CORA Cork Open Research Archive
Cartlann Taighde Oscailte Chorcaí

Title	The gut microbiome in inflammatory bowel disease and its confounders
Authors	Eckenberger, Julia
Publication date	2023
Original Citation	Eckenberger, J. 2023. The gut microbiome in inflammatory bowel disease and its confounders. PhD Thesis, University College Cork.
Type of publication	Doctoral thesis
Rights	© 2023, Julia Eckenberger. - https://creativecommons.org/licenses/by-nc-sa/4.0/
Download date	2026-11-03 07:31:54
Item downloaded from	https://hdl.handle.net/10468/15960



UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

Ollscoil na hÉireann, Corcaigh

National University of Ireland, Cork



The Gut Microbiome in Inflammatory Bowel Disease and its Confounders

Thesis presented by

Julia Eckenberger, M.Sc.

For the degree of

Doctor of Philosophy

University College Cork

School of Microbiology

Head of School: Dr. David Clarke

Supervisor: Assoc. Prof. Marcus Claesson

October 2023

Table of Contents

Declaration....3

Publications....4

Abstract....6

Chapter I: Literature Review....8

Section I: The human gut microbiome in IBD....9

Section II: Bioinformatic Platforms for Metagenomics....50

Aim....89

Chapter II: Ranking microbiome variance in inflammatory bowel disease:
A large longitudinal intercontinental study....90

Chapter III: Interactions between medications and the gut microbiome in
inflammatory bowel disease....125

Chapter IV: Interactions between farnesoid X receptor associated metabolites
and gut microbiota in patients with Crohn's disease....149

Chapter V: Comparison of different ML algorithms for Classification of IBD
from Amplicon Sequencing....193

General discussion....246

Bibliography....259

Acknowledgements....300

Declaration

This is to certify that the work I am submitting is my own and has not been submitted for another degree, either at University College Cork or elsewhere. All external references and sources are clearly acknowledged and identified within the contents. I have read and understood the regulations of University College Cork concerning plagiarism and intellectual property.

Signed: 

Julia Eckenberger

October 2023

Publications

Eckenberger, J*, Walker, SP*, and Claesson, MJ, *1.05 - Bioinformatic Platforms for Metagenomics*. Comprehensive Foodomics, 2021, Elsevier: Oxford. p. 91-112, DOI: 10.1016/B978-0-12-816395-5.00009-5

Clooney, AG*, **Eckenberger, J.***, Laserna-Mendieta, E, Sexton, KA, Bernstein, MT, Vagianos, K, Sargent, M, Ryan, FJ, Moran, C, Sheehan, D, Sleator, RD, Targownik LE, Bernstein, CN, Shanahan, F, and Claesson, MJ, *Ranking microbiome variance in inflammatory bowel disease: a large longitudinal intercontinental study.* Gut, 2021. 70(3): p. 499-510, DOI: 10.1136/gutjnl-2020-321106

Lyte, JM, Keane, J*, **Eckenberger, J***, Anthony, N, Shrestha, S, Marasini, D, Daniels, KM, Caputi, V, Donoghue, AM, and Lyte, M, *Japanese quail (Coturnix japonica) as a novel model to study the relationship between the avian microbiome and microbial endocrinology-based host-microbe interactions.* Microbiome, 2021. 9(1): p. 38, DOI: 10.1186/s40168-020-00962-2

Hogan, G, **Eckenberger, J**, Narayanen, N, Walker, SP, Claesson, MJ, Corrigan, M, O'Hanlon, D, and Tangney, M, *Biopsy bacterial signature can predict patient tissue malignancy.* Scientific Reports, 2021. 11(1): p. 18535, DOI: 10.1038/s41598-021-98089-3

Bannon, CD, **Eckenberger, J**, Snelling, WJ, Huseyin, CE, Allsopp, P, Strain, C, Ramnani, P, Chitarrari, R, Grant, J, Hotchkiss, S, Philp, K, Campbell, R, Tuohy, KM, Claesson, MJ, Ternan, NG, Dooley, JSG, Sleator, RD, Rowland, I,and Gill, CIR, *Low-Molecular-Weight Seaweed-Derived Polysaccharides Lead to Increased Faecal Bulk but Do Not Alter Human Gut Health Markers*. Foods, 2021. 10, DOI: 10.3390/foods10122988

Eckenberger, J, Butler, JC, Bernstein, CN, Shanahan, F, and Claesson, MJ, *Interactions between Medications and the Gut Microbiome in Inflammatory Bowel Disease*. Microorganisms, 2022. 10, DOI: 10.3390/microorganisms10101963.

FitzGerald, J, Patel, S, **Eckenberger, J**, Guillemard, E, Veiga, P, Schäfer, F, Walter, J, Claesson, MJ, and Derrien, M., *Improved gut microbiome recovery following drug therapy is linked to abundance and replication of probiotic strains.* Gut Microbes, 2022. 14(1): p. 2094664, DOI: 10.1080/19490976.2022.2094664

Papoutsoglou, G, Tarazona, S, Lopes, MB, Klammsteiner, T, Ibrahimi, E, **Eckenberger, J**, Novielli, P, Tonda, A, Simeon, A, Shigdel, R, Béreux, S, Vitali, G, Tangaro, S, Lahti, L, Temko, A, Claesson, MJ, and Berland, M, *Machine Learning Approaches in Microbiome Research: Challenges and Best Practices*. Frontiers in Microbiology, 2023. 14:1261889. DOI: 10.3389/fmicb.2023.1261889

Lyte, JM, **Eckenberger, J***, Keane, J*, Robinson, K, Bacon, T, Facchetti, AL, Assumpcao, V, Donoghue, AM, Liyanage, R, Daniels, KM, Caputi, V and Lyte, M, *Cold stress initiates distinct catecholaminergic and serotonergic responses in the chicken gut that are associated with functional shifts in the cecal microbiome*. Poultry Science, 2024*, 103393, DOI: 10.1016/j.psj.2023.103393*

* Authors contributed equally

“Prediction is very difficult, especially about the future”

- Niels Bohr

Abstract

Inflammatory bowel disease (IBD), encompassing Crohn's disease (CD) and ulcerative colitis (UC), is a chronic remittent-relapsing inflammatory disorder of the gastrointestinal tract that affects millions of people worldwide. Despite progress in disentangling the pathogenesis of this disease, the exact cause of IBD remains unknown. As with other chronic inflammatory disorders, the tissue damage is immune mediated and arises from an interaction of genetic susceptibility factors, environmental triggers and indigenous gut microbiota. The gut microbiota play a central role in IBD pathogenesis. However, despite consistent reports of alterations in gut microbial composition, a coherent microbial signature for IBD remains elusive.

Therefore, this work investigated the influence of IBD on the gut microbiota with a particular focus on the multitude of factors, both external and internal, that can potentially confound the distinctions between healthy and diseased states. Our investigation uncovered significant compositional disparities, especially in the case of CD, when comparing individuals with diseases to the control group. Furthermore, longitudinal analyses revealed reduced temporal stability in the microbiota of IBD patients, especially those experiencing fluctuations in disease activity. Geographic location emerged as one of the strongest drivers of microbiota variance, only second to a diagnosis with CD, followed by a history of surgical resection and a diagnosis with UC. Other life style factors also exerted an influence, however, the majority of the compositional variance remained either unexplained or was stochastic in nature.

In view of the increasing evidence that commonly prescribed, non-antibiotic drugs interact with the gut microbiome, we re-examined the microbiota variance in IBD to determine the degree to which medications might account for compositional differences between disease-subtypes and geographic location. Although there were variations in medication profiles among individuals from different countries, treatments accounted for a relatively small proportion of the geographic contribution to microbiome. With that said, the cumulative effects of multiple medications significantly contributed to the microbiome differences between patients with UC and CD.

Cognizant of the crucial role that microbial metabolites play as molecular messengers facilitating communication between the gut microbiota and the host, we next conducted an investigation into the role of microbial metabolites in patients with CD within the context of the liver-bile acid-microbiota axis. Our findings indicate that the typical signalling from the gut to the liver is disrupted in patients with CD compared to healthy controls, which led to excessive hepatic bile acid (BA) synthesis in a subset of patients. Moreover, variations in hepatic BA synthesis and BA reabsorption within the CD patient group were associated with the resection status. As a result of this disruption, we observed specific microbial changes among CD patients marked by an increase in bile-resistant and a decrease in bile-sensitive genera. This suggests that changes in BA metabolites significantly contribute to the observed differences in microbial composition between health and disease as well as between patients with CD.

