen cohort in group B vs. group A samples, with absolute activation z-scores ≥ 2. Chemical upstream regulators are not shown.

Upstream Regulator	Molecule Type	Activation z-score	p-value of overlap	Activation z-score	p-value of overlap
BAX	Transporter	3.223	0.0103	3.471	9.67E-06
CD28	transmembrane receptor	-6.198	0.0313	-3.741	4.54E-13
KDM5B	transcription regulator	-5.879	3.53E-06	-4.241	3.05E-07
TSC22D1	transcription regulator	2.236	0.00386	2.449	0.00112
E2F2	transcription regulator	2.5	2.75E-07	2.63	0.00495
SREBF1	transcription regulator	3.145	0.00983	2.371	0.00225
ATF4	transcription regulator	4.22	0.000396	3.736	0.000108
FOXM1	transcription regulator	4.498	1.65E-05	3.503	0.000101
E2F1	transcription regulator	5	1.91E-15	3.969	3.23E-08
NFE2L2	transcription regulator	6.705	3.33E-10	4.295	5.58E-07
XBP1	transcription regulator	7.864	5.63E-14	7.336	1.58E-10
Irgm1	Other	-4.101	0.00292	-4.208	5.15E-06
BID	Other	2.121	0.0136	3.302	0.000496

APP	Other	2.403	0.00363	4.657	1.50E-19
PRNP	Other	2.714	0.034	2.283	2.34E-05
RAB1B	Other	2.864	0.00401	2.469	0.00657
SCAP	Other	3.59	0.00525	2.652	0.0429
GAST	Other	4.8	0.0156	3.616	0.00011
CD24	Other	5.385	0.0145	2.853	4.10E-10
let-7	microRNA	-7.67	2.62E-13	-2.755	5.97E-07
miR-124-3p (and other miRNAs w/seed AAGGCAC)	mature microRNA	-8.019	3.20E-05	-7.563	3.17E-12
miR-1 (and other miRNAs w/seed GGAAUGU)	mature microRNA	-7.472	1.92E-07	-6.004	3.74E-12
miR-16-5p (and other miRNAs w/seed AGCAGCA)	mature microRNA	-7.363	8.80E-06	-4.226	1.12E-08
miR-155-5p (miRNAs w/seed UAAUGCU)	mature microRNA	-5.855	0.00501	-5.971	4.12E-15
miR-30c-5p (and other miRNAs w/seed GUAAACA)	mature microRNA	-5.81	0.00016	-4.682	4.68E-09
let-7a-5p (and other miRNAs w/seed GAGGUAG)	mature microRNA	-5.531	0.0158	-2.946	0.00713
miR-291a-3p (and other miRNAs w/seed AAGUGCU)	mature microRNA	-4.991	0.0457	-2.811	0.000176
miR-145-5p (and other miRNAs w/seed UCCAGUU)	mature microRNA	-3.312	0.0471	-4.289	8.14E-05
miR-34a-5p (and other miRNAs w/seed GGCAGUG)	mature microRNA	-2.707	0.0207	-2.88	2.14E-06
PIM1	Kinase	2.113	0.0221	2.175	0.0018
ATM	Kinase	2.685	0.000793	2.635	1.93E-06
ANGPT2	growth factor	5.468	0.000382	5.213	2.70E-10
HGF	growth factor	7.825	0.00285	5.829	1.13E-19

caspase	Group	2.538	0.016	2.55	0.0373
Jnk	Group	3.323	0.0293	5.8	8.19E-12
PTGER2	G-protein coupled receptor	5.092	4.28E-09	2.702	0.0102
CAT	Enzyme	-2.745	0.025	-3.046	7.31E-12
PLA2G2A	Enzyme	2.228	0.0108	2.559	0.00036
PIN1	Enzyme	2.534	0.0128	2.107	0.00232
CD38	Enzyme	4.713	0.0189	5.897	5.13E-07
IL3	Cytokine	3.316	0.0132	4.617	4.88E-08
IL5	Cytokine	5.279	0.0306	7.804	1.99E-14
CSF2	Cytokine	8.068	0.000137	8	3.87E-18
CD3	Complex	-7.035	0.00874	-4.992	5.04E-21

One transcription regulator that may regulate the DNA damage response noted in the canonical pathways results is forkhead box protein M1 (FOXM1), which is predicted to be activated in group B of both cohorts. This is supported at the mRNA level in our cohort where FOXM1 is upregulated 2.3 fold (FDR=3.5e-4). FOXM1 is upregulated in response to oxidative stress (via the TNF-α/reactive oxygen species/HIF-1 pathway)[403], and triggers upregulation of ROS scavenger genes including superoxide dismutase (SOD), PRDX3 and catalase (CAT)[404], which is confirmed by gene expression data in both cohorts (Table 7).

Table 7: Upregulation of oxygen scavenger genes in group B of both cohorts.

	FC (GeneST)	FDR (GeneST)	FC (Jorissen)	FDR (Jorissen)
SOD1	1.5	4.3E-03	1.2	5.7E-06
CAT	N/A	N/A	1.2	6.9E-05
PRDX3	1.3	5.0E-02	1.4	3.6E-08

Interestingly however, catalase was predicted to be a deactivated upstream regulator in B group CRCs of both cohorts (Table 6c), which may suggest increased oxidative damage, despite FOXM1 activation. Through its response to oxidative stress, FOXM1 facilitates tumour cell survival[405–407] by inducing proliferation and increasing resistance to apoptosis[403]. FOXM1 expression in CRC is associated with poor prognosis[408] and tumour metastasis is promoted through upregulation of MMPs and pro-angiogenic VEGF[406].

Another transcription factor related to oxidative stress and inflammation that might be relevant to the pathogenesis of group B cancers is X-box binding protein 1 (XBP1). XBP1 is a mediator of endoplasmic reticulum (ER) stress, which is induced by the accumulation of misfolded proteins. Cells with secretory functions, including Paneth and Goblet cells, are therefore particularly dependent on a competent unfolded protein response (UPR)[409]; its role in Paneth- and Goblet-cell function is underscored by the occurrence of unresolved colitis in XBP1 deletion mice, as well as the increased risk of IBD conferred by hypomorphic variants of XBP1[410]. Activation of XBP1 in B group CRCs therefore suggests that these experience increased ER stress, which is mediated by XBP1-related mechanisms, which is in contrast with the overall inflammatory features of this group.

Additional predicted upstream regulators of interest, that are related to tumour initiation and progression in group B cancers, include PIN1 (correlated with CRC progression[411]), ANGPT2 (induces angiogenesis[412]), PIM1 (a proto-oncogene[413]) and Jnk kinases[414].

Interestingly, several miRNAs, including the tumour suppressor let-7, and miRNA145, which suppresses the oncogene ANGPT2[415], are predicted to be deactivated in group B cancers.

Diseases and functions analyses

In our cohort, 110 diseases and functions showed significant activation/deactivation between subgroups ($p \leq 0.05$, $|\text{activation z-score}| \geq 2$). The top 20 scoring diseases and functions by p-value in our cohort were dominated by *DNA Replication*, *Recombination*, *Repair*, *Cell Cycle* and *Infectious Disease* categories (Table 8a). Meanwhile, in the Jorissen cohort the top 20 (of 261) significant diseases and functions categories

included *Cellular Growth and Proliferation*, *Infectious Disease*, and *Cancer*, while the more detailed annotations (column 2, Table 8b) revealed functions related to tumour progression and metastasis in B group CRC (Table 8b).

Table 8a: Top 20 most significantly different IPA diseases and functions group B vs. group A samples in our cohort. A p-value of overlap of 0.05 and |z-score| cutoff of 2 was used.

Categories	Diseases or Functions Annotation	p-Value	Activation z-score
Cell Cycle, DNA Replication, Recombination, and Repair	checkpoint control	7.43E-16	3.38
DNA Replication, Recombination, and Repair	DNA replication	5.96E-14	2.439
DNA Replication, Recombination, and Repair	repair of DNA	1.05E-11	3.485
Cell Cycle	cell cycle progression	5.97E-11	3.578
Infectious Disease	infection of cells	7.81E-09	9.811
Gene Expression, Protein Synthesis	translation of RNA	1.02E-08	-3.348
Cell Cycle	M phase	1.28E-08	3.17
Gene Expression, Protein Synthesis	translation of mRNA	1.93E-08	-3.348
Cellular Compromise, DNA Replication, Recombination, and Repair	damage of chromosomes	2.05E-08	-3.861
DNA Replication, Recombination, and Repair	metabolism of DNA	6.24E-08	3.059
Cellular Compromise, DNA Replication, Recombination, and Repair	breakage of chromosomes	1.74E-07	-3.422
Cell Cycle	interphase	2.32E-07	3.233
Cell Cycle	cycling of centrosome	5.55E-07	2.759
Infectious Disease	infection by RNA virus	5.75E-07	10.075
Infectious Disease	infection by HIV-1	5.77E-07	9.176
Infectious Disease	infection of tumor cell lines	6.62E-07	7.739
Infectious Disease, Reproductive System Disease	infection of cervical cancer cell lines	6.76E-07	7.737
Cell Cycle, DNA Replication, Recombination, and Repair	S phase checkpoint control	6.85E-07	2.23
Cell Cycle, DNA Replication, Recombination, and Repair	checkpoint control of tumor cell lines	7.64E-07	2.059
Cell Cycle	senescence of cells	1.17E-06	-3.239

Table 8b: Top 20 most significantly different IPA diseases and functions group B vs. group A samples in the Jorissen cohort. A p-value of overlap of 0.05 and |z-score| cutoff of 2 was used.

Categories	Diseases or Functions Annotation	p-Value	Predicted Activation State	Activation z-score
Cellular Growth and Proliferation	proliferation of cells	6.09E-46	Increased	6.123
Cellular Movement	cell movement	2.13E-35	Increased	7.194
Cellular Movement	migration of cells	1.42E-33	Increased	7.233
Cellular Movement	invasion of cells	3.58E-32	Increased	4.663
Hematological System Development and Function, Tissue Morphology	quantity of leukocytes	1.71E-31	Increased	5.323
Cell Death and Survival	cell death of immune cells	8.35E-30	Increased	3.67
Hematological System Development and Function, Tissue Morphology	quantity of blood cells	1.50E-28	Increased	4.982
Cell Death and Survival	cell death of blood cells	1.52E-28	Increased	3.468
Cancer	advanced malignant tumor	2.44E-28	Increased	2.61
Cancer	growth of tumor	6.45E-28	Increased	3.179
Infectious Disease	Viral Infection	9.93E-28	Increased	5.851
Cancer	metastasis	3.59E-27	Increased	2.61
Cellular Function and Maintenance	function of blood cells	7.51E-26	Increased	2.927
Cellular Development, Cellular Growth and Proliferation, Hematological System Development and Function	proliferation of immune cells	1.34E-25	Increased	2.709
Cellular Function and Maintenance	function of leukocytes	2.27E-25	Increased	2.715
Cellular Development, Cellular Growth and Proliferation	proliferation of tumor cell lines	8.51E-25	Increased	3.402
Cellular Movement	cell movement of tumor cell lines	4.11E-24	Increased	4.435
Cellular Movement	cell movement of	5.40E-24	Increased	8.081

	blood cells			
Cellular Development, Cellular Growth and Proliferation	proliferation of blood cells	6.64E-24	Increased	2.627
Cellular Movement	migration of blood cells	9.14E-24	Increased	8.181

Next, cohort comparison identified 11 diseases and functions that overlapped between the two data sets, six of which fell into the *Infectious Disease* category (Table 8c).

Table 8c: IPA diseases and functions categories shared between subgroups in our cohort (columns 3&4) and that of the Jorissen cohort (columns 5&6). A p-value of overlap of 0.05 and |z-score| cutoff of 2 was used.

Categories	Diseases or functions annotation	p-value	Activation z score	p-value (Jorissen)	Activation z score. (Jorissen)
Protein Synthesis	metabolism of protein	0.000524	2.722	4.07E-09	2.824
Molecular Transport, Protein Trafficking	transport of protein	0.0022	3.648	1.25E-11	2.364
Cell Death and Survival	cell viability of tumor cell	0.000651	7.672	3.49E-14	4.475
Cell Death and Survival	cell survival	0.00393	8.121	6.58E-22	7.171
Cellular Growth and Proliferation	proliferation of cells	0.0019	8.856	6.09E-46	6.123
Infectious Disease	infection by lentivirus	3.47E-06	9.31	4.27E-08	7.254
Infectious Disease	HIV infection	4.09E-06	9.419	6.95E-08	7.356
Infectious Disease	infection by Retroviridae	4.78E-06	9.423	5.37E-08	7.442
Infectious Disease	infection of cells	7.81E-09	9.811	1.49E-11	7.583
Infectious Disease	infection by RNA virus	5.75E-07	10.075	2.01E-10	7.491
Infectious Disease	Viral Infection	9.10E-05	10.224	9.93E-28	5.851

Interestingly increased *Viral Infection* was indicated in B group CRCs of both cohorts as well as decreased *Bacterial Infection* in the Jorissen cohort (p = 2e-10, z-score = -3). It is not immediately apparent whether the predicted increase in the *Viral Infection* category is based on a) a relative decrease in viral response mechanisms, which translates to increased viral infection, or b) a relative increase in viral response mechanisms, which is indicative of increased viral infection. Heatmaps of the

underlying gene expression data show a general increase in genes implicated in *Viral Infection* (which is predicted to be increased in group B) and in *Bacterial Infection* (which is predicted to be decreased in group B) in B group CRCs. The *Viral* and *Bacterial Infection* functions were based on 912 and 218 genes, respectively that were differentially expressed between group A and B cases of the Jorissen cohort, and 124 genes overlapped between *Viral* and *Bacterial Infection* functions.

Additionally, *cellular proliferation* and *cell viability*, as well as *metabolism and transport of proteins*, were predicted to be increased in B group CRCs of both cohorts.

PARADIGM analysis

In addition to using IPA to compare CRC subtypes at the pathway-level, PARADIGM was applied in a similar manner since PARADIGM has the advantage of allowing integration of multiple omics data types, as well as providing results on a per-patient basis.

Regarding our cohort, differential analysis of PARADIGM IPLs produced 1464 IPLs that were differentially activated (FDR $\leq$ 0.05, |difference in group medians| $\geq$ 0.25) between group A and B samples, including 712 genes or gene families (Appendix C, Table 1), 711 complexes (Appendix C, Table 2), and 41 abstract processes (Appendix C, Table 3). In the Jorissen cohort 3619 IPLs were differentially activated between groups (FDR $\leq$ 0.05, absolute difference in group medians $\geq$ 0.25), including 1922 genes or gene families (Appendix C, Table 4), 1609 complexes (Appendix C, Table 5), and 88 abstract processes (Appendix C, Table 6).

Of the 1464 IPLs differentially activated between subtypes in our cohort, 570 were also present in the Jorissen cohort analysis, 499 (88%) of which had a consistent direction of activation- or deactivation between the two cohorts. Of the 499 IPLs, the shared abstract processes included DNA damage response-related pathways, *activation of caspase activity by cytochrome c* and *prostaglandin biosynthesis* (Table 9). The results from PARADIGM therefore support the results from IPA, and provide an additional layer of validation for the subtypes defined here.

Table 9. PARADIGM-derived abstract processes significantly altered (FDR ≤ 0.05, absolute median difference between subgroups ≥ 0.25) between RPMM-subgroups in our cohort as well as the Jorissen cohort.

Integrated pathway level (IPL)	FDR	Subgroup diff.	FDR (Jorissen)	Subgroup diff. (Jorissen)
negative_regulation_of_DNA_binding_(abstract)	6.31E-03	2.7	7.76E-16	2.6
anoikis_(abstract)	3.66E-02	2.2	3.68E-16	3.3
protein_catabolic_process_(abstract)	8.22E-04	1.9	3.81E-15	1.2
G2/M_transition_DNA_damage_checkpoint_(abstract)	2.26E-02	1.9	7.77E-15	1.6
response_to_radiation_(abstract)	4.74E-02	1.7	8.62E-17	2.3
DNA_damage_checkpoint_(abstract)	1.27E-03	1.6	1.99E-04	0.9
regulation_of_transcription_(abstract)	1.21E-03	1.6	7.59E-04	0.6
prostaglandin_biosynthetic_process_(abstract)	2.36E-02	1.3	3.58E-07	0.7
spindle_assembly_(abstract)	1.99E-02	1.2	1.37E-02	0.6
regulation_of_mitotic_centrosome_separation_(abstract)	3.72E-03	1.2	3.22E-02	0.3
protein_folding_(abstract)	2.24E-03	1.2	7.27E-04	0.9
G1/S_transition_checkpoint_(abstract)	8.83E-03	1.0	1.59E-03	0.4
Golgi_organization_(abstract)	8.32E-03	1.0	4.35E-02	0.6
activation_of_caspase_activity_by_cytochrome_c_(abstract)	2.64E-02	0.8	3.34E-16	1.0
cell_cycle_arrest_(abstract)	3.09E-02	-1.2	2.08E-06	-0.5
ribosome_biogenesis_(abstract)	3.24E-02	-2.9	8.79E-13	-2.7

Subgroup diff.: median difference in PARADIGM activity score between each subgroup. The first two columns represent the FDR and difference in medians for our cohort

DNA and protein damage response: exogenous or endogenous triggers?

Since previous studies have shown that *E. faecalis* induces DNA damage and chromosomal instability *in vitro* through ROS generation[223,416,417], the contribution of *E. faecalis* to the upregulation of DNA and protein damage response pathways identified using pathway analyses was assessed. Heatmaps of the canonical pathways including *Role of BRCA1 in DNA damage response, Cell cycle G2/M DNA damage checkpoint regulation,* and *Protein Ubiquitination* were produced using the list of genes predicted

to be altered in group B samples for each pathway in IPA. Figure 8 shows the result for the *Cell cycle G2/M DNA damage checkpoint regulation* pathway, which was the top scoring canonical pathway when comparing *E. faecalis*+ to *E. faecalis*–CRCs, using genes differentially expressed with a p-value ≤ 0.05 (results not shown) as input.

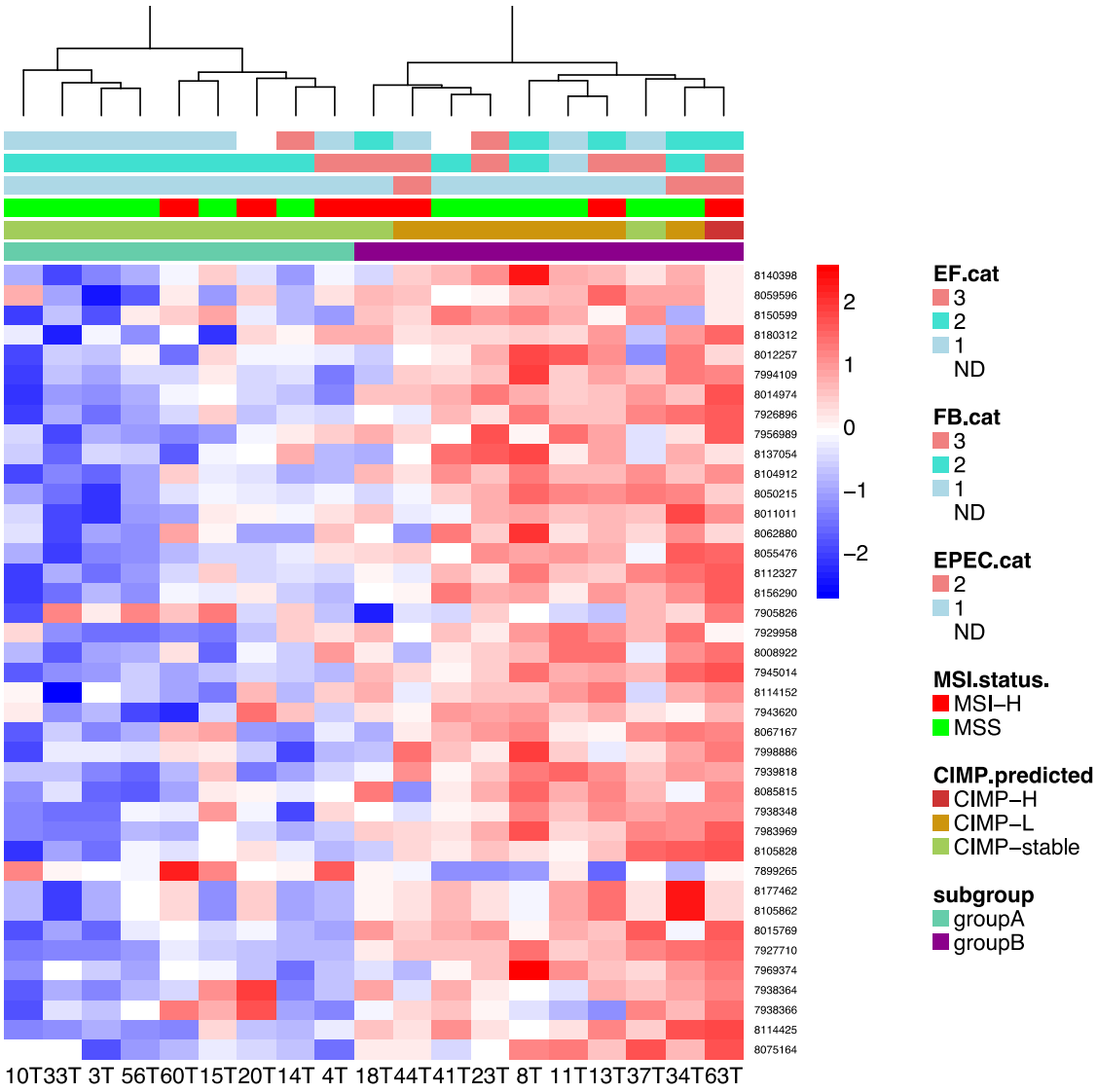


Figure 8: Hierarchical clustering of genes implicated in the *Cell cycle G2/M DNA damage checkpoint regulation* pathway in IPA. The legend categories on the right are presented in the same order as the row annotations at the top of the graph. The scale on the right represents row-scaled expression values. EF.cat: *E. faecalis* colonisation level category (1=negative; 2=low-level; 3=high-level); FB: Fusobacterium colonisation level category (1=negative; 2=low-level; 3=high-level); EPEC.cat: EPEC colonization (1=negative; 2=positive); ND: not determined.

Clearly, although six of the seven *E. faecalis*+ samples fall in the cluster where the Cell cycle G2/M DNA damage checkpoint regulation pathway is upregulated, this response

is not exclusive to *E. faecalis*+ samples, but rather to group B samples in general. It is thus conceivable that ROS might instead be produced by neutrophils and macrophages through the induction of oxidant-generating enzymes including the NOX (NADPH oxidase) and DUOX (Dual oxidase) enzymes, as well as myeloperoxidase (MPO) and inducible nitric oxide synthase (iNOS)[418]. However, no up-regulation of oxidant-generating enzymes was seen in B group CRCs, and in fact *NOX1* was significantly downregulated in Jorissen G group cancers (FC=–2.3) and *NOXA1* was significantly downregulated in B group CRCs of both cohorts (both with FCs of ~ –1.3), which could support the contribution of exogenous ROS by bacteria in B group CRCs. However, since there is no direct evidence of increased ROS and/or NOS in these cancers, it is also possible that the DNA damage response is activated by factors other than increased ROS and/or NOS.

Do MSI-specific antigens trigger the biological response seen in B group cancers?

Compared to their microsatellite-stable counterparts, MSI+ CRCs have a heightened immune response that is evident macroscopically and at the molecular level[419]. This tumour-specific immune response is caused by antigenic MSI-induced frameshift mRNAs and/or peptides[420,421]. The contribution of MSI to subtype-specific alterations in immune-related functions were therefore investigated by comparing canonical pathways of particular biological interest in B group CRCs (related to inflammation, infection and oxidative stress) between a) MSI vs. MSS cancers (using genes differentially expressed between MSI vs. MSS cancers as input ($|FC| \geq 1.25$, FDR $\leq$ 0.05) and b) B group vs. A group cancers of the Jorissen cohort ($|FC| \geq 1.25$, FDR $\leq$ 0.05). The results (presented in Table 10) clearly demonstrate that the increase in inflammation, oxidative stress response, and antigen presentation in B group cancers is not specific to MSI+ cancers, since stronger evidence for upregulation of these pathways is seen in B group CRCs as a whole (of which 21% are MSS).

Table 10: comparison of canonical pathways with particular biological interest in B group vs. A group CRCs compared to MSI vs. MSS status in the same cohort.

Ingenuity Canonical Pathways	-log(p-value) (MSI vs. MSS)	Down-regulated (MSI vs. MSS)	Up-regulated (MSI vs. MSS)	-log(p-value) (group B vs. A)	Down-regulated (group B vs. A)	Up-regulated (group B vs. A)
Antigen Presentation Pathway	6.28	1/34 (3%)	22/34 (65%)	9.28	1/34 (3%)	28/34 (82%)
Colorectal Cancer Metastasis Signaling	4.65	46/230 (20%)	67/230 (29%)	5.27	20/230 (9%)	89/230 (39%)
Production of Nitric Oxide and Reactive Oxygen Species in Macrophages	3.61	36/179 (20%)	64/179 (36%)	3.52	18/179 (10%)	64/179 (36%)
VEGF Signaling	2.73	19/89 (21%)	25/89 (28%)	3.61	10/89 (11%)	36/89 (40%)
Role of Pattern Recognition Receptors in Recognition of Bacteria and Viruses	2.53	17/118 (14%)	42/118 (36%)	2.51	8/118 (7%)	46/118 (39%)
IL-17 Signaling	2.47	12/72 (17%)	29/72 (40%)	5.94	4/72 (6%)	40/72 (56%)
IL-8 Signaling	2.19	32/183 (17%)	55/183 (30%)	4.14	15/183 (8%)	71/183 (39%)
Protein Ubiquitination Pathway	1.32	40/251 (16%)	90/251 (36%)	6.54	21/251 (8%)	101/251 (40%)
Role of BRCA1 in DNA Damage Response	1.13	12/61 (20%)	16/61 (26%)	1.85	12/61 (20%)	17/61 (28%)
Cell Cycle: G2/M DNA Damage Checkpoint Regulation	7.53E-01	13/49 (27%)	13/49 (27%)	1.79	7/49 (14%)	17/49 (35%)
ATM Signaling	4.80E-01	9/59 (15%)	18/59 (31%)	2.77	6/59 (10%)	25/59 (42%)

An important consequence of the anti-tumoural immune response in MSI+ CRCs is selective pressure towards immune evasion[421]. Mechanisms that contribute towards immune evasion in MSI+ CRCs include alteration in antigen-presentation machinery, specifically in HLA class I-mediated antigen presentation, which can be compromised

by mutations in *B2M* (30–60% of MSI+ CRCs) or through loss or downregulation of HLA class I heavy chains (~60% of MSI+ CRCs)[421]. Further, alterations in antigen processing machinery may also facilitate immune evasion[421]. *HLA* gene expression was therefore compared between a) MSI vs. MSS cancers, and b) B group vs. A group CRCs in the Jorissen cohort. Strikingly, multiple *HLA* class II (*HLA-D*) genes were preferentially upregulated in B group CRCs, while no *HLA* class I genes were differentially expressed in either comparison (Table 11). These results suggest induction of MHC class II antigen presentation in B group CRCs (most likely by tumour infiltrating lymphocytes), which is not specific to MSI+ cancers. These findings provide further support for the presence of foreign antigens in B group CRCs, which may well be of microbial origin.

Table 11: Comparison of HLA genes significantly differentially expressed in a) MSI vs. MSS cancers and b) group B vs. group A cancers of the Jorissen cohort.

Probeset ID	Gene symbol	FDR (MSI vs. MSS)	FC (MSI vs. MSS)	FDR (group B vs. group A)	FC (group B vs. group A)	Relative subtype-specific effect
201137_s_at	HLA-DPB1	3.0E-06	2.4	4.0E-14	3.6	increased
203932_at	HLA-DMB	3.3E-09	2.3	3.1E-10	2.4	increased
205671_s_at	HLA-DOB	3.7E-05	1.6	1.4E-02	1.3	decreased
208894_at	HLA-DRA	4.0E-06	3.1	9.9E-13	5.2	increased
209480_at	HLA-DQB1	9.9E-03	1.8	6.4E-03	1.8	decreased
209823_x_at	HLA-DQB1	2.3E-06	2.3	1.2E-12	3.3	increased
210982_s_at	HLA-DRA	2.0E-06	3.0	5.8E-13	4.7	increased
211654_x_at	HLA-DQB1	2.1E-05	2.1	5.5E-07	2.3	increased
211656_x_at	HLA-DQB1	1.1E-05	1.9	1.8E-10	2.4	increased
211990_at	HLA-DPA1	1.7E-06	2.8	4.0E-10	3.6	increased
211991_s_at	HLA-DPA1	5.7E-07	3.1	2.9E-12	4.4	increased
212998_x_at	HLA-DQB1	1.2E-05	2.7	1.4E-08	3.4	increased
212999_x_at	HLA-DQB1	1.5E-03	1.7	4.2E-04	1.7	increased
213537_at	HLA-DPA1	4.0E-06	2.6	1.9E-07	2.8	increased

213831_at	HLA-DQA1	8.2E-03	2.1	1.3E-05	3.0	increased
217362_x_at	HLA-DRB6	5.2E-07	1.7	4.0E-10	1.9	increased
217478_s_at	HLA-DMA	7.8E-08	3.0	4.4E-10	3.4	increased
226878_at	HLA-DOA	1.5E-05	1.9	2.2E-06	1.9	increased
236203_at	HLA-DQA1	8.2E-04	1.6	3.3E-03	1.5	decreased

Relative subtype-specific effect indicates a relative increase or decrease in the result comparing the CRC subtype-specific effect to the MSI-specific effect.

Discussion

Specific bacterial infection, inflammation and DNA and protein damage responses underlie a distinct genomic subtype of CRC

A subtype of CRC that is associated with a relative increase in the frequency of colonisation by *E. faecalis* and high-level colonisation with *Fusobacterium*, CpG island methylation (predicted using a CIMP-specific array-based marker panel) and MSI (as judged from the Jorissen cohort) was identified. These cancers are most similar to de Sousa et al. CCS2/3 subtypes and the inflammatory subtype of Sadanandam et al., which is supported by clinical observations in these patients. Moreover, pathway analyses of gene expression data in our cohort, as well as a larger publically available cohort (GSE13294), revealed increased response to DNA and protein damage, infection, inflammation and proliferation in these cancers. Various genes and pathways linked to increased metastasis and CRC progression were also highlighted in B group CRCs.

One clinically relevant feature of inflammatory group B cancers is the significant upregulation of the *COX-2* in both cohorts, which suggests that aspirin (which inhibits *COX-2*) might be particularly useful in preventing these cancers—an idea supported by pathway analysis of the Jorissen cohort, which lists aspirin as a 'deactivated' upstream regulator (z-score –4.4 p-val 9.6e-8). Moreover, regular prophylactic aspirin use specifically reduces the risk of developing CRCs that overexpress *COX-2*[422]. On the other hand, aspirin apparently helps prevent *BRAF* wild type but not *BRAF* mutant CRCs[423], and the latter is strongly associated with CIMP-H cancers, which may be more frequent in group B cancers. Nevertheless, the level of protection conferred by aspirin in relation to CRC subtypes is certainly intriguing and warrants further investigation,

especially since recent studies offer compelling evidence for the use of aspirin as an adjuvant therapy for CRC—aspirin use (before or after diagnosis) improved patient survival and reduced the risk of metastatic CRC[131].

Given the occurrence of MSI-specific antigens, which is accompanied by an increased immune response in these cancers, the contribution of MSI to the activation of pathways related to antigen presentation, inflammation, DNA and protein damage response, and infection (seen in group B cancers) was investigated. Numerous *HLA* class II genes were preferentially upregulated in B group CRCs, indicating the presence of antigens unrelated to MSI-status in these tumours, which may be of microbial origin, given the increased subtype-specific response to inflammation and infection.

In addition to increased inflammation and response to infection in B group CRCs, these cancers exhibit an increased DNA and protein damage response at the pathway level. The transcription factors FOXM1 and XBP1 likely play important roles in regulating oxidative damage and inflammation in these samples, and overexpression of FOXM1 may facilitate tumour survival and progression in spite of DNA damage. The DNA and protein damage response seen in B group CRCs was not specific to *E. faecalis*+ samples, where one might expect ROS-induced DNA damage. However, the finding that oxidant-generating enzymes were not upregulated in B group CRCs supports the possibility of exogenous bacterially-produced ROS. However, it is also possible that the DNA damage response is activated by factors other than ROS and/or NOS. The DNA damage response and protein ubiquitination was preferentially activated in B group CRCs, as opposed to MSI+ cancers, and it is therefore unlikely that MSI-induced mutations alone cause this response. One intriguing alternative is the well-established link between viral infection and the DNA damage response, which is exploited by several viruses to facilitate incorporation into the host genome[424]. However, a study that examined the distribution of JC polyoma virus (JCV), human adenovirus (AdV), Epstein–Barr virus (EBV), Kaposi sarcoma-associated herpesvirus (KSHV/HHV8) and human papillomavirus (HPV) across a cohort of 185 sporadic CRCs found no association between CIMP status and any of the viruses[425]. Since CIMP+ cancers are

overrepresented in B group CRCs, it seems less likely that the DNA damage response is virally-induced, at least not by the viruses examined in the Karpinski et al. study.

While there appears to be a large-scale transcriptional response that links inflammation, DNA damage response, protein ubiquitination and immune response in group B cancers, it is unclear whether this translates to increased or decreased inflammation/oxidative-induced damage in group B relative to group A cancers. Since oxidative stress can cause an array of DNA damage, including MSI (through mRNA or protein-level downregulation of the MMR system)[137,138], mutations in non-microsatellite regions, DNA double strand breaks and CIN[418,426], it is tempting to speculate that oxidative stress might contribute to mutations and/or chromosomal instability in group B cancers. However, this requires further investigation.

Collectively, the role of infection, inflammation and oxidative stress in B group CRCs, bears resemblance to the pathogenesis of IBD-associated CRCs, where host susceptibility factors underlie an inappropriate response to bacteria, which results in increased inflammation and oxidative stress and an overrepresentation of specific bacterial families and species. Further, several genes previously linked to IBD were consistently upregulated in B group CRCs. Although there are distinct differences between IBD-associated and sporadic CRC[427], there are clearly also similarities at the level of infection, bacterial colonisation and oxidative stress, at least in a subset of cancers. Presumably, there is a continuum in the degree of chronic colonic inflammation across a population, in which case it is reasonable to assume that, similar to IBD-associated CRC, subclinical inflammation accompanied by shifts in microbial composition also confer an increased risk of developing CRC.

Importantly, the increase in IL-8 signaling and *COX-2* expression in group B CRCs closely resembles the NF-κB pro-inflammatory signature (which included COX-2, TNF-α and IL-8) described by Kostic et al. that is amplified in human *Fusobacterium*-infected CRC tissue[218]. They also found, using IPA, that the *Fusobacterium*-associated gene expression signature in human CRCs was highly enriched for the inflammatory response gene ontology category[218]. These findings provide support for the role of Fusobacterium in B group CRCs.

In Chapter 4, high-level colonisation by Fusobacterium was associated with both *pks*-positive *E. coli*, and Enteropathogenic *E. coli*. Interestingly Warren et al. revealed a subset of microbes (including *Leptotrichia* and *Campylobacter* spp.) that were significantly associated with *F. nucleatum* in CRC biopsies, using co-occurrence network analysis of metagenomic signatures[428,429]. Importantly, this polymicrobial signature was associated with over-expression of numerous host genes, including IL-8[428]. *F. nucleatum* is known to provide a scaffold for secondary bacterial colonisers in dental plaque, resulting in a structured biofilm[429–431]. Notably, *H. pylori* is known to selectively adhere to Fusobacterium[432], and *E. faecalis* co-aggregates with *F. nucleatum* in endodontic infections[433]. These findings suggest that Fusobacterium could facilitate colonisation of potentially oncogenic pathogens. Incidentally, *F. nucleatum* is able to adapt to oxidative stress[434,435] and in fact shows enhanced pathogenicity in mice under these conditions[436], with *F. nucleatum* strains originating from inflamed biopsy tissue of IBD patients being significantly more invasive in a Caco-2 cell invasion assay than strains that were isolated from healthy tissue from either IBD patients or control patients[262]. Moreover, this adaptation of *F. nucleatum* to oxidative stress can be conferred to other more strict anaerobes[437].

Study limitations and future recommendations

The major limitation of our study is the relatively small sample size, which limits statistical confidence in our findings. This issue was addressed retrospectively by applying our data analysis pipeline to a large publically available cohort where our method produced broadly similar results to the classifications of de Sousa et al. and Sadanandam et al. Further, closely related RPMM clusters were merged for ease of data analysis, but we recognize that there are likely more than two valid CRC subtypes. However, certain biologically relevant features appear to be shared by more than one subtype, as demonstrated by gene expression and pathway analyses between group A and B cancers.

There does not seem to be consensus in the literature regarding a variability cutoff to use to select the subset of genes used for clustering. This parameter affects the resulting clusters to a certain extent and the decision of using the top 25% most variable genes

(by median absolute deviation) was based on the results from MDS of the Jorissen cohort, which demonstrated better separation of clusters when using the top 25% of clusters compared to using a very stringent cutoff of the 1000 most variable probesets.

Nevertheless, biologically important features that distinguish the CRC subtypes defined here were identified that are likely applicable to the published subtypes of de Sousa et al. and Sadanandam et al.

Regarding the establishment of CIMP-status in our cohort, a published array-based marker panel, for which high sensitivity and specificity has been reported[339], was used. However, the gold standard for CIMP classification is by methylation-specific PCR and the agreement between array-based vs. molecular classification was not confirmed here.

Future work should thus be aimed at validating these findings in a larger cohort, which includes more rectal samples (since only 5/19 from our study and 25/155 of the Jorissen study were rectal samples). Investigating the level of ROS and NOS by CRC subtype, and the association between subgroups and CRC risk factors that have inflammatory or pro-oxidative potential other than bacteria (including viral infection, cigarette smoking, alcohol- and red meat-consumption), may shed further light on the pathogenesis of B group CRCs. In addition, screening patients for single nucleotide polymorphisms known to modify IBD risk may shed light on the cursory resemblance between type B CRCs and IBD observed here, particularly regarding the host's susceptibility to infection and the ability to control inflammation and oxidative stress. Any significant associations could have clinical utility for identifying individuals with an increased risk of developing type B CRC.

Conclusion

Here, direct evidence of altered CRC pathogenesis associated with increased bacterial colonisation by high levels of Fusobacterium and by *E. faecalis* in our group B cohort is provided. The relative importance of infection in B group CRCs is also supported by pathway analysis in the Jorissen cohort where an elevated host-response to infection, which is accompanied by an increased inflammatory response, is predicted at the pathway level. Further, although not exclusively linked to *E. faecalis*, this bacterium

may contribute to oxidative stress in these samples by production of superoxide. This is supported by the lack of evidence for overexpression of oxidant generating host enzymes in B group CRCs.

Given the positive association between colonisation with Fusobacterium and several other bacteria that was identified in this study and elsewhere, we hypothesize that colonisation by Fusobacterium (particularly at high levels) could facilitate colonisation with secondary (potentially oncogenic) pathogens in the colon; moreover, a pro-oxidative environment–which is likely found in group B CRCs–could increase the pathogenic potential of Fusobacterium, as indicated by previous reports.

We propose that a subtype of CRC (which generally corresponds to the de Sousa CCS2/3 subtypes) is associated with a relative increase in colonisation by *E. faecalis*, high-level Fusobacterium (and likely other bacteria not examined here), inflammation, CpG island methylation and microsatellite instability, as well as an increased response to oxidative DNA and protein damage. This subtype is further characterised by a significant upregulation of *COX-2*, which suggests that aspirin in combination with selective antibiotics might be particularly useful in preventing and treating these cancers.

This is the first study to link colonisation by specific bacteria to a transcriptomic subtype of colorectal cancer. The enrichment with both *E. faecalis* and Fusobacterium, together with the relative increase in inflammatory pathways in this subtype, suggests that polymicrobial colonisation of the colonic epithelium and/or tumour may be an important aspect of colonic tumourigenesis that warrants further investigation. Our findings underscore the value of quantifying suspected oncogenic bacteria alongside transcriptomic data in elucidating molecular origins and optimal therapeutic strategies by colorectal cancer subtype.

Chapter 8: Gene expression analyses reveals *E. faecalis*- and *Fusobacterium*-associated genomic alterations in colorectal cancer

Abstract

In addition to characterising the genomic subtypes of CRC, and investigating the relevant patterns of bacterial colonisation, described in Chapter 7, we were interested in identifying transcriptomic and pathway-level changes that were specific to CRCs colonised by each of the bacteria studied here.

Whole-genome differential gene expression analysis was therefore performed for nineteen pairs of tumour and adjacent normal colorectal cancer (CRC) samples and bacteria-associated gene expression profiles were investigated for Fusobacterium, *Enterococcus faecalis*, Enterotoxigenic *Bacteroides fragilis*, Enteropathogenic *Escherichia coli*, ClB+ *E. coli* and afaC+ *E. coli*.

A relatively large subset of genes was differentially expressed in *E. faecalis*-infected CRCs (489 at an FDR ≤ 0.05); included in this subset were various small leucine-rich proteoglycans (*SLRP*s), *CXCL10*, and *BMI1*.

Meanwhile comparison between samples with high-level Fusobacterium infection compared to low-level or no colonisation by Fusobacterium revealed differential expression of the regenerating islet-derived-family genes REG1A, REG3A and REG1P (FDRs ≤ 0.05)–these were the only genes differentially expressed in association with high-level colonisation by Fusobacterium.

Pathway analysis of genes differentially expressed in association with *E. faecalis* colonisation uncovered pathways related to immune function (in particular antigen presentation and microbial pattern recognition), inflammation and CRC progression, which may indicate an important role in the pathogenesis of a subset of CRCs.

Introduction

Previous studies have successfully demonstrated pathogen-induced alterations of the host transcriptome *in vitro*[438] and *in vivo*[439], using microarray-based gene expression analysis. Upon infection of the host, one might expect a) a virulence factor-specific

response in a particular host signaling pathway (e.g. as a result of injected bacterial effectors) or b) a more general immune/inflammatory response in the host that is not necessarily limited to the presence of a single pathogen, and may be linked to dysbiosis.

In *in vitro* models of infection[438], or in mono-associated mouse models[440], differential expression analysis together with pathway analysis can be used to identify any genes/pathways that are specifically altered in response to the bacterium. Potentially confounding factors in clinical CRC samples include the plethora of interspecies interactions in the gut microbiome, cancer-specific gene expression changes (that may overlap with bacterially-induced immune-related response), inter-individual variation, site-specific variation, as well as sampling and technical variation. Therefore, although one may not be able to determine bacterially-*induced* alterations in clinical CRC samples, any bacterially-*associated* signatures obtained provide a true reflection of the system in question that cannot be readily mimicked *in vitro* or in mouse models. These bacterially associated signatures will hopefully provide novel insights and drive hypotheses regarding the poorly understood role of CRC-associated bacteria in the pathogenesis of the disease.

In this study, building a profile of CRC-associated bacteria across a single CRC cohort comes at the cost of variable statistical power to detect species-specific gene expression signatures. In a small cohort such as ours (N=19), one would expect to miss smaller fold change effects, especially for species that are underrepresented in our cohort.

The putative oncogenic mechanisms of CRC-associated pathogens were discussed in Chapter 4. Here, host gene expression changes previously reported for *E. faecalis* and Fusobacterium will be briefly discussed.

Enterococcus faecalis: an overview

Enterococci are normal human commensals that inhabit the gastrointestinal tract, the oral cavity and the vagina[441], yet are ranked among the top three nosocomial (hospital-acquired) bacterial pathogens. Up to 90% of Enterococci infections are caused by *E. faecalis* and antibiotic resistance poses a major problem in the treatment of these infections[442]. Moreover, *E. faecalis* is known for its capacity to transfer antibiotic

resistance to other pathogens[442]. Enterococci are incredibly resilient pathogens that can withstand and adapt to a variety of harsh environmental conditions that may provide them with a competitive advantage[441].

Several virulence factors are more frequently detected in patients with pathogenic *E. faecalis* infection compared to faecal isolates from healthy controls; these include aggregation substance (AS), surface adhesins, sex pheromones, lipoteichoic acid, extracellular superoxide, gelatinase, hyaluronidase, and cytolysin[441]. AS facilitates adhesion and phagocytosis by macrophages and neutrophils, allowing intracellular survival of *E. faecalis*.

As discussed in Chapter 7, *E. faecalis* produces DNA-damaging superoxide, the level of which is strain-specific[441]. Additionally, *E. faecalis* infection of neutrophils or macrophages results in an increase or decrease in host-derived superoxide production, respectively[441].

Cell extracts of AS- and bacteriocin-positive *E. faecalis* induce T-cell proliferation, with subsequent release of TNF-β and IFN-γ; these cell extracts also activate macrophages to release tumor necrosis factor alpha TNF-α. Similarly, lipoteichoic acid, a structural cell-wall component present in many gram-positive bacteria including *E. faecalis*, stimulates leukocytes to release inflammatory TNF-α, IL-1β, IL-6 and IL-8[441].

In IL-10$^{-/-}$ mice, *Enterococcus* spp. *(E. faecalis* and *E. faecium)* induce an inflammatory response similar to that seen in IBD[440]. Compared to control mice, genes differentially expressed in response to *Enterococcus* spp. infection were associated with pathways related to inflammatory disease, immune response, antigen presentation (particularly major histocompatability complex Class II), fatty acid metabolism and detoxification[440].

E. faecalis is found at significantly higher levels in stool samples from CRC patients compared to healthy controls[198] and oncogenic potential has been suggested based on its production of extracellular superoxide which leads to inflammation, DNA damage and CRC in IL-10 knockout mice[199,202,223]. *E. faecalis* also induces aneuploidy and tetraploidy in vitro[223].

