MICROBIAL ETIOLOGY OF INFLAMMATORY BOWEL DISEASE:

MICROBIAL DIVERSITY AND THE ROLE OF *ESCHERICHIA COLI*

BY

SHADI SEPEHRI

A Thesis

Submitted to the Faculty of Graduate Studies

In Partial Fulfillment of the Requirements for the Degree of

DOCTOR OF PHILOSOPHY

Department of Medical Microbiology and Infectious Diseases

University of Manitoba

Winnipeg, Manitoba, Canada

ABSTRACT

Inflammatory bowel disease (IBD), comprises Crohn's disease (CD) and ulcerative colitis (UC), and is a chronic relapsing inflammation of gastrointestinal tract without any known cause or cure. Currently, it is accepted that IBD is a result of a dysfunctional immune response to commensal bacteria in a genetically susceptible host, and that environmental factors can trigger the onset or reactivation of the disease. This thesis considers the possibility of a specific pathogenic agent as well as an imbalance in the composition of the normal microflora in the pathogenesis of IBD.

Gut biopsy tissues were taken from a population-based case-control tissue bank held at the University of Manitoba. Automated ribosomal intergenic spacer analysis (ARISA) and terminal restriction fragment length polymorphisms (T-RFLP) were employed to assess the diversity of gut microbiota. The phylogenetic, virulence and biochemical characteristics of *Escherichia coli* isolated from IBD biopsies were examined using multi-locus sequence typing (MLST), DNA microarray technology and API 20E system.

Utilizing ARISA and T-RFLP, a remarkable increase in the order of unclassified *Clostridia* was detected in inflamed tissues, particularly in CD patients ($P < 0.05$). Moreover, species richness and diversity were the highest in non-inflamed IBD biopsies. Culture-based quantification detected a significantly higher number of *E. coli* in IBD tissues ($P < 0.05$). Phylogenetic analysis revealed the tendency of *E. coli* isolated from IBD patients to be grouped into separate clonal clusters based on their allelic profiles ($P = 0.02$). A link was detected between uropathogenic *E. coli* (UPEC) CFT073 and strains

isolated from IBD, with regards to gene distribution and virulence, using microarray technology. Amino acid substitutions N91S and S99N in FimH, the adhesive subunit of *E. coli* type I fimbria, were significantly associated to IBD ($P < 0.05$).

This study demonstrated an increase in the microbial diversity of non-inflamed IBD tissues and suggested a recruitment phase of bacterial adherence and colonization, before the inflammation sets in. Furthermore, *E. coli* isolated from IBD tissues were distinct from commensal strains in both clonal and virulence characteristics and shared remarkable traits with extraintestinal pathogenic *E. coli*. Features involved in bacterial adhesion to epithelial cells may hold the key to *E. coli* pathogenesis in IBD.

ACKNOWLEDGEMENTS

DEDICATION

To my dearly loved parents and brothers, for their endless support and love. You have always been there for me and belived in me.

And to my loved and adored husband, and my best friend, Ehsan, for all his patience and encouragements. You inspired me to do it all.

FOREWORDS

Parts of this dissertation have been published in the form of three articles in peer reviewed journals. Manuscript fourth has been submitted to the journal of "Inflammatory Bowel Diseases" and is under revision. The fifth manuscript has yet to be submitted for publication.

1. Kotlowski R, Bernstein CN, **Sepehri S** and Krause DO. **2007.** High prevalence of *Escherichia coli* belonging to the B2+D phylogenetic group in inflammatory bowel disease. Gut;56;669-675.

  [URL: http://gut.bmj.com/content/56/5/669.abstract]

2. Sepehri S, Kotlowski R, Bernstein CN, and Krause DO. **2007**. Microbial diversity of inflamed and non-inflamed gut biopsy tissues in inflammatory bowel disease. Inflamm Bowel Dis;13:675-683.

  [URL: http://www3.interscience.wiley.com/journal/114099828/abstract]

3. Sepehri S, Kotlowski R, Bernstein CN, and Krause DO. **2009**. Phylogenetic analysis of inflammatory bowl disease associated *Escherichia coli* and the FimH virulence determinant. Inflamm Bowel Dis;15:1737–1745.

  [URL: http://www3.interscience.wiley.com/journal/122400027/abstract]

4. Sepehri S, Bernstein CN, Coombes BK, Karmali M, Ziebell K, and Krause DO. Virulence characterization of the *Escherichia coli* isolated from gut biopsies of newly diagnosed patients with inflammatory bowel disease. Inflamm Bowel Dis. Submitted July 17, 09. (Accepted with revisions)

5. Sepehri S, Kotlowski R, Bernstein CN, and Krause DO. Comparative analysis of the gene content of *Escherichia coli* isolated from inflammatory bowel disease gut biopsies using microarray technology. (To be submitted)

LIST OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

A = alanine

AAF = aggregative adherence fimbrieae

ADH = arginine dihydrolase

A/E = attaching and effacing

AFLP = amplified fragment length polymorphism

Ag43 = antigen 43

AIEC = adherent invasive *Escherichia coli*

AMY = amygdalin

APEC = avian pathogenic *Escherichia coli*

ARA = arabinose

ARISA = automated ribosomal intergenic spacer analysis

ASCA = anti-Saccharomyces cerevisiae antibodies

ATG16L1 = autophagy-related 16-like 1

bp = base pair

CD = Crohn's disease

CEACAM = carcinoembryonic antigen-related cellular adhesin molecule

CIT = citrate

CRP = C-reactive protein

D = aspartic acid

DAEC = diffusely adherent *Escherichia coli*

DAF = decay-accelerating factor

DGGE = denaturing gradient gel electrophoresis

DLV = double locus variants

DNA = deoxyribonucleic acid

DR = direct repeat

E = glutamic acid

E. coli = *Escherichia coli*

EAEC = enteroaggregative *Escherichia coli*

EAST1 = enteroaggregative heat-stable enterotoxin

ECM1 = extracellular matrix protein 1

EIEC = enteroinvasive *Escherichia coli*

EHEC = enterohemorrhagic *Escherichia coli*

EPEC = enteropathogenic *Escherichia coli*

ESR = erythrocyte sedimentation rate

ETEC = enterotoxigenic *Escherichia coli*

ExPEC = extraintestinal pathogenic *Escherichia coli*

F = phenylalanine

FREQ = frequency

G + C = guanine and cytosine

GEL = gelatinase

GI = gastrointestinal

GLU = glucose

H = histidine

IBD = inflammatory bowel disease

IL23R = interleukin 23 receptor

IND = indole

INO = inositol

IRGM = immunity-related guanosine triphosphates

IS = insertion sequence

kb = kilo base pair

LB = Luria Bertani

LDC = lysine decarboxylase

LEE = locus of enterocyte effacement

LTs = heat-labile enterotoxins

MAN = manitol

MAP = *Mycobacterium avium* subspecies *paratuberculosis*

MEL = melibiose

MiCA = microbial community analysis

MLLE = multi-locus enzyme electrophoresis

MLST = multi-locus sequence typing

N = asparagine

ODC = ornithine decarboxylase

ONPG = ortho nitrophenyl-ß-D-galactopyranosidase

OTU = operational taxonomic unit

P-ANCA = perinuclear anti-neutrophilic cytoplasmic antibody

PAI = pathogenicity island

PCR = polymerase chain reaction

PSC = primary sclerosing cholangitis

PTGER4 = prostaglandin E receptor 4

R = arginine

RAPD = random amplified polymorphic DNA

rDNA = ribosomal DNA

RDP = ribosomal database project

RFLP = restriction fragment length polymorphism

RHA = rhamnose

RISA = ribosomal intergenic spacer analysis

RNA = ribonucleic acid

S = serine

SAC = sucrose

SAT = satellites

ShET1 = Shigella enterotoxin 1

SLV = single locus variants

SOR = sorbitol

SPATE = serine protease autotransporter

ST = sequence type

STAT3 = signal transducer and activator of transcription 3

STs = heat-stable enterotoxins

Stx = shiga toxin

TDA = tryptophane deaminase

TGGE = temperature gradient gel electrophoresis

Tir = translocated intimin receptor

TNF-α = tumor necrosis factor alpha

T-RFLP = terminal restriction fragment length polymorphism

tRNA = transfer RNA

TTGE = temporal temperature gradient gel electrophoresis

UC = ulcerative colitis

UPEC = uropathogenic *Escherichia coli*

URE = urease

V = valine

VP = acetone production

VT = verocytotoxin

Y = tyrosine

INTRODUCTION

Crohn's disease (CD) and ulcerative colitis (UC) are two main types of inflammatory bowel disease (IBD), and are recognized as important causes of chronic gastrointestinal inflammation in children and adults. Although IBD occurs worldwide, it is more common in Western countries. Canada has one of the highest incidence and prevalence rates of IBD around the world with about 200,000 Canadians living with the disease (Bernstein et al. 2006). Both CD and UD are chronic and relapsing disorders and cause substantial reduction in the state of health and well-being of those affected. Moreover, because the peak for the onset of IBD is in early adulthood, when patients are in their prime working years, IBD does not only affect patients physically, but also socially, emotionally and even financially. IBD also has a big burden on the health-care system since patients have to be hospitalized on a regular basis (Longobardi et al. 2003).

While remarkable advances have improved our understanding of IBD, the causative etiology of the disease is still unknown. There are several theories about the pathogenesis of IBD. Currently, the most accepted hypothesis is that IBD is a result of dysfunctional immune response to commensal bacteria in a genetically predisposed host, and that environmental factors can trigger the onset or reactivation of the disease (Sartor 2006). This hypothesis considers the possibility of a specific pathogenic agent as well as an imbalance in the composition of the normal microflora.

The purpose of this study was to evaluate the microbial composition of intestinal mucosa in IBD patients and compare it to healthy controls; and to possibly identify and

characterize specific microorganism in association to IBD. In particular, the role of *E. coli* in the pathogenesis of IBD was tested with regards to its phylogenetic, virulence, biochemical features and genomic content. To eliminate the effect of IBD treatment and inflammation process on the gut microbial composition, *E. coli* isolated from newly diagnosed IBD patients have also been studied.

LITERATURE REVIEW

1. Overview of inflammatory bowel disease

IBD, comprising CD and UC, is the chronic life-long inflammation of gastrointestinal tract in children and adults. IBD is most commonly diagnosed in the 3rd and 4th decades of life, although about 20% of all IBD patients develop their symptoms during childhood (Hendrickson et al. 2002). IBD occurs worldwide but it is more common in North America and Western Europe. To date, there is no known cause or cure for the disease.

1.1. Definition and clinical features of Crohn's disease and ulcerative colitis

In UC, the inflammation is limited to the mucosa and consists of continuous involvement of the tissue with variable degrees of severity from ulcer and edema to hemorrhage. UC inflammation is restricted to colon and rectum (Hendrickson et al. 2002). CD in contrast, can involve any part of the digestive tract, from oropharynx to anus. Inflamed segments in CD are separated by normal bowel creating "skip areas." In CD, inflammation is transmural, often resulting in fistula formation. The most common location for CD is ileocecal region, followed by the terminal ileum (Hendrickson et al. 2002).

UC is mostly presented by lower abdominal cramps, accompanied by the presence of blood and mucus in the stool. Pain is present in the left lower quadrant of abdomen

with distal disease and extends to the entire abdomen with pancolitis (Hendrickson et al. 2002). The presentation of CD however is more subtle, and the clinical symptoms depend on the location, extent, and the severity of involvement. Clinical manifestations of CD range from general symptoms like anorexia, weight loss, and growth failure (in children) to more specific gastrointestinal symptoms such as nausea, dysphagia, diarrhea abdominal pain and perianal diseases such as fissures and fistulas (Hendrickson et al. 2002). In either disease, symptoms may be continuous or patients may go through relapse (flare-up) and remission (symptom-free) periods.

About 40% of all IBD patients experiment extraintestinal manifestations of IBD which are significant causes of morbidity and distress in the patients. Extraintestinal manifestations of IBD can almost affect every organ and include but are not limited to peripheral and axial arthritis, primary sclerosing cholangitis (PSC), psoriasis, and erythema nodosum (Williams et al. 2008).

1.2. Epidemiology

IBD has long been considered a disease that predominantly affects Western populations and its incidence and prevalence rates for developing countries used to be much lower in comparison. The recognition and early reports of the disease were originally from North America and Scandinavia and in white patients (Goh et al. 2009). More recent data however, have shown significantly higher rates of IBD in developing countries, and the trend of changes is very similar to what Western countries experienced about half a century ago. These epidemiological changes seem to occur in parallel with the modernization and westernization taking place in the developing countries (Goh et al.

2009). Canada has one of the highest incidence rate of IBD in the world (Gismera et al. 2008).

1.3. Etiology

Several theories have been suggested regarding the etiology of IBD including genetic predisposing factors, environmental triggers, microbial pathogens and dysfunctional/dysregulated immune response.

1.3.1. Genetic background

Epidemiological and family studies demonstrate that genetic factors play a role in the susceptibility to IBD. However, this genetic predisposition is complex and could not be explained by a single gene model. It is hypothesized that UC and CD are heterogeneous polygenic disorders and that the disease phenotype is determined by factors such as the interaction between allelic variants and environmental influences (Hanauer 2006).

Some racial groups are more susceptible to IBD than others (Goh et al. 2009). The fact that higher rates of IBD occur in Ashkenazic Jews and Caucasians suggests that genetic factors may play a role in the pathogenesis of IBD (Hanauer 2006). It has been suggested that environmental factors such as dietry and socioeconomic aspects may influence genetic factors in the pathogenesis of IBD in the Ashkenazi Jewish group.

There is an increase in the prevalence of IBD in first- and second-degree relatives with a higher risk among siblings. One American study showed that 35% of CD patients and 29% of UC patients had a positive family history of IBD (Farmer et al. 1980). Across all studies, the risk for a sibling of a CD or UC patient developing the disease compared

to the general population is 20-35 fold and 8-15 fold respectively (Zhang et al. 2008). Studies on twins indicated higher rate of concordance in monozygotic versus dizygotic twins for both CD and UC; however, the genetic influence was higher in CD (Hanauer 2006).

Mutation in the *NOD2* (*CARD15*) gene locus of chromosome 16 was the first and to date the most noted genetic association with IBD. Individuals who are homozygous for *NOD2* variants have a > 20-fold increased risk of developing CD. The defect in *NOD2* is present in 15-25% of all CD patients (Hanauer 2006). However, researchers form China and Japan were unable to associate *NOD2* variants to CD in their populations, highlighting significant ethnic heterogeneity of IBD (Knight et al. 2008; Zhang et al. 2008; Brant 2009). Today, more than 30 IBD susceptibility genes have been identified including genes specific for CD (*NOD2*, *ATG16L1*, *IRGM*, *PTGER4*), genes specific for UC (*ECM1*) and genes that are common in both CD and UC (*IL23R*, *STAT3*) (Xavier et al. 2007; Sollid et al. 2008; Zhang et al. 2008; Kaser et al. 2010).

1.3.2. Environmental factors

The changing global incidence patterns for IBD support the role of environmental factors in the pathogenesis of the disease (Ali et al. 2008). This critical role has been demonstrated where second-generation migrants had been studied. A recent study from British Columbia, Canada showed a higher prevalence of both UC and CD in the South Asian children who are second-generation migrants compared with other ethnic groups (Pinsk et al. 2007). Similarly studies from Israel have shown that, while originally the prevalence of UC and CD was highest amongst Western-born Jews compared to Jews from Asia and Africa, this difference has now been reduced and the IBD rates are now

comparable between the two groups (Odes et al. 1994; Niv et al. 2000). Results of such studies could not be explained by genetic and ethnic differenced and highlight the important role of environmental factors in the pathogenesis of IBD.

1.3.2.1. Geographic

IBD is more prevalent in North America and Scandinavia than southern European countries and Asia and the prevalence of IBD in northern states of America is higher than westerns and southern states (Hanauer 2006; Economou et al. 2008). This can be explained not only by differences in climate but also by higher frequency of urban communities compare to rural areas (Hanauer 2006)

1.3.2.2. Socioeconomics

IBD is a disease of cleanliness and it is more frequent in higher socioeconomic groups. It is postulated that improved hygiene alters the microbiota by decreasing the exposure to certain bacteria, particularly in the first few years of life (Hanauer 2006). The "hygiene hypothesis" suggests that increased hygiene in infancy and childhood leads to decrease exposure to microorganisms (both pathogenic and non-pathogenic) and results in aberrant development of immune system, leading to IBD and other immune-mediated diseases such as asthma (Ali et al. 2008; Mazmanian et al. 2008). The explanation for hegiene hypothesis relies on the role of regulatory T cells in maintaining immune system homeostasis. Regulatory T cells (also known as suppressor T cells) are a specialized population of immune cells that act to suppress activation of the immune system and provide tolerance to self-antigens. The developing immune system requires stimuli from microorganisms and environmental triggers in order to adequately develop regulatory T

cells. Lack of exposure to such stimuli, particularly during the first few years of life, could result in insufficient development of regulatory T cells and impared repression of immune response towards commensal stimuli (Hill et al. 2010). It has been also proposed that the reduce exposure to intestinal helminthes in developed countries may be responsible for the rise in IBD incidence (Weinstock et al. 2002). On recents study suggest that symbiotic bacteria residing in the mammalian gut produce molecules that mediate healthy immune responses and protect the host from inflammatory disease (Mazmanian et al. 2008).

1.3.2.3. Life style

IBD used to be the disease of developed regions of the world including Western Europe and North America (Sonnenberg 1990b; Sonnenberg 1990a; Hanauer 2006). Interestingly, the gap has been closing and an increase in IBD incidence is being noted in developing countries as well as in second generation migrants to developed countries. This has been postulated to be the result of "Westernization" of lifestyle, such as changes in diet, stress, and exposure to pollutions and industrial chemicals (Hanauer 2006).

Diet has been one of the environmental factors implicated in the etiopathogenesis of IBD. Several studies have suggested that a change to a more Westernized diet (frequent fast-food intake) may be responsible for increase in IBD rates in developing countries (Hanauer 2006; Goh et al. 2009). One epidemiological study in Japan indicated that the increased intake of meat and dairy products paralleled the increasing trend of UC in Japan (Kitahora et al. 1995). In another study from Japan, the consumption of sweets and a high fat diet has been associated with UC and CD (Sakamoto et al. 2005; Asakura

et al. 2008). A recent study has also shown that an increased intake of red meat and alcohol may contribute to UC flare-ups (Jowett et al. 2004).

Both UC and CD are more prevalent in white-collar compared to blue-collar occupations. One study suggests that employment involving outdoor air and physical activity is protective against IBD (Sonnenberg 1990a). Higher mortality from IBD has also been reported in managerial, clerical and sales positions compared to lower mortality among farmers and construction workers (Sonnenberg 1990b).

1.3.3. Infection by microbial pathogens

The clinical presentation of IBD mimics infections by some known enteric pathogens. For instance, CD can appear like gastroenteritis due to *Yersinia* or *Mycobacterium tuberculosis* whereas UC resembles *Shigella* and *Campylobacter* infections (Hendrickson et al. 2002).

During recent years, several microorganisms have been postulated as infectious explanations for IBD, the most notable being *Mycobacterium avium* subspecies *paratuberculosis* (MAP), the causative agent of Johne's disease in ruminants (Bernstein et al. 2008). Despite considerable efforts, no one particular organism has been yet identified as etiologic agent of IBD. However, the possibility still exists that as yet unidentified and difficult to detect microorganism be responsible for the pathogenesis of IBD (Hendrickson et al. 2002).

1.3.4. Dysfunctional immune response

IBD have long been considered to be a disorder of immune dysfunction. Understanding the nature of disregulated relationship between the gut immune system

and intestinal microflora is essential in the pathogenesis of IBD (Robertson et al. 2008). In fact, variants of several genes involved in the innate immunity pathway have been recognized as IBD susceptibility factors (Kaser et al. 2010).

Although microbial pathogens had been originally thought to be the cause of IBD, today it is generally accepted that the constant and pathogenic stimulation of gut mucosa by commensal flora is the cause of chronic intestinal damage in IBD (Xavier et al. 2007; Sartor 2008). The fact that rodent models of IBD do not show clinical and histological signs of inflammation when they are kept in a germ-free environment, implicates the role of commensal intestinal bacteria in the pathogenesis of IBD (Hendrickson et al. 2002; Sartor 2006; Mazmanian et al. 2008; Kaser et al. 2010). One hypothesis is that in IBD there is loss of tolerance to enteric commensal bacteria which leads to the chronic inflammation of the gastrointestinal tract (Sartor 2006). In other word, in IBD petients, gut-associated lymphoid tissue is unable to control the basal level of information (Frank et al. 2007; Hill et al. 2010). Defective mucosal barrier function, inability of immune system in microbial clearance, and abnormal immunoregulation in forms of defective immunosupression or hyperactive immune stimulation are other suggested mechanisms of immune dysfunction in IBD (Sartor 2006; Knight et al. 2008; Mazmanian et al. 2008; Sartor 2008; Kaser et al. 2010).

1.4. Diagnosis

Diagnosis of IBD continues to be based on history and physical exam, laboratory findings and endoscopic and histological evidences (Ali et al. 2008). One crucial step in the diagnosis of IBD is to exclude gastroenteritis due to enteric pathogens that mimic IBD manifestations. These include *Salmonella*, *Shigella*, *E. coli* O157:H7, *Clostridium*

difficile, *Giardia lamblia*, and *Entamoeba histolytica*. Patients with IBD may have concomitant infections at first, but their symptoms fail to resolve with treatment of the infectious agent.

Once enteric infections are ruled out, further work-up is initiated. IBD screening tests include a complete blood count, inflammatory markers, metabolic profile, and serologic tests. Anemia, leukocytosis (increase in white blood cells), elevation in erythrocyte sedimentation rate (ESR) and C-reactive protein (CRP) levels, hypoalbuminemia and low serum iron levels are indications of IBD (Hendrickson et al. 2002). Serological tests are used to support the diagnosis of IBD, to distinguish between UC and CD, and to predict the disease behavior (Dotan 2007). Anti-Saccharomyces cerevisiae antibodies (ASCA) is detected in about 48-80% of CD patients but also exist in 10-15% of UC patients and healthy controls (Moscandrew et al. 2009). Although the sensitivity of ASCA is variable, when present, it is highly specific (80-95%). However, the prevalence of ASCA seropositivity is low in patients with isolated colonic CD. This is a diagnosis drawback for this marker since distinguishing this type of CD from UC is difficult (Moscandrew et al. 2009). Perinuclear anti-neutrophilic cytoplasmic antibody (pANCA) is detected in 50-85% of UC patients and about 5-28% of CD patients and non-IBD controls. Combined with a negative ASCA, the sensitivity and specificity of a positive pANCA is 44-57% and 97% respectively (Hendrickson et al. 2002; Dotan 2007). ASCA positivity or high ASCA titers are associated with more progressive CD with more complications such as perforating and stricturing. pANCA positive UC patients are more resistant to the treatment and have a higher chance of developing pouchitis (Dotan 2007). It should be noted that the serology results must always be interpreted with cautious

given the overlap of positive values with other gastrointestinal disorders such as celiac disease (Moscandrew et al. 2009).

Once the diagnosis of IBD is considered, endoscopic examination with biopsies is indicated to establish the final diagnosis. In most cases the clinical endoscopic diagnosis is provisional which can be confirmed by histological examination of the biopsies.

1.5. Treatment

To date there is no cure for IBD. Treatments are mostly supportive and aim to reduce the symptoms and complications of the disease which ultimately leads to lesser need for hospitalization and surgery (Sandborn 2008). Sergical resection usually cures UC but CD often returs following the removal of affected intestine. The chance that an IBD patient responds to medications and remain in remission for more than one year is as low as 25% (Mutlu et al. 2008).

Approximately 40% of IBD patients believe that diet can control their symptoms, while about 80% believe that diet is an important component in the overall management of their disease. A diet that is rich in antioxidants, omega-3 fatty acids, vitamin A, vitamin D and Zinc and is limited in processed sugar, saturated fat, and spices can subside IBD symptoms and improve the patients' quality of life (Bernstein et al. 2008; Mutlu et al. 2008).

Medical management of IBD includes oral 5-aminosalicylic acid medications (sulfasalazine, mesalamine, and olsalazine) and corticosteroids (1mg/kg/day) for mild to moderate cases. Immunosupressive therapy with cyclosporine and tacrolimus are used to induce remission; however, 6-mercaptopurine and azathioprine are suggested for maintaining long-term remission (Hendrickson et al. 2002).

The use of antibiotics metronidazole and ciprofloxacin proved to be useful in the treatment of pouchitis as well as mild to moderate CD (Hendrickson et al. 2002)

Tumor necrosis factor alpha (TNF-α) promotes the inflammatory response in various diseases including CD and UC (Rutgeerts et al. 2009). An increase in the production of inflammatory cytokines, especially TNF-α has been reported in the mucosa of IBD patients (Hendrickson et al. 2002). Anti-TNF biologic therapy with infliximab, adalimumab and certilizumab is reserved for the patients who fail other therapeutic approaches. These biological agents have also proven effective in achieving mucosal healing and maintenance of steroid-free remission (Sandborn 2008; Rutgeerts et al. 2009).

1.6. Prognosis

CD is associated with a 54-78% rate of relapse within 2 years following the initial diagnosis. Many CD patients experiment complications and up to 75% of them have to undergo surgery at some point in their life (Hanauer 2006). However, the incidence of recurrence is high with a 34-86% relapse rate in the 3 years following the surgery (Hendrickson et al. 2002). Patients with CD have a 40-fold increased risk of developing small bowel malignancy, however, these tumors are very rare and the absolute rate of such malignancies is still far lower than colon cancer. The risk of colorectal cancer in CD patients with colon involvement is just slightly higher than general population (Ahmadi et al. 2009).

Up to 40% of patients with severe UC will require colectomy. UC patients are also at the risk of developing colorectal cancer. Disease duration, age at the time of diagnosis, anatomic extent of the disease, severity of inflammation, family history of colon cancer

and concomitant PSC are risk factors for developing colorectal cancer (Ahmadi et al. 2009). Described in terms of relative risk, UC increases the risk of colorectal cancer by 6-fold compared to general population (Velayos 2008).

2. Microbes and inflammatory bowel disease

2.1. Gut normal flora

The human digestive tract contains an organized and complex microbial ecosystem known as microflora. These bacteria, viruses and fungi make up most of the flora in colon and about 60% of the mass of feces. Colon has the greatest number of bacteria as well as the most different number of the species (Guarner 2005; Sartor 2008). Gut microflora in human consists of over 10^{12-14} obligate and facultative anaerobes from over 1,800 genera and 36,000 species (Guarner 2005; Manson et al. 2008; Sartor 2008). However, it has been reported that only 11 out of 55 known bacterial divisions are represented in gut (Manson et al. 2008) and that more than 98% of the gut microflora belong to only four bacterial phyla: *Firmicutes* (64%), *Bacteroidetes* (23%), *Proteobacteria* (8%), and *Actinobacteria* (3%) (Eckburg et al. 2005; Ley et al. 2006; Frank et al. 2007; Sartor 2008).

Many studies have demonstrated that intestinal bacteria contribute to the health of human host. The gut microflora has an essential role in preventing the colonization of harmful bacteria as well as developing and stimulating the immune response. Commensal bacteria compete with harmful bacteria for nutrients and adhesion sites and therefore can prevent the growth of pathogenic species (Hooper et al. 2002). Commensal bacteria can also produce bacteriocins which ultimately results in killing the pathogenic bacteria.

Microbiota is also critical in the development of lymphoid structures such as Peyer's patches and in the phenomenon of tolerance in which the immune system is unresponsive to the normal microbiota (Sartor 2008).

2.1.1. Commensal *Escherichia coli*

Escherichia coli typically colonize the gastrointestinal tract of human infants within a few hours after birth and usually coexist within its human host for decades. These commensal *E. coli* strains rarely cause disease except in immunocompromised hosts or where the normal gastrointestinal barriers are breached. The niche of commensal *E. coli* is the mucous layer of the mammalian colon. *E. coli* is the most abundant facultative anaerobe of the human gut microflora (Kaper et al. 2004).

Still, there are several *E. coli* strains that have acquired specific virulence determinants and have the ability to adapt to new niches and cause a broad spectrum of diseases in healthy individuals (Kaper et al. 2004). These pathogenic clones will be later discussed under intestinal and extraintestinal pathogenic *E. coli* (sections 3.3 and 3.4 of this chapter).

2.2. Gut microbial composition in IBD

One IBD theory which has been gaining more attention in recent years is that a shift in the normal balance of microflora, or dysbiosis, leads to the colonization of the gut by bacteria that are more proinflammatory, resulting in the chronic gut inflammation associated with IBD (Salzman et al. 2008; Tannock 2008). Although dysbiosis, an imbalance between protective and harmful intestinal flora, has been reported in several

different IBD studies, it is still not clear whether dysbiosis is actually the cause of IBD or just a secondary phenomenon of IBD (Tamboli et al. 2004).

Intestinal microflora in IBD has been analyzed repeatedly. Although methodologies and results may differ, some generalizations are possible. Earlier studies using conventional culturing methods often showed an increase in the population of anaerobes, particularly Gram negative anaerobes such as *Bacteroides* in IBD. Other studies found increase in the number of *Clostridia* and *Enterobacteriaceae* in both CD and UC (Tamboli et al. 2004). Several studies also noted a reduction in beneficial bacteria, such as *Bifidobacteria* and *Lactobacillus* in IBD (Tamboli et al. 2004; Sartor 2006).

Imbalances in the composition of intestinal microbiota of IBD patients have been also implicated in numerous molecular studies. Fujita et al. reported greater numbers of *E. coli* and *Bacteroides vulgatus* in association to IBD mucosa using real-time PCR (Fujita et al. 2002). The increase in the abundance of *Bacteroides vulgatus* and *Bacteroides ovatus* was also reperted in a recent study on identical twins with CD (Dicksved et al. 2008). In one study by Milonaki et al. less *Bifidobacteria* but more *E. coli* and *Clostridia* were observed in IBD biopsies (Mylonaki et al. 2005). Luke et al. conducted clone libraries and found more *Bacteroides* and *Provotella* in UC compare to controls (Lucke et al. 2006). The increase in the population of *Bacteroides* in IBD was also reported by Bibiloni et al. (Bibiloni et al. 2006). In one study by Gophna et al., significant increase in the population of both *Bacteroides* and *Proteobacteria* was reported in CD tissues compare to controls. Most of the *Proteobacteria* in the last study were of the *Gammaproteobacteria* class and *E. coli* was the most common species

(Gophna et al. 2006). A recent study by Sokol et al. indicated an anti-inflammatory role for *Faecalibacterium prausnitzii* in gut and identified a decrease in the population of this commensal bacterium in CD patients (Sokol et al. 2008b).

2.3. Methods for determining gut microbial composition

The human digestive tract contains an organized and complex microbial ecosystem known as microflora. It is estimated that more than 30,000 species of bacteria inhabit gastrointestinal tract (Guarner 2005; Manson et al. 2008; Sartor 2008). Prior to advances in molecular biology, these microorganisms were studied using culture-dependent techniques. However, the fact that more than 70% of gut microorganisms are either uncultivable or unidentified makes culture-dependent methods limited (Dorigo et al. 2005).

Today, molecular approaches are widely used to assess the dynamics of the microbial compositions of different ecosystems. These culture-independent techniques are more sensitive and more accurate tools than culture-dependent methods and often allow both monitoring of the dynamics of the microbial composition in a defined environment, as well as the monitoring of the impact of the specific factor on microbial composition (Dorigo et al. 2005).

2.3.1. Real-time PCR

Real-time PCR or quantitative PCR (qPCR) is a technique based on the polymerase chain reaction, which is used to amplify and simultaneously quantify a targeted DNA molecule. Conventional PCR assays only detect the final product of amplification at the end of the PCR reaction, where exponential amplification is no

longer achieved (Denman et al. 2006). In contrast, using real-time PCR or qPCR, quantification of the product is done at the end of each cycle during the exponential phase, where theoretically each cycle will result in a doubling of the amplicon. This allows viewing the entire reaction, monitoring the product being generated, and more accurate quantification of the amplicon (Denman et al. 2006). Real-time PCR can be used to assess the changes in the gut microbiota as a whole (using universal primers), or in a specific species or genus of gut microbiota (using specific primers).

2.3.2. Community fingerprinting techniques

Community fingerprinting methods are DNA-based techniques used to assess the microbial composition of a community without the need for sequencing. For studying communities, usually a universal target of the total community DNA is amplified and then separated into fragments based on their size or sequence. Fingerprinting of the 16S rRNA genes has proved to be a suitable approach for studying microbial diversity, for monitoring the microbial shifts in a defined community, and for comparing different communities (Kocherginskaya et al. 2005; Mackie et al. 2007).

2.3.2.1. DGGE and TGGE

Denaturing gradient gel electrophoresis (DGGE) and thermal gradient gel electrophoresis (TGGE) are two commonly used fingerprinting approaches based on the same principals. These routinely used methods involve amplification of 16S rRNA genes followed by the separation of these fragments in a polyacrylamide gel. The DGGE and TGGE were developed based on the fact that the amplicons of the same size that differ by at least one nucleotide can be separated by electrophoresis through a gradient of

increasing chemical denaturants of urea and formamide (DGGE), or through a temperature gradient (TGGE) (Dorigo et al. 2005). The resulted DGGE/TGGE pattern comprises a profile of multiple bands representing the sequence diversity within the bacterial community (Mackie et al. 2007). Most DGGE/TGGE studies focus on the number of the different bands to get an estimate of the community's richness, but there have been few studies that also take into account the intensity of each band as an indicator for the relative abundance of that bands population (Dorigo et al. 2005). Likewise, the similarity indices, which are indicator of similarity of species between communities, can be calculated and the clustering analysis of the profiles can be performed (Mackie et al. 2007). The main limitations of these techniques are the optimization of the gel running conditions and the difficulty of comparing patterns across the gels (Dorigo et al. 2005). Also, DGGE and TGGE are not sensitive enough to represent the bacteria that make up less than 1% of the total bacterial community, which means that only most dominant bacteria will be detected by these approaches (Mackie et al. 2007).

2.3.2.2. RFLP and T-RFLP

Restriction fragment length polymorphism (RFLP) and terminal restriction fragment length polymorphism (T-RFLP) are two rapid screening methods that present the ribosomal diversity patterns of complex communities, in a relatively easy and reproducible manner (Denman et al. 2006). In RFLP, a portion of the 16S rRNA gene is amplified from total bacterial DNA and then subjected to restriction digestion. Resulting products are separated on the gel and the pattern of the bands is a representative of the phylogenetic diversity within microbial populations (Denman et al. 2006). T-RFLP is an

advancement of the RFLP technology, which implies the fluorescent labeling of one or both of the primers, followed by digestion and detection of terminal restriction fragment (T-RF) through a capillary electrophoresis. T-RFLP is a high-throughput and reproducible method of community structure analyses, both qualitatively and quantitatively (Dorigo et al. 2005). On the resulting dendrogram, the number of the peaks represents the community richness and the area under each peak indicates the relative abundance of that fragment (Dorigo et al. 2005). One advantage of T-RFLP is its ability to detect the less frequent members of the microbial community. Moreover, the comparison of T-RF sizes of the subjects to the web datasets that predict T-RF sizes of known bacteria can lead to the phylogenetic assignment of these fragments (Dorigo et al. 2005). Overall, T-RFLP seems to be a useful tool for assessing richness and diversity in microbial communities that are not extremely complex (Dorigo et al. 2005).

2.3.2.3. RISA and ARISA

Ribosomal intergenic spacer analysis (RISA) involves the PCR amplification of the intergenic spacer region (ITS) located between the small (16S) and large (23S) subunits of the ribosomal gene. This region is highly variable in both size and sequence (Dorigo et al. 2005; Denman et al. 2006). Primers are designed to target the conserved regions of 16S and 23S genes. The amplification products are separated on the gel based on their size heterogeneity. The RISA is a particularly powerful tool to discriminate between closely relates strains (Denman et al. 2006). The use of fluorescently labeled primers allows the analysis to be carried out on an automated capillary electrophoresis. This automated ribosomal intergenic spacer analysis (ARISA) is a more rapid, accurate and efficient version of the RISA (Dorigo et al. 2005). The number of fluorescent peaks

detected by capillary electrophoresis is an estimate of species richness within a given sample (Dorigo et al. 2005). Additionally, the sizes of the detected fragments can be compared to the ones available in databases, leading to potential identification of the amplicon's population (Dorigo et al. 2005).

2.3.2.4. RAPD and AFLP

Random amplified polymorphic DNA (RAPD) and amplified fragment length polymorphism (AFLP) are two molecular techniques not relying on the conserved 16S ribosomal gene for phylogenetic studies, but rather focusing on amplification of random genomic sequences (Denman et al. 2006). RAPD analysis is different from ribosomal-based approaches in that it is based on random amplification of the genetic material and therefore no prior genetic information is required. This is in contrast with T-RFLP and other 16S-based methods where genetic information is required for adequate primer design/ restriction enzyme choice and proper interpretation of the results. To perform RAPD, a small random nucleotide primer of about 10 bp is employed under low stringency conditions for PCR amplification. Usually, several primers will be tested and the one which produces the highest level of community discrimination will be chosen (Denman et al. 2006).

The AFLP is a DNA fingerprinting technique, which detects multiple DNA restriction fragments by the mean of PCR amplification. In this method, the community DNA is digested, ligated with the adapters, and selectively amplified using two primers complementary to the adapter and restriction site sequences. Although, only a subset of fragments is amplified using this technique, a typical AFLP profile contains 50-100 different bands. As was the case with RAPD, the application of the AFLP requires no

prior sequence information. The AFLP profiles are analyzed for pattern similarities (Denman et al. 2006). Depending on the degree of polymorphism between samples, individual samples can be fingerprinted using a wide range of restriction enzymes in combination with different primers.

2.3.3. Hybridization methods

The analysis of molecular fingerprinting techniques such as ARISA and T-RFLP is based on comparison of the size of a DNA fragment against a database. These comparisons provide a list of probable organisms matching particular size, but do not clarify how many of those microorganisms are actually present in each specimen. Another downside of the DNA-based molecular techniques is that although they provide a DNA evidence of the microorganism, they are unable to offer any spatial and conformational characterization of the organisms. While data indicating the presence of microorganisms and their quantity in gut is essential in studying this ecosystem, identification of the structural organization and interaction of these organisms is also critical.

2.3.3.1. Fluorescence *in situ* hybridization (FISH)

FISH is a promising tool for studying the distribution of bacterial populations and their dynamics by fluorescence microscopy. The technique combines the molecular identification of bacteria with the direct visualization of the relationships between the bacteria and the mucosa, providing a significant advantage over other conventional and molecular techniques (Swidsinski et al. 2005). FISH is based on hybridization of the fixed tissue with fluorescently labeled 20-40 bp oligonucleotide probes (universal, genus-

specific, or species-specific) followed by tissue's staining and imaging. Evaluation of the images results in identification and quantification of targeted organisms (Swidsinski et al. 2005).

2.3.4. Sequence-based techniques

2.3.4.1. Clone libraries

In this approach, 16S ribosomal genes of the total community bacteria are amplified (using universally conserved primers), cloned and randomly sequenced (Dorigo et al. 2005). Cloning and sequencing of 16S rRNA genes has become the standard procedure in studying bacterial composition of several ecosystems (Mackie et al. 2007). Analysis of the resulting sequences leads to the identification of the dominant microorganisms. Comparison of these sequences with those available in databases, reveals the information about the identity or relatedness of new sequences to known species (Dorigo et al. 2005). The sequence information can also be used to compare community composition and species richness and diversity (Dorigo et al. 2005). However, in a high diversified ecosystem, this approach could be very laborious, time-consuming and expensive, and therefore, not suited for monitoring the dynamics of the microbial communities (Dorigo et al. 2005; Mackie et al. 2007). Also, both PCR and cloning procedures will introduce biases towards detection of the most common species and therefore, constructing clone libraries is not suitable for studying rare species (Mackie et al. 2007).

2.3.4.2. High-throughput sequencing of phylogenetic markers

Microbial community profiling (determining the abundance of each and every strain of microbes in a community) through direct and high-throughput sequencing techniques has become more accessible recently. Deep views into microbial community profiles using 16S rRNA gene sequences are being achieved increasingly by the means of high-throughput techniques such as barcode pyrosequencing (Hamady et al. 2009). Pyrosequencing is a method of real-time DNA sequencing which relies on detection of pyrophosphate release on nucleotide incorporation. This method allows sequencing of a single strand of DNA (targeted gene or whole genome) by synthesizing the complementary strand one base pair at a time, and detecting which base was actually added at each step (real-time).