Collectively, these findings underscore the dynamic nature of microbiomes, highlighting their capacity to adapt to changing environmental conditions while also being subject to host-driven regulation. This added complexity, as well as the increasing data volumes, underscores the need for innovative analytical methodologies that can effectively capture all available information, especially considering the unique characteristics of microbiome data. Hence, we assessed machine learning algorithms, including Support Vector Machines (SVM), Extreme Gradient Boosting (XGB), and Random Forest (RF), for their ability to classify IBD phenotypes using gut microbiome data. All of the tested models successfully differentiated between IBD and non-IBD controls and, to a lesser degree, between IBD subtypes across studies conducted in diverse geographic locations. Importantly, all three algorithms exhibited variations in the selection of taxonomic features considered significant for the classification task, underscoring the need for caution when applying machine learning to tasks aimed at understanding underlying biological aspects rather than solely achieving precise phenotype predictions.

In conclusion, the multifaceted approaches undertaken in this body of work yielded valuable insights into the complex interplay of lifestyle, medication, and microbial metabolites in the context of IBD, emphasising the importance of personalized approaches for host heterogeneity and environmental factors in the pursuit of precision medicine.

Chapter I

Literature Review

Section I: IBD and the human gut microbiome

Section II: Bioinformatic Platforms for Metagenomics

Section II of the literature review was published as a book chapter in Comprehensive Foodomics:

1.05 - Bioinformatic Platforms for Metagenomics

Eckenberger, J., Walker, S.P.*, and Claesson, M.J*

Elsevier: Oxford. p. 91-112, DOI: 10.1016/B978-0-12-816395-5.00009-5

Author contributions: Amplicon sequencing: JE, Whole Genome Shotgun Sequence Analysis: SPW, review and editing: JE, SPW, MJC

Section I: The human gut microbiome in IBD

1. IBD 10

1.1. Epidemiology 11

1.2. Diagnosis 12

1.3. Treatment and Management 14

1.4. Surgery 18

2. Pathogenesis of IBD 19

2.1. Genetic Risk Factors 20

2.2. Immunological abnormalities 21

2.3. Environmental triggers 24

3. Gut Microbiota in IBD 33

3.1. Microbial composition in IBD 35

Faecalibacterium prausnitzii 39

Akkermansia muciniphila 40

Bacteroides fragilis 41

Escherichia coli 42

Ruminococcus gnavus 43

3.2. Microbial metabolites in IBD 45

Short-chain fatty acids 46

Tryptophan 47

Bile acids 48

1. IBD

Inflammatory bowel disease (IBD) encompasses a group of non-infectious chronic inflammatory disorders primarily affecting the gastrointestinal tract (GIT). Its primary constituents are Crohn's disease (CD) and ulcerative colitis (UC) (Sairenji *et al.*, 2017). However, around 7-10% of patients show clear signs of IBD without definite features of CD or UC, which are termed IBD unclassified (IBDU) or colitis of uncertain type or etiology (CUTE) (Geboes *et al.*, 2008). Crohn's disease can affect any segment of the gastrointestinal tract (GIT), and is characterized by a discontinuous pattern of transmural inflammation that may impact other organs through fistulation (Wilkins *et al.*, 2011). In contrast, UC primarily involves a diffuse and continuous mucosal inflammation, typically beginning in the rectum and potentially spreading proximally (*Figure 1*) (Segal *et al.*, 2021). These relapsing-remitting conditions, CD in particular, display significant heterogeneity with distinct underlying mechanisms contributing to a similar clinical presentation which is characterized by abdominal pain, chronic diarrhoea, rectal bleeding, and weight loss, frequently accompanied by systemic manifestations like fatigue, malnutrition, fever, and anaemia (Knowles *et al.*, 2018). In 25% to 40% of patients with IBD the disease causes extra-intestinal manifestations, which theoretically can affect any organ but primarily involve eyes, liver, skin and joints (Sairenji *et al.*, 2017). After 10 years of disease, patients with IBD have a significantly increased colorectal cancer risk compared to the normal population (Porter *et al.*, 2021) and while the cumulative mortality of patients with UC is comparable with that of the population without IBD, the mortality of patients with CD is twice as high (Aniwan *et al.*, 2018, Bitton *et al.*, 2016). These diseases can manifest at any age, but it most commonly emerges in early adulthood, with a peak incidence occurring between the second and fourth decade of life and a second smaller peak in the elderly (Thia *et al.*, 2008, Molodecky *et al.*, 2012). However, an increase of paediatric-onset IBD, is defined as disease onset before the age of 18 (Benchimol *et al.*, 2014).

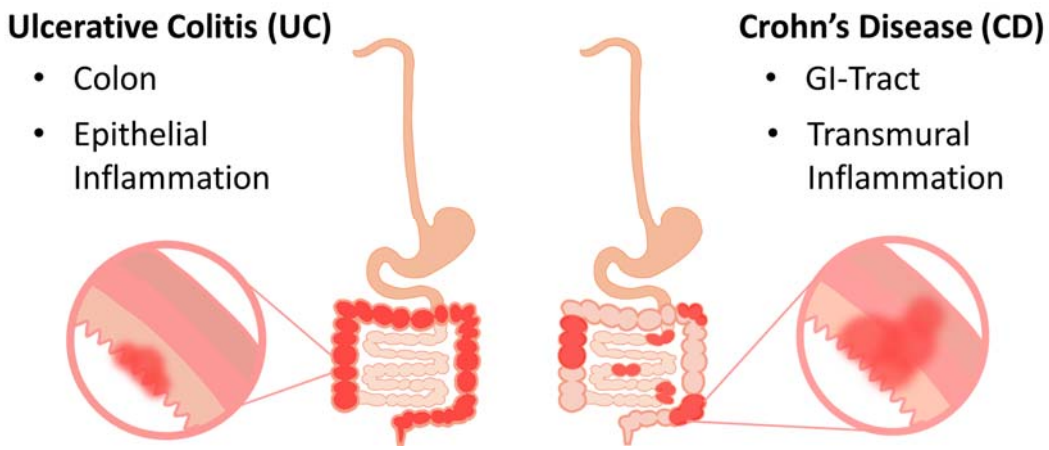


Figure 1: IBD is a chronic relapsing inflammatory condition encompassing two distinct disorders with similar clinical presentation but different phenotype. Ulcerative colitis is characterized by a continuous inflammation that is usually constrained to the innermost lining of the colonic mucosa. Crohn's disease on the other hand can occur along the whole digestive tract and is typified by patchy inflammation that also involves deeper layers of the affected tissues.

1.1. Epidemiology

Epidemiological data indicate that in Western countries CD is male dominated in childhood and adolescence, while it is slightly more prevalent in females in adulthood. On the other hand, there are no differences in UC incidence rates until 45 years of age, from which point onwards there is a male dominance (Shah *et al.*, 2018). In Asian-Pacific cohorts, the incidence of CD and UC is predominantly male from adolescence to middle/late-middle age (Shah *et al.*, 2019) which highlights the for need future IBD studies to consider gender-specific disease modifiers (Chan, 2019). Variations in disease phenotype and outcomes have also been documented among different racial groups, emphasizing the importance of understanding how genetics and environmental factors interact within diverse populations (Barnes *et al.*, 2021). For instance, individuals of Ashkenazi Jewish descent, have a notably higher risk of developing IBD (Bernstein *et al.*, 2006).

In the past, IBD was primarily associated with the so-called Western world. However, recent studies spanning the last two decades have revealed a significant surge in its incidence in newly industrialized regions across the Middle East, Asia, and South America. Conversely, IBD incidence rates in the West appear to have reached a plateau during the 21st century (Ng *et al.*, 2017). In 2019, the global burden of IBD was substantial, with an estimated 4.9 million cases worldwide. A recent study that examined IBD incidence rates showed that 147 out of 204 countries and territories

noted a significant increase in age-standardised prevalence rates, which correlated with a higher socio-demographic index (Wang *et al.*, 2023). The direct medical cost associated with IBD was calculated to exceed US$6.8 billion in the US per annum (M'Koma, 2022), while the annual cost in Canada are estimated to be CDN$1.2 billion (Rocchi *et al.*, 2012) and in Europe €4.6–5.6 billion (Burisch *et al.*, 2013). Nevertheless, it is noteworthy that indirect healthcare costs, including reduced work productivity or mental health issues, are reported to surpass the direct financial burdens of IBD (Kaplan, 2015, van Gennep *et al.*, 2021, Sciberras *et al.*, 2022).