Fusobacterium: an overview

Fusobacteria are well known for their role in periodontal disease. More recently, Fusobacterium spp., and in particular *F. nucleatum,* has been identified as a common bacterium in CRC patients, with a marked increase in the colonisation of tumour compared to adjacent normal mucosal tissue [16,194,219]. This relationship was confirmed in our cohort, where Fusobacterium occurred at significantly higher levels in tumour samples (p=0.0003). In Chapter 7, the ability of *F. nucleatum* to provide a scaffold for secondary bacterial colonisers[429–431], including *E. faecalis*[433] was discussed. Moreover, co-aggregation structures may be relevant in the pathogenesis of CRC, since a subset of microbes are significantly associated with *F. nucleatum* in CRC biopsies[428,429]. In addition to providing a structural scaffold for secondary colonisers, *F. nucleatum* is also able to invade and replicate in epithelial cells[443].

Regarding its oncogenic potential, $APC^{Min/+}$ mice infected with *F. nucleatum* show accelerated tumourigenesis, characterised by infiltration of specific myeloid cell subsets into tumours and an NF-kB pro-inflammatory signature (including COX-2, TNF-α and IL-8); these findings were confirmed by Kostic et al. in human Fusobacterium-infected CRCs[218]. They utilized a deep transcriptome sequencing data set (i.e., RNA-seq) including 133 colon tumors, generated by The Cancer Genome Atlas, to profile both Fusobacterium spp. colonisation and host gene expression profiles. They identified Fusobacterium-associated genes by calculating the Spearman's rank correlation coefficient of the relative abundance of Fusobacterium transcripts with host gene expression; these included *COX-2, IL-1β, IL-6, IL-8, TNF-α*, and *MMP3*[218].

Methods

Sample preparation and microarray-based gene expression analysis

Nineteen paired tumour and normal samples were selected for gene expression analysis, based on (i) the availability of tumour, normal and blood samples, as well as complete consent forms for each patient and on (ii) site of disease, radiotherapy-status, ethnicity, tumour stage and recurrence. Patients who had preoperative radiotherapy were excluded in order to avoid any additional confounding factors.

RNA preparation and gene expression analysis on Affymetrix Human Gene 1.0 ST arrays, as well as data preprocessing and quality control, are described in Chapter 5.

Differential gene expression analysis

ComBat-based batch and quality correction (described in detail in Chapter 5) was performed for each bacterial comparison individually, by specifying the comparison of interest in the model. Differential expression analysis was then performed on the subset of transcriptclusters that mapped to Entrez Gene Symbols, using the R package limma[229]. This left 21934 of the original 33297 transcriptclusters for analysis, which excluded control probes and transcripts with poor annotation. Differential analyses in tumour and normal samples were conducted separately and we compared a) samples with vs. without colonisation by a particular bacterium and b) samples with high vs. low/no-colonisation by a particular bacterium. Comparisons were only made where at least three samples per group were available, as summarized in Tables 1a and 1b.

Pathway analysis

Ingenuity Pathway Analysis (IPA) was applied to the subset of genes significantly altered between groups of interest (FDR $\leq$ 0.05; |FC| $\geq$ 1.25). The IPA categories: canonical pathways, upstream regulators and diseases and functions are included here.

PARADIGM analysis was conducted on tumour samples using whole-genome gene expression and methylation data as input, as described in Chapter 7. To assess the difference between groups of interest using PARADIGM pathway analysis results, the R package limma[229] was used to identify IPLs that showed differential activity between groups of interest; IPLs with an FDR $\leq$ 0.05 and an absolute difference in median activity score between groups of at least 0.25 were deemed significant.

Results

Bacteria-associated gene expression analysis

To investigate the putative effect of CRC-associated bacteria on host gene expression, differential analyses were conducted separately in tumour and normal samples to

compare, for each bacterium a) samples with vs. without colonisation and b) samples with high vs. low/no-infection. The comparisons made, and the number of transcript clusters differentially expressed for each comparison, are listed in Table 1a (tumour samples) and Table 1b (normal samples).

After correcting for multiple testing, only the analysis by *E. faecalis* infection-status in tumour samples produced significant results where 509 transcript clusters corresponding to 489 unique genes, were differentially expressed at an FDR ≤ 0.05. At a more relaxed FDR cutoff of 0.25, the afaC+ vs. afaC– comparison in normal samples produced 3811 transcript clusters, and Fusbacterium-high vs. Fusobacterium-low/neg in normal samples produced 210 transcript clusters; in tumour samples Fusbacterium-high vs. Fusobacterium-low/neg and EPEC+ vs. EPEC– both produced 13 transcript clusters. However, downstream analyses and interpretations were only conducted on results with an FDR ≤ 0.05, and an absolute fold-change of at least 1.25.

Table 1a: Summary of differential gene expression analyses conducted for *tumour* samples showing the number of transcript clusters differentially expressed for each comparison made.

Model specified	Number of samples/category	p ≤ 0.05	FDR ≤ 0.25	FDR ≤ 0.05
EF+ vs. EF–	7 vs. 10	6032	4216	509
FB-H vs. FB-L/N	7 vs. 12	2184	13	3
afaC-H vs. afaC-L/N	5 vs. 14	860	7	0
afaC+ vs. afaC–	13 vs. 8	1596	0	0
ClB+ vs. ClB–	6 vs.13	1719	0	0
ETBF-H vs. ETBF L/N	3 vs. 16	1000	1	0
ETBF+ vs. ETBF–	10 vs. 9	1584	1	0
EPEC+ vs. EPEC–	3 vs. 16	1720	13	0

H: high-level infection; L: low-level infection; N: no infection; FB: *Fusobacterium*; EF: *E. faecalis*

Table 1b: Summary of differential gene expression analyses conducted for *normal* samples showing the number of transcript clusters differentially expressed for each comparison made.

Model specified	Number of samples/category	p ≤ 0.05	FDR ≤ 0.25	FDR ≤ 0.05
ETBF+ vs. ETBF-	10 vs. 9	611	0	0
ETBF-H vs. ETBF-L/N	4 vs. 20	400	0	0
ClB+ vs. ClB-	6 vs. 13	1140	0	0
afaC+ vs. afaC-	10 vs. 9	4185	3811	1
afaC-H vs. afaC-L/N	3 vs. 16	718	0	0
EF+ vs. EF-	4 vs. 10	1503	0	0

FB-H vs. FB-L/N	3 vs. 16	2461	210	0
FB+ vs. FB-	16 vs. 3	797	0	0

H: high-level infection; L: low-level infection; N: no infection; FB: *Fusobacterium*; EF: *E. faecalis*

Enterococcus faecalis-associated genomic alterations in CRCs

Genes differentially expressed in association with E. faecalis colonisation

The top 50 most significantly differentially expressed genes by *E. faecalis* colonisation status are presented in Table 2; given the overrepresentation of *E. faecalis* in group B CRCs (Chapter 7), genes that were also significant in the comparison between group A vs. group B CRCs are highlighted in boldface for clarity.

Table 2: The top 50 most significantly differentially expressed genes by *E. faecalis* colonisation status. Boldface entries were also significant in the comparison between group A vs. group B CRCs in Chapter 7.

Gene ID	Gene symbol	GeneName	P value	FDR	FC (EF+ vs. EF–)
8166906	GPR34	G protein-coupled receptor 34	2.6E-07	3.9E-03	2.4
8126784	PLA2G7	phospholipase A2, group VII (platelet-activating factor acetylhydrolase, plasma)	4.3E-07	3.9E-03	2.7
8138289	ETV1	ets variant 1	5.3E-07	3.9E-03	2.3
7965410	DCN	decorin	2.5E-06	8.6E-03	3.9
8100541	IGFBP7	insulin-like growth factor binding protein 7	2.6E-06	8.6E-03	2.9
8101126	CXCL10	chemokine (C-X-C motif) ligand 10	2.6E-06	8.6E-03	5.8
8102792	PCDH18	protocadherin 18	3.1E-06	8.6E-03	3.7
7957737	TMPO	thymopoietin	3.1E-06	8.6E-03	2.1
7965403	**LUM**	**lumican**	**4.1E-06**	**9.4E-03**	**6.1**
8076292	DNAJB7	DnaJ (Hsp40) homolog, subfamily B, member 7	4.9E-06	9.4E-03	-1.8
8125556	HLA-DPA1	major histocompatibility complex, class II, DP alpha 1	5.0E-06	9.4E-03	3.6
8017210	AP1S2	adaptor-related protein complex 1, sigma 2 subunit	5.2E-06	9.4E-03	2.2
7898988	CLIC4	chloride intracellular channel 4	6.3E-06	1.0E-02	2.1
7919815	CTSK	cathepsin K	6.7E-06	1.0E-02	3.5

8178891	HLA-DPA1	major histocompatibility complex, class II, DP alpha 1	7.4E-06	1.1E-02	3.5
8001800	CDH11	cadherin 11, type 2, OB-cadherin (osteoblast)	7.9E-06	1.1E-02	3.1
8094625	KLHL5	kelch-like family member 5	9.3E-06	1.1E-02	2.6
8160238	**PSIP1**	**PC4 and SFRS1 interacting protein 1**	**9.4E-06**	**1.1E-02**	**1.7**
8157890	PBX3	pre-B-cell leukemia homeobox 3	1.0E-05	1.2E-02	1.9
7926609	BMI1	BMI1 polycomb ring finger oncogene	1.1E-05	1.2E-02	2.0
8145470	DPYSL2	dihydropyrimidinase-like 2	1.2E-05	1.2E-02	2.1
8105229	PELO	pelota homolog (Drosophila)	1.4E-05	1.4E-02	2.8
8180100	HLA-DPA1	major histocompatibility complex, class II, DP alpha 1	1.5E-05	1.4E-02	2.8
8140840	STEAP4	STEAP family member 4	1.6E-05	1.4E-02	2.4
8127563	**COL12A1**	**collagen, type XII, alpha 1**	**1.9E-05**	**1.6E-02**	**4.1**
8128007	GJB7	gap junction protein, beta 7, 25kDa	2.0E-05	1.7E-02	-1.6
8091032	FOXL2	forkhead box L2	2.3E-05	1.8E-02	-1.7
8046895	FAM171B	family with sequence similarity 171, member B	2.5E-05	1.9E-02	2.9
8174322	**MORC4**	**MORC family CW-type zinc finger 4**	**2.7E-05**	**2.0E-02**	**2.3**
7981377	ANKRD9	ankyrin repeat domain 9	2.8E-05	2.0E-02	-1.5
7957260	GLIPR1	GLI pathogenesis-related 1	2.8E-05	2.0E-02	2.8
8036324	**ZNF260**	**zinc finger protein 260**	**3.1E-05**	**2.0E-02**	**2.3**
8053882	DUSP2	dual specificity phosphatase 2	3.1E-05	2.0E-02	-1.6
8121319	SOBP	sine oculis binding protein homolog (Drosophila)	3.1E-05	2.0E-02	2.1
8173732	TAF9B	TAF9B RNA polymerase II, TATA box binding protein (TBP)-associated factor, 31kDa	3.2E-05	2.0E-02	2.1
8176263	TAF9B	TAF9B RNA polymerase II, TATA box binding protein	3.2E-05	2.0E-02	2.1

		(TBP)-associated factor, 31kDa			
8143040	SLC35B4	solute carrier family 35, member B4	3.8E-05	2.1E-02	1.7
8143054	AKR1B1	aldo-keto reductase family 1, member B1 (aldose reductase)	3.9E-05	2.1E-02	2.1
8157605	STOM	stomatin	3.9E-05	2.1E-02	1.5
7936322	GPAM	glycerol-3-phosphate acyltransferase, mitochondrial	3.9E-05	2.1E-02	1.6
7959761	FAM101A	family with sequence similarity 101, member A	4.0E-05	2.2E-02	-1.6
8042439	ANTXR1	anthrax toxin receptor 1	4.4E-05	2.3E-02	3.3
7930833	KCNK18	potassium channel, subfamily K, member 18	4.6E-05	2.3E-02	-1.6
8089714	LSAMP	limbic system-associated membrane protein	4.7E-05	2.3E-02	2.1
7958913	OAS2	2'-5'-oligoadenylate synthetase 2, 69/71kDa	5.0E-05	2.3E-02	2.3
7912852	**EIF1AX**	**eukaryotic translation initiation factor 1A, X-linked**	**5.0E-05**	**2.3E-02**	**1.7**
8115234	ANXA6	annexin A6	5.1E-05	2.3E-02	1.7
8169473	PLS3	plastin 3	5.1E-05	2.3E-02	3.1
8138805	CPVL	carboxypeptidase, vitellogenic-like	5.6E-05	2.4E-02	2.0
8135734	CPED1	cadherin-like and PC-esterase domain containing 1	5.6E-05	2.4E-02	2.4

EF: *E. faecalis*

The chemokine *CXCL10* was upregulated 5.8-fold (FDR=0.009) in *E. faecalis*+ CRCs. Importantly, *Enterococcus* spp. have been shown to upregulate *CXCL10* in IL-10$^{-/-}$ mice[440,444], which may imply *E. faecalis*-induced transcription of *CXCL10* in our cohort. *CXCL10* is secreted by cells stimulated with type I and II interferons or lipopolysaccharide (LPS) and plays an important role in the recruitment of T cells to sites of inflammation[445].

In CRC, *CXCL10* exerts antiangiogenic and antiproliferative effects[397], yet is also associated with advanced cancer[397] and may promote invasion through the induction of cell migration[398].

Another interesting observation is the upregulation of various SLRPs–a family of secreted proteoglycans that are involved in regulating matrix assembly and cellular growth[446]. Co-regulation for 12 of the 13 known SLRPs have been predicted on the basis of a common HOX-Runx module in these genes[446], which might explain the significant and concurrent upregulation of lumican, decorin, biglycan (FDR=0.06), and asporin in *E. faecalis*+ CRCs. Of these SLRPs, only lumican was also significantly differentially expressed in group B cancers, at a slightly lower p-value (p=0.003), but with a lower fold change of 4, compared to the fold change of 6.1 in *E. faecalis* + CRCs.

Both decorin and biglycan act as endogenous ligands of TLR4 and TLR2 following tissue damage and their release from the extracellular matrix[391]. TLR4/2 activation stimulates an inflammatory response through NF-κB, with downstream induction of IL-1B and IL-10. Lumican, on the other hand, is not an endogenous TLR4/2 ligand, but presents LPS to the TLR4/2-adaptor molecule CD14, whereby it induces an immune response to LPS and increased bacterial phagocytosis[391].

Certain pathogens express adhesins that bind to glycosaminoglycans which are linked to proteoglycans—this interaction can a) facilitate cell invasion, and/or b) enhance virulence via shedding of proteoglycans that lead to the release of effectors that weaken host defenses[447]. Notably, *E. faecalis* utilizes both these mechanisms, a) to gain entry into macrophages[448] and b) to weaken host defense through the degradation of the proteoglycan decorin, which leads to the inactivation of α-defensins[449].

Lastly, it is interesting to note that *BMI1* polycomb ring finger oncogene was upregulated 2-fold in *E. faecalis*-infected CRCs. BMI1 is an intestinal stem cell marker[450] that is overexpressed in various cancers[451]. The oncogenic potential of BMI1 has been attributed to its potent negative regulation of the tumour suppressor p16INK4, which suppresses senescence in primary cells[451]. High expression of BMI1 is significantly associated with metastasis[452,453] and poor survival[454] in CRC patients.

Importantly, aberrant BMI1 expression has been found in premalignant gastrointestinal lesions[451], which points to its possible role in cancer initiation.

Pathway-level alterations associated with E. faecalis infection

Pathway analyses were conducted using IPA and PARADIGM. These methods are discussed in broader detail in Chapter 7. IPA takes a list of differentially expressed genes (together with their FDRs and FCs) as input, whereas PARADIGM uses the actual gene expression and methylation data as input and therefore infers activity on a per-patient basis. Both methods are very useful in facilitating biological interpretation of large-scale omics data. Regarding the output, IPA produces statistical measures of confidence in the differential overrepresentation (represented as a p-value) and/or change in activity (represented by a z-score) of canonical pathways, upstream regulators and diseases and functions. Meanwhile, PARADIGM provides an activity score for each integrated pathway level (IPL), which consists of genes, complexes and abstract processes, on a per-patient basis.

The IPA results that were significantly altered in *E. faecalis*+ CRCs ($p \leq 0.05$, $|z\text{-score}| \geq 2$) are presented in Tables 3a–3c for canonical pathways, upstream regulators, and diseases and functions, respectively. Meanwhile, differential analysis of PARADIGM IPLs by *E. faecalis* colonisation status did not provide any significantly altered genes, complexes or abstract processes.

In order to identify pathways that are specifically altered in *E. faecalis*+ CRCs as opposed to group B cancers as a whole (described in Chapter 7), the overlap between pathway analyses for these two sets of results was established—overlapping entries are italicized in Tables 3a–3c. Although *E. faecalis* may significantly contribute to these overlapping pathways, the focus in this chapter is on the subset of pathways for which alteration was more evident in *E. faecalis*+ CRCs, as opposed to group B CRCs in general.

IPA results revealed an increase in immune-related canonical pathways in *E. faecalis*+ CRCs (Table 3a).

Table 3a: Ingenuity Canonical Pathways significantly associated with *E. faecalis* colonisation in CRCs ($p \leq 0.05$). Boldface entries were also significant in the comparison between group A vs. group B CRCs in Chapter 7.

Ingenuity Canonical Pathways	-log(p-value)	Downregulated	Upregulated
Antigen Presentation Pathway	2.88	0/33 (0%)	15/33 (45%)
Hepatic Fibrosis / Hepatic Stellate Cell Activation	2.48	22/191 (12%)	44/191 (23%)
Human Embryonic Stem Cell Pluripotency	2.27	23/129 (18%)	29/129 (22%)
Leukocyte Extravasation Signaling	2.06	15/190 (8%)	48/190 (25%)
Agranulocyte Adhesion and Diapedesis	1.98	17/169 (10%)	38/169 (22%)
Allograft Rejection Signaling	1.91	0/36 (0%)	11/36 (31%)
T Helper Cell Differentiation	1.71	4/62 (6%)	13/62 (21%)
Aryl Hydrocarbon Receptor Signaling	**1.68**	**7/135 (5%)**	**41/135 (30%)**
Chondroitin Sulfate Biosynthesis (Late Stages)	1.65	6/43 (14%)	5/43 (12%)
L-dopachrome Biosynthesis	1.60	1/1 (100%)	0/1 (0%)
Glucocorticoid Receptor Signaling	1.58	18/252 (7%)	67/252 (27%)
Antiproliferative Role of TOB in T Cell Signaling	**1.57**	**0/26 (0%)**	**14/26 (54%)**
Role of Pattern Recognition Receptors in Recognition of Bacteria and Viruses	1.56	11/116 (9%)	28/116 (24%)
OX40 Signaling Pathway	1.55	0/46 (0%)	17/46 (37%)
Growth Hormone Signaling	1.53	6/69 (9%)	15/69 (22%)
Colorectal Cancer Metastasis Signaling	1.53	27/228 (12%)	50/228 (22%)
Cdc42 Signaling	1.48	11/121 (9%)	32/121 (26%)
Chondroitin Sulfate Biosynthesis	1.41	8/51 (16%)	6/51 (12%)
Axonal Guidance Signaling	1.39	57/424 (13%)	81/424 (19%)
Calcium-induced T Lymphocyte Apoptosis	1.36	3/53 (6%)	18/53 (34%)
Dermatan Sulfate Biosynthesis	1.36	8/53 (15%)	7/53 (13%)
Role of JAK2 in Hormone-like Cytokine Signaling	1.34	2/32 (6%)	5/32 (16%)
Granulocyte Adhesion and Diapedesis	1.32	12/159 (8%)	35/159 (22%)
Hypusine Biosynthesis	1.30	1/2 (50%)	0/2 (0%)
Cardiolipin Biosynthesis II	1.30	1/2 (50%)	0/2 (0%)

The most significant finding was the activation of the *Antigen Presentation Pathway*, which was largely based on the increased expression of several HLA type II genes—this is in agreement with the findings by Barnett et al. regarding *E. faecalis* and *E. faecium* infection of IL10$^{-/-}$ mice[440]. Together with the overrepresentation of genes related to the *Role of Pattern Recognition Receptors in Recognition of Bacteria and Viruses* pathway, these data provide evidence for an elevated host immune response to foreign antigens in these tumours that could be due to microbial- or cancer-specific antigens. This is supported by the fact that LPS is predicted to be an active upstream regulator in *E. faecalis*-infected CRCs, based on the differential expression of *CXCL10*, proteoglycans and various other genes (Table 3b).

In accordance with the predicted active role of LPS, SOCS1 (a negative regulator of host response to LPS), is predicted to be deactivated in *E. faecalis*+ CRCs (Table 3b). SOCS-1 plays a crucial role in dampening host response to LPS to avoid excessive tissue damage. Furthermore, SOCS-1 deficient mice are highly sensitive to LPS-induced shock and produce increased levels of inflammatory cytokines[455].

LPS is found in the outer membrane of gram-*negative* bacteria, whereas *E. faecalis* is gram-*positive*. Plausible scenarios that could explain the LPS-like response seen here include the presence of lipoteichoic acid (an immunogenic cell wall component of gram-positive bacteria[456]) activation of TLR2/4 by endogenous ligands, such as decorin and biglycan[391] or the presence of increased levels of mucosally associated gram-negative bacteria.

In accordance with an increased immune response to foreign antigens, an inflammatory response is implied by the predicted increase in *Leukocyte Extravasation Signaling*–the process responsible for transporting leukocytes from the blood to tissue at sites of inflammation. Moreover, various cytokines were predicted to be activated in association with *E. faecalis* infection, including TNF, CD40LG, IFN-γ, IL-1β, INF-α2 and IL-6 (Table 3b).

Table 3b: Upstream regulators predicted to be altered in EF CRCs ($p \leq 0.05$, $|z\text{-score}| \geq 2$), sorted by molecule type. Italicised entries were also significant in the comparison between group A vs. group B CRCs in Chapter 7.

Upstream Regulator	Molecule Type	Predicted Activation State	Activation z-score	p-value of overlap
tretinoin	chemical - endogenous mammalian	Activated	3.2	7.4E-05
hyaluronic acid	chemical - endogenous mammalian	Activated	2.3	4.9E-03
D-glucose	*chemical - endogenous mammalian*	*Activated*	*2.5*	*3.1E-02*
SB203580	chemical - kinase inhibitor	Inhibited	-3.2	8.8E-03
PD98059	chemical - kinase inhibitor	Inhibited	-2.4	3.7E-02
lipopolysaccharide	chemical drug	Activated	4.4	1.3E-03
decitabine	chemical drug	Activated	3.9	4.9E-03
bleomycin	chemical drug	Activated	2.1	6.3E-03
phorbol myristate acetate	chemical drug	Activated	3.3	7.3E-03
epigallocatechin-gallate	chemical drug	Inhibited	-3.1	8.1E-03
mifepristone	chemical drug	Inhibited	-2.5	4.4E-02
N-acetyl-L-cysteine	chemical drug	Inhibited	-2.2	4.5E-02
trichostatin A	chemical drug	Activated	2.2	4.7E-02
SB-431542	chemical reagent	Inhibited	-2.6	8.5E-04
metribolone	chemical reagent	Activated	2.9	1.9E-03
hexachlorobenzene	chemical toxicant	Activated	2.0	3.8E-02
TNF	cytokine	Activated	3.7	6.1E-03
CD40LG	cytokine	Activated	2.9	7.6E-03
CSF2	*cytokine*	*Activated*	*2.9*	*1.0E-02*
IFNG	cytokine	Activated	2.9	1.2E-02
IL1B	cytokine	Activated	3.1	2.6E-02
IFNA2	cytokine	Activated	2.6	3.1E-02
IL6	cytokine	Activated	2.0	3.6E-02
IL5	*cytokine*	*Activated*	*2.7*	*3.9E-02*
TGM2	enzyme	Activated	2.4	8.3E-03
PTGER2	*G-protein coupled receptor*	*Activated*	*2.6*	*1.1E-02*
Alpha catenin	group	Inhibited	-3.8	1.7E-07

Ifnar	group	Activated	2.6	1.9E-03
estrogen receptor	group	Inhibited	-2.7	9.5E-03
IFN Beta	group	Activated	2.2	2.8E-02
TGFB1	growth factor	Activated	3.4	1.4E-05
FGF2	growth factor	Activated	2.4	3.7E-03
WISP2	growth factor	Inhibited	-2.0	1.1E-02
AGT	growth factor	Activated	2.8	2.7E-02
miR-124-3p (and other miRNAs w/seed AAGGCAC)	*mature microRNA*	*Inhibited*	*-3.6*	*3.7E-03*
miR-16-5p (and other miRNAs w/seed AGCAGCA)	*mature microRNA*	*Inhibited*	*-2.6*	*1.9E-02*
let-7	*microRNA*	*Inhibited*	*-2.6*	*3.7E-02*
SOCS1	other	Inhibited	-2.6	9.0E-03
Irgm1	*other*	*Inhibited*	*-2.0*	*1.3E-02*
INSIG1	other	Inhibited	-2.4	3.3E-02
MYD88	other	Activated	2.4	3.6E-02
CD24	*other*	*Activated*	*2.0*	*4.4E-02*
NUPR1	*transcription regulator*	*Inhibited*	*-5.0*	*1.9E-05*
TWIST1	transcription regulator	Activated	2.5	4.1E-05
SMAD7	transcription regulator	Inhibited	-2.2	9.2E-05
MYC	*transcription regulator*	*Activated*	*2.1*	*1.4E-04*
SOX11	transcription regulator	Inhibited	-2.1	2.7E-04
FOXM1	*transcription regulator*	*Activated*	*2.6*	*9.0E-04*
TBX2	*transcription regulator*	*Activated*	*2.4*	*2.6E-03*
CCND1	*transcription regulator*	*Activated*	*2.0*	*4.2E-03*
TRIM24	transcription regulator	Inhibited	-2.6	4.6E-03
SPDEF	transcription regulator	Inhibited	-2.4	5.1E-03
IRF3	transcription regulator	Activated	2.0	5.5E-03
SOX1	transcription regulator	Inhibited	-2.4	6.3E-03
XBP1	*transcription regulator*	*Activated*	*3.0*	*6.4E-03*
GMNN	transcription regulator	Inhibited	-2.4	8.9E-03
SOX3	transcription regulator	Inhibited	-2.4	8.9E-03

IRF5	transcription regulator	Activated	2.2	1.2E-02
JUN	transcription regulator	Activated	2.6	1.7E-02
CEBPB	transcription regulator	Activated	2.2	2.0E-02
POU5F1	transcription regulator	Activated	2.5	2.5E-02
SOX2	transcription regulator	Inhibited	-2.2	2.9E-02
CDKN2A	*transcription regulator*	*Inhibited*	*-2.8*	*3.1E-02*
ISL1	transcription regulator	Activated	2.2	3.5E-02
MTPN	transcription regulator	Activated	2.2	3.5E-02
BTNL2	transmembrane receptor	Activated	2.4	6.8E-03

The pathways and genes significantly associated *E. faecalis*+ CRCs are strikingly similar to those seen following infection of IL10$^{-/-}$ mice with *Enterococcus* spp. (*E. faecalis* and *E. faecium*), including a significant increase in IL-6, TNF and IFN-γ, the *Antigen Presentation Pathway* and pathways related to inflammatory disease and immune response[440]. Given the role of the *E. faecalis* virulence factors including lipoteichoic acid, AS and bacteriocin in stimulating production of TNF-β, IFN-γ and TNF-α[441], strains that possess these virulence factors may be particularly relevant in the pathogenesis of *E. faecalis*+ CRCs.

In Chapter 7 several DNA and protein damage response-related processes were shown to be significantly increased in B group B CRCs. Since *E. faecalis* is known to induce DNA damage through the production of ROS, it is tempting to speculate that *E. faecalis* causes these responses in *E. faecalis*+ CRCs. However, these pathways were not exclusively altered in *E. faecalis*+ CRCs. Nevertheless, given that group B CRCs are significantly enriched with *E. faecalis,* and that Barnett et al. reported *Enterococcus*-specific alterations in the *Cell Cycle: G2/M DNA Damage Checkpoint Regulation* pathway in IL10$^{-/-}$ mice[440], it seems plausible that *E. faecalis* contributes to a ROS-induced DNA damage response in *E. faecalis*+ group B cancers.

An increase in the canonical pathway *CRC Metastasis Signaling* (Table 3a) and in the diseases and functions categories: *metastasis*, *cell movement of colon cancer cell lines* and *migration of colon cancer cell lines* (Table 3c) were found in *E. faecalis*+ CRCs. These results may suggest a more aggressive phenotype for *E. faecalis*-infected CRCs, irrespective of the involvement of *E. faecalis* therein (Table 3c).

Table 3c. Diseases and functions associated with *E. faecalis* colonisation ($p \leq 0.05$, $|z\text{-score}| \geq 2$). Italicised entries were also significant in the comparison between group A vs. group B CRCs in Chapter 7.

Categories	Diseases or Functions Annotation	p-Value	Activation z-score
Connective Tissue Disorders, Developmental Disorder, Skeletal and Muscular Disorders	craniofacial abnormality	2.8E-03	-2.4
Inflammatory Disease, Respiratory Disease	pulmonary emphysema	1.5E-02	-2.2
Developmental Disorder, Immunological Disease	hypoplasia of thymus gland	1.8E-02	-2.6
Cellular Growth and Proliferation	*proliferation of cells*	*1.7E-04*	*3.9*
Cellular Development, Skeletal and Muscular System Development and Function, Tissue Development	differentiation of smooth muscle cells	3.9E-04	2.2
Cancer	metastasis	9.5E-04	3.1
Cellular Movement	invasion of tumor cell lines	9.7E-04	2.9
Cellular Movement	invasion of cells	6.9E-03	2.8
Cellular Development	epithelial-mesenchymal transition	7.0E-03	2.6
Cellular Movement	invasion of breast cancer cell lines	7.4E-03	2.1
Cardiovascular System Development and Function	neovascularization	8.6E-03	2.0
Cancer	neoplasia of cells	1.1E-02	2.2
Organismal Development	size of body	1.3E-02	3.9
Cellular Development	*epithelial-mesenchymal transition of tumor cell lines*	*1.5E-02*	*2.3*
Cellular Movement	cell movement of colon cancer cell lines	1.8E-02	2.6
Cellular Movement	migration of colon cancer cell lines	1.9E-02	2.4
Cellular Development, Skeletal and Muscular System Development and Function, Tissue Development	differentiation of muscle cells	2.1E-02	2.2
Cell-To-Cell Signaling and Interaction, Tissue Development	adhesion of epithelial cells	2.3E-02	2.0

Genomic alterations associated with high-level colonisation by Fusobacterium in CRCs

In the comparison between CRCs colonized by high levels of Fusobacterium (FB-H) vs. those with low/no colonisation by Fusobacterium (as defined in Chapter 4), only three genes were significantly differentially expressed. Interestingly all three belonged to the regenerating islet-derived (REG) family of genes. These were *REG1A*, *REG3A* and *REG1P*—all were highly and significantly overexpressed 23-, 15-fold, and 7-fold, respectively in CRCs with high-level colonisation by Fusobacterium, while *REG1B* showed a 12-fold increase, but was not significant after multiple testing correction (Table 4).

Table 4: Genes differentially expressed at an FDR ≤ 0.25 in *Fusobacterium*-high vs. *Fusobacterium*-low/negative tumours.

Gene ID	Gene symbol	Gene name	P value	FDR	FC (FB-H vs. FB-L/N)
8053341	REG3A	regenerating islet-derived 3 alpha	1.1E-07	2.3E-03	15.4
8042986	REG1A	regenerating islet-derived 1 alpha	6.5E-07	7.2E-03	22.8
8053337	REG1P	regenerating islet-derived 1 pseudogene	1.6E-06	1.2E-02	6.7
8097335	HSPA4L	heat shock 70kDa protein 4-like	1.4E-05	7.6E-02	4.1
7957338	SYT1	synaptotagmin I	2.3E-05	1.0E-01	-1.7
7985213	CHRNA 5	cholinergic receptor, nicotinic, alpha 5 (neuronal)	3.6E-05	1.3E-01	3.1
8156935	ZNF189	zinc finger protein 189	5.0E-05	1.6E-01	1.7
8053330	REG1B	regenerating islet-derived 1 beta	6.2E-05	1.7E-01	12.2
7918026	EXTL2	exostosin-like glycosyltransferase 2	8.8E-05	2.1E-01	1.7
8096704	NPNT	nephronectin	1.1E-04	2.4E-01	2.4
7902913	CDC7	cell division cycle 7	1.4E-04	2.5E-01	1.8
8129947	NMBR	neuromedin B receptor	1.4E-04	2.5E-01	-1.5
8138381	AGR2	anterior gradient 2 homolog (Xenopus laevis)	1.5E-04	2.5E-01	2.0

FB: Fusobacterium; H: high-level infection; L: low-level infection; N: no infection

REG proteins are members of the C-type lectin superfamily and have important roles in proliferation and differentiation in a range of cell types. Of the REGs, only *REG4* is constitutively expressed in the colon but several REG proteins are aberrantly expressed in inflammatory pathologies including IBD where REG1A, REG1B and REG3A are all expressed at the intestinal crypt base by metastatic Paneth cells[384]. REG1A and REG1B have also been found to be concomitantly upregulated in CRC[457].

Considering that both REG1A and REG3A are expressed in metaplastic Paneth cells of inflamed IBD crypts[384]; that REG1A is a downstream target of Wnt pathway activation[458]; that *H. pylori* induces REG1A expression through the induction of IL-8[374] and that Fusobacterium spp. are associated with increased IL-8 *in vitro*[196] and *in vivo*[218], suggests that the upregulation of REG genes seen here may be mediated by increased penetrance by pro-inflammatory bacteria such as Fusobacterium of colonic crypts at the tumour interface. This idea is supported by the finding that REG1 is highly expressed at the tumour interface (the zone between the tumour and the adjacent normal mucosa)—more so than in the tumour itself[459] and the finding that Fusobacterium do enter colonic crypts in the normal mucosa of CRC patients[195].

From a pathological perspective, REG1A is associated with poor prognosis in CRC[460], and correlates with recurrence and/or distant metastasis as well as a short median survival in patients with MSI+ tumors[401]. Importantly, the upregulation of REG1A and REG3A seen here appear to be specific to Fusobacterium-high tumours as opposed to MSI+ tumours, since these genes were not significantly differentially expressed in MSI+ vs. MSI- tumours in our cohort (data not shown).

In normal tissue, although only three normal samples were classified as Fusobacterium-high, REG1B ($p=0.03$, FC=2.5) and REG1P ($p=0.002$, FC=2.7) were also associated with Fusobacterium-high, albeit to a lesser degree compared to tumour samples. Taken together, these results reveal a link between high-level colonisation by Fusobacterium spp. and the expression of several of the REG-family genes, which may be important in the pathogenesis of a subset of CRCs. The potential role of Fusobacterium in regulating the expression of these genes warrants further investigation.

Discussion and conclusions

Although *E. faecalis* is a normal constituent of the human microbiome, it is also a common source of infection, a disparity most likely explained by strain-specific virulence factors. Gene expression and pathway analysis of CRCs stratified by *E. faecalis* infection, revealed genes and pathways that had previously been associated with *E. faecalis* colonisation *in vitro* or in mouse models, as well as novel genes (particularly SLRPs and BMI1) and pathways related to CRC progression.

Pathway analysis of genes differentially expressed in association with *E. faecalis* colonisation uncovered pathways related to immune function (in particular antigen presentation and microbial pattern recognition), inflammation and CRC progression, which may indicate an important role in the pathogenesis of a subset of CRCs.

A substantial increase in the expression of REG-family gene members, particularly in REG1A and REG3A, was seen in CRCs infected with high-level Fusobacterium, which may have important repercussions for progression and metastases of these cancers.

Given the ability of Fusobacterium to enter colonic crypts[195] in CRC patients, and the relative increase in REG1 expression at the tumour interface compared to the tumour itself[459], the increased expression of these genes seen here could plausibly be mediated by increased penetrance of colonic crypts at the tumour interface by pro-inflammatory bacteria such as Fusobacterium.

Recently, Kostic et al. identified a pro-inflammatory signature of Fusobacterium colonisation in CRCs, which included *COX-2, IL-1β, IL-6, IL-8, TNF-α,* and *MMP3*. Here, these genes were not significantly differentially expressed in CRCs with high-level Fusobacterium. However, this signature shows significant overlap with the gene- and/or pathway-level profiles of B group B CRCs (*COX-2, IL-8, IL-1β, IL-6* and *TNF*) and with *E. faecalis*+ CRCs (*IL-1β, IL-6, TNF-α*), which may suggest a more complex pattern of bacterial involvement associated with this pro-inflammatory signature that is not limited to Fusobacterium colonisation but rather represents a signature of dysbiosis that may be characterised by increased colonisation of the host mucosa by CRC-associated pathogens.

Chapter 9: Summary, conclusions and future perspectives

Sporadic CRC has been linked to various lifestyle factors and is one of most extensively characterised cancers, both at a molecular and 'omic level; nevertheless, the precise mechanism driving CRC initiation remains unknown. Importantly, the relationship between currently known CRC-associated bacteria, clinicopathological features of CRC and genomic subtypes of CRC has not been concurrently investigated. Given the poor outcome of late-stage CRCs and the poor rates of colonoscopy-based screening, it is imperative to focus on novel diagnostic and preventative strategies—a process that will be greatly aided by a better understanding of the etiological basis of different CRC subtypes. In this regard, and given the apparent multitude of factors, both exogenous and endogenous to the host that may drive CRC, a systems biology approach, where clinical, 'omic and molecular data are considered in parallel for each patient, will likely play a key role in detailing the yet unknown molecular origins of CRCs, or at least in generating new hypotheses. In this study the relationships between CRC-associated bacteria and transcriptomic and methylation profiles of CRC patients was therefore determined with the aim of gaining insight, and driving future research regarding the potential contribution of these bacteria in the aetiopathogenesis of CRC.

Here, a transcriptomic subtype of colorectal cancer, characterised by an increase in CpG island methylation displays an increased frequency of colonisation by *E. faecalis* and by high levels of Fusobacterium was identified. At the pathway-level, this subtype is enriched for pathways related to DNA and protein damage response, infection, inflammation and cellular proliferation; notably, these findings were confirmed in a well-defined publically available CRC gene expression dataset (GSE13294) of colorectal adenocarcinomas (N=155). These findings suggest that specific bacterial colonisation underlie a distinct genomic subtype of colorectal cancer that is characterised by inflammatory-related gene expression changes.

One of the major technical challenges (which commonly arises when using clinical samples) of this study was variable RNA integrity between samples; this issue was

addressed by establishing a thorough quality control and data analysis pipeline whereby low-integrity samples were either excluded from downstream analysis, or included following suitable adjustment of the data to account for known or unknown sources of variation, such as array quality- and batch- effects using ComBat or Surrogate Variable Analysis. This approach will be useful for transcriptomic studies conducted on the Affymetrix Gene 1.0 ST array platform using samples with variable RNA integrity scores.

In addition to characterising the aforementioned genomic subtypes of CRC, we were interested in identifying transcriptome, methylome and pathway-level profiles that were specifically altered in CRCs colonised by each of the bacteria studied here. Regarding DNA methylation, EPEC-colonised CRCs had intriguing alterations in CpG island methylation, and notably, hypermethylation of the *MLH1* promoter was exclusively identified in these CRCs. Meanwhile, specific genes and pathways that had previously been associated with *E. faecalis* infection *in vitro* or in mouse models were confirmed in *E. faecalis*-infected CRCs– specifically upregulation of CXCL10 and predicted upstream regulation by IL-6, TNF and IFN-γ as well as the *Antigen presentation pathway* was found*;* novel genes significantly associated with *E. faecalis* colonisation in CRCs included SLRPs (lumican, decorin, biglycan and asporin) and BMI1. In CRCs infected with high-level Fusobacterium a substantial increase in the expression of REG-family gene members, particularly in REG1A and REG3A, was noted.

Further, the associations of each bacterium studied here, with various clinicopathological features of CRC was determined–observations from previous studies such as the significant enrichment with Fusobacterium in CRC compared to adjacent normal mucosa samples, as well as the relative increase in the level of Fusobacterium in MSI-H CRCs were confirmed; additionally, novel observations were made, including the increased frequency of ETBF and Fusobacterium in late stage cancers; the increased frequency of ETBF and afaC+ *E. coli* in the colon compared to the rectum; and the association between increased levels of colonisation by Fusobacterium in black patients and in younger patients.

Conclusion

This is the first study to quantify CRC-associated pathogens across a single cohort and to link colonisation by specific bacteria to a transcriptomic subtype of colorectal cancer, and, although we cannot yet prove causality, the enrichment with both *E. faecalis* and Fusobacterium, together with the relative increase in MSI+, CIMP+ and inflamed samples in this subtype, suggests that these bacteria are likely an important aspect of colonic tumourigenesis. Moreover, our findings support previous studies suggesting that polymicrobial colonisation of the colonic epithelium and/or tumour may be an important aspect of colonic tumourigenesis (at least in a subset of CRCs) that warrants further investigation.

Our findings underscore the value of quantifying suspected oncogenic bacteria in parallel with transcriptomic and clinicopathological data in elucidating molecular origins and optimal therapeutic strategies by colorectal cancer subtype. Importantly, the subtype-specific pathway level alterations seen in our cohort (N=19) were confirmed in a large publically available cohort (N=155), which suggests broader applicability of our findings.

Taken together, these findings provide additional support for the role of specific bacteria in the aetiopathogenesis of CRC and provide important clues that will hopefully drive future research on the role of host-pathogen interaction in CRC.

Future perspectives

This study (and previous studies) demonstrates that the level of colonisation by Fusobacterium may be relevant to the pathogenesis of CRC; a consistent method of defining high-level vs. low-level Fusobacterium colonisation will be required for reproducibility and comparibility between studies. Future studies should be aimed at validating our findings in a larger cohort, with a higher proportion of MSI+; black; and young patients since high-level colonisation by Fusobacterium (and by EPEC with regards to MSI-status) appear to be particularly relevant to these patients. In addition, because high-level colonisation by Fusobacterium could plausibly facilitate colonisation

by other potential oncopathogens, species-level metagenomic profiling of these CRCs will be of particular interest.

Given the pathway-level evidence for an increased response to DNA and protein damage in group B CRCs, investigating the level of ROS and NOS by CRC subtype, as well as CRC risk factors other than bacteria that have inflammatory or pro-oxidative potential, including viral infection, cigarette smoking, alcohol- and red meat-consumption, may shed further light on the pathogenesis of these CRCs.

Patients developing group B CRCs have seemingly distinct clinical, microbial and genomic characteristics. However, the evidence presented here is not sufficient to prove that these bacteria cause CRC initiation and/or progression, and further research should be conducted to validate and extend the findings presented here. If these bacteria do prove to play a causal role in CRC initiation/progression, individuals should be screened for these bacteria alongside other CRC screening tools such as colonoscopies and treated with appropriate prophylactic agents.

References

1. Jemal A, Bray F, Ferlay J. Global Cancer Statistics. *CA A Cancer J Clin*. 2011;61(2):69-90. doi:10.3322/caac.20107.Available.

2. Huxley RR, Woodward M, Clifton P. The Epidemiologic Evidence and Potential Biological Mechanisms for a Protective Effect of Dietary Fiber on the Risk of Colorectal Cancer. *Curr Nutr Rep*. 2012;2(1):63-70. doi:10.1007/s13668-012-0030-2.

3. *Cancer in South Africa 2007 Full Report*.; 2007. Available at: http://www.nioh.ac.za/assets/files/NCR_Final_Tables_2007.pdf.

4. Mager DL. Bacteria and cancer: cause, coincidence or cure? A review. *J Transl Med*. 2006;4:14. doi:10.1186/1479-5876-4-14.

5. Garcia M, Jemal A, Ward E, et al. *Global Cancer Facts & Figures 2007*. Atlanta, GA; 2007.

6. *Colorectal Cancer Facts & Figures 2008-2010*. Atlanta; 2010.

7. Chan SSM, Luben R, Bergmann MM, et al. Aspirin in the aetiology of Crohn's disease and ulcerative colitis: a European prospective cohort study. *Aliment Pharmacol Ther*. 2011;34(6):649-55. doi:10.1111/j.1365-2036.2011.04784.x.

8. Ollberding NJ, Nomura AMY, Wilkens LR, Henderson BE, Kolonel LN. Racial/ethnic differences in colorectal cancer risk: the multiethnic cohort study. *Int J Cancer*. 2011;129(8):1899-906. doi:10.1002/ijc.25822.

9. Maddocks ODK, Short AJ, Donnenberg MS, Bader S, Harrison DJ. Attaching and effacing Escherichia coli downregulate DNA mismatch repair protein in vitro and are associated with colorectal adenocarcinomas in humans. *PLoS One*. 2009;4(5):e5517. doi:10.1371/journal.pone.0005517.

10. David O, Maddocks K, Scanlon KM, Donnenberg MS. An Escherichia coli Effector Protein Promotes Host Mutation via Depletion of DNA Mismatch Repair Proteins. *MBio*. 2013;4(3). doi:10.1128/mBio.00152-13.Updated.

11. Cuevas-ramos G, Petit CR, Marcq I, Boury M, Oswald E. Escherichia coli induces DNA damage in vivo and triggers genomic instability in mammalian cells. *PNAS*. 2010;107(25). doi:10.1073/pnas.1001261107.

12. Abdulamir AS, Hafidh RR, Bakar FA. Molecular detection , quantification , and isolation of Streptococcus gallolyticus bacteria colonizing colorectal tumors : inflammation-driven potential. *Mol Cancer*. 2010;9(249):1-18.

13. Boleij A, van Gelder MMHJ, Swinkels DW, Tjalsma H. Clinical Importance of Streptococcus gallolyticus infection among colorectal cancer patients: systematic review and meta-analysis. *Clin Infect Dis*. 2011;53(9):870-8. doi:10.1093/cid/cir609.