2.3.4.3. Metagenomics

Metagenomics involves sampling the genome sequences of a community of organisms inhabiting an environment by high-throughput sequencing. This is different from genomics, in which the complete genetic code of one organism is determined (Hugenholtz et al. 2008). Since metagenomics is trough direct sequencing of DNA obtained from the environment, it does not carry the biases introduced by cultivation or PCR amplification. Metagenomics provides an overview of not only the community's structure (richness and diversity), but also the functional (metabolic) potential of that community. Therefore, metagenomics offers a mean to link diversity to functionality (Booijink et al. 2007).

Metagenomics and high-throughput sequencing techniques are still relatively new and the analysis of the enormous amount of complex data produced trough these approaches, remains the major challenge.

2.4. Biodiversity indices

In its simplest form, biological diversity is the variety of different types of organisms present and interacting in an ecosystem. Recognition of the extent of diversity in each ecosystem is essential for understanding the association between that community's structure and function (Curtis et al. 2004). High species diversity indicates a healthy environment; yet, more species does not simply translate to more diversity. In fact, there are many more factors beyond a simple count of species that determine whether biodiversity is higher or lower in any given ecosystem (Washington 1984). Today, it is generally accepted that a minimum level of diversity is necessary for maintaining a stable ecosystem and that more diverse ecosystems are more productive. Likewise, it has been documented that disturbance of an ecosystem would lower its diversity.

Biodiversity can be expressed in numbers. Some indices require species abundance data (counts) for each species in each sample. This is called abundance data. Other estimators require only presence/absence (occurrence) data for each species in each sample. This is called incidence data (Colwell 2005). Species richness simply presents the number of the species in an environment. Evenness measures how similar the abundances of different species are in a community.

The large number of mathematical methods for calculating diversity indices is proof that no single definition of diversity has yet been established. EstimateS (Colwell

2005) and SPADE (Species Prediction and Diversity Estimation) (Chao et al. 2003) are two widely used statistical packages for measuring biodiversity.

2.4.1. Richness estimators

2.4.1.1. Chao1 and Chao2 richness estimators

These classic richness indices estimate species richness for one community based on the observed abundance (Chao1) or incidence (Chao2) data. These approaches use the numbers of species with a total abundance of 1 (singletons) or 2 (doubletons) to estimate the number of missing/undiscovered species on the assumption that missing/undiscovered species information is mostly concentrated on those low frequency (Chao et al. 2003).

2.4.1.2. ACE and ICE coverage-based richness estimators

The species richness estimators, ACE (Abundance-base Coverage Estimator) and ICE (Incidence-based Coverage Estimator) are modifications of the Chao1 and Chao2 estimators. These estimators divide the observed species into rare and abundant groups. Only the rare group is used to estimate the number of missing species (Chao et al. 2003).

2.4.1.3. MMRuns and MMMeans estimators of total species richness

Michaelis-Menten (MM) richness estimators predict total species richness by functional extrapolation (Colwell 2005). A species accumulation curve is a plot of cumulative species richness against sampling effort with the curve approaching the total species richness asymptote as sampling effort increases. The difference between MMruns and MMMeans is that in MMRuns estimated species richness is calculated for each sample size and then averaged, whereas in MMMeans, the mean is calculated analytically

and then used to estimate species richness for a sample of given size (Walther et al. 1998). Therefore, MMMeans curve is less erratic than the MMRuns method and is generally recommended over MMRuns (Colwell 2005).

2.4.2. Diversity indices

These indices have been developed to mathematically combine the effects of richness and evenness. Each index has its merits, and may put more or less emphasis upon richness or evenness.

2.4.2.1. Simpson's index

Simpson's index measures the probability that two randomly selected individuals belong to the same species. Simpson's index considered a dominance index because it weights toward the abundance of the most common species.

2.4.2.2. Shannon index

Shannon-Weiner index is measuring the order/disorder in a particular system. This order is characterized by the number of individuals found for each species in the sample. Shannon-Weiner index is considered an information index.

2.4.3. Similarity indices

Similarity is a measure of shared species in two communities. Two communities can have identical species richness, but zero similarity, because the same species are not present in both communities. Although similarity indices are measures of community

structure, unlike diversity indices they cannot calculate a value for one sample/community alone (Washington 1984).

2.4.3.1. Classic Jaccard and Sorensen Incidence-based estimators

Classic Jaccard and Sorenson indices are two simplest and mostly used estimators of similarity between two communities. In both approaches, only shared presence contributes to similarity and the shared absences or 0-0 matches are ignored. Values of these indices range from 0 (no shared species) to 1 (identical species) (Wolda 1981; Chao et al. 2005).

2.4.3.2. Bray-Curtis and Morisita-Horn abundance-based estimators

Bary-Curtis index reflects similarities between two samples/communities based on both community composition and species total abundance. Moristia-Horn Similarity estimator is based on the probability that two randomly chosen individuals, one from each of two samples/communities, both belong to the same shared species (Chao et al. 2005).

2.4.3.3. Abundance-based Jaccard and Sorenson indices

Two new abundance-based Jaccard and Sorenson indices have also been recently developed. These estimators calculate the probability that two randomly chosen individuals, one from each of two samples/communities both belong to species shared by both samples/communities (but not necessarily to the same shared species) (Colwell 2005). The estimators for these indices take into account the species that are actually present at both sites, but not detected in one or both samples (Colwell 2005).

2.5. Traditional pathogens

The clinical presentation and histological findings of IBD has similarities to gastrointestinal infections by know pathogenic organisms. For example, CD resembles the intestinal infection due to *Maycobacteria* or *Yersinia,* while the clinical appearance of UC is similar to *Shigella* or *Campylobacter* enteritis (Hendrickson et al. 2002). Numerous bacteria from different genera and species have been associated with IBD. Those include but are not limited to *Helicobacter pylori*, *Enterococcus faecalis*, *Listeria monocytogenes*, *Yersinia pseudotuberculosis*, *Campylobacter* species, *Mycoplasma* species, *Chlamydia* species, *Pseudomonas* species and even Epstein Barr and measles viruses (Guarner 2005; Eckburg et al. 2007; Ingram et al. 2008; McKinley et al. 2008; Pineton de Chambrun et al. 2008; Wagner et al. 2008; Zhang et al. 2009). The focus in recent years however, has been on *Mycobacterium avium* subspecies *paratuberculosis* (MAP) and adherent invasive *E. coli* (AIEC).

Johne's disease is a granulomatous enterocolitis in ruminants with clinical manifestations similar to CD (Feller et al. 2007; Bernstein et al. 2008). Since MAP has been identified as a causative agent of Johne's disease, several studies have been conducted to discover the role of MAP in the pathogenesis of IBD (Chiodini et al. 1996; Hendrickson et al. 2002; Sartor 2008). Potential transmission of MAP to human through infected milk, meat, and water was proposed and MAP was recovered from commercial milk samples and city water supplies (Sartor 2006). Although some groups reported increased recovery of MAP from CD tissues, the results were not conclusive among the studies and the reported rates varied widely (from 0% to 100%) depending to the study

design and/or methodology. Today, the etiologic role of MAP in the development of CD is still controversial (Sartor 2008).

3. *Escherichia coli* **in inflammatory bowel disease**

E. coli is the major cause of intestinal and extraintestinal infections such as pyelonephritis and neonatal meningitis. Paradoxically, it also is the predominant facultative member of the normal human microbiota (Johnson et al. 2000; Kaper et al. 2004). Pathogenic and commensal *E. coli* typically differ in phylogenetic background and virulence features.

A quick review of the literature will reveal that many different microorganisms have been linked to IBD. In recent years however, the role of *E. coli* in the pathogenesis of IBD has been the focus of investigations. Several studies reported higher frequencies of *E. coli* in association to IBD intestinal tissues (Fujita et al. 2002; Ryan et al. 2004; Mylonaki et al. 2005; Swidsinski et al. 2005). Darfeuille-Michaud et al. observed significant increase in the number of *E. coli* isolated from CD biopsies in comparison to the ones isolated from controls. Darfeuille-Michaud and Colombel have been able to recover a new pathovar of *E. coli*, AIEC, from the ileum of 22% of CD patients. These strains appear to have all of the features to colonize intestinal mucosa, cross epithelial barrier, infect epithelial cells and macrophages, and induce pro-inflammatory cytokines (Darfeuille-Michaud et al. 2004; Tamboli et al. 2004). Baumgart et al. conducted series of cultur-independent bacterial analysis on ileal microflora of CD patients and showed that ileal mucosa was enriched in *E. coli* and that the number of *E. coli* was corealated to the severity of ileal disease (Baumgart et al. 2007). The results of Baumgart study also

indicated that *E. coli* isolated from ileum were predominantly novel in phylogeny and *E. coli* with invasive features was restricted to inflamed mucosa (Baumgart et al. 2007). Recently, Sasaki et al. reported that about half of all CD patients harbored invasive *E. coli* strains (Nguyen 2008).

3.1. *E. coli* phylotypes

E. coli is one of the most studied bacteria and has been used as a model system in the development of molecular biology. The *E. coli* reference collection (ECOR) of 72 strains from diverse natural origins was thought to represent the genetic diversity within this species (Ochman et al. 1984; Lecointre et al. 1998). Several comparative analysis including multi-locus enzyme electrophoresis (MLEE) (Herzer et al. 1990), RAPD, and RFLP (Desjardins et al. 1995) divided ECOR strains to four main A, B1, B2 and D phylogenetic groups (Lecointre et al. 1998).

Pathogenic *E. coli* strains derive mainly from phylogenetic group B2 and to a lesser extent group D (Lecointre et al. 1998; Ewers et al. 2007). Commensal *E. coli* by contrast are characteristically from phylogenetic groups A and B1 (Lecointre et al. 1998; Johnson et al. 2000). Groups B2 and D comprise diverse evolutionary lineages that, because of their consistent association with infections, are called "virulent clones" (Picard et al. 1999; Johnson et al. 2000). A hallmark of such "virulent clones" is their possession of virulence factors that holds pathogenic characteristics that are infrequent in commensal *E. coli* (Johnson et al. 2000). B2 strains are rare in the feces in commensal situations and they represent 7% (in Western countries) to 3% (in developing countries) of the commensal *E. coli* population (Picard et al. 1999). Similarly, phylogenetic group D strains are underrepresented in the human commensal strains. In one study on 82

commensal and extraintestinal pathogenic *E. coli*, commensal strains were exclusively from A and B1 phylotypes while extraintestinal pathogens belonged to all A, B1, B2, and D phylogenetic groups although B2 was the predominant virulent phylotype (Picard et al. 1999). However, in the same study, where the lethality of isolates in mice were assessed, B2 was the deadliest followed by D, B1 and A phylotypes (Picard et al. 1999). This observed link between virulence and phylogeny could correspond to the necessity of having virulence determinants into the right genetic background for the emergence of a virulent clone. Selection could have favored virulent strains, since they compensate for the loss of transmission opportunities that result from killing or debilitating their hosts, by increased fitness in the presence of host defenses or other bacterial strains (Picard et al. 1999).

The most basal taxon within *E. coli* in terms of branching patterns is the B2 group rather than the A group, as generally thought. The D group then emerges as the sister group of the B2. Finally, the A and B1 groups are the most closely related groups. Thus, in *E. coli* strain tree, the more basal the branching of a given group, the more virulent it is (Lecointre et al. 1998).

3.2. Intestinal pathogenic *E. coli*

Based on clinical reports and virulence features, *E. coli* have been categorized into intestinal pathogenic, extraintestinal pathogenic (ExPEC), and commensal *E. coli*. Among the intestinal pathogens there are six well-described categories: enteropathogenic *E. coli* (EPEC), enterohaemorrhagic *E. coli* (EHEC), enterotoxigenic *E. coli* (ETEC), enteroaggregative *E. coli* (EAEC), enteroinvasive *E. coli* (EIEC) and diffusely adherent *E. coli* (DAEC) (Kaper et al. 2004; Servin 2005). Pathogenic *E. coli* have evolved several

mechanisms by which they can colonize the mucosa, evade host defense, multiply, damage the host, and cause disease.

EPEC was the first pathotype of *E. coli* to be described. EPEC was responsible for large outbreaks of infant diarrhea in United Kingdom and was characterized by its "attaching and effacing, A/E" histopathologic pattern. The ability of EPEC to induce this A/E pathology is encoded by a pathogenicity island called the locus of enterocyte effacement (LEE) (Kaper et al. 2004). The LEE encodes a protein called intimin, which mediates the intimate attachment of EPEC to epithelial cells, and stimulates mucosal immune responses. Tir (translocated intimin receptor), is one of type III secretion system effector proteins encoded by LEE and inserted into the host cell membrane to function as a receptor for intimin. This is a great example of a pathogen that provides its own receptor for binding to host cells (Kaper et al. 2004).

EHEC causes bloody and non-bloody diarrhea, and haemolytic uremic syndrome. The principal reservoir of EHEC is the bovine intestinal tract and initial outbreaks were associated with consumption of undercooked hamburgers (Kaper et al. 2004). Subsequently, a wide variety of food items have been associated with disease, including sausages, unpasteurized milk, lettuce, cantaloupe melon, apple juice, radish sprout and even drinking water. Although EHEC strain O157:H7 is the most important pathogens in North America and United Kingdom, other serogroups such as O26 and O111 are more prominent in many other countries. Stx1 and Stx2, which are also known as verocytotoxin1 (VT1) and verocytotoxin2 (VT2), are EHEC's key virulence features. Stx is produced in the colon and travels by the bloodstream to the kidney, where it damages renal endothelial cells, induces cytokine and chemokine production and results in renal

inflammation (haemolytic uremic syndrome). Stx also induces apoptosis in intestinal epithelial cells. In addition to Stx, most EHEC strains contain the LEE pathogenicity island that encodes type III secretion system and effector proteins that are homologous to those that are produced by EPEC (Kaper et al. 2004). The genome sequence of O157:H7 revealed additional potential virulence factors, including novel fimbriae, iron uptake and utilization systems, and a urease that is similar to those produced by *Klebsiella* and other urinary tract pathogens (Perna et al. 2001; Heimer et al. 2002).

ETEC is the important cause of childhood diarrhea in the developing countries and is the main cause of travelers' diarrhea. ETEC colonizes the surface of the small intestine and release enterotoxins, which give rise to intestinal secretion causing watery diarrhea. ETEC strains might express only an LT, only an ST (heat-stable enterotoxin), or both LTs and STs. LTs are a class of enterotoxins that are structurally and functionally related with cholera enterotoxin, which is expressed by *Vibrio cholerae*. The ST family seems to represent a case of molecular mimicry since its structure is similar to guanylin for which there is an intestinal receptor (Kaper et al. 2004).

EAEC are cause of persistent diarrhea in children and adults in both developing and developed countries. EAEC adhere to enterocytes in a "stacked-brick" pattern, colonize the intestinal mucosa and secrete enterotoxins and cytotoxins, inducing mild but significant mucosal damage. The process of adhering to enterocyte is facilitated by fimbrial structures known as aggregative adherence fimbriae (AAFs). Several toxins have been also described for EAEC including *Shigella* enterotoxin 1 (ShET1) and enteroaggregative heat-stable enterotoxin (EAST1) (Kaper et al. 2004).

EIEC and *Shigella* spp. are genetically and pathogenically closely related. EIEC are distinguished from *Shigella* by a few minor biochemical tests, but these pathotypes share essential virulence factors. Much of EIEC/*Shigella* pathogenesis is the result of the type III secretion system encoded by virulence plasmid of *S. flexneri*, which is essential for the invasiveness characteristic of EIEC and *Shigella* species. EIEC elicits watery diarrhea that is indistinguishable from that due to infection by other *E. coli* pathogens (Kaper et al. 2004).

DAEC are defined by the presence of a characteristic, diffuse pattern of adherence to host cells. Majority of DAEC produce a fimbrial adhesion belonged to the Dr family of adhesins which uses decay-accelerating factor (DAF, also known as CD55) as receptor. Recognition of Dr family adhesions by DAF leads to the development of long cellular extension which wrap around the adherent bacteria (Kaper et al. 2004; Servin 2005). The first class of DAEC strains includes *E. coli* strains that harbor Afa/Dr adhesion. Adhesin involved in diffuse adherence (AIDA) is expressed by second class of DAEC (Servin 2005).

3.3. Extraintestinal pathogenic *E. coli* (ExPEC)

In contrast to intestinal pathotypes which are well defined by the possession of distinct combination of virulence factors and molecular pathways, pathotypes have yet to be identified for extraintestinal disease as these strains share a lot of virulence features (Ewers et al. 2007). Avian pathogenic *E. coli* (APEC), uropathogenic *E. coli* (UPEC), and newborn meningitis-causing *E. coli* (NMEC) are better characterized extraintestinal pathogens that are able to establish infections in extraintestinal habitats of different hosts (Kaper et al. 2004; Ewers et al. 2007).

UPEC is a subset of *E. coli* that causes uncomplicated cystitis and acute pyelonephritis in human host. Several virulent determinants have been associated to UPEC including hemolysin, CNFs (cytotoxic necrotizing factors 1 and 2), P fimbriae, type I fimbriae, S fimbriae, F1C fimbriae, and Dr family of adhesions (Kaper et al. 2004). UPEC are known for their possession of large and small pathogenicity islands containing genes that are not found in intestinal pathogens. Type 1 fimbriae attach to mannose component of the urinary tract epithelial receptors and initiate colonization and biofilm formation in bladder. P fimbriae on the other hand have receptors in kidney and is responsible for pyelonephritis (Kaper et al. 2004). According to fecal-vaginal-urethral hypothesis, *E. coli* strains causing urinary tract infection are usually derived from the host's own fecal flora (Ewers et al. 2009).

MNEC is the most common cause of Gram negative neonatal meningitis with very high rates of mortality and morbidity. K1 capsule is the most essential virulence determinant of MNEC because of its serum resistance and anti-phagocytic properties. Other virulence factors used by MNEC include OmpA and S fimbriae for luminal attachments and CNF1 for invasion (Kaper et al. 2004).

As a clearly defined genotype of APEC has not been determined, the term APEC is generally used for strains isolated from poultry with clinical signs of localized or systemic *E. coli* infection (Ewers et al. 2009). Infection by APEC usually starts in respiratory tract of domesticated and wild birds, leading to a systematic infection of internal organs with progression to sepsis. A variety of APEC serogroups have been involved in human urinary tract infections and neonatal meningitis (Ewers et al. 2007).

Results of ECOR typing have confirmed that human ExPEC trains mostly belong to group B2, followed by phylotype D (Ewers et al. 2007). In APEC however, phylotype A made up the majority of strains, although a considerable portion of strains were still belonged to B2 (Ewers et al. 2007).

The common presence of a set of virulence features, in addition to similar disease pattern and phylogenetic background in ExPEC, indicates a significant zoonotic risk of avian-derived *E. coli* strains (Ewers et al. 2007). As APEC are epidemiologically highly abundant compare to UPEC and NMEC, many speculate that poultry may be a reservoir for *E. coli* that is capable of causing adult urinary tract infection and newborn meningitis (Ewers et al. 2007).

3.4. Adherent invasive *E. coli* (AIEC) and IBD

As mentioned before, AIEC strains can invade and replicate within macrophages and can induce the release of high amounts of TNF-α, which could lead to the intestinal inflammation that is characteristic of CD (Kaper et al. 2004). No unique genetic sequence has been yet described for AIEC strains (Kaper et al. 2004).

3.5. *E. coli* virulence factors

Pathogens are different from their non-pathogenic relatives in that they have evolved the ability to cause disease in another organism (Gal-Mor et al. 2006). The barrier between commensalisms and virulence results from the balance between the presence and the expression of virulence factors and the status of the host (Picard et al. 1999).

In general, virulence factors facilitate colonization and invasion of the host, avoid or disrupt host defense mechanisms, injure the host tissue, and stimulate inflammatory response (Johnson et al. 2000).

3.5.1. Pathogenicity islands

Pathogenicity island (PAI) consists of mobile genetic elements that play an essential role in virulence of bacterial pathogens (Schmidt et al. 2004). PAIs are relatively large genomic regions (10-200 kb) that are present in the genome of pathogenic bacteria but are absent from the genome of the non-pathogenic bacteria of the same species. PAIs carry virulence genes and other mobile genetic elements (Schneider et al. 2004; Gal-Mor et al. 2006). As a result of their mobility, PAIs often differ from the core genome in their G + C (guanine and cytosine) composition. Genome of each bacterial species carries a specific composition of bases, expresses as a percentage of guanine and cytosine bases. G + C composition of bacteria ranges from 25 to 75% and is generally conserved at the genus or species level. It is believed that the horizontally acquired PAI has the G + C composition of the donor species (Schmidt et al. 2004).

Since the introduction of PAIs in 1990 (Gal-Mor et al. 2006), they have been identified in a wide range of bacterial pathogens. Characterization of these PAIs not only led to the identification of many virulence genes but changed our vision of the evolution of bacterial pathogenicity (see section 3.6.2 of this chapter) (Gal-Mor et al. 2006).

3.5.2. Adhesins

Many key *E. coli* virulence factors are surface structures involved in adherence. Pathogenic *E. coli* strains possess specific adherence factors with the ability to bind to

host receptors which allow them to colonize sites that *E. coli* does not normally inhabit, such as small intestine and urethra (Holden et al. 2004; Kaper et al. 2004; Servin 2005) Sequencing of *E. coli* has shown that they contain multiple fimbrial and non-fimbrial adhesion gene clusters (Holden et al. 2004). Fimbrial adhesions are hair-like surface structures that are different from flagella (Kaper et al. 2004).

3.5.2.1. Fimbriae

3.5.2.1.1. Type 1 fimbriae and FimH subunit

Type 1 fimbriae are encoded by the *fim* gene cluster and consist of a major protein, FimA, associated with ancillary proteins, FimF, FimG and the adhesin FimH (Dho-Moulin et al. 1999). FimH recognizes D-mannose receptors (Holden et al. 2004).

3.5.2.1.2. S fimbriae

S fimbrial adhesin (Sfa) mediates adhesion to sialic acid-containing receptors (Schmoll et al. 1990; Hacker et al. 2000; Holden et al. 2004). Sfa gene cluster contains nine genes with s*faA* encoding major pilin subunit and s*faG*, s*faS*, and s*faH* encoding minor subunits of fimbrial. s*faB* and s*faC* are regulatory genes that have also strong homology to *pap* regulatory genes (Schmoll et al. 1990). Sfa/foc operons are found only in the phylogenetic B2 group (Picard et al. 1999).

3.5.2.1.3. P fimbriae

Pyelonephritis-associated fimbriae (P fimbriae) are known to play an essential role in the pathogenesis of APEC and UPEC infections by means of their PapG adhesin

(Ewers et al. 2009). P fimbriae recognize digalactoside-containing receptors (Holden et al. 2004).

3.5.2.2. Afa/Dr adhesins

The *afa* gene cluster encodes afimbrial adhesins (Afas) that are expressed by pathogenic strains of *E .coli*. The structural genes coding for Afa/Dr adhesins consist of genes *A* to *D*, which encode accessory proteins and are highly conserved, and gene *E*, which encodes adhesin molecule itself, and is more divergent (Servin 2005). It has been proposed that the Afa/Dr family of adhesins includes 13 human adhesins, including AfaE-III which is expressed by both uropathogenic and diarrheaogenic strains of *E. coli*. Carcinoembryonic antigen–related cell adhesion molecules 1 and 6 (CEACAM1 and CEACAM6) are the receptors for afaE-III adhesins (Servin 2005).

3.5.2.3. AIDA-I

Adhesin involved in diffuse adherence I (AIDA-I) is a member of autotransporters family and is responsible for diffuse-adhering phenotype of *E. coli*. AIDA-I is an autotransporter in that it is processed into a passenger domain in a manner consistent with autotransporter's secretion (Henderson et al. 2004).

3.5.3. Aggregation factors

3.5.3.1. Antigen 43

Antigen 43 (Ag43) is an autotransporter protein encoded by *agn43* gene (also known as *flu* gene) and is a cell surface adhesin with autoaggregating ability (Henderson et al. 2004; Hatt et al. 2008). Many strains of pathogenic *E. coli* possess multiple copies

of Ag43 (Henderson et al. 2004). Ag43 is able to enhance *E. coli* biofilm formation by inducing microcoloni structures. Moreover, the expression of Ag43 mediates interaction between different bacterial species on a surface and under continuous flow conditions indicating Ag43's ability to provide both cell-to cell and cell-to-surface contacts (Hatt et al. 2008). Ag43 expression allows both inter- and intraspecies cell aggregation, which is ideal in a biofilm setting (Henderson et al. 2004). Ag43 is present in about 60% of UPEC strains and has shown to be involved in both abiotic and living tissue biofilm formation, similar to type I pili (Hatt et al. 2008).

3.5.4. Toxins

Secreted toxins are more frequent in diarrheagenic *E. coli* than adhesins. The heat-labile enterotoxin (LT), heat-stable enterotoxin a (STa) and heat-stable enterotoxin b (STb), which are all produced by different strains of ETEC, have the ability to increase the intracellular concentrations of cyclic AMP, cyclic GMP and Ca2+, respectively (Kaper et al. 2004). The Shiga toxin (Stx) of EHEC cleaves ribosomal RNA and disrupts protein synthesis which results in destruction of the intoxicated epithelial or endothelial cells. The cytotoxic necrotizing factors (CNF 1 and CNF 2), lock the RhoA signaling molecules in the 'on' position, leading to cytoskeletal alterations, multinucleation with cellular enlargement, and finally necrosis (Kaper et al. 2004).

A variety of toxins are transported from the bacterial cytoplasm to the host cells by different mechanisms (Type I-V secretion systems). The UPEC hemolysin is the prototype of the type I secretion mechanism which is a simple system using only three protein subunits, the ABC protein, membrane fusion protein, and outer membrane protein to export molecules from the cell (Henderson et al. 2004; Kaper et al. 2004). LT is a

classic A–B subunit toxin that is secreted to the extracellular milieu by a type II secretion system. EPEC, EHEC and EIEC contain type III secretion systems, which are complex protein structures forming a 'needle and syringe' gadget that allows effector proteins to be injected directly into the host cell (Kaper et al. 2004). No type IV secretion systems have been described for pathogenic *E. coli*. Several toxins, such as Sat and Pic, are called autotransporters (type V secretion system) because part of these proteins forms a pore in the outer membrane that allows the extracellular access of the other part of the protein. Type V secretion system is known as the simplest mechanism of protein secretion. (Henderson et al. 2004).

3.5.5. Serine protease autotransporters (SPATEs)

The SPATEs are a subfamily of serine protease autotransporters that are produced by diarrheagenic and uropathogenic *E. coli* and Shigella strains (Kaper et al. 2004). Despite description of SPATEs in several pathotypes of *E. coli*, their full contribution to the pathogenesis is unknown. However, no SPATE has been yet identified in non-pathogenic organisms, and each SPATE member is among the predominant secreted proteins of their respective pathogens (Henderson et al. 2004). Temperature-sensitive hemagglutinin (Tsh) and Vat of APEC, *Shigella* extracellular protein A (SepA) and SigA of *Shigella*, EPEC-secreted protein C (EspC), extracellular serine protease plasmid-encoded (EspP) of EHEC, Pic and Pet of EAEC, and secreted autotransporter toxin (Sat) of UPEC are among the better known SPATEs (Henderson et al. 2004).

3.6. Evolution of IBD associated *E. coli*

Bacterial evolution occurs mainly by clonal divergence through the modification of existing genetic information and by acquisition of new sequences through horizontal gene transfer, followed by a periodic selection (Le Gall et al. 2007). This has led to the concept that bacterial genome are composed of conserved "core" genome containing the genetic information essential for cellular functions, and of "flexible" gene pool which is able to fit to different environments (Hacker et al. 2001).

The process in which the content and organization of genetic information of a species changes over time is known as genome evolution (Schmidt et al. 2004). In bacteria, this process includes point mutations, rearrangements, insertions, and deletion of DNA segments (Hacker et al. 2001; Schmidt et al. 2004). This loss and acquisition of gene content can rapidly change the life style of a bacterium and seems to be the primary force by which bacteria fit to novel environments and diverge evolutionary (Schmidt et al. 2004).

The path of evolution is determined by an increase in the fitness of an organism toward its environment. Fitness is considered to be a set of properties that enhance the transmission, spread and the survival of an organism within a specific ecological niche (Preston et al. 1998; Hacker et al. 2001). Fitness depends on several criteria including the genetic composition of the mobile element, genetic background of the bacterial host and the ecological habitat of the microorganism (Hacker et al. 2001). The data on *E. coli* supports the idea that the fitness of an isolate is optimum when there is a fine tuning between the chromosomal backbone and the genes newly acquired through horizontal transfer (Escobar-Paramo et al. 2004a; Le Gall et al. 2007).

3.6.1. Mutations, insertions and deletions, "conventional evolution"

At the genetic level, change constantly occurs in the form of mutation, substitution, insertion and deletion. However, most of the changes are not naturally selective (silent mutations, synonymous mutations) and they become fixed randomly in time (Woese 1987). These silent mutations usually results in different versions of a sequence, all being functionally equal, and only reflect a lineage's mutation rate (Woese 1987). In contrast, non-synonymous mutations happen under selective pressure, and contribute to the fitness of organism in a specific niche.

3.6.2. Horizontal gene transfer and "evolution in quantum leaps"

Pathogenic and non-pathogenic species of the same bacteria often only differ with respect to certain virulence factors which may be packaged into large blocks such as PAIs (Hacker et al. 2003). The process of developing different pathotypes through acquisition or loss of large and mobile genetic elements, e.g. plasmids, integrons, or PAIs is known as "evolution in quantum leaps" (Hacker et al. 2003). This horizontal genetic transfer can change the composition of bacterial genome rapidly and dramatically over relatively short time periods and though is critical to bacterial evolution (Hacker et al. 2001).

3.7. Sequence-based methods for phylogenetic analysis of microorganisms

Phylogenetics is the study of the ancestry of organisms over long time scales, involving many thousands or even millions of years. In other words, Phylogenetics studies evolutionary relatedness among various groups of organisms and tracks the origin of them (Ludwig 2007). In the past, phylogenetic analysis was heavily based on phenotypical and biochemical data. With many whole genome sequencing projects being

completed and many more genome projects on the horizon, construction of molecular trees based on nucleotides and derived amino acid data has become an important tool in phylogenetic studies.

3.7.1. Phylogenetic markers

In 1965, Zuckerkandl and Pauling introduced the idea that macromolecules such as nucleic acids and proteins can act as documents of evolutionary history, containing a record of changes since divergence from a common ancestor (Zuckerkandl et al. 1965). Central to this approach was the selection of a suitable molecule that was not restricted to particular taxonomic levels or bacterial groups, and had not undergone subtle functional changes (Owen 2004). Not all sequences are of equal value in determining phylogenetic relationships. The criteria for choosing such adequate genes have been defined in 1987 by Woese (Woese 1987). To be useful in phylogenetic studies, a molecule has to meet certain specifications:

(i) The molecule must be universally distributed among the line of the descent (Uilenberg et al. 2004). This is essential since this molecule should be representative of all organisms under the study.

(ii) Changes in its sequence have to occur as randomly as possible. Virulence genes and regions that are prone to lateral gene transfer are obviously not suitable for phylogenetic measures (Uilenberg et al. 2004) since they are under huge selective pressure. The more useful molecules for phylogenetic measurement are the ones that represent highly enforced functions. Some sequences of this type change slowly enough to span the full evolutionary spectrum.

(iii) Rates of change have to be commensurate with the spectrum of evolutionary distances being measured. Bacteria have evolved for much longer time compare to Eukaryotes, therefore, the range of the molecules used to measure phylogenetic relationships among bacteria needs to be considerably greater than what is optimal for Eukaryotes.

(iv) The molecule has to be large enough to provide an adequate amount of information. In addition to the obvious need for large size (good statistics), the molecule needs to have fairly large number of domains (functional units) that are somewhat independent of one another in an evolutionary sense (Woese 1987). In this case, selective changes affecting one of the units will not substantially affect the other domains. Therefore, when one domain of the molecule becomes altered by introduction of selected changes, the other domains remain practically unaffected. The more units of this kind a molecule contains, the less sensitive its evolutionary measure is to selective changes in one domain (Woese 1987).

3.7.1.1. 16S ribosomal RNA

There are three types of rRNA present in all bacteria, namely 16S rRNA, approximately 1500 bp and found in the small (30S) ribosomal subunit, and 5S and 23S rRNAs, about 115 bp and 2900 bp respectively, and both found in the large (50S) ribosomal subunit. These macromolecules are easy to isolate, could be sequenced directly, and in the case of 16S and 23S rRNA, contained considerable taxonomic information that is typically present in multiple copies throughout the genome (Owen 2004).

16S rRNAs are still the most useful and the most used of molecules in phylogenetic studies. They occur in all organisms, and different positions in their sequences change at very different rates, allowing most phylogenetic relationships to be measured. Their sizes are large (1500 bp) and they consist of many domains. 16S rRNA molecules show a high degree of functional constancy, which assures relatively low selective behavior. Comparisons among the full sequences of 16S rRNA allows high resolution phylogenetic groupings to be identified at various taxonomic levels. This tool has proved to be particularly useful for the identification and classification of unculturable organisms, as 16S rRNA gene can be amplified by polymerase chain reaction (PCR) of genomic DNA using universal primers (Uilenberg et al. 2004).

Using the sequence of 16S rRNA for constructing phylogenetic trees has some limitations, particularly in the discrimination of closely related taxa, and there is a growing interest in the use of alternative loci such as *cpn60*, *gyrB* and RNAase P sequences (Owen 2004; Riesenfeld et al. 2004). These alternative macromolecules are universally distributed among bacteria, their genes do not transmit horizontally, and their molecular evolution rate is comparable to or higher than that of 16S rDNA, which makes them more suitable for differentiation of closely related organisms (Vandamme et al. 1996). Comparison of the degree of similarity between phylogenetic trees derived from different genes provides a valuable test of the extent they represent the phylogeny of the species (Owen 2004).

3.7.1.2. Heat shock protein 60 (HSP60, CPN60)

Type I chaperonins are molecular chaperones present in virtually all bacteria, some archaea and the plastids and mitochondria of eukaryotes. Sequences of *cpn60*

genes, encoding 60-kDa chaperonin protein subunits (CPN60, also known as GroEL, MopA, or HSP60), are useful for phylogenetic studies and as targets for detection and identification of organisms (Hill et al. 2004). Multiple functions have been described for CPN60. While the primary intracellular role of CPN60 is thought to be as a molecular chaperone in the processes of post-translational protein folding and assembly of protein complexes, CPN60 also appears to function as an intercellular signaling molecule. The universal nature of *cpn60* genes makes them attractive targets for phylogenetic studies (Hill et al. 2002). An analysis of the *cpn60* sequences from a variety of bacterial and eukaryotic species led to the design of universal PCR primers, which can amplify a 549 to 567 bp region of *cpn60* corresponding to nucleotides 274–828 (Hill et al. 2004).

It has been demonstrated that the *cpn60* universal region generally provides more discriminating and phylogenetically informative data than the 16S rDNA target. A number of *cpn60* characteristics offer significant advantages over 16S rDNA for studying complex microbial populations and for quantitative assays. As protein-coding genes, *cpn60* sequences are less constrained from sequence variation than are structural RNA-encoding genes. Furthermore, sequence variation extends quite uniformly throughout the *cpn60* coding region, whereas variable regions of 16S rRNA genes are dispersed between regions of highly conserved sequence. Highly stable secondary structure that is associated with 16S rRNA is not present in *cpn60* genes or transcripts. Generally, *cpn60* genes are single copy in prokaryotic genomes, and the relatively small size of the universal target facilitates high-throughput sequencing approaches (Hill et al. 2004). A comparative study of 35 serotypes of *Streptococcus suis* showed that the chaperonin 60 gene sequences for *S. suis* strains were significantly more distant from each other than the 16S rRNA

sequences were. Similarities between 16S rRNA sequences were typically more than 96% whereas the similarity of *cpn60* sequences were distributed between 70 to 100% (Brousseau et al. 2001). A collection of *cpn60* sequence data is available at http://cpndb.cbr.nrc.ca.

3.7.1.3. Other phylogenetic markers

23S rRNA, gyrase B subunit (*gyrB*), *ureI*, ribonuclease P (RNase P), *recA* elongation factors, the gene encoding ATPase subunits, *rpoβ* (β subunit of the RNA polymerase), glutamate synthase *gltA*, and translation initiation factor (*infB*) have been also investigated with regards to their potential as phylogenetic markers (Ludwig et al. 1994; Yamamoto et al. 1996; La Scola et al. 2003; Owen 2004).

3.7.2. Multi-locus sequence typing (MLST)

In an attempt to overcome some of the limitations of 16S rRNA phylogeny, MLST analysis based on conserved sequences of housekeeping genes is used to group closely related isolates (Gevers et al. 2005). Housekeeping genes are the genes expressed in the cell constitutively and encode products that are essential for maintaining bacterial function (Hanage et al. 2006). Therefore, these genes believed to undergo slower evolution as a result of selective pressure (Hanage et al. 2006). Using sequences of several loci reduces the biases that may happen as a result of recombination in single gene. MLST is a powerful tool for phylogenetic analysis and evolutionary reconstruction since: (i) it is highly discriminative and has the ability to detect mutations (substitutions, insertions, and deletions) at the level of nucleotide; (ii) the digital format of the data facilitates the establishment of a world-wide web-accessible database; (iii) the sequence

data is easily transferable and comparable between labs; and (iv) the process is highly reproducible (Tartof et al. 2005). The rapidly expanding datasets of multi-locus sequence types (MLST) offer an important source for identifying polymorphisms in housekeeping gene sequences, which is linked to nonsynonymous mutations and could provide valuable phylogenetic information (Owen 2004).

3.8. Link between virulence and phylogeny

UPEC, NMEC, and APEC strains sharing the same virulence gene combination also share the same phylogenetic group but not always the same serogroups (Ewers et al. 2007). Many studies showed that ExPEC from B2 and D phylogroups have greater frequency and diversity of virulence features than groups A and B1 (Boyd et al. 1998; Johnson et al. 2001; Le Gall et al. 2007; Ewers et al. 2009). Ewers et al. screened 526 medical and veterinary isolates of APEC, UPEC and NMEC and identified an average of 15 virulence genes for each strain with highest and lowest number of genes being found in B2 and A phylotypes respectively (Ewers et al. 2007). MLST analysis of ExPEC revealed that most strains that were virulent in chicken belonged to sequence types that were almost exclusively associated with human extraintestinal disease, supporting the hypothesis that the phylogeny of the *E. coli* and not the ecohabitat determines virulence (Ewers et al. 2009). On the other hand, B2 phylogroup can be the most frequent strain isolated from the feces of asymptomatic humans (Zhang et al. 2002; Escobar-Paramo et al. 2004b). Moreover, it seems that the prevalence of B2 strains isolated from human feces have increased substantially over the last few decades (Nowrouzian et al. 2006; Le Gall et al. 2007).

Virulence factors that are only present in a few B2 or D phylotypes are likely to have been acquired recently by those phylogenetic groups. The confinement of such virulence factors to pathogenic phylotypes is either due to their recent arrival in the B2 or D phylotype or to barriers to their horizontal movement into other groups (Johnson et al. 2000).

Those virulence genes that are present in most but not all B2 or D strains probably have entered the B2 or D groups after their differentiation from other *E. coli*. Prevalence of such genes in other phylotypes strongly indicates their horizontal transfer to other phylogenetic groups (Johnson et al. 2000). In contrast, those B2 associated virulence factors that are nearly universal among B2 strains and are also prominent among other phylotypes may have been acquired by *E. coli* prior to the branching off of the B2 group; or alternatively, may have been acquired very early by B2 cluster with subsequent horizontal transfer into other lineages (Lecointre et al. 1998; Johnson et al. 2000).

When a virulence factor is highly prevalent throughout the population, it suggest that they either entered *E. coli* early during its evolutionary history, hence now are present in all members of the species (*fimH*), or, if more recently acquired, are highly horizontally mobile and are strongly selected for in the context of the particular disease (Johnson et al. 2000). On the contrary, the virulence factors that are concentrated outside of B2 may have been entered non-B2 lineages subsequent to the branching off of phylogroup B2 (Lecointre et al. 1998).