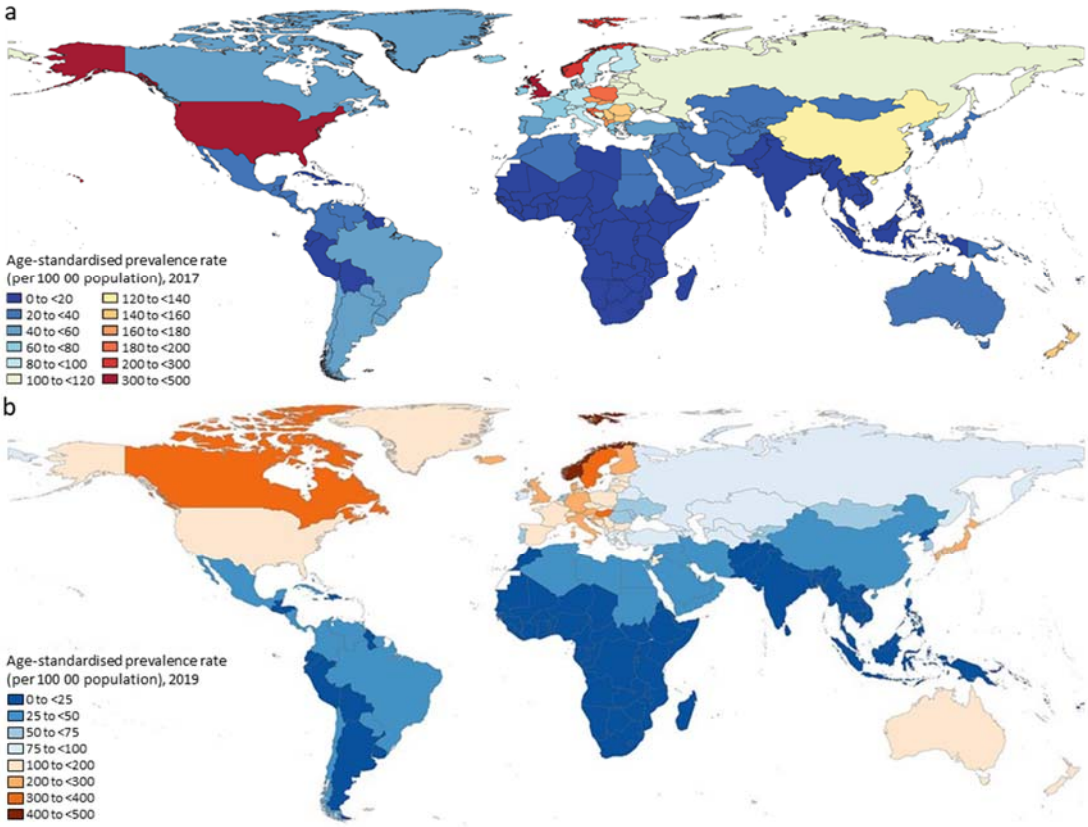


Figure 2: Changes in IBD prevalence between 2017 and 2019. While incidence rates in established 'Western' regions appear to have stabilized, emerging industrialized regions across the Middle East, Asia, and South America have witnessed a notable surge in age-standardized prevalence rates (Figure adapted from (Ng et al., 2017) and (Wang et al., 2023)).

1.2. Diagnosis

The diagnostic landscape for IBD currently lacks a standardized tool, necessitating a multifaceted approach for accurate evaluation which can be quite complex and lengthy (*Figure 3a*). A Swiss study from 2012 reported that only 75% patients with IBDU received the correct diagnosis after 6 months, while patients with UC needed 12 months and patients with CD waited as long as 24 months (Vavricka *et al.*, 2012).

A more recent study that surveyed 19,555 cases of IBD and 78,114 controls found that less than 50% of patients received specialist care within 18 months of presenting with chronic gastro-intestinal (GI) symptoms and that 10% of patients had symptoms for as long as 10 years before IBD diagnosis (Blackwell *et al.*, 2020). Considering the significant impact of untreated IBD on patients' quality of life (Ghosh *et al.*, 2007), as well as the association between delayed diagnosis and an increased risk of complications and the need for subsequent bowel resection, there is a pressing need to develop more accessible methods for diagnosing and monitoring IBD (Thia *et al.*, 2010, Schoepfer *et al.*, 2013).

The state-of-the-art diagnostic process integrates clinical, endoscopic, radiological, histological, and biochemical assessments (Gomollón *et al.*, 2017). The aim is not only to diagnose IBD, but also to distinguish between UC, CD, and other diseases such as colorectal cancer or coeliac disease, as well as to establish the extent, activity and site of the disease. Patients presenting with symptoms such as abdominal pain and diarrhoea, after excluding possible viral or bacterial infections, may undergo diagnostic tests that reveal anaemia and an elevated platelet count in their blood morphology, suggesting potential IBD (Rampton *et al.*, 2020). For those undergoing evaluation, a negative result for endomysial or transglutaminase antibodies typically excludes celiac disease (Bernstein *et al.*, 2016). Elevated levels of C-reactive protein (CRP) and reduced serum albumin levels are indicative of active disease in individuals with established disease, as well as suggestive of IBD in patients that haven’t been diagnosed yet. (Wilkins *et al.*, 2011, Segal *et al.*, 2021). CRP is primarily synthesized in the liver by hepatocytes and is triggered by pro-inflammatory cytokines at the site of inflammation. Due to its short half-life, it serves as a valuable indicator for detecting and monitoring disease activity in CD. In contrast, UC often shows a limited or negligible CRP response, even with active inflammation (Vermeire *et al.*, 2004). Another marker that serves as a reliable indicator of intestinal inflammatory activity and aids in monitoring disease activity in established IBD patients is faecal calprotectin (fCAL). Conversely, in newly presenting cases with diarrhoea, a normal fCAL result effectively rules out IBD (Ma *et al.*, 2019). Calprotectin constitutes a major portion of neutrophil cytosolic proteins, with lower amounts found in monocytes and

macrophages (Johne *et al.*, 1997). It functions as a positive acute phase protein, influencing neutrophil migration, adhesion, and phagocytosis, facilitating chemotaxis, and contributing to innate immunity. The presence of calprotectin in faeces occurs as a result of neutrophil migration into intestinal tissues during inflammation (Pathirana *et al.*, 2018). Notably, neither CRP nor fCAL are specific to IBD, necessitating more invasive approaches for the definitive diagnosis of IBD (Chen *et al.*, 2020).

The current gold standard for a definitive diagnose of IBD as well as disease activity monitoring comprises of colonoscopy and endoscopy with biopsies, which provide direct visualization and tissue samples (Parigi *et al.*, 2022). However, as the procedures are costly, burdensome for the patient and require a specialist to perform (Moran *et al.*, 2016), the diagnostic landscape has recently expanded to encompass supplementary methodologies. These include video capsule endoscopy (Limpias Kamiya *et al.*, 2022), magnetic resonance enterography, and computed tomography enterography (Shaban *et al.*, 2022). Notably, the diagnostic images generated by these modalities can be harnessed by advanced technologies such as machine learning (ML) and deep learning (DL), i.e. a subset of ML employing artificial neural networks with multiple layers. These cutting-edge computational models hold the potential to guide the management and treatment strategies for individuals affected by IBD (Takenaka *et al.*, 2022).

1.3. Treatment and Management

While there is currently no cure, the management of IBD has improved substantially, owing to new formulations and better use of conventional drugs, such as corticosteroids, amino salicylates, antibiotics, and immunomodulatory agents (*Figure 3b*). For severe and treatment resistant cases the introduction of biologics i.e. drugs produced from and/or containing components of living organisms, such as proteins, was a major advancement, reducing both hospitalisation and surgery rates (Dulai *et al.*, 2016b, Gomollón *et al.*, 2017).