14. Toprak NU, Yagci A, Gulluoglu BM, Akin ML, Demirkalem P, Celenk T. A possible role of Bacteroides fragilis enterotoxin in the aetiology of colorectal cancer. *Microbiology*. 2006. doi:10.1111/j.1469-0691.2006.01494.x.

15. Wu S, Rhee KJ, Albesiano E, et al. A human colonic commensal promotes colon tumorigenesis via activation of T helper type 17 T cell responses. *Nat Med*. 2009;15(9):1016-22. doi:10.1038/nm.2015.

16. Kostic AD, Gevers D, Pedamallu CS, et al. Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. *Genome Res*. 2012;22(2):292-8. doi:10.1101/gr.126573.111.

17. Wang T, Cai G, Qiu Y, Fei N, Zhang M, Pang X. Structural segregation of gut microbiota between colorectal cancer patients and healthy volunteers. *Life Sci*. 2011:1-10. doi:10.1038/ismej.2011.109.

18. Scanlan PD, Shanahan F, Clune Y, et al. Culture-independent analysis of the gut microbiota in colorectal cancer and polyposis. *Environ Microbiol*. 2008;10(3):789-98. doi:10.1111/j.1462-2920.2007.01503.x.

19. Garrett WS, Gordon JI, Glimcher LH. Homeostasis and inflammation in the intestine. *Cell*. 2010;140(6):859-70. doi:10.1016/j.cell.2010.01.023.

20. Parkin DM. The global health burden of infection-associated cancers in the year 2002. *Int J Cancer*. 2006;118(12):3030-44. doi:10.1002/ijc.21731.

21. Ott JJ, Ullrich A, Mascarenhas M, Stevens G. Global cancer incidence and mortality caused by behavior and infection. *J Public Health (Oxf)*. 2011;33(2):223-33. doi:10.1093/pubmed/fdq076.

22. De Martel C, Ferlay J, Franceschi S, et al. Global burden of cancers attributable to infections in 2008: a review and synthetic analysis. *Lancet Oncol*. 2012;13(6):607-15. doi:10.1016/S1470-2045(12)70137-7.

23. O'Keefe SJD, Ou J, Aufreiter S, et al. Products of the Colonic Microbiota Mediate the Effects of Diet on Colon Cancer Risk 1 , 2. *J Nutr*. 2009;139:2044-2048. doi:10.3945/jn.109.104380.

24. Anderson WF, Umar A, Brawley OW. Colorectal carcinoma in black and white race. *Cancer Metastasis Rev*. 2003;22(1):67-82.

25. Bremner C., Ackerman V. Polyps and carcinoma of the large bowel in the South African Bantu. *Cancer*. 1970;26:991-999.

26. Segal I, Cooke S, Hamilton D, Ou Tim L. Polyps and colorectal cancer in South African Blacks. *Gut*. 1981;22(8):653-7.

27. Ou J, Carbonero F, Zoetendal EG, et al. Diet , microbiota , and microbial metabolites in colon cancer risk in rural Africans and African Americans. *Am J Nutr*. 2013;98:111-120. doi:10.3945/ajcn.112.056689.

28. O'Keefe SJD, Chung D, Mahmoud N, et al. Why do African Americans get more colon cancer than Native Africans? *J Nutr*. 2007;137(1 Suppl):175S-182S.

29. Hamer HM, Jonkers D, Venema K, Vanhoutvin S, Troost FJ, Brummer R-J. Review article: the role of butyrate on colonic function. *Aliment Pharmacol Ther*. 2008;27(2):104-19. doi:10.1111/j.1365-2036.2007.03562.x.

30. Moore WE, Moore LH. Intestinal floras of populations that have a high risk of colon cancer . *Appl Environ Microbiol*. 1995;61(9).

31. Cronjé L, Paterson A, Becker PJ. Colorectal cancer in South Africa: a heritable cause suspected in many young black patients. *S Afr Med J*. 2009;99(2):103-6.

32. Degiannis E, Sliwa K, Levy R, Hale M, Saadia R. Clinicopathological trends in colorectal carcinoma in a Black South African population. *Trop Gastroenterol*. 1995;16(4):55-61.

33. Cronjé L, Becker PJ, Paterson AC, Ramsay M. Hereditary non-polyposis colorectal cancer is predicted to contribute towards colorectal cancer in young South African blacks. *S Afr J Sci*. 2009;105(1-2). doi:10.1590/S0038-23532009000100023.

34. Sangeeta A, Bhupinderjit A, Bhutani M, et al. Colorectal cancer in African Americans. *Am J Gastroenterol*. 2005;3:515-523.

35. Ibrahim NK, Abdul-Karim FW. Colorectal Adenocarcinoma in Young Lebanese Adults. *Cancer*. 1986;58:816-820.

36. Liang JT, Huang KC, Cheng AL, Jeng YM, Wu MS, Wang SM. Clinicopathological and molecular biological features of colorectal cancer in patients less than 40 years of age. *Br J Surg*. 2003;90(2):205-14. doi:10.1002/bjs.4015.

37. Dijxhoorn D, Boutall A, Mulder C, et al. Colorectal cancer in patients form Uganda: A histopathological study. *East Cent African J Surg*. 2014;19(April):112-119.

38. Irabor D, Adedeji O. Colorectal cancer in Nigeria: 40 years on. A review. *Eur J Cancer Care (Engl)*. 2009;18(2):110-5. doi:10.1111/j.1365-2354.2008.00982.x.

39. S. A. The frequency of large bowel cancer as seen in Addis Ababa University, Pathology Department. *Ethiop Med J*. 2000;38(4):277-282.

40. De Silva M V., Fernando MS, Fernando D. Comparison of some clinical and histological features of colorectal carcinoma occurring in patients below and above 40 years. *Ceylon Med J*. 2000;45(4):166-168.

41. Guraya SY, Eltinay OE. Higher prevalence in young population and rigthward shift of colorectal carcinoma. *Saudi Med J*. 2006;966(May):1391-1393.

42. Laskar RS, Talukdar FR, Mondal R, Kannan R, Ghosh SK. High frequency of young age rectal cancer in a tertiary care centre of southern Assam, North East India. *Indian J Med Res*. 2014;139(2):314-8.

43. Abou-Zeid AA, Khafagy W, Marzouk DM, Alaa A, Mostafa I, Ela MA. Colorectal cancer in Egypt. *Dis Colon Rectum*. 2002;45(9):1255-60. doi:10.1097/01.DCR.0000027122.04363.74.

44. Chan TL, Yuen ST, Chung LP, et al. Frequent microsatellite instability and mismatch repair gene mutations in young Chinese patients with colorectal cancer. *J Natl Cancer Inst*. 1999;91(14):1221-6.

45. Liu B, Farrington SM, Petersen GM, et al. Genetic instability occurs in the majority of young patients with colorectal cancer. *Nat Med*. 1995;1(4):348-352.

46. Tortora, Gerard J. and BHD. *Principles of anatomy and physiology*. John Wiley & Sons; 2008.

47. Staff B co. Blausen gallery 2014. doi:DOI:10.15347/wjm/2014.010.

48. LaPointe LC, Dunne R, Brown GS, et al. Map of differential transcript expression in the normal human large intestine. *Physiol Genomics*. 2008;33(1):50-64. doi:10.1152/physiolgenomics.00185.2006.

49. Yamauchi M, Lochhead P, Morikawa T, et al. Colorectal cancer: a tale of two sides or a continuum? *Gut*. 2012;61(6):794-7. doi:10.1136/gutjnl-2012-302014.

50. Yamauchi M, Morikawa T, Kuchiba A, et al. Assessment of colorectal cancer molecular features along bowel subsites challenges the conception of distinct dichotomy of proximal versus distal colorectum. *Gut*. 2012;61(6):847-54. doi:10.1136/gutjnl-2011-300865.

51. Szmulowicz UM, Hull TL. The ASCRS Textbook of Colon and Rectal Surgery. Beck DE, Roberts PL, Saclarides TJ, Senagore AJ, Stamos MJ, Wexner SD, eds. 2011. doi:10.1007/978-1-4419-1584-9.

52. LeBlanc JG, Milani C, de Giori GS, Sesma F, van Sinderen D, Ventura M. Bacteria as vitamin suppliers to their host: a gut microbiota perspective. *Curr Opin Biotechnol*. 2013;24(2):160-8. doi:10.1016/j.copbio.2012.08.005.

53. Hooper L V, Littman DR, Macpherson AJ. Interactions between the microbiota and the immune system. *Science*. 2012;336(6086):1268-73. doi:10.1126/science.1223490.

54. Goto Y, Kiyono H. Epithelial barrier : an interface for the cross-communication between gut flora and immune system. *Immunol Rev*. 2012;245:147-163.

55. Johansson ME V, Larsson JMH, Hansson GC. The two mucus layers of colon are organized by the MUC2 mucin , whereas the outer layer is a legislator of host – microbial interactions. *PNAS*. 2011;108:4659-4665. doi:10.1073/pnas.1006451107.

56. Carvalho FA, Aitken JD, Vijay-Kumar M, Gewirtz AT. Toll-like receptor-gut microbiota interactions: perturb at your own risk! *Annu Rev Physiol*. 2012;74:177-98. doi:10.1146/annurev-physiol-020911-153330.

57. Johansson ME V, Phillipson M, Petersson J, Velcich A, Holm L, Hansson GC. The inner of the two Muc2 mucin-dependent mucus layers in colon is devoid of bacteria. *Proc Natl Acad Sci U S A*. 2008;105(39):15064-9. doi:10.1073/pnas.0803124105.

58. Van der Sluis M, De Koning B a E, De Bruijn ACJM, et al. Muc2-deficient mice spontaneously develop colitis, indicating that MUC2 is critical for colonic protection. *Gastroenterology*. 2006;131(1):117-29. doi:10.1053/j.gastro.2006.04.020.

59. Velcich A, Yang W, Heyer J, et al. Colorectal cancer in mice genetically deficient in the mucin Muc2. *Science*. 2002;295(5560):1726-9. doi:10.1126/science.1069094.

60. Rescigno M. The pathogenic role of intestinal flora in IBD and colon cancer. *Curr Drug Targets*. 2008;9(5):395-403.

61. Groschwitz KR, Hogan SP. Intestinal barrier function: molecular regulation and disease pathogenesis. *J Allergy Clin Immunol*. 2009;124(1):3-20; quiz 21-2. doi:10.1016/j.jaci.2009.05.038.

62. Barker N, Bartfeld S, Clevers H. Tissue-resident adult stem cell populations of rapidly self-renewing organs. *Cell Stem Cell*. 2010;7(6):656-70. doi:10.1016/j.stem.2010.11.016.

63. Culpepper BST, Mai V. Evidence for Contributions of Gut Microbiota to Colorectal Carcinogenesis. *Curr Nutr Rep*. 2012;2(1):10-18. doi:10.1007/s13668-012-0032-0.

64. Carulli AJ, Samuelson LC, Schnell S. Unraveling intestinal stem cell behavior with models of crypt dynamics. *Integr Biol (Camb)*. 2014;6(3):243-57. doi:10.1039/c3ib40163d.

65. Corr SC, Gahan CCGM, Hill C. M-cells: origin, morphology and role in mucosal immunity and microbial pathogenesis. *FEMS Immunol Med Microbiol*. 2008;52(1):2-12. doi:10.1111/j.1574-695X.2007.00359.x.

66. Murgas Torrazza R, Neu J. The developing intestinal microbiome and its relationship to health and disease in the neonate. *J Perinatol*. 2011;31 Suppl 1(S1):S29-34. doi:10.1038/jp.2010.172.

67. Duerkop BA, Vaishnava S, Hooper L V. Immune responses to the microbiota at the intestinal mucosal surface. *Immunity*. 2009;31(3):368-76. doi:10.1016/j.immuni.2009.08.009.

68. Muniz LR, Knosp C, Yeretssian G. Intestinal antimicrobial peptides during homeostasis, infection, and disease. *Front Immunol*. 2012;3(October):310. doi:10.3389/fimmu.2012.00310.

69. Kim YS, Ho SB. Intestinal goblet cells and mucins in health and disease: recent insights and progress. *Curr Gastroenterol Rep*. 2010;12(5):319-30. doi:10.1007/s11894-010-0131-2.

70. Byrd JC, Bresalier RS. Mucins and mucin binding proteins in colorectal cancer. *Cancer*. 2004:77-99.

71. Raja SB, Murali MR, Devaraj H, Devaraj SN. Differential expression of gastric MUC5AC in colonic epithelial cells: TFF3-wired IL1 β/Akt crosstalk-induced mucosal immune response against Shigella dysenteriae infection. *J Cell Sci*. 2012;125(Pt 3):703-13. doi:10.1242/jcs.092148.

72. Mall AS. Analysis of mucins: role in laboratory diagnosis. *J Clin Pathol*. 2008;61(9):1018-24. doi:10.1136/jcp.2008.058057.

73. Ashida H, Ogawa M, Kim M, Mimuro H, Sasakawa C. Bacteria and host interactions in the gut epithelial barrier. *Nat Chem Biol*. 2012;8(1):36-45. doi:10.1038/nchembio.741.

74. Joo M, Shahsafaei A, Odze RD. Paneth cell differentiation in colonic epithelial neoplasms: evidence for the role of the Apc/beta-catenin/Tcf pathway. *Hum Pathol*. 2009;40(6):872-80. doi:10.1016/j.humpath.2008.12.003.

75. Bevins CL, Salzman NH. Paneth cells , antimicrobial peptides and maintenance of intestinal homeostasis. *Nat Publ Gr*. 2011;9(5):356-368. doi:10.1038/nrmicro2546.

76. Cobo E, Chadee K. Antimicrobial Human β-Defensins in the Colon and Their Role in Infectious and Non-Infectious Diseases. *Pathogens*. 2013;2(1):177-192. doi:10.3390/pathogens2010177.

77. Jensen M. Different age and the large bowel relationship for cancer of subsites of. 1984:825-829.

78. Bufill JA. Colorectal cancer: evidence for distinct genetic categories based on proximal or distal tumor location. *Ann Intern Med*. 1990;113(10):779-788.

79. Iacopetta B. Are there two sides to colorectal cancer? *Int J Cancer*. 2002;101(5):403-8. doi:10.1002/ijc.10635.

80. Azzoni C, Bottarelli L, Campanini N, et al. Distinct molecular patterns based on proximal and distal sporadic colorectal cancer: arguments for different mechanisms in the tumorigenesis. *Int J Colorectal Dis*. 2007;22(2):115-26. doi:10.1007/s00384-006-0093-x.

81. Gervaz P, Bucher P, Morel P. Two colons-two cancers: paradigm shift and clinical implications. *J Surg Oncol*. 2004;88(4):261-6. doi:10.1002/jso.20156.

82. Limsui D, Vierkant RA, Tillmans LS, et al. Cigarette smoking and colorectal cancer risk by molecularly defined subtypes. *J Natl Cancer Inst*. 2010;102(14):1012-22. doi:10.1093/jnci/djq201.

83. Larsson SC, Rafter J, Holmberg L, Bergkvist L, Wolk A. Red meat consumption and risk of cancers of the proximal colon, distal colon and rectum: the Swedish Mammography Cohort. *Int J Cancer*. 2005;113(5):829-34. doi:10.1002/ijc.20658.

84. Benedix F, Schmidt U, Mroczkowski P, Gastinger I, Lippert H, Kube R. Colon carcinoma--classification into right and left sided cancer or according to colonic subsite?--Analysis of 29,568 patients. *Eur J Surg Oncol*. 2011;37(2):134-9. doi:10.1016/j.ejso.2010.12.004.

85. Missiaglia E, Jacobs B, D'Ario G, et al. Distal and proximal colon cancers differ in terms of molecular, pathological and clinical features. *Ann Oncol*. 2014;(mdu275).

86. Li LS, Kim N-G, Kim SH, et al. Chromosomal imbalances in the colorectal carcinomas with microsatellite instability. *Am J Pathol*. 2003;163(4):1429-36. doi:10.1016/S0002-9440(10)63500-6.

87. Grady WM, Carethers JM. Genomic and epigenetic instability in colorectal cancer pathogenesis. *Gastroenterology*. 2008;135(4):1079-99. doi:10.1053/j.gastro.2008.07.076.

88. Rodriguez J, Frigola J, Vendrell E, et al. Chromosomal instability correlates with genome-wide DNA demethylation in human primary colorectal cancers. *Cancer Res*. 2006;66(17):8462-9468. doi:10.1158/0008-5472.CAN-06-0293.

89. Dodge JE, Okano M, Dick F, et al. Inactivation of Dnmt3b in mouse embryonic fibroblasts results in DNA hypomethylation, chromosomal instability, and spontaneous immortalization. *J Biol Chem*. 2005;280(18):17986-91. doi:10.1074/jbc.M413246200.

90. Bond CE, Nancarrow DJ, Wockner LF, et al. Microsatellite stable colorectal cancers stratified by the BRAF V600E mutation show distinct patterns of chromosomal instability. *PLoS One*. 2014;9(3):e91739. doi:10.1371/journal.pone.0091739.

91. Markowitz SD, Bertagnolli MM. Molecular Basis of Colorectal Cancer. *N Engl J Med*. 2009;361(25):2449-2460.

92. Ward R, Meagher A, Tomlinson I, et al. Microsatellite instability and the clinicopathological features of sporadic colorectal cancer. *Gut*. 2001;48(6):821-9.

93. Lothe RA, Peltomäki P, Meling GI, et al. Genomic Instability in Colorectal Cancer : Relationship to Clinicopathological Variables and Family History. *Cancer Res*. 1993:5849-5852.

94. Merok MA, Ahlquist T, Røyrvik EC, et al. Microsatellite instability has a positive prognostic impact on stage II colorectal cancer after complete resection: results from a large, consecutive Norwegian series. *Ann Oncol*. 2013;24(5):1274-82. doi:10.1093/annonc/mds614.

95. Yoon YS, Yu CS, Kim TW, et al. Mismatch repair status in sporadic colorectal cancer: immunohistochemistry and microsatellite instability analyses. *J Gastroenterol Hepatol*. 2011;26(12):1733-9. doi:10.1111/j.1440-1746.2011.06784.x.

96. Jass JR, Do KA, Simms LA, et al. Morphology of sporadic colorectal cancer with DNA replication errors. *Gut*. 1998;42(5):673-9.

97. Laiho P, Launonen V, Lahermo P, et al. Low-level microsatellite instability in most colorectal carcinomas. *Cancer Res*. 2002;62(4):1166-70.

98. Aaltonen LA, Salovaara R, Kristo P, et al. Incidence of hereditary nonpolyposis colorectal cancer and the feasibilty of molecular screening for the disease. *N Engl J Med*. 1998;338(21):1481-1487.

99. Salovaara BR, Loukola A, Kristo P, et al. Population-Based Molecular Detection of Hereditary Nonpolyposis Colorectal Cancer. *J Clin Oncol*. 2000;18(11):2193-2200.

100. Chung DC, Rustgi AK. The Hereditary Nonpolyposis Colorectal Cancer Syndrome: Genetics and Clinical Implications. *Ann Intern Med*. 2003;138:560-570.

101. Geary J, Sasieni P, Houlston R, et al. Gene-related cancer spectrum in families with hereditary non-polyposis colorectal cancer (HNPCC). *Fam Cancer*. 2008;7(2):163-72. doi:10.1007/s10689-007-9164-6.

102. Poynter JN, Siegmund KD, Weisenberger DJ, et al. Molecular characterization of MSI-H colorectal cancer by MLHI promoter methylation, immunohistochemistry, and mismatch repair germline mutation screening. *Cancer Epidemiol Biomarkers Prev*. 2008;17(11):3208-15. doi:10.1158/1055-9965.EPI-08-0512.

103. Suzuki H, Watkins DN, Jair K-W, et al. Epigenetic inactivation of SFRP genes allows constitutive WNT signaling in colorectal cancer. *Nat Genet*. 2004;36(4):417-22. doi:10.1038/ng1330.

104. TCGA. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*. 2012;487(7407):330-7. doi:10.1038/nature11252.

105. Wend P, Holland JD, Ziebold U, Birchmeier W. Wnt signaling in stem and cancer stem cells. *Semin Cell Dev Biol*. 2010;21(8):855-63. doi:10.1016/j.semcdb.2010.09.004.

106. Miyaki M, Konishi M, Kikuchi-yanoshita R, et al. Characteristics of Somatic Mutation of the Adenomatous Polyposis Coli Gene in Colorectal Tumors. *Cancer Res*. 1994:3011-3020.

107. Samowitz WS, Slattery ML, Sweeney C, Herrick J, Wolff RK, Albertsen H. APC mutations and other genetic and epigenetic changes in colon cancer. *Mol Cancer Res*. 2007;5(2):165-70. doi:10.1158/1541-7786.MCR-06-0398.

108. Henry T. Lynch, M.D., and Albert de la Chapelle, M.D. P. Hereditary Colorectal Cancer. *N Engl J Med*. 2003;348:919-932.

109. Grivennikov SI. Inflammation and colorectal cancer: colitis-associated neoplasia. *Semin Immunopathol*. 2013;35(2):229-44. doi:10.1007/s00281-012-0352-6.

110. Fu X, Li L, Peng Y. Wnt signalling pathway in the serrated neoplastic pathway of the colorectum: possible roles and epigenetic regulatory mechanisms. *J Clin Pathol*. 2012;65(8):675-9. doi:10.1136/jclinpath-2011-200602.

111. Oshima M, Oshima H, Kitagawa K, Kobayashi M, Itakura C, Taketo M. Loss of Apc heterozygosity and abnormal tissue building in nascent intestinal polyps in mice carrying a truncated Apc gene. *Proc Natl Acad Sci U S A*. 1995;92(10):4482-6.

112. Bienz M, Clevers H. Linking Colorectal Cancer to Wnt Signaling. *Cell*. 2000;103:311-320.

113. Barker N, Ridgway RA, van Es JH, et al. Crypt stem cells as the cells-of-origin of intestinal cancer. *Nature*. 2009;457(7229):608-11. doi:10.1038/nature07602.

114. Hartnett L, Egan LJ. Inflammation, DNA methylation and colitis-associated cancer. *Carcinogenesis*. 2012;33(4):723-31. doi:10.1093/carcin/bgs006.

115. Balkwill F, Mantovani A. Inflammation and cancer: back to Virchow? *Lancet*. 2001;357(9255):539-45. doi:10.1016/S0140-6736(00)04046-0.

116. Trinchieri G. Cancer and inflammation: an old intuition with rapidly evolving new concepts. *Annu Rev Immunol*. 2012;30:677-706. doi:10.1146/annurev-immunol-020711-075008.

117. Dvorak H. Tumors: wounds that do not heal. Similarities between tumor stroma generation and wound healing. *N Engl J Med*. 1986;315:1650-59.

118. Colotta F, Allavena P, Sica A, Garlanda C, Mantovani A. Cancer-related inflammation , the seventh hallmark of cancer : links to genetic instability. *Carcinogenesis*. 2009;30(7):1073-1081. doi:10.1093/carcin/bgp127.

119. Fantini MC, Wirtz S, Nikolaev A, Lehr HA. IL-6 Signaling Promotes Tumor Growth in Colorectal Cancer. *Cell Cycle*. 2005;(February):217-220.

120. Grivennikov S, Karin E, Terzic J, et al. IL-6 and Stat3 are required for survival of intestinal epithelial cells and development of colitis-associated cancer. *Cancer Cell*. 2009;15(2):103-13. doi:10.1016/j.ccr.2009.01.001.

121. Wang S, Liu Z, Wang L, Zhang X. NF-kappaB signaling pathway, inflammation and colorectal cancer. *Cell Mol Immunol*. 2009;6(5):327-34. doi:10.1038/cmi.2009.43.

122. Neurath MF, Pettersson S, Karl-Hermann Meyer Zum Büschenfelde and WS. Local administration of antisense phosphorothioate oligonucleotides to the p65 subunit of NF-kB abrogates established experimental colitis in mice. *Nat Med*. 1996;2(9):998-1004.

123. Greten FR, Eckmann L, Greten TF, et al. IKKbeta links inflammation and tumorigenesis in a mouse model of colitis-associated cancer. *Cell*. 2004;118(3):285-96. doi:10.1016/j.cell.2004.07.013.

124. Hanahan D, Weinberg R a. Hallmarks of cancer: the next generation. *Cell*. 2011;144(5):646-74. doi:10.1016/j.cell.2011.02.013.

125. Xavier RJ, Podolsky DK. Unravelling the pathogenesis of inflammatory bowel disease. *Nature*. 2007;448(7152):427-34. doi:10.1038/nature06005.

126. Kotlowski R, Bernstein CN, Sepehri S, Krause DO. High prevalence of Escherichia coli belonging the B2+D phylogenetic group in inflammtory bowel disease. *Gut*. 2007;56:669-675. doi:10.1136/gut.2006.099796.

127. Martin H. Enhanced adherence and invasion in Crohn's disease and colon cancer. *Gastroenterology*. 2004;127(1):80-93. doi:10.1053/j.gastro.2004.03.054.

128. Honda K, Littman DR. The Microbiome in Infectious Disease and Inflammation. *Annu Rev Immunol*. 2012;30:759-95. doi:10.1146/annurev-immunol-020711-074937.

129. Hong PY, Croix JA, Greenberg E, Gaskins HR, Mackie RI. Pyrosequencing-based analysis of the mucosal microbiota in healthy individuals reveals ubiquitous bacterial groups and micro-heterogeneity. *PLoS One*. 2011;6(9):e25042. doi:10.1371/journal.pone.0025042.

130. Rothwell PM, Wilson M, Elwin C-E, et al. Long-term effect of aspirin on colorectal cancer incidence and mortality: 20-year follow-up of five randomised trials. *Lancet*. 2010:1741-1750. doi:10.1016/S0140-6736(10)61543-7.

131. Wang D, DuBois RN. The role of anti-inflammatory drugs in colorectal cancer. *Annu Rev Med*. 2013;64:131-44. doi:10.1146/annurev-med-112211-154330.

132. Burn J, Gerdes A-M, Macrae F, et al. Long-term effect of aspirin on cancer risk in carriers of hereditary colorectal cancer: an analysis from the CAPP2 randomised controlled trial. *Lancet*. 2011;378(9809):2081-7. doi:10.1016/S0140-6736(11)61049-0.

133. McIlhatton MA, Tyler J, Kerepesi LA, et al. Aspirin and low-dose nitric oxide-donating aspirin increase life span in a Lynch syndrome mouse model. *Cancer Prev Res*. 2011;4(5):684-93. doi:10.1158/1940-6207.CAPR-10-0319.

134. Flossmann E, Rothwell PM. Effect of aspirin on long-term risk of colorectal cancer: consistent evidence from randomised and observational studies. *Lancet*. 2007;369(9573):1603-13. doi:10.1016/S0140-6736(07)60747-8.

135. Ferrández A, Piazuelo E, Castells A. Aspirin and the prevention of colorectal cancer. *Best Pract Res Clin Gastroenterol*. 2012;26(2):185-95. doi:10.1016/j.bpg.2012.01.009.

136. Rüschoff J, Wallinger S, Dietmaier W, et al. Aspirin suppresses the mutator phenotype associated with hereditary nonpolyposis colorectal cancer by genetic selection. *Proc Natl Acad Sci U S A*. 1998;95(19):11301-6.

137. Jackson AL, Chen R, Loeb LA. Induction of microsatellite instability by oxidative DNA damage. *Proc Natl Acad Sci U S A*. 1998;95(21):12468-73.

138. Chang CL, Marra G, Chauhan DP, et al. Oxidative stress inactivates the human DNA mismatch repair system. *Am J Physiol Cell Physiol*. 2002;283(1):C148-54. doi:10.1152/ajpcell.00422.2001.

139. Ishitsuka T. Microsatellite instability in inflamed and neoplastic epithelium in ulcerative colitis. *J Clin Pathol*. 2001;54(7):526-532. doi:10.1136/jcp.54.7.526.

140. Kloor M, Huth C, Voigt AY, et al. Prevalence of mismatch repair-deficient crypt foci in Lynch syndrome: a pathological study. *Lancet Oncol*. 2012;13(6):598-606. doi:10.1016/S1470-2045(12)70109-2.

141. Ten Kate GL, Kleibeuker JH, Nagengast FM, et al. Is surveillance of the small bowel indicated for Lynch syndrome families? *Gut*. 2007;56(9):1198-201. doi:10.1136/gut.2006.118299.

142. Eaden J. Review article: the data supporting a role for aminosalicylates in the chemoprevention of colorectal cancer in patients with inflammatory bowel disease. *Aliment Pharmacol Ther*. 2003:15-21.

143. Ananthakrishnan AN, Higuchi LM, Huang ES, Khalili H, Richter JM, Charles S. Fuchs and ATC. Aspirin, Nonsteroidal Anti-inflammatory Drug Use, and Risk for Crohn Disease Ulcerative Colitis: A Cohort Study. *Ann Intern Med*. 2012;156(5):350-359. doi:10.1059/0003-4819-156-5-201203060-00007.Aspirin.

144. Dubeau M, Iacucci M, Beck PL, et al. Drug-induced inflammatory bowel disease and IBD-like conditions. *Inflamm Bowel Dis*. 2012.

145. Geramizadeh B, Taghavi A, Banan B. Clinical, endoscopic and pathologic spectrum of non-steroidal anti-inflammatory drug-induced colitis. *Indian J Gastroenterol*. 2009;28(4):150-3. doi:10.1007/s12664-009-0053-9.

146. Samadder NJ, Mukherjee B, Huang S-C, et al. Risk of colorectal cancer in self-reported inflammatory bowel disease and modification of risk by statin and NSAID use. *Cancer*. 2011;117(8):1640-8. doi:10.1002/cncr.25731.

147. Christensen BC, Houseman EA, Marsit CJ, et al. Aging and environmental exposures alter tissue-specific DNA methylation dependent upon CpG island context. *PLoS Genet*. 2009;5(8):e1000602. doi:10.1371/journal.pgen.1000602.

148. Cortessis VK, Thomas DC, Levine AJ, et al. Environmental epigenetics: prospects for studying epigenetic mediation of exposure-response relationships. *Hum Genet*. 2012;131(10):1565-89. doi:10.1007/s00439-012-1189-8.

149. Madrigano J, Baccarelli A, Mittleman MA, et al. Longitudinal changes in gene-specific DNA methylation. *Epigenetics*. 2012;7(1):63-70.

150. Niwa T, Tsukamoto T, Toyoda T, et al. Inflammatory processes triggered by Helicobacter pylori infection cause aberrant DNA methylation in gastric epithelial cells. *Cancer Res*. 2010;70(4):1430-40. doi:10.1158/0008-5472.CAN-09-2755.

151. Matsusaka K, Kaneda A, Nagae G, et al. Classification of Epstein-Barr virus-positive gastric cancers by definition of DNA methylation epigenotypes. *Cancer Res*. 2011;71(23):7187-97. doi:10.1158/0008-5472.CAN-11-1349.

152. Ogino S, Kawasaki T, Nosho K, et al. LINE-1 hypomethylation is inversely associated with microsatellite instability and CpG island methylator phenotype in colorectal cancer. *Int J Cancer*. 2008;122(12):2767-73. doi:10.1002/ijc.23470.

153. Matsuzaki K, Deng G, Tanaka H, Kakar S, Miura S, Kim YS. The relationship between global methylation level, loss of heterozygosity, and microsatellite instability in sporadic colorectal cancer. *Clin Cancer Res*. 2005;11(24 Pt 1):8564-9. doi:10.1158/1078-0432.CCR-05-0859.

154. Deng G, Nguyen A, Tanaka H, et al. Regional hypermethylation and global hypomethylation are associated with altered chromatin conformation and histone acetylation in colorectal cancer. *Int J Cancer*. 2006;118(12):2999-3005. doi:10.1002/ijc.21740.

155. Nosho K, Irahara N, Shima K, et al. Comprehensive biostatistical analysis of CpG island methylator phenotype in colorectal cancer using a large population-based sample. *PLoS One*. 2008;3(11):e3698. doi:10.1371/journal.pone.0003698.

156. Van Engeland M, Derks S, Smits KM, Meijer G a, Herman JG. Colorectal Cancer Epigenetics: Complex Simplicity. *J Clin Oncol*. 2011:1-10. doi:10.1200/JCO.2010.28.2319.

157. Feinberg AP, Ohlsson R, Henikoff S. The epigenetic progenitor origin of human cancer. *Nat Rev Genet*. 2006;7(1):21-33. doi:10.1038/nrg1748.

158. De Martel C, Franceschi S. Infections and cancer: established associations and new hypotheses. *Crit Rev Oncol Hematol*. 2009;70(3):183-94. doi:10.1016/j.critrevonc.2008.07.021.

159. Cummins J, Tangney M. Bacteria and tumours: causative agents or opportunistic inhabitants? *Infect Agent Cancer*. 2013;8(1):11. doi:10.1186/1750-9378-8-11.

160. Wei MQ, Mengesha A, Good D, Anné J. Bacterial targeted tumour therapy-dawn of a new era. *Cancer Lett*. 2008;259(1):16-27. doi:10.1016/j.canlet.2007.10.034.

161. Hajishengallis G, Darveau RP, Curtis M a. The keystone-pathogen hypothesis. *Nat Rev Microbiol*. 2012;10(10):717-25. doi:10.1038/nrmicro2873.

162. Rook GAW, Dalgleish A. Infection, immunoregulation, and cancer. *Immunol Rev*. 2011;240:141-159.

163. Ma J-L, Zhang L, Brown LM, et al. Fifteen-year effects of Helicobacter pylori, garlic, and vitamin treatments on gastric cancer incidence and mortality. *J Natl Cancer Inst*. 2012;104(6):488-92. doi:10.1093/jnci/djs003.

164. Schiffman M, Castle PE, Jeronimo J, Rodriguez AC, Wacholder S. Human papillomavirus and cervical cancer. *Lancet*. 2007;370:890-907.

165. Stein M, Ruggiero P, Rappuoli R, Bagnoli F. Helicobacter pylori CagA: From Pathogenic Mechanisms to Its Use as an Anti-Cancer Vaccine. *Front Immunol*. 2013;4:328. doi:10.3389/fimmu.2013.00328.

166. Ernst PB, Peura D a, Crowe SE. The translation of Helicobacter pylori basic research to patient care. *Gastroenterology*. 2006;130(1):188-206; quiz 212-3. doi:10.1053/j.gastro.2005.06.032.

167. Berg RD. The indigenous gastrointestinal microflora. *Trends Microbiol*. 1996;4(11):430-5.

168. Erturk-Hasdemir D, Kasper DL. Resident commensals shaping immunity. *Curr Opin Immunol*. 2013;25(4):450-5. doi:10.1016/j.coi.2013.06.001.

169. Hold GL. Western lifestyle : a " master " manipulator of the intestinal microbiota ? *Gut*. 2013;0:1-2. doi:10.1136/gutjnl-2013-304969.

170. Hawrelak JA, Myers SP. The causes of intestinal dysbiosis: a review. *Altern Med Rev*. 2004;9(2):180-97.

171. Wu GD, Chen J, Hoffmann C, et al. Linking long-term dietary patterns with gut microbial enterotypes. *Science*. 2011;334(6052):105-8. doi:10.1126/science.1208344.

172. Bailey MT, Dowd SE, Galley JD, Hufnagle AR, Allen RG, Lyte M. Exposure to a social stressor alters the structure of the intestinal microbiota: implications for stressor-induced immunomodulation. *Brain Behav Immun*. 2011;25(3):397-407. doi:10.1016/j.bbi.2010.10.023.

173. Stevens MP. *Microbial Endocrinology*. (Lyte M, Freestone PPE, eds.). New York, NY: Springer New York; 2010:111-134. doi:10.1007/978-1-4419-5576-0.

174. Qin J, Li Y, Cai Z, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*. 2012;490(7418):55-60. doi:10.1038/nature11450.

175. Greenblum S, Turnbaugh PJ, Borenstein E. Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease. *PNAS*. 2012;109(2). doi:10.1073/pnas.1116053109/-/DCSupplemental.www.pnas.org/cgi/doi/10.1073/pnas.1116053109.

176. Geng J, Fan H, Tang X, Zhai H, Zhang Z. Diversified pattern of the human colorectal cancer microbiome. *Gut Pathog*. 2013;5(1):2. doi:10.1186/1757-4749-5-2.

177. Weir TL, Manter DK, Sheflin AM, Barnett B a, Heuberger AL, Ryan EP. Stool microbiome and metabolome differences between colorectal cancer patients and healthy adults. *PLoS One*. 2013;8(8):e70803. doi:10.1371/journal.pone.0070803.

178. Tjalsma H, Boleij A. Intestinal Host-Microbiome Interactions. *Oncology*. 2011.

179. Ohigashi S, Sudo K, Kobayashi D, et al. Changes of the intestinal microbiota, short chain fatty acids, and fecal pH in patients with colorectal cancer. *Dig Dis Sci*. 2013;58(6):1717-26. doi:10.1007/s10620-012-2526-4.

180. Spor A, Koren O, Ley R. Unravelling the effects of the environment and host genotype on the gut microbiome. *Nat Rev Microbiol*. 2011;9(4):279-90. doi:10.1038/nrmicro2540.

181. Arumugam M, Raes J, Pelletier E, et al. Enterotypes of the human gut microbiome. *Nature*. 2011:1-7. doi:10.1038/nature09944.

182. Candela M, Guidotti M, Fabbri A, Brigidi P, Franceschi C, Fiorentini C. Human intestinal microbiota : cross-talk with the host and its potential role in colorectal cancer. *Crit Rev Microbiol*. 2010;(March):1-14. doi:10.3109/1040841X.2010.501760.

183. Martinez-Medina M, Denizot J, Dreux N, et al. Western diet induces dysbiosis with increased E coli in CEABAC10 mice, alters host barrier function favouring AIEC colonisation. *Gut*. 2014;63(1):116-24. doi:10.1136/gutjnl-2012-304119.

184. Söderholm JD, Perdue. MH. Stress and intestinal barrier function. *Am J Physiol Liver Physiol*. 2001;1.

185. Luo B, Xiang D, Nieman DC, Chen P. The effects of moderate exercise on chronic stress-induced intestinal barrier dysfunction and antimicrobial defense. *Brain Behav Immun*. 2013. doi:10.1016/j.bbi.2013.11.013.

186. Reber SO, Peters S, Slattery DA, et al. Mucosal immunosuppression and epithelial barrier defects are key events in murine psychosocial stress-induced colitis. *Brain Behav Immun*. 2011;25(6):1153-61. doi:10.1016/j.bbi.2011.03.004.

187. Shigeshiro M, Tanabe S, Suzuki T. Repeated exposure to water immersion stress reduces the Muc2 gene level in the rat colon via two distinct mechanisms. *Brain Behav Immun*. 2012;26:1061-1065.

188. Freestone PPE, Haigh RD, Williams PH, Lyte M. Stimulation of bacterial growth by heat-stable , norepinephrine-induced autoinducers. *FEMS Microbiol Lett*. 1999;172:53-60.

189. Lyte M, Erickson AK, Arulanandam BP, Frank CD, Crawford MA, Francis DH. Norepinephrine-induced expression of the K99 pilus adhesin of enterotoxigenic Escherichia coli. *Biochem Biophys Res Commun*. 1997;232(3):682-6. doi:10.1006/bbrc.1997.6356.

190. Carvalho FA, Barnich N, Sivignon A, et al. Crohn's disease adherent-invasive Escherichia coli colonize and induce strong gut inflammation in transgenic mice expressing human CEACAM. *J Exp Med*. 2009;206(10):2179-89. doi:10.1084/jem.20090741.

191. Couturier-maillard A, Secher T, Rehman A, et al. NOD2-mediated dysbiosis predisposes mice to transmissible colitis and colorectal cancer. *J Clin Invest*. 2012. doi:10.1172/JCI62236DS1.

192. Tong M, McHardy I, Ruegger P, et al. Reprograming of gut microbiome energy metabolism by the FUT2 Crohn's disease risk polymorphism. *ISME J*. 2014:1-14. doi:10.1038/ismej.2014.64.

193. Kashyap PC, Marcobal A, Ursell LK, et al. Genetically dictated change in host mucus carbohydrate landscape exerts a diet-dependent effect on the gut microbiota. *Proc Natl Acad Sci U S A*. 2013;110(42):17059-64. doi:10.1073/pnas.1306070110.

194. Warren L, Freeman JD, Dreolini L, et al. Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma. *Genome Res*. 2012;22:299-306. doi:10.1101/gr.126516.111.

195. McCoy AN, Araújo-Pérez F, Azcárate-Peril A, Yeh JJ, Sandler RS, Keku TO. Fusobacterium is associated with colorectal adenomas. *PLoS One*. 2013;8(1):e53653. doi:10.1371/journal.pone.0053653.

196. Rubinstein MR, Wang X, Liu W, Hao Y, Cai G, Han YW. Fusobacterium nucleatum promotes colorectal carcinogenesis by modulating E-cadherin/β-catenin signaling via its FadA adhesin. *Cell Host Microbe*. 2013;14(2):195-206. doi:10.1016/j.chom.2013.07.012.

197. Gupta A, Madani R, Mukhtar H. Streptococcus bovis endocarditis , a silent sign for colonic tumour. *Color Dis*. 2010;12:164-171. doi:10.1111/j.1463-1318.2009.01814.x.

198. Balamurugan R, Rajendiran E, George S, Samuel GV, Ramakrishna BS. Real-time polymerase chain reaction quantification of specific butyrate-producing bacteria, Desulfovibrio and Enterococcus faecalis in the feces of patients with colorectal cancer. *J Gastroenterol Hepatol*. 2008;23(8 Pt 1):1298-303. doi:10.1111/j.1440-1746.2008.05490.x.

199. Huycke MM, Abrams V, Moore DR. Enterococcus faecalis produces extracellular superoxide and hydrogen peroxide that damages colonic epithelial cell DNA. *Carcinogenesis*. 2002;23(3):529-536.

200. Balish E, Warner T. Enterococcus faecalis induces inflammatory bowel disease in interleukin-10 knockout mice. *Am J Pathol*. 2002;160(6):2253-7. doi:10.1016/S0002-9440(10)61172-8.

201. Kim SC, Tonkonogy SL, Albright CA, et al. Variable phenotypes of enterocolitis in interleukin 10-deficient mice monoassociated with two different commensal bacteria. *Gastroenterology*. 2005;128(4):891-906. doi:10.1053/j.gastro.2005.02.009.

202. Wang X, Huycke MM. Extracellular superoxide production by Enterococcus faecalis promotes chromosomal instability in mammalian cells. *Gastroenterology*. 2007;132(2):551-61. doi:10.1053/j.gastro.2006.11.040.

203. Wu S, Morin PJ, Maouyo D, Sears CL. Bacteroides fragilis enterotoxin induces c-Myc expression and cellular proliferation. *Gastroenterology*. 2003;124(2):392-400. doi:10.1053/gast.2003.50047.

204. Burnett-Hartman AN, Newcomb PA, Potter JD. Infectious agents and colorectal cancer: a review of Helicobacter pylori, Streptococcus bovis, JC virus, and human papillomavirus. *Cancer Epidemiol Biomarkers Prev*. 2008;17(11):2970-9. doi:10.1158/1055-9965.EPI-08-0571.

205. Collins D, Hogan AM, Winter DC. Microbial and viral pathogens in colorectal cancer. *Lancet Oncol*. 2010;2045(10). doi:10.1016/S1470-2045(10)70186-8.

206. Tenaillon O, Skurnik D, Picard B, Denamur E. The population genetics of commensal Escherichia coli. *Nat Rev Microbiol*. 2010;8(3):207-17. doi:10.1038/nrmicro2298.

207. Swidsinski A, Khilkin M, Kerjaschki D, et al. Association Between Intraepithelial Escerichia coli and Colorectal Cancer. *Gastroenterology*. 1998;115:281-286.

208. Kotlowski R, Bernstein CN, Sepehri S, Krause DO. High prevalence of Escherichia coli belonging to the B2+D phylogenetic group in inflammatory bowel disease. *Gut*. 2007;56(5):669-75. doi:10.1136/gut.2006.099796.

209. Boudeau J, Glasser A, Masseret E, Joly B, Darfeuille-michaud A. Invasive Ability of an Escherichia coli Strain Isolated from the Ileal Mucosa of a Patient with Crohn ' s Disease. *Infect Immun*. 1999;67(9):4499-4509.

210. Miquel S, Peyretaillade E, Claret L, et al. Complete genome sequence of Crohn's disease-associated adherent-invasive E. coli strain LF82. *PLoS One*. 2010;5(9). doi:10.1371/journal.pone.0012714.

211. Arthur JC, Perez-Chanona E, Mühlbauer M, et al. Intestinal inflammation targets cancer-inducing activity of the microbiota. *Science*. 2012;338(6103):120-3. doi:10.1126/science.1224820.

212. Buc E, Dubois D, Sauvanet P, Raisch J, Delmas J. High Prevalence of Mucosa-Associated E . coli Producing Cyclomodulin and Genotoxin in Colon Cancer. *PLoS One*. 2013;8(2). doi:10.1371/journal.pone.0056964.

213. Prorok-Hamon M, Friswell MK, Alswied A, et al. Colonic mucosa-associated diffusely adherent afaC+ Escherichia coli expressing lpfA and pks are increased in inflammatory bowel disease and colon cancer. *Gut*. 2014;63(5):761-70. doi:10.1136/gutjnl-2013-304739.

214. Yamamoto D, Hernandes RT, Blanco M, et al. Invasiveness as a putative additional virulence mechanism of some atypical Enteropathogenic Escherichia coli strains with different uncommon intimin types. *BMC Microbiol*. 2009;9:146. doi:10.1186/1471-2180-9-146.

215. Schauer DB, Zabel BA, Pedraza IF, et al. Genetic and biochemical characterization of Citrobacter rodentium sp . nov . *J Clin Microbiol*. 1995;33(8):2064-2068.

216. Hardwidge PR, Rodriguez-Escudero I, Goode D, et al. Proteomic analysis of the intestinal epithelial cell response to enteropathogenic Escherichia coli. *J Biol Chem*. 2004;279(19):20127-36. doi:10.1074/jbc.M401228200.