MATERIALS AND METHODS

1. Microbial diversity of gut biopsy tissues in IBD

Study subjects. Biopsies were drawn from a population-based case-control study undertaken at the University of Manitoba (Bernstein et al. 2003; Eckburg et al. 2005). For creating a population-based database, subjects were identified through the administrative database of Manitoba Health, the single provincial health insurer, and were mailed questionnaires regarding their histories of IBD and willingness to participate in future studies. Those who returned the questionnaires were logged in the University of Manitoba Research Registry (Bernstein et al. 2003) and enrolled in the study.

As part of a case-control study, a population-based set of controls was selected from the Manitoba Health population registry. This registry contains demographic information on all individuals registered with the Manitoba Health public insurance system. The registry is regularly updated with vital registrations and information from medical and hospital transactions, and closely matches population estimates derived from the Canadian census (Statistics Canada) (Roos et al. 1993). A random sample of registered individuals was selected with stratification for age and gender to achieve balance with the case series for those two variables. Using this stratification, Manitoba Health's Information Services generated a list of eligible controls and sent an information package prepared by the investigators explaining the study and requesting participation. For a second control group, patients with IBD were asked to refer one or more of their siblings.

All cases and controls who agreed to participate completed a questionnaire. Controls were invited to participate in the colonoscopy plus biopsy study, and those who agreed were paid an honorarium. Cases were asked to contact the study personnel when they were to undergo their next colonoscopy for clinical reasons. All cases who were to undergo colonoscopies agreed to tissue collection for study purposes.

At colonoscopy, biopsy specimens were obtained from cecum, colon, and rectum, using a standard colonoscopy preparation procedure with Fleet[®] Phospho-soda[®] oral saline laxative. Biopsies were taken from both inflamed and non-inflamed sites. For subjects with a previous cecal resection, biopsy specimens were taken from the right colon distal to the ileocolonic anastomosis. All biopsy specimens were snap-frozen in liquid nitrogen and stored at $-70\,°C$ (Bernstein et al. 2003).

For this study, 58 biopsies were taken from 16 healthy controls (HC), 10 CD patients, and 15 UC patients (Table 1). Biopsies from more than one site were obtained from some subjects. CD and UC diagnoses were verified by chart review and based on standard endoscopic, histologic, and where applicable surgical criteria. None of the subjects had undergone antibiotic treatment within 3 month prior to the procedure (Sepehri et al. 2007). A subset of biopsies was subject to standard histological staining with haematoxylin and eosin to verify the inflammation state. A site was considered inflamed if it had histological evidence of inflammation and was considered non-inflamed if it was histologically normal (Sepehri et al. 2007).

Table 1. Biopsy samples used for microbial diversity analysis

Disease state	Total	Location[a]		Histological diagnosis[b]	
		Cecum	Rectum	Normal	Inflamed
HC	20 (20)	12 (12)	8 (8)	20 (20)	0 (0)
CD	16 (10)	9 (9)	7 (7)	8 (7)	8 (6)
UC	22 (16)	11 (11)	11 (11)	10 (10)	12 (9)

HC, healthy controls; CD, Crohn's disease; UC, ulcerative colitis
Numbers in parentheses represent the number of patients
[a] Samples taken from the ascending colon (2) and descending colon (4) of UC patients were grouped with cecum and rectum respectively.
[b] Biopsies were also taken from non-inflamed sites of CD and UC.

DNA extraction. Biopsy tissues were suspended in 150 ml lysis buffer (10 mM Tris-HCl, pH 8.0; 5 mM EDTA, pH 8.0; 4 M guanidinium isothiocyanate, pH 7.5; 50 g sarcosyl/l, 2.5 g SDS/l, 5 g sodium citrate/l and 5 g Triton X-100/l), followed by the addition of 300 ml of chloroform and Tris-saturated phenol (pH 6.9). Tubes were placed at –20 °C for 1 h. Samples were then centrifuged at 4 °C for 20 min at 10,000 g and supernatants were transferred to fresh tubes. Isopropanol was added to 1/4 volume of the supernatants and the mixtures loaded on to silica-cellulose membranes in columns. Samples were allowed to filter through the membrane by gravity. The membranes were washed twice with 300 ml of 95% ethanol. DNA was eluted with 400 ml Tris-EDTA buffer and precipitated with two portions of 95% ethanol. The resulting pellets were suspended in 25 ml 0.5× Tris-EDTA buffer (pH 8.0) and stored at –20 °C until further analysis (Kotlowski et al. 2007).

DNA amplification, digestion and fragment sizing. For ARISA, previously reported ITSF and ITSReub primer sets (Table 2) were used to amplify the ribosomal intergenic spacer (ITS) region from community DNA (Cardinale et al. 2004). Primers 27f and 342r (Table 2) were applied in order to amplify a fraction of 16S rDNA gene (Lane 1991). Forward primers were fluorescently labeled (WellRED D4dye, Sigma-Proligo, USA) to allow detection of the fragments by capillary electrophoresis. The PCR reaction was as follows: 94 °C for 1 min; 36 cycles at 94 °C for 1 min; 55 °C for 1 min; 72 °C for 2 min; and a final extension at 72 °C for 5 min. To produce terminal restriction fragments (T-RF), the 27-342 region of 16S DNA was digested with *Hha*I restriction enzyme (10 μl of PCR product, 10 units of *Hha*I, 1× *Hha*I buffer and 20 μg of bovine

serum). The mix was adjusted to a final volume of 20 μl with MilliQ water and the DNA was digested at 37 °C for 3 hours. The precise length of ITS and T-RF amplicons were determined by performing capillary electrophoresis with a CEQ^{TM} 8800 Genetic Analysis System (Beckman Coulter Inc.). Two μl of fluorescently labeled fragments (ITS or T-RF), 26μl of sample loading solution, and 0.5 μl of DNA size standard (600 bp for ARISA and 400 bp for T-RFLP) were mixed and separated. An electropherogram with peaks of different sizes was obtained for each biopsy sample. Each peak represented an operational taxonomic unit (OTU) and was identified by its fragment size.

Fragment analysis. CEQ software version 9.0 was used to analyze the fragment data. Binning of 3bp or 2bp (Fisher et al. 1999; Brown et al. 2005; Ruan et al. 2006) was conducted when constructing OTU profiles of ARISA and T-RFLP data respectively. Only peaks with relative abundances higher than 1% were included (Brown et al. 2005). The incidence (presence/absence) data derived from the profiles were used for numerical analysis. OTU profiles were applied to JMP IN 5.1 (SAS Institute Inc.) and hierarchical clusters were built using Ward's method.

Table 2. Primers used

Target	Primer	Primer sequence (5' to 3')	Fragment size (bp)	Anneal. temp (°C)	Reference
Intergenic spacer primers for ARISA analysis					
ITS	ITSF	GTCGTAACAAGGTAGCCGTA	28-1558	55	(Cardinale et al. 2004)
	ITSReub	GCCAAGGCATCCACC			
Universal ribosomal primers for T-RFLP analysis					
16S	27f	GAAGAGTTTGATCATGGCTCAG	352	55	(Lane 1991)
	342r	CTGCTGCCTCCCGTAG			
Chaperonin 60 universal primers					
cpn60	H729	CGCCAGGGTTTTCCCAGTCACGACGAIIIIGCIGGIGAYGGIACIACIAC	549-567	50	(Hill et al. 2004)
	H730	AGCGGATAACAATTTCACACAGGAYKIYKITCICCRAAICCIGGIGCYTT			
	H1594	CGCCAGGGTTTTCCCAGTCACGACGACGTCGCCGGTGACGGCACCACCAC	656	58	(Hill et al. 2004)
	H1595	AGCGGATAACAATTTCACACAGGACGACGGTCGCCGAAGCCCGGGGCCTT			
	M13f*	CGCCAGGGTTTTCCCAGTCACGAC	-	-	(Hill et al. 2004)
	M13r	AGCGGATAACAATTTCACACAGGA			
Ribosomal sequencing					
16S	27f	AGAGTTTGATCMTGGCTCAG	1108	58	(Miyoshi et al. 2005)
	1100r	AGGGTTGCGCTCGTTG			
E. coli ABD typing					
chuA	ChuAf	CGGACGAACCAACGGTCAGGAT	281	70	(Kotlowski et al. 2007)
	ChuAr	TGCCGCCAGTACCAAAGACACG			
tsp	Tsp1	GGGAGTAATGTCGGGGCATTCAG	161	70	(Kotlowski et al. 2007)
	Tsp2	CATCGCGCCAACAAAGTATTACGCAG			
yja	Yjaf	CGTGAAGTGTCAGGAGACGCTGC	226	70	(Kotlowski et al. 2007)
	Yjar	TGCGTTCCTCAACCTGTGACAAACC			
Housekeeping genes for *E. coli* MLST					
adk	adkF	ATTCTGCTTGGCGCTCCGGG	583 (536)[†]	54	(Wirth et al. 2006)
	adkR	CCGTCAACTTTCGCGTATTT			
fumC	fumCF	TCACAGGTCGCCAGCGCTTC	806 (469)	65	(Wirth et al. 2006)
	fumCR	GTACGCAGCGAAAAAGATTC			

Table 2. Primers used

Target	Primer	Primer sequence (5' to 3')	Fragment size (bp)	Anneal. temp (°C)	Reference
	fumCR1[‡]	TCCCGGCAGATAAGCTGTGG			
gyrB	gyrBF	TCGGCGACACGGATGACGGC	911 (460)	68	(Wirth et al. 2006)
	gyrBR	ATCAGGCCTTCACGCGCATC			
	gyrBR1	GTCCATGTAGGCGTTCAGGG			
icd	icdF	ATGGAAAGTAAAGTAGTTGTTCCGGCACA	878 (518)	58	(Wirth et al. 2006)
	icdR	GGACGCAGCAGGATCTGTT			
mdh	mdhF	ATGAAAGTCGCAGTCCTCGGCGCTGCTGGCGG	932 (452)	68	(Wirth et al. 2006)
	mdhR	TTAACGAACTCCTGCCCCAGAGCGATATCTTTCTT			
purA	purAF	TCGGTAACGGTGTTGTGCTG	816 (478)	65	(Wirth et al. 2006)
	purAF1	CGCGCTGATGAAAGAGATGA			
	purAR	CATACGGTAAGCCACGCAGA			
recA	recAF	CGCATTCGCTTTACCCTGACC	780 (510)	62	(Wirth et al. 2006)
	recAF1	ACCTTTGTAGCTGTACCACG			
	recAR	TCGTCGAAATCTACGGACCGGA			
	recAR1	AGCGTGAAGGTAAAACCTGTG			
E. coli virulence factors					
SPATE	SPATE1	GAGGTCAACAACCTGAACAAACGTATGGG	617	57	(Kotlowski et al. 2007)
	SPATE2	CCGGCACGGGCTGTCACTTTCCAG			
agn43	Ag43F	TGACACAGGCAATGGACTATGACCG	317	67	(Kotlowski et al. 2007)
	Ag43R	GGCATCATCCCGGACCGTGC			
aidA	AIDA1	TATGCCACCTGGTATGCCGATGAC	545	69	(Kotlowski et al. 2007)
	AIDA2	ACGCCCACATTCCCCCAGAC			
PAI I	PAI1	TAGCTCAGACGCCAGGATTTTCCCTG	736	61	(Kotlowski et al. 2007)
	PAI2	CCTGGCGCCTGCGGGCTGACTATCAGGG			
Pap	PapF	CCGGCGTTCAGGCTGTAGCTG	97	65	(Kotlowski et al. 2007)
	PapR	GCTACAGTGGCAGTATGAGTAATGACCGTTA			
sfaD-E	SfaF	CGGAGGAGTAATTACAAACCTGGCA	408	64	(Martin et al. 2004)
	SfaR	CTCCGGAGAACTGGGTGCATCTTAC			
AfaE	AfaF	TATGGTGAGTTGGCGGGGATGTACAGTTACA	271	58	(Kotlowski et

Table 2. Primers used

Target	Primer	Primer sequence (5' to 3')	Fragment size (bp)	Anneal. temp (°C)	Reference
	AfaR	CCGGGAAAGTTGTCGGATCCAGTGT			al. 2007)
cnf	CnfF	AGTACTGACACTCACTCAAGCCGC	930	62	(Kotlowski et
	CnfR	GCAGAACGACGTTCTTCATAAGTATCACC			al. 2007)
vt1	Vt1F	CGCATAGTGGAACCTCACTGACGC	91	64	(Kotlowski et
	Vt1R	CATCCCCGTACGACTGATCCC			al. 2007)
vt2	Vt2F	CGGAATGCAAATCAGTCGTCACTCAC	265	65	(Kotlowski et
	Vt2R	TCCCCGATACTCCGGAAGCAC			al. 2007)
hlyA	HlyAF	TGCAGCCTCCAGTGCATCCCTC	355	63	(Kotlowski et
	HlyAR	CTTACCACTCTGACTGCGATCAGC			al. 2007)
eae	EaeF	CCAGGCTTCGTCACAGTTGCAGGC	300	66	(Kotlowski et
	EaeR	CGCCAGTATTCGCCACCAATACC			al. 2007)
bmaE	BmaEF	CTAACTTGCCATGCTGTGACAGTA	302	59	(Martin et al.
	BmaER	TTATCCCCTGCGTAGTTGTGAATC			2004)
E. coli fimbria subunit					
fimH	FimHf	CTGGTCATTCGCCTGTAAAACCGCCA	846 (679)	68	(Kotlowski et
	FimHr	GTCACGCCAATAATCGATTGCACATTCCCT			al. 2007)

ARISA, automated ribosomal intergenic spacer analysis; T-RFLP, terminal restriction fragment length polymorphism; ITS, ribosomal intergenic spacer; MLST, multi-locus sequence typing; SPATE, genes encoding serine protease autotransporter; *agn43*, the gene coding for antigen involved in *E. coli* autoaggregation; *aidA*, adhesin involved in diffuse adherence; PAI I, pathogenicity island I; *Pap*, *E. coli* Pap pili subunit and a part of pathogenicity island II; *sfa*, the gene coding for S-fimbria minor subunit; *afaE*, gene encoding AfaE-III afimbreal adhesin involved in *E. coli* diffuse adherence; *cnf*, cytotoxic necrotising factors Cnf1 and Cnf2; *vt1* and *vt2*, verocytotoxins 1 and 2; *hlyA*, α hemolysin; *eae*, intimin; *bmaE*, gene for M-agglutinin subunit
I=inosine, Y=C or T, R=G or A, K=G or T
[*] M13f and M13r were used to sequence the cpn60 fragment.
[†] Numbers in parenthesis are the size of the sequenced fragment.
[‡] MLST primers ending with "1" are the ones used for sequencing.

Richness, diversity, and similarity measures. Incidence-based richness indicators; Chao2, ICE (Incidence-based Coverage Estimator) and MM (Michaelis-Menten function) Mean, along with Shannon and Simpson diversity indices, were calculated using EstimateS 7.5 (Colwell 2005). Several estimators were selected because if indices follow the same trend regardless of the calculation method, the results are likely to be robust. An upper abundance limit of 5 was used to determine rare or infrequent species. The order of the samples was randomized 500 times for each run, to reduce the effect of sample order. Tukey's multiple comparison test (SAS Institute Inc.) was applied to detect significant differences among experimental groups. Multiple incidence-based similarity indices; Bray-curtis and Jaccard, were calculated using SPADE 2.1 (Chao et al. 2005).

Bioinformatic analysis of T-RFLP data. MiCA (Microbial Community Analysis, version 3; Department of Biological Sciences, University of Idaho [http://mica.ibest.uidaho.edu/]) was used to build a putative reference database of probable T-RFs of the gut. For this purpose we incorporated 16S rDNA clone libraries of near complete sequence of gut microorganisms found in human (Eckburg et al. 2005), swine (Leser et al. 2002), mouse (Ley et al. 2005), and ruminants (Nelson et al. 2003; Ozutsumi et al. 2005a; Ozutsumi et al. 2005b) into the MiCA which we called the H.Q. database. This greatly facilitates analysis by excluding the T-RFs that are unlikely to occur in the gut (Rappe et al. 2003; Eckburg et al. 2005). Primers 27f and 342r plus *Hha*I restriction digestion were applied to the H.Q. database of MiCA in a virtual digest (ISPaR). A reference library for our study was constructed, and exported to PAT (Phylogenetic Assignment Tool) (Kent et al. 2003). Concurrently, using T-RFLP data

obtained from CEQ software (fragment sizes and peak areas), various profiles of interest were developed with reference to disease condition, site, and the inflammation state of the biopsies. Each of these specific profiles was compared to the assigned reference library through the T-RFLP phylogenetic assignment tool, and a library of probable accession numbers was obtained for each profile. These libraries were entered into the hierarchical browser of RDP-II (Ribosomal Database Project) (Cole et al. 2005) and converted to Genbank format. The resulting libraries were then assigned to the library compare tool of RDP-II. T-RFs of the same size were in many cases ambiguous in their assignment of taxonomic rank. Such T-RFs with multiple accession numbers were assigned to taxonomic rank according to phylum, class, order, and family. Based on this analysis, reported values were expressed as a proportion of phylogenetic lineage for each library. Statistical significance were calculated using the LSD multiple comparison test (SAS Institute Inc.).

2. Characterization of *E. coli* isolated from IBD gut biopsy tissues

2.1. Bacterial isolation

All bacterial cultivation was done with untreated biopsy tissues, and no procedures were used to wash the tissues or to remove the mucus. Each biopsy specimen was resuscitated in 1 ml of 100 mM buffered peptone water (pH 7.2, Difco Laboratories, Detroit, MI). One hundred µl of the suspension were transferred onto the Chromogenic *E. coli*/coliform Agar (Oxoid CM0956, Nepean, Ontario, Canada) directly and after overnight incubation at 37 °C. Agar plates were dried and incubated at 37 °C for 18-24 h. Single bacterial colonies that appeared in different colors (purple, pink, white, straw, blue, green, etc) or shapes (round, irregular, rough, etc) on chromogenic agar were isolated and stored in glycerol at −80 °C.

2.2. Bacterial DNA extraction

Bacterial isolates were suspended in 1.5 ml Luria Bertani (LB) broth and grown for 24 h at 37 ºC with shaking. Following centrifugation at 2,000 × *g* for 10 min, the supernatant was discarded and the pellet was suspended in 500 µl 1X TE buffer [pH 8.0], 30 µl SDS [10 %], and 15 µl proteinase K [20mg/ml]. The mixture was incubated in a water bath at 37 ºC for 1 h followed by the addition of 100 µl of 5M NaCl and 80 µl of cetyltrimethylammonium bromide (CTAB). After 10 min of incubation at 65 °C, the cell

lysate was extracted with an equal volume of phenol:chloroform:isoamyl alcohol (25:24:1) followed by a second extraction with chloroform:isoamyl alcohol (24:1). The aqueous phase, containing the DNA was mixed with 0.6 volume isopropanol. The precipitate was centrifuged for 10 min at $20{,}000 \times g$. The DNA pellet was washed twice with 600 µl of 70 % ethanol. The resulting pellet was dried, suspended in 100 µl 1X TE buffer (pH 8.0) and stored at -20 °C.

2.3. Bacterial identification

16S rDNA amplification. Partial fragments of 16S ribosomal DNA were amplified using primers 27f and 1100r (Table 2). Each PCR reaction contained 50 ng of genomic DNA, 5 µl of 10× PCR buffer, 1.5 µl Mgcl$_2$ (50mM), 1 µl dNTPs (10mM), 1 µl of each primer (25 pmol/µl), 2 units of Taq polymerase and water to 50 µl. Thermal cycling conditions were: One cycle of initial denaturation (94 °C, 2 min); 35 cycles of denaturation (94 °C, 1 min), annealing (58 °C, 1 min), and extension (72 °C, 1.5 min); followed by a final extension (72 °C, 5min). The amplified fragments were subjected to agarose gel electerophoresis to ensure the amplicon is pure, single and of adequate quantity.

Amplification of *cpn60* gene. A partial fragment of gene encoding 60-kDa type I chaperonin protein subunit (CPN60) were amplified using universal primers listed in Table 2. Primers H1594 and H1595 were employed when amplification with H729 and H730 was unsuccessful (Hill et al. 2004). PCR conditions for the amplification of *cpn*60 were set as has been previously described (Hill et al. 2002): each 50 µl reaction mixture

consisted of 50 ng of chromosomal DNA, 5 µl of 10× buffer without Mg, 2 µl Mgcl$_2$ (50mM), 1 µl dNTPs (10mM), 2 µl of each primer (10 pmol/µl), 2 units of Taq polymerase and water to 50 µl. The following thermal cycling condition was used: denaturation for 5 min at 94 °C, followed by 35 cycles of 30 sec at 94 °C, 30 sec at annealing temperature (see Table 2), and 45 sec at 72 °C followed by the final extension of 5 min at 72 °C. The quality and purity of amplified fragments were check by agarose gel electrophoresis.

Sequence analysis. The amplicons were sequenced in both directions at the McGill University Genome Quebec Innovation Centre (Montréal, Québec, Canada). Primers M13 forward and reverse (Table 2) were used for sequence determination of *cpn60*. The forward and reverse sequences were aligned using ClustalW (Larkin et al. 2007). The consensus sequence was identified using sequence comparison with the RDP (Ribosomal Database Project) SEQ MATCH (Cole et al. 2009) and NCBI (National Center for Biotechnology Information) nucleotide BLAST for 16S rDNA, and cpnDB (chaperonin database) (Hill et al. 2004) and NCBI nucleotide BLAST for *cpn*60 sequence.

2.4. Enumeration of *E. coli*

Study subjects. Enumeration was performed on *E. coli* isolated from 16 CD, 22 UC and 20 HC biopsy specimens described in section 1 of the materials and methods.

Enumeration. To ensure that as many *E. coli* cells were detected as possible, resuscitation was performed by incubation of the biopsy specimen in 1 ml of 100 mM

buffered peptone water for 16 h at 37 °C. Buffered peptone water was also used to make decimal serial dilutions in 2 ml deep-well plates. Plates were covered with aluminum foil and sterilized at 121ºC and 15 psi for 15 min. A 100 µl of homogenized resuscitated biopsy was inoculated into 900 µl of 2% buffered peptone water and serially diluted to 10^{-7}. Dilutions from 10^{-1} to 10^{-7} were plated by dispensing of 10 droplets (10 µl each) onto solid Chromogenic agar. Following absorption of the droplets into the medium, plates were inverted and incubated at 37 °C for 18 h. *E. coli* were counted in the dilutions that gave 3 to 30 colonies and an appropriate dilution factor was applied to determine the number of *E. coli* attached to gut mucosa.

2.5. Biochemical profiling of *E. coli*

Study subjects. All *E. coli* isolated from 58 biopsies described in section 1 of the materials and methods were subjected to biochemical profiling.

Biochemical profiling. *E. coli* strains were phenotypically characterized using the API 20E system (bioMérieux, St. Laurent, Québec) according to the manufacturer's guidelines. The API 20E system consists of a plastic strip of 20 individual, miniaturized tests tubes (cupules) each containing a different reagent used to determine isolate's metabolic capabilities. The reagents in the API 20E cupules are purposely designed to test for the presence of products of bacterial metabolism specific to the family *Enterobacteraceae*. *E. coli* isolates were tested for amino acid hydration (ADH, arginine dihydrolase), amino acid deamination (TDA, tryptophane deaminase), amino acid decarboxylation (LDC, lysine decarboxylase; ODC, ornithine decarboxylase), citrate utilization (CIT), enzymatic activity (ONPG, β-galactosidase; URE, urease; and GEL,

gelatinase), H_2S production (H_2S), indole production (IND), acetoin production (VP), and carbohydrate fermentation and oxidation characteristics (GLU, glucose; MAN, manitol; INO, inositol; SOR, sorbitol; RHA, rhamnose; SAC, sucrose; MEL, melibiose; AMY, amygdalin; and ARA, arabinose). The inoculum was prepared by suspending a single well-isolated fresh (18-24 h) *E. coli* colony in 5 ml of sterile distilled water. Each cupule was inoculated with a suspension of a pure bacterial culture, rehydrating the dried reagent in each tube. Some of the tubes were completely filled (CIT, VP and GEL), whereas others were topped off with mineral oil so that anaerobic reactions could be carried out (ADH, LDC, ODC, H2S, URE). Each strip was then incubated in a small, plastic humidity chamber for 18-24 hours at 37 °C. After incubation, each tube was assessed for a specific color change indicating the presence of a metabolic reaction (Figure 1). Some of the cupule contents changed color due to pH differences; others (TDA, IND, and VP) contained end products that had to be identified using additional reagents (Figure 1). The IND test was performed the last since the gaseous product could interfere with the interpretation of other tests.
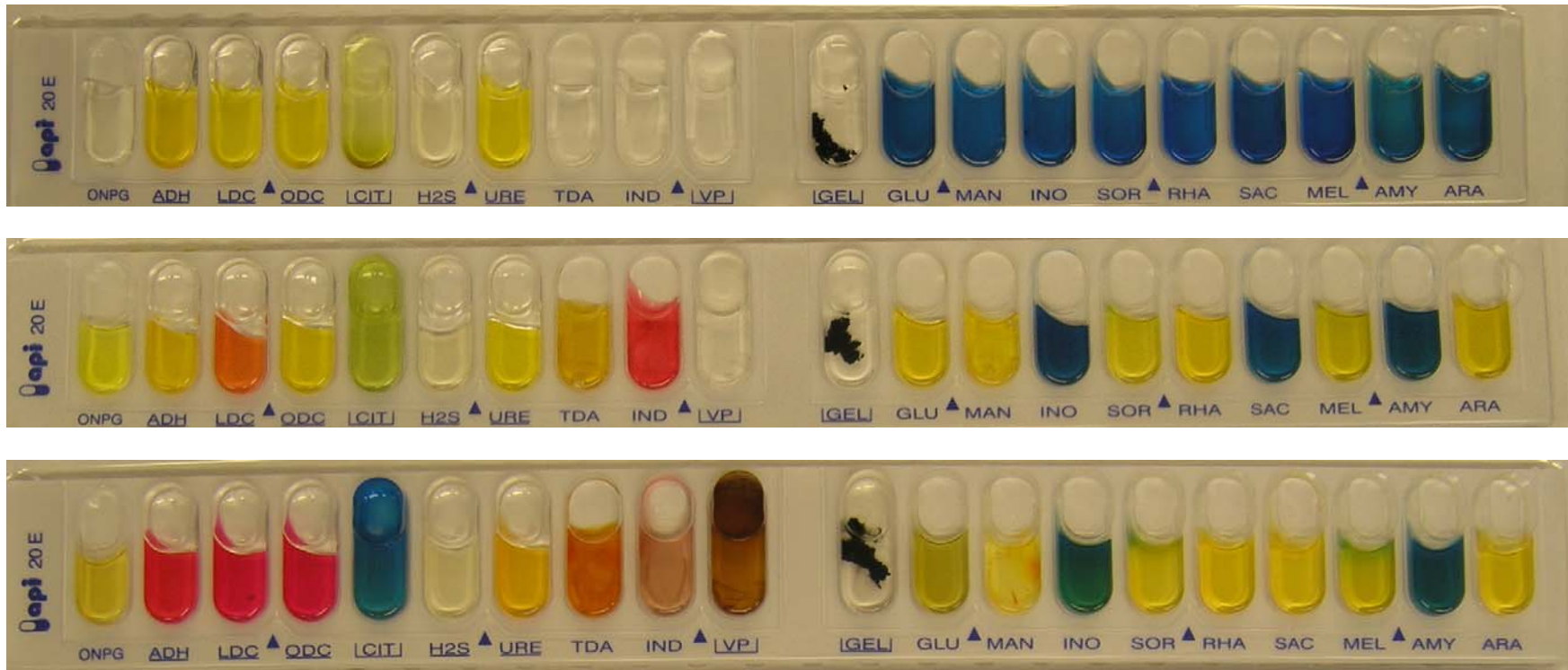
Figure 1. Biochemical profiling of *E. coli* using API 20E system. The picture shows color changes at three different stages of the experiment: After inoculation with bacterial solution (stripe on the top), after 18-24 h incubation at 37 °C (strip in the middle), and after the addition of TDA, IND, and VP reagents (strip at the bottom).

2.6. *E. coli* serotyping

Study subjects. Biopsies were collected at first diagnosis from IBD patients at sites across Canada, as a part of the Crohn's and Colitis Foundation of Canada tissue bank project. Subjects with suspected IBD signed an informed consent prior to undergoing endoscopy plus biopsy. Once subjects were enrolled and their endoscopies and biopsies were consistent with a diagnosis of either CD or UC, their clinical data were reviewed to confirm that they had a diagnosis of definite IBD. HC included those undergoing colonoscopy for other reasons such as screening. A total of 59 biopsies were taken from the right colon of 23 CD patients, 11 UC patients and 25 HC. All the biopsies were stored at −80 °C until further analysis. Inflammation was assessed both endoscopically and histologically.

Serotyping. All *E. coli* strains were serotyped for their O (lipopolysaccharide) and H (flagellar) antigens at the Public Health Agency of Canada, Laboratory for Foodborne Zoonoses, Reference Laboratories, Guelph, Ontario.

2.7. *E. coli* phylotyping

E. coli isolates were typed into A, B1, B2, and D phylogenetic types using modified triplex PCR method (Clermont et al. 2000; Kotlowski et al. 2007). Primers listed in Table 2 were used to amplify *chuA*, *yjaA* and TSPE4.C2 from *E. coli* genomic DNA. PCR reactions and thermal cycling were as expressed for 16S rDNA amplification, section 2.3 of materials and methods. Isolates belonging to A or B1 groups were negative for the *chuA* gene. These isolates were classified as A if they were negative for

TSPE4.C2 fragment or B1, if they were positive for TSPE4.C2. In contrast, B2 and D isolates were positive for the *chuA* gene and was separated into B2 and D groups based on the presence or absence of the *yjaA* gene, respectively (Clermont et al. 2000).

2.8. Phylogenetic analysis of *E. coli* using MLST technique

Study subjects. *E. coli* isolated from 22 UC, 16 CD and 20 HC biopsies described in section 1 of the materials and methods were used for phylogenetic analysis. *E. coli* strains not associated with IBD were also included as controls: 2 strains from cattle feces; 2 strains from swine feces; the probiotic *E. coli* Nissle 1917 (Schultz 2008), the nonpathogenic *E. coli* K12; and enterohaemorrhagic *E. coli* (EHEC) O157:H7.

Amplification and sequencing of the housekeeping genes. *E. coli* genomic DNA was extracted as described in section 2.2 of the materials and methods. Internal fragments of seven housekeeping genes (*adk*, *fumC*, *gyrB*, *icd*, *mdh*, *purA* and *recA*) (Wirth et al. 2006) were amplified using the primers listed in Table 2. Each PCR reactions contained 50 ng of genomic DNA, 5 µl of 10× PCR buffer, 2 µl Mgcl$_2$ (50mM), 1 µl dNTPs, 1 µl of each primer (25 pmol), 2 units of Taq polymerase and water to 50 µl. Thermal cycling conditions were: One cycle of denaturation (95 °C, 2 min); 30 cycles of denaturation (94 °C, 1 min), annealing (see Table 2, 1 min), and extension (72 °C, 2 min); followed by a final extension (72 °C, 5min). PCR products were cleaned by repeated (two times) ethanol precipitation and sequenced in both directions at the McGill University Genome Quebec Innovation Centre (Montréal, Québec, Canada).

Sequence analysis. The forward and reverse sequences were aligned using ClustalW (Chenna et al. 2003) and the consensus sequence was entered into the MLST.net *Escherichia coli* database (Wirth et al. 2006). Each sequence that was not homologous to an already existing sequence in the database was assigned a new allelic number. The allelic profile of each *E. coli* isolate was then generated, which consisted of seven allele numbers, each associated with one housekeeping gene. Unique combinations of allelic numbers were assigned a sequenced type (ST). Related STs were grouped into ST complexes, and each ST complex was defined as two or more STs that share at least six alleles. STs that did not match the inclusion criteria were referred to only by their ST number (Wirth et al. 2006). Sequences of housekeeping genes of UPEC CFT073 (NC004431) and APEC O1:K1:H7 (NC008563) were also included for comparison.

Statistical analysis. eBURST analysis (Feil et al. 2004) developed for MLST analysis was performed. Strains were grouped together if two of the seven alleles were homologous. STs that differed from each other by one or two alleles were designated single locus variants (SLV) or double locus variants (DLV) respectively. Satellite (SAT) were referred to STs that differ by more than two alleles. Phylogenetic analysis was carried out using MEGA4 (Tamura et al. 2007). Concatenated sequences of seven loci were clustered using the unweighted pair group method with arithmetic means (UPGMA) and the robustness of clusters was tested by bootstrapping (1,000 as default). Discriminant multivariate analysis was used to examine possible relationships between new alleles and STs and IBD using JMP IN 5.1 (SAS Institute Inc.). Chi-square analysis based on Mantel Haenszel method (Epi Info 6.04, CDC, Atlanta, Georgia, USA) was

performed to examine the relationship between amino acid substitutions in housekeeping genes and the disease state or *E. coli* phylotype.

2.9. Virulence characterization of the *E. coli*

Study subjects. *E. coli* isolated from newly diagnosed IBD patients and HC described in section 2.6 of the materials and methods were assessed for virulence factors.

***E. coli* virulence factor identification.** Single well-isolated *E. coli* colonies were isolated from biopsy specimens as described in section 2.1 of this chapter. The genomic DNA of *E. coli* was extracted following the method explained in section 2.2. The presence of 14 virulence factors (SPATE, *agn43*, *aida*, PAI I, *pap*, *sfa*, *afaE*, *cnf*, *vt1*, *vt2*, *hlyA*, *eae*, *bmaE*, and *fimH*) was tested using primers listed in Table 2. Primers for serine protease autotransporters (SPATE) and pathogenicity island I (PAI I) were designed based on conserved regions of major groups of those elements (Kotlowski et al. 2007). PCR reactions and thermal cycling were as expressed for 16S rDNA amplification, section 2.7 of materials and methods. The amplified products were subjected to electrophoresis using agarose gel.

Statistical analysis. Chi-square analysis based on Mantel Haenszel method (Epi Info 6.04, CDC, Atlanta, Georgia, USA) was performed to detect significant relationships between the disease state, *E. coli* phylotypes, and virulence factors.

2.10. Analysis of *fimH* subunit of *E. coli* fimbria gene

Study subjects. Sequences of *fimH* from all *E. coli* strains, isolated from patients with established IBD (materials and methods section 2.8), newly diagnosed IBD patients (materials and methods section 2.6), and HC were studied.

***fimH* amplification and sequencing.** *E. coli* cultivation and DNA extraction were done as previously described (materials and methods section 2.1 and 2.2). Genes encoding FimH subunit of type I fimbriae were amplified using primers listed in Table 2 and sequenced in both directions at the McGill University Genome Quebec Innovation Centre (Montréal, Québec, Canada). PCR was carried out as described in section 2.7 of this chapter.

Sequence analysis. Both forward and reverse sequences were aligned using ClustalW (Chenna et al. 2003) and the consensus sequence was used for further analysis. The AIEC LF82 (AF288194) sequences for *fimH* were obtained from the literature (Boudeau et al. 2001) and included to the analysis for comparison purposes.

Statistical analysis. Phylogenetic analysis was carried out using UPGMA algorithm and MEGA4 program (Tamura et al. 2007). Chi-square analysis based on Mantel Haenszel method (Epi Info 6.04) was performed to examine the relationship between amino acid substitutions in FimH subunit of fimbriae and the disease state or *E. coli* phylotype.

2.11. *E. coli* gene content analysis using microarray technology

Study subjects. A total of 16 *E. coli* strains from CD (n=6), UC (n=5), and HC (n=5) were selected for this study. These strains were isolated from both established and newly diagnosed IBD patients (materials and methods sections 1 and 2.6). UPEC CFT073 and *E. coli* Nissle 1917 were also included to the study as controls.

DNA extraction. For this study, the genomic DNA of *E. coli* was isolated using QIAamp® DNA mini kit (catalog no. 51304) (QIAGEN®, Mississauga, ON, Canada). *E. coli* isolates were recultivated from their glycerol stocks into 1.5 ml of LB broth and were incubated overnight at 37 °C. Microcentrifuge tubes containing *E. coli* cultures were centrifuged for 5 min at 5,000 × *g*. The supernatant was discarded and the pellet was suspended in 180 μl of buffer ATL. Twenty μl of proteinase K (provided with kit) was added to the mixture, mixed by vortexing and incubated at 56 °C for 1 h. After the addition of 20 μl of RNase (20 mg/ml) and incubation at room temperature for 5 min, 200 μl of buffer AL was added to the sample. Following incubation at 70 °C for 10 min, samples were washed with 200 μl of ethanol (96-100%) at 6,000 × *g* through QIAamp mini spin column. The mixture was washed two more times through column with 500 μl of buffers AW1 and AW2 at 6,000 × *g* for 1 min and 20,000 × *g* for 3 min respectively. To precipitate the DNA, column was incubated with 60 μl buffer AE for 10 min at room temperature before 1 min centrifugation at 6,000 × *g*, followed by second elution step with 60 μl of buffer AE. DNA eluted in buffer AE (120 μl) was stored at −20 °C until further analysis.

Enzymatic digestion of genomic DNA. DNA lysis buffer (40 mM Tris-HCl pH 7.5, 6 mM MgCl$_2$) was prepared. Digestion of *E. coli* genomic DNA was carried out by

incubating 120 μl of genomic DNA, 30 μl of lysis buffer and 1.5 μl of deoxyribonuclease I (DNase I 10,000 u/ml RNase-free, GE Healthcare Life Sciences) at 37 °C for 5 min. Fragmented DNA was immediately placed on ice. Successful fragmentation was confirmed by running three μl of fragmented DNA along with three μl of original genomic DNA on a 2% agarose gel to verify the presence of a smear of less than 1,000 bp in size (Figure 2). Fragmented DNA was stored at –20 °C.

Fragmented DNA cleanup. Fragmented DNA was cleaned from buffers, enzymes, and inhibitors as follows: DNA was mixed with 1/10 volume of 3 M NaAc pH 5.2, precipitated by two portions of cold ethanol (95%), and placed at –20 °C overnight. After centrifugation at $20,000 \times g$ for 15 min at 4 °C, the supernatant was removed and the pellet was washed two times with 70% ethanol. The pellet was then dried and suspended in 55 μl of AE buffer.
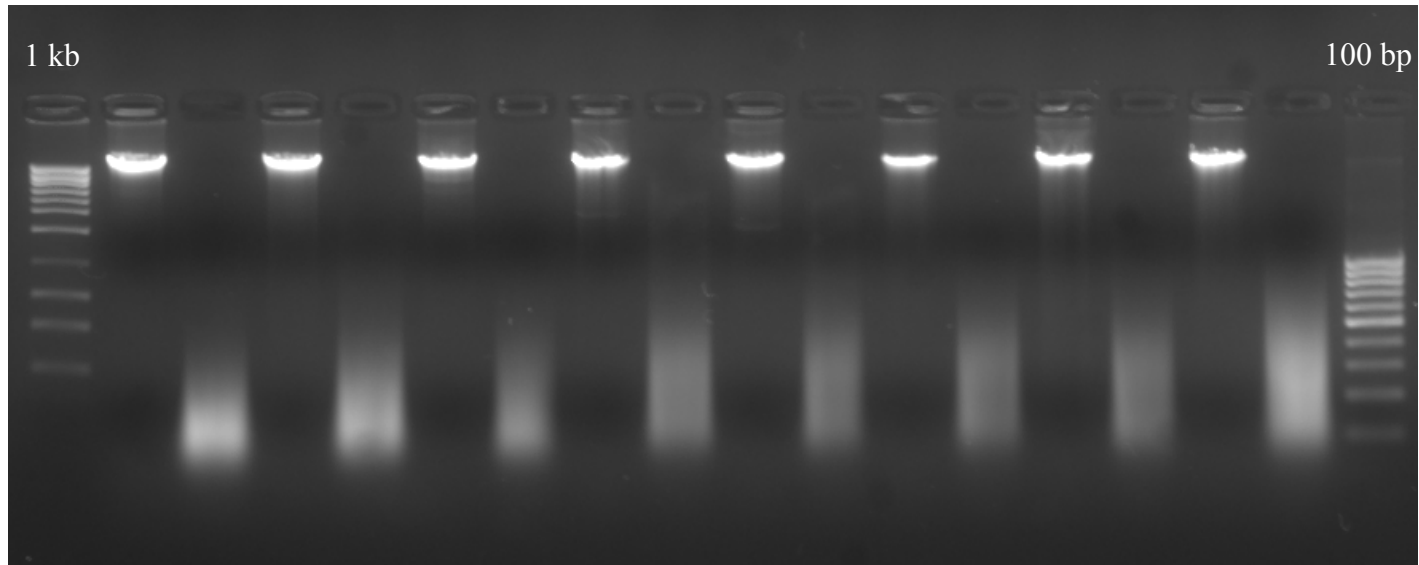
Figure 2. Typical example of fragmented DNA on 2% agarose gel. The original genomic DNA and fragmented DNA for each *E. coli* were run side by side. The smear confirms the fragmentation of DNA to pieces smaller than 1,000 bp.