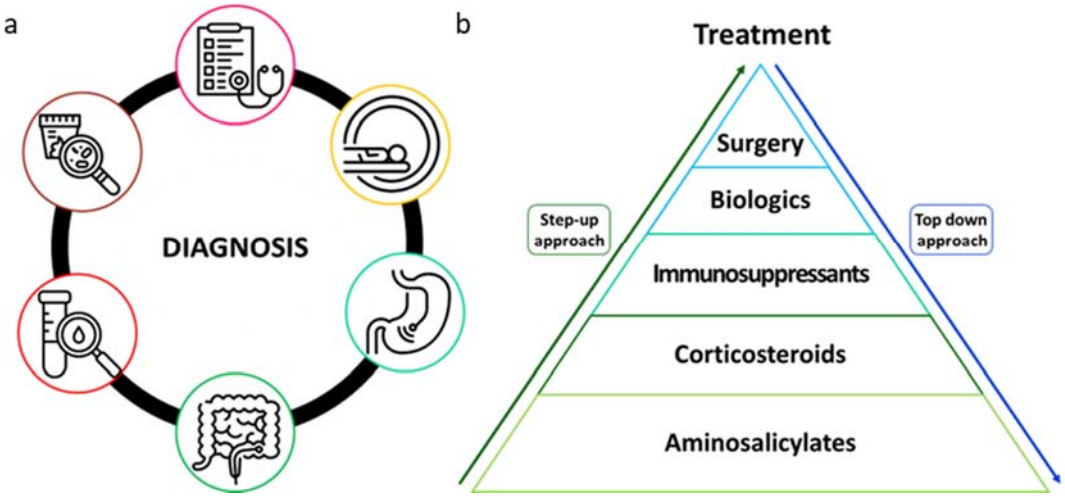


Figure 3: Diagnosis and treatment strategies for IBD. a) Diagnosis typically requires a combination of tools and tests, as no single test can definitively diagnose or rule out the conditions. b) Treatment pyramid: Mild inflammation is initially managed with antibiotics or aminosalicylates, while escalating disease severity necessitates stronger medications (step-up approach). Recent studies propose a top-down approach to potentially mitigate bowel damage.

The primary goals in managing IBD are to achieve full clinical and endoscopic remission, alleviate symptoms, and prevent disease progression and complications like strictures and colorectal cancer. Simultaneously, the aim is to improve the patient's overall quality of life (Peyrin-Biroulet *et al.*, 2015). In severe cases or when complications arise, surgical interventions may be necessary to remove affected sections of the gastrointestinal tract (Gomollón *et al.*, 2017).

Corticosteroids, known for their rapid action, are typically used to induce initial remission and manage acute flare-ups in both UC and CD. They work by inhibiting the production of pro-inflammatory molecules and immune cells. However, corticosteroids are generally unsuitable for long-term maintenance therapy due to their potential for serious side effects, including osteoporosis, diabetes, weight gain, mood disturbances, and increased susceptibility to infections (Hussenbux *et al.*, 2021). In the paediatric population, a specialized dietary therapy called exclusive enteral nutrition (EEN) is preferred over corticosteroids as the first-line therapy. It involves consuming a liquid formula diet as the sole source of nutrition and has shown rapid clinical and biochemical responses (Banerjee *et al.*, 2018), with 85% of paediatric patients achieving remission (Rubio *et al.*, 2011). However, it is less widely used for adult patients due to tolerability and adherence problems (Hansen *et al.*, 2018).

Anti-inflammatory 5-aminosalicylic acid (5-ASA) drugs are suitable for treating moderately active UC or preventing relapse in patients with inactive disease. These drugs have fewer systemic side effects compared to systemic corticosteroids, making them a viable long-term option for UC management. Various formulations are available, differing in their intestinal release profiles (Rampton *et al.*, 2020). Thiopurines, a class of immunomodulatory drugs, play a valuable role in managing both steroid-dependent and steroid-refractory CD and UC, as well as fistulating CD. They work by inhibiting DNA synthesis in T cells, and inducing apoptosis of the same. However, side effects, including nausea, vomiting, headaches, increased infection risk, pancreatitis, and bone marrow suppression, may lead to discontinuation in up to a third of patients (Singh *et al.*, 2022, Hanauer *et al.*, 2019).

In severe or refractory cases of IBD, the therapeutic arsenal includes biologics, a class of medications encompassing various antibodies and agents designed to counteract abnormal cytokine expression and impede the migration of immune cells into the gut wall. Currently, three major categories of biologic therapies are available for managing IBD. Anti-TNFα agents, such as infliximab and adalimumab, function by binding to both free and surface-bound TNFα on activated T cells, instigating their apoptosis (Juillerat *et al.*, 2022). Vedolizumab, on the other hand, targets the α4β7 integrin, effectively impeding immune cell migration within the gastrointestinal tract (Sandborn *et al.*, 2013). Ustekinumab operates on the p40 subunit of pro-inflammatory cytokines, notably IL-12 and IL-23, thus curtailing the activation of T lymphocytes. Meanwhile, Tofacitinib, a Janus kinase inhibitor, disrupts intracellular signalling pathways involved in the perpetuation of inflammatory processes (Sandborn *et al.*, 2017). In addition to these, biologics with novel targets are emerging, such as monoclonal antibodies targeting the β7 subunit of integrins α4β7 and αEβ7 (Etrolizumab) (Danese *et al.*, 2022), or the p19 subunit of interleukin 23 (Guselkumab) (Sandborn *et al.*, 2022), further expanding the array of potential treatments in development for IBD management (Ungar *et al.*, 2016).

Antibiotics are occasionally used to manage infections and complications in IBD patients. Metronidazole with or without ciprofloxacin, is employed to reduce the risk of recurrence in CD and is also used in UC patients with pouchitis post-colectomy

(Nitzan *et al.*, 2016). Broad-spectrum antibiotics may be reserved for cases with severe colitis, while antibiotics such as amoxicillin, trimethoprim, or ciprofloxacin may address issues like fatty stools or diarrhoea due to bacterial overgrowth (Nitzan *et al.*, 2016, Rampton *et al.*, 2020). Additionally, there is an increasing use of complementary and alternative medicine (CAM), with estimated numbers ranging from 21-60% of patients (Lin *et al.*, 2018). CAMs include herbal remedies (Triantafyllidi *et al.*, 2015), traditional Chinese medicines (Yang *et al.*, 2021), and acupuncture (Song *et al.*, 2019a), with limited data suggesting benefit for some IBD patients (Cheifetz *et al.*, 2017).

Our increasing understanding of the pathogenesis provides an expanding arsenal of medications, however, especially biologics come with a hefty price tag (Degli Esposti *et al.*, 2023) and not all patients respond to certain medications or therapies stop working or show significant side effects (Stallmach et al., 2010, Cohen, 2001, Click et al., 2022, Quezada et al., 2018). Application of precision medicine could help improve IBD patient outcome. The aim of precision medicine is to create a tailored therapy for individual patients and to apply the most suitable treatment option based on patient characteristics. However, only few biomarkers are currently available to lead the choice or dosing of treatment. A recent study found evidence that copies of the HLA-DQA1*05 allele, carried by approximately 40% of Europeans, significantly increase the rate of immunogenicity of anti-TNFα therapies, which could be predictive of treatment failure (Sazonovs *et al.*, 2020). Another example of a biomarker for guided individual dosing is the measurement of thiopurine methyltransferase (TPMT; enzyme responsible for the metabolism of AZA and MP, homozygous deficiency occurs in about 0.2% of the population) levels prior to starting therapy with azathioprine (AZA) or mercaptopurine (MP). The majority of cases of TPMT-deficiency are associated with three alleles, TPMT*2, TPMT*3A and TPMT*3C. Individuals with complete TPMT deficiency who receive standard doses of AZA or MP are highly likely to develop severe and potentially fatal myelosuppression. Heterozygotes are also at risk of toxicity at standard doses but this is prevented by initiation at 50% of the conventional dose (Hawwa *et al.*, 2008, Warner *et al.*, 2018). The Increasing awareness of the microbial influence of the gut microbiota on both

the development of the disease and its bidirectional interaction with drugs highlights the microbiota as a promising target. In particular baseline microbiota has been observed to influence the response to treatments (Rashidi *et al.*, 2021, Zhang *et al.*, 2015, Pouncey *et al.*, 2018).

1.4. Surgery

Surgery plays a significant role in the management of IBD, even though surgery rates have decreased in recent decades for both UC and CD (Frolkis *et al.*, 2013, Burisch *et al.*, 2013) due to advancements in medical therapies, including biologics and immunomodulators (Ungar *et al.*, 2016). Surgery in IBD may still be necessary for various reasons, such as severe disease that does not respond to medical treatment, complications like strictures, abscesses, or fistulae, and to manage certain extra-intestinal manifestations (Kalman, 2021, DeLeon *et al.*, 2022). In both UC and CD, the long-term risk of colorectal cancer increases, especially in those with extensive and long-standing disease. Surgical removal of the colon (colectomy) may be recommended in some high-risk UC patients as a preventive measure against the development of cancer (Mattar *et al.*, 2011, Dulai *et al.*, 2016a).