217. Wong ARC, Pearson JS, Bright MD, et al. Enteropathogenic and enterohaemorrhagic Escherichia coli: even more subversive elements. *Mol Microbiol*. 2011;80(6):1420-38. doi:10.1111/j.1365-2958.2011.07661.x.

218. Kostic AD, Chun E, Robertson L, et al. Fusobacterium nucleatum potentiates intestinal tumorigenesis and modulates the tumor-immune microenvironment. *Cell Host Microbe*. 2013;14(2):207-15. doi:10.1016/j.chom.2013.07.007.

219. Tahara T, Yamamoto E, Suzuki H, et al. Fusobacterium in colonic flora and molecular features of colorectal carcinoma. *Cancer Res*. 2014;74(5):1311-8. doi:10.1158/0008-5472.CAN-13-1865.

220. McCoy, W. C. and JMM 3rd. Enterococcal endocarditis associated with carcinoma of the sigmoid; report of a case. *J Med Assoc State Ala*. 1951;21(6).

221. Schlegel L, Grimont F, Ageron E, Grimont, Patrick AD, Bouvet. A. Reappraisal of the taxonomy of the Streptococcus bovis/Streptococcus equinus complex and related species: description of Streptococcus gallolyticus subsp. gallolyticus subsp. nov., S. gallolyticus subsp. macedonicus subsp. nov. and S. gallolyticus subsp. . *Int J Syst Evol Microbiol*. 2003;53(3):631-645.

222. Solheim M, Aakra A, Snipen LG, Brede D a, Nes IF. Comparative genomics of Enterococcus faecalis from healthy Norwegian infants. *BMC Genomics*. 2009;10:194. doi:10.1186/1471-2164-10-194.

223. Wang X, Allen TD, May RJ, Lightfoot S, Houchen CW, Huycke MM. Enterococcus faecalis induces aneuploidy and tetraploidy in colonic epithelial cells through a bystander effect. *Cancer Res*. 2008;68(23):9909-17. doi:10.1158/0008-5472.CAN-08-1551.

224. Zhang G, Svenungsson B, Karnell A, Weibtraub A. Prevalence of Enterotoxigenic Bacteroides fragilis in Adult Patients with Diarrhea and Healthy Controls. *Clin Infect Dis*. 1999;29:590-594.

225. Sears CL. Enterotoxigenic Bacteroides fragilis: a rogue among symbiotes. *Clin Microbiol Rev*. 2009;22(2):349-69, Table of Contents. doi:10.1128/CMR.00053-08.

226. Qiu W, Lee M-LT, Whitmore GA. Sample Size and Power Calculation in Microarray Studies Using the sizepower package. 2013.

227. Simon RM, Dobbin K. Experimental design of DNA microarray experiments. *Biotechniques*. 2003;Suppl:16-21. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12664680.

228. StatsToDo. Available at: https://www.statstodo.com/SSizUnequal_Pgm.php. Accessed June 16, 2014.

229. Smyth GK. Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. *Stat Appl Genet Mol Biol*. 2004;3(1).

230. Imai K, Yamamoto H. Carcinogenesis and microsatellite instability: the interrelationship between genetics and epigenetics. *Carcinogenesis*. 2008;29(4):673-80. doi:10.1093/carcin/bgm228.

231. Loukola A, Eklin K, Laiho P, et al. Microsatellite marker analysis in screening for hereditary nonpolyposis colorectal cancer (HNPCC). *Cancer Res*. 2001;61(11):4545-9.

232. Cicek MS, Lindor NM, Gallinger S, et al. Quality assessment and correlation of microsatellite instability and immunohistochemical markers among population- and clinic-based colorectal

tumors results from the Colon Cancer Family Registry. *J Mol Diagn*. 2011;13(3):271-81. doi:10.1016/j.jmoldx.2010.12.004.

233. Schiff M, Castle PE, Jeronimo J, Rodriguez AC, Wacholder S. Human papillomavirus and cervical cancer. *Lancet*. 2007;370:890-907.

234. Manuel A, Machado D, Figueiredo C, Seruca R, Juel L. Helicobacter pylori infection generates genetic instability in gastric cells. *Biochim Biophys Acta*. 2010. doi:10.1016/j.bbcan.2010.01.007.

235. Vogelmann R, Amieva MR. The role of bacterial pathogens in cancer. *Curr Opin Microbiol*. 2007;10(1):76-81. doi:10.1016/j.mib.2006.12.004.

236. McCoy, W. C. and JMM 3rd. Enterococcal endocarditis associated with carcinoma of the sigmoid; report of a case. *ournal Med Assoc State Alabama*. 1951;21(6).

237. Kostic AD, Gevers D, Pedamallu CS, et al. Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. *Genome Res*. 2012:292-298. doi:10.1101/gr.126573.111.

238. Toprak NU, Yagci A, Gulluoglu BM, et al. A possible role of Bacteroides fragilis enterotoxin in the aetiology of colorectal cancer. *Clin Microbiol Infect*. 2006;12(8):782-6. doi:10.1111/j.1469-0691.2006.01494.x.

239. Umar S, Wang Y, Morris AP, Sellin JH. Dual alterations in casein kinase I-E and GSK-3β modulate β-catenin stability in hyperproliferating colonic epithelia. *Am J Physiol Gastrointest Liver Physiol*. 2007;292:599-607. doi:10.1152/ajpgi.00343.2006.

240. Ellmerich S, Schöller M, Duranton B, et al. Promotion of intestinal carcinogenesis by Streptococcus bovis. *Carcinogenesis*. 2000;21(4):753-6.

241. Biarc J, Nguyen IS, Pini A, et al. Carcinogenic properties of proteins with pro-inflammatory activity from Streptococcus infantarius (formerly S.bovis). *Carcinogenesis*. 2004;25(8):1477-84. doi:10.1093/carcin/bgh091.

242. Martin HM, Campbell BJ, Hart CA, et al. Enhanced Escherichia coli adherence and invasion in Crohn's disease and colon cancer. *Gastroenterology*. 2004;127(1):80-93. doi:10.1053/j.gastro.2004.03.054.

243. Corredoira-Sánchez J, García-Garrote F, Rabuñal R, et al. Association between bacteremia due to Streptococcus gallolyticus subsp. gallolyticus (Streptococcus bovis I) and colorectal neoplasia: a case-control study. *Clin Infect Dis*. 2012;55(4):491-6. doi:10.1093/cid/cis434.

244. Ye J, Coulouris G, Zaretskaya I, Cutcutache I, Rozen S, Madden TL. Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics*. 2012;13:134. doi:10.1186/1471-2105-13-134.

245. Vidal R, Vidal M, Lagos R, Levine M, Prado V. Multiplex PCR for Diagnosis of Enteric Infections Associated with Diarrheagenic Escherichia coli. *Society*. 2004;42(4):1787-1789. doi:10.1128/JCM.42.4.1787.

246. Derzelle S, Grine A, Madic J, et al. A quantitative PCR assay for the detection and quantification of Shiga toxin-producing Escherichia coli (STEC) in minced beef and dairy products. *Int J Food Microbiol*. 2011;151(1):44-51. doi:10.1016/j.ijfoodmicro.2011.07.039.

247. Chiodini RJ, Dowd SE, Davis B, et al. Crohn's disease may be differentiated into 2 distinct biotypes based on the detection of bacterial genomic sequences and virulence genes within submucosal tissues. *J Clin Gastroenterol*. 2013;47(7):612-20. doi:10.1097/MCG.0b013e31827b4f94.

248. Sedgley CM, Nagel AC, Shelburne CE, Clewell DB, Appelbe O, Molander A. Quantitative real-time PCR detection of oral Enterococcus faecalis in humans. *Arch Oral Biol*. 2005;50(6):575-83. doi:10.1016/j.archoralbio.2004.10.017.

249. Pantosti A, Malpeli M, Wilks M, Menozzi MG, D'Ambrosio F. Detection of enterotoxigenic Bacteroides fragilis by PCR. *J Clin Microbiol*. 1997;35(10):2482-6.

250. Walter J, Margosch D, Hammes WP, Hertel C. Detection of Fusobacterium Species in Human Feces Using Genus-Specific PCR Primers and Denaturing Gradient Gel Electrophoresis. *Microb Ecol Health Dis*. 2002;14:129-132.

251. Don RH, Cox PT, Wainwright BJ, Baker K, Mattick JS. "Touchdown" PCR to circumvent spurious priming during gene amplification. *Nucleic Acids Res*. 1991;19(14):4008.

252. Korbie DJ, Mattick JS. Touchdown PCR for increased specificity and sensitivity in PCR amplification. *Online*. 2008:13-15. doi:10.1038/nprot.2008.133.

253. Dolezel J, Bartos J, Voglmayr H, Greilhuber J. Nuclear DNA content and genome size of trout and human. *Cytometry A*. 2003;51(2):127-8. doi:10.1002/cyto.a.10013.

254. Rusniok C, Couvé E, Da Cunha V, et al. Genome sequence of Streptococcus gallolyticus: insights into its adaptation to the bovine rumen and its ability to cause endocarditis. *J Bacteriol*. 2010;192(8):2266-76. doi:10.1128/JB.01659-09.

255. Van Beers EH, Joosse SA, Ligtenberg MJ, et al. A multiplex PCR predictor for aCGH success of FFPE samples. *Br J Cancer*. 2006;94(2):333-7. doi:10.1038/sj.bjc.6602889.

256. Buc E, Dubois D, Sauvanet P, et al. High prevalence of mucosa-associated E. coli producing cyclomodulin and genotoxin in colon cancer. *PLoS One*. 2013;8(2):e56964. doi:10.1371/journal.pone.0056964.

257. Xue-han Z, Qing Y, Ya-dong L, Bin L, Renata I, Kong-wang H. Development of a LAMP assay for rapid detection of different intimin variants of attaching and effacing microbial pathogens. *J Med Microbiol*. 2013;62(11):1665-1672.

258. Phillips AD, Navabpour S, Hicks S, et al. Enterohaemorrhagic Escherichia coli O157:H7 target Peyer's patches in humans and cause attaching/effacing lesions in both human and bovine intestine. 2000:377-381.

259. R J Fitzhenry, D J Pickard, E L Hartland, S Reece, G Dougan, A D Phillips GF. Intimin type influences the site of human intestinal mucosal colonisation by enterohaemorrhagic Escherichia coli O157:H7. *Gut*. 2002;50:180-185.

260. Mundy R, Schüller S, Girard F, Fairbrother JM, Phillips AD, Frankel G. Functional studies of intimin in vivo and ex vivo: implications for host specificity and tissue tropism. *Microbiology*. 2007;153(Pt 4):959-67. doi:10.1099/mic.0.2006/003467-0.

261. Bonnet M, Buc E, Sauvanet P, et al. Colonization of the human gut by E. coli and colorectal cancer risk. *Clin Cancer Res*. 2014;20(4):859-67. doi:10.1158/1078-0432.CCR-13-1343.

262. Strauss J, Kaplan GG, Beck PL, et al. Invasive potential of gut mucosa-derived Fusobacterium nucleatum positively correlates with IBD status of the host. *Inflamm Bowel Dis*. 2011;17(9):1971-8. doi:10.1002/ibd.21606.

263. Kaz A, Kim YH, Dzieciatkowski S, et al. Evidence for the role of aberrant DNA methylation in the pathogenesis of Lynch syndrome adenomas. *Int J Cancer*. 2007;120(9):1922-9. doi:10.1002/ijc.22544.

264. Speake D, O'Sullivan J, Evans DG, Lalloo F, Hill J, McMahon RFT. Hyperplastic polyps are innocuous lesions in hereditary nonpolyposis colorectal cancers. *Int J Surg Oncol*. 2011;2011:653163. doi:10.1155/2011/653163.

265. Parsons MT, Buchanan DD, Thompson B, Young JP, Spurdle AB. Correlation of tumour BRAF mutations and MLH1 methylation with germline mismatch repair (MMR) gene mutation status: a literature review assessing utility of tumour features for MMR variant classification. *J Med Genet*. 2012;49(3):151-7. doi:10.1136/jmedgenet-2011-100714.

266. Bettstetter M, Dechant S, Ruemmele P, et al. Distinction of hereditary nonpolyposis colorectal cancer and sporadic microsatellite-unstable colorectal cancer through quantification of MLH1 methylation by real-time PCR. *Clin Cancer Res*. 2007;13(11):3221-8. doi:10.1158/1078-0432.CCR-06-3064.

267. Ollikainen M, Hannelius U, Lindgren CM, Abdel-Rahman WM, Kere J, Peltomäki P. Mechanisms of inactivation of MLH1 in hereditary nonpolyposis colorectal carcinoma: a novel approach. *Oncogene*. 2007;26(31):4541-9. doi:10.1038/sj.onc.1210236.

268. Zhang J, Lindroos A, Ollila S, et al. Gene conversion is a frequent mechanism of inactivation of the wild-type allele in cancers from MLH1/MSH2 deletion carriers. *Cancer Res*. 2006;66(2):659-64. doi:10.1158/0008-5472.CAN-05-4043.

269. Wheeler JM, Loukola A, Aaltonen LA, Mortensen NJ, Bodmer WF. The role of hypermethylation of the hMLH1 promoter region in HNPCC versus MSI+ sporadic colorectal cancers. *J Med Genet*. 2000;37(8):588-92.

270. Jass JR. HNPCC and sporadic MSI-H colorectal cancer : a review of the morphological similarities and differences. *Am J Pathol*. 2004;3:93-100.

271. Patil DT, Shadrach BL, Rybicki LA, Leach BH, Pai RK. Proximal colon cancers and the serrated pathway: a systematic analysis of precursor histology and BRAF mutation status. *Mod Pathol*. 2012;25(10):1423-31. doi:10.1038/modpathol.2012.98.

272. Gebert J, Kloor M, Lee J, et al. Colonic carcinogenesis along different genetic routes: glycophenotyping of tumor cases separated by microsatellite instability/stability. *Histochem Cell Biol*. 2012;138(2):339-50. doi:10.1007/s00418-012-0957-9.

273. Campbell BJ, Yu LG, Rhodes JM. Altered glycosylation in inflammatory bowel disease: a possible role in cancer development. *Glycoconj J*. 2003;18(11-12):851-8.

274. Kostic AD, Chun E, Meyerson M, Garrett WS. Microbes and inflammation in colorectal cancer. *Cancer Immunol Res*. 2013;1(3):150-7. doi:10.1158/2326-6066.CIR-13-0101.

275. Dutilh BE, Backus L, van Hijum SAFT, Tjalsma H. Screening metatranscriptomes for toxin genes as functional drivers of human colorectal cancer. *Best Pract Res Clin Gastroenterol*. 2013;27:85-99.

276. Viljoen KS, Blackburn JM. Quality assessment and data handling methods for Affymetrix Gene 1.0 ST arrays with variable RNA integrity. *BMC Genomics*. 2013;14(1):14. doi:10.1186/1471-2164-14-14.

277. Tomita H, Vawter MP, Walsh DM, et al. Effect of agonal and postmortem factors on gene expression profile: quality control in microarray analyses of postmortem human brain. *Biol Psychiatry*. 2004;55(4):346-52. doi:10.1016/j.biopsych.2003.10.013.

278. Mengual L, Burset M, Ars E, et al. Partially Degraded RNA from Bladder Washing is a Suitable Sample for Studying Gene Expression Profiles in Bladder Cancer. *Eur Urol*. 2006;50:1347-1356. doi:10.1016/j.eururo.2006.05.039.

279. Linton KM, Hey Y, Saunders E, et al. Acquisition of biologically relevant gene expression data by Affymetrix microarray analysis of archival formalin-fixed paraffin-embedded tumours. *Br J Cancer*. 2008;98(8):1403-14. doi:10.1038/sj.bjc.6604316.

280. Linton K, Hey Y, Dibben S, et al. Methods comparison for high-resolution transcriptional analysis of archival material on Affymetrix Plus 2.0 and Exon 1.0 microarrays. *Biotechniques*. 2009;47(July):587-96. doi:10.2144/000113169.

281. April C, Klotzle B, Royce T, et al. Whole-Genome Gene Expression Profiling of Formalin- Fixed , Paraffin-Embedded Tissue Samples. *PLoS One*. 2009;4(12):1-10. doi:10.1371/journal.pone.0008162.

282. Opitz L, Salinas-riester G, Grade M, et al. Impact of RNA degradation on gene expression profiling. *BMC Med Genomics*. 2010;3(36):1-14. doi:10.1186/1755-8794-3-36.

283. Fleige S, Pfaffl MW. RNA integrity and the effect on the real-time qRT-PCR performance. *Mol Aspects Med*. 2006;27:126-139. doi:10.1016/j.mam.2005.12.003.

284. Lassmann S, Kreutz C, Schoepflin A, Hopt U, Timmer J, Werner M. A novel approach for reliable microarray analysis of microdissected tumor cells from formalin-fixed and paraffin-embedded colorectal cancer resection specimens. *J Mol Med*. 2009;87:211-24. doi:10.1007/s00109-008-0419-y.

285. Antonov J, Goldstein DR, Oberli A, et al. Reliable gene expression measurements from degraded RNA by quantitative real-time PCR depend on short amplicons and a proper normalization. *Lab Invest*. 2005;85(8):1040-50. doi:10.1038/labinvest.3700303.

286. Binder H, Krohn K, Preibisch S. “Hook”-calibration of GeneChip-microarrays: chip characteristics and expression measures. *Algorithms Mol Biol*. 2008;3(11). doi:10.1186/1748-7188-3-11.

287. Chow ML, Winn ME, Li H-R, et al. Preprocessing and quality control strategies for Illumina DASL assay-based brain gene expression studies with semi-degraded samples. *Front Genet*. 2012;3. doi:10.3389/fgene.2012.00011.

288. Johnson WE, Li C. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*. 2007;8(1):118-127. doi:10.1093/biostatistics/kxj037.

289. Leek JT, Storey JD. Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. *PLoS Genet*. 2007;3(9):1724-1735. doi:10.1371/journal.pgen.0030161.

290. Salazar R, Roepman P, Capella G, et al. Gene expression signature to improve prognosis prediction of stage II and III colorectal cancer. *J Clin Oncol*. 2011;29(1):17-24. doi:10.1200/JCO.2010.30.1077.

291. O'Connell MJ, Lavery I, Yothers G, et al. Relationship between tumor gene expression and recurrence in four independent studies of patients with stage II/III colon cancer treated with surgery alone or surgery plus adjuvant fluorouracil plus leucovorin. *J Clin Oncol*. 2010;28(25):3937-3944. doi:10.1200/JCO.2010.28.9538.

292. Dydensborg AB, Herring E, Auclair J, Tremblay E, Beaulieu J-F. Normalizing genes for quantitative RT-PCR in differentiating human intestinal epithelial cells and adenocarcinomas of the colon. *Am J Gastrointest Liver Physiol*. 2006;290:G1067-G1074. doi:10.1152/ajpgi.00234.2005.

293. Rubie C, Kempf K, Hans J, et al. Housekeeping gene variability in normal and cancerous colorectal, pancreatic, esophageal, gastric and hepatic tissues. *Mol Cell Probes*. 2005;19:101-109. doi:10.1016/j.mcp.2004.10.001.

294. Kheirelseid EAH, Chang KH, Newell J, Kerin MJ, Miller N. Identification of endogenous control genes for normalisation of real-time quantitative PCR data in colorectal cancer. *BMC Mol Biol*. 2010;11(12). doi:10.1186/1471-2199-11-12.

295. 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(15):5245-5250. doi:10.1158/0008-5472.CAN-04-0496.

296. Vandesompele J, Preter K De, Poppe B, Roy N Van, Paepe A De. Accurate normalization of real-time quantitative RT -PCR data by geometric averaging of multiple internal control genes. *Genome*. 2002;3(7).

297. Mccall MN, Bolstad BM, Irizarry RA. Frozen robust multiarray analysis (fRMA). *Biostatistics*. 2010;11(2):242-253. doi:10.1093/biostatistics/kxp059.

298. Carvalho B, Irizarry RA, Scharpf RB, Carey VJ. Processing and Analyzing Affymetrix SNP Chips with Bioconductor. *Stat Biosci*. 2009;1:160-180. doi:10.1007/s12561-009-9015-0.

299. McCall MN, Murakami PN, Lukk M, Huber W, Irizarry R a. Assessing affymetrix GeneChip microarray quality. *BMC Bioinformatics*. 2011;12(1):137. doi:10.1186/1471-2105-12-137.

300. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *J R Stat Soc*. 1995;57(1):289-300.

301. Young MD, Wakefield MJ, Smyth GK, Oshlack A. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol*. 2010;11(2):R14. doi:10.1186/gb-2010-11-2-r14.

302. Gibbons FD, Roth FP. Judging the Quality of Gene Expression-Based Clustering Methods Using Gene Annotation. *Genome Res*. 2002;(617):1574-1581. doi:10.1101/gr.397002.1574.

303. Dalman MR, Deeter A, Nimishakavi G, Duan Z-H. Fold change and p-value cutoffs significantly alter microarray interpretations. *BMC Bioinformatics*. 2012;13 Suppl 2(Suppl 2):S11. doi:10.1186/1471-2105-13-S2-S11.

304. Jung M, Schaefer A, Steiner I, et al. Robust microRNA stability in degraded RNA preparations from human tissue and cell samples. *Clin Chem*. 2010;56(6):998-1006. doi:10.1373/clinchem.2009.141580.

305. Hinoue T, Weisenberger DJ, Lange CPE, et al. Genome-scale analysis of aberrant DNA methylation in colorectal cancer Genome-scale analysis of aberrant DNA methylation in colorectal cancer. *Genome Res*. 2011. doi:10.1101/gr.117523.110.

306. Weisenberger DJ, Siegmund KD, Campan M, et al. CpG island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with BRAF mutation in colorectal cancer. *Nat Genet*. 2006;38(7):787-93. doi:10.1038/ng1834.

307. Bettington M, Walker N, Clouston A, Brown I, Leggett B, Whitehall V. The serrated pathway to colorectal carcinoma: current concepts and challenges. *Histopathology*. 2013;62(3):367-86. doi:10.1111/his.12055.

308. Gardiner-Garden M, Frommer M. CpG islands in vertebrate genomes. *J Mol Biol*. 1987;196:261-282. doi:10.1016/0022-2836(87)90689-9.

309. Zhao Z, Han L. CpG islands: Algorithms and applications in methylation studies. *Biochem Biophys Res Commun*. 2009;382(4):643-645. doi:10.1016/j.bbrc.2009.03.076.

310. Bergman Y, Cedar H. DNA methylation dynamics in health and disease. *Nat Struct Mol Biol*. 2013;20(3):274-81. doi:10.1038/nsmb.2518.

311. Slotkin RK, Martienssen R. Transposable elements and the epigenetic regulation of the genome. *Nat Rev Genet*. 2007;8(4):272-85. doi:10.1038/nrg2072.

312. Yu DH, Ware C, Waterland RA, et al. Developmentally programmed 3' CpG island methylation confers tissue- and cell-type-specific transcriptional activation. *Mol Cell Biol*. 2013;33(9):1845-58. doi:10.1128/MCB.01124-12.

313. Shukla S, Kavak E, Gregory M, et al. CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature*. 2011;479(7371):74-9. doi:10.1038/nature10442.

314. Toyota M, Ho C, Ahuja N. Identification of Differentially Methylated Sequences in Colorectal Cancer by Methylated CpG Island Amplification. *Cancer Res*. 1999;(59):2307-2312.

315. Toyota M, Ahuja N, Ohe-Toyota M, Herman JG, Baylin SB, Issa JJ. CpG island methylator phenotype in colorectal cancer. *Proc Natl Acad Sci USA*. 1999;96(July):8681-8686.

316. Suzuki H, Yamamoto E, Maruyama R, Niinuma T, Kai M. Biological significance of the CpG island methylator phenotype. *Biochem Biophys Res Commun*. 2014. doi:http://dx.doi.org/10.1016/j.bbrc.2014.07.007.

317. Hinoue T, Weisenberger DJ, Lange CPE, et al. Genome-scale analysis of aberrant DNA methylation in colorectal cancer. *Genome Res*. 2011. doi:10.1101/gr.117523.110.

318. Noreen F, Röösli M, Gaj P, et al. Modulation of age- and cancer-associated DNA methylation change in the healthy colon by aspirin and lifestyle. *J Natl Cancer Inst*. 2014;106(7):1-9. doi:10.1093/jnci/dju161.

319. Stein RA. Epigenetics-the link between infectious diseases and cancer. *JAMA*. 2011;305(14):1484-5. doi:10.1001/jama.2011.446.

320. Shin CM, Kim N, Jung Y, Park JH, Kang GH, Kim JS. Role of Helicobacter pylori infection in aberrant DNA methylation along multistep gastric carcinogenesis. *Cancer Sci*. 2010;101(6):1337-1346. doi:10.1111/j.1349-7006.2010.01535.x.

321. Gómez-Díaz E, Jordà M, Peinado MA, Rivero A. Epigenetics of host-pathogen interactions: the road ahead and the road behind. *PLoS Pathog*. 2012;8(11):e1003007. doi:10.1371/journal.ppat.1003007.

322. Bobetsis YA, Barros SP, Lin DM, et al. Bacterial infection promotes DNA hypermethylation. *J Dent Res*. 2007;86(2):169-74.

323. Bock C. Analysing and interpreting DNA methylation data. *Nat Rev Genet*. 2012;13(10):705-19. doi:10.1038/nrg3273.

324. Laird PW. Principles and challenges of genomewide DNA methylation analysis. *Nat Rev Genet*. 2010;11(3):191-203. doi:10.1038/nrg2732.

325. Illumina. Infinium HumanMethylation450 BeadChip. 2012. Available at: http://res.illumina.com/documents/products/datasheets/datasheet_humanmethylation450.pdf.

326. Bibikova M, Barnes B, Tsan C, et al. High density DNA methylation array with single CpG site resolution. *Genomics*. 2011;98(4):288-95. doi:10.1016/j.ygeno.2011.07.007.

327. Du P, Zhang X, Huang C, et al. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics*. 2010;11(1):587. doi:10.1186/1471-2105-11-587.

328. Triche TJ, Weisenberger DJ, Van Den Berg D, Laird PW, Siegmund KD. Low-level processing of Illumina Infinium DNA Methylation BeadArrays. *Nucleic Acids Res*. 2013;41(7):e90. doi:10.1093/nar/gkt090.

329. Wilhelm-Benartzi CS, Koestler DC, Karagas MR, et al. Review of processing and analysis methods for DNA methylation array data. *Br J Cancer*. 2013;109(6):1394-402. doi:10.1038/bjc.2013.496.

330. Illumina. Infinium HD Assay Methylation Protocol Guide. 2011:1-247.

331. Aryee MJ, Jaffe AE, Corrada-Bravo H, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics*. 2014;30(10):1363-9. doi:10.1093/bioinformatics/btu049.

332. Teschendorff AE, Marabita F, Lechner M, et al. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics*. 2013;29(2):189-96. doi:10.1093/bioinformatics/bts680.

333. Pidsley R, Y Wong CC, Volta M, Lunnon K, Mill J, Schalkwyk LC. A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC Genomics*. 2013;14:293. doi:10.1186/1471-2164-14-293.

334. Wang D, Yan L, Hu Q, et al. IMA : An R package for high-throughput analysis of Illumina ' s 450K Infinium methylation data. *Oxford Univ Press*. 2012:1-3.

335. Barfield RT, Kilaru V, Smith AK, Conneely KN. CpGassoc: an R function for analysis of DNA methylation microarray data. *Bioinformatics*. 2012;28(9):1280-1. doi:10.1093/bioinformatics/bts124.

336. Naeem H, Wong NC, Chatterton Z, et al. Reducing the risk of false discovery enabling identification of biologically significant genome-wide methylation status using the HumanMethylation450 array. *BMC Genomics*. 2014;15(1):51. doi:10.1186/1471-2164-15-51.

337. Houseman EA, Christensen BC, Yeh R-F, et al. Model-based clustering of DNA methylation array data: a recursive-partitioning algorithm for high-dimensional data arising as a mixture of beta distributions. *BMC Bioinformatics*. 2008;9:365. doi:10.1186/1471-2105-9-365.

338. Emmert-Streib F. *Statistical Diagnostics For Cancer: Analyzing High-Dimensional Data*. (Matthias Dehmer, ed.). John Wiley & Sons; 2012.

339. Hinoue T, Weisenberger DJ, Pan F, et al. Analysis of the association between CIMP and BRAF in colorectal cancer by DNA methylation profiling. *PLoS One*. 2009;4(12):e8357. doi:10.1371/journal.pone.0008357.

340. Hahsler M. Getting Things in Order : An Introduction to the R Package seriation. 2008;25(3).

341. Gaujoux R. Generating heatmaps for Nonnegative Matrix Factorization. 2014:1-11.

342. Bettstetter M, Dechant S, Ruemmele P, et al. Distinction of hereditary nonpolyposis colorectal cancer and sporadic microsatellite-unstable colorectal cancer through quantification of MLH1 methylation by real-time PCR. *Clin Cancer Res*. 2007;13(11):3221-8. doi:10.1158/1078-0432.CCR-06-3064.

343. Van Roon EH, Boot A, Dihal AA, et al. BRAF mutation-specific promoter methylation of FOX genes in colorectal cancer. *Clin Epigenetics*. 2013;5(1):2. doi:10.1186/1868-7083-5-2.

344. Hellebrekers DMEI, Lentjes MHFM, van den Bosch SM, et al. GATA4 and GATA5 are potential tumor suppressors and biomarkers in colorectal cancer. *Clin Cancer Res*. 2009;15(12):3990-7. doi:10.1158/1078-0432.CCR-09-0055.

345. Wendt MK, Cooper AN, Dwinell MB. Epigenetic silencing of CXCL12 increases the metastatic potential of mammary carcinoma cells. *Oncogene*. 2008;27(10):1461-71. doi:10.1038/sj.onc.1210751.

346. Suzuki M, Mohamed S, Nakajima T, et al. Aberrant methylation of CXCL12 in non-small cell lung cancer is associated with an unfavorable prognosis. *Int J Oncol*. 2008;33(1):113-9. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18575756.

347. Wendt MK, Johanesen PA, Kang-Decker N, Binion DG, Shah V, Dwinell MB. Silencing of epithelial CXCL12 expression by DNA hypermethylation promotes colonic carcinoma metastasis. *Oncogene*. 2006;25(36):4986-97. doi:10.1038/sj.onc.1209505.

348. Cheng YY, Jin H, Liu X, et al. Fibulin 1 is downregulated through promoter hypermethylation in gastric cancer. *Br J Cancer*. 2008;99(12):2083-7. doi:10.1038/sj.bjc.6604760.

349. Xiao W, Wang J, Li H, et al. Fibulin-1 is down-regulated through promoter hypermethylation and suppresses renal cell carcinoma progression. *J Urol*. 2013;190(1):291-301. doi:10.1016/j.juro.2013.01.098.

350. Kanda M, Nomoto S, Okamura Y, et al. Promoter hypermethylation of fibulin 1 gene is associated with tumor progression in hepatocellular carcinoma. *Mol Carcinog*. 2011;50(8):571-9. doi:10.1002/mc.20735.

351. Sartor MA, Dolinoy DC, Jones TR, et al. Genome-wide methylation and expression differences in HPV(+) and HPV(-) squamous cell carcinoma cell lines are consistent with divergent mechanisms of carcinogenesis. *Epigenetics*. 2011;6(6):777-787. doi:10.4161/epi.6.6.16216.

352. Jithesh P V, Risk JM, Schache AG, et al. The epigenetic landscape of oral squamous cell carcinoma. *Br J Cancer*. 2013;108(2):370-9. doi:10.1038/bjc.2012.568.

353. Vanuytsel T, Senger S, Fasano A, Shea-Donohue T. Major signaling pathways in intestinal stem cells. *Biochim Biophys Acta*. 2013;1830(2):2410-26. doi:10.1016/j.bbagen.2012.08.006.

354. Stresemann C, Lyko F. Modes of action of the DNA methyltransferase inhibitors azacytidine and decitabine. *Int J Cancer*. 2008;123(1):8-13. doi:10.1002/ijc.23607.

355. Coppedè F. Epigenetic biomarkers of colorectal cancer: Focus on DNA methylation. *Cancer Lett*. 2014;342(2):238-47. doi:10.1016/j.canlet.2011.12.030.

356. Gyparaki M-T, Basdra EK, Papavassiliou AG. DNA methylation biomarkers as diagnostic and prognostic tools in colorectal cancer. *J Mol Med (Berl)*. 2013;91(11):1249-56. doi:10.1007/s00109-013-1088-z.

357. De Sousa E Melo F, Wang X, Jansen M, et al. Poor-prognosis colon cancer is defined by a molecularly distinct subtype and develops from serrated precursor lesions. *Nat Med*. 2013;19(5):614-8. doi:10.1038/nm.3174.

358. Sadanandam A, Lyssiotis CA, Homicsko K, et al. A colorectal cancer classification system that associates cellular phenotype and responses to therapy. *Nat Med*. 2013;19(5):619-25. doi:10.1038/nm.3175.

359. Sadanandam A, Wang X, Melo FDSE, Gray JW, Vermeulen L, Hanahan D. Reconciliation of classification systems defining molecular subtypes of colorectal cancer. *Cell Cycle*. 2014;13(3):353-357.

360. Roepman P, Schlicker A, Tabernero J, et al. Colorectal cancer intrinsic subtypes predict chemotherapy benefit, deficient mismatch repair and epithelial-to-mesenchymal transition. *Int J Cancer*. 2014;134(3):552-62. doi:10.1002/ijc.28387.

361. Jorissen RN, Lipton L, Gibbs P, et al. DNA Copy-Number Alterations Underlie Gene Expression Differences between Microsatellite Stable and Unstable Colorectal Cancers. *Clin Cancer Res*. 2008;14(24):8061-8069. doi:10.1158/1078-0432.CCR-08-1431.

362. Budinska E, Popovici V, Tejpar S, et al. Gene expression patterns unveil a new level of molecular heterogeneity in colorectal cancer. *J Pathol*. 2013;231(1):63-76. doi:10.1002/path.4212.

363. Vaske CJ, Benz SC, Sanborn JZ, et al. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics*. 2010;26(12):i237-45. doi:10.1093/bioinformatics/btq182.

364. Smyth GK, Speed TP. Normalization of cDNA microarray data. *Methods*. 2003;31(4):265-273.

365. 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(6):882-3. doi:10.1093/bioinformatics/bts034.

366. MacDonald J. hugene10sttranscriptcluster.db: Affymetrix hugene10 annotation data (chip hugene10sttranscriptcluster).

367. Aguilera O, Fraga MF, Ballestar E, et al. Epigenetic inactivation of the Wnt antagonist DICKKOPF-1 (DKK-1) gene in human colorectal cancer. *Oncogene*. 2006;25(29):4116-21. doi:10.1038/sj.onc.1209439.

368. Matsui A, Yamaguchi T, Maekawa S, et al. DICKKOPF-4 and -2 genes are upregulated in human colorectal cancer. *Cancer Sci*. 2009;100(10):1923-30. doi:10.1111/j.1349-7006.2009.01272.x.

369. Lee B Bin, Lee EJ, Jung EH, et al. Aberrant methylation of APC, MGMT, RASSF2A, and Wif-1 genes in plasma as a biomarker for early detection of colorectal cancer. *Clin Cancer Res*. 2009;15(19):6185-91. doi:10.1158/1078-0432.CCR-09-0111.

370. Ying J, Li H, Yu J, et al. WNT5A exhibits tumor-suppressive activity through antagonizing the Wnt/beta-catenin signaling, and is frequently methylated in colorectal cancer. *Clin Cancer Res*. 2008;14(1):55-61. doi:10.1158/1078-0432.CCR-07-1644.

371. Zhang W, Glöckner SC, Guo M, et al. Epigenetic inactivation of the canonical Wnt antagonist SRY-box containing gene 17 in colorectal cancer. *Cancer Res*. 2008;68(8):2764-72. doi:10.1158/0008-5472.CAN-07-6349.

372. Dhir M, Montgomery E a, Glöckner SC, et al. Epigenetic regulation of WNT signaling pathway genes in inflammatory bowel disease (IBD) associated neoplasia. *J Gastrointest Surg*. 2008;12(10):1745-53. doi:10.1007/s11605-008-0633-5.

373. Kao A-P, Wang K-H, Long C-Y, et al. Interleukin-1β induces cyclooxygenase-2 expression and promotes the invasive ability of human mesenchymal stem cells derived from ovarian endometrioma. *Fertil Steril*. 2011;96(3):678-684.e1. doi:10.1016/j.fertnstert.2011.06.041.

374. Yoshino N, Ishihara S, Rumi MAK, et al. Interleukin-8 regulates expression of Reg protein in Helicobacter pylori-infected gastric mucosa. *Am J Gastroenterol*. 2005;100(10):2157-66. doi:10.1111/j.1572-0241.2005.41915.x.

375. Holla VR, Mann JR, Shi Q, DuBois RN. Prostaglandin E2 regulates the nuclear receptor NR4A2 in colorectal cancer. *J Biol Chem*. 2006;281(5):2676-82. doi:10.1074/jbc.M507752200.

376. Li J, Ma W, Wang P, Hurley PJ, Bunz F, Hwang PM. Polo-like kinase 2 activates an antioxidant pathway to promote the survival of cells with mitochondrial dysfunction. *Free Radic Biol Med*. 2014;73:270-277.

377. Waugh DJJ, Wilson C. The interleukin-8 pathway in cancer. *Clin Cancer Res*. 2008;14(21):6735-41. doi:10.1158/1078-0432.CCR-07-4843.

378. Yu Y, Gong R, Mu Y, et al. Hepatitis B virus induces a novel inflammation network involving three inflammatory factors, IL-29, IL-8, and cyclooxygenase-2. *J Immunol*. 2011;187(9):4844-60. doi:10.4049/jimmunol.1100998.

379. Costa H, Nascimento R, Sinclair J, Parkhouse RME. Human cytomegalovirus gene UL76 induces IL-8 expression through activation of the DNA damage response. *PLoS Pathog*. 2013;9(9):e1003609. doi:10.1371/journal.ppat.1003609.

380. Gursoy U, Kononen E, Uitto V. Stimulation of epithelial cell matrix metalloproteinase (MMP-2,-9, -13) and interleukin-8 secretion by fusobacteria. *Oral*. 2008;3(17):432-434.

381. Huang GT, Zhang HB, Dang HN, Haake SK. Differential regulation of cytokine genes in gingival epithelial cells challenged by Fusobacterium nucleatum and Porphyromonas gingivalis. *Microb Pathog*. 2004;37(6):303-12. doi:10.1016/j.micpath.2004.10.003.

382. Weersma RK, Crusius JBA, Roberts RL, et al. Association of FcgR2a, but not FcgR3a, with inflammatory bowel diseases across three Caucasian populations. *Inflamm Bowel Dis*. 2010;16(12):2080-9. doi:10.1002/ibd.21342.

383. Banks C, Bateman A, Payne R, Johnson P, Sheron N. Chemokine expression in IBD. Mucosal chemokine expression is unselectively increased in both ulcerative colitis and Crohn's disease. *J Pathol*. 2003;199(1):28-35. doi:10.1002/path.1245.

384. Van Beelen Granlund A, Østvik AE, Brenna Ø, Torp SH, Gustafsson BI, Sandvik AK. REG gene expression in inflamed and healthy colon mucosa explored by in situ hybridisation. *Cell Tissue Res*. 2013;352(3):639-46. doi:10.1007/s00441-013-1592-z.

385. Dionne S, Hiscott J, D'Agata I, Duhaime A, Seidman EG. Quantitative PCR analysis of TNF-alpha and IL-1 beta mRNA levels in pediatric IBD mucosal biopsies. *Dig Dis Sci*. 1997;42(7):1557-66.

386. McKaig BC, McWilliams D, Watson S a, Mahida YR. Expression and regulation of tissue inhibitor of metalloproteinase-1 and matrix metalloproteinases by intestinal myofibroblasts in inflammatory bowel disease. *Am J Pathol*. 2003;162(4):1355-60. doi:10.1016/S0002-9440(10)63931-4.

387. Lawrance IC, Fiocchi C, Chakravarti S. Ulcerative colitis and Crohn ' s disease : distinctive gene expression profiles and novel susceptibility candidate genes. *Hum Mol Genet*. 2001;10(5):445-456.

388. Pender SLF, Li CKF, Di Sabatino a, Sabatino a DI, MacDonald TT, Buckley MG. Role of macrophage metalloelastase in gut inflammation. *Ann N Y Acad Sci*. 2006;1072:386-8. doi:10.1196/annals.1326.019.

389. Han N-Y, Choi W, Park J-M, Kim EH, Lee H, Hahm K-B. Label-free quantification for discovering novel biomarkers in the diagnosis and assessment of disease activity in inflammatory bowel disease. *J Dig Dis*. 2013;14(4):166-74. doi:10.1111/1751-2980.12035.

390. Iozzo R V, Sanderson RD. Proteoglycans in cancer biology, tumour microenvironment and angiogenesis. *J Cell Mol Med*. 2011;15(5):1013-31. doi:10.1111/j.1582-4934.2010.01236.x.

391. Frey H, Schroeder N, Manon-Jensen T, Iozzo R V, Schaefer L. Biological interplay between proteoglycans and their innate immune receptors in inflammation. *FEBS J*. 2013;280(10):2165-79. doi:10.1111/febs.12145.

392. De Wit M, Belt EJT, Delis-van Diemen PM, et al. Lumican and versican are associated with good outcome in stage II and III colon cancer. *Ann Surg Oncol*. 2013;20 Suppl 3:S348-59. doi:10.1245/s10434-012-2441-0.

393. Wight TN. Versican: a versatile extracellular matrix proteoglycan in cell biology. *Curr Opin Cell Biol*. 2002;14(5):617-623. doi:10.1016/S0955-0674(02)00375-7.

394. Kamhi E, Joo EJ, Dordick JS, Linhardt RJ. Glycosaminoglycans in infectious disease. *Biol Rev Camb Philos Soc*. 2013;88(4):928-43. doi:10.1111/brv.12034.

395. Zucker S, Vacirca J. Role of matrix metalloproteinases (MMPs) in colorectal cancer. *Cancer Metastasis Rev*. 2004;23(1-2):101-17.

396. Decock J, Thirkettle S, Wagstaff L, Edwards DR. Matrix metalloproteinases: protective roles in cancer. *J Cell Mol Med*. 2011;15(6):1254-65. doi:10.1111/j.1582-4934.2011.01302.x.

397. Liu M, Guo S, Stiles JK. The emerging role of CXCL10 in cancer (Review). *Oncol Lett*. 2011;2(4):583-589. doi:10.3892/ol.2011.300.

398. Zipin-Roitman A, Meshel T, Sagi-Assif O, et al. CXCL10 promotes invasion-related properties in human colorectal carcinoma cells. *Cancer Res*. 2007;67(7):3396-405. doi:10.1158/0008-5472.CAN-06-3087.

399. Bao S, Ouyang G, Bai X, et al. Periostin potently promotes metastatic growth of colon cancer by augmenting cell survival via the Akt/PKB pathway. *Cancer Cell*. 2004;5(4):329-39. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15093540.

400. Li a., Dubey S, Varney ML, Dave BJ, Singh RK. IL-8 Directly Enhanced Endothelial Cell Survival, Proliferation, and Matrix Metalloproteinases Production and Regulated Angiogenesis. *J Immunol*. 2003;170(6):3369-3376. doi:10.4049/jimmunol.170.6.3369.

401. Ak S, Tunca B, Yilmazlar T, et al. Microsatellite instability status affects gene expression profiles in early onset colorectal cancer patients. *J Surg Res*. 2013;185(2):626-37. doi:10.1016/j.jss.2013.07.014.

402. Pilotte L, Larrieu P, Stroobant V, et al. Reversal of tumoral immune resistance by inhibition of tryptophan 2,3-dioxygenase. *Proc Natl Acad Sci U S A*. 2012;109(7):2497-502. doi:10.1073/pnas.1113873109.

403. Xia L, Mo P, Huang W, et al. The TNF-α/ROS/HIF-1-induced upregulation of FoxMI expression promotes HCC proliferation and resistance to apoptosis. *Carcinogenesis*. 2012;33(11):2250-9. doi:10.1093/carcin/bgs249.

404. Park S, Yoo EJ, Cho N, Kim N, Kang GH. Comparison of CpG island hypermethylation and repetitive DNA hypomethylation in premalignant stages of gastric cancer , stratified for Helicobacter pylori infection. *J Pathol*. 2009;219:410-416. doi:10.1002/path.

405. Yoshida Y, Wang IC, Yoder HM, Davidson NO, Costa RH. The forkhead box M1 transcription factor contributes to the development and growth of mouse colorectal cancer. *Gastroenterology*. 2007;132(4):1420-31. doi:10.1053/j.gastro.2007.01.036.

406. Raychaudhuri P, Park HJ. FoxM1: a master regulator of tumor metastasis. *Cancer Res*. 2011;71(13):4329-33. doi:10.1158/0008-5472.CAN-11-0640.

407. Park HJ, Carr JR, Wang Z, et al. FoxM1, a critical regulator of oxidative stress during oncogenesis. *EMBO J*. 2009;28(19):2908-18. doi:10.1038/emboj.2009.239.

408. Chu XY, Zhu ZM, Chen LB, et al. FOXM1 expression correlates with tumor invasion and a poor prognosis of colorectal cancer. *Acta Histochem*. 2012;114(8):755-62. doi:10.1016/j.acthis.2012.01.002.

409. Kaser A, Blumberg RS. Autophagy, microbial sensing, endoplasmic reticulum stress, and epithelial function in inflammatory bowel disease. *Gastroenterology*. 2011;140(6):1738-47. doi:10.1053/j.gastro.2011.02.048.

410. Kaser A, Lee AH, Franke A, et al. XBP1 links ER stress to intestinal inflammation and confers genetic risk for human inflammatory bowel disease. *Cell*. 2008;134(5):743-56. doi:10.1016/j.cell.2008.07.021.