DNA labeling. During this stage, the fragmented samples were labeled using the GeneChip® DNA Labeling Reagent (Affymetrix Inc., Santa Clara, CA, USA). Labeling mixture was prepared for each sample by combining 7 μl of GeneChip® DNA Labeling Reagent (7.5mM), 7 μl of 10× NE buffer 4 (New England Biolabs), 7 μl of $CoCl_2$ (2.5mM, New England Biolabs), 1 μl of terminal deoxynucleotidyl transferase (TdT 20,000 u/ml, New England Biolabs), and 48 μl of fragmented DNA to the total volume of 70 μl. The solution was mixed by repeated pipetting, placed in a pre-heated thermal cycler block (PTC-100, MJ research), and run at 37 °C for 30 min and 70 °C for 10 min. The mixture was hold at 4 °C.

Target hybridization. Hybridization mixture was prepared as follows: 140 μl of Tetramethyl Ammonium Chloride solution (TMACL 5M, Sigma-Aldrich Inc., St. Louis, MO, USA), 13 μl of DMSO (100%), 13 μl of Denhardt's solution 50× (1% solution of BSA, Ficoll® and PVP, Sigma), 12 μl of MES 12× (1.25 M MES and 0.89 M [Na^+], Sigma), 3 μl of ethylenediaminetetraacetic acid disodium salt (EDTA 0.5 M, 0.2 μm filtered, DNase, RNase, and protease free, Sigma), 3 μl of Herring Sperm DNA (HSDNA 10 mg/ml, Promega, Madison, WI, USA), 3 μl of Human Cot-1 DNA® (1.0 mg/ml, Invitrogen, Eugene, OR, USA), 2 μl of Oligo Control Reagent (OCR) 0100 (Control Oligo B2 3 nM, Affymetrix), and 1 μl of Tween-20 3% (surfact_Amps® 20, Pierce Chemical, Rockford, IL, USA) were mixed to the total volume of 190 μl. The hybridization mix (190 μl) was added to the labeled DNA sample (70 μl) to the total volume of 260 μl, mixed properly and run at 95 °C for 10 min in a thermal cycler, and hold at 49 °C. Ninety five μl of each sample were loaded onto a chip (GeneChip® *E. coli* Genome 2.0 array, Affymetrix) allowing one bubble. The injection sites of the array

(septa) were covered with Tough-Spot label dots (USA scientific, Orlando, FL) and the chip was immediately placed in the hybridization oven (GeneChip Hybridization Oven 640, Affymetrix) and allowed to rotate at 60 rpm for 16-18 h at 49 °C. The remaining of the cocktails were stored at –80 °C.

Affymetrix GeneChip® *E. coli* Genome 2.0 arrays. The *E. coli* genome 2.0 array includes about 10,000 probes for all 20,366 genes present in four strains of *E. coli*, namely non-pathogenic *E. coli* strain K12 MG1655 and pathogenic strains UPEC CFT073, EHEC O157:H7 EDL933, and EHEC O157:H7 Sakai. The Affymetrix GeneChip® *E. coli* Genome 2.0 array tiles probes over the entire open reading frame (ORF) of *E. coli* including 700 intergenic regions.

Washing and staining the arrays. For both washing and staining stages, $1\times$ array holding buffer (final concentration was 100 mM MES, 1 M [Na$^+$], and 0.01% Tween-20) was prepared. Ten ml of the buffer contained 7.31 ml of nuclease-free water (autoclaved and 0.2 μm filtered, Ambion®, Austin, TX, USA), 1.85 ml of NaCl 5M (autoclaved and 0.2 μm filtered, Ambion®), 830 μl of MES $12\times$, and 10 μl of Tween-20 10%. After 16-18 h of hybridization, the cocktail was removed from the array and the array was completely filled with $1\times$ array holding buffer.

Wash A (non-stringent wash buffer) and Wash B (stringent wash buffer) were prepared as follows: For 1000 ml of Wash A ($6\times$ SSPE, 0.01% Tween-20), 300 ml of $20\times$ SSPE (0.2 M phosphate buffer pH 7.4, 2.98 M NaCl, 0.02 EDTA, Sigma) and 1 ml of 10% Tween-20 were added to 699 ml of nuclease-free water. For Wash B ($0.6\times$ SSPE, 0.01% Tween-20, pH 8.0), 30 ml of $20\times$ SSPE and 1 ml of 10% Tween-20 were mixed in 969 ml of nuclease-free water. Both Washes A and B were stored at room temperature

until their further use as intake buffers for fluidics station (GeneChip® Fluidics station 450, Affymetrix).

For each sample, 1188 μl of staining buffer was prepared by combining 800 μl of nuclease-free water, 360 μl of 20× SSPE, 24 μl of 50× Denhardt's solution, and 4 μl of 3% Tween-20. Three staining solutions were arranged for each array. Staining solutions 1 and two were made by adding 6 μl of streptavidin phycoerythrin (SAPE 1 mg/ml, Invitrogen) and 6 μl of biotinylated anti-streptavidin antibody (0.5 mg/ml, Vector laboratories, Burlingame, CA, USA) to 594 μl of staining buffer respectively. Eight hundred and twenty μl of 1× array holding buffer was used as staining solution 3.

The fluidics station 450 was used to wash and stain the arrays. The station was operated using GeneChip Operating Software (Affymetrix). The fluidics station was turned on and Washes A and B were positioned into the intake buffer reservoirs A and B respectively. Priming the fluidics station was done by running Prime_450 protocol under the fluidics menu of GeneChip operating software. After the priming was finished, a new experiment was opened from the file menu and the experiment's name, sample information and chip barcode were entered and saved. The station was then loaded with array chips as well as vials containing staining solutions 1, 2, and 3. Mini_prok2v1_450 protocol was run and the instructions on the fluidics display were followed.

Scanning of arrays. GeneChip® scanner 3000 G7 (Affymetrix) was used to scan the arrays. The scanner was turned on 15 min before use to allow the warm up of the laser. Label dots were applied to each septa and array cartridges were inserted into the scanner. Arrays were scanned and array images (.dat file) were recorded for analysis.

Image analysis. Saved array images were analyzed using GeneChip Operating Software. The B2 oligo, a component of the OCR 0100, served as a positive hybridization control and was used by the software to place a grid over the image. The B2 oligo was responsible for the alternative patterns of intensity on the boarders and the checkerboard pattern at each corner (Figure 3). Presence of each array gene in sample was determined by comparing each probe's signal intensity to the average background and average noise calculated for each image.

Statistical analysis. A list of identified virulence genes of *E. coli* was obtainted from literature (Johnson et al. 2000; Kaper et al. 2004; Afset et al. 2006; Bruant et al. 2006; Ewers et al. 2007). Chi-square analysis based on Mantel Haenszel method (Epi Info 6.04) was performed to examine the significant associations between virulence characteristics of UPEC and EHEC strains and *E. coli* isolated from IBD, as well as the association of non-pathogenic characteristics of K12 MG1655 and *E. coli* from HC. Chi-squares and probabilities were calculated separately for each gene/probe and therefore biases that are usually introduced by multiple comparisons were avoided.

Figure 3. A hybridized array. The array's name, hybridized dots, alternating patterns of intensities on the borders, and the checkerboard pattern at the corner are depicted.

RESULTS

1. Microbial diversity of gut biopsy tissues in IBD

A total of 58 biopsies isolated from 41 individuals, including controls were analyzed (Table 1). Diagnosis of CD or UC was based on histological examination, and eight non-inflamed CD and 10 non-inflamed UC tissues were included in the sample collection. For 3 CD and 3 UC patients, biopsies were obtained from both macroscopically inflamed and non-inflamed GI locations whereas for the remaining of the patients biopsies were taken from either inflamed or non-inflamed sites.

Universal primers amplified the intergenic spacer region between the 16S and 23S rDNA genes and separation by capillary electrophoresis (ARISA) indicated that non-inflamed and inflamed tissues had different profiles (Figure 4). A total of 114 different size fragments, with the average of 11 peaks per profile, were observed. The presence or absence of unique peaks from ARISA was used to construct a matrix of incidence values (0 or 1) that were numerically analyzed to produce a dendrogram of relationships (Figure 6). ARISA profiles from HC, UC (including non-inflamed UC tissues), and CD (including non-inflamed CD tissue), did not cluster as tightly as non-inflamed versus inflamed. If a clustering distance of 80% was used, the majority of inflamed or non-inflamed tissues aggregated into a single cluster for 14 of the 17 clusters. However, for the largest cluster, constituting 29 observations, a match of only 55.1% was obtained (Figure 5).
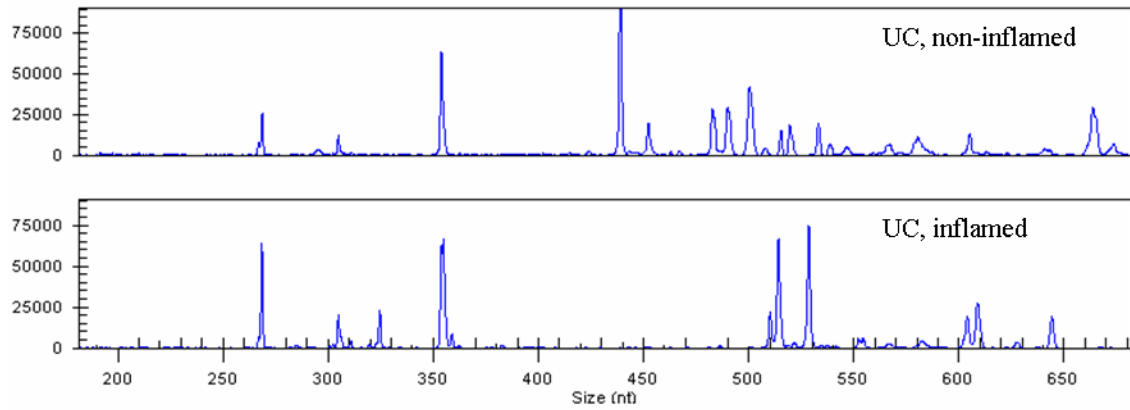
Figure 4. Typical automated ribosomal intergenic spacer analysis (ARISA) electropherogram of microbial communities from inflamed and non-inflamed ulcerative colitis tissues. The *x*-axis represents the size of the intergenic spacer (nt) whilst the *y*-axis represents the fluorescent intensity.
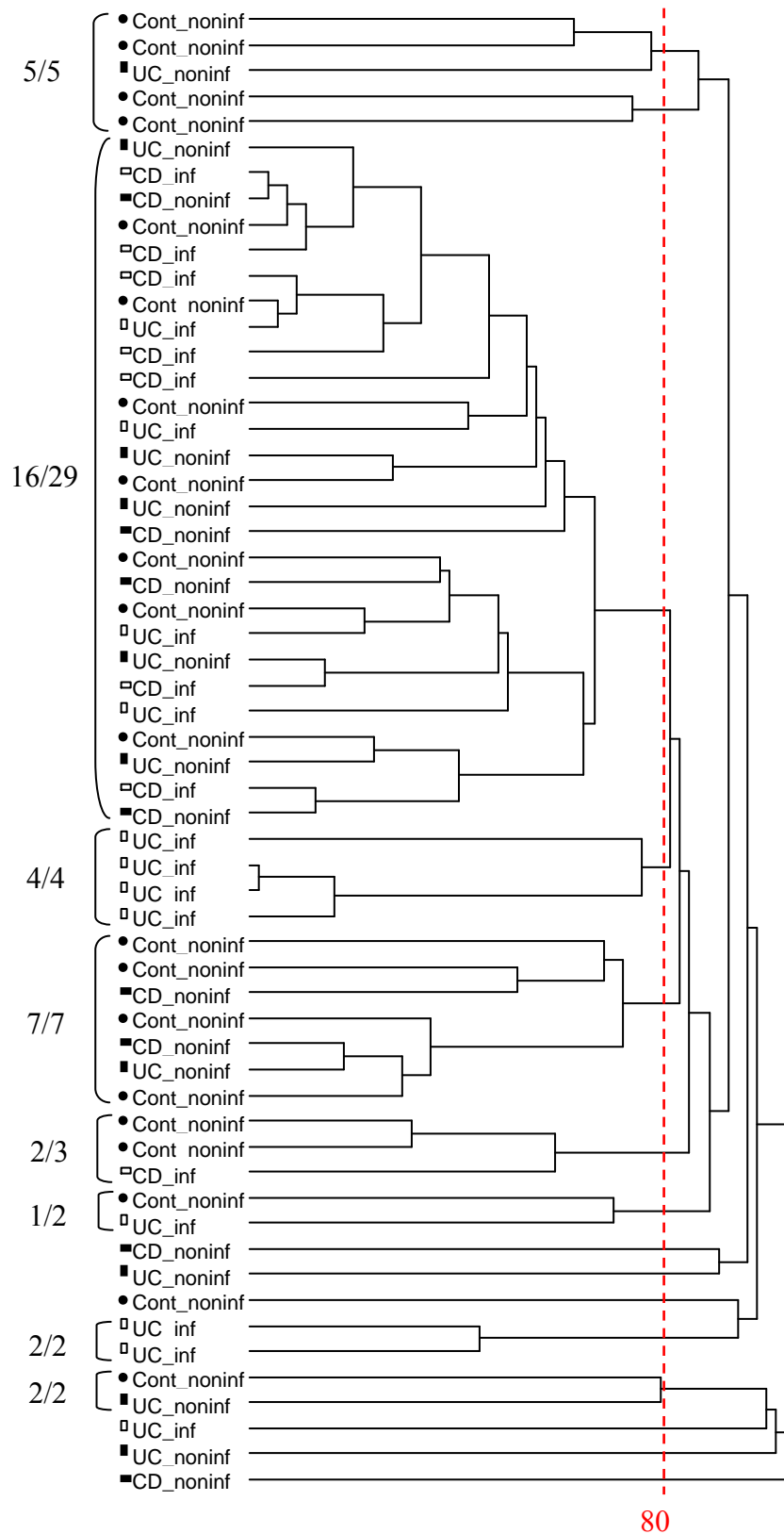
Figure 5. Phylogenetic tree based on ARISA incidence profiles. The values on the y-axis indicate the proportion of the total number of observations in the cluster that come from inflamed or non-inflamed tissue.

T-RFLP profiles were generated when 16S rDNA fragments amplified with universal primers (27f and 342r), digested (*Hha*I) and separated by capillary electrophoresis. Sixty eight distinct fragments were obtained in total, with the average of 9 fragments per sample. Non-inflamed tissues and diseased tissues had different profiles (Figure 6). As was the case with ARISA, clustering was not robust when tissues were grouped by their clinical diagnosis, and a much higher level of clustering was attained if tissues were separated based on inflammation state (Figure 7). At a 70% distance cut-off, more than 90% of tissues grouped into inflamed or non-inflamed clusters.
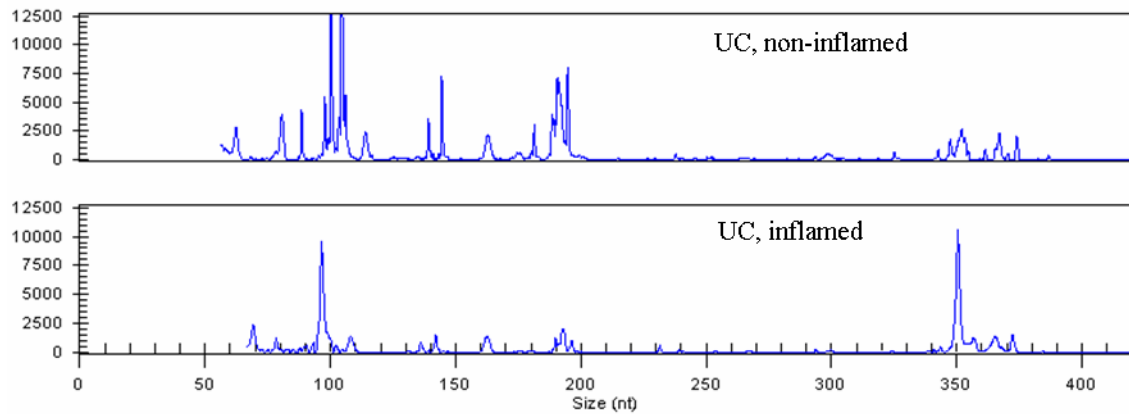
Figure 6. Typical terminal restriction fragment length polymorphisms (T-RFLP) electropherogram of microbial communities from inflamed and non-inflamed ulcerative colitis tissues. The *x*-axis represents the size of the intergenic spacer (nt) whilst the *y*-axis represents the fluorescent intensity.
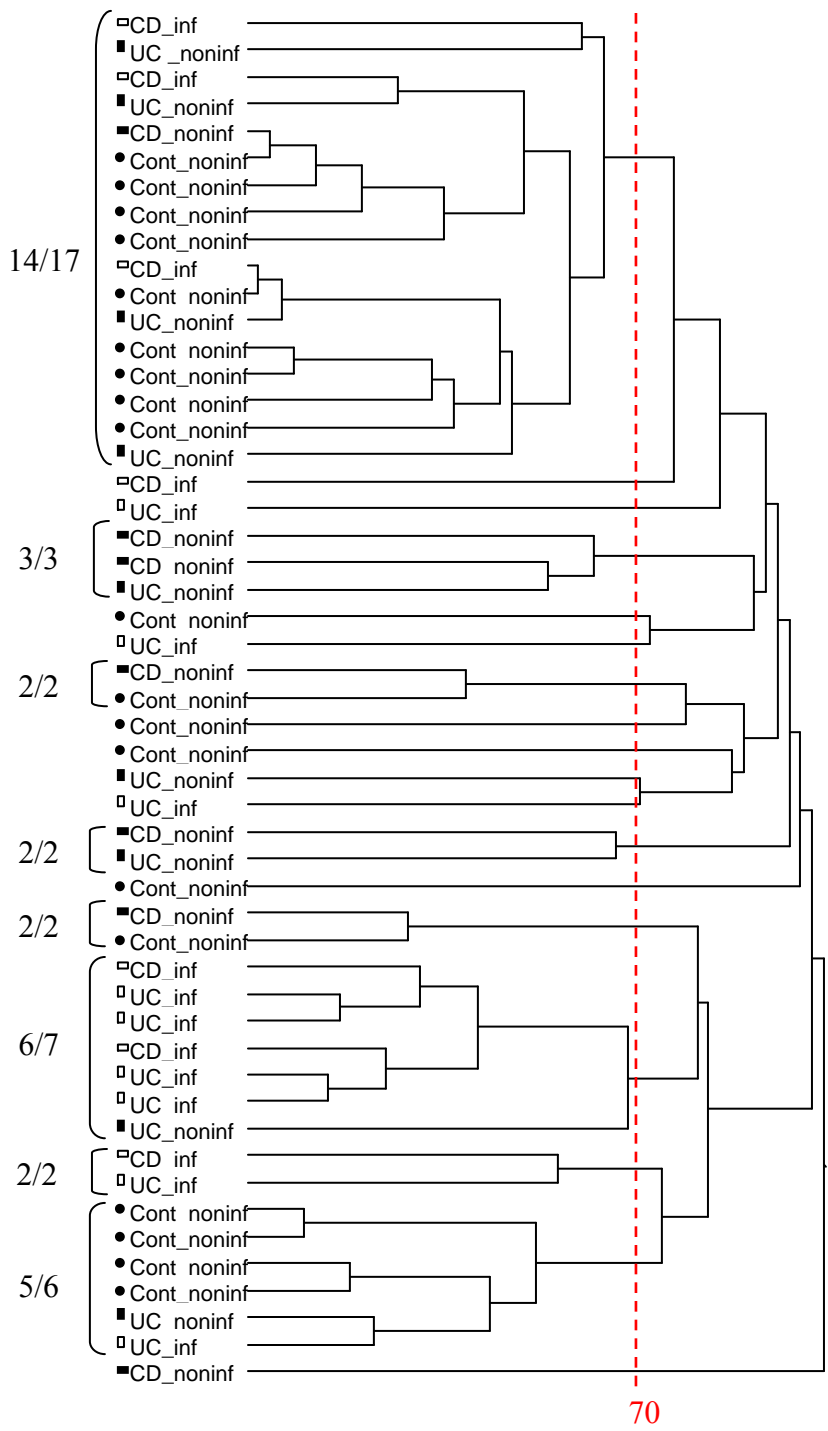
Figure 7. Phylogenetic tree based on T-RFLP incidence profiles. The values on the y-axis indicate the proportion of the total number of observations in the cluster that come from inflamed or non-inflamed tissue.

Species richness, diversity, and similarity indices were used to describe the microbial population in the biopsy tissue. When HC, CD, and UC tissues were compared using ARISA and T-RFLP output data, richness tended to be lower in the rectum than the cecum (Table 3). However, a much higher level of discrimination was obtained with the same calculations when designating tissues as inflamed or non-inflamed (Table 4). When diversity was calculated no significant difference were obtained with the Shannon index. Using the Simpson index diversity with ARISA data was control > CD > UC, but with T-RFLP data the diversity was CD > control > UC. Classifying tissues according to their inflammation state, the diversity was non-inflamed > control > inflamed, regardless of the algorithm used for calculating the indices (Table 4). The similarity indices were also calculated comparing normal tissues of controls and patients, as well as non-inflamed and inflamed tissues of the IBD patients (Table 5).

Table 3. Richness and diversity indices by disease state and biopsy site, calculated from ARISA and T-RFLP incidence data

Index†	Control			CD			UC		
	Cecum	Rectum	Total	Cecum	Rectum	Total	Cecum	Rectum	Total
				----ARISA----					
Richness									
Chao2	80.7^{bx}	64.1^a	86.3^b	83.8^{bx}	51.2^a	95.1^b	99.7^{by}	65.8^a	92.5^b
ICE	108.1	94.9	101	129.3^b	83.4^a	97.3^{ab}	131.7^b	80.2^a	104.9^{ab}
MMMean	103.1	75.2	101	110.2	82.5	92.9	106.8^b	63.8^a	98.4^b
Diversity									
Shannon	3.5	3.2	3.7	3.2	2.9	3.4	3.6	3.3	3.7
Simpson	73.3^b	55.9^{ax}	62.6^{ay}	79.8^b	73.8^{by}	58.3^{axy}	71.9^b	44.5^{ax}	53.1^{ax}
				----T-RFLP----					
Richness									
Chao2	48.3	38.6	48.0^{xy}	53.9^b	27.7^a	54.1^{by}	45.5	32.6	46.0^x
ICE	58.1	57.5	54.2	71.9^b	40.5^a	64.4^b	59.1	39.4	51.4
MMMean	50.0	39.2	49.5	62.9	29.4	55.5	48.2	26.7	43.6
Diversity									
Shannon	2.9	2.9	3.1	3.2	2.6	3.2	3.0	2.9	3.2
Simpson	31.2	39.1	31.1	51.4^b	30.4^a	37.8^a	34.8	30.0	28.8

† Values of indices are relative and do not represent absolute numbers of operational taxonomic units (OTU)

abc Values without a common superscript are different. Superscripts refer to comparisons of cecum vs. rectum vs. total within disease state (P < 0.05)

xyz Values without a common superscript are different. Superscripts refer to comparisons of the cecum, rectum, or total, between disease states (P < 0.05)

Table 4. Richness and diversity indices by disease and inflammation state, calculated from ARISA and T-RFLP incidence profiles

Index	Control Total	IBD Non-inflamed	Inflamed
	--------------------ARISA--------------------		
Richness			
Chao2	86.3^b	108.5^c	70.9^a
ICE	101^b	131.2^c	81.8^a
MMMeans	101^b	121.2^c	75.3^a
Diversity			
Shannon	3.7	3.7	3.5
Simpson	62.6^b	71.9^c	43.9^a
	--------------------T-RFLP--------------------		
Richness			
Chao2	48.1^b	53.4^c	40.9^a
ICE	54.2^a	63.4^b	48.2^a
MMMeans	49.5^{ab}	58.5^b	41.4^a
Diversity			
Shannon	3.1	3.2	3.1
Simpson	31.1^b	36.3^c	27.6^a

[abc] Values without a common superscript are statistically different using multiple comparison analysis. ($P < 0.05$)

Table 5. Multiple incidence-based similarity indices by disease and inflammation state, calculated from ARISA and T-RFLP profiles

Index	% of similarity ± SE	
	Normal (Cont) vs Non-inf. (IBD)	Inflamed (IBD) vs Non-inf. (IBD)
	------------------ARISA-----------------	
Bray-Curtis	71.1 ± 3.7	63.9 ± 3.8
Jaccard	77.0 ± 4.5	70.9 ± 5.1
	------------------T-RFLP----------------	
Bray-Curtis	74.8 ± 3.4	68.8 ± 4.3
Jaccard	79.9 ± 5.0	76.5 ± 5.8

T-RFLP fragments generated and matched to Genbank accession numbers produced peak identities that in many cases fell into multiple phyla (Table 6). To resolve this conflict we matched multiple accession numbers associated with single peaks at the phylum, class, order, family, and genus level based on Bergey's taxonomy. At the order level more than 95% of accession numbers were matched to a single taxonomic order (Table 6). Thus, we assigned fragments generated from T-RFLP only to the order level. Using this approach we observed less ($P < 0.05$) members of the *Bacteroidetes*, but more *Firmicutes* in the inflamed tissue (Table 7). The increase was particulary significant in the order unclassified *Clostridia*. Likewise, a decrease in the population of *Bacteroidetes* and an increase in the populatin of *Firmicutes* were detected in CD compared to UC (Table 7). Although these results indicated an increase in the population of *Proteobacteria* in inflamed tisusses, as well as CD and UC biopsies, compared to non-inflamed tissues and healthy controls (Table 7), the changes were not statistically significant.

Table 6. Distribution of T-RFs at different levels of taxonomic complexity

Terminal Fragment size (bp)[a]	Accession numbers[b]	Phylogeny Level[c]			
		Phylum	Class	Order	Family
37	36	36	30 (83.3)	30 (83.3)	26 (72.2)
38	5	5	4 (80)	4 (80)	4 (80)
66	4	4	4 (100)	4 (100)	2 (50)
67	26	26	24 (92.3)	24 (92.3)	20 (76.9)
68	9	9	9 (100)	9 (100)	6 (66.6)
71	4	4	4 (100)	4 (100)	4 (100)
72	3	3	3 (100)	3 (100)	3 (100)
93	2	2	1 (50)	1 (50)	1 (50)
94	3	3	1 (33.3)	1 (33.3)	1 (33.3)
96	5	5	3 (60)	3 (60)	3 (60)
98	3	3	3 (100)	3 (100)	2 (66.6)
99	2	2	2 (100)	2 (100)	2 (100)
100	2	2	2 (100)	2 (100)	2 (100)
102	46	46	44 (95.6)	44 (95.6)	43 (93.5)
103	11	11	11 (100)	11 (100)	11 (100)
104	6	6	6 (100)	6 (100)	6 (100)
179	3	3	3 (100)	3 (100)	3 (100)
180	6	6	6 (100)	6 (100)	4 (66.6)
189	9	9	9 (100)	9 (100)	5 (55.5)
190	9	9	9 (100)	9 (100)	5 (55.5)
191	10	10	10 (100)	10 (100)	10 (100)
192	23	23	22 (95.6)	22 (95.6)	18 (78.3)
193	4	4	4 (100)	4 (100)	4 (100)
205	3	3	3 (100)	3 (100)	3 (100)
206	4	4	4 (100)	4 (100)	4 (100)
211	2	2	1 (50)	1 (50)	1 (50)
231	4	4	4 (100)	4 (100)	4 (100)
233	3	3	3 (100)	3 (100)	3 (100)
262	3	3	3 (100)	3 (100)	3 (100)
338	2	2	2 (100)	2 (100)	2 (100)

[a] Fragments that only matched to one accession number (n=23) were excluded.
[b] Number of different accession numbers match to the same size fragment.
[c] Number and percent (in brackets) of the accession numbers fit into the same phylogenetic level. In case the accession numbers fall into more than one level, only the highest percentage has been reported.

Table 7. Comparison of putative microbial distribution generated from T-RF libraries according to disease condition and inflammation state

Microbial level	Taxonomic rank (%)[†]				
	Inflammation state			Disease state	
	Normal	Non-inflamed	Inflamed	CD	UC
Phylum *Bacteroidetes*	55.6^a	52.4^a	42.4^b	41.4^b	50.9^{ab}
Class *Bacteroidetes*	53.5^a	50.5^a	41.2^b	39.9^b	49.1^{ab}
Order *Bacteroidales*	53.5^a	50.5^a	41.2^b	39.9^b	49.1^{ab}
Class unclassified *Bacteroidetes*	2.0	1.9	1.2	1.5	1.9
Phylum *Firmicutes*	41.4^a	45.2^{ab}	53.3^b	54.5^b	43.9^{ab}
Class *Bacilli*	1.0	1.0	0.8	1.1	0.5
Order *Lactobacillales*	0.5	1.0	0.4	0.7	0.5
Order *Bacillales*	0.5	0.0	0.4	0.4	0.0
Class *Clostridia*	39.9^a	43.8^a	52.1^b	53^b	43^a
Order *Clostridiales*	39.4^a	43.3^{ab}	45.5^b	46.3^b	42.5^{ab}
Order unclassified *Clostridia*	0.5^a	0.5^a	6.6^b	6.7^b	0.5^a
Class unclassified *Firmicutes*	0.5	0.5	0.5	0.4	0.5
Phylum *Proteobacteria*	3.0	2.4	4.3	4.1	5.1

[†] Values are a proportion of the library for each taxonomic rank.
[ab] Statistical significance using LSD multiple comparison ($P < 0.05$). Normal tissues from controls were compared to inflamed and non-inflamed IBD tissues, as well as CD and UC biopsies.

2. Characterization of *E. coli* isolated from IBD gut biopsy tissues

2.1. Isolation and enumeration of *E. coli*

A total of 36 *E. coli* were isolated from IBD patients (CD = 14; UC = 14) and healthy controls (HC = 8). Although resuscitation of biopsies in buffered peptone water was performed, *E. coli* was only obtained from 46.7% of HC, 69.2% of patients with CD and 63.2% of patients with UC.

Serial dilution and plating on chromogenic agar allowed the quantification of total *E. coli* (Figure 8). The number of *E coli* was higher in CD (6.42×10^6/ml) and UC (5.32×10^5/ml) than controls (2.69×10^2/ml) ($p < 0.05$).
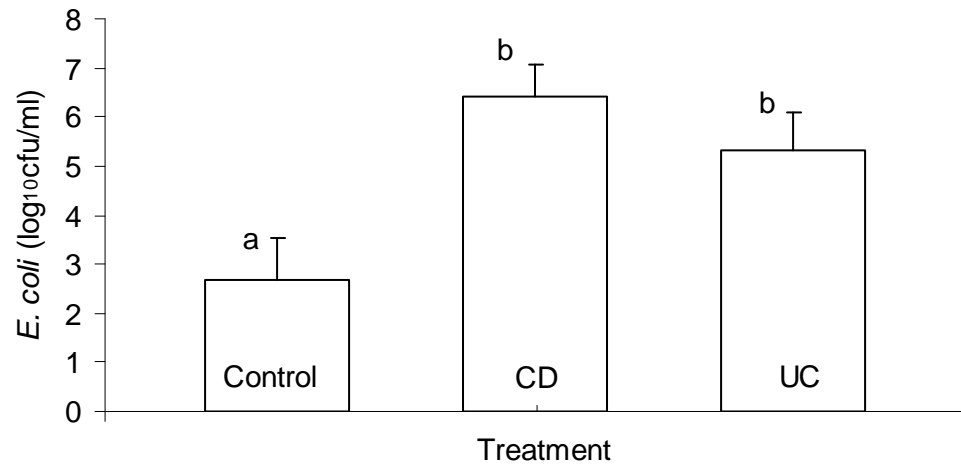
Figure 8. Number of *E. coli* isolated from gut biopsies of healthy controls and patients with CD and UC.

2.2. Biochemical profiling of *E. coli*

Phenotypic characteristics of 36 *E. coli* isolates from IBD patients (CD = 14 , UC = 14) and healthy controls (HC = 8), as well as two strains of *E. coli* isolated from cattle feces, two strains from swine feces, the probiotic *E. coli* Nissle 1917, the non-pathogenic *E. coli* K12, and pathogenic EHEC O157:H7, UPEC CFT073, and APEC O1:K1:H7 were studied using API 20E system.

Identification of all *E .coli* strains were confirmed using API 20E strips (Table 8). All *E. coli* isolates fermented glucose (GLU) and mannitol (MAN) but not inositol (INO). All strains were positive for ortho-nitrophenyl-ß-D-galactopyranosidase (ONPG), and negative for urease (URE), tryptophane deaminase (TDA), and gelatinase (GEL). No isolates produced H_2S or acetone (VP). Differences were detected between isolates in arginine dehydration (ADH), lysine decarboxylation (LDC), ornithine decarboxylation (ODC), citrate utilization (CIT), indole production (IND), and fermentation of sorbitol (SOR), rhamnose (RHA), sucrose (SAC), melibiose (MEL), amygdalin (AMY), and arabinose (ARA). No significant association was found between any of biochemical characteristics of *E. coli* and the disease state.

Table 8. API 20E biochemical profiles of *E. coli* isolates

Origin	Sample No.	ABD type*	Amino acid hydration/decarboxylation			Citrate utilization	Indole prod.	Carbohydrate fermentation/oxidation					
			ADH	LDC	ODC	CIT	IND	SOR	RHA	SAC	MEL	AMY	ARA
HC	17B	A	-	-	-	-	+	+	+	-	+	-	+
HC	73A	A	-	+	+	-	+	+	+	+	+	-	+
HC	88	A	-	+	-	-	+	+	+	-	+	-	+
HC	90A	A	-	+	+	-	+	+	+	+	+	-	+
HC	90B	A	-	+	+	-	+	+	+	+	+	-	+
HC	43A	B2	-	+	-	-	+	+	+	-	+	-	+
HC	48B	B2	-	+	+	-	+	+	+	+	-	-	+
HC	50A	B2	-	+	+	-	+	+	+	-	+	-	+
CD	91A	B1	-	+	+	-	+	+	+	+	+	-	+
CD	120A	B1	-	+	+	-	-	+	+	+	+	-	+
CD	120C	B1	-	+	+	-	-	+	+	+	+	-	+
CD	15B	B2	-	-	+	-	+	-	+	+	+	-	-
CD	79	B2	-	+	+	-	+	+	+	-	-	-	+
CD	126C	B2	+	-	-	+	+	+	+	+	+	+	+
CD	132A	B2	-	+	-	-	+	-	+	-	+	-	+
CD	132B	B2	-	+	-	-	+	-	+	-	+	-	+
CD	143A	B2	-	+	+	-	+	+	+	+	+	-	+
CD	143B	B2	-	+	+	-	+	+	+	+	+	-	+
CD	146A	B2	-	+	+	-	+	+	+	-	+	-	+
CD	146B	B2	-	+	+	-	+	+	+	-	+	-	+
CD	149A	B2	-	+	+	-	+	+	+	-	+	-	+
CD	149B	B2	-	+	+	-	+	+	-	-	+	-	+
UC	121	B1	-	+	+	-	+	-	+	+	+	-	+
UC	135B	B1	-	+	+	-	+	+	+	+	+	-	+
UC	128A	B2	-	+	+	-	+	+	+	+	-	-	+
UC	128B	B2	-	+	+	-	+	+	+	+	-	-	+
UC	131A	B2	-	+	+	-	+	+	+	-	-	-	+
UC	136	B2	-	+	+	-	+	+	+	-	+	-	+
UC	141B	B2	-	+	+	-	+	+	+	-	+	-	+
UC	142B	B2	-	+	+	-	+	+	+	+	-	-	+
UC	145B	B2	-	+	+	-	+	+	+	+	+	-	+
UC	147A	B2	-	+	+	-	+	+	+	-	+	-	+
UC	117A	D	-	+	-	-	+	+	+	+	+	-	+
UC	127A	D	-	+	-	-	+	+	-	+	-	-	-
UC	139B	D	-	+	+	-	+	+	+	+	+	-	+
UC	139C	D	-	+	+	-	+	+	+	+	+	-	+
Cattle	C24	A	-	+	-	-	+	+	+	-	-	-	+
Cattle	C64	B1	-	+	+	-	+	+	+	+	+	-	+
Swine	P6	B1	-	+	+	-	+	+	-	+	+	-	-
Swine	P10	A	-	+	+	-	+	+	+	+	+	-	+
Nissle	Nissle	B2	+	+	+	-	+	+	+	-	+	-	+
K12	K12	A	-	+	-	-	+	+	+	-	+	-	+
EHEC	EHEC	B2	-	+	+	-	+	-	+	+	+	-	+

*ABD is a system for typing *E. coli* isolates based on their phylogenetic relationships. A and B1 are less pathogenic and B2 and D are more phatogenic *E. coli* phylotypes

ADH, arginine dihydrolase; LDC, lysine decarboxylase; ODC, ornithine decarboxylase; CIT, citrate; IND, indole; SOR, sorbitol; RHA, rhamnose; SAC, sucrose; MEL, melibiose; AMY, amygdalin; ARA, arabinose; CD, Crohn's disease; UC, ulcerative colitis. All the *E.coli* strains were positive for ortho Nitrophenyl-ß-D-galactopyranosidase (ONPG), glucose (GLU) fermentation, manitol (MAN) fermentation, and negative for urease (URE), tryptophane deaminase (TDA), gelatinase (GEL), and inositol (INO) fermentation. None of the *E. coli* isolates produced H_2S or acetone (VP).

2.3. Phylogenetic analysis of *E. coli* using MLST technique

The allelic profiles of all isolates are shown in the Table 9. The *fumC* and *gyrB* exhibited the greatest allelic diversity (17 alleles) and *recA* the least (12 alleles) (Figure 9). Of 100 alleles obtained from this study, 9 were newly identified and added to the MLST.net database. All of the new alleles were associated with IBD. A total of 26 sequence types (ST) were identified and 11 of them were new STs (Table 9).

Table 10 shows all the STs, the number of isolates related with each ST (FREQ), the number of STs in the group that differ at a single locus (SLVs), double loci (DLVs), and those that are more distantly related (SATs). Isolates were placed into four different groups with only one isolate designated as singleton. All *E. coli* strains fell into separate clonal groups based on their phylotype except one strain (strain 128B). Group 1 primarily encompassed A and B1 phylotypes while groups 2 and 4 only included B2. Group 3 was only made up of phylotype D. Chi-squared analysis indicated that disease state (IBD) was significantly related with clonal groups ($P < 0.05$).

Concatenated sequences (total of 3,423 bp) of housekeeping genes were used to construct a radiation tree which consisted of three major clusters, one B2 cluster, one D cluster, and one hybrid cluster of AB (Figure 10).

Discriminant analysis using STs as the independent variable and disease state as the dependent variable associated the new STs to IBD (Figure 11).

When amino acid sequences were compared (Table 11), no substitutions were detected for *icd*, *purA*, and *recA* genes, and only a few substitutions were found for the rest of the genes. Detected amino acid replacements were associated to phylotypes (Table 11) but no association was found to be significant with IBD.