The incidence of surgery varies depending on several factors, including the type of IBD, disease severity, and individual patient characteristics. Despite surgery offering a long-term solution for UC, and the risk of recurrence of CD after resection, surgery is more common in patients with CD than in UC. Approximately 10-30% of patients with UC will eventually require surgery during their lifetime (Andersson *et al.*, 2009). Surgery often involves a total proctocolectomy (removal of colon and rectum) followed by, ileostomy or an internal ileal pouch. Surgery is considered a cure for UC as the disease is limited to the colon and thus cannot flare-up again after its removal, even though some extra-intestinal symptoms may remain (Bernstein *et al.*, 2010). However, the potential need for living with a stoma, long-term follow-up care, and complications occurring in a significant amount of patients need to be balanced when considering surgery (Peyrin-Biroulet *et al.*, 2016, Fradet *et al.*, 2020). Additionally major mal-effects such as depression, body image and sexual function issues have been reported after surgery (Brown *et al.*, 2015a). In comparison, a significant

proportion of patients with CD, about 70-75%, will undergo surgery at some point in their life (Bernell *et al.*, 2000). As CD can occur anywhere along the GI tract, complete removal is not an option, and thus surgery is usually not curative. Common surgical procedures for CD include drainage of abscesses, strictureplasty (widening of narrowed segments of the intestine) and bowel resection (removing diseased portions of the intestine) (Strong *et al.*, 2015). Surgery can greatly improve a patient's quality of life by alleviating symptoms and reducing medication dependency, but as postoperative recurrence rates in CD are high, about half of patients need repeat surgery after 10 years (Gklavas *et al.*, 2017).

2. Pathogenesis of IBD

While there is progress in disentangling the pathogenesis of this disease, the exact cause of IBD remains largely unknown. As with other chronic inflammatory disorders, the tissue damage is immune mediated and arises from an interaction of genetic susceptibility factors, environmental factors and indigenous gut microbiota (*Figure* 4).

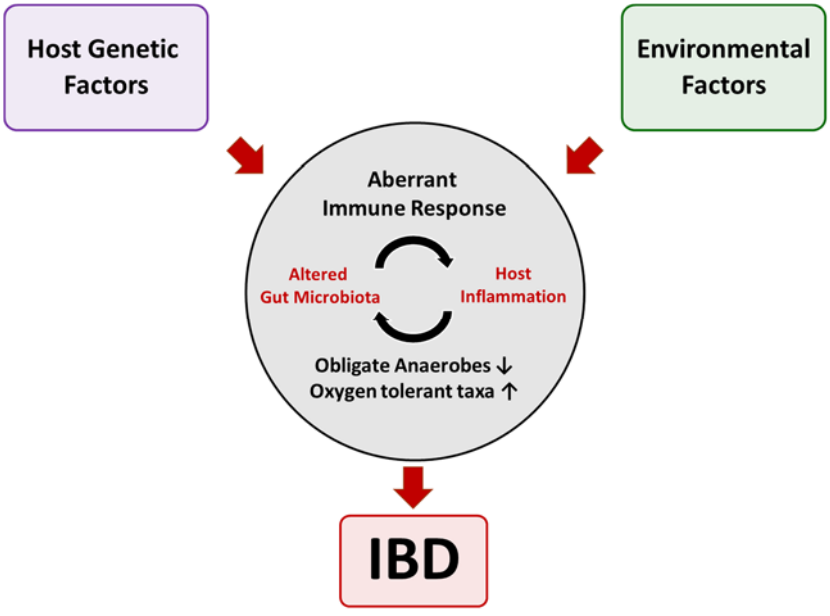


Figure 4: Pathogenesis of IBD. The pathogenesis of IBD involves a multifaceted interplay of genetic, microbial, and environmental factors that contribute to the disruption of the intestinal barrier. This disruption in mucosal integrity initiates a complex cycle, which ultimately exacerbates and perpetuates the progression of IBD.

2.1. Genetic Risk Factors

Support for a role of genetic factors in IBD were originally derived from twin studies and family-based studies of disease clustering. Between 4% and 14% of patients have a family history of CD and 8% to 14% of patients with UC. Furthermore, concordance rates for CD and UC in monozygotic twins are higher (20-50% and 16%, respectively) than in dizygotic twins (10% and 4%) (Ananthakrishnan, 2015, Moller *et al.*, 2015). Meanwhile, Genome-wide association studies (GWAS) identified more than 240 non overlapping genetic susceptibility loci. The majority of these are shared between the IBD subtypes, implying comparable susceptibility and mutual signalling pathways. It is important to note that no single risk gene is both necessary and sufficient to cause IBD, indicating a complex multifactorial aetiology, which polygenetic risk scores might shed new light on (Sun *et al.*, 2023). Given this heterogeneous complexity and the observation that genetic determinants yields are relatively modest, IBD is thought to align with the common variant hypothesis which posits that the condition arises from cumulative effect of multiple common genetic risk factors (de Lange *et al.*, 2017, Turpin *et al.*, 2018, Mirkov *et al.*, 2017). A substantial portion of these genetic risk factors revealed strong associations with immune function, and specifically with the immune system's interaction with the microbiota such as pathways for epithelial barrier function, innate mucosal defence and autophagy (Aschard *et al.*, 2019, Jostins *et al.*, 2012, Liu *et al.*, 2015, Lavoie *et al.*, 2019).

Mutations in the nucleotide-binding oligomerization domain 2 (NOD2) are the strongest genetic risk factor for IBD. They occur in about a third of patients with CD (Strober *et al.*, 2006) and carrying one of the three main SNPs in leucine rich repeats, increases the risk for developing CD 2-4 fold and up to 40 fold if individuals carry two or more of the same NOD2 variants (Horowitz *et al.*, 2021). The role of NOD2 variants with UC is less clear, but there is increasing evidence that they also play a role in this IBD subtype (Liu *et al.*, 2022). NOD2 belongs to the cytosolic Nod-like receptor (NLR) family, which plays a critical role in detecting microbial invaders and initiating immune responses (Eckmann *et al.*, 2005). Unlike the membrane-bound toll-like receptors (TLRs), NOD2 is located in the cytoplasmic compartment. The encoded protein functions as an intracellular sensor of bacterial components, specifically

recognizing muramyl dipeptide (MDP), a fragment of peptidoglycans found in the cell walls of both Gram-negative and Gram-positive bacteria (Strober *et al.*, 2011). Upon binding on MDP, NOD2 forms an active oligomer that recruits adaptor proteins to initiate a downstream signalling cascade that activates nuclear factor κB (NFκB). This ultimately results in transcription and production of pro-inflammatory cytokines, or in association with ATG16L1 in stimulation of autophagy (Strober *et al.*, 2011). In IBD, three of the most frequently observed NOD2 variants, namely R702W, G908R, and 1007fs, are located within the leucine-rich repeat domain responsible for binding to MDP (Lesage *et al.*, 2002). This suggests that the increased risk of IBD may be attributed to defective bacterial sensing (Eckmann *et al.*, 2005). Moreover, NOD2 variants have been linked to a decreased production of α-defensins and antimicrobial peptides in Paneth cells (Wehkamp *et al.*, 2004), potentially due to the resulting abnormal morphology of Paneth cells observed in CD patients (Lala *et al.*, 2003). Autophagy is a cellular recycling process involving the degradation of cytosolic components, organelles, and intracellular microbes (Kaser *et al.*, 2011). Stimulation by MDP triggers this process through the interaction of ATG16L1 and NOD2. The common T300A mutation in the ATG16L1 gene increases susceptibility to caspase-3 cleavage, leading to reduced autophagic function as well as increased cytokine release, and decreased intracellular bacterial clearance (Cohen *et al.*, 2019, Lavoie *et al.*, 2019). Although the majority of risk-associated loci are common across populations, it's noteworthy that while NOD2 variants are prevalent among most Europeans, they are notably absent in East Asian populations (Liu *et al.*, 2015).

Despite the substantial number of single nucleotide polymorphisms (SNPs) associated with the disease phenotype, they can only account for approximately 25% of the observed heritability (Khor *et al.*, 2011, Loddo *et al.*, 2015), implying the involvement of other contributing factors.

2.2. Immunological abnormalities

The gut serves as a crucial immunological interface, maintaining a balance by appropriately recognizing and tolerating commensal bacteria, dietary antigens, and self-antigens (a tolerogenic response). Simultaneously, it identifies and responds to harmful pathogens (an immunogenic response) (Xavier *et al.*, 2007). Dysregulated

immune responses within the gut disrupt the balance between pro-inflammatory and anti-inflammatory pathways in both innate (immediate, nonspecific defence against pathogens) and adaptive (highly specific response against pathogens) immunity. These disruptions are believed to be central to the development and persistence of inflammation in IBD (Park *et al.*, 2017). Contributing factors include compromised epithelial barrier function, neutrophil accumulation in the epithelium, reduced antigen clearance by macrophages, and impaired conditioning of dendritic cells within the innate immune system (de Souza *et al.*, 2016).