411. Kuramochi J, Arai T, Ikeda S, Kamagai J, Uetake H, Sugihara K. High Pin1 Expression Is Associated With Tumor Progression in Colorectal Cancer. *J Surg Oncol*. 2006;94(January):155-160. doi:10.1002/jso.

412. Fagiani E, Christofori G. Angiopoietins in angiogenesis. *Cancer Lett*. 2013;328(1):18-26. doi:10.1016/j.canlet.2012.08.018.

413. Carcinoma C, Weirauch U, Beckmann N, et al. Functional Role and Therapeutic Potential of the Pim-1 Kinase. *Neoplasia*. 2013;15(7):783-794, IN28. doi:10.1593/neo.13172.

414. Dhanasekaran DN. JNK Signaling Network and Cancer. *Genes Cancer*. 2013;4(9-10):332-3. doi:10.1177/1947601913507949.

415. Lu R, Ji Z, Li X, et al. miR-145 functions as tumor suppressor and targets two oncogenes, ANGPT2 and NEDD9, in renal cell carcinoma. *J Cancer Res Clin Oncol*. 2014;140(3):387-97. doi:10.1007/s00432-013-1577-z.

416. Strickertsson JAB, Desler C, Martin-Bertelsen T, et al. Enterococcus faecalis infection causes inflammation, intracellular oxphos-independent ROS production, and DNA damage in human gastric cancer cells. *PLoS One*. 2013;8(4):e63147. doi:10.1371/journal.pone.0063147.

417. Wang X, Yang Y, Huycke MM. Commensal bacteria drive endogenous transformation and tumour stem cell marker expression through a bystander effect. *Gut*. 2014:1-10. doi:10.1136/gutjnl-2014-307213.

418. Roessner A, Kuester D, Malfertheiner P, Schneider-Stock R. Oxidative stress in ulcerative colitis-associated carcinogenesis. *Pathol Res Pract*. 2008;204(7):511-24. doi:10.1016/j.prp.2008.04.011.

419. Banerjea A, Ahmed S, Hands RE, et al. Colorectal cancers with microsatellite instability display mRNA expression signatures characteristic of increased immunogenicity. *Mol Cancer*. 2004;3:21. doi:10.1186/1476-4598-3-21.

420. Schwitalle Y, Kloor M, Eiermann S, et al. Immune response against frameshift-induced neopeptides in HNPCC patients and healthy HNPCC mutation carriers. *Gastroenterology*. 2008;134(4):988-97. doi:10.1053/j.gastro.2008.01.015.

421. Kloor M, Michel S, von Knebel Doeberitz M. Immune evasion of microsatellite unstable colorectal cancers. *Int J Cancer*. 2010;127(5):1001-10. doi:10.1002/ijc.25283.

422. Chan AT, Ogino S, Fuchs CS. Aspirin and the risk of colorectal cancer in relation to expression of COX-2. *N Engl J Med*. 2007;356(21):2131-2142.

423. Nishihara R, Lochhead P, Kuchiba A, et al. Aspirin use and risk of colorectal cancer according to BRAF mutation status. *JAMA*. 2013;309(24):2563-71. doi:10.1001/jama.2013.6599.

424. Lilley CE, Schwartz R a, Weitzman MD. Using or abusing: viruses and the cellular DNA damage response. *Trends Microbiol*. 2007;15(3):119-26. doi:10.1016/j.tim.2007.01.003.

425. Karpinski P, Myszka A, Ramsey D, Kielan W, Sasiadek MM. Detection of viral DNA sequences in sporadic colorectal cancers in relation to CpG island methylation and methylator phenotype. *Tumour Biol*. 2011;32(4):653-9. doi:10.1007/s13277-011-0165-6.

426. Limoli CL, Giedzinski E. Induction of Chromosomal Instability by Chronic Oxidative Stress. *Neoplasia*. 2003;5(4):339-346. doi:10.1016/S1476-5586(03)80027-1.

427. Rhodes JM, Campbell BJ. Inflammation and colorectal cancer: IBD-associated and sporadic cancer compared. *Trends Mol Med*. 2002;8(1):10-16.

428. Warren RL, Freeman DJ, Pleasance S, et al. Co-occurrence of anaerobic bacteria in colorectal carcinomas. *Microbiome*. 2013;1(1):16. doi:10.1186/2049-2618-1-16.

429. Allen-Vercoe E, Jobin C. Fusobacterium and Enterobacteriaceae: Important players for CRC? *Immunol Lett*. 2014:1-8. doi:10.1016/j.imlet.2014.05.014.

430. Kolenbrander PE, London J. Adhere Today, Here Tomorrow : Oral Bacterial Adherence. *J Bacteriol*. 1993;175(11):3247-3252.

431. Rickard AH, Gilbert P, High NJ, Kolenbrander PE, Handley PS. Bacterial coaggregation: an integral process in the development of multi-species biofilms. *Trends Microbiol*. 2003;11(2):94-100. doi:10.1016/S0966-842X(02)00034-3.

432. Rn A, Ganeshkurnar N, Pe K. Helicobacter pylor adherese selectively to Fusobacterium spp. *Oral Microbiol Immunol*. 1998;(13):51-54.

433. Johnson EM, Flannagan SE, Sedgley CM. Coaggregation Interactions Between Oral and Endodontic Enterococcus faecalis and Bacterial Species Isolated From Persistent Apical Periodontitis. *J Endod*. 2006;32(10):946-950.

434. Diaz PI, Zilm PS, Rogers a H. The response to oxidative stress of Fusobacterium nucleatum grown in continuous culture. *FEMS Microbiol Lett*. 2000;187(1):31-4.

435. Steeves CH, Potrykus J, Barnett DA, Bearne SL. Oxidative stress response in the opportunistic oral pathogen Fusobacterium nucleatum. *Proteomics*. 2011;11(10):2027-2037.

436. Silva VL, Diniz CG, Cara DC, et al. Enhanced pathogenicity of Fusobacterium nucleatum adapted to oxidative stress. *Microb Pathog*. 2005;39(4):131-8. doi:10.1016/j.micpath.2005.07.002.

437. Diaz PI, Zilm PS, Rogers AH. Fusobacterium nucleatum supports the growth of Porphyromonas gingivalis in oxygenated and carbon-dioxide-depleted environments. *Microbiology*. 2002;148(Pt 2):467-72.

438. Tailleux L, Waddell SJ, Pelizzola M, et al. Probing host pathogen cross-talk by transcriptional profiling of both Mycobacterium tuberculosis and infected human dendritic cells and macrophages. *PLoS One*. 2008;3(1):e1403. doi:10.1371/journal.pone.0001403.

439. Resnick MB, Sabo E, Meitner P a, et al. Global analysis of the human gastric epithelial transcriptome altered by Helicobacter pylori eradication in vivo. *Gut*. 2006;55(12):1717-24. doi:10.1136/gut.2006.095646.

440. Barnett MPG, McNabb WC, Cookson AL, et al. Changes in colon gene expression associated with increased colon inflammation in interleukin-10 gene-deficient mice inoculated with Enterococcus species. *BMC Immunol*. 2010;11:39. doi:10.1186/1471-2172-11-39.

441. Kayaoglu G, Orstavik D. Virulence Factors of Enterococcus Faecalis: Relationship To Endodontic Disease. *Crit Rev Oral Biol Med*. 2004;15(5):308-320. doi:10.1177/154411130401500506.

442. Nallapareddy SR, Wenxiang H, George M, Murray BE, Weinstock GM. Molecular Characterization of a Widespread, Pathogenic, and Antibiotic Resistance-Receptive Enterococcus faecalis Lineage and Dissemination of Its Putative Pathogenicity Island. *J Bacteriol*. 2005;187(16):5709-5718. doi:10.1128/JB.187.16.5709.

443. Gursoy UK, Könönen E, Uitto V-J. Intracellular replication of fusobacteria requires new actin filament formation of epithelial cells. *APMIS*. 2008;116(12):1063-70. doi:10.1111/j.1600-0463.2008.00868.x.

444. Hoffmann M, Kim SC, Sartor RB, Haller D. Enterococcus faecalis Strains Differentially Regulate Alix / AIP1 Protein Expression and ERK 1 / 2 Activation in Intestinal Epithelial Cells in the Context of Chronic Experimental Colitis. *J Proteome Res*. 2009;8:1183-1192.

445. Dufour JH, Dziejman M, Liu MT, Leung JH, Lane TE, Luster a. D. IFN- -Inducible Protein 10 (IP-10; CXCL10)-Deficient Mice Reveal a Role for IP-10 in Effector T Cell Generation and Trafficking. *J Immunol*. 2002;168(7):3195-3204. doi:10.4049/jimmunol.168.7.3195.

446. Tasheva ES, Klocke B, Conrad GW. Analysis of transcriptional regulation of the small leucine rich proteoglycans. *Mol Vis*. 2004;10(August):758-72.

447. Menozzi FD, Pethe K, Bifani P, Soncin F, Brennan MJ, Locht C. Enhanced bacterial virulence through exploitation of host glycosaminoglycans. *Mol Microbiol*. 2002;43(6):1379-86.

448. Baldassarri L, Bertuccini L, Creti R, et al. Glycosaminoglycans mediate invasion and survival of Enterococcus faecalis into macrophages. *J Infect Dis*. 2005;191(8):1253-62. doi:10.1086/428778.

449. Schmidtchen A, Frick IM, Björck L. Dermatan sulphate is released by proteinases of common pathogenic bacteria and inactivates antibacterial alpha-defensin. *Mol Microbiol*. 2001;39(3):708-13.

450. Sangiorgi E, Capecchi MR. Bmi1 is expressed in vivo in intestinal stem cells. *Nat Genet*. 2008;40(7):915-920.

451. Tateishi K, Ohta M, Kanai F, et al. Dysregulated expression of stem cell factor Bmi1 in precancerous lesions of the gastrointestinal tract. *Clin Cancer Res*. 2006;12(23):6960-6. doi:10.1158/1078-0432.CCR-06-0449.

452. Li D, Tang H, Fan J, et al. Expression level of Bmi-1 oncoprotein is associated with progression and prognosis in colon cancer. *J Cancer Res Clin Oncol*. 2010;136(7):997-1006. doi:10.1007/s00432-009-0745-7.

453. Pun JCS, Chan JYJ, Chun BKM, et al. Plasma Bmi1 mRNA as a potential prognostic biomarker for distant metastasis in colorectal cancer patients. *Mol Clin Oncol*. 2014;2(5):817-820. doi:10.3892/mco.2014.321.

454. Du J, Li Y, Li J, Zheng J. Polycomb group protein Bmi1 expression in colon cancers predicts the survival. *Med Oncol*. 2010;27(4):1273-6. doi:10.1007/s12032-009-9373-y.

455. Nakagawa R, Naka T, Tsutsui H, et al. SOCS-1 participates in negative regulation of LPS responses. *Immunity*. 2002;17(5):677-87.

456. Zídek Z, Farghali H, Kmoníčková E. Intrinsic nitric oxide-stimulatory activity of lipoteichoic acids from different Gram-positive bacteria. *Nitric Oxide*. 2010;23(4):300-10. doi:10.1016/j.niox.2010.09.001.

457. Rechreche H, Montalto G, Mallo G V, et al. pap, reg Ialpha and reg Ibeta mRNAs are concomitantly up-regulated during human colorectal carcinogenesis. *Int J Cancer*. 1999;81(5):688-94.

458. Cavard C, Terris B, Grimber G, et al. Overexpression of regenerating islet-derived 1 alpha and 3 alpha genes in human primary liver tumors with beta-catenin mutations. *Oncogene*. 2006;25(4):599-608. doi:10.1038/sj.onc.1208860.

459. Zenilman ME, Kim S, Levine BA, Lee C, Steinberg JJ. Ectopic expression of reg protein: A marker of colorectal mucosa at risk for neoplasia. *J Gastrointest Surg*. 1997;1(2):194-201.

460. Astrosini C, Roeefzaad C, Dai Y-Y, Dieckgraefe BK, Jöns T, Kemmner W. REG1A expression is a prognostic marker in colorectal cancer and associated with peritoneal carcinomatosis. *Int J Cancer*. 2008;123(2):409-13. doi:10.1002/ijc.23466.

Appendix A

Table 1: Participant-level characteristics table 1/2. FF = fresh-frozen; FFPE = formalin-fixed paraffin embedded; Tissue type: N=matched normal mucosa, T = tumour tissue; Ethnicity: MA = mixed ancestry, C = caucasian, B = black, I = indian; Gender: M = male, F = female; Stage: Dukes stage of tumour tissue.

Sample	Specimen Type	Tumour Type	Tissue type	Patient	Age	Ethnicity	RT	Gender	BMI	Stage	Location	Site
10N	FF	Sporadic	N	10	63	MA	N	M	24.7	NA	Distal	Descending colon
10T	FF	Sporadic	T	10	63	MA	N	M	24.7	III	Distal	Descending colon
11N	FF	Sporadic	N	11	84	C	N	M	28.7	NA	Proximal	Transverse colon
11T	FF	Sporadic	T	11	84	C	N	M	28.7	I	Proximal	Transverse colon
13N	FF	HNPCC	N	13	46	MA	N	F	22.3	NA	Proximal	Ceacum
13T	FF	HNPCC	T	13	46	MA	N	F	22.3	II	Proximal	Ceacum
14N	FF	Sporadic	N	14	80	MA	N	M	23.5	NA	Distal	Sigmoid colon
14T	FF	Sporadic	T	14	80	MA	N	M	23.5	IV	Distal	Sigmoid colon
15N	FF	Sporadic	N	15	74	I	N	M	24.2	NA	Distal	Rectum
15T	FF	Sporadic	T	15	74	I	N	M	24.2	II	Distal	Rectum
16N	FF	Sporadic	N	16	76	C	N	M	NA	NA	Proximal	Ascending colon
16T	FF	Sporadic	T	16	76	C	N	M	NA	I	Proximal	Ascending colon
17N	FF	Sporadic	N	17	79	C	N	M	22.7	NA	Distal	Sigmoid colon
17T	FF	Sporadic	T	17	79	C	N	M	22.7	III	Distal	Sigmoid colon
18N	FF	HNPCC	N	18	44	MA	N	F	NA	NA	Proximal	Transverse colon
18T	FF	HNPCC	T	18	44	MA	N	F	NA	I	Proximal	Transverse colon
1N	FF	HNPCC	N	1	58	MA	N	F	29.6	NA	Proximal	Transverse colon
1T	FF	HNPCC	T	1	58	MA	N	F	29.6	IV	Proximal	Transverse colon
20N	FF	HNPCC	N	20	NA	MA	N	F	NA	NA	Proximal	Ceacum

20T	FF	HNPCC	T	20	NA	MA	N	F	NA	NA	Proximal	Ceacum
23N	FF	Sporadic	N	23	70	MA	N	F	NA	NA	Distal	Splenic flexure
23T	FF	Sporadic	T	23	70	MA	N	F	NA	III	Distal	Splenic flexure
33N	FF	Sporadic	N	33	69	MA	N	F	28.7	NA	Distal	Rectum
33T	FF	Sporadic	T	33	69	MA	N	F	28.7	III	Distal	Rectum
34N	FF	Sporadic	N	34	70	MA	N	F	25.4	NA	Distal	Rectum
34T	FF	Sporadic	T	34	70	MA	N	F	25.4	III	Distal	Rectum
37N	FF	Sporadic	N	37	49	B	N	M	20	NA	Distal	Proximal descending colon
37T	FF	Sporadic	T	37	49	B	N	M	20	III	Distal	Proximal descending colon
3N	FF	Sporadic	N	3	70	MA	N	M	29.5	NA	Distal	RSJ
3T	FF	Sporadic	T	3	70	MA	N	M	29.5	III	Distal	RSJ
41N	FF	Sporadic	N	41	71	MA	N	F	31.6	NA	Distal	Rectum
41T	FF	Sporadic	T	41	71	MA	N	F	31.6	II	Distal	Rectum
44N	FF	Sporadic	N	44	36	B	N	M	18.2	NA	Proximal	Ceacum
44T	FF	Sporadic	T	44	36	B	N	M	18.2	III	Proximal	Ceacum
48N	FF	Sporadic	N	48	37	MA	N	M	30.7	NA	Distal	NA
48T	FF	Sporadic	T	48	37	MA	N	M	30.7	IV	Distal	NA
4N	FF	HNPCC	N	4	44	MA	N	F	26.7	NA	Proximal	Ceacum
4T	FF	HNPCC	T	4	44	MA	N	F	26.7	III	Proximal	Ceacum
55N	FF	Sporadic	N	55	64	MA	N	M	25.8	NA	Distal	NA
55T	FF	Sporadic	T	55	64	MA	N	M	25.8	III	Distal	NA
56N	FF	Sporadic	N	56	54	MA	N	F	26.9	NA	Distal	RSJ
56T	FF	Sporadic	T	56	54	MA	N	F	26.9	III	Distal	RSJ
60N	FF	Sporadic	N	60	65	MA	N	M	NA	NA	Distal	RSJ
60T	FF	Sporadic	T	60	65	MA	N	M	NA	II	Distal	RSJ

63N	FF	Sporadic	N	63	78	C	N	F	26.8	NA	Proximal	Hepatic flexure
63T	FF	Sporadic	T	63	78	C	N	F	26.8	II	Proximal	Hepatic flexure
8N	FF	Sporadic	N	8	61	C	N	M	25	NA	Distal	Rectum
8T	FF	Sporadic	T	8	61	C	N	M	25	III	Distal	Rectum
19N	FF	Sporadic	N	19	62	MA	Y	F	NA	NA	Distal	Rectum
19T	FF	Sporadic	T	19	62	MA	Y	F	NA	I	Distal	Rectum
21N	FF	Sporadic	N	21	61	MA	N	F	40.5	NA	Distal	Rectum
21T	FF	Sporadic	T	21	61	MA	N	F	40.5	NA	Distal	Rectum
24N	FF	Sporadic	N	24	42	MA	Y	M	NA	NA	Distal	Rectum
24T	FF	Sporadic	T	24	42	MA	Y	M	NA	III	Distal	Rectum
25N	FF	Sporadic	N	25	67	MA	Y	F	NA	NA	Distal	Rectum
25T	FF	Sporadic	T	25	67	MA	Y	F	NA	I	Distal	Rectum
26N	FF	Sporadic	N	26	73	C	N	F	NA	NA	Distal	Rectum
26T	FF	Sporadic	T	26	73	C	N	F	NA	I	Distal	Rectum
28N	FF	Sporadic	N	28	79	I	Y	M	NA	NA	Distal	Rectum
28T	FF	Sporadic	T	28	79	I	Y	M	NA	II	Distal	Rectum
29N	FF	Sporadic	N	29	73	B	Y	M	25.8	NA	Distal	Rectum
29T	FF	Sporadic	T	29	73	B	Y	M	25.8	III	Distal	Rectum
2N	FF	Sporadic	N	2	67	MA	Y	M	NA	NA	Distal	Rectum
2T	FF	Sporadic	T	2	67	MA	Y	M	NA	II	Distal	Rectum
30N	FF	Sporadic	N	30	25	B	Y	M	NA	NA	Distal	Rectum
30T	FF	Sporadic	T	30	25	B	Y	M	NA	II	Distal	Rectum
35N	FF	Sporadic	N	35	68	MA	Y	M	NA	NA	Distal	Rectum
35T	FF	Sporadic	T	35	68	MA	Y	M	NA	II	Distal	Rectum
39N	FF	Sporadic	N	39	51	MA	Y	F	36.3	NA	Distal	Rectum
39T	FF	Sporadic	T	39	51	MA	Y	F	36.3	II	Distal	Rectum

45N	FF	Sporadic	N	45	36	B	Y	F	NA	NA	Distal	Rectum
45T	FF	Sporadic	T	45	36	B	Y	F	NA	II	Distal	Rectum
47N	FF	HNPCC	N	47	25	MA	Y	M	NA	NA	Distal	Descending colon
47T	FF	HNPCC	T	47	25	MA	Y	M	NA	III	Distal	Descending colon
54N	FF	Sporadic	N	54	60	MA	N	M	NA	NA	NA	NA
54T	FF	Sporadic	T	54	60	MA	N	M	NA	NA	NA	NA
58N	FF	Sporadic	N	58	71	MA	Y	NA	NA	NA	Distal	Rectum
58T	FF	Sporadic	T	58	71	MA	Y	NA	NA	NA	Distal	Rectum
61N	FF	Sporadic	N	61	61	MA	N	F	NA	NA	Distal	Sigmoid colon
61T	FF	Sporadic	T	61	61	MA	N	F	NA	NA	Distal	Sigmoid colon
62N	FF	Sporadic	N	62	40	MA	Y	F	NA	NA	Distal	Rectum
62T	FF	Sporadic	T	62	40	MA	Y	F	NA	II	Distal	Rectum
65N	FF	Sporadic	N	65	73	C	Y	M	NA	NA	Distal	Rectum
65T	FF	Sporadic	T	65	73	C	Y	M	NA	III	Distal	Rectum
66N	FF	Sporadic	N	66	59	MA	Y	F	NA	NA	Distal	Rectum
66T	FF	Sporadic	T	66	59	MA	Y	F	NA	III	Distal	Rectum
67N	FF	Sporadic	N	67	64	MA	N	M	23.8	NA	Distal	Rectum
67T	FF	Sporadic	T	67	64	MA	N	M	23.8	III	Distal	Rectum
69N	FF	Sporadic	N	69	NA	NA	NA	F	NA	NA	NA	NA
69T	FF	Sporadic	T	69	NA	NA	NA	F	NA	NA	NA	NA
6N	FF	Sporadic	N	6	23	B	Y	F	NA	NA	Distal	Rectum
6T	FF	Sporadic	T	6	23	B	Y	F	NA	III	Distal	Rectum
7N	FF	Sporadic	N	7	52	MA	N	M	27.9	NA	Distal	Rectum
7T	FF	Sporadic	T	7	52	MA	N	M	27.9	III	Distal	Rectum
22N	FF	Sporadic	N	22	49	MA	Y	M	NA	NA	Distal	Rectum
22T	FF	Sporadic	T	22	49	MA	Y	M	NA	III	Distal	Rectum

40N	FF	Sporadic	N	40	47	MA	Y	F	NA	NA	Distal	Rectum
40T	FF	Sporadic	T	40	47	MA	Y	F	NA	II	Distal	Rectum
43N	FF	Sporadic	N	43	67	MA	Y	F	NA	NA	Distal	Rectum
43T	FF	Sporadic	T	43	67	MA	Y	F	NA	II	Distal	Rectum
51N	FF	Sporadic	N	51	47	MA	Y	M	NA	NA	Distal	Rectum
51T	FF	Sporadic	T	51	47	MA	Y	M	NA	II	Distal	Rectum
57N	FF	Sporadic	N	57	45	MA	N	M	25.7	NA	Distal	RSJ
57T	FF	Sporadic	T	57	45	MA	N	M	25.7	III	Distal	RSJ
59N	FF	Sporadic	N	59	79	C	Y	F	NA	NA	NA	NA
59T	FF	Sporadic	T	59	79	C	Y	F	NA	II	NA	NA
64N	FF	Sporadic	N	64	69	MA	Y	F	NA	NA	Distal	Rectum
64T	FF	Sporadic	T	64	69	MA	Y	F	NA	II	Distal	Rectum
31N	FF	Sporadic	N	31	54	MA	Y	F	NA	NA	Distal	Rectum
31T	FF	Sporadic	T	31	54	MA	Y	F	NA	II	Distal	Rectum
71T	FFPE	Sporadic	T	71	41	NA	N	F	NA	II	Proximal	Hepatic flexure
71N	FFPE	Sporadic	N	71	41	NA	N	F	NA	NA	Proximal	Hepatic flexure
72T	FFPE	Sporadic	T	72	43	B	N	M	NA	III	Proximal	Ceacum
72N	FFPE	Sporadic	N	72	43	B	N	M	NA	NA	Proximal	Ceacum
73T	FFPE	Sporadic	T	73	30	NA	N	F	NA	III	Proximal	Ceacum
73N	FFPE	Sporadic	N	73	30	NA	N	F	NA	NA	Proximal	Ceacum
74T	FFPE	Sporadic	T	74	41	NA	Y	F	NA	IV	Distal	Rectum
74N	FFPE	Sporadic	N	74	41	NA	Y	F	NA	NA	Distal	Rectum
75T	FFPE	Sporadic	T	75	NA	NA	NA	NA	NA	NA	NA	NA
76T	FFPE	Sporadic	T	76	41	B	N	M	NA	II	Proximal	Ceacum
76N	FFPE	Sporadic	N	76	41	B	N	M	NA	NA	Proximal	Ceacum
77T	FFPE	Sporadic	T	77	NA	NA	NA	NA	NA	NA	NA	NA

78T	FFPE	Sporadic	T	78	NA	NA	NA	NA	NA	NA	NA	NA
79T	FFPE	HNPCC	T	79	NA	NA	NA	NA	NA	NA	NA	NA
80T	FFPE	Sporadic	T	80	NA	NA	NA	NA	NA	NA	NA	NA
81T	FFPE	Sporadic	T	81	NA	NA	NA	NA	NA	III	NA	NA
82T	FFPE	Sporadic	T	82	NA	NA	NA	NA	NA	NA	NA	NA
83T	FFPE	Sporadic	T	83	NA	NA	NA	NA	NA	NA	NA	NA
84T	FFPE	Sporadic	T	84	NA	NA	NA	NA	NA	NA	NA	NA
85T	FFPE	Sporadic	T	85	35	NA	N	M	NA	III	Distal	Sigmoid colon
85N	FFPE	Sporadic	N	85	35	NA	N	M	NA	NA	Distal	Sigmoid colon
86T	FFPE	Sporadic	T	86	NA	NA	NA	NA	NA	NA	NA	NA
87T	FFPE	Sporadic	T	87	NA	NA	NA	NA	NA	NA	NA	NA
88T	FFPE	HNPCC	T	88	NA	NA	NA	NA	NA	NA	NA	NA

Table 2: Participant characteristics, table 2/2. Bacterial quantitation data expressed as bacteria/50ng human DNA; EPEC limit of detection (LOD) & *Fusobacterium* LOD: for FFPE tissue these are the estimated LOD's based on normalisation against COX1; MSI method: PCR = Bethesda panel of markers; MLH1 meth. = MLH1 methylation testing by methylation-specific PCR; MMR prot. = MMR protein(s) with known methylation or absence of staining by immunohistochemistry (IHC) of MLH1, MSH2 and MSH6; FB = *Fusobacterium*; EF = *E. faecalis*. NA no data for that entry available, whereas 0 means that 0 bacterial copies were detected for the relevant species noted

Sample	**MSI**	**MSI method**	**MLH1 meth.**	**MMR prot.**	**Inflammation noted (pathology report)**	**ETBF**	**EPEC**	**EPEC LOD**	**EF**	**FB**	**FB LOD**	**afaC+ *E. coli***	**ClB+ *E. coli***
10N	MSS	PCR	NA	NA	N	12	0	10	NA	8	2	2177	0
10T	MSS	PCR	NA	NA	N	139	0	10	0	1	2	170	0
11N	MSS	PCR	N	NA	N	0	0	10	0	0	2	0	0

11T	MSS	PCR	N	NA	N	0	0	10	0	0	2	0	0
13N	MSI-H	PCR	N	NA	Y	374	0	10	0	65	2	3787	0
13T	MSI-H	PCR	N	MLH1	Y	3186	0	10	1	3273	2	19200	0
14N	MSS	PCR	N	NA	N	3423	0	10	331	67	2	261	0
14T	MSS	PCR	N	NA	N	987	0	10	151	271	2	729	0
15N	MSS	PCR	N	NA	N	2	0	10	0	9	2	4787	1172
15T	MSS	PCR	N	NA	N	0	0	10	0	17	2	9887	1707
16N	MSS	PCR	N	NA	Y	0	0	10	0	1953	2	319000	1780
16T	MSS	PCR	N	NA	Y	0	0	10	0	21	2	2043	0
17N	MSS	PCR	N	NA	N	130	0	10	0	48	2	15	0
17T	MSS	PCR	N	NA	N	486	0	10	0	833	2	81	0
18N	MSI-H	PCR	N	NA	NA	0	0	10	5	377	2	1937	0
18T	MSI-H	PCR	N	NA	NA	0	0	10	3	2610	2	7593	994
1N	MSS	PCR	N	NA	N	4	0	10	0	12	2	12	0
1T	MSS	PCR	N	NA	N	2	0	10	0	276	2	47	0
20N	MSI-H	PCR	N	NA	NA	3	0	10	NA	0	2	0	0
20T	MSI-H	PCR	N	NA	NA	13	0	10	NA	2	2	0	0
23N	MSS	PCR	N	NA	Y	5	0	10	1	26	2	0	0
23T	MSS	PCR	N	NA	Y	3	0	10	250	4820	2	0	0
33N	MSS	PCR	N	NA	N	0	0	10	0	3	2	0	677
33T	MSS	PCR	N	NA	N	0	0	10	0	14	2	0	622

34N	MSS	PCR	N	NA	N	0	52	10	0	25	2	30900	1263
34T	MSS	PCR	N	NA	N	3	63	10	1	289	2	33433	1400
37N	MSS	PCR	N	NA	Y	1620	0	10	NA	1897	2	0	3693
37T	MSS	PCR	N	NA	Y	35300	0	10	0	4773	2	0	1800
3N	MSS	PCR	N	NA	N	0	0	10	0	1	2	0	0
3T	MSS	PCR	N	NA	N	0	0	10	0	3	2	0	0
41N	MSS	PCR	N	NA	N	0	0	10	NA	69	2	0	0
41T	MSS	PCR	N	NA	N	0	0	10	NA	213	2	2	0
44N	MSI-H	PCR	N	NA	Y	0	3037	10	0	1110	2	1197	36
44T	MSI-H	PCR	Y	NA	Y	0	1111	10	0	68700	2	156	0
48N	MSS	PCR	N	NA	N	2328	0	10	0	1	2	0	0
48T	MSS	PCR	N	NA	N	1106	0	10	0	44	2	0	0
4N	MSI-H	PCR	N	NA	N	0	0	10	0	20	2	195	0
4T	MSI-H	PCR	N	NA	N	3	0	10	0	60767	2	6537	0
55N	MSI-L	PCR	N	NA	N	4	0	10	0	99	2	0	5777
55T	MSI-L	PCR	N	NA	N	3	0	10	0	9	2	0	536
56N	MSS	PCR	N	NA	N	698	0	10	0	46	2	0	0
56T	MSS	PCR	N	NA	N	648	0	10	0	378	2	0	0
60N	MSI-H	PCR	N	NA	N	20	0	10	0	3	2	2750	78
60T	MSI-H	PCR	N	NA	N	0	0	10	0	22	2	5237	69
63N	MSI-H	PCR	N	NA	N	0	0	10	NA	73	2	0	0

63T	MSI-H	PCR	Y	NA	N	0	13	10	4	2730	2	0	0
8N	MSS	PCR	N	NA	N	12	0	10	4	0	2	592	0
8T	MSS	PCR	N	NA	N	14	0	10	2	37	2	619	0
19N	NA	NA	NA	NA	NA	0	0	10	0	0	2	0	0
19T	NA	NA	NA	NA	NA	0	0	10	0	0	2	0	0
21N	MSS	PCR	NA	NA	NA	0	0	10	NA	0	2	0	0
21T	MSS	PCR	N	NA	NA	0	0	10	NA	0	2	1	0
24N	NA	NA	NA	NA	NA	0	0	10	0	21	2	0	0
24T	NA	NA	NA	NA	NA	0	0	10	NA	57	2	0	0
25N	NA	NA	NA	NA	NA	0	0	10	0	8	2	0	0
25T	NA	NA	NA	NA	NA	0	0	10	0	0	2	0	0
26N	MSI-L	PCR	N	NA	N	0	0	10	4	333	2	0	34
26T	MSI-L	PCR	N	NA	N	0	0	10	NA	40	2	0	0
28N	NA	NA	NA	NA	NA	0	0	10	0	1	2	0	0
28T	NA	NA	NA	NA	NA	0	0	10	0	0	2	0	0
29N	MSS	PCR	NA	NA	N	0	14	10	0	69	2	0	0
29T	MSS	PCR	NA	NA	N	0	31	10	0	6678	2	19	0
2N	NA	NA	NA	NA	NA	0	0	10	0	71	2	0	0
2T	NA	NA	NA	NA	NA	0	0	10	0	1026	2	0	0
30N	NA	NA	NA	NA	NA	0	0	10	4	1543	2	0	22800
30T	NA	NA	NA	NA	NA	0	0	10	0	106	2	0	0
35N	NA	NA	NA	NA	NA	0	0	10	0	422	2	0	0
35T	NA	NA	NA	NA	NA	0	0	10	1	1353	2	0	0

39N	NA	NA	NA	NA	N	0	0	10	NA	11	2	0	0
39T	NA	NA	NA	NA	N	0	0	10	NA	49	2	0	0
45N	NA	NA	NA	NA	NA	0	0	10	0	0	2	0	0
45T	NA	NA	NA	NA	NA	0	51	10	0	5	2	0	0
47N	NA	NA	NA	NA	Y	5	0	10	0	187	2	0	0
47T	NA	NA	NA	NA	Y	0	0	10	0	125	2	0	0
54N	MSS	PCR	N	NA	NA	0	0	10	0	0	2	0	516
54T	MSS	PCR	N	NA	NA	0	0	10	0	8	2	0	316
58N	NA	NA	NA	NA	NA	0	0	10	0	1	2	2	0
58T	NA	NA	NA	NA	NA	0	0	10	0	0	2	2	0
61N	MSI-L	PCR	N	NA	Y	0	0	10	0	36	2	25	0
61T	MSI-L	PCR	N	NA	Y	0	0	10	0	14	2	20	60
62N	NA	NA	NA	NA	NA	0	0	10	NA	5	2	0	0
62T	NA	NA	NA	NA	NA	0	0	10	NA	78	2	0	0
65N	NA	NA	NA	NA	NA	0	0	10	8	340	2	10	3617
65T	NA	NA	NA	NA	NA	0	0	10	131	1353	2	0	3497
66N	NA	NA	N	NA	NA	0	0	10	0	38	2	0	0
66T	NA	NA	N	NA	NA	0	0	10	1	8990	2	0	0
67N	MSS	PCR	N	NA	N	0	0	10	NA	217	2	0	0
67T	MSS	PCR	N	NA	N	0	0	10	12	1483	2	0	0
69N	NA	NA	NA	NA	NA	0	0	10	0	0	2	29	0
69T	NA	NA	NA	NA	NA	0	0	10	0	240	2	23	0
6N	MSS	PCR	NA	NA	NA	0	0	10	0	273	2	0	0
6T	MSS	PCR	NA	NA	NA	0	0	10	0	2417	2	0	0
7N	MSS	PCR	N	NA	N	0	0	10	0	1397	2	0	44
7T	MSS	PCR	N	NA	N	0	0	10	0	8497	2	0	62
22N	NA	NA	NA	NA	NA	0	0	10	NA	4	2	0	0

22T	NA	NA	NA	NA	NA	0	41	10	NA	5740	2	0	7373
40N	NA	NA	NA	NA	NA	0	0	10	NA	5	2	0	0
40T	NA	NA	NA	NA	NA	0	0	10	NA	3	2	0	0
43N	NA	NA	NA	NA	NA	NA	0	10	NA	4	2	0	0
43T	NA	NA	NA	NA	NA	0	0	10	NA	9	2	0	0
51N	NA	NA	NA	NA	NA	0	0	10	NA	NA	2	0	0
51T	NA	NA	NA	NA	NA	0	0	10	NA	9	2	0	0
57N	NA	NA	NA	NA	N	NA	NA	10	NA	2	2	NA	NA
57T	NA	NA	NA	NA	N	NA	NA	10	NA	NA	2	NA	NA
59N	NA	NA	NA	NA	NA	0	0	10	NA	28	2	0	0
59T	NA	NA	NA	NA	NA	0	0	10	NA	47	2	0	0
64N	NA	NA	NA	NA	NA	0	0	10	NA	8	2	0	0
64T	NA	NA	NA	NA	NA	0	0	10	NA	10	2	0	0
31N	NA	NA	NA	NA	NA	0	0	10	NA	0	2	0	0
31T	NA	NA	NA	NA	NA	0	0	10	NA	4	2	NA	0
71T	MSI-H	IHC	NA	MSH6, MSH2	Y	NA	NA	36	NA	107	164	NA	NA
71N	MSI-H	IHC	NA		Y	NA	NA	56	NA	NA	252	NA	NA
72T	MSI-H	IHC	NA	MSH2	N	NA	NA	13	NA	260	59	NA	NA
72N	MSI-H	IHC	NA		N	NA	NA	35	NA	4091	156	NA	NA
73T	MSI-H	IHC	NA	MSH6	N	NA	NA	19	NA	NA	87	NA	NA
73N	MSI-H	IHC	NA		N	NA	NA	18	NA	NA	79	NA	NA
74T	MSI-H	IHC	NA	MSH6	NA	NA	NA	3	NA	NA	14	NA	NA
74N	MSI-H	IHC	NA		NA	NA	NA	9	NA	NA	39	NA	NA
75T	MSI-H	IHC	NA	MSH6	NA	NA	NA	86	NA	526	385	NA	NA
76T	MSI-H	IHC	NA	MLH1	N	NA	NA	42	NA	124	188	NA	NA

76N	MSI-H	IHC	NA		N	NA	NA	61	NA	NA	276	NA	NA
77T	MSI-H	IHC	NA	MLH1	NA	NA	NA	77	NA	NA	346	NA	NA
78T	MSI-H	IHC	NA	MLH1	NA	NA	NA	5	NA	27291	24	NA	NA
79T	MSI-H	IHC	NA	MLH1	NA	NA	NA	55	NA	1888	248	NA	NA
80T	MSI-H	IHC	NA	MLH1	NA	NA	NA	457	NA	1587	2056	NA	NA
81T	MSI-H	IHC	NA	MLH1	NA	NA	NA	82	NA	221	367	NA	NA
82T	MSI-H	IHC	NA	MLH1, MSH2	NA	NA	NA	2968	NA	NA	13356	NA	NA
83T	MSI-H	IHC	NA	MSH2, MSH6	NA	NA	NA	242	NA	NA	1088	NA	NA
84T	MSI-H	IHC	NA	MSH2	NA	NA	NA	26	NA	262	117	NA	NA
85T	MSI-H	IHC	NA	MSH2	Y	NA	NA	43	NA	6642	194	NA	NA
85N	MSI-H	IHC	NA	NA	Y	NA	NA	91	NA	22818	410	NA	NA
86T	MSI-H	IHC	NA	MSH2, MLH1	NA	NA	NA	74	NA	NA	331	NA	NA
87T	MSS	IHC	NA	NA	Y	NA	NA	21	NA	NA	94	NA	NA
88T	MSI-H	IHC	NA	MLH1	NA	NA	NA	25	NA	342	112	NA	NA

Table 3: qPCR conditions used.

Bacteria	Reagent	qPCR conditions	Number of cycles
ETBF		2 min @ 95°C	
	SensiFAST SYBR No Rox Mix (1x)	1 s @ 95°C	x45
	F (200 nM final)	5 s @ 65-55°C; decrease 0.5 °C/cycle	
	R (200 nM final)	acquire 1 s @ 80°C	
	template: 50 ng human DNA		
	H_20		
	total = 20 µl		
E. faecalis		10 min @ 95°C	
	Maxima SYBR qPCR Master Mix (1x)	15 s @ 95°C	x50
	F (900 nM final)	60 s @ 60°C	
	R (900 nM final)	acquire @ 72°C	
	template: 50 ng human DNA		
	H_20		
	total = 25 µl		
S. gallolyticus		2 min @ 95°C	
	SensiFAST SYBR No Rox Mix (1x)	5 s @ 95°C	x60
	F (200 nM final)	10 s @ 60-50°C, decrease 0.5 °C/cycle	
	R (200 nM final)	acquire @ 80°C	
	template: 50 ng human DNA		
	H_20		
	total = 20 µl		
EPEC (*bfpA*)		2 min @ 95°C	
	SensiFAST SYBR No Rox Mix (1x)	5 s @ 95°C	x60
	F (200 nM final)	7 s @ 70-64°C; decrease 0.5 °C/cycle	
	R (200 nM final)	acquire @ 80°C	
	template: 50 ng human DNA		
	H_20		
	total = 20 µl		
EPEC (*eae*)	PCR buffer	2 min @ 95°C	
	SensiFAST SYBR No Rox Mix (1x)	5 s @ 95°C	x55
	F (200 nM final)	7 s @ 70-60°C; decrease 0.5 °C/cycle	

	R (200 nM final)	acquire @ 80°C	
	template: 50 ng human DNA		
	H_20		
	total = 20 µl		
EHEC (*stx1*)		2 min @ 95°C	
	SensiFAST SYBR No Rox Mix (1x)	5 s @ 95°C	x60
	F (600 nM final)	10 s @ 70-60°C; decrease 0.5 °C/cycle	
	R (600 nM final)	acquire @ 80°C	
	template: 50 ng human DNA		
	H_20		
	total = 20 µl		
EHEC (*stx2*)		1 min @ 95°C	
	SensiFAST SYBR No Rox Mix (1x)	1 s @ 95°C	x55
	F (400 nM final)	5 s @ 70-60°C; decrease 0.5 °C/cycle	
	R (400 nM final)	acquire @ 80°C	
	template: 50 ng human DNA		
	H_20		
	total = 20 µl		
Fusobacterium spp.		10 min @ 95	
	Maxima SYBR qPCR Master Mix (1x)	15 s @ 95°C	x50
	F (300 nM final)	45 s @ 60°C	
	R (300 nM final)	acquire @ 72°C	
	template: 50 ng human DNA		
	H_20		
	total = 25 µl		
AIEC (*afaC*)		10 min @ 95	
	Maxima SYBR qPCR Master Mix (1x)	15 s @ 95°C	x50
	F (900 nM final)	60 s @ 65-60 decrease 0.5 °C/cycle.	
	R (900 nM final)	acquire at 72°C	
	template: 50 ng human DNA		
	H_20		
	total = 25 µl		
AIEC (*ClB*)		10 min @ 95	
	Maxima SYBR qPCR Master Mix (1x)	15 s @ 95°C	x60
	F (600 nM final)	60 s @ 60°C	
	R (600 nM final)	acquire at 72°C	
	template: 50 ng human DNA		
	H_20		
	total = 25 µl		

Appendix B

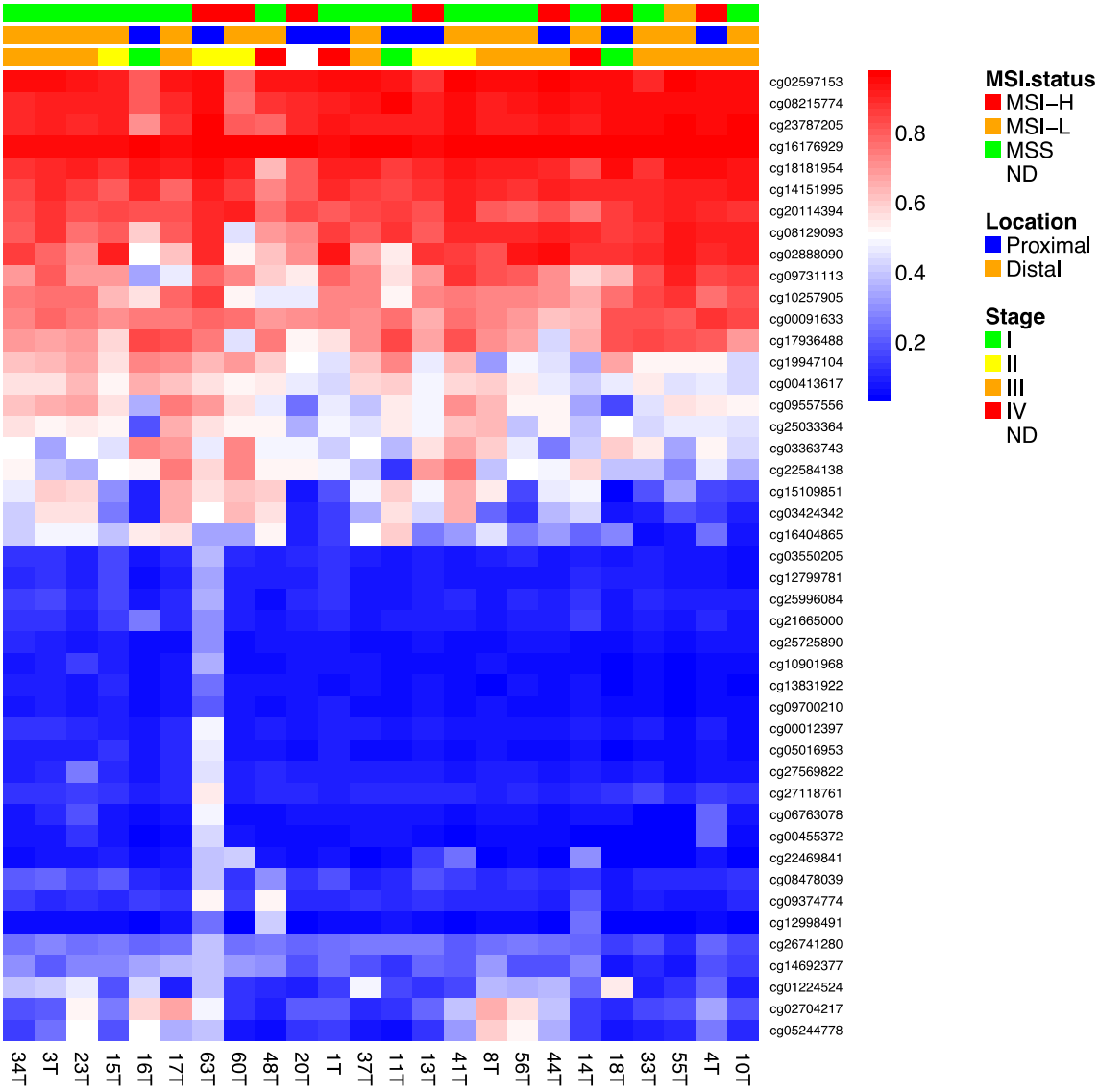


Figure 1: Predicting **CIMP-H** using an array-based marker panel. RPMM-based clustering of probes mapping to CpG islands in the Hinoue CIMP-H marker panel (FAM78A, FSTL1, KCNC1, MYOCD, and SLC6A4). Sample 63T is considered to be CIMP-H. The scale on the right represents beta values (0–1)

Table 1: Region-specific detail of the 35 genes that are differentially methylated and differentially expressed in EPEC+ CRCs. CpG islands = 1; non-CpG islands = 0; EPEC+/- median: median beta values in each group; EPEC+ vs. EPEC- diff.: difference in median beta values between groups; Target: Refseq gene target and region(s) to which each probe map.