Table 9. Detailed MLST information including allelic data of seven housekeeping genes and ST and ST complex information of 45 *E. coli* strains

Sample No.	Origin	MLST allelic numbers							ST	ST complex	ABD type
		adk	*fumC*	*gyrB*	*icd*	*mdh*	*purA*	*recA*			
17B	HC	10	11	57	8	7	18	6	216	None	A
90B	HC	6	4	12	1	20	12	7	88	23	A
90A	HC	6	4	12	1	20	12	7	88	23	A
48B	HC	13	40	13	13	23	25	66	357	None	B2
88	HC	10	11	4	8	8	8	2	10	10	A
50A	HC	36	24	9	13	17	11	25	73	73	B2
73A	HC	6	107	1	95	69	8	7	**635**	399	A
43A	HC	37	38	19	37	17	11	26	95	95	B2
128A	UC	13	108	10	97	18	68	*93*	**636**	None	B2
147A	UC	37	38	19	37	17	*100*	26	**647**	95	B2
128B	UC	13	108	10	97	18	68	*93*	**636**	None	B2
131A	UC	76	24	9	13	17	11	25	**638**	73	B2
139C	UC	62	100	17	31	5	5	4	362	None	D
127A	UC	21	35	27	6	5	5	4	69	69	D
117A	UC	21	35	27	6	5	5	4	69	69	D
139B	UC	62	100	17	31	5	5	4	362	None	D
136	UC	37	38	19	37	17	11	26	95	95	B2
141B	UC	37	38	19	37	17	11	26	95	95	B2
135B	UC	43	41	15	18	11	7	6	101	101	B1
145B	UC	13	*147*	93	13	17	28	30	**640**	None	B2
121	UC	9	6	33	*131*	24	8	7	**641**	86	B1
142B	UC	14	14	10	14	17	92	10	550	14	B2
132B	CD	10	11	4	8	8	8	2	10	10	B2
91A	CD	43	41	15	18	11	7	6	101	101	B1
120A	CD	9	23	33	18	11	8	6	**642**	278	B1
79	CD	36	24	9	13	17	11	25	73	73	B2
146A	CD	37	38	19	37	*95*	11	26	**643**	95	B2
15B	CD	13	40	19	13	36	11	30	**644**	538	B2
132A	CD	10	11	4	8	8	8	2	10	10	B2
126C	CD	*117*	*148*	*120*	12	80	1	2	**645**	None	B2
146B	CD	37	38	19	37	*95*	11	26	**643**	95	B2
143B	CD	53	40	47	13	36	28	29	131	None	B2
149A	CD	13	96	*121*	13	17	11	10	**646**	None	B2
120C	CD	9	23	33	18	11	8	6	**642**	278	B1
143A	CD	53	40	47	13	36	28	29	131	None	B2
149B	CD	13	96	*121*	13	17	11	10	**646**	None	B2
Nissle	Probiotic	36	24	9	13	17	11	25	73	73	B2
C24	Cattle	10	11	4	8	8	8	2	10	10	A
C64	Cattle	6	4	4	16	24	8	14	58	155	B1
P6	Swine	6	4	4	18	24	8	14	223	155	B1
P10	Swine	6	4	12	1	20	8	7	90	23	A
K12	K12	10	11	4	8	8	8	2	10	10	A
EHEC	EHEC	12	12	8	12	15	2	2	11	11	B2
UPEC	Database	36	24	9	13	17	11	25	73	73	B2
APEC	Database	37	38	19	37	17	11	26	95	95	B2

Allelic numbers in ***bold and italic*** indicate new alleles identified in this study.
STs (sequence type) in **bold** belong to the new sequence types observed in this study.

Allelic diversity of seven housekeeping genes
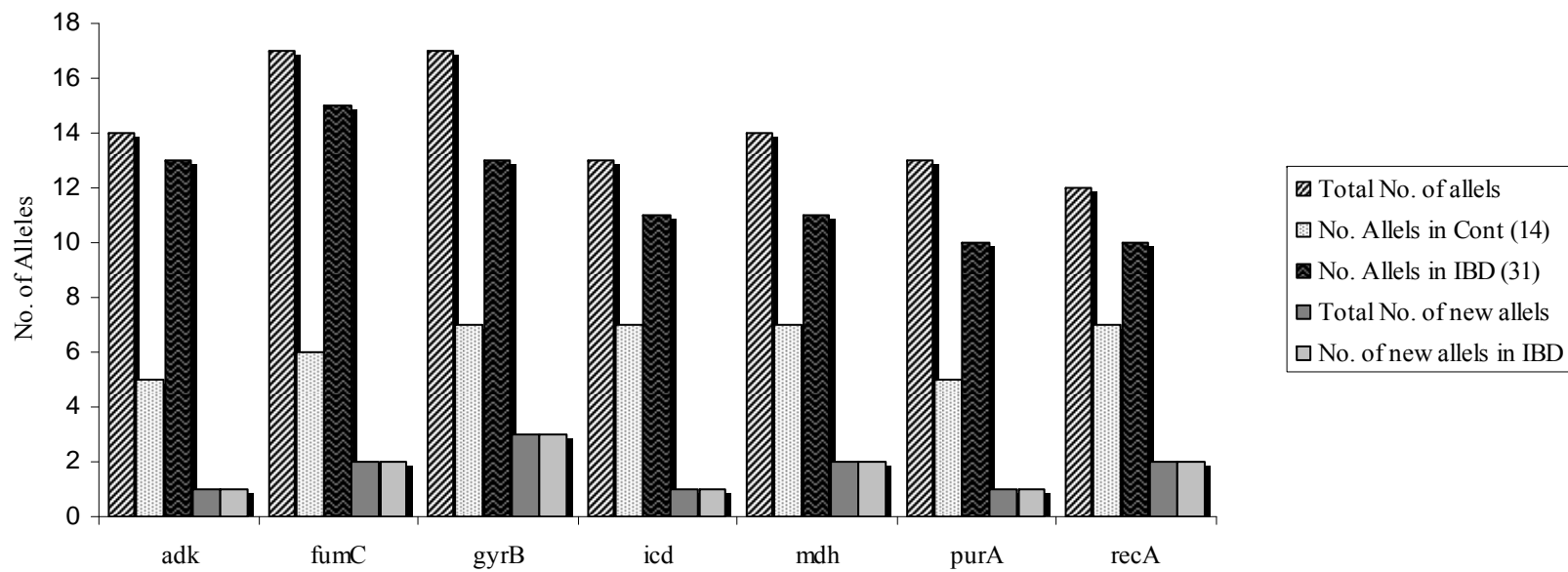
Figure 9. Allelic diversity of seven housekeeping genes in 45 *E. coli* strains under study.

Table 10. MLST allelic profile evaluation with eBURST analysis using a group definition of two

ST	FREQ	SLV	DLV	SAT	Source	Type
Group 1						
216	1	0	0	9	HC	A
88	2	1	0	8	HC	A
10	5	0	0	9	HC, CD, Cattle, K12	A, B2
635	1	0	0	9	HC	A
101	2	0	0	9	CD, UC	B1
641	1	0	0	9	UC	B1
642	2	0	0	9	CD	B1
58	1	1	0	8	Cattle	B1
223	1	1	0	8	Swine	B1
90	1	1	0	8	Swine	A
Group 2						
357	1	0	0	10	HC	B2
73	4	1	0	9	HC, CD, Nissle, UPEC CFT073	B2
131	2	0	0	10	CD	B2
638	1	1	0	9	UC	B2
95	4	2	0	8	HC, UC, APEC O1:K1:H7	B2
640	1	0	0	10	UC	B2
550	1	0	0	10	UC	B2
643	2	1	1	8	CD	B2
644	1	0	0	10	CD	B2
646	2	0	0	10	CD	B2
647	1	1	1	8	UC	B2
Group 3						
362	2	0	0	1	UC	D
69	2	0	0	1	UC	D
Group 4						
645	1	0	0	1	CD	B2
11	1	0	0	1	EHEC O157:H7	B2
Singleton						
636	2				UC	B2

ST, sequence type; FREQ, frequency; SLV, single locus variants; DLV, double locus variants; SAT, satellites.

eBURST analysis suggested a significant relationship between *E. coli* isolated from IBD patients and clonal groups ($p = 0.02$)
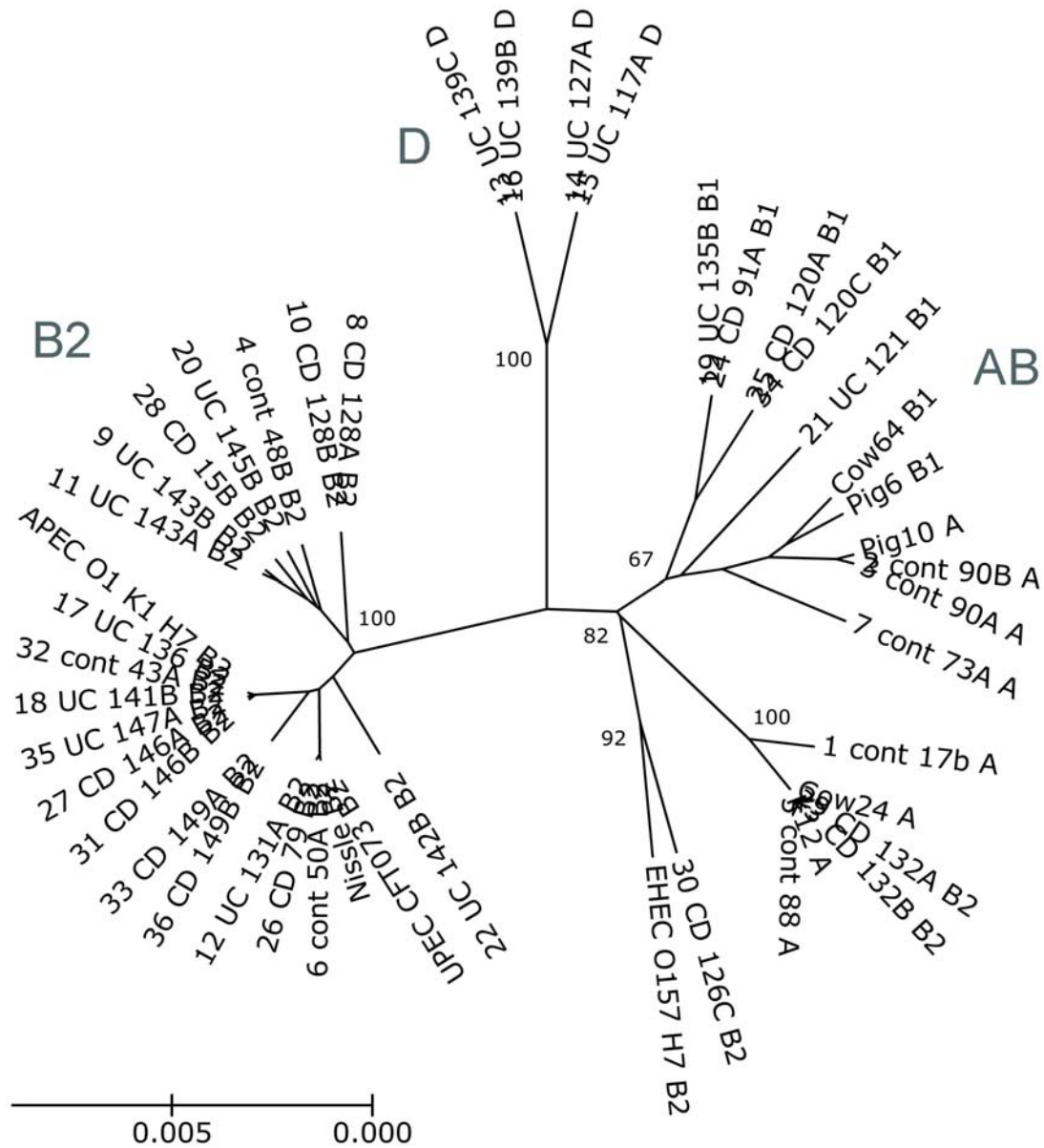
Figure 10. Radiation tree of MLST sequences of *E. coli* evaluated in this study illustrating three major clusters containing B2, D, and hybrid AB phylotypes. Sequences of reference strains (UPEC CFT073 and APEC O1:K1:H7) were obtained from MLST.net and included.
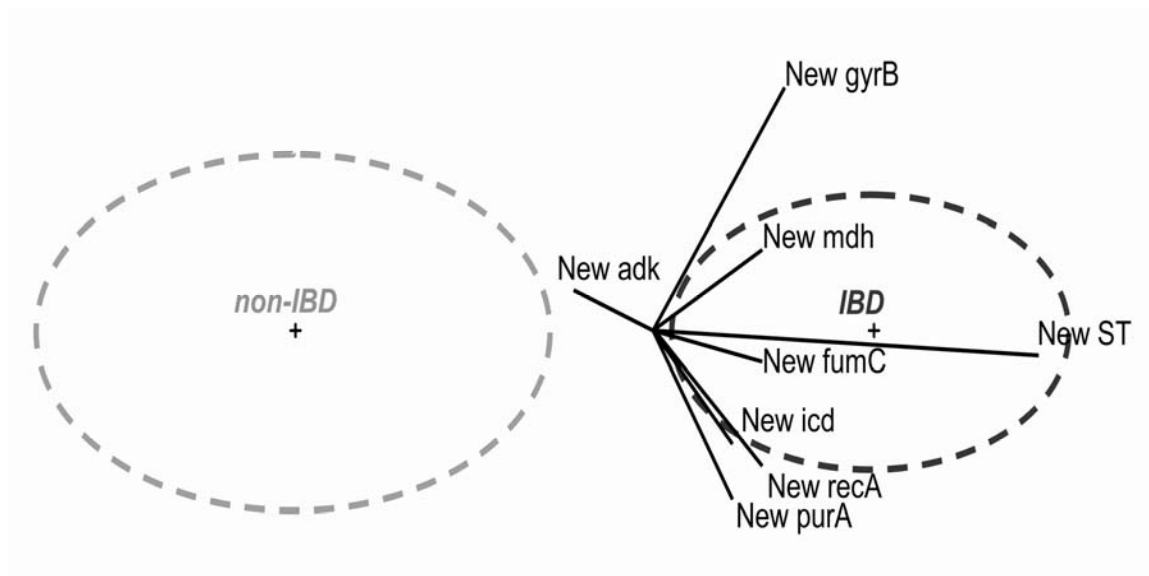
Figure 11. Discriminant analysis of new MLST alleles and STs of 45 *E. coli* strains under this study.

Table 11. Relationship between amino acid substitution in housekeeping genes and disease state or *E. coli* phylotype

Gene	Amino acid substitution	Disease state	P–value	*E. coli* phylotype	P–value
adk	A44V	IBD	0.33	B1	$0.0008^{\dagger}$
	R67H	IBD	0.23	B1	$0.00003^{\dagger}$
	N68Y	IBD	0.30	B2	0.01^{*}
fumC	S102N	HC	0.04^{*}	A	$0.0008^{\dagger}$
gyrB	E72D	IBD	0.26	B2	$0.00006^{\dagger}$
mdh	A29S	IBD	0.16	B1	$0.000001^{\dagger}$

IBD, inflammatory bowel disease; HC, healthy controls
A, Alanine; V, Valine; R, Arginine; H, Histidine; N, Asparagine; Y, Tyrosine; S, Serine; E, Glutamic acid; D, aspartic acid
No amino acid substitution was detected for *icd*, *purA*, and *recA* genes.
* $p < 0.05$, † $p < 0.01$

2.4. Virulence characterization of the *E. coli*

A total of 65 *E. coli* isolates were obtained from 59 subjects. Of the 23 CD patients 14 had *E. coli*, of the 11 UC patients 6 had *E. coli*, and 11 of the 25 healthy controls contained *E. coli* (Table 12). Although more *E. coli* were isolated from IBD tissue compared to HC, the relationship was not statistically significant (P = 0.20). Also, there appeared to be no significant association between the presence of *E. coli* and inflammation in IBD tissues (P = 0.30).

When *E. coli* strains were typed for their O and H antigens, no dominant serotype was observed. *E. coli* isolates belonged to 27 different serotypes (Table 12) and no significant association was detected between any of O and H antigens and the disease. We also identified the isolated other than *E. coli* (that grow on chromogenic agar plates) with 16S rDNA and cpn60 sequencing. There was a predominance of streptococci and staphylococci in all tissues but particularly in the HC (Table 12).

Table 12. Characterization of *E. coli* and other bacteria isolated from HC and newly diagnosed CD, and UC biopsy tissues

Disease	Infl.	Tissue	Phylotype	Serotype	Virulence factors	Other strains[*]
HC	No	20				*Corynebacterium coyleae*
HC	No	41	B2	O4:H5	SPATE, Ag43, PAI I, Pap	*Streptococcus anginosus*
HC	No	43				*Staphylococcus epidermidis*
HC	No	47				*Staphylococcus muscae*
HC	No	108	B2, D	O77:H18	SPATE, Ag43, AIDA-I	*Streptococcus salivarius*
HC	No	109	B2	O77:H18	Ag43, AIDA-I	*Streptococcus salivarius, Staphylococcus aureus*
HC	No	111	B2	O2:H18	SPATE, Ag43, AIDA-I, Sfa	*Streptococcus parasanguinis*
HC	No	118				*Streptococcus parasanguinis*
HC	No	127	B2	O25:H1	SPATE, Ag43, AIDA-I, Sfa	
HC	No	128	B2	O25:H4	SPATE, Ag43	
HC	No	130				*Micrococcus sp.*
HC	No	138				*Lactobacillus casei, Streptococcus parasanguinis, Staphylococcus aureus*
HC	No	142	B2, D	O166:H15, O1:H34	Ag43, AIDA	
HC	No	143	A, B1	O2:NM, O88:H8	SPATE, Ag43, AIDA	*Klebsiella ornithinolytica*
HC	No	145	A	O8:NM	AIDA	
HC	No	146				*Streptococcus salivarius*
HC	No	147				*Haemophilus sp.*
HC	No	174				*Dermabacter hominis*
HC	No	197				*Paenibacillus lactis*
HC	No	198	A		AIDA-I	
HC	No	209	B1	O153:H8	AIDA-I, PAI I, Pap	
CD	No	21	A		Ag43, AIDA-I	
CD	No	22	A, B1	O10:H12, O82:H8	AIDA-I	*Streptococcus sanguinis*
CD	Yes	24				*Klebsiella sp.*
CD	No	31	A			*Staphylococcus aureus, Enterobacter hormaechei, Bacillus sp.*
CD	Yes	44	D	O166:H15	Ag43, AIDA-I	*Proteus mirabilis*
CD	No	45	B2	O175:H5	SPATE	*Staphylococcus aureus*
CD	No	112				*Streptococcus salivarius*

Table 12. Characterization of *E. coli* and other bacteria isolated from HC and newly diagnosed CD, and UC biopsy tissues

Disease	Infl.	Tissue	Phylotype	Serotype	Virulence factors	Other strains[*]
CD	Yes	134	A, B2	O152:H23, O77:H18	AIDA-I, PAI I, Pap	
CD	Yes	136				*Staphylococcus captis*
CD	Yes	155	B2	O6:H31	SPATE, Ag43, PAI I, Pap, Sfa	
CD	Yes	162	A	O141:H30	AIDA-I	*Streptococcus salivarius, Paenibacillus lautus*
CD	No	165	B1	O58:H25	Ag43, AIDA-I , AfaE, BmaE	
CD	No	166	B2	O123:H6	SPATE, Ag43, Pap, AfaE	
CD	Yes	167				*Staphylococcus epidermidis, Enterococcus faecalis*
CD	No	168	B2	O1:H7	SPATE, Ag43, PAI I, Pap	
CD	No	171	B1	O4:H11	Ag43, AIDA-I, PAI I, Pap	*Enterobacter sp., Klebsiella pneumoniae*
CD	Yes	172				*Staphylococcus aureus*
CD	No	173	B2, D	O1:H7	SPATE, Ag43, AIDA-I, PAI I, Pap, AfaE	*Lactococcus lactis*
CD	Yes	175				*Staphylococcus aureus, Candida albicans*
CD	Yes	199				*Streptococcus sp., Klebsiella oxytoca*
CD	Yes	206	B2	O25:H4	SPATE, Ag43, AIDA-I	
UC	No	150	B2	O73:H41	AIDA-I	
UC	No	164	A	O9:NM	AIDA-I	*Lactococcus lactis*
UC	No	170	B2	O75:NM	SPATE, Ag43, AfaE	*Klebsiella pneumoniae*
UC	No	177				*Corynebacterium aurimucosum, Streptococcus sanguinis*
UC	No	201	B1	O88:H8	AIDA-I	
UC	Yes	202	B2, D	O18ac:H1, O163:H9	SPATE, Ag43, AIDA-I, PAI I, Pap	*Staphylococcus sp.*
UC	No	204	B2	O126:H20	Ag43, AIDA-I	
UC	No	205				*Enterococcus faecium*
UC	No	214				*Paenibacillus macerans, Bacillus sp.*

SPATE, serine protease autotransporter; Ag43, antigen involved in *E. coli* autoaggregation; AIDA-I, adhesin involved in diffuse adherence; PAI I, pathogenicity island I; PAP, *E. coli* Pap pili subunit and a part of pathogenicity island II; Sfa, minor subunit of S-fimbriae; AfaE, afimbrial adhesin involved in *E. coli* diffuse adherence; BmaE, M-agglutinin subunit.
[*]These isolates were identified by sequencing the 16S ribosomal DNA and chaperonin 60 genes.

Typing *E. coli* with the ABD system indicated that B2 was the most dominant and made up 60% of *E. coli* isolates from UC tissues. Together B2 and D made up the majority of isolates (Table 13). The virulence genes assessed were widely distributed and ranged from 0% for *eae*, *cnf*, *hlyA*, *vt1*, and *vt2* to 100% for *fimH* (Table 13). In addition to *fimH*, the most predominant virulence factor for HC was *aida* (72%), for CD were *agn43* (63%) and *aida* (67%), and for UC was *aida* (60%).

When the statistical relationships between virulence genes, ABD typing, and the disease state were assessed (Table 14), a highly significant relationship was found ($P = 0.01$) between *afaE* and IBD, and this strong relationship was influenced primarily by the strong association between CD and *afaE* ($P = 0.005$). There were significant statistical associations between B2 and *sfa* ($P = 0.04$) and B2 and *agn43* ($P = 0.05$) but not with phylotype D. On the other hand there was a strong association between B2+D and inflammation ($P = 0.04$) and B2+D and SPATE ($P = 0.00001$).

Table 13. Distribution of the phylotypes and virulence factors of *E. coli* isolated from HC and first diagnosed CD and UC patients.

Item	HC (25)	CD (30)	UC (10)
Phylotype			
A	5(20%)	7(23%)	2(20%)
B1	5(20%)	7(23%)	1(10%)
B2	11(44%)	12(40%)	6(60%)
D	4(16%)	4(14%)	1(10%)
Virulence factors			
SPATE	9(36%)	13(43%)	4(40%)
Ag43	10(40%)	19(63%)	4(40%)
AIDA	18(72%)	20(67%)	6(60%)
PAI I	5(20%)	10(33%)	1(10%)
Pap	6(24%)	12(40%)	1(10%)
Sfa	2(8%)	1(3%)	0
AfaE	0	8(27%)	1(10%)
BmaE	0	3(10%)	0

Values in parentheses represent the frequency of items in patients as a percentage.

SPATE, serine protease autotransporter; Ag43, antigen involved in *E. coli* autoaggregation; AIDA-I, adhesin involved in diffuse adherence ; PAI I, pathogenicity island I; Pap, *E. coli* Pap pili subunit and a part of pathogenicity island II; Sfa, minor subunit of S-fimbriae; AfaE, afimbrial adhesin involved in *E. coli* diffuse adherence; ; BmaE, M-agglutinin subunit.

All *E. coli* isolated were negative for *cnf* (cytotoxic necrotizing factors Cnf1 and Cnf2), *vt1* and *vt2* (verocytotoxins 1 and 2), *hlyA* (α hemolysin), and *eae* (intimin) toxins.

Table 14. Statistical relationship between different characteristics of *E. coli* isolated from HC and first diagnosed CD and UC patients.

Positive correlation		
Disease	Factor	P–value
IBD	B2+D	0.84
IBD	AfaE	0.01*
CD	AfaE	0.005[†]
CD	BmaE	0.15
Phylotype	Factor	P–value
B2+D	Inflammation (Clin or Hist)	0.04*
B2+D	SPATE	0.00001[†]
B2	Ag43	0.05*
B2	Sfa	0.04*
Negative correlation		
Phylotype	Factor	P–value
A	PAI I	0.01*
A	PAP	0.007[†]
B2+D	AIDA-I	0.0004[†]

IBD, inflammatory bowel disease; SPATE, serine protease autotransporter; Ag43, antigen involved in *E. coli* autoaggregation; AIDA-I, adhesin involved in diffuse adherence ; PAI I, pathogenicity island I; Pap, *E. coli* Pap pili subunit and a part of pathogenicity island II; Sfa, minor subunit of S-fimbriae; AfaE, afimbrial adhesin involved in *E. coli* diffuse adherence; BmaE, M-agglutinin subunit.
* $p < 0.05$, [†] $p < 0.01$

2.5. Analysis of *fimH* subunit of *E. coli* fimbria gene

Genes encoding FimH, a minor subunit of *E. coli* pilus located at the tip of type I fimbriae, were sequenced to examine their potential role in the pathogenesis of *E. coli* in patients with established IBD as well as patients who were just diagnosed with IBD.

When relationship of *E. coli* isolated from patients with established IBD were examined using radiation tree, *fimH* sequences were not exclusively clustered based on the IBD state or *E. coli* phylotype (Figure 12). However, one cluster contained 12 *E. coli* strains which were all from IBD patients and belonged to B2, including AIEC LF82 strain, UPEC CFT073 and APEC O1:K1:H7. *E. coli* isolated from cattle and swine, non-pathogenic *E. coli* K12, probiotic Nissle 1917, and EHEC O157:H7 were clustered separately based on their *fimH* sequences (Figure 12). When relationship was evaluated at amino acid level (Table 15), amino acid substitutions of alanine for valine at 48 and 140 residues of FimH were associated to HC ($P < 0.05$) and amino acid substitutions of serine for asparagine at 91 and asparagine for serine at 99 FimH residues were significantly associated to IBD ($p < 0.05$).

Phylogenetic analysis based on *fimH* of *E. coli* from newly diagnosed patients (Figure 13), followed the same trend as the analysis of patients with established disease, and a separate cluster consisting of B1 and B2 isolates, including the adherent invasive strain LF82, uro-pathogenic *E. coli* CFT073 and the avian-pathogenic *E. coli* O1:K1:H7 was formed.   The amino acid substitutions N91S ($P = 0.004$), S99N ($P = 0.004$), and A223V ($P = 0.05$) in FimH were significantly associated with IBD (Table 16).   Two of these amino acid substitutions, N91S and S99N, were also highly associated with the B2 phylotype (Table 16).
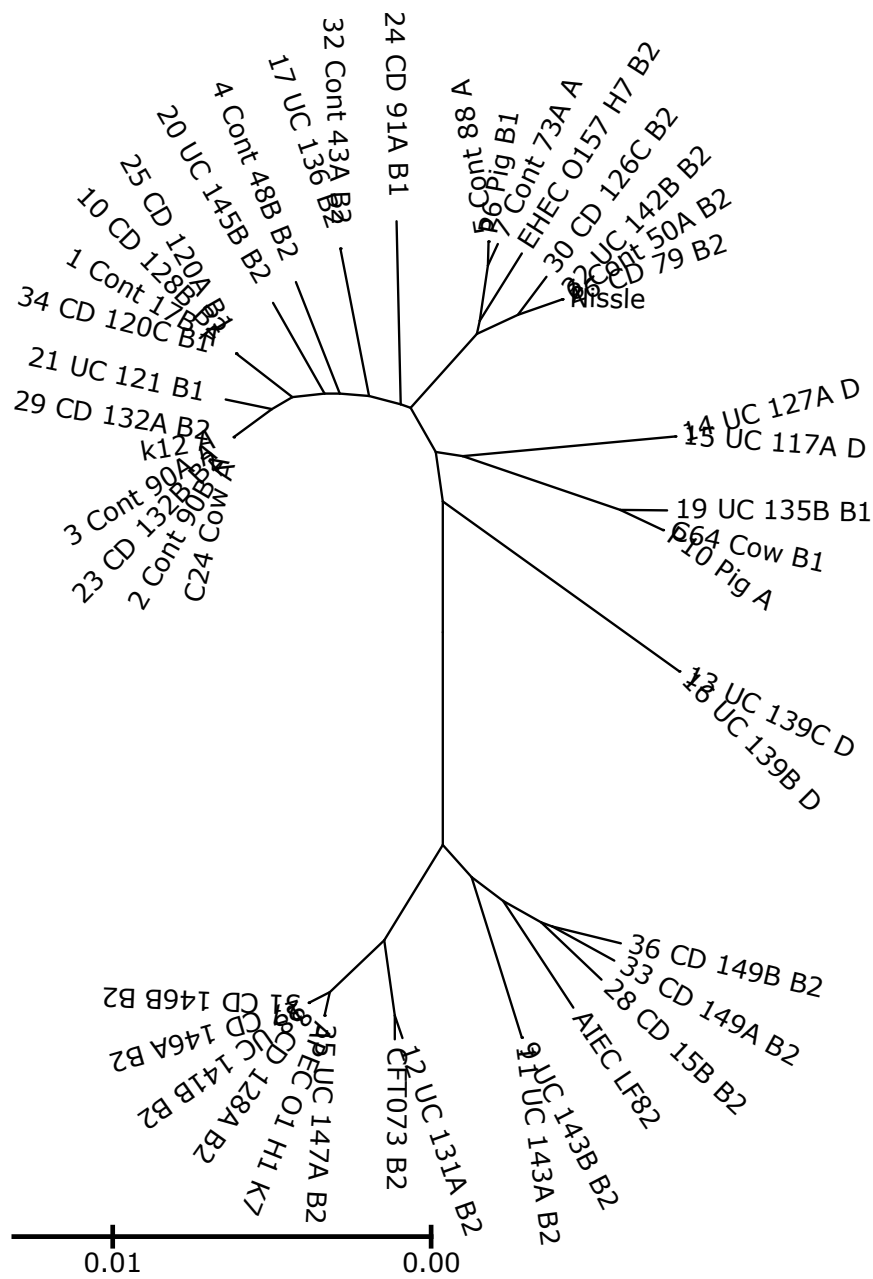
Figure 12. Radiation tree of *fimH* sequences of *E. coli* isolated from healthy controls (Cont = 8) and patients with established IBD (CD = 14; UC = 14). *fimH* sequences of *E. coli* isolated from cattle (n = 2), swine (n = 2), probiotic Nissle 1917, and reference strains K12 MG1655, AIEC (adherent invasive *E. coli*) LF82, EHEC (enterohaemorrhagic *E. coli*) O157:H7, APEC (avian pathogenic *E. coli*) O1:K1:H7, UPEC (uropathogenic *E. coli*) CFT073 were also added to the study as controls.

Table 15. Relationship between amino acid substitution in *fimH* subunit of fimbria gene of *E. coli* isolated from patients with established IBD and disease state, or *E. coli* phylotype

Amino acid substitution	Disease state	P–value	Phylotype	P–value
A48V	HC	0.01*	A	0.00005[†]
S83A	IBD	0.15	B2	0.07
N91S	IBD	0.03*	B2	0.007*
S99N	IBD	0.006[†]	B2	0.0004[†]
A140V	HC	0.03*	A+B1	0.04*
R187H	IBD	0.76	B2	0.07
Y216F	IBD	0.32	D	0.000006[†]

IBD, inflammatory bowel disease; HC, healthy controls
A, Alanine; V, Valine; S, Serine; N, Asparagine; R, Arginine; H, Histidine; Y, Tyrosine; F, Phenylalanine
For *fimH* analysis, the sequence of AIEC LF82 fimbrial gene was also added to the study.
* $p < 0.05$, [†] $p < 0.01$
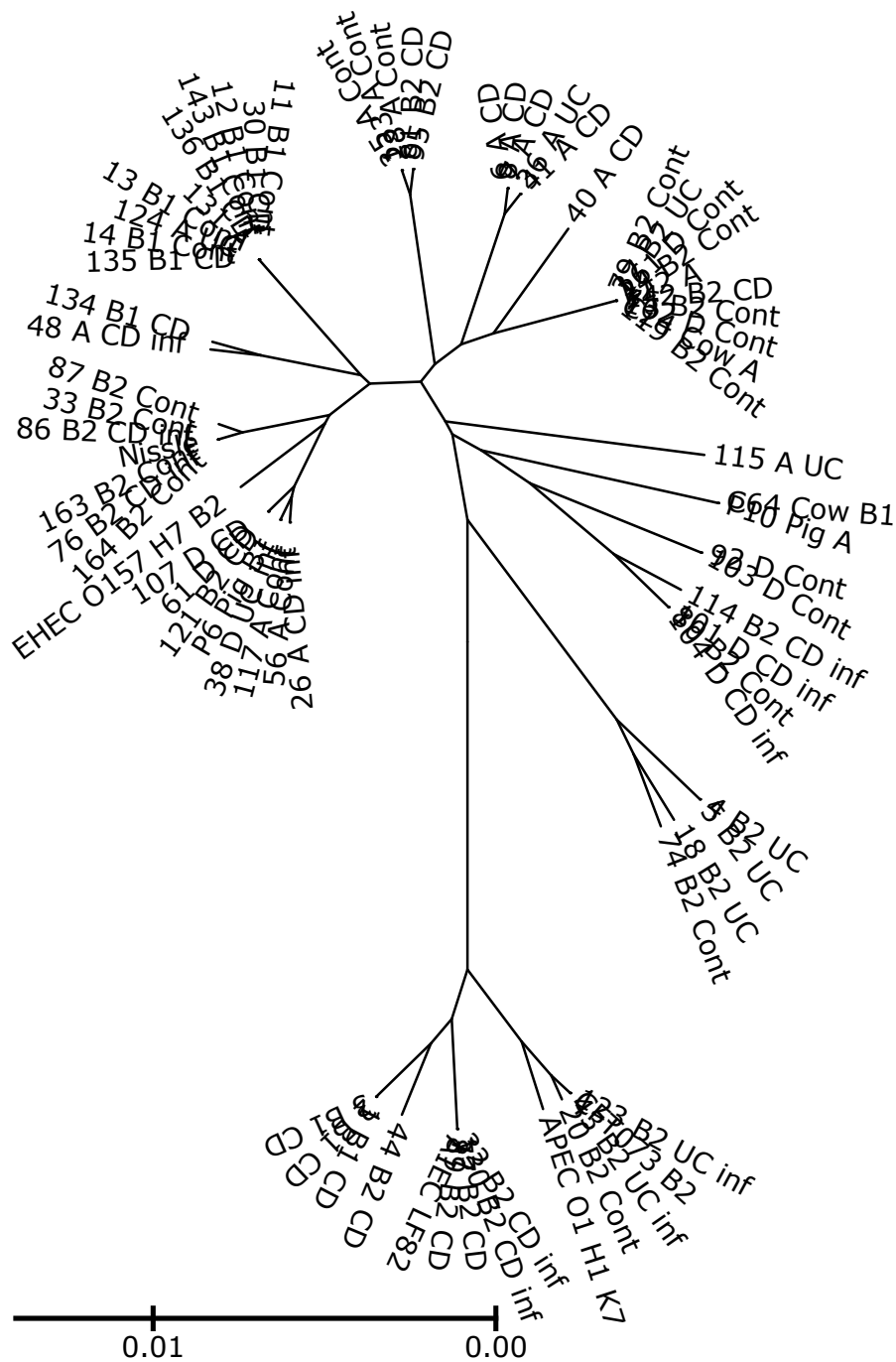
Figure 13. Radiation tree of *fimH* sequences of *E. coli* isolated from healthy controls (Cont = 25) and newly diagnosed IBD patients (CD = 30; UC = 10). *fimH* sequences of *E. coli* isolated from cattle (n = 2), swine (n = 2), probiotic Nissle 1917, and reference strains K12 MG1655, AIEC (adherent invasive *E. coli*) LF82, EHEC (enterohaemorrhagic *E. coli*) O157:H7, APEC (avian pathogenic *E. coli*) O1:K1:H7, and UPEC (uropathogenic *E. coli*) CFT073 were also added to the study as controls.

114

Table 16. Relationship between amino acid substitution in *fimH* subunit of fimbria gene of *E. coli* isolated from newly diagnosed IBD patients and disease state, or *E. coli* phylotype

Amino acid substitution	Disease state	P–value	Phylotype	P–value
A48V	HC	0.04*	A	0.35
N91S	IBD	0.004†	B2	0.005†
S99N	IBD	0.004†	B2	0.005†
A140V	HC	0.09	A+B1	0.10
V184A	IBD	0.14	B2	0.04*
R187H	HC	0.09	B2	0.003†
Y216F	IBD	0.80	B2	0.13
A223V	IBD	0.04*	A	0.00002†
A263V	IBD	0.14	B1	0.0004†

IBD, inflammatory bowel disease; HC, healthy controls
A, Alanine; V, Valine; R, Arginine; H, Histidine; N, Asparagine; Y, Tyrosine; S, Serine; E, Glutamic acid; D, aspartic acid; F, Phenylalanine
* $p < 0.05$, † $p < 0.01$

2.6. *E. coli* gene content analysis using microarray technology

The genetic content of randomly selected *E. coli* from 5 UC patients, 6 CD patients and 5 HC were evaluated using Affymetrix GeneChip *E. coli* Genome 2.0 array. Out of 10208 genes that their probes were included in the array, 2861 (28%) genes were present and 1102 (11%) were absent in all *E. coli* isolates (Table 17). All *E. coli* strains carried 58% of non-pathogenic K12 MG1655 genes, compared to 4% of EHEC EDL933, 7% of EHEC Sakai, and 8% of UPEC CFT073 genes. About 40% of EHEC O157:H7 genes were not present in any of *E. coli* strains (Table 17). Genes from UPEC CFT073 showed the highest diversity with 97% of them being present in at least one *E. coli* strain.

When the association of genes to healthy or diseased state was measured (Table 18), 876 (35.2%) of UPEC CFT073 genes were significantly ($P < 0.05$) associated to IBD. None of UPEC CFT073 genes were associated to HC. Only 59 (3.3%) of EHEC EDL933 genes and 13 (3.5%) of EHEC Sakai genes were associated to IBD (Table 18). On the other hand, about 333 (8.1%) of K12 MG1655 genes were significantly related with HC. Only less than 1% of intergenic regions were associated to either of HC or IBD. In total, 962 (9.4%) of array genes were significantly ($P < 0.05$) associated to IBD and 408 (4%) of array genes were significantly related with healthy state.

From all the genes associated to IBD (962), 91% belonged to UPEC CFT073 (Figure 14). In contrast, 80% of all genes associated to healthy state belonged to K12 MG1655 (Figure 15).