The mucus layer and intestinal epithelium segregate the gut lumen from the host, forming both a physical and a biochemical barrier against bacteria and food antigens (Peterson *et al.*, 2014). In the context of IBD, mucosal barrier abnormalities that lead to heightened permeability have been observed (Kaistha *et al.*, 2014). These abnormalities including impaired expression of junctional proteins, which are crucial for maintaining epithelial tightness and stability, as well as antimicrobial peptides that play a role in nonspecific mucosal defence (Wehkamp *et al.*, 2003, Gassler *et al.*, 2001). Neutrophils, comprising 50-70% of leukocytes, are vital in the immediate immune response at infection sites, where they phagocytose pathogens and also generate reactive oxygen species (ROS) with antimicrobial properties. Besides limiting microbial invasion, they promote the recruitment of other immune cells and facilitate inflammation resolution via mediator release essential for mucosal healing (Fournier *et al.*, 2012). Dysregulated control can trigger harmful positive feedback loops that amplify neutrophil activation and lead to tissue damage (Wéra *et al.*, 2016). Neutrophil roles in IBD may vary between CD and UC (Fournier *et al.*, 2012); in UC, overactive neutrophils contribute to the disease (Zhang *et al.*, 2023a), while in CD, neutrophil dysfunction may hinder microbe elimination and result in uncontrolled secondary immune responses (Vazeille *et al.*, 2015). Macrophages, the main tissue mononuclear phagocytes, play a central role in initiating and resolving intestinal inflammation by efficiently engulfing and breaking down foreign substances. They originate from circulating monocytes or resident intestinal macrophages and possess two distinct activation states: pro-inflammatory (M1-like) and anti-inflammatory (M2-like) (Han *et al.*, 2021). Under normal conditions,

monocytes mature into M2-like macrophages, which promote anti-inflammatory cytokine secretion (e.g., IL-10 and TGF-β), exhibit hypo-responsiveness to TLR stimulation, aid the clearance of apoptotic cells, and promote epithelial stem cell renewal (Pan *et al.*, 2022). In IBD, this M2 differentiation is disrupted. Monocytes are drawn to inflamed areas, where they differentiate into M1-like macrophages that secrete pro-inflammatory cytokines like IL-12, IL-23, and IL-1β, leading to Th-1 and Th-17 immune responses and epithelial damage (Yip *et al. 2022,* Isidro *et al.*, 2016).

Dendritic cells (DCs) play a crucial role in mediating communication between innate and adaptive immunity, influencing whether T and B cells develop tolerance or become immunogenic, thus affecting the balance between mucosal homeostasis and inflammation (Rescigno *et al.*, 2009). Intestinal epithelial cells (IEC) respond to signals from commensal bacteria by producing substances like retinoic acid, thymic stromal lymphopoietin (TSLP), and transforming growth factor beta (TGF-β), which promote the development of tolerogenic DCs (Peterson *et al.*, 2014). These conditioned DCs, when exposed to Th-1-inducing pathogens, stimulate the differentiation of naive CD4+ T cells into T_{REG} cells and the maturation of B cells into IgA-secreting plasma cells (Rimoldi *et al.*, 2005). However, in about 70% of CD patients, epithelial cells do not express TSLP, preventing the induction of anti-inflammatory DCs which leads to the release of IL-12, driving Th1 responses to bacteria (Rimoldi *et al.*, 2005).

Antigen presentation by immune cells like DCs, macrophages, and B cells triggers the maturation of naïve T helper cells, which then differentiate into Th-1, Th-2 or Th-17. Th-1 cells release IFN-γ and combat intracellular pathogens, while Th-2 cells mainly assist B cells by producing IL-4. Th-17 cells defend against bacteria and fungi but can worsen inflammation under autoimmune conditions while T_{REG} inhibit effector T cell function and counteract inflammatory responses (Geremia *et al.*, 2014). Maintaining a balanced T cell profile is crucial for immune system homeostasis. In IBD, aberrantly activated T cells release excessive cytokines and chemokines such as Tumour Necrosis Factor-α (TNF-α), Interferon-γ (IFN-γ), Interleukin 1β (IL-1β), IL-6, IL-9, IL-12, IL-17, and IL-23, affecting both innate and adaptive immunity (Choy *et al.*, 2017).

2.3. Environmental triggers

Epidemiological evidence strongly suggests a significant impact of environmental factors on IBD incidence. Higher incidence rates are observed in developed countries, whereas there is a notable surge in IBD incidence in developing nations within just one or two generations following socioeconomic development (Ng *et al.*, 2017). Furthermore, children of migrants into developing nations face a greater risk of developing IBD compared to non-migrant children, with the risk increasing the earlier they left their home country (Brant, 2013). In fact, migrants to developed countries assume a similar risk of developing IBD as the native population, irrespective of their ethnicity (Kaplan, 2015). This transition is accompanied by significant microbial changes, marked by a reduction in microbial diversity, a decline in ancestral microorganisms, and the loss of enzymes responsible for plant degradation (Vangay *et al.*, 2018). Potential environmental factors associated with industrialization encompass urbanization, shifts in dietary habits, increased antibiotic usage, modifications in hygiene practices, and heightened environmental pollution (Ananthakrishnan, 2015). These factors have all demonstrated their capacity to influence the gut microbiome and given the pivotal role of the gut microbiota in the development and progression of inflammation in IBD (Kostic *et al.*, 2014), it underscores the idea that environmental factors and their interactions with microbiota have a substantial impact on the pathogenesis of IBD.

The plasticity of the gut microbiota is largest at birth and is becoming continuously less malleable with age (Thriene *et al.*, 2023). The indigenous microbiota gradually forming during the initial three years of life, coinciding with the maturation of the immune system i.e. the process of developing its ability to accurately sense and respond to environmental threats (Derrien *et al.*, 2019, Gensollen *et al.*, 2016). Disturbances in the microbiota, particularly in early life, can potentially disrupt the maturation of the immune system and predispose individuals to immune-related issues triggered by microbes (Tamburini *et al.*, 2016). Consequently, exposure to antibiotics, which have been found to devastate the gut microbiome (Modi *et al.*, 2014, Patangia *et al.*, 2022), especially if recurrent in the first year of life, has been associated with an increased likelihood of developing IBD. For instance, Shaw and

colleagues' research indicated that children diagnosed with IBD had a higher history of early-life antibiotic use when compared to control groups (Shaw *et al.*, 2010). Moreover, those who received antibiotics until the age of 5 exhibited a threefold higher risk of developing IBD than children that weren't exposed (Shaw *et al.*, 2013). Adults diagnosed with IBD were also more likely to have been prescribed antibiotics 2–5 years before their diagnosis (Shaw *et al.*, 2011). An UK study observed a dose-response relationship between childhood antibiotic exposure and the development of IBD, with this association diminishing as the age of exposure increased (Kronman *et al.*, 2012). A meta-analysis of 11 studies revealed a significant link between antibiotic exposure and CD, but this relationship was not significant for UC and this relationship was more marked in children (Ungaro *et al.*, 2014). Curiously, a large population-based analysis conducted in Asia contradicted these findings and instead suggested a protective effect against IBD with antibiotic exposure (Lok *et al.*, 2008).

Breastfeeding is another influential early-life factor that significantly shapes the development of both the gut microbiome and the immune system (Zuurveld *et al.*, 2020). Human breast milk contains a diverse array of structurally distinct human milk oligosaccharides and milk microbiota, which possess pre- and probiotic properties, impacting the assembly of the microbial community (Moossavi *et al.*, 2018).These components enhance both innate and adaptive immunity by engaging directly with pathogens or interacting with the intestinal epithelium and immune cells (Yu *et al.*, 2016, Kurakevich *et al.*, 2013, He *et al.*, 2016). Although the relationship between breastfeeding and chronic non-communicable diseases remains a topic of debate, an increasing body of evidence suggests that breastfeeding may play a protective role (Kelishadi *et al.*, 2014). A potential connection between breastfeeding and the incidence of UC was recognised as early as 1961 when a consistently higher proportion of children, that were weaned within the first two weeks of their life, had UC when stratified by age, social class or family rank (Acheson *et al.*, 1961). In 2004, a meta-analysis of 17 relevant articles showed a clear inverse association between breastfeeding and the occurrence of both CD and UC (Klement *et al.*, 2004). This finding was supported by a 2009 systematic review conducted by Barclay, that analysed 79 articles focusing on paediatric IBD, which suggested a protective effect

of breastfeeding against early-onset IBD (Barclay *et al.*, 2009). Additionally, another systematic review by Xu and colleagues reported a dose-dependent relationship, indicating that breastfeeding for at least 12 months provided the strongest protection against the development of IBD when compared to shorter durations of 3 or 6 months (Xu *et al.*, 2017).