Gene ID	Gene Symbol	Chromosome	Chromosome start	Chromosome end	CpG island	P value	FDR	EPEC- median	EPEC+ median	EPEC + vs. EPEC - diff.	Target
cg02128087	C12orf68	12	48577366	48579709	1	3.10E-02	0.87	0.07	0.31	0.24	NM_001013635:TSS1500
cg03272292	C12orf68	12	48577366	48579709	1	1.60E-02	0.87	0.07	0.3	0.23	NM_001013635:TSS200
cg05376611	C12orf68	12	48577366	48579709	1	5.50E-03	0.71	0.1	0.33	0.23	NM_001013635:TSS1500
cg11201894	C6orf223	6	43968337	43973694	0	4.00E-02	0.87	0.7	0.43	-0.27	NM_153246:3'UTR
cg14525935	C7orf50	7	-1036623	-1177893	1	1.90E-02	0.87	0.11	0.45	0.33	NM_001134395:Body; NM_032350:Body; NM_001134396:Body
cg07266431	CDK6	7	-92234235; -92234235	-92465941; -92463231	0	4.90E-03	0.7	0.64	0.39	-0.25	NM_001259:Body; NM_001145306:Body
cg25032595	CLDN10	13	96085853; 96204947	96232010; 96232010	1	1.20E-02	0.84	0.6	0.38	-0.22	NM_006984:5'UTR; NM_006984:1stExon; NM_182848:Body; NM_001160100:Body
cg27318087	CPPED1	16	-12753656	-12897744	0	3.50E-03	0.65	0.43	0.68	0.25	NM_001099455:Body; NM_018340:Body

cg23830540	CRMP1	4	-5822491; -5822491	-5894785; -5890315	0	3.60E-03	0.65	0.43	0.64	0.21	NM_001014809:Body; NM_001313:Body
cg17011964	EFNB2	13	-107142079	-107187388	0	4.40E-02	0.87	0.61	0.85	0.23	NM_004093:3'UTR
cg04403917	EPDR1	7	37960163; 37960921	37991542; 37991542	1	1.60E-02	0.87	0.12	0.46	0.35	NM_017549:Body
cg04499325	EPDR1	7	37960163; 37960921	37991542; 37991542	1	2.00E-02	0.87	0.08	0.55	0.46	NM_017549:1stExon
cg08608193	EPDR1	7	37960163; 37960921	37991542; 37991542	1	4.40E-03	0.68	0.07	0.4	0.32	NM_017549:Body
cg10876076	EPDR1	7	37960163; 37960921	37991542; 37991542	1	1.20E-02	0.84	0.09	0.56	0.48	NM_017549:Body
cg19468504	EPDR1	7	37960163; 37960921	37991542; 37991542	1	9.50E-03	0.79	0.12	0.57	0.46	NM_017549:1stExon
cg01915994	EPHB1	3	134514099	134979307	1	2.40E-02	0.87	0.11	0.38	0.27	NM_004441:1stExon; NM_004441:5'UTR
cg02934930	EPHB1	3	134514099	134979307	1	6.90E-06	0.02	0.09	0.48	0.4	NM_004441:TSS1500
cg04891921	EPHB1	3	134514099	134979307	1	2.60E-04	0.18	0.1	0.4	0.31	NM_004441:TSS200
cg17769793	FAM13A	4	-89647105; -89647105	-89978346; -89744512	0	7.10E-04	0.32	0.72	0.5	-0.21	NM_014883:Body
cg02098413	HACE1	6	-105175968	-105307794	0	2.20E-02	0.87	0.58	0.85	0.27	NM_020771:TSS1500
cg04892643	HDAC9	7	18126572; 18535369; 18535369; 18535369; 18535885; 18535885; 18548900	18708466; 18708465; 18708466; 19036992; 18993939; 19036992; 18708466	0	3.40E-02	0.87	0.48	0.79	0.32	NM_178425:Body; NM_178423:Body; NM_058176:Body
cg24761195	LRP11	6	-150139894	-150185480	1	4.80E-02	0.87	0.06	0.81	0.75	NM_032832:Body
cg02096492	MCMDC2	8	67782984; 67782984; 67783737	67814014; 67834283; 67817599	1	3.30E-02	0.87	0.09	0.32	0.23	NM_001136160:TSS1500; NM_0011361

											61:TSS200; NM_173518:TSS200
cg26030713	MCMDC2	8	67782984; 67782984; 67783737	67814014; 67834283; 67817599	0	5.30E-03	0.71	0.11	0.44	0.32	NM_001136160:TSS1500; NM_001136161:TSS200; NM_173518:TSS200
cg23076591	MN1	22	-28144265	-28197486	1	4.30E-02	0.87	0.08	0.33	0.25	NM_002430:TSS200
cg25803927	MN1	22	-28144265	-28197486	1	3.40E-06	0.01	0.06	0.44	0.38	NM_002430:TSS1500
cg20612002	MYEF2	15	-48431629	-48470558	1	2.10E-02	0.87	0.09	0.35	0.26	NM_016132:Body
cg09358973	NAV1	1	201617450; 201708963	201796102; 201796102	1	1.10E-02	0.81	0.28	0.5	0.21	NM_020443:1stExon
cg27441486	NAV1	1	201617450; 201708963	201796102; 201796102	0	3.10E-02	0.87	0.48	0.74	0.26	NM_001167738:TSS1500; NM_020443:Body
cg24085946	NFASC	1	204797782; 204797782; 204913354	204946274; 204991950; 204991950	1	4.40E-03	0.68	0.16	0.41	0.25	NM_001160333:1stExon; NM_001005388:5'UTR; NM_001005388:1stExon; NM_001160332:1stExon; NM_015090:1stExon; NM_001160332:5'UTR; NM_015090:5'UTR; NM_001160333:5'UTR; NM_001005389:1stExon;

											NM_001005389:5'UTR
cg00936907	NOG	17	54671060	54672951	1	6.10E-03	0.73	0.08	0.35	0.27	NM_005450:1stExon
cg12537619	OSBPL5	11	-3108346	-3186582	0	3.10E-02	0.87	0.58	0.8	0.22	NM_001144063:TSS1500; NM_020896:TSS1500; NM_145638:TSS1500
cg22629987	QPCT	2	37571753	37600465	0	3.50E-03	0.65	0.2	0.46	0.26	NM_012413:TSS200
cg02230017	RBFOX1	16	6069132; 6823810; 7382751	7763340; 7763340; 7763340	1	2.40E-02	0.87	0.14	0.35	0.22	NM_018723:TSS200; NM_001142333:TSS200
cg19378133	RBFOX1	16	6069132; 6823810; 7382751	7763340; 7763340; 7763340	0	6.40E-04	0.3	0.21	0.58	0.37	NM_018723:TSS1500; NM_001142333:TSS1500
cg09987129	RPS6KA2	6	-166822854; -166822854	-167275771; -167040726	1	1.50E-02	0.87	0.1	0.49	0.39	NM_001006932:TSS1500
cg20206437	RPS6KA2	6	-166822854; -166822854	-167275771; -167040726	1	2.00E-02	0.87	0.11	0.48	0.37	NM_001006932:TSS1500
cg20515377	RPS6KA2	6	-166822854; -166822854	-167275771; -167040726	0	2.00E-03	0.53	0.18	0.43	0.25	NM_001006932:TSS1500
cg13364903	SCUBE3	6	35182190	35218609	1	1.10E-05	0.03	0.11	0.34	0.24	NM_152753:TSS1500
cg21604042	SCUBE3	6	35182190	35218609	1	1.30E-02	0.84	0.21	0.42	0.21	NM_152753:TSS1500
cg27301032	SOX1	13	112721913	112726020	1	2.40E-02	0.87	0.11	0.42	0.31	NM_005986:TSS1500
cg12346504	SPAG16	2	214149103;	214182709; 215275225	0	2.60E-02	0.87	0.05	0.28	0.23	NM_001025436:TSS1500;

			214149103								NM_024532:TSS1500
cg22674613	SPATS1	6	44310397	44344904	1	2.00E-02	0.87	0.17	0.4	0.23	NM_145026:TSS200
cg00522893	TLX2	2	74741596	74744275	1	2.00E-02	0.87	0.13	0.53	0.4	NM_016170:5'UTR; NM_016170:1stExon
cg12964144	WSCD1	17	5973934	6027747	1	2.20E-04	0.16	0.1	0.4	0.29	NM_015253:5'UTR
cg14097019	ZNF287	17	-16453631	-16472520	1	3.20E-02	0.87	0.08	0.32	0.24	NM_020653:TSS200
cg12501868	ZNF419	19	57999079	58006048	1	5.90E-03	0.72	0.08	0.47	0.39	NM_001098493:TSS200; NM_001098496:TSS200; NM_024691:TSS200; NM_001098495:TSS200; NM_001098491:TSS200; NM_001098492:TSS200; NM_001098494:TSS200
cg15745619	ZNF419	19	57999079	58006048	1	1.20E-02	0.84	0.1	0.42	0.32	NM_001098493:TSS200; NM_001098496:TSS200; NM_024691:TSS200; NM_001098495:TSS200; NM_001098491:TSS200; NM_001098492:TSS200; NM_001098494:TSS200

cg23190994	ZNF546	19	40502943	40523514	1	3.50E-02	0.87	0.13	0.46	0.32	NM_178544:5'UTR
cg08397818	ZNF606	19	-58488441	-58514714	1	1.60E-03	0.49	0.17	0.51	0.33	NM_025027:5'UTR
cg04386405	ZNF83	19	-53115618; -53115618	-53193886; -53141644	1	3.70E-05	0.05	0.18	0.49	0.32	NR_003936:Body; NM_018300:5'UTR; NM_001105553:5'UTR; NM_001105550:5'UTR; NM_001105554:5'UTR; NM_001105549:5'UTR; NM_001105552:5'UTR; NM_001105551:5'UTR

Appendix C

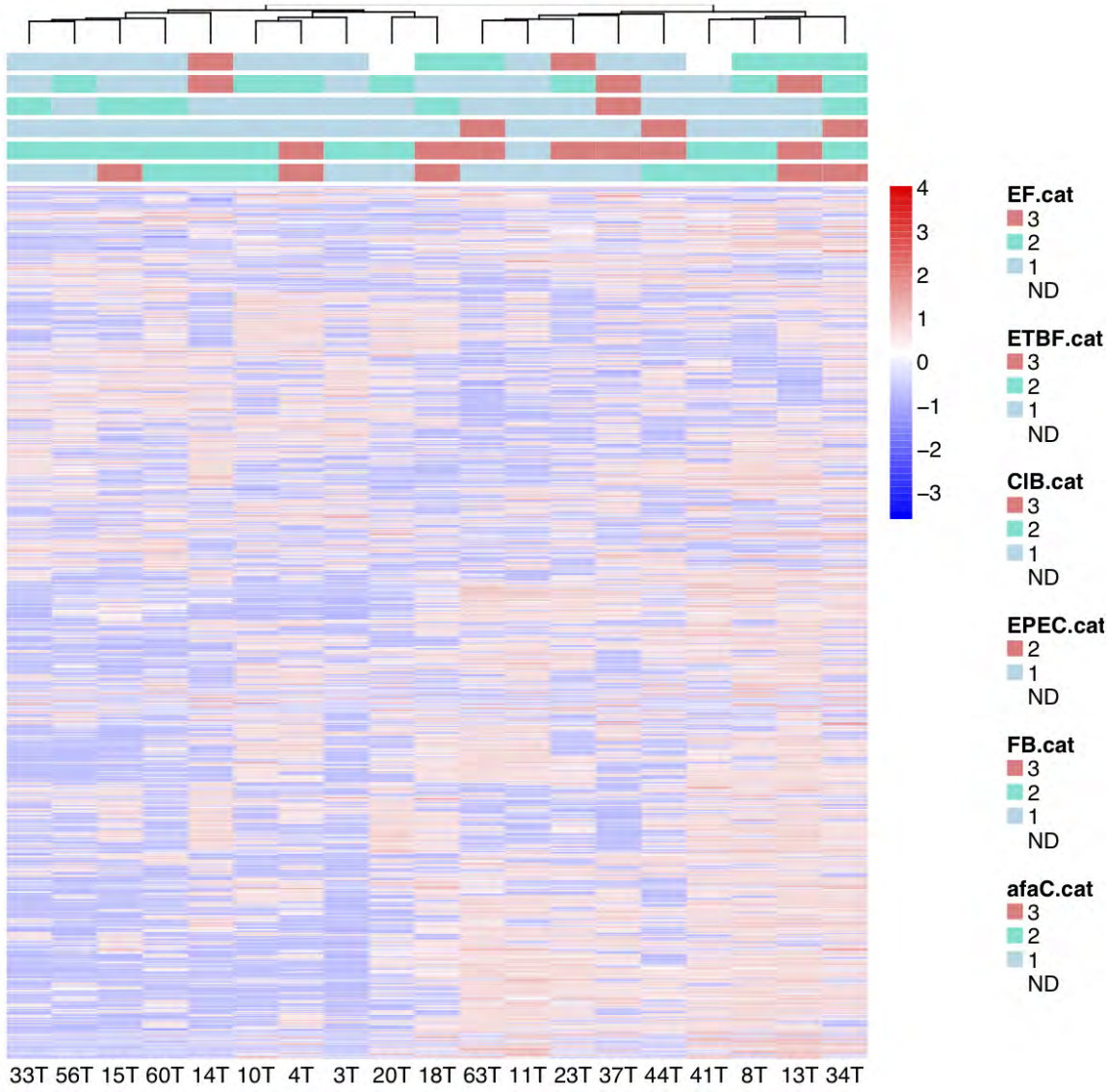


Figure 1: Hierarchical clustering of the 5334 most variable PARADIGM IPLs. Two main clusters can be distinguished that are identical to the RPMM gene-expression clusters except for 18T. The scale on the right represents row-scaled expression values.

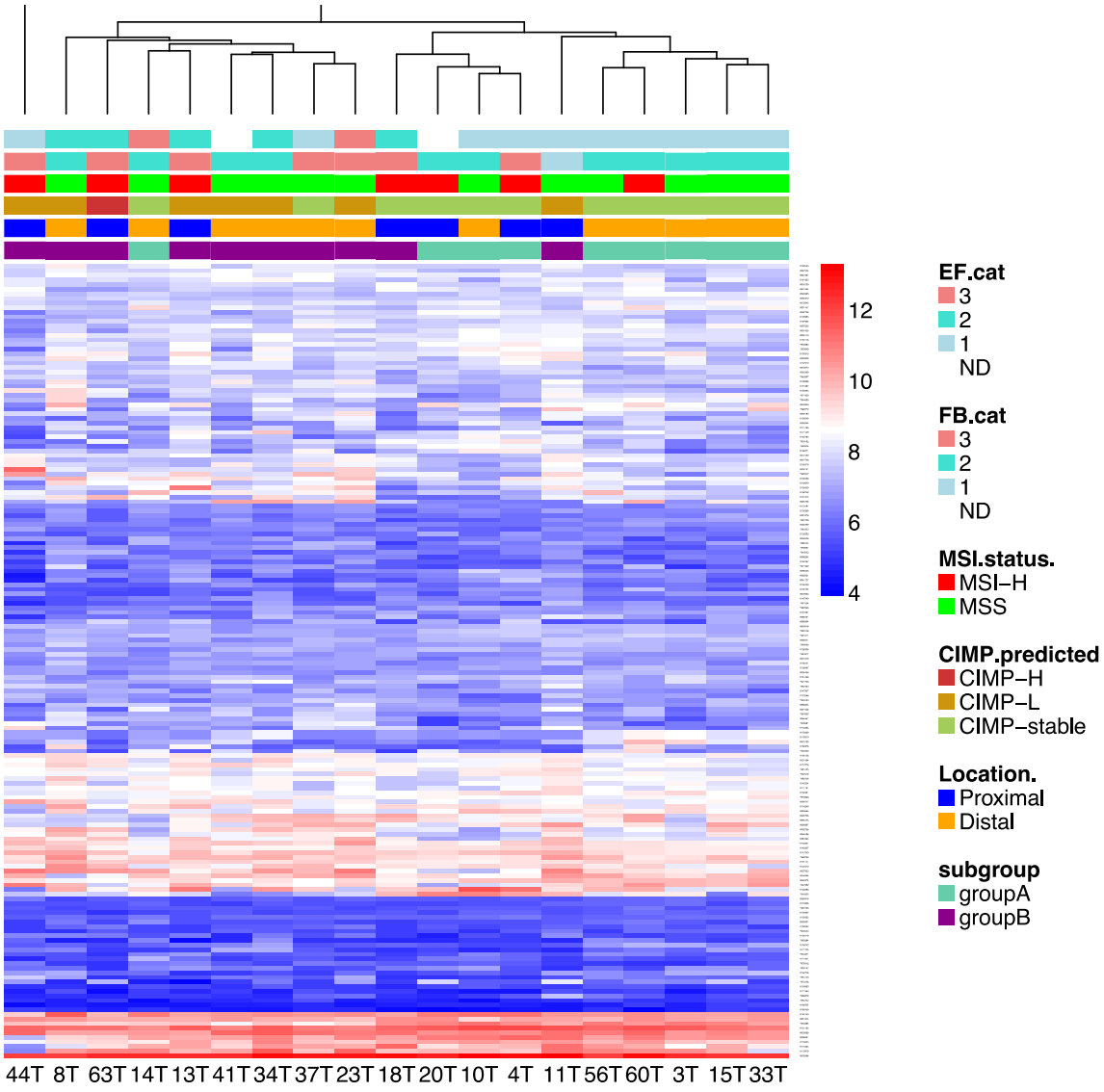


Figure 2: RPMM clustering of gene expression of Wnt pathway antagonists known to be methylated in CRC. The scale on the right represents log2 expression values.

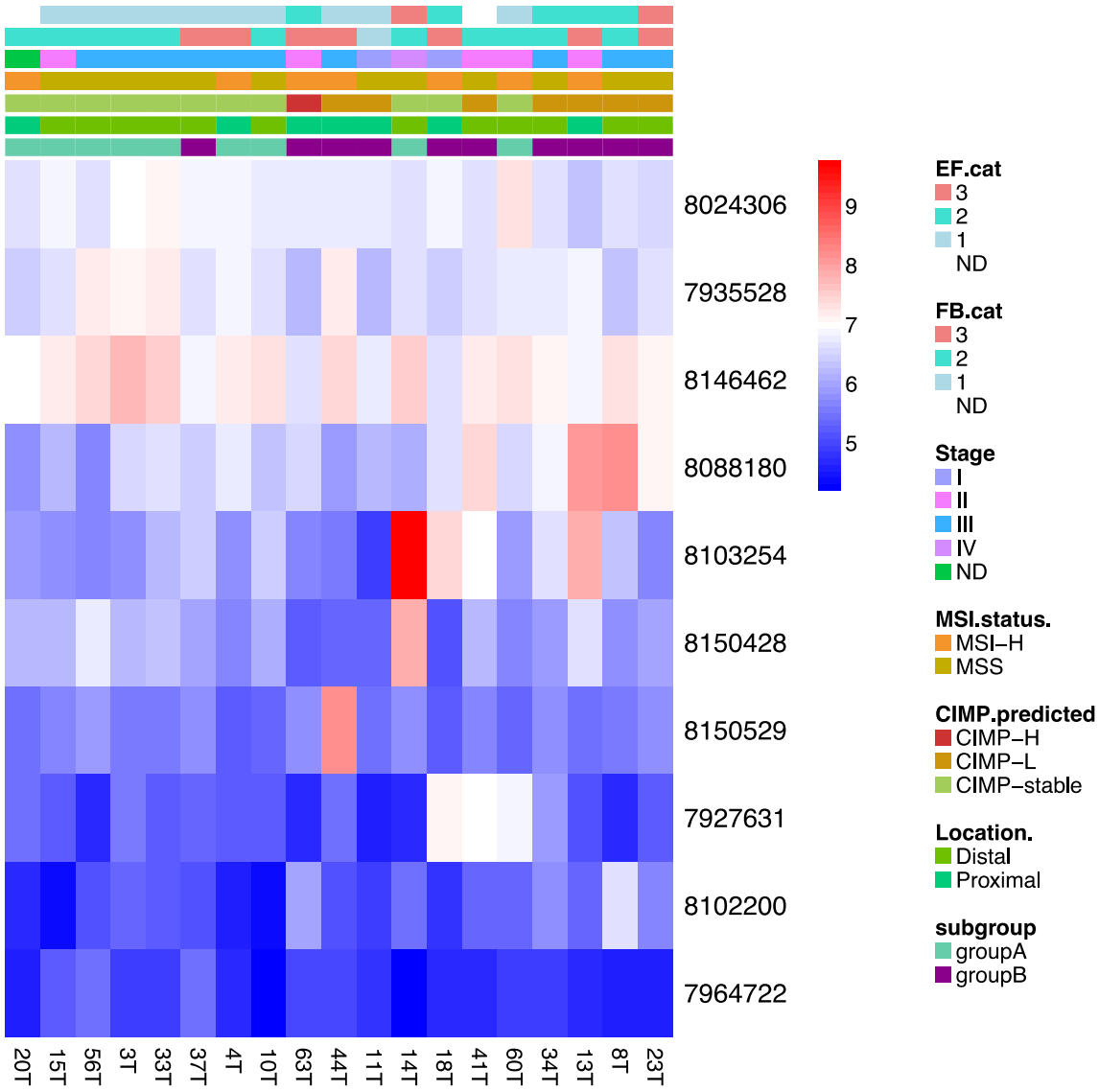


Figure 3: CRC classification for genes from the CRCassigner-786 classifier associated with the transit-amplifying subtype according to Sadanandam et al. The scale on the right represents log2 expression values.

Table 1: The 100 most significant PARADIGM genes or gene families of the 712 differentially activated between group A and B CRCs in our cohort. Genes have been sorted by the difference in activation scores between group B and A.

ID	P-value	adj. P-value	Group B median	Group A median	Group B vs. group A
FOXM1	1.7E-06	1.8E-04	3.92	-5.06	8.97
MIR17_(miRNA)	1.8E-08	5.2E-06	3.70	-2.32	6.02

MIR9-3_(miRNA)	1.8E-08	5.2E-06	3.70	-2.32	6.02
MIR141_(miRNA)	1.8E-08	5.2E-06	3.70	-2.32	6.02
MIR338_(miRNA)	1.8E-08	5.2E-06	3.70	-2.32	6.02
MIR429_(miRNA)	1.8E-08	5.2E-06	3.70	-2.32	6.02
MIR200A_(miRNA)	1.8E-08	5.2E-06	3.70	-2.32	6.02
MYC	8.1E-07	1.1E-04	2.68	-1.16	3.84
MAX	1.6E-06	1.8E-04	2.37	-1.25	3.62
ATF6	2.6E-07	4.9E-05	1.83	-1.73	3.56
CCNB2	3.1E-10	2.3E-07	0.56	-1.16	1.72
H2AFZ	4.1E-06	3.7E-04	0.54	-1.18	1.72
SSR1	3.8E-06	3.4E-04	0.48	-1.23	1.72
CALR	2.1E-06	2.2E-04	0.41	-1.27	1.68
HSP90B1	4.3E-06	3.8E-04	0.37	-1.30	1.67
HELLS	3.9E-07	7.2E-05	0.49	-1.15	1.64
CDK1	2.2E-16	1.9E-12	0.51	-1.11	1.63
E2F1_targets_(family)	3.5E-06	3.2E-04	0.49	-1.13	1.62
DUSP10	2.6E-08	6.2E-06	0.48	-1.14	1.62
TOP2A	4.2E-12	3.8E-09	0.49	-1.11	1.60
ECT2	5.3E-15	1.8E-11	0.43	-1.15	1.58
ACTL6A	2.4E-06	2.4E-04	0.36	-1.14	1.50
ME1	2.1E-08	5.3E-06	0.36	-1.13	1.49
EBNA1BP2	4.3E-06	3.8E-04	0.34	-1.15	1.49
FANCI	2.2E-13	2.4E-10	0.39	-1.10	1.48
YWHAE	4.5E-06	3.9E-04	0.33	-1.15	1.48
VBP1	4.6E-14	1.3E-10	0.39	-1.08	1.48
HNRNPC	6.6E-14	1.6E-10	0.40	-1.07	1.48
RPL6	4.5E-06	3.9E-04	0.33	-1.15	1.48
POT1	1.9E-13	2.4E-10	0.39	-1.09	1.48
HSP90AB1	2.1E-08	5.3E-06	0.37	-1.11	1.47
NUF2	2.0E-08	5.3E-06	0.37	-1.10	1.47
RPS15A	9.7E-09	4.0E-06	0.35	-1.12	1.47
SUZ12	2.1E-08	5.3E-06	0.36	-1.11	1.47
MND1	2.6E-06	2.5E-04	0.34	-1.13	1.47
PSMB6	4.5E-06	3.9E-04	0.33	-1.14	1.47
TUBGCP4	4.5E-06	3.9E-04	0.32	-1.14	1.46
ZNF484	9.7E-09	4.0E-06	0.35	-1.12	1.46
PSMD14	1.8E-06	1.9E-04	0.32	-1.15	1.46
PSMB5	2.7E-06	2.6E-04	0.33	-1.13	1.46
SUMO1	2.1E-08	5.3E-06	0.37	-1.10	1.46
C1QBP	2.1E-08	5.3E-06	0.35	-1.11	1.46

PSMD1	2.1E-08	5.3E-06	0.35	-1.11	1.46
CHAF1B	2.8E-06	2.6E-04	0.33	-1.13	1.46
EEF2	2.1E-08	5.3E-06	0.35	-1.10	1.46
CENPN	2.1E-08	5.3E-06	0.36	-1.10	1.46
BUB1	9.9E-18	1.7E-13	0.37	-1.08	1.45
MED21	1.5E-06	1.7E-04	0.36	-1.10	1.45
KIF20A	3.3E-15	1.4E-11	0.39	-1.06	1.45
NCAPG	2.9E-15	1.4E-11	0.39	-1.06	1.45
SSB	2.8E-06	2.7E-04	0.32	-1.13	1.45
SKA1	3.9E-09	2.0E-06	0.34	-1.10	1.45
NEDD1	9.2E-09	4.0E-06	0.36	-1.09	1.45
LAMB1	2.8E-06	2.6E-04	0.32	-1.12	1.44
CENPQ	9.5E-09	4.0E-06	0.35	-1.09	1.44
HNRNPK	1.6E-06	1.8E-04	0.33	-1.11	1.43
NDC80	1.6E-06	1.8E-04	0.32	-1.11	1.43
CENPE	7.5E-06	5.7E-04	0.38	-1.04	1.43
NDUFAB1	1.7E-06	1.8E-04	0.32	-1.11	1.43
HIST1H1B	6.1E-06	5.1E-04	0.38	-1.04	1.43
CCT2	1.6E-06	1.8E-04	0.33	-1.10	1.43
COPS2	7.5E-06	5.7E-04	0.38	-1.04	1.43
KIF15	1.7E-06	1.9E-04	0.32	-1.11	1.42
NUP43	9.5E-09	4.0E-06	0.35	-1.08	1.42
KIF11	3.6E-09	2.0E-06	0.35	-1.07	1.42
KIF18A	3.7E-09	2.0E-06	0.35	-1.07	1.42
SMC2	6.2E-06	5.1E-04	0.38	-1.03	1.42
PARP2	5.1E-07	8.1E-05	0.35	-1.07	1.41
MED17	5.3E-07	8.3E-05	0.35	-1.07	1.41
NUDT21	3.8E-09	2.0E-06	0.34	-1.07	1.41
XPO1	8.3E-07	1.1E-04	0.32	-1.08	1.41
OIP5	8.6E-07	1.2E-04	0.32	-1.09	1.41
STT3A	8.7E-07	1.2E-04	0.32	-1.09	1.41
MSH2	3.3E-06	3.1E-04	0.36	-1.05	1.40
SNUPN	4.9E-07	8.1E-05	0.34	-1.06	1.40
PLK4	9.0E-07	1.2E-04	0.32	-1.08	1.40
GAPDH	6.6E-06	5.3E-04	0.19	-1.21	1.39
CCT6A	6.7E-06	5.3E-04	0.17	-1.17	1.35
UBE2R2	6.7E-06	5.3E-04	0.17	-1.17	1.35
EXOC6	6.7E-06	5.3E-04	0.16	-1.17	1.34
CKAP5	2.3E-06	2.4E-04	0.19	-1.13	1.31
UBE2T	2.1E-08	5.3E-06	0.35	-0.96	1.31

RMI1	2.3E-06	2.4E-04	0.18	-1.12	1.30
SLC11A2	7.0E-08	1.6E-05	0.13	-1.15	1.27
GDI2	4.1E-06	3.7E-04	0.00	-1.18	1.18
CHEK1	3.5E-07	6.4E-05	0.15	-0.36	0.50
CHK1-2_(family)	6.0E-06	5.0E-04	0.15	-0.32	0.47
NEB	6.6E-06	5.3E-04	-1.17	0.00	-1.17
TLX2	6.5E-06	5.3E-04	-1.18	0.00	-1.18
AMN	1.3E-06	1.5E-04	-1.10	0.35	-1.45
LOC440461	2.4E-06	2.4E-04	-1.12	0.36	-1.49
MIR26A2_(miRNA)	1.8E-08	5.2E-06	-3.70	2.32	-6.02
DLEU2_(rna)	1.8E-08	5.2E-06	-3.70	2.32	-6.02
MIR26B_(miRNA)	1.8E-08	5.2E-06	-3.70	2.32	-6.02
DLEU1_(rna)	1.8E-08	5.2E-06	-3.70	2.32	-6.02
MIR23B_(miRNA)	1.8E-08	5.2E-06	-3.70	2.32	-6.02
MIR26A1_(miRNA)	1.8E-08	5.2E-06	-3.70	2.32	-6.02
MIRLET7G_(miRNA)	1.8E-08	5.2E-06	-3.70	2.32	-6.02
MIR22_(miRNA)	1.8E-08	5.2E-06	-3.70	2.32	-6.02
MIR146A_(miRNA)	1.8E-08	5.2E-06	-3.70	2.32	-6.02

Table 2: The 100 most significant PARADIGM complexes of the 711 differentially activated between group A and B CRCs in our cohort. Complexes have been sorted by the difference in activation scores between group B and A.

ID	P-value	FDR	Group B median	Group A median	Group B vs. group A
MYC/Max_(complex)	6.0E-09	2.9E-06	7.06	-6.34	13.40
MYC/Max/HDAC3_(complex)	9.7E-07	1.3E-04	2.01	-1.11	3.12
Myc/Max_heterodimer_(complex)	3.1E-09	1.8E-06	1.60	-1.43	3.02
MYC/Max/RPL11_(complex)	5.8E-07	8.8E-05	1.77	-0.67	2.45
MYC/Max/DNA_replication_preinitiation_complex_(complex)	8.2E-08	1.8E-05	1.60	-0.81	2.42
MYC/Max/HBP1_(complex)	6.5E-06	5.3E-04	1.73	-0.68	2.40
ATF6-alpha/BiP_(complex)	6.8E-06	5.4E-04	0.70	-1.66	2.37
MYC/Max/P-TEFb_(complex)	3.2E-07	6.1E-05	1.54	-0.78	2.33
MYC/Max/NF-Y_(complex)	3.0E-06	2.8E-04	1.44	-0.89	2.33
MYC/Max/PML4_(complex)	8.7E-08	1.9E-05	1.54	-0.66	2.20
FANCD2/FANCI_(complex)	1.2E-12	1.1E-09	0.50	-1.58	2.08
Cyclin_B2/phospho-Cdc2(Thr_14_Thr_161)_(complex)	4.9E-11	3.9E-08	0.71	-1.36	2.06
Cyclin_B2/phospho-Cdc2(Thr_161)_(complex)	4.9E-11	3.9E-08	0.71	-1.36	2.06

Cyclin_A2/phospho-Cdc2(Thr_161)_(complex)	7.5E-09	3.5E-06	0.61	-1.42	2.02
Cyclin_B/phospho-Cdc2(Thr_14)_(complex)	2.6E-09	1.6E-06	0.63	-1.39	2.02
Cyclin_B/Cdk1_complex_(complex)	2.6E-09	1.6E-06	0.63	-1.39	2.02
Cyclin_B/Cdc2_complex_(complex)	2.6E-09	1.6E-06	0.63	-1.39	2.02
Cyclin_A/CDK1-2_(complex)	1.9E-06	2.0E-04	0.50	-1.44	1.94
calnexin/calreticulin_(complex)	4.8E-07	8.1E-05	0.51	-1.38	1.88
BUB1/BUB3_(complex)	6.3E-08	1.4E-05	0.49	-1.37	1.85
nuclear_Cyclin_B1/Cdc2_complexes_(family)	4.1E-07	7.4E-05	0.67	-1.16	1.83
cytoplasmic_Cyclin_B1/Cdc2_complexes_(family)	4.1E-07	7.4E-05	0.67	-1.16	1.83
EIF2A/14-3-3_E_(complex)	6.1E-07	9.2E-05	0.40	-1.43	1.82
MYC/Max/p14ARF_(complex)	1.0E-07	2.2E-05	1.44	-0.39	1.82
MAD2*CDC20_complex_(complex)	3.4E-06	3.1E-04	0.48	-1.33	1.81
Centralspindlin_(complex)	1.2E-06	1.4E-04	0.39	-1.41	1.81
active_nuclear_Cyclin_B1/Cdc2_complexes_(family)	4.5E-07	7.9E-05	0.66	-1.14	1.80
RAD51/BRCA2_(complex)	1.1E-06	1.3E-04	0.49	-1.28	1.77
RAD51/BRCA2_complex_(complex)	1.1E-06	1.3E-04	0.49	-1.28	1.77
BRCA2/Rad51_(complex)	1.1E-06	1.3E-04	0.49	-1.28	1.77
Cyclin_B1/CDK1_(complex)	2.2E-06	2.3E-04	0.35	-1.35	1.70
Cyclin_B1/phospho-Cdc2_(_Thr_14)_(complex)	2.4E-06	2.4E-04	0.36	-1.34	1.70
phospho-Cyclin_B1/phospho-Cdc2(Thr_161)_(complex)	2.4E-06	2.4E-04	0.36	-1.34	1.70
nuclear_Cyclin_B1/phospho-Cdc2_(_Thr_14)_complexes_(complex)	2.4E-06	2.4E-04	0.36	-1.34	1.70
phospho-Cyclin_B1(CRS)/phospho-Cdc2_(Thr_161)_(complex)	2.4E-06	2.4E-04	0.36	-1.34	1.70
phospho-cyclin_B1(CRS)/phosph-Cdc2(Thr_161)_(complex)	2.4E-06	2.4E-04	0.36	-1.34	1.70
Cyclin_B1/phospho-Cdc2(Thr_161)_(complex)	2.4E-06	2.4E-04	0.36	-1.34	1.70
Phospho-Cyclin_B1_(CRS)/phospho-Cdc2(Thr_161)_(complex)	2.4E-06	2.4E-04	0.36	-1.34	1.70
Cyclin_B1/phospho-Cdc2_(Thr_14_Thr_161)_(complex)	2.4E-06	2.4E-04	0.36	-1.34	1.70
ORC_complex_bound_to_origin_(complex)	2.8E-08	6.5E-06	0.65	-0.99	1.64
Cyclin_B1/phospho-Cdc2(Thr_161_Thr_14_Tyr_15)_(complex)	2.7E-06	2.6E-04	0.36	-1.22	1.59
SKP2/p27Kip1_(complex)	4.6E-06	3.9E-04	0.60	-0.92	1.52
Centromeric_Chromatin/CENPH-I_Complex_(complex)	2.0E-07	3.9E-05	0.82	-0.69	1.51
unfolded_protein/(Glc)1_(GlcNAc)2_(	5.0E-07	8.1E-05	0.47	-1.03	1.50

Man)9_(Asn)1/chaperone_(complex)					
Laminin-1_(EHS_laminin)_(complex)	5.9E-06	5.0E-04	0.44	-1.05	1.49
Laminin_1_(complex)	5.9E-06	5.0E-04	0.44	-1.05	1.49
Cyclin_A/Cdc2_(complex)	2.0E-13	2.4E-10	0.46	-1.01	1.47
Cyclin_A/phospho-Cdc2(Thr_14)_(complex)	2.0E-13	2.4E-10	0.46	-1.01	1.47
Cyclin_A/phospho-Cdc2(Thr_161)_complex_(complex)	2.0E-13	2.4E-10	0.46	-1.01	1.47
Cyclin_A/phospho-Cdc2(Thr_161)_(complex)	2.0E-13	2.4E-10	0.46	-1.01	1.47
phospho-Cdc2(Thr_14)/Cyclin_A_(complex)	2.0E-13	2.4E-10	0.46	-1.01	1.47
Cyclin_A/phospho-Cdc2(Thr_161_Thr_14_Tyr_15)_(complex)	2.0E-13	2.4E-10	0.46	-1.01	1.46
Cyclin_A/phospho-Cdc2(Thr_14_Thr_161)_(complex)	2.0E-13	2.4E-10	0.46	-1.01	1.46
ORC/origin_(complex)	9.1E-07	1.2E-04	0.68	-0.75	1.43
ORC_(complex)	1.2E-06	1.5E-04	0.54	-0.86	1.41
MKLP2/PLK1_(complex)	2.7E-08	6.2E-06	0.56	-0.83	1.39
EIF-2-alpha(P)/EIF-2-gamma/EIF-2-beta_(complex)	4.5E-07	7.9E-05	0.89	-0.48	1.38
PICH/PLK1_(complex)	3.4E-06	3.1E-04	0.50	-0.86	1.36
eIF2/GDP_(complex)	4.1E-06	3.7E-04	0.87	-0.45	1.33
eIF2/GTP_(complex)	4.1E-06	3.7E-04	0.87	-0.45	1.33
14-3-3E_homodimer_(complex)	5.4E-06	4.6E-04	0.25	-1.07	1.32
AZIN1_bound_OAZ/ODC_complex_(complex)	5.2E-06	4.4E-04	0.57	-0.75	1.32
C1q_binding_protein_tetramer_(complex)	2.2E-08	5.3E-06	0.28	-1.03	1.30
eEF2/GDP_(complex)	2.2E-08	5.3E-06	0.28	-1.03	1.30
eEF2/GTP_(complex)	2.2E-08	5.3E-06	0.28	-1.03	1.30
KIF18A_dimer_(complex)	5.0E-09	2.5E-06	0.28	-0.99	1.27
Kinesin-5_homotetramer_(complex)	4.9E-09	2.5E-06	0.28	-0.99	1.27
KIF15_dimer_(complex)	2.4E-06	2.4E-04	0.25	-1.03	1.27
HuR/APRIL/pp32/SET/SETalpha/Nup214/SETbeta/CRM1_Complex_(complex)	1.8E-07	3.5E-05	0.48	-0.78	1.27
PML/SUMO1_(complex)	1.7E-07	3.3E-05	0.54	-0.73	1.26
ORC/CDC6/CDT1_(complex)	7.2E-08	1.6E-05	0.78	-0.44	1.22
Emi1/Cdc20_complex_(complex)	5.5E-07	8.5E-05	0.43	-0.78	1.21
DNA_Pol_alpha/primase_(complex)	1.0E-06	1.3E-04	0.70	-0.49	1.19
CCT/TriC(ATP)/unfolded_tubulin_complex_(complex)	7.5E-07	1.0E-04	0.63	-0.52	1.15
Sumo_target_protein/SUMO_1_(complex)	2.2E-08	5.3E-06	0.24	-0.89	1.13

CCT/TriC(ADP)_(complex)	1.3E-06	1.5E-04	0.66	-0.46	1.12
CCT/TriC(ATP)_(complex)	1.3E-06	1.5E-04	0.66	-0.46	1.12
FANCD2/FANCI/H2AX_(complex)	2.3E-09	1.6E-06	0.84	-0.26	1.10
SMN_complex_(complex)	1.4E-06	1.7E-04	0.36	-0.66	1.01
Nup107-160_complex_(complex)	6.1E-07	9.2E-05	0.45	-0.56	1.01
Cdt1/geminin_(complex)	1.3E-06	1.5E-04	0.21	-0.78	0.99
p27Kip1/KPNA1_(complex)	1.9E-06	2.1E-04	0.48	-0.44	0.92
p27Kip1/14-3-3_family_(complex)	1.9E-06	2.1E-04	0.48	-0.44	0.92
Shelterin_Complex/Apollo_(complex)	6.7E-06	5.3E-04	0.51	-0.41	0.91
Rev/importin-beta/B23/Ran-GTP_complex_(complex)	1.3E-06	1.5E-04	0.38	-0.52	0.90
FANCD2/FANCI/BRCA2/PALB2_(complex)	1.4E-08	5.2E-06	0.78	-0.11	0.89
mono-ubiquitinated_FANCD2_(complex)	6.4E-07	9.2E-05	0.19	-0.66	0.85
RNA_primer/origin_duplex/DNA_polymerase_alpha/primase_complex_(complex)	1.1E-07	2.2E-05	0.58	-0.27	0.85
RNA_primer/G-strand_extended_telomere_end/DNA_polymerase_alpha/primase_complex_(complex)	1.1E-07	2.2E-05	0.58	-0.27	0.85
mono-ubiquitinated_FANCI_(complex)	4.6E-11	3.9E-08	0.20	-0.61	0.82
BRCA1/BRCA2/PALB2_(complex)	1.1E-07	2.3E-05	0.48	-0.29	0.77
Rev_multimer-bound_HIV-1_mRNA/CRM1_complex_(complex)	7.2E-07	9.9E-05	0.18	-0.57	0.75
Ku/Artemis/DNA-PKcs_(complex)	5.7E-06	4.8E-04	0.46	-0.25	0.71
alpha7X1/beta1D_Integrin/Laminin_1_(complex)	6.3E-06	5.2E-04	0.35	-0.31	0.66
FANCD_and_ub-FANCI-bound_chromatin_(complex)	4.0E-13	4.1E-10	0.29	-0.36	0.65
RNA_primer-DNA_primer/origin_duplex_(complex)	1.1E-07	2.3E-05	0.47	-0.16	0.63
RNA_primer-DNA_primer/G-strand_extended_telomere_(complex)	1.1E-07	2.3E-05	0.47	-0.16	0.63
43s_Ribosome_subunit_(complex)	1.2E-06	1.5E-04	0.40	-0.07	0.47
SUMO-1/ubiquitin_(complex)	1.1E-06	1.4E-04	0.04	-0.42	0.47
Viral_RNA_dependent_RNA_polymerase_(complex)	1.5E-06	1.7E-04	0.14	-0.20	0.35

Table 3: The 41 PARADIGM abstract processes differentially activated between group A and B CRCs in our cohort (FDR ≤ 0.05). IDs have been sorted by the magnitude of the difference between group B vs. group A samples.

ID	P value	FDR	Group B median	Group A median	Group B vs. Group A

positive_regulation_of_Wnt_receptor_signaling_pathway_(abstract)	1.4E-03	2.3E-02	2.14	-1.51	3.65
negative_regulation_of_DNA_binding_(abstract)	1.9E-04	6.3E-03	1.37	-1.32	2.68
G1/S_transition_of_mitotic_cell_cycle_(abstract)	2.1E-03	3.0E-02	1.15	-1.33	2.48
anoikis_(abstract)	2.9E-03	3.7E-02	0.62	-1.57	2.20
regulation_of_centriole_replication_(abstract)	2.0E-06	2.1E-04	0.43	-1.64	2.07
protein_catabolic_process_(abstract)	1.2E-05	8.2E-04	1.32	-0.63	1.95
G2/M_transition_DNA_damage_checkpoint_(abstract)	1.4E-03	2.3E-02	1.68	-0.24	1.92
response_to_radiation_(abstract)	4.2E-03	4.7E-02	0.62	-1.05	1.67
DNA_damage_checkpoint_(abstract)	2.1E-05	1.3E-03	0.39	-1.26	1.65
regulation_of_transcription_(abstract)	1.9E-05	1.2E-03	0.43	-1.13	1.56
regulation_of_attachment_of_spindle_microtubules_to_kinetochore_(abstract)	1.9E-05	1.2E-03	0.39	-1.14	1.53
positive_regulation_of_telomere_maintenance_(abstract)	2.4E-03	3.2E-02	0.18	-1.27	1.45
MRN_complex_relocalizes_to_nuclear_foci_(abstract)	1.6E-03	2.6E-02	0.26	-1.12	1.38
prostaglandin_biosynthetic_process_(abstract)	1.5E-03	2.4E-02	0.43	-0.92	1.35
centrosome_localization_(abstract)	3.7E-04	9.8E-03	0.58	-0.76	1.34
chromosome_segregation_(abstract)	2.6E-03	3.4E-02	0.52	-0.70	1.22
microtubule-based_process_(abstract)	5.2E-08	1.2E-05	0.55	-0.67	1.22
spindle_assembly_(abstract)	1.1E-03	2.0E-02	0.48	-0.72	1.20
regulation_of_mitotic_centrosome_separation_(abstract)	9.1E-05	3.7E-03	0.39	-0.80	1.19
protein_folding_(abstract)	4.7E-05	2.2E-03	0.00	-1.17	1.17
transcription_from_RNA_polymerase_III_promoter_(abstract)	5.8E-04	1.3E-02	0.96	-0.09	1.04
G1/S_transition_checkpoint_(abstract)	3.2E-04	8.8E-03	0.14	-0.89	1.04
Golgi_organization_(abstract)	2.9E-04	8.3E-03	0.31	-0.72	1.03
positive_T_cell_selection_(abstract)	3.3E-04	9.0E-03	-0.09	-0.92	0.83
activation_of_caspase_activity_by_cytochrome_c_(abstract)	1.7E-03	2.6E-02	0.22	-0.57	0.79
DNA_replication_termination_(abstract)	3.3E-05	1.7E-03	0.66	-0.04	0.69
G2/M_transition_of_mitotic_cell_cycle_(abstract)	1.4E-03	2.3E-02	0.46	-0.12	0.58
degradation_(abstract)	1.7E-03	2.6E-02	0.12	-0.45	0.57
negative_regulation_of_cell_proliferation_(abstract)	6.9E-04	1.5E-02	0.49	0.00	0.49
negative_regulation_of_transcription_during_mitosis_(abstract)	4.3E-07	7.6E-05	0.14	-0.33	0.47
transcription_termination_(abstract)	1.7E-05	1.1E-03	0.14	-0.30	0.44

regulation_of_DNA_replication_(abstract)	1.1E-06	1.4E-04	0.30	-0.11	0.41
chromatin_modification_(abstract)	5.3E-04	1.2E-02	0.23	-0.10	0.33
regulation_of_DNA_replication_initiation_(abstract)	1.6E-03	2.6E-02	0.29	-0.04	0.33
translational_initiation_(abstract)	3.3E-04	9.0E-03	0.23	-0.08	0.30
DNA_repair_(abstract)	9.9E-04	1.9E-02	0.27	-0.04	0.30
mitotic_spindle_organization_(abstract)	3.0E-03	3.8E-02	0.11	-0.17	0.28
S_phase_of_mitotic_cell_cycle_(abstract)	9.3E-04	1.8E-02	0.14	-0.12	0.27
neural_crest_cell_migration_(abstract)	2.5E-04	7.6E-03	-0.83	0.27	-1.10
cell_cycle_arrest_(abstract)	2.2E-03	3.1E-02	-1.14	0.06	-1.20
ribosome_biogenesis_(abstract)	2.4E-03	3.2E-02	-2.69	0.22	-2.91

Table 4: The top 100 of 1922 most significantly differentially activated PARADIGM genes and gene families between group A and B CRCs of the Jorissen cohort (FDR ≤ 0.05). IDs have been sorted by the magnitude of the difference between group B vs. group A samples.