Table 17. Distribution of genes in 16 *E. coli* isolated from 5 HC, 5 UC and 6 CD patients

	Total number of genes	Number of genes present in all *E. coli* isolates	Number of genes absent in all *E. coli* isolates
E. coli K12 MG1655	4070	2376 (58%)	89 (2%)
E. coli O157:H7 EDL933	1787	66 (4%)	597 (33%)
E. coli O157:H7 Sakai	373	27 (7%)	175 (47%)
E. coli CFT073	2486	189 (8%)	73 (3%)
Intergenic regions	1297	215 (17%)	39 (3%)
Other	193	13 (7%)	129 (67%)
Total	10208	2861 (28%)	1,102 (11%)

Numbers in parenthesis present the percentage of total number of genes from each strain

Table 18. Association of array genes to healthy and diseased states in 16 *E. coli* isolated from 5 HC, 5 UC and 6 CD patients

	Total number of genes	Number of genes associated to IBD	Number of genes associated to HC
E. coli K12 MG1655	4070	5 (0.1%)	333 (8.1%)
E. coli O157:H7 EDL933	1787	59 (3.3%)	45 (2.5%)
E. coli O157:H7 Sakai	373	13 (3.5%)	17 (4.5%)
E. coli CFT073	2486	876 (35.2%)	0 (0.0%)
Intergenic regions	1297	5 (0.4%)	11 (0.8%)
Other	193	4 (2.1%)	2 (1.0%)
Total	10208	962 (9.4%)	408 (4.0%)

Comparing IBD vs HC, P < 0.05 was considered significant.
Numbers in parenthesis present the percentage of total number of genes from each strain.
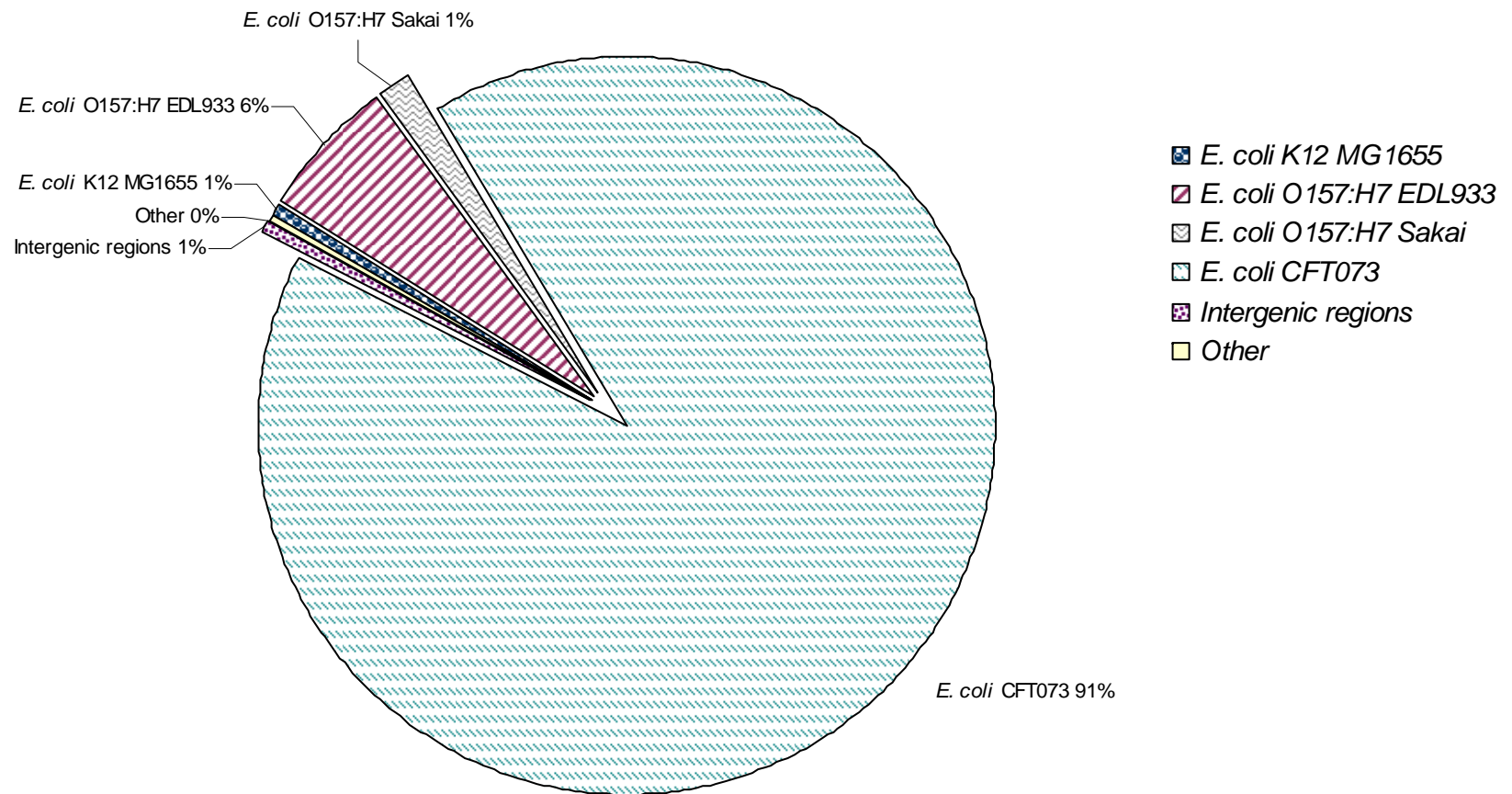
Figure 14. Distribution of *E. coli* genes associated with IBD (P < 0.05).
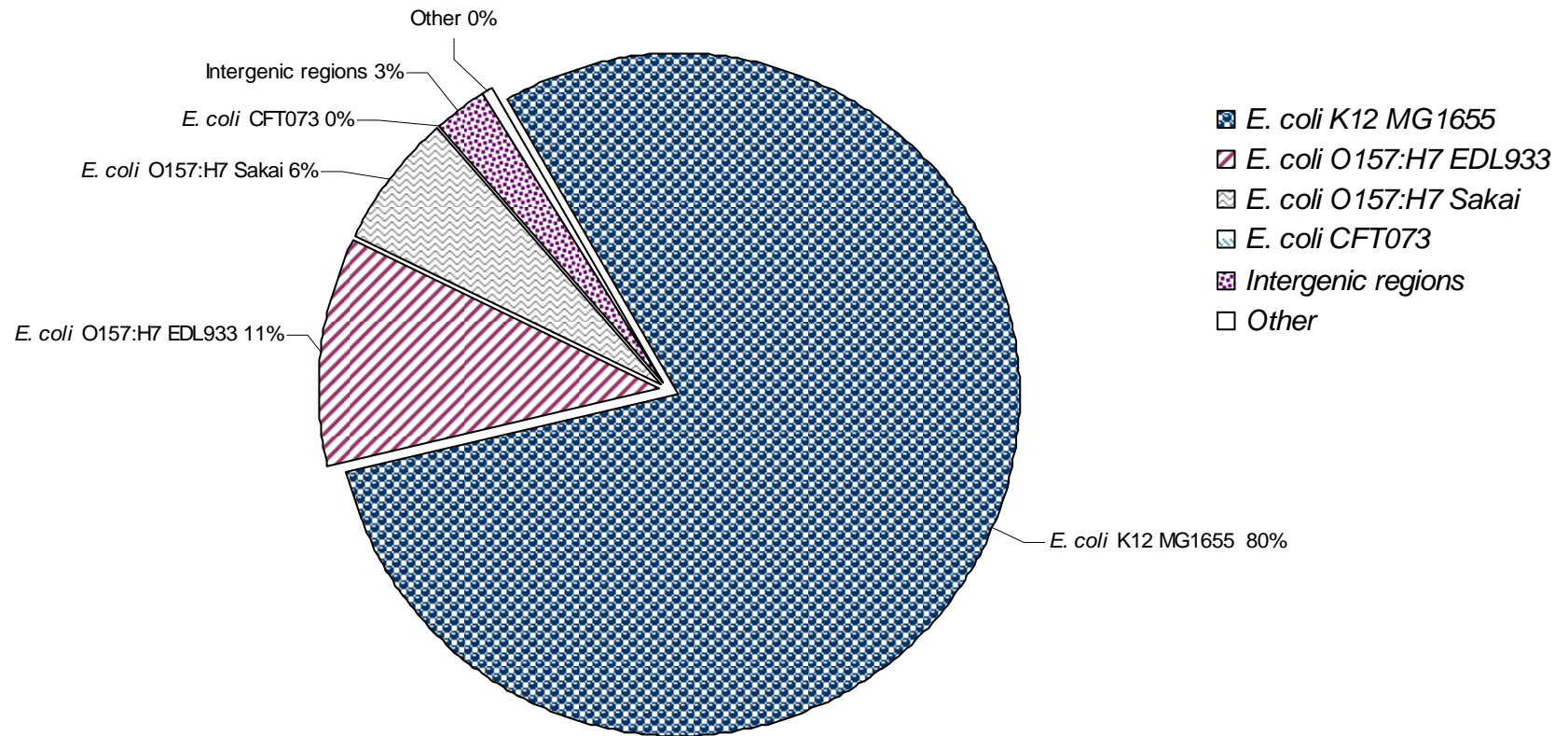
Figure 15. Distribution of *E. coli* genes associated with healthy state (P < 0.05).

The complete list of genes associated with UPEC CFT073, EHEC O157:H7 EDL933, and EHEC O157:H7 Sakai were browsed for known virulence factors. Other recognized virulence factors of *E. coli* were also looked for, regardless of their significant association to IBD. The virulence/ putative virulence genes of UPEC CFT073, EHEC O157:H7 EDL933 and Sakai are listed in tables 19 and 20 respectively.

When genes from UPEC CFT073 were assessed (Table 19), *chuA*, the gene encoding outer membrane haemoglobin receptor which is also a marker for B2 phylotype, was highly related with IBD ($P < 0.01$). Capsule synthesis genes were only detected in IBD isolates and the association was significant for *kpsC*, *kpsD*, *kpsE*, and *kpsS* ($P < 0.05$) (Table 19). Genes involved in F1C fimbrial assembly, *focC*, *focD*, and *focF*, were associated with UC ($P < 0.05$). While putative F1C and S fimbrial proteins (*sfaB*, *sfaC*, *sfaD*) were only present in IBD isolates, the relationship was not statistically significant (Table 19). A number of hypothetical fimbrial-like protein precursors (*yadN*, *ycfZ*, *yfcP*, *yfcQ*, *yfcS*, *yfcV*, *ygiL*, and *yqiH*) were also significantly and exclusively related with IBD ($P < 0.05$).

Well-known EHEC O157:H7 EDL933 genes were either absent in all *E. coli* strains or present in only a few strains with no significant association to IBD (Table 20). Genes encoding putative holing and putative minor tail protein in EHEC O157:H7 Sakai were associated to IBD whereas genes responsible for exclusion of phage super-infection were related with healthy state ($P < 0.05$) (Table 20).

Studying genes from *E. coli* K12 MG1655, significant relations were detected between *csgA* (gene encoding curlin major subunit), *hrsA* (responsible for ompC induction), *sfmA*, *sfmF*, *sfmH*, *yadK*, *yadM*, *yadN*, and *yraK* (encoding putative fimbrial-

like proteins), and *ycfZ* (hypothetical protein) and healthy state (Table 21). Only *fliD*

from *E. coli* K12 MG1655, filament capping protein responsible for filament assembly,

was associated to IBD (P < 0.05).

Table 19. Incidence of UPEC CFT073 virulence/putative virulence genes within the genome of *E. coli* isolated from IBD patients and HC.

Gene	Description	HC (n = 5)	UC (n = 5)	CD (n = 6)	P - value IBD vs HC	P - value CD vs UC vs HC
-	putative major fimbrial subunit precursor	-	10, 12, 18, 35	25, 26, 27, 33	0.009[†]	0.02[*]
-	putative fimbrial adhesin precursor	-	10, 12, 18, 35	26, 27, 33	0.02[*]	0.04[*]
-	putative fimbrial chaperone precursor	-	10, 12, 18, 35	26, 27, 33	0.02[*]	0.04[*]
-	putative minor fimbrial subunit precursor	-	10, 12, 18, 35	26, 27, 33	0.02[*]	0.04[*]
-	putative minor fimbrial subunit precursor	-	10, 12, 18, 35	26, 27, 33	0.02[*]	0.04[*]
chuA	Outer membrane haemoglobin receptor	-	10, 12, 14, 18, 35	26, 27, 33	0.009[†]	0.007[†]
csgA	Major curlin subunit precursor	162	10, 12, 14, 18, 35	26, 27, 33	0.06	0.04[*]
eaeH	putative adhesin	2	10, 12, 14, 18, 35	26, 27, 33	0.06	0.04[*]
entE	Enterobactin synthetase component E	162	10, 12, 14, 18, 35	26, 27, 30, 33	0.02[*]	0.03[*]
entF	Enterobactin synthetase component F	-	10, 12, 18, 35	26, 27, 33	0.02[*]	0.04[*]
fepA	Ferrienterobactin receptor precursor	-	10, 12, 18, 35	26, 27, 33	0.02[*]	0.04[*]
fepG	Ferric enterobactin transport system permease protein fepG	-	10, 12, 18, 35	26, 27, 33	0.02[*]	0.04[*]
fhuB	Ferrichrome transport system permease protein fhuB	2	10, 12, 14, 18, 35	23, 26, 27, 33	0.02[*]	0.03[*]
fhuF	Ferric iron reductase protein fhuF	-	12, 14, 18, 35	27, 30, 33	0.02[*]	0.04[*]
focC	F1C periplasmic chaperone	-	12, 18, 35	-	0.21	0.02[*]
focD	F1C fimbrial usher	-	12, 18, 35	-	0.21	0.02[*]
focF	F1C minor fimbrial subunit F precursor	-	12, 18, 35	-	0.21	0.02[*]
kpsC	KpsC protein	-	12, 18, 35	26, 27, 33	0.04[*]	0.11
kpsD	KpsD protein	-	12, 18, 35	23, 26, 27, 33	0.02[*]	0.06
kpsE	KpsE protein	-	12, 18, 35	23, 26, 27, 33	0.02[*]	0.06
kpsS	KpsS protein	-	12, 18, 35	23, 26, 27, 33	0.02[*]	0.06
lacZ	Beta-galactosidase	-	10, 12, 18, 35	26, 27, 33	0.02[*]	0.04[*]
malX	PTS system, maltose and glucose-specific IIABC component	-	10, 12, 18, 35	26, 27, 33	0.02[*]	0.04[*]

Table 19. Incidence of UPEC CFT073 virulence/putative virulence genes within the genome of *E. coli* isolated from IBD patients and HC.

Gene	Description	HC (n = 5)	UC (n = 5)	CD (n = 6)	P - value IBD vs HC	P - value CD vs UC vs HC
ompA	Outer membrane protein A precursor	2, 7, 143	12	26	0.11	0.25
ompC	Outer membrane protein C precursor	-	12	23, 26	0.21	0.37
papA1	pilus associated with pyelonephritis A1	-	-	-	-	-
papA2	pilus associated with pyelonephritis A2	-	-	-	-	-
papC	papC protein	-	12, 18, 35	23, 25	0.08	0.12
papD1	pilus associated with pyelonephritis D1	-	12, 18, 35	27	0.13	0.08
papD2	pilus associated with pyelonephritis D2	-	12, 18, 35	27	0.13	0.08
papE	pilus associated with pyelonephritis E	-	12, 18, 35	-	0.21	0.02[*]
papF	pilus associated with pyelonephritis F	-	12, 18, 35	27	0.13	0.08
papG	pilus associated with pyelonephritis G	-	12	-	0.5	0.31
papH	pilus associated with pyelonephritis H	-	12, 18, 35	27	0.13	0.08
PapJ	pilus associated with pyelonephritis J	-	12, 18, 35	27	0.13	0.08
PapK	pilus associated with pyelonephritis K	-	12, 18, 35	27	0.13	0.08
pic	serine protease precursor	-	12	26, 33	0.21	0.37
sat	secreted autotransporter toxin	-	12	23, 26	0.21	0.37
sfaB	putative F1C and S fimbrial switch Regulatory protein	-	12, 18, 35	27, 35	0.08	0.12
sfaC	putative F1C and S fimbrial switch Regulatory protein	-	12, 18, 35	27, 34	0.08	0.12
sfaD	putative minor F1C fimbrial subunit precursor	-	12, 18, 35	27, 33	0.08	0.12
sitA	SitA protein	-	10, 12, 14, 18, 35	23, 26, 27, 33	0.003[†]	0.005[†]
sitB	SitB protein	-	10, 12, 14, 18, 35	23, 26, 27, 33	0.003[†]	0.005[†]
sitD	Salmonella iron transport gene	-	10, 12, 14, 18, 35	23, 26, 27, 33	0.003[†]	0.005[†]
yadN	Hypothetical fimbrial-like protein yadN precursor	-	10, 12, 18, 35	26, 27, 33	0.02[*]	0.04[*]
ycfZ	hypothetical protein, homolog of virulence factor	-	10, 12, 18, 35	26, 27, 33	0.02[*]	0.04[*]

Table 19. Incidence of UPEC CFT073 virulence/putative virulence genes within the genome of *E. coli* isolated from IBD patients and HC.

Gene	Description	HC (n = 5)	UC (n = 5)	CD (n = 6)	P - value IBD vs HC	P - value CD vs UC vs HC
yfcP	Hypothetical fimbrial-like protein yfcp precursor	-	12, 18, 35	26, 27, 33	0.04*	0.11
yfcQ	Hypothetical fimbrial-like protein yfcQ precursor	-	10, 12, 18, 35	26, 27, 33	0.02*	0.04*
yfcS	Hypothetical fimbrial chaperone yfcS precursor	-	10, 12, 18, 35	26, 27, 33	0.02*	0.04*
yfcV	Hypothetical fimbrial-like protein yfcV precursor	-	10, 12, 18, 35	26, 27, 33	0.02*	0.04*
ygiL	Hypothetical fimbrial-like protein ygiL precursor	-	12, 18, 35	23, 26, 27, 33	0.02*	0.06
yqiH	Hypothetical fimbrial chaperone yqiH precursor	-	12, 18, 35	26, 27, 33	0.04*	0.11

* $p < 0.05$, † $p < 0.01$

Table 20. Incidence of EHEC O157:H7 EDL933 and Sakai virulence/putative virulence genes within the genome of *E. coli* isolated from IBD patients and HC

Gene	Description	HC (n = 5)	UC (n = 5)	CD (n = 6)	P – value IBD vs HC	P – value CD vs UC vs HC
EHEC O157:H7 EDL933						
-	putative RTX family exoprotein	-	14	30	0.32	0.59
artJ	arginine 3rd transport system binding protein, putative virulence	-	-	-	-	-
chuA	outer membrane haemoglobin receptor	-	14, 35	30	0.21	0.26
csgA	curlin major subunit, coiled surface structures	-	-	30	0.5	0.41
eae	Intimin adherence protein	-	-	-	-	-
eaeH	putative adhesin	-	14	25, 30	0.21	0.37
entD	enterobactin synthetase component D	143	-	25, 30	0.93	0.37
espA	EspA protein secreted by the type III secretion system of the LEE, involved in the formation of surface appendages	-	-	-	-	-
espB	EspB protein secreted by the type III secretion system of the LEE	-	-	-	-	-
espD	*E. coli* secreted protein EspD	-	-	-	-	-
espF	*E. coli* secreted protein EspF	-	-	-	-	-
exbB	uptake of B colicins	-	-	30	0.5	0.41
fepC	Gene encoding ferric enterobactin transport ATP-binding protein	-	-	-	-	-
fliC	Gene encoding *E. coli* flagellin, major subunit, flagellar biosynthesis; flagellin, filament structural protein	-	10, 18, 35	27	0.13	0.08
fliD	flagellar biosynthesis; filament capping protein; filament assembly	-	-	-	-	-
ompC	outer membrane protein 1b (Ib;c)	2, 7,	14, 18, 35	25, 27	0.21	0.29

Table 20. Incidence of EHEC O157:H7 EDL933 and Sakai virulence/putative virulence genes within the genome of *E. coli* isolated from IBD patients and HC

Gene	Description	HC (n = 5)	UC (n = 5)	CD (n = 6)	P – value IBD vs HC	P – value CD vs UC vs HC
		143, 162				
per	Gene encoding perosamine synthetase, O157 antigen	-	-	-	-	-
stx1A	shiga-like toxin 1 subunit A	-	-	-	-	-
stx1B	shiga-like toxin 1 subunit B	-	-	-	-	-
stx2A	shiga-like toxin II A subunit	-	-	-	-	-
stx2B	shiga-like toxin II B subunit	-	-	-	-	-
tir	putative translocated intimin receptor protein	-	-	-	-	-
ureD	Gene encoding the urease-associated protein UreD	-	-	-	-	-
EHEC O157:H7 Sakai						
-	putative complement resistance protein	-	-	25, 26	0.32	0.15
-	BfpM-like protein	-	-	33	0.5	0.41
-	putative holin	-	10, 12, 14, 18, 35	23, 25, 26, 27	0.003[†]	0.005[†]
-	putative minor tail protein	-	10, 12, 18, 35	23, 25, 27	0.03[*]	0.04[*]
-	phage superinfection exclusion protein	5, 7	-	-	0.03[*]	0.08
Aida-I	adhesin involve in diffuse adherence	-	18, 35	26, 27	0.13	0.29
KdpF	protein of high-affinity K^+ transport system	7	14, 18, 35	23, 26, 27, 30, 33	0.06	0.11

[*] $p < 0.05$, [†] $p < 0.01$

Table 21. Incidence of *E. coli* K12 MG1655 virulence/putative virulence genes within the genome of *E. coli* isolated from IBD patients and HC

Gene	Description	HC (n = 5)	UC (n = 5)	CD (n = 6)	P – value IBD vs HC	P – value CD vs UC vs HC
agn43	antigen 43, phase-variable bipartite outer membrane fluffing protein, autoaggregation and biofilm formation factor (aka *flu*)	-	-	-	-	-
artJ	arginine 3rd transport system periplasmic binding protein, putative virulence	all	all	all	-	-
crl	transcriptional regulator of *csgA* gene for curli surface fibers	all	all	all	-	-
csgA	curlin major subunit, coiled surface structures	2, 5, 7, 143	-	23, 25	0.02[*]	0.03[*]
csgB	curlin minor subunit precursor, similar to CsgA	all	all	all	-	-
csgE	assembly/transport component in curli production	all	all	all	-	-
eaeH	putative adhesin	2, 5, 143	14	23, 25, 30	0.39	0.41
feoA, B	ferrous iron transport protein A, B	all	all	all	-	-
fepA	outer membrane receptor for ferric enterobactin and colicins B, D	all	all	all	-	-
fepB	ferric enterobactin (enterochelin) binding protein; periplasmic	all	all	all	-	-
fepC	Gene encoding ferric enterobactin transport ATP-binding protein	all	all	all	-	-
fepD	ferric enterobactin (enterochelin) transport protein	all	all	all	-	-
fepE	ferric enterobactin (enterochelin) transport protein	2, 5, 7	14, 35	23, 25	0.39	0.66
fepG	ferric enterobactin transport protein	2, 5, 7	14	23, 25, 30	0.39	0.41
fimA	major type 1 subunit fimbrin (pilin)	2, 5, 7, 162	14	23, 25, 26, 30	0.21	0.13
fimH	Gene encoding D-mannose specific adhesin of type 1 fimbriae	all	all	all	-	-

Table 21. Incidence of *E. coli* K12 MG1655 virulence/putative virulence genes within the genome of *E. coli* isolated from IBD patients and HC

Gene	Description	HC (n = 5)	UC (n = 5)	CD (n = 6)	P – value IBD vs HC	P – value CD vs UC vs HC
fimZ	fimbrial Z protein; probable signal transducer	2, 5, 7, 143, 162	14	23, 25, 30	0.02*	0.04*
fliC	flagellin, filament structural protein, flagellar major subunit	-	-	-	-	-
fliD	flagellar biosynthesis; filament capping protein; filament assembly	-	10, 12, 18, 35	26, 27, 33	0.02*	0.04*
hlyE	hemolysin E	2, 5, 143, 162	14	23, 25, 30	0.12	0.17
hrsA	protein modification enzyme, induction of *ompC*	2, 5, 7, 143	-	25	0.006†	0.02*
ompA	outer membrane protein 3a	5, 162	10, 14, 18, 35	23, 25, 27, 30, 33	0.11	0.24
ompC	outer membrane protein 1b	5	-	23	0.55	0.59
ompT	outer membrane protein 3b (a), protease VII	5, 143, 162	10, 12, 18, 35	26, 27, 33	0.89	0.59
sfmA	putative fimbrial-like protein	2, 5, 7, 143, 162	14	23, 25, 30	0.02*	0.04*
sfmF	putative fimbrial-like protein	2, 5, 7, 143, 162	14	23, 25, 30	0.02*	0.04*
sfmH	involved in fimbrial assembly	2, 5, 7, 143, 162	14	23, 25, 30	0.02*	0.04*
yadK	putative fimbrial protein	2, 5, 7, 143	-	23, 25	0.02*	0.03*
yadM	putative fimbrial-like protein	2, 5, 7, 143	-	23, 25	0.02*	0.03*
yadN	putative fimbrial-like protein	2, 5, 7, 143	-	23, 25	0.02*	0.03*
ycfZ	hypothetical protein, homolog of virulence factor	2, 5, 7, 143		23, 25	0.02*	0.03*
yjaA	Gene encoding hypothetical protein, phylogenetic marker gene	2, 5	10, 12, 18, 35	23, 26, 27, 30, 33	0.11	0.24
yraK	putative fimbrial protein	2, 5, 7, 143	-	23, 25	0.02*	0.03*

* $p < 0.05$, † $p < 0.01$

Comparing genetic content of *E. coli* Nissle 1917 and UPEC CFT073 revealed about 10% of difference in genes included in the array. In total, 5831 genes were present and 3341 genes were absent in both strains, compare to 575 genes that were specific for UPEC CFT073 and 460 genes that were only present in *E. coli* Nissle 1917. The distribution of array genes in *E. coli* Nissle 1917 and UPEC CFT073 strain is presented in Table 22. Differences in genes belonged to UPEC CFT073 (564 genes) accounted for the majority of the dissimilarities between the two strains. UPEC CFT073 strain also lacked about 105 of *E. coli* K12 MG1655 genes present in *E. coli* Nissle 1917 (Table 22).

Table 23 shows a selected list of genes that differ between *E. coli* Nissle 1917 and UPEC CFT073. Expectedly, lack of many virulence genes in probiotic *E. coli* Nissle 1917 was responsible for a great part of the differences. These virulence factors include Pic and Sat serine proteases, Ag 43, Type I fimbriae, hemolysin, flagellin, F1C fimbriae, capsule proteins, outer membrane proteins A and C, and P fimbriae. In contrast, *E. coli* Nissle 1917 carried some of K12 and EHEC O157:H7 genes including *aida-I* from EHEC O157:H7 Sakai (Table 23).

Table 22. Comparing the distribution of array genes in probiotic Nissle 1917 and UPEC CFT073

	Total number of genes	Number of genes present in *E. coli* Nissle 1917	Number of genes present in UPEC CFT073
E. coli K12 MG1655	4070	2990 (73%)	2885 (71%)
E. coli O157:H7 EDL933	1787	318 (18%)	291 (16%)
E. coli O157:H7 Sakai	373	68 (18%)	55 (15%)
E. coli CFT073	2486	1901 (76%)	2465 (99%)
Intergenic regions	1297	851 (66%)	704 (54%)
Other	193	34 (18%)	31 (16%)
Total	10208	6162 (60%)	6431 (63%)

Numbers in parenthesis present the percentage of total number of genes from each strain.

Table 23. A selected list of the genes that are present in uropathogenic *E. coli* CFT073 but absent in probiotic *E. coli* Nissle 1917 (A) and the genes that are present in *E. coli* Nissle 1917 but absent in UPEC CFT073 (B)

AFFY	Strain_Gene	Description
A. CFT073		
1768916	CFT073_c0308	Haemolysin expression modulating protein
1760743	CFT073_c0350	Pic serine protease precursor
1769263	CFT073_c1273	Antigen 43 precursor
1760775	CFT073_c2897	Type 1 fimbriae Regulatory protein fimB
1760979	CFT073_c2898	Type 1 fimbriae Regulatory protein fimB
1762852	CFT073_c3556	Prophage P4 integrase
1762642	CFT073_c3649	Haemolysin expression modulating protein
1768656	CFT073_c3655	Antigen 43 precursor
1762239	CFT073_c5216	Prophage P4 integrase
1759116	CFT073_c5371	Prophage P4 integrase
1767374	CFT073_emrB	Multidrug resistance protein B
1767582	CFT073_fhuA	Ferrichrome-iron receptor precursor
1768034	CFT073_fimA	Type-1 fimbrial protein, A chain precursor
1763858	CFT073_fliC	Flagellin
1762790	CFT073_focA	F1C major fimbrial subunit precursor
1759482	CFT073_focG	F1C minor fimbrial subunit protein G presursor
1759111	CFT073_kpsM	KpsM protein (capsule protein)
1764875	CFT073_kpsT	KpsT protein (capsule protein)
1766301	CFT073_mchE	Microcin H47 secretion protein
1768172	CFT073_ompA	Outer membrane protein A precursor
1760091_s	CFT073_ompC	Outer membrane protein C precursor
1769066	CFT073_papA	PapA protein
1759535	CFT073_papA	PapA protein
1761392_s	CFT073_papE	PapE protein
1764925_s	CFT073_papG	PapG protein
1768090	CFT073_sat	Secreted auto transpoter toxin
1769039	CFT073_yadC	Hypothetical fimbrial-like protein yadC precursor
B. Nissle 1917		
1765832_s	EDL933_fliC	Flagellar biosynthesis; flagellin, filament structural protein
1764034_s	EDL933_ompC	Outer membrane protein 1b (Ib;c)
1769176	MG1655_b1145	Putative phage repressor
1762168_s	MG1655_emrB	Multidrug resistance; probably membrane translocase
1766597_s	MG1655_entD	Enterochelin synthetase, component D
1769102_s	MG1655_flhE	Flagellar protein
1759842_s	MG1655_hemM	Outer-membrane lipoprotein
1765595_s	MG1655_ompA	Outer membrane protein 3a (II*;G;d)
1767586_s	SAKAI_ECs1396	Aida-I

DISCUSSION

1. Microbial diversity of gut biopsy tissues in IBD

The role of microbes in the pathogenesis of IBD has been well documented: (a) dysbiosis, the imbalance between protective and harmful biota has been reported in IBD (Tamboli et al. 2004); (b) IBD patients have a greater number of mucosal-associated bacteria than controls (Swidsinski et al. 2002; Barnich et al. 2007b); (c) clinical symptoms of IBD generally subside following antibiotic administration/probiotics usage/luminal wash (Danese et al. 2004; Barnich et al. 2007b); (d) diversion of the fecal stream through ileo-rectal shunt reduces the symptoms of intestinal inflammation in IBD patients, whereas reintroduction of the fecal stream triggers the inflammation (Danese et al. 2004; Barnich et al. 2007b); and (e) IL-10 knockout mice spontaneously develop colitis, but inflammation is absent when IL-10$^{-/-}$ mice were kept germ free (Sellon et al. 1998; Danese et al. 2004; Sartor 2006).

In molecular analysis of bacterial populations associated with IBD, we are faced with a three main challenges. First, analyses of microbiota should preferably be done on biopsy tissue rather than feces. Although it is much easier to obtain fecal samples, they obscure the differences between the sites (Frank et al. 2001; Eckburg et al. 2005; Lepage et al. 2005; Mentula et al. 2005; Swidsinski et al. 2005; Sokol et al. 2008a). Second, identification of adequate controls is difficult. The ideal control needs to be an individual that voluntarily submits to colonoscopy, but does not have any gastrointestinal symptoms

that could potentially bias the analysis. Third, there is significant microbial variation between individuals, and a major portion of this variation can be accounted for by the individual's genotype (Zoetendal et al. 1998; Zoetendal et al. 2001; Zoetendal et al. 2002). Therefore, to obtain reliable results, sizable numbers of individuals must be sampled.

The first two criteria were met because analyzed biopsy tissues were taken from a case-control population-based IBD tissue bank held at the University of Manitoba (Bernstein et al. 2003). At the time of this study, the highest resolution technique based on 16S rDNA was the use of clone libraries which was an expensive and time-consuming procedure (Eckburg et al. 2005; Manichanh et al. 2006). Consequently, few samples can be processed with clone libraries, so to meet our third criteria, large numbers of samples, we used ARISA and T-RFLP.

ARISA and T-RFLP are both accepted methods for assessing complex microbial communities (Hartmann et al. 2005; Danovaro et al. 2006). Recently, Danovaro et al. showed that ARISA is more efficient than T-RFLP for estimating the diversity of aquatic communities (Danovaro et al. 2006). However, our study indicated that neither ARISA nor T-RFLP could group tissues into phenotypically meaningful clusters. For example, a majority of tissues clustered into inflamed or non-inflamed sets only if a very liberal cutoff (70, 80%) was used (Figures 5, 7). A more conservative cutoff resulted in clustering that according to our definition (based on phenotype of tissue or disease) was not meaningful. Cluster analysis of the ARISA (Figure 5) and TRFLP data (Figure 7) did not clearly differentiate between controls, CD, and UC when the clustering was based on the disease state; however, a far higher level of clustering was obtained using the

inflamed versus non-inflamed criteria. The better clustering with T-RFLP (90%) than ARISA (78%) (Figures 5, 7) could be explained by the fact that T-RFLP is based on the 16S rDNA structural gene, which has a relatively low level of taxonomic resolution compared to the ribosomal intergenic spacer region target of ARISA, which not only has significant sequence variation, but also a length polymorphism (Leckie et al. 2004; Dorigo et al. 2005). The inability to clearly cluster tissues by disease state is most likely the consequence of interindividual microbiota variation (Zoetendal et al. 1998; Zoetendal et al. 2001; Zoetendal et al. 2002). Using more than one restriction enzyme for T-RFLP could potentially increase the specificity of the results.

The effects of host genotype on gut microflora was most clearly demonstrated in homozygotic twins (Zoetendal et al. 2001). The DGGE profiles of the 16S genes indicated that the variation in population was more influenced by host genome than the environment, even though the environmental factors were still important. This concept was further supported by the fact that spouses (different genotypes) sharing the same diet, housing, etc. (same environment) were less similar in gut microbiota composition than the siblings of the spouses (same genome) living in a different environments. Eckburg et al. constructed 16S rDNA clone libraries from biopsy tissues of three healthy individuals, consisting of over 13,000 clones (Eckburg et al. 2005). Significant differences were detected among individuals as well as different sites of the same subject, but the intersubject variance was the greatest. In a more recent study, Turnbaugh at al. characterized the fecal microbial composition of monozygotic and dizygotic twins and their mothers, concordant for their leanness or obesity, to address how genotype, environmental factors and obesity affects the composition of microbiota (Turnbaugh et al.

2009). Their study showed that individuals from the same family had more similar bacterial community structure compared to unrelated individuals, and that all sampled individuals shared a wide range of microbial genes, indicating the existence of "core microbiome" at the gene, rather than lineage level. Turnbaugh et al. also demonstrated an association between obesity and decrease in gut microbial diversity (Turnbaugh et al. 2009).

Recognition of the extent of diversity in gut ecology is essential to our understanding of the association between microbial community's structure and function (Curtis et al. 2004), as well as the relationship between the microbial flora and the human host (Tannock 1999). The gut microbiota, in general, develops over the first few years of life and then remains remarkably stable, except in the case of certain diseases (Rath 2003). Prolonged alteration of this balanced composition may result in the chronic stimulation of the mucosal immune system and, consequently, loss of tolerance to the commensal bacteria (Mahida et al. 2004).

Given the limitations in analyzing the ARISA and T-RFLP data using hierarchical clustering methods, we employed statistical indices of ecological diversity and similarity. Diversity indices and similarity indices are often used interchangeably in the gut microbiology literature and it is useful to define these terms explicitly. Diversity indices are made up of richness, abundance, and evenness. Species richness is simply the number of different species, or operational taxonomic units (OTU) present, and diversity is a means of weighting OTU abundance and evenness. Similarity, on the other hand, is a measure of shared species. For example, two microbial communities can have identical

species richness, but zero similarity, because the same species are not present in both communities.

Previous studies have demonstrated a reduction in diversity in both feces (Manichanh et al. 2006; Scanlan et al. 2006), and mucosal associated (Ott et al. 2004) microbiota of IBD tissues as compared to controls. Using two approaches (ARISA and T-RFLP) we have shown similar results (Table 3), but have gone further and examined non-inflamed tissue in IBD patients (Table 4). Although the results of our study did not show any differences between the diversity of different anatomical sites, which is similar to the results of Frank et al. (Frank et al. 2007), our results indicate that there is a clear difference in both richness and diversity between the microbiota of the inflamed and non-inflamed sites, but this is less clear with disease state as the comparator. This is an interesting finding since one of the clinical features of CD are the skip-regions, which are areas of mucosa that are not inflamed and lie adjacent to inflamed tissue (Swidsinski et al. 2005; Thompson-Chagoyan et al. 2005). In UC, mainly a large bowel disease, there is usually a gradient from non-inflamed to inflamed tissue, and over time the inflammation progresses into the non-inflamed areas. Organisms present in the non-inflamed or pre-inflamed tissue in IBD patients are of interest because biological processes in these regions potentially represent pathogenic entities. Our results clearly indicate that there is an increase in diversity from controls to the non-inflamed tissues, and then when the inflammation sets in, the diversity of microbiota declines. We hypothesize that there is a recruitment phase in the non-inflamed tissue and species involved in this phase may well hold the keys to the pathogenesis of IBD.

Results of other studies suggest similar microbial composition for inflamed and non-inflamed IBD tissues (Seksik et al. 2005; Bibiloni et al. 2006; Gophna et al. 2006). Bibiloni et al. (Bibiloni et al. 2006) and Seksik et al. (Seksik et al. 2005) used DGGE and TTGE to calculate similarity indices, not species richness or diversity, for inflamed and non-inflamed tissues for a given patient. Gophna et al. (Gophna et al. 2006) compared 16S clone libraries from inflamed and non-inflamed sites in the same patient, based on a modified similarity index. In our study, we pooled all inflamed and non-inflamed biopsy data, and then made comparisons regardless of the subject, gut location, or disease. We observed less similarity between the inflamed and non-inflamed libraries (64%–76%, Table 5) than previous studies (90%–97%) (Seksik et al. 2005; Bibiloni et al. 2006).

Assignment of T-RFs to hierarchical taxonomic groups demonstrated statistically significant, or near significant differences, in the phyla *Bacteroidetes* and *Firmicutes* (Table 7). In phylum *Bacteroidetes* there is a putative loss of species as the tissue becomes inflamed, while in the class *Clostridia* of the phylum *Firmicutes* an increase in the number of species is observed. In particular, we observed a highly significant increase in a number of unclassified *Clostridia* in the inflamed tissue. Although the increase in the population of *Proteobacteria* we observed in UC and CD patients compared to healthy controls was not statistically significant, Frank et al. detected more *Proteobacteria* in their IBD samples using rRNA clone libraries (P < 0.001) (Frank et al. 2007). Perhaps ARISA and T-RFLP techniques were not sensitive enough to detect such differences between the two communities.

The composition of feces microbiota of CD patients was assessed in several studies (Manichanh et al. 2006; Scanlan et al. 2006; Sokol et al. 2006). Manichanh et al.

(Manichanh et al. 2006) constructed metagenomics libraries from feces taken from CD patients while Scanlan et al. (Scanlan et al. 2006) generated DGGE profiles for different microbial groups. Both studies found significant changes in the composition of the microbiota of CD patients, mainly in the *Firmicutes* (decreased diversity), and *Bacteroidetes* (less diversity) (Manichanh et al. 2006; Scanlan et al. 2006). A more recent study by Frank et al. also found less members of phylums *Firmecutes* and *Bacteroides* in both CD and UC patients (Frank et al. 2007). In our study we observed a putative increase in the diversity of *Firmicutes*, most significantly in unclassified *Firmicutes*. Given that the unclassified *Firmicutes* represent less well understood bacteria, it is likely that the bacteria represented by these T-RFs are good targets for future studies in IBD.

Recently Sokol et al reviewed a large number of studies that investigated microbial composition of IBD patients and concluded that sampling and technical problems are responsible for much of the disagreements between different research groups (Sokol et al. 2008a). With recent improvements in the quality of microbial investigations of communities such as pyrosequensing, more thoughtful selection of patients (recruitment of newly diagnosed and untreated patients) and greater attention to the way in which specimens are collected are needed to be considered in order to improve future microbiological studies in IBD (Sokol et al. 2008a).

In summary, we have demonstrated that diversity of microbial species in non-inflamed IBD tissue likely forms an intermediary community of organisms in transition to the inflamed state. We believe that further analysis of the organisms involved in this transitional state may lead to significant advances in our understanding of the

pathogenesis of IBD. Future studies should focus on isolating these species and trying to

address their relationship to the pathophysiology of CD and UC.

2. Characterization of *E. coli* isolated from IBD gut biopsy tissues

It is generally accepted that the chronic inflammation characteristic of UC and CD is a result of the inability of gut immune system to control the basal level of inflammation. However, the antigenic stimuli responsible for overreaction of gut immune system are still unknown but are likely to be from bacteria that resides within human gut (Frank et al. 2007). Despite general agreement upon the central role of microorganisms in the pathogenesis of IBD, specific microbial etiologies remain elusive (Korzenik 2005). To date, many different microorganisms have been linked to IBD. Theses include, but are not limited to, *Mycobacterium paratuberculosis*, *E. coli*, *Helicobacter pylori, Listeria monocytogenes*, *Yersinia pseudotuberculosis*, *Mycoplasma* species, *Chlamydia* species, and even measles virus (Guarner 2005; Eckburg et al. 2007). *Faecalibacterium prausnitzii* on the other hand, had been identified as an anti-inflammatory gut bacterium in CD (Sokol et al. 2008b).

Mycobacterium avium spp paratuberculosis (MAP) has been a popular topic of study in IBD for a long time because of its association with Johne's disease in cattle (Hendrickson et al. 2002; Sartor 2006; Barnich et al. 2007b; Feller et al. 2007). However, the evidence supporting the etiologic role of MAP in IBD is inconsistent and inconclusive. In recent years however, *E. coli* have been strongly implicated in IBD (Darfeuille-Michaud et al. 2004; Sartor 2006; Barnich et al. 2007b; Kotlowski et al. 2007; Rhodes 2007).

Darfeuille-Michaud and Colombel have been able to recover a new pathovar of *E. coli*, adherent invasive *E. coli* (AIEC), from the ileum of 22% of CD patients. These strains appear to have all the features necessary to colonize the intestinal mucosa, cross the epithelial barrier, infect epithelial cells and macrophages, and induce pro-inflammatory cytokines (Darfeuille-Michaud et al. 2004; Tamboli et al. 2004). In our previous research (Kotlowski et al. 2007), we revealed the association between certain phylotypes of *E. coli* (B2+D) and IBD. Furthermore, we characterized a number of putative virulence factors in *E. coli*, including some adhesins and serine protease autotransporters (SPATE), and linked them to the disease condition (Kotlowski et al. 2007).