In later life the use of various non-antibiotic medications also has been implicated with the risk of developing or exacerbating IBD. Nonsteroidal anti-inflammatory drugs (NSAIDs) are among the most frequently used drugs (Singh, 2000). Gastrointestinal toxicity is a well-recognized problem, with around 25% of all reported adverse drug reactions being attributed to prescription NSAID use, with GI irritation being the most commonly reported one (Hui *et al.*, 2004, Schoenfeld *et al.*, 1999). NSAID-induced GI injury can be divided into 3 tiers: indigestion, asymptomatic ulceration, and more serious complications such as GI bleeding, perforation and symptomatic ulceration (Cryer, 2004). NSAIDs are thought to induce gastrointestinal harm by either directly harming the gastrointestinal mucosa, or systemically, by reducing prostaglandin production (Schoenfeld *et al.*, 1999). Multiple studies have explored the connection between NSAID usage and its potential impact on the development or progression of IBD. Gleeson and David reported a notably high occurrence of prior NSAID or salicylate exposure among new patients with colitis (Gleeson *et al.*, 2003). A 2012 study found an elevated risk of both UC and CD in frequent NSAID users (>15 days/month) and those with prolonged NSAID use (>6 years) compared to non-users (Ananthakrishnan *et al.*, 2012a). Conversely, another investigation focusing on patients with existing CD or UC revealed associations with active disease in patients with CD, particularly with more frequent NSAID use (>5 days/month), but not in UC cases (Long *et al.*, 2016). Nonselective NSAIDs were associated with a 17%-28% relapse rate within 9 days, coinciding with heightened intestinal inflammatory activity (Takeuchi *et al.*, 2006). On the contrary, low-dose NSAID usage did not exhibit a notable increase in disease activity among outpatients with CD or UC. However, high-dose NSAID use was linked to higher disease activity index scores among CD patients with colonic involvement, although this did not translate into a rise in significant disease flares (Bonner *et al.*, 2004). Interestingly,

since the 1960s, it has been known that NSAID use can lead to significant shifts in gut microbial composition (Kent *et al.*, 1969). Some of these changes may be attributed to reactions to inflammatory processes (Zeng *et al.*, 2017), but certain studies have shown alterations in gut microbiota even in the absence of macroscopic or histological damage or inflammation (Hutka *et al.*, 2021).

Oral contraceptive pills (OCPs) present another example of drugs that have been investigated as potential risk factors for IBD. OCPs are a commonly employed form of contraception. Studies exploring the association between OCP use and the onset of inflammatory bowel disease have yielded conflicting results. Which might be owed to the fact that many of those studies faced limitations such as small sample sizes and a retrospective design, which hindered their ability to account for crucial lifestyle factors (Khalili *et al.*, 2012). However, numerous meta-analyses have consistently revealed a positive association between the use of OCPs and the risk of both CD and UC. In 1995, Godet *et al.* reported a modest link between OCPs and the development of IBD based on an analysis of 9 studies (Godet *et al.*, 1995). Subsequently, Cornish *et al.* extended this observation in 2008, showing that the risk of developing IBD was not only elevated in OCP users but also increased with the duration of exposure, synthesizing findings from 14 studies (Cornish *et al.*, 2008). Furthermore, Ortizo *et al.* confirmed this association in 2017 with data from an analysis of 20 studies (Ortizo *et al.*, 2017). The exact mechanisms underlying these observations are not yet fully elucidated; however, several factors provide biological plausibility to support these findings. Oestrogen, for instance, has demonstrated the capacity to promote macrophage proliferation and strengthen the humoral immune system (Cutolo *et al.*, 2008). Additionally, it participates in altering colonic barrier function (Looijer-van Langen *et al.*, 2011) and might contribute to multifocal gastrointestinal infarction owing to its thrombotic potential (Abou-Ismail *et al.*, 2020). There is also growing evidence suggesting bidirectional interactions between gut microbiota and sex hormones (Flak *et al.*, 2013). For instance, Flores *et al.* associated non-ovarian oestrogen levels with gut microbial richness and diversity (Flores *et al.*, 2012). Additionally, in a systematic review, d'Afflitto *et al.* demonstrated correlations

between sex steroid hormone levels and both gut microbial diversity and composition (d'Afflitto *et al.*, 2022).

Cigarette smoking (CS) is another significant risk factor for gastrointestinal disorders, including peptic ulcers, CD, and several types of cancers. Smoking doubles the risk of developing CD (Mahid *et al.*, 2006), tends to exacerbate the disease's aggressiveness (Cosnes, 2008) and is concomitant with a higher risk of recurrence after surgery (Reese *et al.*, 2008). Paradoxically, CS acts as a protective factor against the onset and progression of UC (Calkins, 1989, Mahid *et al.*, 2006). However, discontinuing smoking increases the risk of developing UC to levels comparable to CD within 2-5 years (Higuchi *et al.*, 2012). And while the elevated risk of UC persists for up to 20 years after smoking cessation (Higuchi *et al.*, 2012), the risk does not remain increased for CD once smoking ceased (Odes *et al.*, 2001). Several mechanisms were proposed to contribute to the negative effects of smoking on the GI system. These include changes in mucus composition, such as increased expression of Muc2, Muc3 (ileum), and Muc4 (colon) (Allais *et al.*, 2016), impaired small intestine barrier function due to villi atrophy and tight junction alterations (Zuo *et al.*, 2014), as well as macrophage-specific immune impairments (O'Leary *et al.*, 2014). Furthermore, CS can hamper ulcer healing due to alterations in microvaculae and impairment of the endothelial growth factor pathway, which impacts epithelial cell activities like proliferation, migration, and mucus production (Zhang *et al.*, 2012). Cigarette smoke contains over 60 carcinogenic and mutagenic compounds and is rich in ROS and free radicals, leading to oxidative stress, tissue damage, and disruption of endothelial function (Berkowitz *et al.*, 2018). The nicotine concentration in gastric fluid has also been reported to be 10 times greater than that in arterial blood and 80 times greater than that in venous blood (Lindell *et al.*, 1993). It is therefore not unexpected that smoking, or the cessation of thereof has been implicated with changes in the gut microbiota. In mice, exposure to smoke led to significant changes in bacterial activity and community structure in the colon (Allais *et al.*, 2016). Additionally, Biedermann *et al.* observed significant changes in the composition of the intestinal microbiota in healthy individuals following smoking cessation. This was characterized by an increase in key representatives from the *Firmicutes* and *Actinobacteria* phyla,

coupled with a decrease in *Bacteroidetes* and *Proteobacteria* (Biedermann *et al.*, 2014). Lee *et al.* also corroborated these trends for *Firmicutes* and *Bacteroidetes* but not *Proteobacteria* (Lee *et al.*, 2018).

Part of the effect of industrialisation and migration on IBD prevalence is believed to be induced by changes in diet. Many studies found associations between dietary intake of macronutrients and IBD, although the directionality of their results do not always match. A meta-analysis of 19 studies found elevated consumption of total fats, polyunsaturated fatty acids (PUFAs), omega-6 fatty acids, and meat was linked to a heightened risk of both CD and UC. Conversely, a diet rich in fibre and fruits was associated with a reduced risk of CD, while increased vegetable intake was linked to a decreased risk of UC. (Hou *et al.*, 2011). A Japanese study found more than a 2-fold increased risk of CD following the consumption of sugars/sweeteners, sweets, fats and oils, and total fat (Sakamoto *et al.*, 2005), while another study didn't find any associations between carbohydrate intake and IBD (Chan *et al.*, 2014b). However, the potential protective role of dietary fibre and fruit was also reported in other studies (Amre *et al.*, 2007, Ananthakrishnan *et al.*, 2013b, Milajerdi *et al.*, 2021). Furthermore, the source of fibres was also found to be important, as the fibre from fruits was associated with a lower CD risk, while the fibre from grains, bran, and cereal did not modify disease risk (Ananthakrishnan *et al.*, 2014). In a 2013 study, the consumption of n-6 PUFA was positively linked to an increased risk of UC, while n-3 PUFAs were inversely correlated with UC risk (Chan *et al.*, 2014a). This inverse correlation is thought to be mediated by the immunomodulatory effects of n-3 PUFAs, which can shift the balance between Th-1 and Th-2 cells by suppressing Th1 cell development (Chapkin *et al.*, 2007). In a large prospective study involving 67,581 middle-aged women, a high intake of total protein, particularly from animal sources, was significantly associated with an increased risk of IBD. Among the various animal protein sources, a high consumption of meat or fish, but not eggs or dairy products, was linked to an elevated risk of IBD (Jantchou *et al.*, 2010). Furthermore, another study found that the consumption of meat, especially red meat, alcohol and sulphated increased the risk the likelihood of a relapse in patients with UC (Jowett *et al.*, 2004).