ID	P value	FDR	Group A	Group B	Group B vs. Group A
TP53	1.9E-20	7.0E-18	-2.84	4.14	6.98
XBP1-2	8.6E-18	1.2E-15	-2.95	2.21	5.16
CAV1	1.1E-18	2.3E-16	-0.66	1.84	2.50
SNAI2	3.5E-19	8.9E-17	-0.71	1.67	2.38
GJA1	1.4E-15	9.8E-14	-0.84	1.43	2.27
VCAN	2.2E-18	3.8E-16	-0.64	1.62	2.26
ENO1	2.5E-19	6.6E-17	-1.42	0.76	2.18
KDELR3	6.8E-19	1.5E-16	-1.58	0.48	2.06
HIF1A	8.7E-17	8.5E-15	-0.47	1.57	2.04
ADM	2.1E-17	2.5E-15	-1.13	0.90	2.02
C19orf10	3.5E-20	1.1E-17	-1.57	0.44	2.02
DNAJB9	1.6E-16	1.5E-14	-1.55	0.47	2.01
COL18A1	1.0E-19	3.1E-17	-0.81	1.19	2.00
CXCR4	1.8E-21	9.9E-19	-1.02	0.98	2.00
PFKFB3	7.5E-16	5.8E-14	-1.10	0.82	1.92
EGLN1	6.7E-17	6.9E-15	-1.17	0.70	1.87
RCHY1	3.0E-23	2.7E-20	-0.88	0.85	1.73
TGFB1	3.0E-17	3.4E-15	-1.02	0.71	1.73
ITGB2	1.1E-28	1.8E-24	-1.01	0.70	1.71
MCL1	4.8E-18	7.7E-16	-1.00	0.60	1.59
DUSP1	5.2E-16	4.1E-14	-0.61	0.92	1.53

FAS	4.1E-28	3.2E-24	-0.71	0.80	1.50
CSF1R	9.8E-21	4.0E-18	-0.66	0.79	1.46
DUSP5	8.9E-17	8.6E-15	-0.58	0.68	1.26
PPP3CA	1.6E-16	1.4E-14	-1.03	0.19	1.22
SLC2A3	6.7E-19	1.5E-16	-0.88	0.32	1.20
TCF7L2	2.2E-16	1.9E-14	-0.71	0.47	1.18
SERPINB5	3.3E-19	8.6E-17	-0.59	0.54	1.13
HIV-1_(rna)	2.2E-19	6.1E-17	-0.61	0.46	1.07
CSRP2	4.9E-16	3.9E-14	-0.90	0.17	1.07
HMOX1	1.2E-15	8.5E-14	-0.69	0.37	1.07
BNIP3L	3.4E-18	5.5E-16	-0.62	0.43	1.05
CCR5	2.0E-16	1.7E-14	-0.79	0.26	1.05
NT5E	4.7E-19	1.1E-16	-0.81	0.22	1.03
TNFRSF10B	1.4E-15	1.0E-13	-0.81	0.19	0.99
NDRG1	1.2E-16	1.1E-14	-0.54	0.45	0.99
ARF4	1.8E-20	6.6E-18	-0.73	0.21	0.94
CTSD	1.1E-17	1.5E-15	-0.57	0.35	0.92
PHF23	2.5E-18	4.2E-16	-0.67	0.25	0.92
CD86	5.7E-17	5.9E-15	-0.65	0.26	0.91
HCLS1	2.0E-18	3.7E-16	-0.60	0.30	0.90
SNX2	2.8E-20	9.0E-18	-0.66	0.24	0.90
TAF13	1.5E-21	8.5E-19	-0.68	0.22	0.90
ARFGAP3	5.4E-18	8.4E-16	-0.62	0.27	0.90
CDC27	3.6E-25	6.2E-22	-0.64	0.25	0.89
FBXO8	1.1E-22	8.8E-20	-0.66	0.23	0.89
TYROBP	5.6E-18	8.6E-16	-0.61	0.28	0.89
CD14	1.0E-20	4.0E-18	-0.61	0.28	0.89
HPSE	1.4E-16	1.3E-14	-0.59	0.30	0.89
FPR3	7.6E-16	5.8E-14	-0.61	0.28	0.88
NRBF2	4.9E-17	5.3E-15	-0.65	0.23	0.88
UBE2D3	1.9E-23	1.8E-20	-0.64	0.25	0.88
FCGR2A	7.7E-17	7.6E-15	-0.65	0.23	0.88
SEC24A	1.5E-17	2.0E-15	-0.62	0.26	0.88
SLC30A7	4.9E-17	5.3E-15	-0.67	0.21	0.88
PAFAH1B1	8.1E-18	1.2E-15	-0.68	0.20	0.88
SSR3	4.4E-21	2.1E-18	-0.61	0.27	0.88
RNF19B	2.8E-18	4.8E-16	-0.61	0.27	0.88
ATP6V1B2	5.3E-18	8.4E-16	-0.63	0.24	0.87
SRGN	2.4E-18	4.2E-16	-0.58	0.29	0.87
GNS	1.3E-20	5.1E-18	-0.60	0.26	0.87

C1QA	5.2E-17	5.6E-15	-0.59	0.28	0.87
NUP50	1.4E-15	9.7E-14	-0.63	0.23	0.86
ATP5A1	4.2E-16	3.3E-14	-0.77	0.09	0.86
KAT2B	1.1E-17	1.5E-15	-0.61	0.25	0.86
ARHGAP30	2.8E-20	9.0E-18	-0.61	0.24	0.86
PTPRC	1.6E-19	4.7E-17	-0.60	0.26	0.86
FCER1G	8.2E-23	6.9E-20	-0.58	0.28	0.85
CSF2RB	7.1E-18	1.1E-15	-0.58	0.27	0.85
C1QB	2.6E-19	6.7E-17	-0.58	0.27	0.85
Ubiquitin_conjugating_enzyme_(family)	1.0E-18	2.2E-16	-0.60	0.21	0.82
TLR8	4.7E-17	5.2E-15	-0.63	0.18	0.81
CDC40	3.8E-16	3.1E-14	-0.62	0.17	0.80
CXCR4_protein_(family)	1.5E-19	4.3E-17	-0.20	0.55	0.75
CXCL8_(family)	8.3E-17	8.1E-15	-0.12	0.60	0.71
VEGF	1.2E-22	8.9E-20	-0.21	0.48	0.69
P55269	4.6E-26	1.6E-22	-0.29	0.36	0.66
MDM2	4.3E-16	3.4E-14	-0.25	0.40	0.65
CCR5/CXCR4_(family)	5.5E-28	3.2E-24	-0.32	0.33	0.65
Active_caspases_(family)	9.8E-17	9.4E-15	-0.31	0.15	0.47
MEK1-2_(family)	7.5E-16	5.8E-14	-0.20	0.22	0.41
IGFBP3	1.8E-16	1.5E-14	-0.11	0.31	0.41
C13orf15	2.5E-25	4.9E-22	-0.21	0.20	0.41
PMS2	2.5E-25	4.9E-22	-0.21	0.20	0.41
MEK1/MEK2_(family)	9.0E-16	6.7E-14	-0.19	0.21	0.40
Caspase-2_p12_subunit_(family)	2.3E-16	1.9E-14	-0.24	0.11	0.35
Caspase-2_precursor_(family)	2.3E-16	1.9E-14	-0.24	0.11	0.35
Caspase-2_p18_subunit_(family)	2.3E-16	1.9E-14	-0.24	0.11	0.35
Procaspase2/3_(family)	1.6E-16	1.5E-14	-0.22	0.10	0.32
PKM2	7.6E-17	7.6E-15	-0.05	0.27	0.31
PLA2G4B	1.7E-17	2.1E-15	-0.15	0.13	0.28
SULT1A3	1.7E-17	2.1E-15	-0.15	0.13	0.28
RORA-4	5.1E-17	5.4E-15	-0.03	0.25	0.28
GPX2	1.1E-19	3.1E-17	0.29	-0.55	-0.84
CD9	8.9E-21	4.0E-18	0.23	-0.61	-0.84
ZMIZ2	8.0E-18	1.2E-15	0.19	-0.67	-0.86
TRIM24	3.9E-17	4.3E-15	0.23	-0.68	-0.90
RPA4	2.3E-16	1.9E-14	0.29	-0.62	-0.91
SCAP	9.8E-18	1.4E-15	0.20	-0.71	-0.92
USP7	1.3E-16	1.2E-14	0.20	-0.78	-0.98

Table 5: The top 100 of 1609 significantly differentially activated PARADIGM complexes between group A and B CRCs of the Jorissen cohort (FDR ≤ 0.05). IDs have been sorted by the magnitude of the difference between group B vs. group A samples.

ID	P value	FDR	Group A median	Group B median	Group B vs. group A
p53_(tetramer)_(complex)	2.1E-20	7.6E-18	-3.58	5.96	9.54
HIF1A/ARNT_(complex)	6.7E-22	3.9E-19	-3.67	2.74	6.42
JUN/FOS_(complex)	4.6E-19	1.1E-16	-2.48	3.88	6.35
p53_tetramer_(complex)	3.9E-19	9.7E-17	-0.90	2.21	3.10
MYC/Max/MIZ-1_(complex)	2.0E-19	5.5E-17	-1.68	1.27	2.95
p53/BCL2_subfamily_(complex)	2.0E-19	5.6E-17	-0.62	1.92	2.55
p53/mSin3A_(complex)	1.8E-18	3.3E-16	-0.43	1.93	2.37
p53/PIN1_(complex)	1.5E-17	2.0E-15	-0.48	1.86	2.33
p53_(tetramer)/SP1_(complex)	6.1E-18	9.2E-16	-0.01	1.89	1.90
enolase_1_dimer_(alpha)_(complex)	1.3E-17	1.7E-15	-1.22	0.67	1.90
p53/SIRT1_(complex)	4.6E-19	1.1E-16	-0.39	1.50	1.89
Glucocorticoid_receptor/Dexamethasone_Complex_(complex)	1.7E-17	2.2E-15	-0.72	1.10	1.82
HIF1A/ARNT/Cbp/p300_(complex)	3.0E-16	2.5E-14	-0.82	0.99	1.80
p53/NEDD8_(complex)	8.8E-19	1.9E-16	-0.23	1.57	1.80
Caveolin-1_bound_to_Basigin_(complex)	2.0E-17	2.5E-15	-0.27	1.52	1.79
p53/SP1_(complex)	8.2E-18	1.2E-15	-0.33	1.46	1.79
CXCR4_(dimer)_(complex)	5.9E-21	2.8E-18	-0.92	0.87	1.79
Dexamethasone/glucocorticoid_receptor_(complex)	2.3E-17	2.7E-15	-0.72	1.06	1.78
p53_(tetramer)/NF-Y_(complex)	3.4E-17	3.9E-15	-0.04	1.74	1.78
glucocorticoid/glucocorticoid_receptor_(complex)	1.9E-17	2.3E-15	-0.72	1.05	1.78
ARNT/IPAS_(complex)	6.3E-17	6.5E-15	-0.57	1.17	1.73
HIF-1-alpha/ARNT/CREB/p300/JAB1/c-JUN_(complex)	4.1E-22	2.4E-19	-0.83	0.84	1.67
MYC/Max/MIZ-1/DNMT3A/GFI1_(complex)	1.2E-16	1.1E-14	-0.73	0.93	1.66
HIF1A/JAB1_(complex)	2.4E-20	8.2E-18	-0.22	1.43	1.66
HIF1A/p53_(complex)	9.5E-21	4.0E-18	-0.25	1.40	1.64
JUN/FOS/GATA2_(complex)	4.5E-16	3.6E-14	-0.03	1.59	1.62
BAK/p53_(complex)	3.1E-18	5.2E-16	-0.24	1.34	1.59
MIZ-1/IRF8_(complex)	1.8E-16	1.6E-14	-1.24	0.34	1.58
p53_(tetramer)/Drosha/DGCR8/p68_helicase_(complex)	1.3E-18	2.6E-16	-0.04	1.47	1.51
AR/GR_(complex)	4.1E-16	3.3E-14	-0.56	0.88	1.44

HIF-1-alpha/ARNT_(complex)	7.6E-19	1.7E-16	-0.52	0.91	1.43
FAS_(trimer)_(complex)	3.7E-27	1.6E-23	-0.65	0.74	1.39
Phospho-COP1(Ser-387)/p53_complex_(complex)	1.1E-16	1.0E-14	-0.14	1.16	1.30
TLR1/MD2_(complex)	4.1E-16	3.3E-14	-0.93	0.34	1.27
ERM/c-JUN_(complex)	4.9E-17	5.3E-15	-0.83	0.34	1.17
SMAD2-3/SMAD4/MYC/Max/MIZ-1_(complex)	2.8E-17	3.2E-15	-0.50	0.63	1.13
GLUT3_tetramer_(complex)	7.1E-19	1.6E-16	-0.81	0.29	1.10
C1q_(complex)	1.5E-16	1.4E-14	-0.54	0.56	1.10
C1Q_subunit_(C1QA/C1QB/C1QC_heterotrimer)_(complex)	1.5E-16	1.4E-14	-0.54	0.56	1.10
alphaD/beta2_Integrin_(complex)	3.6E-22	2.2E-19	-0.67	0.41	1.08
Integrin_alphaDbeta2_(complex)	3.6E-22	2.2E-19	-0.67	0.41	1.08
SDF1/CXCR4_(dimer)_(complex)	1.2E-18	2.4E-16	-0.15	0.87	1.03
BAK/MCL1_(complex)	2.8E-16	2.3E-14	-0.64	0.38	1.02
Integrin_alphaLbeta2_(LFA-1)_(complex)	1.5E-25	3.6E-22	-0.65	0.36	1.01
alphaL/beta2_Integrin_(complex)	1.5E-25	3.6E-22	-0.65	0.36	1.01
alphaX/beta2_Integrin_(complex)	3.0E-22	2.0E-19	-0.64	0.37	1.01
alphaX/beta2_Integrin/heparin_(complex)	3.0E-22	2.0E-19	-0.64	0.37	1.01
Integrin_alphaXbeta2_(complex)	3.0E-22	2.0E-19	-0.64	0.37	1.01
alphaM/beta2_Integrin_(complex)	1.4E-18	2.7E-16	-0.65	0.33	0.98
Integrin_alphaMbeta2_(complex)	2.0E-21	1.1E-18	-0.64	0.33	0.98
5prime-nucleotidase_ecto_(CD73)_holoenzyme_(complex)	4.9E-19	1.2E-16	-0.75	0.20	0.95
TRAIL_receptor-2/TRAIL_Trimer_(complex)	3.3E-21	1.7E-18	-0.57	0.37	0.94
PAFAH/LIS1_(complex)	1.4E-23	1.4E-20	-0.93	0.00	0.93
MAD/Max_(complex)	5.1E-17	5.4E-15	-0.62	0.27	0.89
Mad/Max_(complex)	4.9E-17	5.3E-15	-0.62	0.27	0.89
LFA-1/ICAM_1-4_(complex)	6.7E-25	1.1E-21	-0.43	0.45	0.88
ARF4/GTP_(complex)	2.5E-20	8.2E-18	-0.70	0.16	0.86
ARF4/GDP_(complex)	2.2E-20	7.8E-18	-0.68	0.17	0.85
Cathepsin_D/ceramide_(complex)	1.0E-17	1.4E-15	-0.54	0.30	0.84
alphaM/beta2_Integrin/P-Selectin/PSGL1_(complex)	5.9E-16	4.6E-14	-0.28	0.53	0.82
LPS_complexed_with_secreted_CD14_(complex)	1.0E-20	4.0E-18	-0.57	0.24	0.81
LPS_complexed_with_GPI-anchored_CD14_(complex)	1.0E-20	4.0E-18	-0.57	0.24	0.81
CXCR4/g-alpha-q_(complex)	6.2E-20	1.9E-17	-0.62	0.18	0.79
FceRI_gamma_dimer_(complex)	8.3E-23	6.9E-20	-0.54	0.25	0.79

SDF1/CXCR4_(complex)	7.3E-21	3.4E-18	-0.34	0.44	0.79
CSF2RB_(dimer)_(complex)	7.8E-18	1.2E-15	-0.55	0.24	0.78
signal_recognition_particle_endoplasmic_reticulum_targeting_(complex)	1.1E-17	1.6E-15	-0.47	0.31	0.78
C1Q_(complex)	3.3E-16	2.7E-14	-0.31	0.46	0.77
LPS_complexed_with_CD14_(family)	1.0E-20	4.0E-18	-0.55	0.22	0.77
p-PECAM1/p-PECAM1_(complex)	6.8E-16	5.3E-14	-0.52	0.23	0.74
Integrin_alphaXbeta2/AMICA1_(complex)	2.2E-18	3.9E-16	-0.37	0.37	0.74
Integrin-alpha/Integrin-beta/Caveolin-1/FYN_(complex)	3.4E-16	2.8E-14	-0.07	0.67	0.74
full_length_TLR8_(complex)	1.7E-16	1.5E-14	-0.58	0.13	0.72
KIR2DS2_complexed_with_DAP12_(complex)	5.6E-18	8.6E-16	-0.50	0.21	0.72
TRAIL_receptor-2/TRAIL_complex_(complex)	1.9E-18	3.4E-16	-0.71	0.00	0.71
p53/MDM2_(complex)	3.3E-18	5.4E-16	-0.01	0.69	0.70
CD4/CD4/CXCR4_(complex)	1.2E-17	1.7E-15	-0.70	0.00	0.70
HIF1A/p14ARF_(complex)	1.2E-18	2.4E-16	-0.29	0.40	0.69
RNF125/E2_enzyme_(UBE2K_UbcH5a-c)/K48-polyubiquitin_(complex)	1.3E-18	2.5E-16	-0.44	0.25	0.69
LIS1/Poliovirus_Protein_3A_(complex)	1.3E-17	1.7E-15	-0.55	0.14	0.69
VEGF/VEGF_R_(complex)	3.4E-21	1.7E-18	-0.12	0.56	0.69
FPRL2/FPRL2_ligands_(complex)	2.7E-17	3.2E-15	-0.50	0.18	0.68
alphaM/beta2_Integrin/JAM-B/JAM-C_(complex)	1.9E-16	1.7E-14	-0.31	0.37	0.68
IL3RB/Jak2_(complex)	8.2E-18	1.2E-15	-0.48	0.20	0.68
TRAIL/TRAILR2/FADD/TRADD/RIP_(complex)	2.9E-18	4.8E-16	-0.34	0.34	0.68
HIV-1/VIF_(complex)	7.1E-17	7.3E-15	-0.29	0.39	0.68
Ca2+/CaM_(complex)	1.4E-16	1.3E-14	-0.18	0.47	0.65
LFA-1/JAM-A_(complex)	1.2E-17	1.6E-15	-0.31	0.26	0.57
TRAIL/TRAILR2_(complex)	6.1E-17	6.4E-15	-0.08	0.48	0.57
CCR5/g-alpha-q_(complex)	1.7E-18	3.3E-16	-0.56	0.00	0.56
active_Caspase-2_(complex)	7.3E-17	7.4E-15	-0.36	0.18	0.53
C1_complex_(complex)	8.8E-18	1.3E-15	-0.06	0.47	0.53
p-SLP-76/NCK1_(complex)	1.1E-17	1.5E-15	-0.44	0.00	0.44
SNX1/SNX2_(complex)	4.3E-17	4.7E-15	-0.43	0.00	0.43
alphaD/beta2_Integrin/ICAM3_(complex)	3.2E-17	3.7E-15	-0.21	0.22	0.43
VCAN/TLR2/TLR1_(complex)	2.4E-16	2.0E-14	-0.05	0.38	0.43
cIAP1/UbcH5C_(complex)	6.4E-20	2.0E-17	-0.41	0.00	0.41
Complement_activator/C1_complex_	2.1E-17	2.5E-15	-0.04	0.36	0.40

(complex)					
pyruvate_kinase_M2_complex_(complex)	7.3E-17	7.4E-15	-0.04	0.24	0.28
TL_phenol_transferase_1A3_homodimer_(complex)	2.1E-17	2.5E-15	-0.14	0.12	0.25

Table 6: The 88 significantly differentially activated PARADIGM abstract processes between group A and B CRCs of the Jorissen cohort (FDR ≤ 0.05). IDs have been sorted by the magnitude of the difference between group B vs. group A samples.

ID	P value	FDR	Group B median	Group A median	Group B vs. Group A
DNA_damage_(abstract)	5.3E-09	6.9E-08	-3.53	2.90	6.44
anoikis_(abstract)	2.1E-18	3.7E-16	-1.11	2.16	3.27
negative_regulation_of_DNA_binding_(abstract)	4.9E-18	7.8E-16	-0.68	1.95	2.63
response_to_radiation_(abstract)	3.4E-19	8.6E-17	-0.63	1.66	2.29
gap_junction_assembly_(abstract)	1.3E-15	9.1E-14	-0.79	1.43	2.22
G2/M_transition_DNA_damage_checkpoint_(abstract)	7.8E-17	7.8E-15	-0.02	1.59	1.61
osteoclast_differentiation_(abstract)	5.6E-10	8.9E-09	-0.46	1.14	1.59
primary_microRNA_processing_(abstract)	4.9E-18	7.8E-16	-0.03	1.32	1.34
extracellular_matrix_organization_(abstract)	1.3E-11	3.2E-10	-0.59	0.70	1.29
protein_catabolic_process_(abstract)	3.4E-17	3.8E-15	-0.22	1.00	1.21
regulation_of_cell_cycle_(abstract)	1.1E-07	1.1E-06	-0.70	0.51	1.21
activation_of_caspase_activity_by_cytochrome_c_(abstract)	1.8E-18	3.3E-16	-0.05	0.99	1.04
T-helper_1_type_immune_response_(abstract)	4.8E-09	6.2E-08	-0.72	0.20	0.92
DNA_damage_checkpoint_(abstract)	3.9E-05	2.0E-04	-0.74	0.17	0.92
phagocytosis_triggered_by_activation_of_immune_response_cell_surface_activating_receptor_(abstract)	5.5E-19	1.3E-16	-0.58	0.32	0.91
monocyte_activation_(abstract)	1.3E-08	1.6E-07	-0.59	0.29	0.88
dendritic_cell_antigen_processing_and_presentation_(abstract)	5.9E-17	6.2E-15	-0.55	0.32	0.87
cytotoxic_T_cell_degranulation_(abstract)	3.1E-08	3.4E-07	-0.62	0.25	0.87
protein_folding_(abstract)	1.7E-04	7.3E-04	-0.62	0.25	0.86
pseudopodium_formation_(abstract)	1.3E-09	1.9E-08	-0.61	0.23	0.84
lipid_biosynthetic_process_(abstract)	1.8E-03	5.6E-03	-0.65	0.19	0.84
Antiviral_Response_(abstract)	5.0E-06	3.2E-05	-0.65	0.15	0.79
Antibacterial_Response_(abstract)	5.0E-06	3.2E-05	-0.65	0.15	0.79
neutrophil_chemotaxis_(abstract)	5.6E-12	1.5E-10	-0.59	0.20	0.79

DNA_biosynthetic_process_(abstract)	1.9E-05	1.0E-04	-0.77	0.00	0.77
positive_regulation_of_leukocyte_migration_(abstract)	4.9E-15	3.0E-13	-0.37	0.39	0.76
antigen_processing_and_presentation_of_peptide_antigen_via_MHC_class_II_(abstract)	7.8E-06	4.8E-05	-0.65	0.07	0.71
antigen_processing_and_presentation_of_peptide_antigen_via_MHC_class_I_(abstract)	7.8E-06	4.8E-05	-0.65	0.07	0.71
Immunoregulation_(abstract)	7.8E-06	4.8E-05	-0.65	0.07	0.71
prostaglandin_biosynthetic_process_(abstract)	3.3E-08	3.6E-07	-0.74	-0.04	0.70
phosphatidic_acid_metabolic_process_(abstract)	1.3E-02	3.2E-02	-0.70	0.00	0.70
DNA_damage_response_signal_transduction_by_p53_class_mediator_resulting_in_induction_of_apoptosis_(abstract)	2.2E-11	5.0E-10	-0.49	0.18	0.67
stress_fiber_assembly_(abstract)	1.4E-04	5.9E-04	-0.45	0.21	0.66
neuron_apoptosis_(abstract)	2.9E-07	2.5E-06	-0.41	0.22	0.63
negative_regulation_of_T_cell_proliferation_(abstract)	5.2E-04	1.9E-03	-0.33	0.25	0.58
Golgi_organization_(abstract)	1.9E-02	4.3E-02	-0.26	0.32	0.57
cholesterol_biosynthetic_process_(abstract)	5.2E-03	1.4E-02	-0.55	0.03	0.57
spindle_assembly_(abstract)	5.0E-03	1.4E-02	-0.34	0.23	0.57
regulation_of_granulocyte_colony-stimulating_factor_production_(abstract)	2.1E-10	3.7E-09	-0.28	0.29	0.57
regulation_of_transcription_(abstract)	1.8E-04	7.6E-04	-0.47	0.10	0.57
necrosis_(abstract)	1.0E-15	7.5E-14	-0.27	0.25	0.53
Schwann_cell_development_(abstract)	1.3E-12	4.2E-11	-0.25	0.19	0.44
Metastasis_(abstract)	4.9E-03	1.4E-02	-0.08	0.33	0.42
positive_regulation_of_phagocytosis_(abstract)	1.9E-12	5.9E-11	-0.24	0.17	0.41
skeletal_muscle_tissue_development_(abstract)	1.2E-05	6.9E-05	0.22	0.63	0.41
actin_cytoskeleton_organization_(abstract)	5.0E-08	5.2E-07	-0.44	-0.03	0.41
activation-induced_cell_death_of_T_cells_(abstract)	5.0E-04	1.8E-03	-0.03	0.38	0.41
JNK_cascade_(abstract)	6.0E-07	4.8E-06	-0.11	0.25	0.37
G1/S_transition_checkpoint_(abstract)	4.2E-04	1.6E-03	-0.51	-0.15	0.36
neutrophil_activation_(abstract)	1.1E-05	6.5E-05	-0.35	0.00	0.35
regulation_of_interleukin-6_production_(abstract)	3.9E-15	2.4E-13	-0.04	0.28	0.32

respiratory_burst_involved_in_inflammatory_response_(abstract)	1.2E-07	1.2E-06	-0.32	0.00	0.32
G1_phase_of_mitotic_cell_cycle_(abstract)	3.5E-04	1.4E-03	-0.25	0.07	0.32
positive_regulation_of_cell_migration_(abstract)	1.2E-04	5.3E-04	-0.47	-0.16	0.31
natural_killer_cell_activation_(abstract)	1.2E-08	1.4E-07	-0.13	0.18	0.31
clathrin-independent_pinocytosis_(abstract)	9.7E-06	5.8E-05	-0.18	0.12	0.30
positive_regulation_of_endocytosis_(abstract)	9.7E-06	5.8E-05	-0.18	0.12	0.30
liver_development_(abstract)	9.7E-06	5.8E-05	-0.18	0.12	0.30
regulation_of_epithelial_cell_migration_(abstract)	9.7E-06	5.8E-05	-0.18	0.12	0.30
regulation_of_mitotic_centrosome_separation_(abstract)	1.3E-02	3.2E-02	-0.29	0.00	0.29
The_NLRP3_inflammasome_(abstract)	1.1E-04	5.0E-04	-0.17	0.12	0.28
positive_regulation_of_JNK_cascade_(abstract)	9.8E-13	3.3E-11	-0.16	0.12	0.28
cell_growth_and/or_maintenance_(abstract)	1.6E-03	5.1E-03	-0.23	0.05	0.28
apoptosis_(abstract)	2.2E-05	1.2E-04	-0.04	0.24	0.27
positive_regulation_of_cell_proliferation_(abstract)	1.2E-03	4.0E-03	-0.32	-0.05	0.26
calcium_ion-dependent_exocytosis_(abstract)	4.0E-02	8.3E-02	-0.12	0.14	0.26
membrane_fusion_(abstract)	4.0E-02	8.3E-02	-0.12	0.14	0.26
tube_development_(abstract)	2.6E-05	1.4E-04	-0.13	0.12	0.25
actin_filament_polymerization_(abstract)	2.4E-03	7.3E-03	0.17	-0.09	-0.26
axonogenesis_(abstract)	3.9E-02	8.2E-02	-0.21	-0.49	-0.27
embryonic_digit_morphogenesis_(abstract)	1.1E-07	1.0E-06	0.03	-0.25	-0.28
glutamate_secretion_(abstract)	1.4E-07	1.3E-06	0.21	-0.11	-0.31
CD4-positive_CD25-positive_alpha-beta_regulatory_T_cell_lineage_commitment_(abstract)	8.3E-04	2.9E-03	0.06	-0.28	-0.34
positive_regulation_of_Wnt_receptor_signaling_pathway_(abstract)	2.5E-02	5.6E-02	0.20	-0.17	-0.37
chromatin_remodeling_(abstract)	2.7E-02	6.0E-02	-0.28	-0.66	-0.38
heart_development_(abstract)	1.8E-03	5.8E-03	0.03	-0.45	-0.48
myoblast_fusion_(abstract)	4.8E-06	3.1E-05	0.52	0.00	-0.52
membrane_budding_(abstract)	2.2E-07	1.9E-06	0.03	-0.48	-0.52
cell_cycle_arrest_(abstract)	2.4E-07	2.1E-06	0.00	-0.52	-0.52
cytokine_production_involved_in_inflammatory_response_(abstract)	2.1E-10	3.7E-09	0.28	-0.29	-0.57

negative_regulation_of_cell_cycle_(abstract)	2.3E-02	5.2E-02	0.43	-0.19	-0.62
BMP_signaling_pathway_(abstract)	1.8E-03	5.6E-03	0.77	0.00	-0.77
regulation_of_isotype_switching_to_IgG_isotypes_(abstract)	5.1E-06	3.3E-05	0.21	-0.59	-0.80
apoptotic_nuclear_changes_(abstract)	1.9E-07	1.7E-06	0.17	-0.63	-0.80
virus_assembly_(abstract)	1.6E-13	6.7E-12	0.59	-0.25	-0.84
megakaryocyte_differentiation_(abstract)	2.0E-05	1.1E-04	0.29	-0.60	-0.89
Bergmann_glial_cell_differentiation_(abstract)	8.4E-11	1.7E-09	0.32	-0.97	-1.29
ribosome_biogenesis_(abstract)	1.7E-14	8.8E-13	0.03	-2.64	-2.67

Appendix D

BMC Genomics

METHODOLOGY ARTICLE Open Access

Quality assessment and data handling methods for Affymetrix Gene 1.0 ST arrays with variable RNA integrity

Katie S Viljoen and Jonathan M Blackburn*

Abstract

Background: RNA and microarray quality assessment form an integral part of gene expression analysis and, although methods such as the RNA integrity number (RIN) algorithm reliably asses RNA integrity, the relevance of RNA integrity in gene expression analysis as well as analysis methods to accommodate the possible effects of degradation requires further investigation. We investigated the relationship between RNA integrity and array quality on the commonly used Affymetrix Gene 1.0 ST array platform using reliable within-array and between-array quality assessment measures. The possibility of a transcript specific bias in the apparent effect of RNA degradation on the measured gene expression signal was evaluated after either excluding quality-flagged arrays or compensation for RNA degradation at different steps in the analysis.

Results: Using probe-level and inter-array quality metrics to assess 34 Gene 1.0 ST array datasets derived from historical, paired tumour and normal primary colorectal cancer samples, 7 arrays (20.6%), with a mean sample RIN of 3.2 (SD = 0.42), were flagged during array quality assessment while 10 arrays from samples with RINs < 7 passed quality assessment, including one sample with a RIN < 3. We detected a transcript length bias in RNA degradation in only 5.8% of annotated transcript clusters (p-value 0.05, FC ≥ |2|), with longer and shorter than average transcripts under- and overrepresented in quality-flagged samples respectively. Applying compensatory measures for RNA degradation performed at least as well as excluding quality-flagged arrays, as judged by hierarchical clustering, gene expression analysis and Ingenuity Pathway Analysis; importantly, use of these compensatory measures had the significant benefit of enabling lower quality array data from irreplaceable clinical samples to be retained in downstream analyses.

Conclusions: Here, we demonstrate an effective array-quality assessment strategy, which will allow the user to recognize lower quality arrays that can be included in the analysis once appropriate measures are applied to account for known or unknown sources of variation, such as array quality- and batch- effects, by implementing ComBat or Surrogate Variable Analysis. This approach of quality control and analysis will be especially useful for clinical samples with variable and low RNA qualities, with RIN scores ≥ 2.

Keywords: Gene expression profiling, Microarray, RNA quality, RNA integrity number, Quality control, ComBat, Surrogate variable analysis, Non-biological experimental variance

* Correspondence: jonathan.blackburn@uct.ac.za
Institute of Infectious Disease and Molecular Medicine, University of Cape Town, Anzio Road, Observatory, Cape Town 7925, South Africa

Background

RNA degradation is a common concern in gene expression analysis, especially for clinical samples where RNA degradation may occur before sample collection [1]. A wealth of archival material, either snap frozen or formalin fixed and paraffin embedded (FFPE), could potentially be used for gene expression analysis, given an appropriate method to evaluate and account for the effect of RNA degradation on the quality of downstream gene expression data. Methods such as the RNA integrity number (RIN) algorithm reliably assesses RNA integrity by extracting features from the RNA electropherogram. The RIN algorithm was developed using learning tools to identify regions (features) indicative of RNA integrity in the electropherogram, which are then used to compile the RNA integrity number on a scale of 1 to 10. However, the relevance of RNA integrity in gene expression analysis, especially when there is large variability between samples, requires further investigation and validation on a platform specific basis. The impact of RNA integrity on gene expression analysis has been investigated on both qRT-PCR and certain microarray platforms [2-7]. Opitz et al investigated the impact of RNA degradation on Agilent 44 k gene expression profiling by subjecting RNA from clinical biopsies to temperature-induced RNA degradation and comparing gene expression to the original, intact samples. Notably, less than 1% of genes were affected, even after substantial RNA degradation, where control and test samples had RINs of 9 and 5 respectively. The affected transcripts were relatively shorter, had lower GC content, or had probes relatively closer to the 5' region of the gene compared to more robust genes [6]. Although the process of RNA degradation is not fully understood, both exonuclease and endonuclease activity is likely to play an important role [6]. Classical oligo-dT based cDNA synthesis, which starts at the poly-A tail, will most certainly be compromised by exonuclease activity. In contrast random priming does not rely on full length mRNA and therefore is in theory at least partially relieved from the affects of RNA degradation [6-9].

When using semi-degraded RNA for gene expression studies, reliable measures of array quality provide valuable information that can be used to guide downstream analysis. Microarray data quality may be defined in terms of accuracy (systematic bias between the true and measured value), precision (the uncertainty in replicated measures), specificity (the selective power of the measurement to respond only to the specific targets) and sensitivity (the expression range potentially covered by the measurement) [10]. Any attempt to utilise array quality results to guide downstream analysis should ideally take into account the possible effects of RNA degradation on sensitivity, specificity and accuracy. In previous work, Binder et al proposed a single-array preprocessing method that allows correction for systematic biases such as RNA degradation by utilising information on the 3'/5'-amplification bias and the sample-specific calling rate [10]. Lassmann et al proposed using a data adjustment method to allow comparative analysis of microarray datasets derived from fresh frozen vs. FFPE samples by centering the log intensities of each probe set independently to a mean of zero in both groups [8]. Chow et al evaluated the suitability of different quality control and preprocessing strategies for use with partially degraded RNA samples on the Illumina DASL-based gene expression assay using mean inter-array correlation and multivariate distance matrix regression (MDMR) as a measure of success [11]. Unfortunately none of these studies are directly applicable to one of the most commonly used human transcriptomic microarray platforms, namely Affymetrix Gene 1.0 ST arrays, either because they do not use a random priming approach or because the design of the microarray platform differs substantially from Gene 1.0 ST arrays. We therefore identified two alternative approaches that might be used as compensatory methods: Firstly, Johnson et al developed an empirical Bayes algorithm, ComBat, to directly adjust for non-biological experimental variation. As the name implies, this method is most often used to adjust for batch effects i.e. when microarrays are processed on different dates [12]. Secondly, Leek et al developed a method called Surrogate Variable Analysis (SVA), which examines the contribution of sources ofsignal due to unknown (surrogate) variables in high-dimensional data sets, which may confound the biological signal of interest [13]. The surrogate variables are constructed directly from the gene expression data where groups of genes that are affected by each source of variation are identified, factors are then estimated for each array which can be included in a linear model to adjust for unknown sources of noise e.g. RNA- or array-quality.

Here, we investigate the relationship between RNA integrity and array quality on Affymetrix Gene 1.0 ST arrays for 34 paired colorectal tumour and adjacent normal biopsies of highly variable RNA integrity. We assume that at a certain point on the RIN scale, RNA will be degraded to the extent where fragments are too small to analyse reliably and for the purpose of this analysis we arbitrarily select a RIN cutoff of 2. We describe the within- and between-array quality control measures and analysis methods that we found most relevant for gene expression analysis of samples with highly variable RINs on Affymetrix Gene 1.0 ST arrays. We then investigate the possibility of a transcript-length dependency in RNA degradation. Finally, we apply array quality information to either exclude quality-flagged arrays, to directly adjust the data using the ComBat algorithm, or to account for unknown sources of variation (such as RNA integrity or array quality) in the model fitting process using SVA. The data discussed, have been submitted to ArrayExpress, with accession number E-MEXP-3715.

Results

Array quality

We assessed array quality using within- and between-array measures – the former to assess raw data quality (Figure 1a & 1b), and the latter to assess the quality of an array relative to a large publically available collection of high quality Gene 1.0 ST arrays (Figure 1c). Raw array quality was investigated at the probe level by calculating the difference between the means of perfect match- and background-probes for each array as well as the coefficient of variation (CV) across all probes for each array. Preprocessed data quality was assessed using the global normalised, unscaled standard error (GNUSE) [14]. See Methods section for details.

The 34 RNA samples used in this study had a mean RIN of 6.3 and a standard deviation of 2.0. Samples that failed all three measures of quality had RINs between 2 and 3.3 as summarised in Table 1. Samples were ranked by GNUSE median and we found a good concordance in terms of ranking between the different quality control metrics. Samples that failed at least two out of the three quality measures were flagged for downstream analysis, resulting in 7 out of 34 samples being flagged (mean RIN = 3.2; SD = 0.42). Interestingly, for one sample with a RIN of 2.6, array quality was not compromised, judged by our quality measures. The possibility of a RIN-independent RNA quality factor, such as chemical purity, was investigated by performing a two-tailed Student's t-test, comparing A260/230 ratios between quality-flagged and quality-passed sample groups but no significant association was found (p-value = 0.14).

Transcript-dependent effects of RNA degradation on accuracy

To investigate a possible probe-positional intensity bias related to RNA integrity, we plotted the mean probe intensity from the 5'- to 3' end of the sequence using 4644/32321 (14.4%) of transcript clusters for Gene 1.0 ST arrays and 54130/54675 (99%) of probesets for HGU133-plus2 arrays. The number of probes per set varies for GeneST arrays, so we selected the largest group (N = 4664), which had exactly 25 probes/set. Interestingly, from the 4644 transcript clusters displayed in Figure 2, Gene ST 1.0 arrays, do not display the same probe-positional intensity bias typically seen in oligo-dT based arrays such as the HGU133-plus2 arrays.

We next investigated which genes were most affected in our quality-flagged category and identified 1994 out of 21943 annotated transcript clusters (with 1172 uniquely identified genes) that were significantly different (fold change ≥ |2|, adjusted p-value ≤ 0.05) between the two quality categories previously discussed. Of the 1172 uniquely identified genes, 1032 and 140 showed decreased or increased intensity in the quality-flagged category respectively (Figure 3a). To investigate transcript characteristics in the genes most affected, we compared transcript lengths (taken as the median cDNA length for each gene) between the different groups. Compared to the unaffected genes, median cDNA lengths of genes that showed increased intensity were significantly shorter (p-value < 2.2 e – 16) while those with decreased intensity significantly longer (p-value = 2.9 e – 9) with regards to quality, judged using the Mann Whitney test (Figure 3b).

Quality dependent methods of data adjustment and analysis

After assigning samples to two categories according to array quality measures, we next assessed the performance of the five preprocessing and analysis methods. Broadly speaking, the data was either directly adjusted for quality effects using ComBat, or quality-flagged samples were

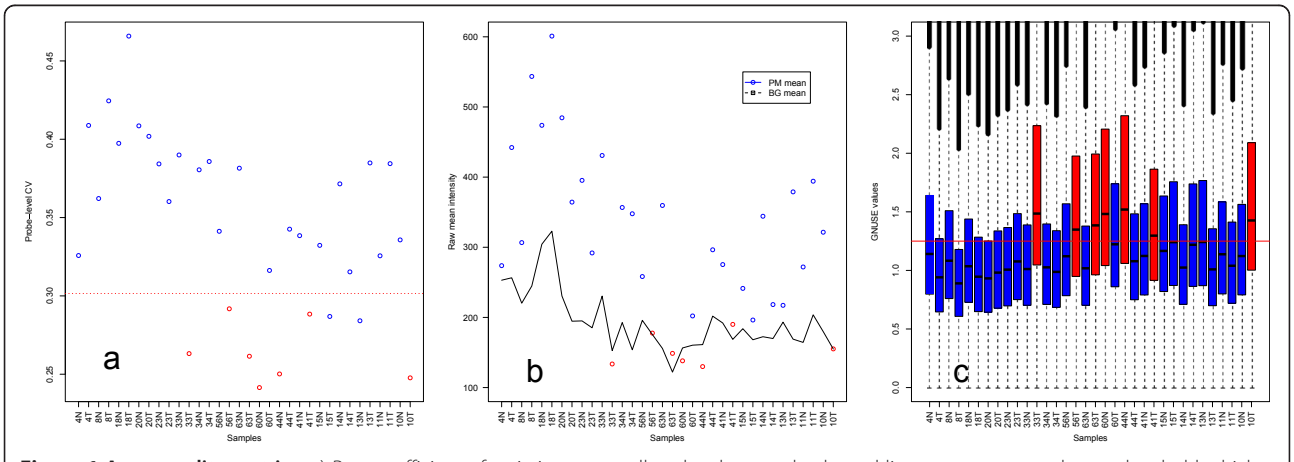


Figure 1 Array quality metrics. a) Raw coefficient of variation across all probes by sample, the red line represents our chosen threshold which is calculated as 2SD from the mean of CVs for arrays with RINs > 6. **b**) Raw perfect match mean - background mean **c**) Global normalised unscaled errors (GNUSE) across probes for each array. Samples that were flagged during quality assessment are highlighted in red.