2.1. Isolation and enumeration of *E. coli*

Several studies reported higher numbers of *E. coli* in association with IBD intestinal tissues (Fujita et al. 2002; Martin et al. 2004; Ryan et al. 2004; Mylonaki et al. 2005; Swidsinski et al. 2005; Rhodes 2007). Martin et al.(Martin et al. 2004) found a significant increase in the numbers of *E. coli* only in specimens from patients with CD, but not in those with UC. Like wise, Darfeuille-Michaud et al. (Darfeuille-Michaud et al. 2004) isolated *E. coli* in higher numbers from tissues of patients with IBD than from controls, but did not enumerate the bacteria on the epithelial tissue. Mylonaki et al. (Mylonaki et al. 2005) used fluorescence microscopy to show that the numbers of *E. coli* were high in the rectal tissue of patients with UC but not of controls, but this methodology did not allow for decimal enumeration of *E. coli*. In this research, we have been able to isolate 1,000-10,000 times more *E. coli* from IBD intestinal biopsies than

biopsies from HC (Figure 8). Conte et al. (Conte et al. 2006) showed that Gram-negative bacteria, including *E. coli*, have increased by 3–4 logs in tissues of patients with IBD, a result strikingly similar to ours. It is thus clear that the numbers of *E. coli* do increase on the epithelial tissue of patients with IBD, irrespective of the differences in techniques used.

2.2. Phylogenetic analysis of *E. coli*

Although AIEC has been associated with IBD on a consistent basis (Darfeuille-Michaud et al. 1998; Darfeuille-Michaud 2002; Rhodes 2007), its phylogenetic position in relation to other pathogenic and non-pathogenic *E. coli* has not been explored. AIEC strains have been isolated from IBD tissue but it is unclear what their relationship is to *E. coli* that resides in the gut but causes extra-intestinal disease. By extension it is not known whether AIEC forms a separate phylogenetic lineage within the *E. coli* species.

The MLST analysis of seven housekeeping genes identified 26 different STs (new and old) among 45 *E. coli* isolates (Table 9) and even though 11 of the 26 ST were novel, this did not place them into a separate lineage (Figure 10). In fact, most were placed into the same cluster as the urinary tract pathogen CFT073 and the avian pathogen APEC 01:K1:H7 (Figure 10). Thess findings were in agreement with the results of a study by Baumgart et al. (Baumgart et al. 2007) which used MLST to demonstrate that *E. coli* isolated from ileum of CD patients were predominantly novel in phylogeny. The same study also showed that *E. coli* isolated from ileal CD contain elements similar to ExPEC particularly UPEC and APEC (Baumgart et al. 2007). In our study, the eBURST analysis of allelic profiles grouped IBD isolates into separate clonal clusters (P=0.02) (Table 10)

suggesting that these isolates may have taken advantage of a specific "IBD microenvironment". Both allelic profiles (Table 10) and phylogenetic tree (Figure 10) separated *E. coli* isolates based on their phylotypes. Actual phenotype differences were not obvious as biochemical profiling did not show any separation (Table 8).

As demonstrated by both phylogenetic and clonal analyses, sequences of seven housekeeping genes of *E. coli* Nissle 1917 and UPEC CFT073 were identical. While Nissle is a probiotic used to change the balance of gut microbiota towards beneficial microbes, UPEC is a virulent strain of *E. coli* causing mainly urinary tract and intestinal infections. It has been previously shown that *E. coli* exhibits considerable intraspecies genomic variation in order to maintain its adaptation with various environments (Ihssen et al. 2007). For instance, the fact that the chromosomal lengths of the natural *E. coli* isolates vary widely (4.5-5.5Mb) indicates some events of foreign DNA insertion and deletion (Ihssen et al. 2007). This phenomenon of lateral genetic transfer includes both pathogenic and non-pathogenic exchanges. So, it is likely that Nissle and UPEC share the same ancestral progeny but have adopted different events of recombination and lateral genetic transfer. This brings up the prospect of Nissle acquiring some virulent features and become a pathogenic strain, putting its position as a safe probiotic under question.

The *E. coli* reference collection (ECOR) is a set of 72 *E. coli* strains isolated from a wide range of natural sources that represent genetic diversity (Ochman et al. 1984). Earlier phylogenetic studies on this reference collection, divided *E. coli* to four phylogenetic groups: A and B1 (less pathogenic types), and B2 and D (more pathogenic types) (Wirth et al. 2006; Moulin-Schouleur et al. 2007). Clermont et al. (Clermont et al. 2000) developed a method to type pathogenic *E. coli* using *chuA*, a gene required for

haem transport in enterohaemorrhagic *E. coli* O157:H7*, yjaA*, a gene identified in *E. coli* K12 but which has no known function, and TSPE4.C2, a cryptic fragment that was identified from subtractive libraries. These genes, when applied to 230 isolates, determined that type B2, and to a lesser extent type D, included virulent extraintestinal strains of *E. coli*, but the prevalence of B2 and D in gastrointestinal isolates was not determined. Analyzing *E. coli* from fecal samples from a range of geographically separate healthy human subjects determined that non-pathogenic groups A and B1 were most prevalent, while group D made up only 15% and group B2 11% of isolates (Duriez et al. 2001).

For long it was believed that A, B1, B2, and D lineages are the ancestral source of polymorphism within *E. coli*. However, a more recent study on a wider range of *E. coli* strains by Wirth *et al*. revealed the presence of hybrid groups (AB1, ABD) which hold ancestry from multiple phylotypes (Wirth et al. 2006). Wirth et al. showed that the hybrid groups account for one-third of total *E. coli* strains and contain *E. coli* that underwent more frequent recombination events. The hybrid groups were also rich in pathogens, indicating a possible link between virulence and recombination (Wirth et al. 2006). According to this model, any commensal *E. coli* has the ability to acquire virulence by horizontal gene transfer. Being more exposed to host's immune system likely derives higher rates of mutation and recombination (Wirth et al. 2006). Based on MLST analysis, we showed that hybrid groups exist in IBD-isolated *E. coli* (Figure 10). Taken together, the MLST analysis and ABD typing support the conclusions of Wirth *et al*.

In this study, we illustrated that 26 different STs exist among the total of 45 *E. coli* isolates (Table 9). This shows that *E. coli* isolated from IBD patients and healthy

individuals come from very broad ancestral backgrounds. The possible explanation for this wide range is that the composition of normal flora is highly individual (Eckburg et al. 2007) and different people harbor different strain of *E. coli* in their digestive tract. However, the fact that 10 of the 11 STs that were reported for the first time were isolated from IBD tissues (Table 9) is another indication of more frequent mutation and recombination events occurring in these strains of *E. coli* (Figure 11).

2.3. Virulence characterization of the *E. coli*

Although AIEC appear to be consistently associated with biopsy tissues in a subset of IBD patients, it is not known whether these strains are the cause of the disease or simply the consequence of the disease process. If AIEC are involved as a primary etiologic cause of IBD then they should be present in biopsy tissues at the initiation of the disease. It is not possible to obtain tissue at the disease onset, but we were able to gain access to tissues that were collected from IBD patients at the time of first diagnosis. Another advantage of these tissues was that they were collected from across Canada, which excludes the epidemiological biases of studying tissues obtained from only one centre.

Our study is the first to investigate the presence of putative AIEC strains at first diagnosis in adults. The only other study that addresses the presence of AIEC in early disease is the pediatric study of Schippa (Schippa et al. 2008) who observed a significant association between AIEC and disease. We found that *E. coli* were ubiquitous in the biopsy tissues and were present in 11/25 healthy controls (44%), 14/23 CD patients (60%), and 6/11 UC patients (54%). Schippa et al. (Schippa et al. 2008), isolated 60

mucosal-associated *E. coli* from 28 CD, 19 UC, and 13 control biopsies from pediatric subjects. When the presence of *E. coli* in each group was compared, *E. coli* were detected in 10/12 CD patients, 7/7 UC patients, and 10/19 controls. The authors (Schippa et al. 2008) conducted pulsed field gel electrophoresis on the isolates and concluded that particular genotypes were associated with the disease. They (Schippa et al. 2008) reported that there was no significant association between phylotypes B2 or D and disease as previously described (Kotlowski et al. 2007), although B2 was only isolated from diseased tissues and not controls. We did however find a statistically significant (P = 0.04) association between B2+D and inflammation (Table 14). The difference may lie in the fact that we had a larger sample size than Schippa et al. (Schippa et al. 2008).

We also assessed these *E. coli* for the presence of virulence genes (Table 12) that had been previously related with IBD (Kotlowski et al. 2007). Afimbrial adhesins (Afa) were initially associated with diffusely adhering *E. coli* (DAEC) and act as virulence factors (Servin 2005). The *afa* genes are organized into a polycystronic transcriptional unit made up of transcriptional regulators (AfaA and F), periplasmic chaperonins (AfaB), anchoring proteins (AfaC), invasins (AfaD), and adhesins (AfaE) (Servin 2005). We chose to study the *afaE* gene because we hypothesized that interaction with the target receptor on epithelial cells would be the most informative and because of the demonstrated interactions between Afa and CEACAM (Servin 2005).

We found a significant association between *afaE* and IBD (P = 0.01) particularly between *afaE* and CD (P = 0.005) (Table 14). The pediatric study of Schippa (Schippa et al. 2008) did not detect any *afaB* and *C* in their *E. coli* isolates. Martinez-Medina (Martinez-Medina et al. 2009) also looked for *afaB* and *C* in AIEC from CD and controls

and found 19 % vs 0 % respectively. Previously, our group did not find a significant association between *afaE* and IBD (Kotlowski et al. 2007). However, both Martinez-Medina et al., (Martinez-Medina et al. 2009) and Kotlowski et al. (Kotlowski et al. 2007) found *E. coli* positive for *afa* only in IBD patients but never in healthy controls. The absence of *afa* in healthy controls is consistent with the current study.

Serine protease autotransporters (SPATE) are peptides that are potential virulence factors and appear to target intracellular proteins like defensins and elastases (Henderson et al. 2004). Our group observed a significant association between *E. coli* from IBD biopsy tissues and SPATE (Kotlowski et al. 2007). In the current study we detected a highly significant association (P = 0.00001) between SPATE and *E. coli* that were B2 or D (Table 14). A similar observation was made with histolytic Boxer dog granulomatous colitis, pic (a SPATE), and AIEC (Simpson et al. 2006).

Ag43 is a member of the family of cell surface adhesins involved in diffuse adherent (AIDA) and mediates autoaggregation in bacteria (Henderson et al. 2004). Recent studies have demonstrated that Ag43 contributes to *E. coli* biofilm formation by mediating inter- and intraspecies cell aggregation (Danese et al. 2000; Kjaergaard et al. 2000a; Kjaergaard et al. 2000b). Our study demonstrated a significant association between B2 and *agn43* (P = 0.05) (Table 14). The *sfa* gene encodes a class of S-fimbrial adhesins which bind to sialic acid-specific receptors on uroepithelial cells (Morschhauser et al. 1993; Hacker et al. 2000). We found a significant association between B2 and *sfa* (P = 0.04) (Table 14).

The ubiquitous distribution of AIDA-I in *E. coli* isolated from CD and UC as well as HC may be explained by the fact that posttranslational processing of AIDA-I is

essential for its diffuse adherence pattern (Benz et al. 2001). Perhaps this posttranslational step is the one affected by selective pressure of diseased vs healthy environment, resulting in diffuse adherence of *E. coli* to gut tissue.

2.4. Analysis of *fimH* subunit of *E. coli* fimbria gene

Although commensal *E. coli* are normal inhabitants of the human gut, pathogenic *E. coli* can cause serious intestinal or extra-intestinal infections (Barnich et al. 2007a). Type 1 fimbriae, which are responsible for bacterial adherence to the host epithelial surfaces, are virulence factors of *E. coli* (Boudeau et al. 2001; Barnich et al. 2007a). FimH, the adhesive subunit of type 1 pili, is located at the tip of the fimbria and thought to mediate *E. coli* colonization (Boudeau et al. 2001; Weissman et al. 2006). Barnich et al. showed that AIEC adhesion is dependent on type 1 pili expression on the bacterial surface and on CEACAM6 expression on the apical surface of ileal epithelial cells (Barnich et al. 2007a). It has also been reported that CEACAM6 acts as a receptor for AIEC adhesion and is abnormally expressed by ileal epithelial cells in CD patients (Barnich et al. 2007a; Carvalho et al. 2009). Although other studies (Schippa et al. 2008; Martinez-Medina et al. 2009) detected *fimH* in their *E. coli* isolates they did not sequence the genes thus were not able to associate the gene structure with the adherence and invasive phenotype.

We sequenced and aligned *fimH* sequences of the 35 *E. coli* isolates from patients with established IBD (Figure 12), 65 *E. coli* isolated from newly diagnosed IBD patients (Figure 13) and 10 strains of *E. coli* as controls. Phylogenetic analysis of *fimH* sequences of patients with established IBD (Figure 12) delineated a tight cluster containing LF82,

the best described AIEC, as well as IBD isolates which we had previously demonstrated to have a number of virulence genes (Kotlowski et al. 2007). However, strains CFT073 and APEC O1:K1:H7 were also in this cluster, and neither of these pathogenic strains has been previously linked to IBD. An important observation is that this cluster does not contain genes from enteric pathogens like *E. coli* O157:H7 or from non-pathogenic *E. coli* K12, probiotic Nissle 1917, or any of *E. coli* isolated from swine or cattle (Figure 12).

This raises the possibility that IBD isolated *E. coli* are members of a general pool of extra-intestinal pathogenic *E. coli* which reside in the gut and have evolved specific potentialities based on the microenvironment within which they find themselves. Bacteria that reside in human gut have the capacity to activate protective mucosal immune responses (tolerance, clearance, etc.) in normal hosts. In at least a subset of IBD patients, the mucosal immunity is dysfunctional (e.g. NOD2 mutation) and is unable to clear microbial pathogens. In these genetically predisposed hosts, the mucosal immune response to the putative enteric pathogens could be detrimental and result in the induction of chronic intestinal inflammation (Sartor 2001). *E. coli* associated to IBD would then by extension be an organism that has accumulated genes and mutations to take advantage of this specific microniche. This hypothesis is supported by the observation that considerable intra-species genomic size variation of *E. coli* is found in different strains (Ihssen et al. 2007). Addition or deletion of genome segments likely represent the accumulation, or loss, of genes encode for an adapted phenotype. Future research should focus on virulence factors that are unique for IBD-associated *E. coli* so that a better understanding of the mechanism of action of this group of bacteria can be achieved.

Very similar results were obtained when sequences of *fimH* from *E. coli* isolated from newly diagnosed IBD patients were studied (Figure 13). The fact that these changes were present in *E. coli* isolated from both inflamed and non-inflamed gut biopsies suggests that the virulence shift in *fimH* is present since the onset of the disease and is not the consequence of microbial ecological changes that occur as a result of the inflammatory process or the IBD treatment.

It has also been shown that amino acid substitutions in FimH can change *E. coli* tropism toward epithelial cells (Weissman et al. 2006). In our study, we determined that amino acid substitutions V48A, N91S, S99N, and V140A were significantly associated with established IBD ($P < 0.05$) (Table 15) whereas V48A, N91S, S99N, and A223V were significantly related with newly diagnosed IBD ($P < 0.05$) (Table 16). Some of these amino acid substitutions have also been reported previously as important in AIEC (Boudeau et al. 2001), and may facilitate the interaction of AIEC with CEACAM 6 in IBD patients (Barnich et al. 2007a). The role of FimH amino acid substitutions in the pathogenesis of *E. coli* has been also confirmed previously by mutation analysis (Boudeau et al. 2001). Although it is believed that these amino acid substitutions change *E. coli* tropism towards gut epithelium, the exact mechanism is unknown and could serve as a topic for future investigations.

2.5. *E. coli* gene content analysis

After the completion of whole genomes of the intestinal and extraintestinal pathogens, the search for potential virulence genes became more feasible. By comparing the pathogenic genome with the genome of non-pathogenic *E. coli* K12 (Afset et al.

2006), many regions unique to pathogenic strains were identified. Although various molecular methods can be employed to identify the virulence determinants, microarrays proved to be powerful tools for the simultaneous detection of high numbers of genes within a given *E. coli* strain (Afset et al. 2006; Bruant et al. 2006). DNA microarrays have been used successfully in various studies involving taxonomy (Cho et al. 2001), microbial genotyping (Dougherty et al. 2002; Grimm et al. 2004), and detection of *E. coli* antimicrobial resistance (Call et al. 2003; Yu et al. 2004; Bruant et al. 2006) and virulence genes (Chizhikov et al. 2001; Bekal et al. 2003; Bruant et al. 2006). Moreover, the direct search for virulence determinants is obviously limited to known factors.

The goal of our last experiment was to search for potential virulence genes in *E. coli* that are linked to IBD in general, or specifically to CD or UC. To achieve this goal, we studied the genetic content of *E. coli* isolated from IBD patients (CD = 6, UC = 5) and healthy controls (HC = 5) using Affymetrix GeneChip *E. coli* Genome 2.0 array. This arrays includes probes for all 20,366 genes present in non-pathogenic *E. coli* K12 MG1655, and pathogenic UPEC CFT073, EHEC O157:H7 EDL933 and EHEC O157:H7 Sakai. The probes cover all open reading frames, including 700 intergenic regions.

Bacterial genomes generally consist of stable regions called a "core gene pool" and variable regions called a "flexible gene pool" (Hacker et al. 2003). The core gene pool comprises of genes that are part of the bacterial chromosome and encode products that are critical for the basic cellular function (Hacker et al. 2001). Elements of the flexible gene pool on the other hand, often have features of mobile genetic elements and encode additional functions that are not essential for the bacterium but provide advantages under particular conditions (Hacker et al. 2001). The fact that all *E. coli*

strains in this study shared 58% of *E. coli* K12 MG1655 genes indicates that this non-pathogenic strain of *E. coli* provides the majority of the core genetic pool of this organism (Table 17). Likewise, 333 of the genes in *E. coli* K12 MG1655 were significantly associated with the *E. coli* from HC. This highlights the role of this strain in maintaining a healthy gut ecosystem (Table 18).

One interesting finding in our study was the significant association of 876 UPEC CFT073-specific genes to IBD (Table 18), which accounted for 35.2% of total UPEC CFT073-specific genes (Table 18) and 91% of all the genes that were significantly associated to IBD (Figure 14). This was in contrast to the findings regarding EHEC strains. Only 3.3 and 3.5% of the *E. coli* O157:H7 EDL933- and *E. coli* O157:H7 Sakai-specific genes were associated to IBD respectively which in total accounted for only 7% of all IBD-associated genes (Table 18) (Figure 14).

According to the fecal-vaginal-urethral hypothesis, *E. coli* strains causing urinary tract infection are usually derived from the host's own fecal flora (Ewers et al. 2009). These are the organisms that reside in the human gut without causing any harm; however, they have developed specific virulence features that make them pathogenic to the urinary system. Type 1 fimbriae of UPEC attach to the mannose component of the urinary tract epithelial receptors and initiate colonization and biofilm formation in the bladder. The P fimbriae on the other hand have receptors in the kidney and are responsible for pyelonephritis (Kaper et al. 2004). Acute pyelonephritis is known to be initiated by the dominance of uropathogenic strains in fecal flora. Although recurrent infections might occasionally be due to a persistent infection, the majority have been thought to be

reinfections caused by the initially infecting strain persisting in the fecal flora (Servin 2005).

It is likely that *E. coli* isolated from IBD biopsies originally belonged to the subset of *E. coli* that resides in human gut but causes infection in urinary system. Although these UPEC strains do not disturb the gut in a healthy individual, they could result in gastrointestinal inflammation in a predisposed individual with dysfunctional immune system. In other words, the same UPEC strain that resides in a healthy gut can result in chronic inflammation and IBD in an individual who is genetically susceptible.

The high association between genes specific for UPEC and IBD could also be explained with the phenomenon of horizontal gene transfer. A study by Lawrence et al. on *E. coli* K12 MG1655 based on its whole genome sequence revealed that about 18% of the total *E. coli* genetic content belongs to the flexible gene pool (Lawrence et al. 1998). This reflects the elasticity of *E. coli* genome compared to other microorganisms such as *Mycoplasma* which has a flexible gene pool of <1% (Hacker et al. 2001). *E. coli*, like many other organisms, has the ability of losing or acquiring genetic content in order to fit into its environment. In this context, either UPEC inhabitants of the gut gained some virulence features which led them to cause gut inflammation, or other commensal strains of *E. coli* acquired some uropathogenic virulence features under the influence of the "IBD microenvironment".

Of the known *E. coli* virulence factors, only a few were significantly associated with IBD. The genes responsible for capsule synthesis in UPEC CFT073 were only present in IBD isolates and the association was significant for *kpsC*, *kpsD*, *kpsE*, and *kpsS* (P < 0.05) (Table 19). Similarly, UPEC CFT073 genes involved in F1C fimbrial

assembly, *focC*, *focD*, and *focF* were highly related with UC (P < 0.05) (Table 19). Also, *sfaB*, *sfaC*, and *sfaD* genes involved in S fimbrial synthesis of UPEC CFT073 were only detected in IBD isolates, but the relationship was not strong enough to be significant (P < 0.08) (Table 19). On the contrary, well-known EHEC O157:H7 genes were either absent in all *E. coli* strains, or were present in only a few strains without any significant association to IBD (Table 20). Such lack of association between EHEC O157:H7-specific virulence factors and IBD is noteworthy since, in the context of this experiment, it rules out this enteric pathogen as a causative agent of IBD.

We previously showed that probiotic *E. coli* Nissle 1917 and UPEC CFT073 have identical housekeeping gene sequences. While UPEC CFT073 is a pathogen causing bladder and kidney infections, *E. coli* Nissle 1917 is a probiotic used to change the balance of gut microbiota towards beneficial microbes. In order to discover the differences between *E. coli* Nissle 1917 and UPEC CFT073, we compared the genetic content of these two strains using Affymetrix GeneChip *E. coli* Genome 2.0 array. Analysis of array data revealed a 10% difference in genetic content of these two strains. Lack of many virulence genes in *E. coli* Nissle 1917 was responsible for a significant proportion of the differences (Table 23). These virulence factors include pic and serine proteases, Ag43, hemolysin, flagellin, type I fimbriae, F1C fimbriae, P fimbriae, outer membrane proteins A and C and capsule proteins (Table 23). In fact, these virulence features may be essential in the pathogenesis of IBD given that their lack allows an *E. coli* strain to be a probiotic.

To our knowledge, this is the first study in which the genomic content of *E. coli* isolated from IBD tissues has been evaluated using DNA microarray technology. Our

study revealed a remarkable association between the genetic content of *E. coli* isolated from IBD and UPEC CFT073. Furthermore, adhesins such as UPEC-specific F1C fimbriae and S fimbriae were exclusively detected in IBD isolates. Uropathogenic specific virulence factors, particularly the ones that are not present in *E. coli* Nissle 1917, may hold the key to our understanding of IBD pathogenesis. Investigating the incidence of urinary tract infections in IBD patients and characterizating the microorganisms involved are also important and could provide clues regarding the role of these bacteria in the pathophysiology of IBD.

CONCLUSION

A substantial body of evidence supports the role of microorganisms in the initiation or reactivation of IBD (Sokol et al. 2008a). In this study we applied ARISA and T-RFLP to explore the microbial composition of IBD gut biopsies and healthy controls. Remarkable features of our study were i) having access to a population-based case-control IBD tissue bank, and ii) utilizing non-inflamed biopsies of IBD patients in addition to inflamed tissues.

Our study has led to an interesting observation of higher species richness and diversity in non-inflamed IBD tissues. An increase from healthy to non-inflamed IBD tissue could indicate the formation of an intermediary community of organisms in transition to the inflamed state. Based on these findings, we hypothesize that microorganisms present in the non-inflamed or pre-inflamed sites of IBD intestinal tissue may hold clues to the pathophysiology of IBD. To further evaluate this hypothesis in future, we propose an experiment in which a gradient of three to five colon biopsies from non-inflamed to inflamed sites will be taken from IBD patients enrolled in the population–based case-control tissue bank. Changes in microbial composition and spatial organization of gut flora through this gradient could be tested using real-time PCR, pyrosequencing, and fluorescent *in situ* hybridization (FISH). We believe such studies eliminate the bias inherent in interindividual differences in gut microflora, and may reveal pathophysiological processes that occur before the establishment of persistent inflammation.

We also studied the particular role of *E. coli* in the pathogenesis of IBD. Although the results of our T-RFLP and ARISA analysis showed an increasing trend in the population of *Proteobacteria* in both CD and UC biopsies, these methods were not sensitive enough to assess specific changes in *E. coli* populations. Still, several investigators have demonstrated an increase in the *E. coli* population in IBD tissues. For that reason, we chose to pursue *E. coli* as a potential etiological agent in IBD. By means of cultivation, we were able to isolate 3-4 logs more *E. coli* from IBD biopsy tissues compared to controls. Our results showed that these *E. coli* have virulence factors that are associated with IBD. We have also demonstrated an association between *E. coli* clonal groups and IBD using phylogenetic techniques. Moreover, by using genome-wide Affymetrix arrays, we demonstrated that approximately 10% of the *E. coli* genes on the array were significantly related with IBD. Based on these findings, we suggest that *E. coli* isolated with "IBD-genotypes" are not members of the normal commensal gut microbiota. However, the question remained whether the increase in certain types of *E. coli* in IBD is the consequence of inflammation or the cause.

We hypothesized that to be involved in IBD as a primary etiologic cause, *E. coli* should be present in the biopsy tissue at the initiation of the disease. Although it was not possible to obtain tissues at disease onset, we were able to gain access to biopsies that were collected from IBD patients at first diagnosis. We studied *E. coli* isolated from these newly diagnosed patients and showed that their virulence and phylogenetic characteristics are very similar to *E. coli* that had been isolated from patients with established IBD, emphasizing the role of *E. coli* at disease onset.

Our results are also in agreement with previous studies and highlight the potential role of *E. coli* adhesins in the pathophysiology of IBD. Not only were we able to associate factors such as AfaE and SPATE to IBD, we were also able to correlate two amino acid substitutions in the FimH adherent subunit of type I fimbriae with the disease. This is a notable observation since CEACAMs act as epithelial receptors for both AfaE and FimH adhesive subunits.

Although *E. coli* have been associated to IBD on a constant basis, the focus has been on AIEC. This is due to observations that primary CD lesions often occur in Peyer's patches. The fact that such lesions also occur in shigellosis, salmonellosis, and yersinial enterocolitis, in which invasiveness is an essential virulence factor, has led to identification of AIEC pathotype in a subset of CD patients. However, we do not necessarily have to look for invasive phenotype of *E. coli* in all IBD patients since these granulomatous lesions only occur in a subset of patients.

To our knowledge our study is the first that demonstrates IBD-isolated *E. coli* and UPEC CFT073 share considerable genomic content and virulence features. Based on MLST phylogenetic analysis, we also demonstrated a link between APEC H1:K1:O7, UPEC CFT073 and *E. coli* from IBD patients. This is a remarkable finding, considering the fact that we were unable to detect a resemblance between IBD-associated *E. coli* and EHEC O157:H7. Hence, we suggest that it might be erroneous to limit the study of *E. coli* in IBD only to AIEC pathotypes. One way forward would be to screen IBD-associated *E. coli* for virulence factors found in known pathotypes of *E. coli*, particularly ExPEC, with hightroughput techniques such as virulence arrays. Of most interest are virulence features that are involved in adherence to epithelial cells.

REFERENCES

1. **Afset, J. E., G. Bruant, R. Brousseau, J. Harel, E. Anderssen, L. Bevanger and K. Bergh** (**2006**). "Identification of virulence genes linked with diarrhea due to atypical enteropathogenic *Escherichia coli* by DNA microarray analysis and PCR." *J Clin Microbiol* 44(10): 3703-11.

2. **Ahmadi, A., S. Polyak and P. V. Draganov** (**2009**). "Colorectal cancer surveillance in inflammatory bowel disease: the search continues." *World J Gastroenterol* 15(1): 61-6.

3. **Ali, S. and C. P. Tamboli** (**2008**). "Advances in epidemiology and diagnosis of inflammatory bowel diseases." *Curr Gastroenterol Rep* 10(6): 576-84.

4. **Asakura, H., K. Suzuki, T. Kitahora and T. Morizane** (**2008**). "Is there a link between food and intestinal microbes and the occurrence of Crohn's disease and ulcerative colitis?" *J Gastroenterol Hepatol* 23(12): 1794-801.

5. **Barnich, N., F. A. Carvalho, A. L. Glasser, C. Darcha, P. Jantscheff, M. Allez, H. Peeters, G. Bommelaer, P. Desreumaux, J. F. Colombel and A. Darfeuille-Michaud** (**2007a**). "CEACAM6 acts as a receptor for adherent-invasive *E. coli*, supporting ileal mucosa colonization in Crohn disease." *J Clin Invest* 117(6): 1566-74.

6. **Barnich, N. and A. Darfeuille-Michaud** (**2007b**). "Role of bacteria in the etiopathogenesis of inflammatory bowel disease." *World J Gastroenterol* 13(42): 5571-6.

7. **Baumgart, M., B. Dogan, M. Rishniw, G. Weitzman, B. Bosworth, R. Yantiss, R. H. Orsi, M. Wiedmann, P. McDonough, S. G. Kim, D. Berg, Y. Schukken, E. Scherl and K. W. Simpson** (**2007**). "Culture independent analysis of ileal mucosa reveals a selective increase in invasive *Escherichia coli* of novel phylogeny relative to depletion of *Clostridiales* in Crohn's disease involving the ileum." *ISME J* 1(5): 403-18.

8. **Bekal, S., R. Brousseau, L. Masson, G. Prefontaine, J. Fairbrother and J. Harel** (**2003**). "Rapid identification of Escherichia coli pathotypes by virulence gene detection with DNA microarrays." *J Clin Microbiol* 41(5): 2113-25.

9. **Benz, I. and M. A. Schmidt** (**2001**). "Glycosylation with heptose residues mediated by the aah gene product is essential for adherence of the AIDA-I adhesin." *Mol Microbiol* 40(6): 1403-13.

10. **Bernstein, C. N., G. Nayar, A. Hamel and J. F. Blanchard** (**2003**). "Study of animal-borne infections in the mucosas of patients with inflammatory bowel disease and population-based controls." *J Clin Microbiol* 41(11): 4986-90.

11. **Bernstein, C. N. and F. Shanahan** (**2008**). "Disorders of a modern lifestyle: reconciling the epidemiology of inflammatory bowel diseases." *Gut* 57(9): 1185-91.

12. **Bernstein, C. N., A. Wajda, L. W. Svenson, A. MacKenzie, M. Koehoorn, M. Jackson, R. Fedorak, D. Israel and J. F. Blanchard** (**2006**). "The epidemiology of inflammatory bowel disease in Canada: a population-based study." *Am J Gastroenterol* 101(7): 1559-68.

13. **Bibiloni, R., M. Mangold, K. L. Madsen, R. N. Fedorak and G. W. Tannock** (**2006**). "The bacteriology of biopsies differs between newly diagnosed, untreated, Crohn's disease and ulcerative colitis patients." *J Med Microbiol* 55(Pt 8): 1141-9.

14. **Booijink, C. C., E. G. Zoetendal, M. Kleerebezem and W. M. de Vos** (**2007**). "Microbial communities in the human small intestine: coupling diversity to metagenomics." *Future Microbiol* 2: 285-95.

15. **Boudeau, J., N. Barnich and A. Darfeuille-Michaud** (**2001**). "Type 1 pili-mediated adherence of *Escherichia coli* strain LF82 isolated from Crohn's disease is involved in bacterial invasion of intestinal epithelial cells." *Mol Microbiol* 39(5): 1272-84.

16. **Boyd, E. F. and D. L. Hartl** (**1998**). "Chromosomal regions specific to pathogenic isolates of *Escherichia coli* have a phylogenetically clustered distribution." *J Bacteriol* 180(5): 1159-65.

17. **Brant, S. R.** (**2009**). "Exposed: the genetic underpinnings of ulcerative colitis relative to Crohn's disease." *Gastroenterology* 136(2): 396-9.

18. **Brousseau, R., J. E. Hill, G. Prefontaine, S. H. Goh, J. Harel and S. M. Hemmingsen** (**2001**). "*Streptococcus suis* serotypes characterized by analysis of chaperonin 60 gene sequences." *Appl Environ Microbiol* 67(10): 4828-33.

19. **Brown, M. V., M. S. Schwalbach, I. Hewson and J. A. Fuhrman** (**2005**). "Coupling 16S-ITS rDNA clone libraries and automated ribosomal intergenic spacer analysis to show marine microbial diversity: development and application to a time series." *Environ Microbiol* 7(9): 1466-79.

20. **Bruant, G., C. Maynard, S. Bekal, I. Gaucher, L. Masson, R. Brousseau and J. Harel** (**2006**). "Development and validation of an oligonucleotide microarray for

detection of multiple virulence and antimicrobial resistance genes in *Escherichia coli*." *Appl Environ Microbiol* 72(5): 3780-4.

21. **Call, D. R., M. K. Bakko, M. J. Krug and M. C. Roberts** (**2003**). "Identifying antimicrobial resistance genes with DNA microarrays." *Antimicrob Agents Chemother* 47(10): 3290-5.

22. **Cardinale, M., L. Brusetti, P. Quatrini, S. Borin, A. M. Puglia, A. Rizzi, E. Zanardini, C. Sorlini, C. Corselli and D. Daffonchio** (**2004**). "Comparison of different primer sets for use in automated ribosomal intergenic spacer analysis of complex bacterial communities." *Appl Environ Microbiol* 70(10): 6147-56.

23. **Carvalho, F. A., N. Barnich, A. Sivignon, C. Darcha, C. H. Chan, C. P. Stanners and A. Darfeuille-Michaud** (**2009**). "Crohn's disease adherent-invasive *Escherichia coli* colonize and induce strong gut inflammation in transgenic mice expressing human CEACAM." *J Exp Med* 206(10): 2179-89.

24. **Chao, A., R. L. Chazdon, R. K. Colwell and T. Shen** (**2005**). "A new statistical approach for assessing similarity of species composition with incidence and abundance data." *Ecology Letters* 8(2): 148-159.

25. Chao, A. and T. Shen. (2003, 2009). "User's Guide for Program SPADE (Species Prediction And Diversity Estimation)." Retrieved 02/08, 2009, from *http://chao.stat.nthu.edu.tw/softwareCE.html*.

26. **Chenna, R., H. Sugawara, T. Koike, R. Lopez, T. J. Gibson, D. G. Higgins and J. D. Thompson** (**2003**). "Multiple sequence alignment with the Clustal series of programs." *Nucleic Acids Res* 31(13): 3497-500.

27. **Chiodini, R. J. and C. A. Rossiter** (**1996**). "Paratuberculosis: a potential zoonosis?" *Vet Clin North Am Food Anim Pract* 12(2): 457-67.

28. **Chizhikov, V., A. Rasooly, K. Chumakov and D. D. Levy** (**2001**). "Microarray analysis of microbial virulence factors." *Appl Environ Microbiol* 67(7): 3258-63.

29. **Cho, J. C. and J. M. Tiedje** (**2001**). "Bacterial species determination from DNA-DNA hybridization by using genome fragments and DNA microarrays." *Appl Environ Microbiol* 67(8): 3677-82.

30. **Clermont, O., S. Bonacorsi and E. Bingen** (**2000**). "Rapid and simple determination of the *Escherichia coli* phylogenetic group." *Appl Environ Microbiol* 66(10): 4555-8.

31. **Cole, J. R., B. Chai, R. J. Farris, Q. Wang, S. A. Kulam, D. M. McGarrell, G. M. Garrity and J. M. Tiedje** (**2005**). "The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis." *Nucleic Acids Res* 33(Database issue): D294-6.

32. **Cole, J. R., Q. Wang, E. Cardenas, J. Fish, B. Chai, R. J. Farris, A. S. Kulam-Syed-Mohideen, D. M. McGarrell, T. Marsh, G. M. Garrity and J. M. Tiedje** (**2009**). "The Ribosomal Database Project: improved alignments and new tools for rRNA analysis." *Nucleic Acids Res* 37(Database issue): D141-5.

33. Colwell, R. K. (2005). "EstimateS: statistical estimation of species richness and shared species from samples. Version 7.5. User's Guide and application." from *http://purl.oclc.org/estimates*.

34. **Conte, M. P., S. Schippa, I. Zamboni, M. Penta, F. Chiarini, L. Seganti, J. Osborn, P. Falconieri, O. Borrelli and S. Cucchiara** (**2006**). "Gut-associated bacterial microbiota in paediatric patients with inflammatory bowel disease." *Gut* 55(12): 1760-7.

35. **Curtis, T. P. and W. T. Sloan** (**2004**). "Prokaryotic diversity and its limits: microbial community structure in nature and implications for microbial ecology." *Curr Opin Microbiol* 7(3): 221-6.

36. **Danese, P. N., L. A. Pratt, S. L. Dove and R. Kolter** (**2000**). "The outer membrane protein, antigen 43, mediates cell-to-cell interactions within *Escherichia coli* biofilms." *Mol Microbiol* 37(2): 424-32.

37. **Danese, S., M. Sans and C. Fiocchi** (**2004**). "Inflammatory bowel disease: the role of environmental factors." *Autoimmun Rev* 3(5): 394-400.

38. **Danovaro, R., G. M. Luna, A. Dell'anno and B. Pietrangeli** (**2006**). "Comparison of two fingerprinting techniques, terminal restriction fragment length polymorphism and automated ribosomal intergenic spacer analysis, for determination of bacterial diversity in aquatic environments." *Appl Environ Microbiol* 72(9): 5982-9.

39. **Darfeuille-Michaud, A.** (**2002**). "Adherent-invasive *Escherichia coli*: a putative new E. coli pathotype associated with Crohn's disease." *Int J Med Microbiol* 292(3-4): 185-93.

40. **Darfeuille-Michaud, A., J. Boudeau, P. Bulois, C. Neut, A. L. Glasser, N. Barnich, M. A. Bringer, A. Swidsinski, L. Beaugerie and J. F. Colombel** (**2004**). "High prevalence of adherent-invasive *Escherichia coli* associated with ileal mucosa in Crohn's disease." *Gastroenterology* 127(2): 412-21.

41.   **Darfeuille-Michaud, A., C. Neut, N. Barnich, E. Lederman, P. Di Martino, P. Desreumaux, L. Gambiez, B. Joly, A. Cortot and J. F. Colombel** (**1998**). "Presence of adherent *Escherichia coli* strains in ileal mucosa of patients with Crohn's disease." *Gastroenterology* 115(6): 1405-13.

42.   **Denman, S. E. and C. S. McSweeney** (**2006**). "Development of a real-time PCR assay for monitoring anaerobic fungal and cellulolytic bacterial populations within the rumen." *FEMS Microbiology Ecology* 58(3): 572-582.

43.   **Desjardins, P., B. Picard, B. Kaltenbock, J. Elion and E. Denamur** (**1995**). "Sex in *Escherichia coli* does not disrupt the clonal structure of the population: evidence from random amplified polymorphic DNA and restriction-fragment-length polymorphism." *J Mol Evol* 41(4): 440-8.

44.   **Dho-Moulin, M. and J. M. Fairbrother** (**1999**). "Avian pathogenic *Escherichia coli* (APEC)." *Vet Res* 30(2-3): 299-316.

45.   **Dicksved, J., J. Halfvarson, M. Rosenquist, G. Jarnerot, C. Tysk, J. Apajalahti, L. Engstrand and J. K. Jansson** (**2008**). "Molecular analysis of the gut microbiota of identical twins with Crohn's disease." *ISME J* 2(7): 716-27.

46.   **Dorigo, U., L. Volatier and J. F. Humbert** (**2005**). "Molecular approaches to the assessment of biodiversity in aquatic microbial communities." *Water Res* 39(11): 2207-18.

47.   **Dotan, I.** (**2007**). "Serologic markers in inflammatory bowel disease: tools for better diagnosis and disease stratification." *Expert Rev Gastroenterol Hepatol* 1(2): 265-74.

48.   **Dougherty, J. D. and D. H. Geschwind** (**2002**). "Subtraction-coupled custom microarray analysis for gene discovery and gene expression studies in the CNS." *Chem Senses* 27(3): 293-8.

49.   **Duriez, P., O. Clermont, S. Bonacorsi, E. Bingen, A. Chaventre, J. Elion, B. Picard and E. Denamur** (**2001**). "Commensal *Escherichia coli* isolates are phylogenetically distributed among geographically distinct human populations." *Microbiology* 147(Pt 6): 1671-6.