Table 1 Array quality assessment summary

Sample ID	RIN	RNA 260/230 ratio	GNUSE	probe-level CV	PM-BG	Array weight
44N	3	2.41	**fail (1)**	**fail (3)**	**fail (1)**	**0.22 (1)**
33T	2.8	2.08	**fail (2)**	**fail (5)**	**fail (2)**	**0.28 (2)**
60N	3.2	2.03	**fail (3)**	**fail (1)**	**fail (3)**	**0.42 (3)**
63T	3	2.2	**fail (4)**	**fail (4)**	**pass**	**0.59 (6)**
10T	3.2	2.18	**fail (5)**	**fail (2)**	**fail (4)**	**0.60 (7)**
56T	3.3	1.87	**fail (6)**	**fail (10)**	**fail (5)**	**0.42 (4)**
41T	4.2	2.21	**fail (7)**	**fail (9)**	**pass**	**0.78 (8)**
13N	4.6	2.24	pass	fail (7)	pass	0.82 (9)
15T	4.8	2.15	pass	fail (8)	pass	1.07 (15)
4N	2.6	1.62	pass	pass	pass	0.44 (5)
18N	7.1	1.66	pass	pass	pass	0.83 (10)
8T	8.5	2.16	pass	pass	pass	0.85 (11)
56N	6.5	1.94	pass	pass	pass	0.95 (12)
20T	7.4	1.6	pass	pass	pass	1.02 (13)
44T	6.9	1.72	pass	pass	pass	1.03 (14)
11T	8.6	2.16	pass	pass	pass	1.07 (16)
60T	6.4	1.64	pass	pass	pass	1.09 (17)
14T	6.4	1.76	pass	pass	pass	1.09 (18)
13T	8.3	2	pass	pass	pass	1.11 (19)
23T	7	2.17	pass	pass	pass	1.18 (20)
8N	7.1	2.22	pass	pass	pass	1.25 (21)
18T	7.4	1.85	pass	pass	pass	1.26 (22)
33N	8.1	1.82	pass	pass	pass	1.45 (23)
34T	8	2.25	pass	pass	pass	1.49 (24)
11N	6.8	1.94	pass	pass	pass	1.50 (25)
20N	7.3	2.11	pass	pass	pass	1.50 (26)
63N	7.5	2.13	pass	pass	pass	1.61 (27)
23N	8.4	2.02	pass	pass	pass	1.61 (28)
34N	8.3	2.21	pass	pass	pass	1.61 (29)
14N	8.1	2.36	pass	pass	pass	1.74 (30)
41N	5.4	2.07	pass	pass	pass	1.76 (31)
10N	7.3	1.78	pass	pass	pass	1.78 (32)
15N	6.9	2.16	pass	pass	pass	1.90 (33)
4T	8.4	2.25	pass	pass	pass	2.14 (34

Array performance is ranked for each measure with 1 considered the worst quality. Samples highlighted in bold were flagged for downstream analysis.

excluded from the analysis, or possible quality effects were addressed by including known or unknown sources of non-biological variance in the linear model fit to assess differential expression.

The five methods of data preprocessing and analysis, further detailed in the Methods section, were: 1) Estimating array quality weights which were then included in the linear model fit; 2) Excluding quality-flagged arrays from the analysis; 3) Applying a batch correction algorithm, ComBat, [12] to directly adjust the data according to quality, where arrays were divided into two categories according to the array quality assessment; 4) "Quality" and "batch" were included as a factors in the linear model together with disease status; 5) Possible unknown sources of non-biological variance, such as quality, was estimated by SVA, with the output incorporated into the linear model fit [13].

To assess the effect of using ComBat for direct data adjustment, hierarchical clustering using Euclidian distance was performed before and after direct adjustment (Figure 4). We chose to use Euclidian distance based on research by Gibbons et al who demonstrated that, for log-transformed expression data, using Euclidian distance is more appropriate than Pearson's correlation coefficients

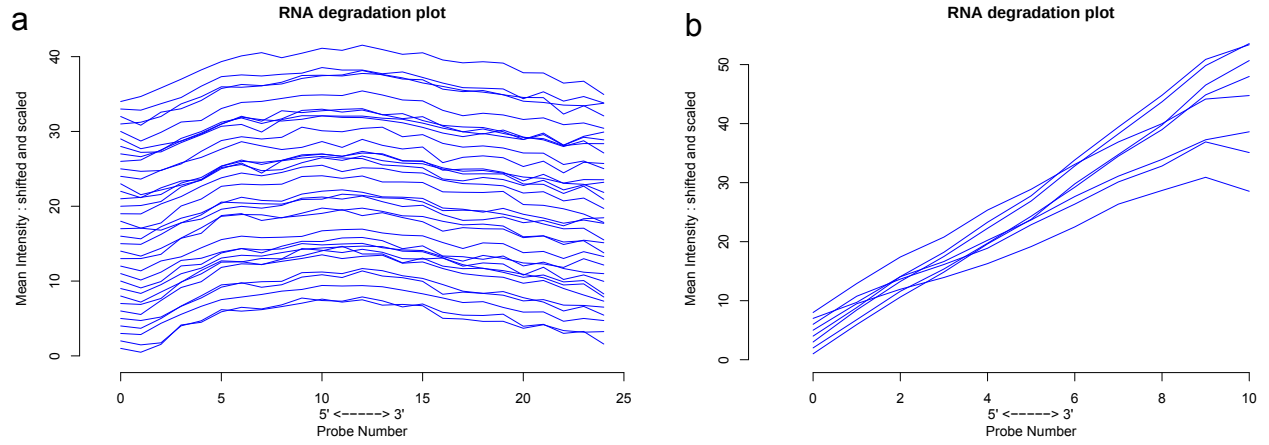


Figure 2 Mean probe intensity by probe position. Each line represents an array for **a**) Gene 1.0 ST arrays: transcript clusters with exactly 25 probes (N = 4644) and **b**) HGU133-plus2 arrays previously analysed with a subset of the cohort: probesets with exactly 11 probes per probeset.

[15]. Before adjustment, samples that were flagged during quality assessment cluster closely together, irrespective of the disease status of the samples. After adjustment, the maximum distance between samples is greatly reduced, and quality-flagged samples no longer cluster together. Also, samples segregate more clearly by disease status after adjustment. Furthermore, applying ComBat clearly has a stabilising effect on the transcript clusters most affected by RNA quality (Figure 5b & 5c).

SVA identified two surrogate variables that were subsequently used in downstream analysis. Plotting the estimates of these surrogate variables for each sample revealed a pattern whereby samples were clearly grouped by batch and quality (Figure 6). Importantly, SVA identified these two variables without supervision.

To evaluate the performance of each method, we first compared the number of differentially expressed genes detected between tumour and normal samples at a stringent p-value of 0.01. For our analysis, we did not use a fold change cutoff since we feel that artificial fold change cutoffs, which exclude subtle changes in the expression of many genes, may result in the loss of valuable biological information, or worse, affect the interpretation of the data – this is particularly true for applications such as network/pathway analysis [16].

SVA and ComBat detected 2137 and 1945 genes (p-value ≤ 0.01), respectively. The top four methods had 1117 differentially expressed genes in common (Figure 7). At the commonly used p-value- and fold change-cutoffs of 0.05 and 2 respectively, SVA,

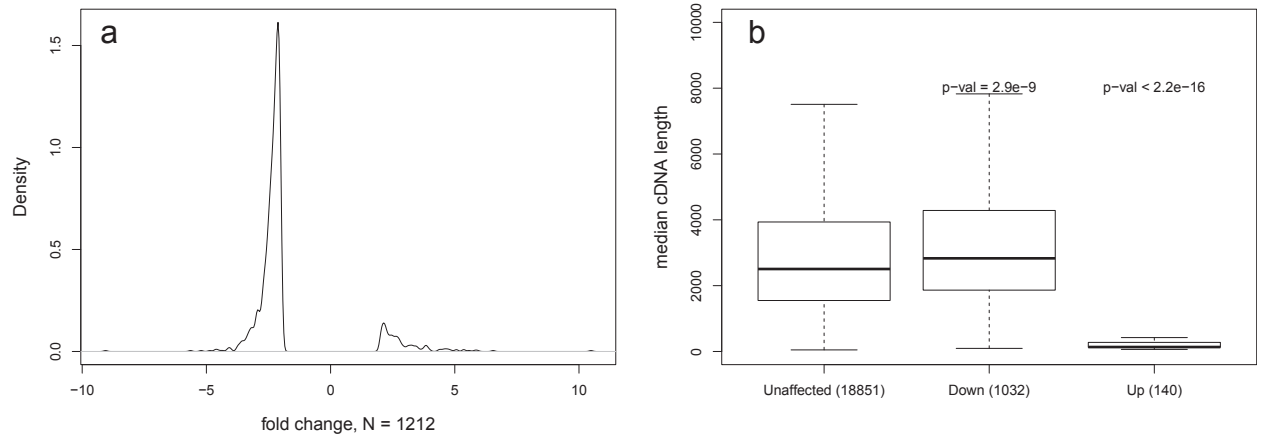


Figure 3 Characteristics of genes most affected by RNA degradation. Comparison of samples that either passed or were flagged during QC. **a**) Fold change distribution of annotated transcript clusters comparing samples that were flagged vs. samples that passed QC **b**) Gene lengths of uniquely identified genes. Expression signal significantly increased (Up) or decreased (Down) with respect to the 'Unaffected' group, judged using a Mann-Whitney test. Adjusted p-value ≤ 0.01, |fold change| > 2.

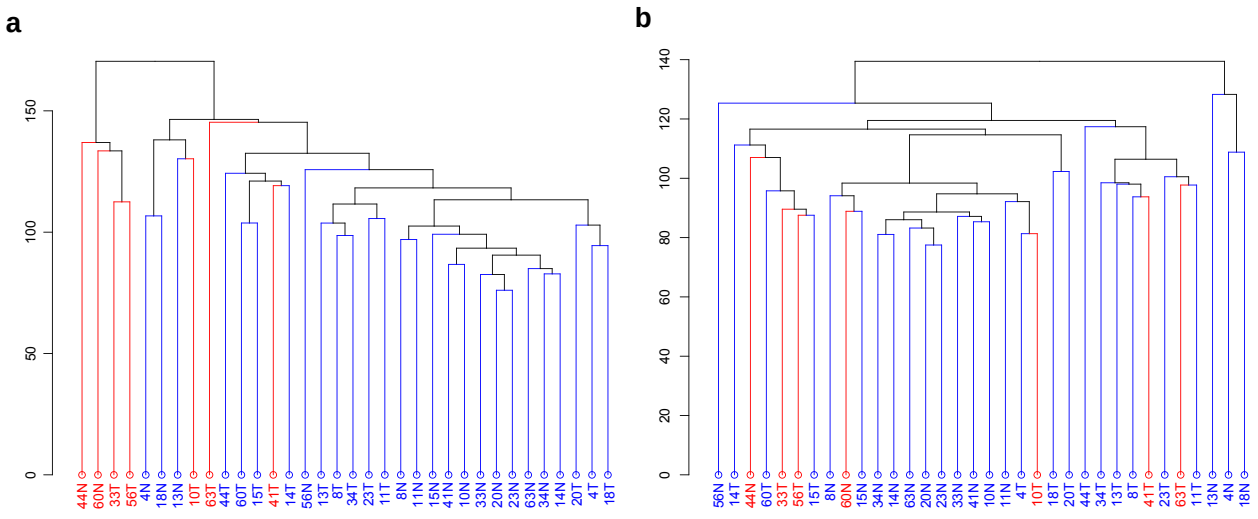


Figure 4 Expression profiles of samples clustered using average linkage hierarchical clustering. a) Sample clustering after preprocessing. **b**) Sample clustering after preprocessing and correction for batch and quality using ComBat. Samples that were flagged during quality assessment are highlighted in red. The dissimilarity measure (height) used was 1- Pearson correlation of the $\log_2$-transformed expression values.

Combat, ArrayWeights and excluding arrays, produced 447, 475, 461 and 521 differentially expressed genes respectively, suggesting similar performance under these criteria. We next assessed the relevance of these differentially expressed genes in colorectal cancer using Ingenuity Pathway Analysis where, statistically significant over-representation of our listed genes in a given process such as "colorectal tumour" or "infection of embryonic cell lines" is scored by p-value.

We considered the top 10 functions for each method (Table 2) from which it was clear that the 615 and 423 additional genes identified as differentially expressed by SVA and ComBat, compared to that obtained when excluding quality-flagged arrays, were certainly relevant to colorectal cancer. Using IPA, we considered the top 10 upstream regulators (highest absolute activation z-scores) when comparing tumour vs. normal samples, to further investigate the utility of SVA or ComBat as suitable analysis methods when including low-RIN samples (Table 3). We found considerable overlap in the identity and direction of activation of these upstream regulators between the methods compared.

qRT-PCR validation of select genes

In order to ascertain whether or not data obtained by microarray analysis with low-RIN samples were comparable to the results obtained using the method designed by Antonov et al for qPCR analysis of low-RIN samples, we selected two genes, dipeptidase 1 (DPEP1) and claudin 1 (CLDN1), for qRT-PCR validation. Given that our microarray data

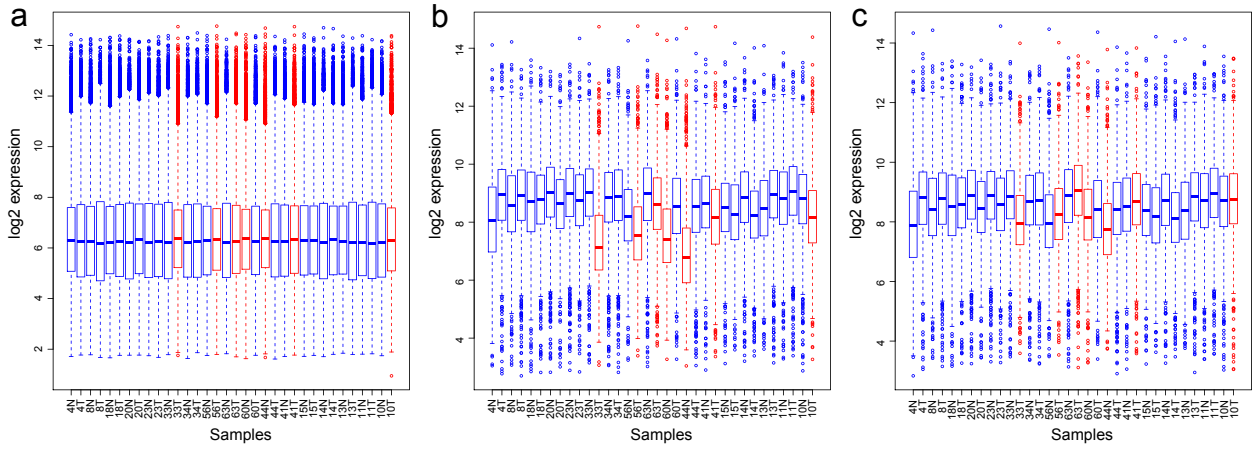


Figure 5 Boxplots of frma expression. a) All transcript clusters. **b**) Genes most affected by quality (adjusted p-value ≤ 0.01, |fold change| > 2. **c**) Samples that were flagged during quality assessment are highlighted in red.

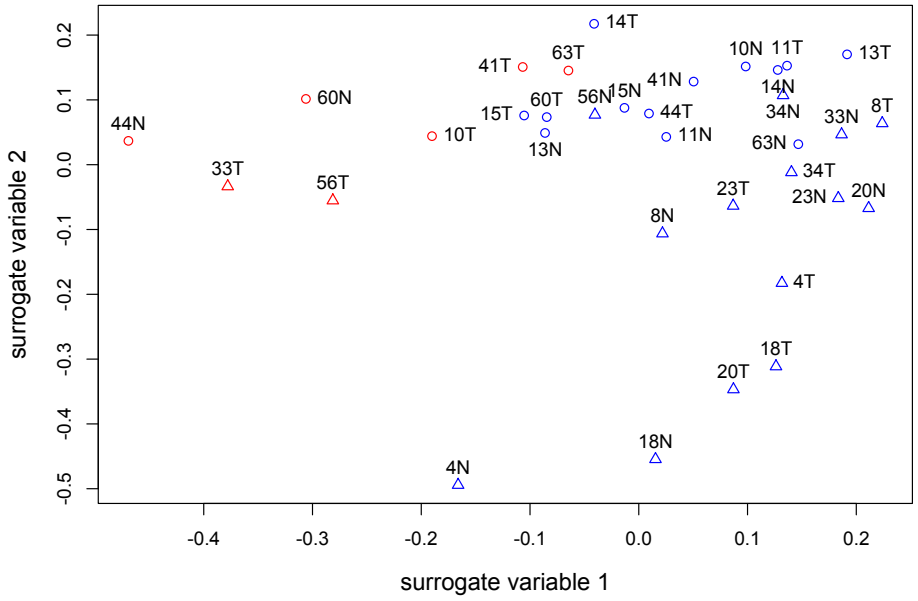


Figure 6 Surrogate variable analysis results. Samples that were flagged during quality assessment are highlighted in red. Two latent variables were identified by SVA. Circles and triangles represent samples from two different batches.

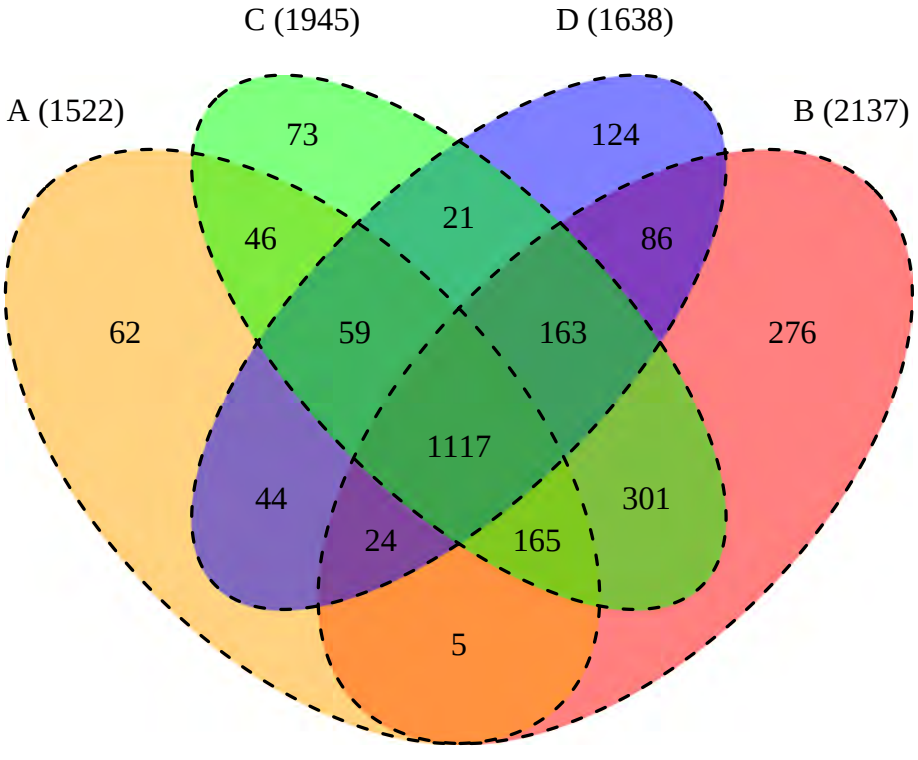


Figure 7 Venn diagram of unique differentially expressed genes (tumour vs. normal) with adjusted p-values ≤ 0.01 for the four best-performing methods. A - removing quality-flagged arrays before analysis. **B** - applying SVA to batch corrected data. **C** - ComBat used to correct for batch and quality. **D** - Array weights included in the linear model.

Table 2 P-values for evidence for overrepresentation in the functions listed for each method

Functions	A	B	C	D	E
Cancer	7.72E-29	NA	8.15E-24	NA	3.38E-23
cancer	NA	2.58E-25	NA	2.70E-26	NA
carcinoma	8.64E-37	2.52E-33	1.87E-34	2.87E-32	5.56E-30
colon cancer	1.30E-26	1.19E-36	1.10E-26	1.99E-21	3.29E-21
colon tumor	1.10E-26	4.31E-37	3.65E-27	1.80E-21	7.28E-22
colorectal cancer	2.27E-26	4.74E-29	2.43E-26	1.11E-21	1.98E-23
colorectal tumor	2.28E-26	6.80E-29	2.97E-26	4.67E-22	1.72E-23
digestive organ tumor	2.68E-31	6.82E-32	1.24E-28	7.27E-27	2.72E-29
epithelial tumor	2.16E-38	NA	2.27E-35	NA	1.11E-30
gastrointestinal tract cancer	2.35E-25	2.42E-28	4.00E-24	3.19E-21	5.31E-22
intestinal cancer	2.02E-26	5.77E-29	2.58E-26	1.03E-21	1.55E-23
neoplasia	NA	1.63E-24	NA	1.10E-25	NA
solid tumor	3.31E-35	8.07E-32	6.80E-33	4.65E-31	3.88E-29
tumorigenesis	NA	1.55E-26	NA	3.31E-28	NA
uterine serous papillary cancer	3.46E-21	1.71E-20	8.61E-25	1.26E-22	1.14E-15

A - excluding quality-flagged arrays from the analysis. B - applying SVA to batch corrected data. C - ComBat used to correct for batch and quality. D - Array weights included in the linear model. E - including batch and quality as factors in the linear model.

analysis suggests ~95% of genes are unaffected by RNA integrity, we wished to compare microarray and qPCR data for genes that were apparently unaffected by RNA integrity; DPEP1 and CLDN1 were found to be significantly differentially expressed in our microarray data by all of the five methods used and, in addition, there is strong literature evidence for their differential expression between tumour and normal samples. From reference genes previously cited as suitable for colorectal cancer studies, we selected those most stably expressed in our cohort using the Normfinder algorithm (UBC, B2M, ATP5E) [17-21]. We found good correlations, for both CLDN1 (Adjusted R^2 = 0.81) and DPEP1 (Adjusted R^2 = 0.83), between qRT-PCR- and microarray-based fold change values (Figure 8), irrespective of RIN score.

Discussion

RNA is extremely vulnerable to degradation and as such has the potential to introduce a systematic bias in gene expression measures. Reliable measures of sample and data quality are therefore essential to evaluate the effects of RNA integrity on accuracy, sensitivity and specificity of gene expression results. From previous studies as well as our own, it is now clear that the level of acceptable RNA degradation within an experiment depends largely on the experimental design, platform and application. Multiple studies have demonstrated an improvement in microarray and qRT-PCR performance by using random priming when RNA integrity is in doubt. Here we observed a direct association between RINs and array quality in the majority of cases. To gauge the consequences of using these arrays in downstream analysis, we compared quality-flagged to quality-passed arrays and found a relatively small subset of genes, 1172/20019, to be significantly affected (p-value 0.05, FC ≥ |2|) in our samples on the Gene 1.0 ST platform. It is of course possible that the exact identity and proportion of the affected genes may differ between studies on Gene 1.0 ST arrays but, based on our data, we suggest that the overall proportion of affected genes is unlikely to be significantly different to that observed here. Depending on the application, this may or may not have an effect on the study outcome. However, the most common microarray applications such as finding differentially expressed genes between two conditions, pathway analysis, and clustering do not rely on interrogating specific genes and appear to be largely robust to the effects of RNA degradation on this platform (Table 2).

Using within- and between-array quality measures, we investigated the relationship between RNA integrity and array quality on Affymetrix Gene 1.0 ST arrays. We found a combination of within- and between-array quality measures useful to rank samples by array quality. However, the single most useful array quality measure appears to be GNUSE, since it provides a more general measure of array quality relative to a large set of publically available arrays. We found that 86% of samples with RINs ≤ 3.3 were flagged by at least two of our quality control measures. One sample with RIN score < 3 passed all three quality measures, although it did have relatively low array quality weight. Furthermore, 10 out of 17 samples with RIN scores ≤ 7 passed at least 2 out of 3 quality measures, suggesting that the widely used RIN cutoff of 7 is too stringent for Gene 1.0 ST arrays.

We then examined the genes most affected by RNA degradation and demonstrated a relationship between

Table 3 Top 10 IPA-derived upstream regulators, by absolute activation z-score

A					
Upstream Regulator	**Log Ratio**	**Molecule Type**	**Predicted Activation State**	**Activation z-score**	**p-value of overlap**
TP53		transcription regulator	Inhibited	−4.88	1.05E-16
CDKN1A	−0.469	kinase	Inhibited	−3.274	4.20E-10
TRAF2		enzyme	Activated	2.804	3.06E-06
CCNK		other	Activated	2.905	3.83E-04
TNF		cytokine	Activated	2.935	7.69E-04
IL1B		cytokine	Activated	2.952	1.76E-01
TP63		transcription regulator	Activated	3.181	8.37E-10
TREM1		other	Activated	3.352	3.69E-05
FOXM1	1.37	transcription regulator	Activated	4.28	3.71E-17
Mek		group	Activated	4.336	2.38E-07
B					
Upstream Regulator	**Log Ratio**	**Molecule Type**	**Predicted Activation State**	**Activation z-score**	**p-value of overlap**
TP53	0.622	transcription regulator	Inhibited	−5.749	6.48E-12
TGM2		enzyme	Inhibited	−4.243	3.64E-02
CDKN1A	−0.485	kinase	Inhibited	−3.548	1.85E-10
KDM5B		transcription regulator	Inhibited	−3.126	3.31E-08
NFkB (complex)		complex	Activated	3.034	3.59E-03
TREM1		other	Activated	3.073	2.18E-05
TP63		transcription regulator	Activated	3.63	6.25E-06
IL1B		cytokine	Activated	3.686	4.13E-01
FOXM1	1.29	transcription regulator	Activated	3.925	5.82E-11
Mek		group	Activated	4.771	7.08E-08
C					
Upstream Regulator	**Log Ratio**	**Molecule Type**	**Predicted Activation State**	**Activation z-score**	**p-value of overlap**
TP53		transcription regulator	Inhibited	−5.126	1.30E-13
CDKN1A	−0.496	kinase	Inhibited	−3.534	5.99E-10
TGM2		enzyme	Inhibited	−3.402	4.25E-02
miR-483-3p		mature microRNA	Inhibited	−3.153	6.49E-03
EGFR		kinase	Activated	3.104	4.43E-03
IL1B		cytokine	Activated	3.281	1.73E-01
TP63		transcription regulator	Activated	3.524	1.48E-09
TREM1		other	Activated	3.845	5.74E-06
FOXM1	1.398	transcription regulator	Activated	4.386	4.18E-16
Mek		group	Activated	4.654	9.72E-08

A - excluding quality-flagged arrays from the analysis. B - applying SVA to batch corrected data. C - ComBat used to correct for batch and quality.

accuracy and length of the original transcript, with both longer than average, and very short transcripts being under- and overrepresented in quality-flagged samples respectively. This is in contrast to the findings by Opitz et al who found that short transcripts were more vulnerable to the perceived effects of degradation, whereas long transcripts were more stable relative to the average length transcript [6]. Interestingly, of the genes that were overrepresented in quality-flagged samples, 70% were small non-protein coding RNAs, including 94 small nucleolar RNAs, and 4 microRNAs, consistent with reports that microRNAs are more robust to RNA degradation compared to mRNA [22], perhaps because they are more thermodynamically stable than mRNAs.

Without excluding any genes, we then compared the orthogonal approaches of either excluding quality-flagged arrays or compensating for RNA degradation at different steps in the analysis. Sample clustering showed that when using ComBat adjustment, quality-flagged samples no longer clustered together. Furthermore, samples tend to segregate more clearly by disease status following adjustment, which suggests that the algorithm is not introducing

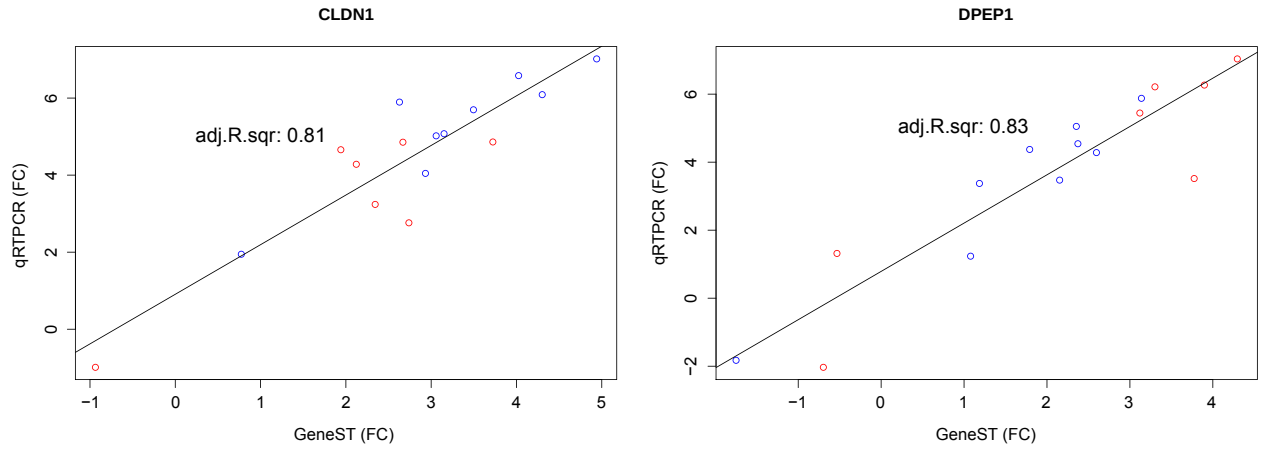


Figure 8 DPEP1 and CLDN1 tumour vs. normal fold change (FC) results for qRT-PCR and microarray results. Samples that were flagged during quality assessment are highlighted in red.

artifacts. It is worth noting that patients 13, 4 and 18 were diagnosed with a hereditary form of CRC (HNPCC) – it is therefore not surprising that the 'normal' samples from these patients form a separate cluster.

Irrespective of sample/array quality, applying compensatory measures for RNA degradation performed at least as well as excluding arrays that were flagged during quality assessment, as judged by gene expression analysis and IPA. At a p-value of 0.01, SVA and Combat detected the highest number of differentially expressed genes between tumour and normal samples and the top four methods applied here had 1117 differentially expressed genes in common. To evaluate the biological plausibility of the genes deemed significantly differentially expressed between tumour and normal samples, we harnessed the results from IPA to show that, in terms of the top scoring biological functions and upstream regulators, there is considerable overlap in the identity and direction of biological activation when comparing analysis methods that either excluded or included quality-flagged arrays. These results suggest that our analysis strategies are biologically sound and not biased by non-biological variance.

The relevance of each method will depend on the downstream application and the proportion of quality-flagged arrays: If a small percentage of arrays are flagged, there might not be much benefit in including them for downstream analysis. However, if a large proportion of the arrays are affected by RNA quality – which is likely to often be the case where the RNA is derived from irreplaceable historical clinical samples – the ability to retain all arrays and to account for these effects in the analysis will be valuable. Here, ComBat may be useful if direct data adjustment is required, e.g. for sample/gene clustering. On the other hand, for analysis of differential expression, especially when the source of non-biological variance is not immediately apparent, SVA may be most useful since it does not require supervision; notably, in our hands SVA was able to identify two surrogate variables which closely corresponded to "batch" and "quality" factors, judged by the grouping of samples. To establish whether the measures used here to compensate for quality-effects are superior to excluding these arrays from the analysis will require a controlled study with known true- and false-positives where the discriminatory power of each method can be objectively investigated. However, the significant overlap observed between the differentially expressed genes identified by the different approaches used here, combined with the considerable overlaps in both biological function and upstream regulators identified by pathway analysis of the resultant data, argues against a simple expansion of false positives when lower quality array data is included in the analyses. The quality assessment and data analysis methods discussed here should in principle be as useful for Affymetrix Exon ST array analysis as well.

Conclusions

In conclusion, array quality measures can be used to set quality thresholds, to provide valuable information that can be used to improve the linear model of differential expression, or to correct expression signal prior to assessing differential expression. We suggest that accounting for known or unknown sources of variation, such as variable RNA integrity and batch, by implementing ComBat or Surrogate Variable Analysis for analysis of differential gene expression enables robust analysis of microarray datasets derived from variable and low quality RNA, thereby extending the range of clinical samples that are suitable for microarray analysis.

Methods

Sample collection and storage

Paired colorectal patient samples (diseased tumour tissue and adjacent healthy gut epithelial tissue) were collected during surgical resection of previously untreated patients at the Groote Schuur Hospital, Cape Town, South Africa. The samples were frozen immediately in liquid nitrogen and stored at -80°C. Ethical consent was obtained (UCT HREC REF 416/2005) and each patient provided written informed consent to donate samples from the tissues left over after surgical resection to subsequent molecular studies.

Sample preparation and quality control

Frozen samples were transitioned to RNA®later-ICE (Ambion), an RNA stabilisation solution, using dry ice to prevent thawing of the tissue at any stage. RNA was extracted using a Dounce homogenizer and the AllPrep DNA/RNA/Protein kit (Qiagen) including DNAse treatment. RNAseZap (Ambion) was used to eliminate RNAse from the work surface, pipettes and glassware. RNA integrity assessment was conducted on an Agilent Bioanalyser 2100.

Quantitative real-time PCR

From a biological perspective, we used the stability of expression of housekeeping genes to investigate the effect of RNA integrity on array- and qRT-PCR performance. Gene candidates were selected from those previously been specifically identified as good reference genes for colorectal cancer [17-21]. Expression stabilities were ranked using the Normfinder algorithm [23] and three genes were selected for use as reference genes. All primers except those for *b2m* [24] were designed using Primer-BLAST - sequences are shown in Table 4. Experiments were performed in triplicate on a Roche LightCycler® 480 Real-Time PCR System in 96-well format. Efficiency was determined for each primer pair using a two-fold dilution series across five points for five patient samples of varying RNA integrity. For each patient, tumour vs. normal fold change was determined based on the method of Antonov et al whereby the Ct of the test gene is normalised by the geometric mean of multiple control genes [9]. Since our efficiencies were quite low in some cases, we adapted the Antonov et al method to include primer efficiency as shown in the equation below:

$$\frac{e^{\Delta Ct(t)}}{\sqrt[n+1]{e_i^{\Delta Ct(i)} \times e_{i+1}^{\Delta Ct(i+1)} \ldots \times e_{i+n}^{\Delta Ct(i+n)}}}$$

where t represents the test gene, e represents efficiency and i represents the control gene(s).

Microarray analysis: Affymetrix HuGene 1.0 ST expression arrays

Thirty-four samples with A260/230 ratios of at least 1.6, RINs of at least 2 and no sign of genomic DNA contamination, were selected for microarray analysis. The samples were amplified from 200ng of total RNA in accordance with the Ambion® WT Expression assay kit and fragmented and end labeled in accordance with the Affymetrix® GeneChip® WT Terminal Labeling protocol. The prepared targets were hybridized overnight to Affymetrix Human Gene 1.0 ST arrays. Following hybridization, the arrays were washed and stained using the GeneChip Fluidics Station 450 and scanned using the GeneChip® Scanner 3000 7G. Arrays were processed in two batches - batch one had 10 arrays, and batch two 24. Individual patient pairs were not split across batches.

Microarray quality assessment and data analysis

Standard Affymetrix quality control was conducted using Expression Console® Software: The quality of cDNA preparation and array hybridisation was assessed using appropriate spike-in controls at each stage.

Raw array quality was investigated at the probe level by 1) the difference between the mean of the perfect match probes and the mean of the background probes for each array as well as 2) the coefficient of variation (CV) across all probes for each array. A threshold for the CV across probes was set as two standard deviations from the mean CV, where the mean was calculated from arrays with RINs > 6. The data was preprocessed in R using the Bioconductor packages frma [25], oligo [26], and the ComBat algorithm for batch correction [12]. Preprocessed data quality

Table 4 Primers used for qRT-PCR analysis

Test genes	Forward primers (5' - 3')	Reverse primers (5' - 3')	Product (bp)
dpep1	GACAACTGGCTGGTGGACA	ACCACACGCTGCCCAAA	74
cldn1	GCTGTCATTGGGGGTGCGAT	GGCAACTAAAATAGCCAGACCTGC	54
Reference genes			
ubc	GGTCGCAGTTCTTGTTTGTGG	CACGAAGATCTGCATTGTCAAG	59
b2m	TGCTGTCTCCATGTTTGATGTATCT	TCTCTGCTCCCCACCTCTAAGT	86
atp5e	CTGGACTCAGCTACATCCGA	GCATCTCTCACTGCTTTTGCAC	55

was assessed using the global normalised, unscaled standard error (GNUSE) [14]. The SE estimates are normalized such that for each probe set, the median standard error across all arrays is equal to 1. Since most genes are not expected to be differentially expressed, boxplots for each array should be centered around 1. Samples with a median GNUSE of greater than 1.25 were flagged for downstream analysis. This threshold is fairly arbitrary and has not been validated for the Gene 1.0 ST platform but roughly equates to having a precision that is on average 25% worse than the average Gene 1.0 ST array [14].

Five comparative methods for analysis of differential expression were individually applied to the preprocessed data: 1) The arrayWeights function in the Bioconductor package limma [27] was used to estimate array quality weights which were then included in the linear model fit; 2) Arrays that were flagged in array quality assessment were excluded from the analysis; 3) The ComBat algorithm [12] for batch correction was applied to directly adjust the data according to quality, where arrays were divided into two categories according to the array quality assessment; 4) "Quality" and "batch" were included as a factors in the linear model together with disease status; 5) Surrogate variable analysis was applied to frma-processed data without any direct adjustment, the output from SVA being incorporated into the linear model fit [13].

To rank genes by evidence for differential expression, the eBayes function in limma was applied to compute moderated t-statistics, moderated F-statistic, and log-odds of differential expression by empirical Bayes shrinkage of the standard errors towards a common value [27]. Next, using the topTable function in limma, p-values were adjusted for multiple hypothesis testing using the Benjamini and Hochberg method [28]. Transcript clusters were annotated in R using the Bioconductor package hugene10sttranscriptcluster.db (Affymetrix Human Gene 1.0-ST Array Transcriptcluster Revision 8 annotation data, assembled using data from public repositories).

The subset of genes differentially affected by RNA quality was similarly obtained, now using array quality for grouping, instead of disease status. Genes with adjusted p-values ≤ 0.05 and FCs ≥ |2| were included in the analysis. Transcript length was obtained for all annotated transcript clusters using the Bioconductor package goseq [29]. Hierarchical clustering with average linkage and Euclidian distance as distance measure was performed in R using the hclust function.

For Ingenuity Pathway Analysis, genes that were found to be significantly differentially expressed for each method (adjusted p-value ≤ 0.01), were used as input for IPAs "Core Analysis". Here, statistically significant over-representation of our listed genes in a given process such as "colorectal tumour" or "infection of embryonic cell lines" is scored by p-value, using the right-tailed Fisher's Exact Test. In the case of upstream regulators, the predicted activation state and activation z-score is based on the direction of fold change values for genes in the input dataset for which an experimentally observed causal relationship has been established. Performance was assessed using the top 10 functions in terms of p-values for each method while taking into account the relevance of the function to colorectal cancer.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

KV carried out the sample preparation and RNA extraction and performed the data analysis. KV and JB conceived and designed the study and wrote the manuscript. Both authors read and approved the final manuscript.

Acknowledgements

We wish to thank the Cancer Association of South Africa, (CANSA) for their financial support of this project. KV wishes to thank the University of Cape Town, the Harry Crossley Foundation, the National Research Foundation (NRF) and the German Academic Exchange Service (DAAD) for their financial support through bursaries. JB thanks the NRF for a South African Research Chair grant. We thank Mr. Ryan Goosen and Ms. Jo McBride for their technical assistance.

Received: 13 September 2012 Accepted: 2 January 2013
Published: 16 January 2013

References

1. Tomita H, Vawter MP, Walsh DM, Evans SJ, Choudary PV, Li J, Overman KM, Atz ME, Myers RM, Jones EG, Watson SJ, Akil H, Bunney WE: **Effect of agonal and postmortem factors on gene expression profile: quality control in microarray analyses of postmortem human brain.** *Biological Psychiatry* 2004, **55**(4):346–352.
2. Mengual L, Burset M, Ars E, Ribal MJ, Lozano JJ, Minana B, Sumoy L, Alcaraz A: **Partially Degraded RNA from Bladder Washing is a Suitable Sample for Studying Gene Expression Profiles in Bladder Cancer.** *European Urology* 2006, **50:**1347–1356.
3. Linton KM, Hey Y, Saunders E, Jeziorska M, Denton J, Wilson CL, Swindell R, Dibben S, Miller CJ, Pepper SD, Radford JA, Freemont AJ: **Acquisition of biologically relevant gene expression data by Affymetrix microarray analysis of archival formalin-fixed paraffin-embedded tumours.** *British Journal of Cancer* 2008, **98:**1403–1414.
4. Linton K, Hey Y, Dibben S, Miller C, Freemont A, Radford J, Pepper S: **Methods comparison for high-resolution transcriptional analysis of archival material on Affymetrix Plus 2.0 and Exon 1.0 microarrays.** *BioTechniques* 2009, **47:**587–596.
5. April C, Klotzle B, Royce T, Wickham-garcia E, Boyaniwsky T, Izzo J, Cox D, Jones W, Rubio R, Holton K, Matulonis U, Quackenbush J, Fan J: **Whole-Genome Gene Expression Profiling of Formalin-Fixed, Paraffin-Embedded Tissue Samples.** *PloS one* 2009, **4**(12):1–10.
6. Opitz L, Salinas-riester G, Grade M, Jung K, Jo P, Emons G, Ghadimi BM, Beiÿ barth T, Gaedcke J: **Impact of RNA degradation on gene expression profiling.** *BMC Medical Genomics* 2010, **3**(36):1–14.
7. Fleige S, Pfaffl MW: **RNA integrity and the effect on the real-time qRT-PCR performance.** *Molecular aspects of medicine* 2006, **27:**126–139.
8. Lassmann S, Kreutz C, Schoepflin A, Hopt U, Timmer J, Werner M: **A novel approach for reliable microarray analysis of microdissected tumor cells from formalin-fixed and paraffin-embedded colorectal cancer resection specimens.** *Journal of molecular medicine* 2009, **87:**211–224.
9. Antonov J, Goldstein DR, Oberli A, Baltzer A, Pirotta M, Fleischmann A, Altermatt HJ, Jaggi R: **Reliable gene expression measurements from degraded RNA by quantitative real-time PCR depend on short amplicons and a proper normalization.** *Laboratory Investigation* 2005, **85:**1040–1050.
10. Binder H, Krohn K, Preibisch S: **"Hook"-calibration of GeneChip-microarrays: chip characteristics and expression measures.** *Algorithms for molecular biology* 2008, **3:**11.

11. Chow ML, Winn ME, Li HR, April C, Wynshaw-Boris A, Fan JB, Fu X, Courchesne E, Schork NJ: **Preprocessing and quality control strategies for Illumina DASL assay-based brain gene expression studies with semi-degraded samples.** *Frontiers in Genetics* 2008, **3:**11.
12. Johnson WE, Li C: **Adjusting batch effects in microarray expression data using empirical Bayes methods.** *Biostatistics* 2007, **8:**118–127.
13. Leek JT, Storey JD: **Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis.** *PLoS Genetics* 2007, **3**(9):1724–1735.
14. McCall MN, Murakami PN, Lukk M, Huber W, Irizarry R: **Assessing affymetrix GeneChip microarray quality.** *BMC bioinformatics* 2011, **12:**137.
15. Gibbons FD, Roth FP: **Judging the Quality of Gene Expression-Based Clustering Methods Using Gene Annotation.** *Genome Research* 2002, **12:**1574–1581.
16. Dalman MR, Anthony D, Gayathri N, Zhong-Hui D: **Fold change and p-value cutoffs significantly alter microarray interpretations.** *BMC Bioinformatics* 2012, **13**(Suppl 2):S11.
17. Salazar R, Roepman P, Capella G, Moreno V, Simon I, Dreezen C, Lopez-Doriga A, Santos C, Marijnen C, Westerga J, Bruin S, Kerr D, Kuppen P, van de Velde C, Morreau H, Van Velthuysen L, Glas AM, Van't Veer LJ, Tollenaar R: **Gene expression signature to improve prognosis prediction of stage II and III colorectal cancer.** *Journal of clinical oncology* 2011, **29:**17–24.
18. O'Connell MJ, Lavery I, Yothers G, Paik S, Clark-Langone KM, Lopatin M, Watson D, Baehner FL, Shak S, Baker J, Cowens JW, Wolmark N: **Relationship between tumor gene expression and recurrence in four independent studies of patients with stage II/III colon cancer treated with surgery alone or surgery plus adjuvant fluorouracil plus leucovorin.** *Journal of clinical oncology* 2010, **28**(25):3937–3944.
19. Dydensborg AB, Herring E, Auclair J, Tremblay E, Beaulieu JF: **Normalizing genes for quantitative RT-PCR in differentiating human intestinal epithelial cells and adenocarcinomas of the colon.** *American Journal of Gastrointestinal and Liver Physiology* 2006, **290:**G1067–G1074.
20. Rubie C, Kempf K, Hans J, Su T, Tilton B, Georg T, Brittner B, Ludwig B, Schilling M: **Housekeeping gene variability in normal and cancerous colorectal, pancreatic, esophageal, gastric and hepatic tissues.** *Molecular and cellular probes* 2005, **19:**101–109.
21. Kheirelseid EH, Chang KH, Newell J, Kerin MJ, Miller N: **Identification of endogenous control genes for normalisation of real-time quantitative PCR data in colorectal cancer.** *BMC molecular biology* 2010, **11:**12.
22. Jung M, Schaefer A, Steiner I, Kempkensteffen C, Stephan C, Erbersdobler A, Jung K: **Robust microRNA stability in degraded RNA preparations from human tissue and cell samples.** *Clinical chemistry* 2010, **56**(6):998–1006.
23. 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 research* 2004, **64**(15):5245–5250.
24. Vandesompele J, Preter KD, Poppe B, Roy NV, Paepe AD: **Accurate normalization of real-time quantitative RT -PCR data by geometric averaging of multiple internal control genes.** *Genome biology* 2002, **3**(7):1–12.
25. Mccall MN, Bolstad BM, Irizarry RA: **Frozen robust multiarray analysis (fRMA).** *Biostatistics* 2010, **11**(2):242–253.
26. Carvalho B, Irizarry RA, Scharpf RB, Carey VJ: **Processing and Analyzing Affymetrix SNP Chips with Bioconductor.** *Stat Biosci* 2009, **1:**160–180.
27. Smyth GK: **Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments.** *Statistical Applications in Genetics and Molecular Biology* 2004, **3:**3.
28. Benjamini Y, Hochberg Y: **Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing.** *Journal of the Royal Statistical Society* 1995, **57:**289–300.
29. Young MD, Wakefield MJ, Smyth GK, Oshlack A: **Gene ontology analysis for RNA-seq: accounting for selection bias.** *Genome Biology* 2010, **11:**R14.

doi:10.1186/1471-2164-14-14
Cite this article as: Viljoen and Blackburn: **Quality assessment and data handling methods for Affymetrix Gene 1.0 ST arrays with variable RNA integrity.** *BMC Genomics* 2013 **14**:14.