50.   **Eckburg, P. B., E. M. Bik, C. N. Bernstein, E. Purdom, L. Dethlefsen, M. Sargent, S. R. Gill, K. E. Nelson and D. A. Relman** (**2005**). "Diversity of the human intestinal microbial flora." *Science* 308(5728): 1635-8.

51.   **Eckburg, P. B. and D. A. Relman** (**2007**). "The role of microbes in Crohn's disease." *Clin Infect Dis* 44(2): 256-62.

52. **Economou, M. and G. Pappas** (**2008**). "New global map of Crohn's disease: Genetic, environmental, and socioeconomic correlations." *Inflamm Bowel Dis* 14(5): 709-20.

53. **Escobar-Paramo, P., O. Clermont, A. B. Blanc-Potard, H. Bui, C. Le Bouguenec and E. Denamur** (**2004a**). "A specific genetic background is required for acquisition and expression of virulence factors in *Escherichia coli*." *Mol Biol Evol* 21(6): 1085-94.

54. **Escobar-Paramo, P., K. Grenet, A. Le Menac'h, L. Rode, E. Salgado, C. Amorin, S. Gouriou, B. Picard, M. C. Rahimy, A. Andremont, E. Denamur and R. Ruimy** (**2004b**). "Large-scale population structure of human commensal *Escherichia coli* isolates." *Appl Environ Microbiol* 70(9): 5698-700.

55. **Ewers, C., E. M. Antao, I. Diehl, H. C. Philipp and L. H. Wieler** (**2009**). "Intestine and environment of the chicken as reservoirs for extraintestinal pathogenic *Escherichia coli* strains with zoonotic potential." *Appl Environ Microbiol* 75(1): 184-92.

56. **Ewers, C., G. Li, H. Wilking, S. Kiessling, K. Alt, E. M. Antao, C. Laturnus, I. Diehl, S. Glodde, T. Homeier, U. Bohnke, H. Steinruck, H. C. Philipp and L. H. Wieler** (**2007**). "Avian pathogenic, uropathogenic, and newborn meningitis-causing *Escherichia coli*: how closely related are they?" *Int J Med Microbiol* 297(3): 163-76.

57. **Farmer, R. G., W. M. Michener and E. A. Mortimer** (**1980**). "Studies of family history among patients with inflammatory bowel disease." *Clin Gastroenterol* 9(2): 271-7.

58. **Feil, E. J., B. C. Li, D. M. Aanensen, W. P. Hanage and B. G. Spratt** (**2004**). "eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data." *J Bacteriol* 186(5): 1518-30.

59. **Feller, M., K. Huwiler, R. Stephan, E. Altpeter, A. Shang, H. Furrer, G. E. Pfyffer, T. Jemmi, A. Baumgartner and M. Egger** (**2007**). "*Mycobacterium avium subspecies paratuberculosis* and Crohn's disease: a systematic review and meta-analysis." *Lancet Infect Dis* 7(9): 607-13.

60. **Fisher, M. M. and E. W. Triplett** (**1999**). "Automated approach for ribosomal intergenic spacer analysis of microbial diversity and its application to freshwater bacterial communities." *Appl Environ Microbiol* 65(10): 4630-6.

61. **Frank, D. N. and N. R. Pace** (**2001**). "Molecular-phylogenetic analyses of human gastrointestinal microbiota." *Curr Opin Gastroenterol* 17(1): 52-7.

62. **Frank, D. N., A. L. St Amand, R. A. Feldman, E. C. Boedeker, N. Harpaz and N. R. Pace** (**2007**). "Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases." *Proc Natl Acad Sci U S A* 104(34): 13780-5.

63. **Fujita, H., Y. Eishi, I. Ishige, K. Saitoh, T. Takizawa, T. Arima and M. Koike** (**2002**). "Quantitative analysis of bacterial DNA from *Mycobacteria* spp., *Bacteroides vulgatus*, and *Escherichia coli* in tissue samples from patients with inflammatory bowel diseases." *J Gastroenterol* 37(7): 509-16.

64. **Gal-Mor, O. and B. B. Finlay** (**2006**). "Pathogenicity islands: a molecular toolbox for bacterial virulence." *Cell Microbiol* 8(11): 1707-19.

65. **Gevers, D., F. M. Cohan, J. G. Lawrence, B. G. Spratt, T. Coenye, E. J. Feil, E. Stackebrandt, Y. Van de Peer, P. Vandamme, F. L. Thompson and J. Swings** (**2005**). "Opinion: Re-evaluating prokaryotic species." *Nat Rev Microbiol* 3(9): 733-9.

66. **Gismera, C. S. and B. S. Aladren** (**2008**). "Inflammatory bowel diseases: a disease (s) of modern times? Is incidence still increasing?" *World J Gastroenterol* 14(36): 5491-8.

67. **Goh, K. and S. D. Xiao** (**2009**). "Inflammatory bowel disease: a survey of the epidemiology in Asia." *J Dig Dis* 10(1): 1-6.

68. **Gophna, U., K. Sommerfeld, S. Gophna, W. F. Doolittle and S. J. Veldhuyzen van Zanten** (**2006**). "Differences between tissue-associated intestinal microfloras of patients with Crohn's disease and ulcerative colitis." *J Clin Microbiol* 44(11): 4136-41.

69. **Grimm, V., S. Ezaki, M. Susa, C. Knabbe, R. D. Schmid and T. T. Bachmann** (**2004**). "Use of DNA microarrays for rapid genotyping of TEM beta-lactamases that confer resistance." *J Clin Microbiol* 42(8): 3766-74.

70. **Guarner, F.** (**2005**). "The intestinal flora in inflammatory bowel disease: normal or abnormal?" *Curr Opin Gastroenterol* 21(4): 414-8.

71. **Hacker, J., G. Blum-Oehler, B. Hochhut and U. Dobrindt** (**2003**). "The molecular basis of infectious diseases: pathogenicity islands and other mobile genetic elements. A review." *Acta Microbiol Immunol Hung* 50(4): 321-30.

72. **Hacker, J. and E. Carniel** (**2001**). "Ecological fitness, genomic islands and bacterial pathogenicity. A Darwinian view of the evolution of microbes." *EMBO Rep* 2(5): 376-81.

73. **Hacker, J. and J. B. Kaper** (**2000**). "Pathogenicity islands and the evolution of microbes." *Annu Rev Microbiol* 54: 641-79.

74. **Hamady, M. and R. Knight** (**2009**). "Microbial community profiling for human microbiome projects: Tools, techniques, and challenges." *Genome Res* 19(7): 1141-52.

75. **Hanage, W. P., C. Fraser and B. G. Spratt** (**2006**). "Sequences, sequence clusters and bacterial species." *Philos Trans R Soc Lond B Biol Sci* 361(1475): 1917-27.

76. **Hanauer, S. B.** (**2006**). "Inflammatory bowel disease: epidemiology, pathogenesis, and therapeutic opportunities." *Inflamm Bowel Dis* 12 Suppl 1: S3-9.

77. **Hartmann, M., B. Frey, R. Kolliker and F. Widmer** (**2005**). "Semi-automated genetic analyses of soil microbial communities: comparison of T-RFLP and RISA based on descriptive and discriminative statistical approaches." *J Microbiol Methods* 61(3): 349-60.

78. **Hatt, J. K. and P. N. Rather** (**2008**). "Role of bacterial biofilms in urinary tract infections." *Curr Top Microbiol Immunol* 322: 163-92.

79. **Heimer, S. R., R. A. Welch, N. T. Perna, G. Posfai, P. S. Evans, J. B. Kaper, F. R. Blattner and H. L. Mobley** (**2002**). "Urease of enterohemorrhagic *Escherichia coli*: evidence for regulation by fur and a trans-acting factor." *Infect Immun* 70(2): 1027-31.

80. **Henderson, I. R., F. Navarro-Garcia, M. Desvaux, R. C. Fernandez and D. Ala'Aldeen** (**2004**). "Type V protein secretion pathway: the autotransporter story." *Microbiol Mol Biol Rev* 68(4): 692-744.

81. **Hendrickson, B. A., R. Gokhale and J. H. Cho** (**2002**). "Clinical aspects and pathophysiology of inflammatory bowel disease." *Clin Microbiol Rev* 15(1): 79-94.

82. **Herzer, P. J., S. Inouye, M. Inouye and T. S. Whittam** (**1990**). "Phylogenetic distribution of branched RNA-linked multicopy single-stranded DNA among natural isolates of *Escherichia coli*." *J Bacteriol* 172(11): 6175-81.

83. **Hill, D. A. and D. Artis** (**2010**). "Intestinal Bacteria and the Regulation of Immune Cell Homeostasis." *Annu Rev Immunol*.

84. **Hill, J. E., S. L. Penny, K. G. Crowell, S. H. Goh and S. M. Hemmingsen** (**2004**). "cpnDB: a chaperonin sequence database." *Genome Res* 14(8): 1669-75.

85. **Hill, J. E., R. P. Seipp, M. Betts, L. Hawkins, A. G. Van Kessel, W. L. Crosby and S. M. Hemmingsen** (**2002**). "Extensive profiling of a complex microbial community by high-throughput sequencing." *Appl Environ Microbiol* 68(6): 3055-66.

86. **Holden, N. J. and D. L. Gally** (**2004**). "Switches, cross-talk and memory in *Escherichia coli* adherence." *J Med Microbiol* 53(Pt 7): 585-93.

87. **Hooper, L. V., T. Midtvedt and J. I. Gordon** (**2002**). "How host-microbial interactions shape the nutrient environment of the mammalian intestine." *Annu Rev Nutr* 22: 283-307.

88. **Hugenholtz, P. and G. W. Tyson** (**2008**). "Microbiology: metagenomics." *Nature* 455(7212): 481-3.

89. **Ihssen, J., E. Grasselli, C. Bassin, P. Francois, J. C. Piffaretti, W. Koster, J. Schrenzel and T. Egli** (**2007**). "Comparative genomic hybridization and physiological characterization of environmental isolates indicate that significant (eco-)physiological properties are highly conserved in the species *Escherichia coli*." *Microbiology* 153(Pt 7): 2052-66.

90. **Ingram, S., J. M. McKinley, F. Vasey, J. D. Carter and P. G. Brady** (**2008**). "Are inflammatory bowel disease (IBD) and pouchitis a reactive enteropathy to group D streptococci (Enterococci)?" *Inflamm Bowel Dis* 15(11): 1609-1610.

91. **Johnson, J. R., P. Delavari, M. Kuskowski and A. L. Stell** (**2001**). "Phylogenetic distribution of extraintestinal virulence-associated traits in *Escherichia coli*." *J Infect Dis* 183(1): 78-88.

92. **Johnson, J. R. and A. L. Stell** (**2000**). "Extended virulence genotypes of *Escherichia coli* strains from patients with urosepsis in relation to phylogeny and host compromise." *J Infect Dis* 181(1): 261-72.

93. **Jowett, S. L., C. J. Seal, M. S. Pearce, E. Phillips, W. Gregory, J. R. Barton and M. R. Welfare** (**2004**). "Influence of dietary factors on the clinical course of ulcerative colitis: a prospective cohort study." *Gut* 53(10): 1479-84.

94. **Kaper, J. B., J. P. Nataro and H. L. Mobley** (**2004**). "Pathogenic *Escherichia coli*." *Nat Rev Microbiol* 2(2): 123-40.

95. **Kaser, A., S. Zeissig and R. S. Blumberg** (**2010**). "Inflammatory Bowel Disease." *Annu Rev Immunol*.

96. **Kent, A. D., D. J. Smith, B. J. Benson and E. W. Triplett** (**2003**). "Web-based phylogenetic assignment tool for analysis of terminal restriction fragment length polymorphism profiles of microbial communities." *Appl Environ Microbiol* 69(11): 6768-76.

97. **Kitahora, T., T. Utsunomiya and A. Yokota** (**1995**). "Epidemiological study of ulcerative colitis in Japan: incidence and familial occurrence. The Epidemiology Group of the Research Committee of Inflammatory Bowel Disease in Japan." *J Gastroenterol* 30 Suppl 8: 5-8.

98. **Kjaergaard, K., M. A. Schembri, H. Hasman and P. Klemm** (**2000a**). "Antigen 43 from *Escherichia coli* induces inter- and intraspecies cell aggregation and changes in colony morphology of Pseudomonas fluorescens." *J Bacteriol* 182(17): 4789-96.

99. **Kjaergaard, K., M. A. Schembri, C. Ramos, S. Molin and P. Klemm** (**2000b**). "Antigen 43 facilitates formation of multispecies biofilms." *Environ Microbiol* 2(6): 695-702.

100. **Knight, P., B. J. Campbell and J. M. Rhodes** (**2008**). "Host-bacteria interaction in inflammatory bowel disease." *Br Med Bull* 88(1): 95-113.

101. Kocherginskaya, S. A., I. K. O. Cann and R. I. Mackie (2005). Denaturing gradient gel electrophoresis. Methods in Gut Microbial Ecology for Ruminants. Dordrecht, The Netherlands, Springer**:** 119-128.

102. **Korzenik, J. R.** (**2005**). "Past and current theories of etiology of IBD: toothpaste, worms, and refrigerators." *J Clin Gastroenterol* 39(4 Suppl 2): S59-65.

103. **Kotlowski, R., C. N. Bernstein, S. Sepehri and D. O. Krause** (**2007**). "High prevalence of *Escherichia coli* belonging to the B2+D phylogenetic group in inflammatory bowel disease." *Gut* 56(5): 669-75.

104. **La Scola, B., Z. Zeaiter, A. Khamis and D. Raoult** (**2003**). "Gene-sequence-based criteria for species definition in bacteriology: the *Bartonella* paradigm." *Trends Microbiol* 11(7): 318-21.

105. Lane, D. J. (1991). 16S/23S rRNA sequencing. Nucleic Acid Techniques in Bacterial Systematics. E. Stackebrandt and M. Goodfellow. New York. , John Wiley & Sons.**:** 115-175.

106. **Larkin, M. A., G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez, J. D. Thompson, T. J.**

Gibson and D. G. Higgins (**2007**). "Clustal W and Clustal X version 2.0." *Bioinformatics* 23(21): 2947-8.

107. **Lawrence, J. G. and H. Ochman** (**1998**). "Molecular archaeology of the *Escherichia coli* genome." *Proc Natl Acad Sci U S A* 95(16): 9413-7.

108. **Le Gall, T., O. Clermont, S. Gouriou, B. Picard, X. Nassif, E. Denamur and O. Tenaillon** (**2007**). "Extraintestinal virulence is a coincidental by-product of commensalism in B2 phylogenetic group *Escherichia coli* strains." *Mol Biol Evol* 24(11): 2373-84.

109. **Leckie, S. E., C. E. Prescott, S. J. Grayston, J. D. Neufeld and W. W. Mohn** (**2004**). "Characterization of humus microbial communities in adjacent forest types that differ in nitrogen availability." *Microb Ecol* 48(1): 29-40.

110. **Lecointre, G., L. Rachdi, P. Darlu and E. Denamur** (**1998**). "*Escherichia coli* molecular phylogeny using the incongruence length difference test." *Mol Biol Evol* 15(12): 1685-95.

111. **Lepage, P., P. Seksik, M. Sutren, M. F. de la Cochetiere, R. Jian, P. Marteau and J. Dore** (**2005**). "Biodiversity of the mucosa-associated microbiota is stable along the distal digestive tract in healthy individuals and patients with IBD." *Inflamm Bowel Dis* 11(5): 473-80.

112. **Leser, T. D., J. Z. Amenuvor, T. K. Jensen, R. H. Lindecrona, M. Boye and K. Moller** (**2002**). "Culture-independent analysis of gut bacteria: the pig gastrointestinal tract microbiota revisited." *Appl Environ Microbiol* 68(2): 673-90.

113. **Ley, R. E., F. Backhed, P. Turnbaugh, C. A. Lozupone, R. D. Knight and J. I. Gordon** (**2005**). "Obesity alters gut microbial ecology." *Proc Natl Acad Sci U S A* 102(31): 11070-5.

114. **Ley, R. E., P. J. Turnbaugh, S. Klein and J. I. Gordon** (**2006**). "Microbial ecology: human gut microbes associated with obesity." *Nature* 444(7122): 1022-3.

115. **Longobardi, T., P. Jacobs, L. Wu and C. N. Bernstein** (**2003**). "Work losses related to inflammatory bowel disease in Canada: results from a National Population Health Survey." *Am J Gastroenterol* 98(4): 844-9.

116. **Lucke, K., S. Miehlke, E. Jacobs and M. Schuppler** (**2006**). "Prevalence of *Bacteroides* and *Prevotella* spp. in ulcerative colitis." *J Med Microbiol* 55(Pt 5): 617-24.

117. **Ludwig, W.** (**2007**). "Nucleic acid techniques in bacterial systematics and identification." *Int J Food Microbiol* 120(3): 225-36.

118. **Ludwig, W. and K. H. Schleifer** (**1994**). "Bacterial phylogeny based on 16S and 23S rRNA sequence analysis." *FEMS Microbiol Rev* 15(2-3): 155-73.

119. Mackie, R. I., I. K. O. Cann, E. Zoetendal and E. Forano (2007). Molecular approaches to study bacterial diversity and functio in the intestinal tract. 7th International Symposium on the Nutrition of Herbivores, Beijing, China, China Agricultural Press.

120. **Mahida, Y. R. and V. E. Rolfe** (**2004**). "Host-bacterial interactions in inflammatory bowel disease." *Clin Sci (Lond)* 107(4): 331-41.

121. **Manichanh, C., L. Rigottier-Gois, E. Bonnaud, K. Gloux, E. Pelletier, L. Frangeul, R. Nalin, C. Jarrin, P. Chardon, P. Marteau, J. Roca and J. Dore** (**2006**). "Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach." *Gut* 55(2): 205-11.

122. **Manson, J. M., M. Rauch and M. S. Gilmore** (**2008**). "The commensal microbiology of the gastrointestinal tract." *Adv Exp Med Biol* 635: 15-28.

123. **Martin, H. M., B. J. Campbell, C. A. Hart, C. Mpofu, M. Nayar, R. Singh, H. Englyst, H. F. Williams and J. M. Rhodes** (**2004**). "Enhanced *Escherichia coli* adherence and invasion in Crohn's disease and colon cancer." *Gastroenterology* 127(1): 80-93.

124. **Martinez-Medina, M., X. Aldeguer, M. Lopez-Siles, F. Gonzalez-Huix, C. Lopez-Oliu, G. Dahbi, J. E. Blanco, J. Blanco, L. J. Garcia-Gil and A. Darfeuille-Michaud** (**2009**). "Molecular diversity of *Escherichia coli* in the human gut: New ecological evidence supporting the role of adherent-invasive *E. coli* (AIEC) in Crohn's disease." *Inflamm Bowel Dis*.

125. **Mazmanian, S. K., J. L. Round and D. L. Kasper** (**2008**). "A microbial symbiosis factor prevents intestinal inflammatory disease." *Nature* 453(7195): 620-5.

126. **McKinley, J., F. Vasey, A. McMillen, J. Carter and P. Brady** (**2008**). "Joint observations with lessons for rheumatologists and gastroenterologists." *Inflamm Bowel Dis* 14(5): 730-1.

127. **Mentula, S., J. Harmoinen, M. Heikkila, E. Westermarck, M. Rautio, P. Huovinen and E. Kononen** (**2005**). "Comparison between cultured small-intestinal and fecal microbiotas in beagle dogs." *Appl Environ Microbiol* 71(8): 4169-75.

128.   **Miyoshi, T., T. Iwatsuki and T. Naganuma** (**2005**). "Phylogenetic characterization of 16S rRNA gene clones from deep-groundwater microorganisms that pass through 0.2-micrometer-pore-size filters." *Appl Environ Microbiol* 71(2): 1084-8.

129. **Morschhauser, J., B. E. Uhlin and J. Hacker** (**1993**). "Transcriptional analysis and regulation of the sfa determinant coding for S fimbriae of pathogenic *Escherichia coli* strains." *Mol Gen Genet* 238(1-2): 97-105.

130.   **Moscandrew, M. E. and E. V. Loftus, Jr.** (**2009**). "Diagnostic advances in inflammatory bowel disease (imaging and laboratory)." *Curr Gastroenterol Rep* 11(6): 488-95.

131. **Moulin-Schouleur, M., M. Reperant, S. Laurent, A. Bree, S. Mignon-Grasteau, P. Germon, D. Rasschaert and C. Schouler** (**2007**). "Extraintestinal pathogenic *Escherichia coli* strains of avian and human origin: link between phylogenetic relationships and common virulence patterns." *J Clin Microbiol* 45(10): 3366-76.

132. **Mutlu, E. A. and N. Gor** (**2008**). "To diet or not if you have inflammatory bowel disease." *Expert Rev Gastroenterol Hepatol* 2(5): 613-6.

133. **Mylonaki, M., N. B. Rayment, D. S. Rampton, B. N. Hudspith and J. Brostoff** (**2005**). "Molecular characterization of rectal mucosa-associated bacterial flora in inflammatory bowel disease." *Inflamm Bowel Dis* 11(5): 481-7.

134. **Nelson, K. E., S. H. Zinder, I. Hance, P. Burr, D. Odongo, D. Wasawo, A. Odenyo and R. Bishop** (**2003**). "Phylogenetic analysis of the microbial populations in the wild herbivore gastrointestinal tract: insights into an unexplored niche." *Environ Microbiol* 5(11): 1212-20.

135. **Nguyen, D. D.** (**2008**). "Are *E. coli* the culprit?" *Inflamm Bowel Dis* 14(4): 575-6.

136. **Niv, Y., G. Abuksis and G. M. Fraser** (**2000**). "Epidemiology of ulcerative colitis in Israel: a survey of Israeli kibbutz settlements." *Am J Gastroenterol* 95(3): 693-8.

137. **Nowrouzian, F. L., I. Adlerberth and A. E. Wold** (**2006**). "Enhanced persistence in the colonic microbiota of *Escherichia coli* strains belonging to phylogenetic group B2: role of virulence factors and adherence to colonic cells." *Microbes Infect* 8(3): 834-40.

138.   **Ochman, H. and R. K. Selander** (**1984**). "Standard reference strains of *Escherichia coli* from natural populations." *J Bacteriol* 157(2): 690-3.

139. **Odes, H. S., C. Locker, L. Neumann, H. J. Zirkin, Z. Weizman, A. D. Sperber, G. M. Fraser, P. Krugliak, N. Gaspar, L. Eidelman and et al.** (**1994**). "Epidemiology of Crohn's disease in southern Israel." *Am J Gastroenterol* 89(10): 1859-62.

140. **Ott, S. J., M. Musfeldt, D. F. Wenderoth, J. Hampe, O. Brant, U. R. Folsch, K. N. Timmis and S. Schreiber** (**2004**). "Reduction in diversity of the colonic mucosa associated bacterial microflora in patients with active inflammatory bowel disease." *Gut* 53(5): 685-93.

141. **Owen, R. J.** (**2004**). "Bacterial taxonomics: finding the wood through the phylogenetic trees." *Methods Mol Biol* 266: 353-83.

142. **Ozutsumi, Y., H. Hayashi, M. Sakamoto, H. Itabashi and Y. Benno** (**2005a**). "Culture-independent analysis of fecal microbiota in cattle." *Biosci Biotechnol Biochem* 69(9): 1793-7.

143. **Ozutsumi, Y., K. Tajima, A. Takenaka and H. Itabashi** (**2005b**). "The effect of protozoa on the composition of rumen bacteria in cattle using 16S rRNA gene clone libraries." *Biosci Biotechnol Biochem* 69(3): 499-506.

144. **Perna, N. T., G. Plunkett, 3rd, V. Burland, B. Mau, J. D. Glasner, D. J. Rose, G. F. Mayhew, P. S. Evans, J. Gregor, H. A. Kirkpatrick, G. Posfai, J. Hackett, S. Klink, A. Boutin, Y. Shao, L. Miller, E. J. Grotbeck, N. W. Davis, A. Lim, E. T. Dimalanta, K. D. Potamousis, J. Apodaca, T. S. Anantharaman, J. Lin, G. Yen, D. C. Schwartz, R. A. Welch and F. R. Blattner** (**2001**). "Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7." *Nature* 409(6819): 529-33.

145. **Picard, B., J. S. Garcia, S. Gouriou, P. Duriez, N. Brahimi, E. Bingen, J. Elion and E. Denamur** (**1999**). "The link between phylogeny and virulence in *Escherichia coli* extraintestinal infection." *Infect Immun* 67(2): 546-53.

146. **Pineton de Chambrun, G., J. F. Colombel, D. Poulain and A. Darfeuille-Michaud** (**2008**). "Pathogenic agents in inflammatory bowel diseases." *Curr Opin Gastroenterol* 24(4): 440-7.

147. **Pinsk, V., D. A. Lemberg, K. Grewal, C. C. Barker, R. A. Schreiber and K. Jacobson** (**2007**). "Inflammatory bowel disease in the South Asian pediatric population of British Columbia." *Am J Gastroenterol* 102(5): 1077-83.

148. **Preston, G. M., B. Haubold and P. B. Rainey** (**1998**). "Bacterial genomics and adaptation to life on plants: implications for the evolution of pathogenicity and symbiosis." *Curr Opin Microbiol* 1(5): 589-97.

149. **Rappe, M. S. and S. J. Giovannoni** (**2003**). "The uncultured microbial majority." *Annu Rev Microbiol* 57: 369-94.

150. **Rath, H. C.** (**2003**). "The role of endogenous bacterial flora: bystander or the necessary prerequisite?" *Eur J Gastroenterol Hepatol* 15(6): 615-20.

151. **Rhodes, J. M.** (**2007**). "The role of *Escherichia coli* in inflammatory bowel disease." *Gut* 56(5): 610-2.

152. **Riesenfeld, C. S., P. D. Schloss and J. Handelsman** (**2004**). "Metagenomics: genomic analysis of microbial communities." *Annu Rev Genet* 38: 525-52.

153. **Robertson, S. P. and R. Poulton** (**2008**). "Longitudinal studies of gene-environment interaction in common diseases--good value for money?" *Novartis Found Symp* 293: 128-37; discussion 138-42, 181-3.

154. **Roos, L. L., A. M. Cameron, J. P. Nicol, D. F. McLerran, D. J. Malenka, T. K. Young and M. M. Cohen** (**1993**). "Registries and Administrative Data: Organization and Accuracy." *Medical Care* 31(3): 201-212.

155. **Ruan, Q., J. A. Steele, M. S. Schwalbach, J. A. Fuhrman and F. Sun** (**2006**). "A dynamic programming algorithm for binning microbial community profiles." *Bioinformatics* 22(12): 1508-14.

156. **Rutgeerts, P., S. Vermeire and G. Van Assche** (**2009**). "Biological therapies for inflammatory bowel diseases." *Gastroenterology* 136(4): 1182-97.

157. **Ryan, P., R. G. Kelly, G. Lee, J. K. Collins, G. C. O'Sullivan, J. O'Connell and F. Shanahan** (**2004**). "Bacterial DNA within granulomas of patients with Crohn's disease--detection by laser capture microdissection and PCR." *Am J Gastroenterol* 99(8): 1539-43.

158. **Sakamoto, N., S. Kono, K. Wakai, Y. Fukuda, M. Satomi, T. Shimoyama, Y. Inaba, Y. Miyake, S. Sasaki, K. Okamoto, G. Kobashi, M. Washio, T. Yokoyama, C. Date and H. Tanaka** (**2005**). "Dietary risk factors for inflammatory bowel disease: a multicenter case-control study in Japan." *Inflamm Bowel Dis* 11(2): 154-63.

159. **Salzman, N. H. and C. L. Bevins** (**2008**). "Negative interactions with the microbiota: IBD." *Adv Exp Med Biol* 635: 67-78.

160. **Sandborn, W. J.** (**2008**). "Current directions in IBD therapy: what goals are feasible with biological modifiers?" *Gastroenterology* 135(5): 1442-7.

161. **Sartor, R. B. (2001)**. "Intestinal microflora in human and experimental inflammatory bowel disease." *Curr Opin Gastroenterol* 17(4): 324-30.

162. **Sartor, R. B. (2006)**. "Mechanisms of disease: pathogenesis of Crohn's disease and ulcerative colitis." *Nat Clin Pract Gastroenterol Hepatol* 3(7): 390-407.

163. **Sartor, R. B. (2008)**. "Microbial influences in inflammatory bowel diseases." *Gastroenterology* 134(2): 577-94.

164. **Scanlan, P. D., F. Shanahan, C. O'Mahony and J. R. Marchesi (2006)**. "Culture-independent analyses of temporal variation of the dominant fecal microbiota and targeted bacterial subgroups in Crohn's disease." *J Clin Microbiol* 44(11): 3980-8.

165. **Schippa, S., M. P. Conte, O. Borrelli, V. Iebba, M. Aleandri, L. Seganti, C. Longhi, F. Chiarini, J. Osborn and S. Cucchiara (2008)**. "Dominant genotypes in mucosa-associated *Escherichia coli* strains from pediatric patients with inflammatory bowel disease." *Inflamm Bowel Dis*.

166. **Schmidt, H. and M. Hensel (2004)**. "Pathogenicity islands in bacterial pathogenesis." *Clin Microbiol Rev* 17(1): 14-56.

167. **Schmoll, T., J. Morschhauser, M. Ott, B. Ludwig, I. van Die and J. Hacker (1990)**. "Complete genetic organization and functional aspects of the *Escherichia coli* S fimbrial adhesion determinant: nucleotide sequence of the genes sfa B, C, D, E, F." *Microb Pathog* 9(5): 331-43.

168. **Schneider, D. and R. E. Lenski (2004)**. "Dynamics of insertion sequence elements during experimental evolution of bacteria." *Res Microbiol* 155(5): 319-27.

169. **Schultz, M. (2008)**. "Clinical use of *E. coli* Nissle 1917 in inflammatory bowel disease." *Inflamm Bowel Dis* 14(7): 1012-8.

170. **Seksik, P., P. Lepage, M. F. de la Cochetiere, A. Bourreille, M. Sutren, J. P. Galmiche, J. Dore and P. Marteau (2005)**. "Search for localized dysbiosis in Crohn's disease ulcerations by temporal temperature gradient gel electrophoresis of 16S rRNA." *J Clin Microbiol* 43(9): 4654-8.

171. **Sellon, R. K., S. Tonkonogy, M. Schultz, L. A. Dieleman, W. Grenther, E. Balish, D. M. Rennick and R. B. Sartor (1998)**. "Resident enteric bacteria are necessary for development of spontaneous colitis and immune system activation in interleukin-10-deficient mice." *Infect Immun* 66(11): 5224-31.

172. **Sepehri, S., R. Kotlowski, C. N. Bernstein and D. O. Krause** (**2007**). "Microbial diversity of inflamed and noninflamed gut biopsy tissues in inflammatory bowel disease." *Inflamm Bowel Dis* 13(6): 675-83.

173. **Servin, A. L.** (**2005**). "Pathogenesis of Afa/Dr diffusely adhering *Escherichia coli*." *Clin Microbiol Rev* 18(2): 264-92.

174. **Simpson, K. W., B. Dogan, M. Rishniw, R. E. Goldstein, S. Klaessig, P. L. McDonough, A. J. German, R. M. Yates, D. G. Russell, S. E. Johnson, D. E. Berg, J. Harel, G. Bruant, S. P. McDonough and Y. H. Schukken** (**2006**). "Adherent and invasive *Escherichia coli* is associated with granulomatous colitis in boxer dogs." *Infect Immun* 74(8): 4778-92.

175. **Sokol, H., C. Lay, P. Seksik and G. W. Tannock** (**2008a**). "Analysis of bacterial bowel communities of IBD patients: what has it revealed?" *Inflamm Bowel Dis* 14(6): 858-67.

176. **Sokol, H., B. Pigneur, L. Watterlot, O. Lakhdari, L. G. Bermudez-Humaran, J. J. Gratadoux, S. Blugeon, C. Bridonneau, J. P. Furet, G. Corthier, C. Grangette, N. Vasquez, P. Pochart, G. Trugnan, G. Thomas, H. M. Blottiere, J. Dore, P. Marteau, P. Seksik and P. Langella** (**2008b**). "*Faecalibacterium prausnitzii* is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients." *Proc Natl Acad Sci U S A* 105(43): 16731-6.

177. **Sokol, H., P. Seksik, L. Rigottier-Gois, C. Lay, P. Lepage, I. Podglajen, P. Marteau and J. Dore** (**2006**). "Specificities of the fecal microbiota in inflammatory bowel disease." *Inflamm Bowel Dis* 12(2): 106-11.

178. **Sollid, L. M. and F. E. Johansen** (**2008**). "Animal models of inflammatory bowel disease at the dawn of the new genetics era." *PLoS Med* 5(9): e198.

179. **Sonnenberg, A.** (**1990a**). "Occupational distribution of inflammatory bowel disease among German employees." *Gut* 31(9): 1037-40.

180. **Sonnenberg, A.** (**1990b**). "Occupational mortality of inflammatory bowel disease." *Digestion* 46(1): 10-8.

181. **Swidsinski, A., A. Ladhoff, A. Pernthaler, S. Swidsinski, V. Loening-Baucke, M. Ortner, J. Weber, U. Hoffmann, S. Schreiber, M. Dietel and H. Lochs** (**2002**). "Mucosal flora in inflammatory bowel disease." *Gastroenterology* 122(1): 44-54.

182. **Swidsinski, A., V. Loening-Baucke, H. Lochs and L. P. Hale** (**2005**). "Spatial organization of bacterial flora in normal and inflamed intestine: a fluorescence in situ hybridization study in mice." *World J Gastroenterol* 11(8): 1131-40.

183. **Tamboli, C. P., C. Neut, P. Desreumaux and J. F. Colombel** (**2004**). "Dysbiosis as a prerequisite for IBD." *Gut* 53(7): 1057.

184. **Tamura, K., J. Dudley, M. Nei and S. Kumar** (**2007**). "MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0." *Mol Biol Evol* 24(8): 1596-9.

185. **Tannock, G. W.** (**1999**). "Analysis of the intestinal microflora: a renaissance." *Antonie Van Leeuwenhoek* 76(1-4): 265-78.

186. **Tannock, G. W.** (**2008**). "The search for disease-associated compositional shifts in bowel bacterial communities of humans." *Trends Microbiol* 16(10): 488-95.

187. **Tartof, S. Y., O. D. Solberg, A. R. Manges and L. W. Riley** (**2005**). "Analysis of a uropathogenic *Escherichia coli* clonal group by multilocus sequence typing." *J Clin Microbiol* 43(12): 5860-4.

188. **Thompson-Chagoyan, O. C., J. Maldonado and A. Gil** (**2005**). "Aetiology of inflammatory bowel disease (IBD): role of intestinal microbiota and gut-associated lymphoid tissue immune response." *Clin Nutr* 24(3): 339-52.

189. **Turnbaugh, P. J., M. Hamady, T. Yatsunenko, B. L. Cantarel, A. Duncan, R. E. Ley, M. L. Sogin, W. J. Jones, B. A. Roe, J. P. Affourtit, M. Egholm, B. Henrissat, A. C. Heath, R. Knight and J. I. Gordon** (**2009**). "A core gut microbiome in obese and lean twins." *Nature* 457(7228): 480-4.

190. **Uilenberg, G., F. Thiaucourt and F. Jongejan** (**2004**). "On molecular taxonomy: what is in a name?" *Exp Appl Acarol* 32(4): 301-12.

191. **Vandamme, P., B. Pot, M. Gillis, P. de Vos, K. Kersters and J. Swings** (**1996**). "Polyphasic taxonomy, a consensus approach to bacterial systematics." *Microbiol Rev* 60(2): 407-38.

192. **Velayos, F.** (**2008**). "Colon cancer surveillance in inflammatory bowel disease patients: current and emerging practices." *Expert Rev Gastroenterol Hepatol* 2(6): 817-25.

193. **Wagner, J., K. Short, A. G. Catto-Smith, D. J. Cameron, R. F. Bishop and C. D. Kirkwood** (**2008**). "Identification and characterisation of Pseudomonas 16S ribosomal DNA from ileal biopsies of children with Crohn's disease." *PLoS One* 3(10): e3578.

194. **Walther, B. A. and S. Morand** (**1998**). "Comparative performance of species richness estimation methods." *Parasitology* 116 (Pt 4): 395-405.

195. **Washington, H. G.** (**1984**). "Diversity, biotic and similarity indices: A review with special relevance to aquatic ecosystems." *Water Research* 18(6): 653-694.

196. **Weinstock, J. V., R. W. Summers, D. E. Elliott, K. Qadir, J. F. Urban, Jr. and R. Thompson** (**2002**). "The possible link between de-worming and the emergence of immunological disease." *J Lab Clin Med* 139(6): 334-8.

197. **Weissman, S. J., S. Chattopadhyay, P. Aprikian, M. Obata-Yasuoka, Y. Yarova-Yarovaya, A. Stapleton, W. Ba-Thein, D. Dykhuizen, J. R. Johnson and E. V. Sokurenko** (**2006**). "Clonal analysis reveals high rate of structural mutations in fimbrial adhesins of extraintestinal pathogenic *Escherichia coli*." *Mol Microbiol* 59(3): 975-88.

198. **Williams, H., D. Walker and T. R. Orchard** (**2008**). "Extraintestinal manifestations of inflammatory bowel disease." *Curr Gastroenterol Rep* 10(6): 597-605.

199. **Wirth, T., D. Falush, R. Lan, F. Colles, P. Mensa, L. H. Wieler, H. Karch, P. R. Reeves, M. C. Maiden, H. Ochman and M. Achtman** (**2006**). "Sex and virulence in *Escherichia coli*: an evolutionary perspective." *Mol Microbiol* 60(5): 1136-51.

200. **Woese, C. R.** (**1987**). "Bacterial evolution." *Microbiol Rev* 51(2): 221-71.

201. **Wolda, H.** (**1981**). "Similarity Indices, Sample Size and Diversity." *Oecologia* 50(3): 296-302.

202. **Xavier, R. J. and D. K. Podolsky** (**2007**). "Unravelling the pathogenesis of inflammatory bowel disease." *Nature* 448(7152): 427-34.

203. **Yamamoto, S. and S. Harayama** (**1996**). "Phylogenetic analysis of *Acinetobacter* strains based on the nucleotide sequences of gyrB genes and on the amino acid sequences of their products." *Int J Syst Bacteriol* 46(2): 506-11.

204. **Yu, X., M. Susa, C. Knabbe, R. D. Schmid and T. T. Bachmann** (**2004**). "Development and validation of a diagnostic DNA microarray to detect quinolone-resistant *Escherichia coli* among clinical isolates." *J Clin Microbiol* 42(9): 4083-91.

205. **Zhang, H., D. Massey, M. Tremelling and M. Parkes** (**2008**). "Genetics of inflammatory bowel disease: clues to pathogenesis." *Br Med Bull* 87: 17-30.

206. **Zhang, L., B. Foxman and C. Marrs** (**2002**). "Both urinary and rectal *Escherichia coli* isolates are dominated by strains of phylogenetic group B2." *J Clin Microbiol* 40(11): 3951-5.

207. **Zhang, L., S. M. Man, A. S. Day, S. T. Leach, D. A. Lemberg, S. Dutt, M. Stormon, A. Otley, E. V. O'Loughlin, A. Magoffin, P. H. Ng and H. Mitchell** (**2009**). "Detection and isolation of *Campylobacter* species other than *C. jejuni* from children with Crohn's disease." *J Clin Microbiol* 47(2): 453-5.

208. **Zoetendal, E. G., A. D. Akkermans and W. M. De Vos** (**1998**). "Temperature gradient gel electrophoresis analysis of 16S rRNA from human fecal samples reveals stable and host-specific communities of active bacteria." *Appl Environ Microbiol* 64(10): 3854-9.

209. **Zoetendal, E. G., A. D. L. Akkermans, W. M. Akkermans-van Vliet, J. A. G. M. de Visser and W. M. de Vos** (**2001**). "The Host Genotype Affects the Bacterial Community in the Human Gastronintestinal Tract." *Microbial Ecology in Health & Disease* 13(3): 129-134.

210. **Zoetendal, E. G., A. von Wright, T. Vilpponen-Salmela, K. Ben-Amor, A. D. Akkermans and W. M. de Vos** (**2002**). "Mucosa-associated bacteria in the human gastrointestinal tract are uniformly distributed along the colon and differ from the community recovered from feces." *Appl Environ Microbiol* 68(7): 3401-7.

211. **Zuckerkandl, E. and L. Pauling** (**1965**). "Molecules as documents of evolutionary history." *J Theor Biol* 8(2): 357-66.