In addition to macronutrients, micronutrients are acknowledged for their role in supporting the integrity of the gut epithelial barrier and influencing immune responses. Deficiencies in micronutrients are prevalent in over half of individuals with IBD, particularly in those with CD as opposed to UC. The most commonly deficient micronutrients include iron, vitamin B12, vitamin D, vitamin K, vitamin B6, vitamin B1, selenium, and zinc (Weisshof *et al.*, 2015). Zinc serves as a cofactor for various metalloproteins and is an integral part of as much as 10% of the proteome. Up to 50% of patients with IBD exhibit zinc deficiency (Zupo *et al.*, 2022) primarily linked to diarrhoea, surgical procedures, and high-output fistulas (Michielan *et al.*, 2015). This deficiency may play a role in promoting mucosal inflammation and exacerbating colitis in affected individuals by increasing the production of pro-inflammatory cytokines (Wong *et al.*, 2015). Furthermore, low zinc levels have been correlated with an elevated risk of hospitalizations, surgeries, and disease complications (Siva *et al.*, 2017). Zinc levels exhibit an inverse association with the risk of CD, although this is not the case for UC (Ananthakrishnan *et al.*, 2015) and zinc supplementation has been demonstrated to enhance intestinal permeability (Sturniolo *et al.*, 2001). Another candidate for a micronutrient with a role in the pathogenesis of IBD is Vitamin D (VitD), which is well-known for its role in calcium homeostasis and bone metabolism. It also plays a crucial role in preserving the integrity of the gastrointestinal barrier, which is achieved through the regulation of proteins related to epithelial cell gap junctions, monitoring the gut microbiota, and modulating inflammatory immune responses, leading to potent anti-inflammatory effects (Fletcher *et al.*, 2019). Genetic association studies indicate a significant role for VD in the development of IBD (Xue *et al.*, 2013). VitD deficiency is common in people with IBD although it is not established if this is a cause or a consequence of the disease (Cantorna, 2006). Women with elevated VitD levels demonstrated a significantly reduced likelihood of developing CD (Ananthakrishnan *et al.*, 2012b). Conversely, individuals with VitD deficiency faced a higher risk of IBD-related surgery, although this risk diminished after achieving normal VitD levels (Ananthakrishnan *et al.*, 2013a). In murine models, the restoration of VitD levels halted disease progression and alleviated symptoms in mice that had already developed IBD (Cantorna *et al.*, 2000). Furthermore, a meta-analysis encompassing 19 randomized

controlled trials (RCTs) revealed that VitD supplementation significantly lowered the relapse rate compared to the control group. Interestingly, no significant differences were observed between low-dose and high-dose VitD treatments (Li *et al.*, 2018). Vitamin A (VitA) is a fat-soluble component with antioxidant properties that must be obtained from the diet. It has a central role in preserving the integrity of epithelial tissues, supporting vision and growth (Gürbüz *et al.*, 2022, D'Ambrosio *et al.*, 2011). Additionally, it also contributes to the regulation of the immune system by either directly influencing immune cells or indirectly by modulation of the gut microbiota (Barbalho *et al.*, 2019). Retinoic acid, the biologically active form of VitA, supports immune tolerance in collaboration with TGF-β by priming dendritic cells to induce anti-inflammatory T_{REG} cells through the promotion of Foxp3 expression and the suppression of pro-inflammatory Th-17 cell differentiation (Brown *et al.*, 2015b, Bakdash *et al.*, 2015). Vitamin A can also inactivate free radicals which are produced by neutrophils and cause mucosal damage (Ramakrishna *et al.*, 1997). A meta-analysis in 2017 with 19 case-control studies showed lower levels of fat soluble vitamins in patients with IBD compared to controls, as well as significant associations between C-reactive protein and VitA status in patients with CD (Fabisiak *et al.*, 2017). Furthermore, VitA supplementation was shown to decrease disease activity in patients with UC (Masnadi Shirazi *et al.*, 2018).

Macro and micronutrients are not only directly influencing the host, they also serve as substrates and cofactors for a variety physiological processes in our gut microbiota. A lack of these nutrients results in the disruption of functions that are required for commensal bacterial life and the epithelial immune system (Mach *et al.*, 2017, Albenberg *et al.*, 2014). A 2017 study conducted by Hibberd *et al.* provided a clear demonstration of how the absence of a single micronutrient (such as VitA, iron, zinc, or folic acid) can lead to disturbances in both the composition and function of the gut microbiota of gnotobiotic mice, which were colonized with human microbiota. Vitamin A showed the largest effect on bacterial community structure in that study, leading to a marked increase of *Bacteroides vulgatus* (Hibberd *et al.*, 2017). However, every microbial species possesses distinct response mechanisms that lead to specific needs for micronutrients, unique genomic activities, and cellular

metabolism. Consequently, they will exhibit varied responses to different micronutrients. For example, a deficiency in VitA encourages the growth of segmented filamentous bacteria (Mach *et al.*, 2017) while magnesium requirements are greater for Gram-positive bacteria compared to Gram-negative bacteria, which leads to decreased abundances of *Bifidobacterium* in mice fed with a magnesium deficient diet (Mach *et al.*, 2017, Pachikian *et al.*, 2010).

Due to the size differences and the fact that the human gut microbiota coevolved with its host (Bäckhed *et al.*, 2005) and form a relative stable community after reaching adulthood (Thriene *et al.*, 2023), the changes in humans are less marked. However, multiple studies have demonstrated that changes in the dietary habits can rapidly lead to significant alterations in the gut microbial composition (David *et al.*, 2014, Wu *et al.*, 2011, Walker *et al.*, 2011). A number of studies have associated increased microbial richness, at both the taxonomic or functional level, with diets higher in fruits, vegetables, and fibre (De Filippis *et al.*, 2016, Martínez *et al.*, 2013, Maier *et al.*, 2017, Cotillard *et al.*, 2013). Notably, gut microbiota possesses the ability to adapt their substrate utilization based on the available substrates. *Bacteroides thetaiotaomicron* for instance can modulate its gene expression in response to a scarcity of dietary sources of specific carbohydrates, enabling it to primarily metabolize host-derived glycans (Sonnenburg *et al.*, 2005). Moreover, some species have not only developed the ability to thrive in environments with limited micronutrient availability, giving them a competitive advantage, but certain pathogens can even manipulate the host's genetic responses to meet their own nutritional needs. For instance, *Salmonella enterica Serovar Typhimurium* increases the expression of the intestinal antimicrobial peptide lipocalin-2, disrupting bacterial iron acquisition and thereby enhancing their competitive advantage over other bacteria (Ohland *et al.*, 2015).

In light of the fact that many genetic risk factors for IBD involve host-microbiota interactions, tissue damage arises from heightened pro-inflammatory immune responses directed at the gut microbiota, exacerbated by environmental triggers that also induce significant alterations in microbial composition and behaviour, it

becomes evident that the gut microbiota play a central role in the pathogenesis of IBD.

3. Gut Microbiota in IBD

In contemporary microbiology research, there has been a significant paradigm shift, acknowledging that all eukaryotes should not be seen as single autonomous organism, but rather be regarded as meta-organisms or holobiont in recognition of the indispensable connection between eukaryotic organisms and their associated microbiota, forming an inseparable and functional unit (Jones, 2013).

The combined microbial population residing within and on the human body surpasses the number of human cells at an approximate ratio of 1.26 microbial cells to every human cell (Sender *et al.*, 2016). This estimation suggests a total of approximately 3.8x10^13 microbial cells and an estimated 500-1000 species inhabiting the human body. Notably, the colon stands out as the most abundant microbial habitat among them, thanks to its vast surface area and the high nutrient availability (Sender *et al.*, 2016, Turnbaugh *et al.*, 2007). More impressively, the collective genetic material of these microorganisms, known as the microbiome, contains a wealth of genetic diversity and functional potential that greatly surpasses the human genome, underscoring the profound impact and importance of the microbiota within the human ecosystem (Qin *et al.*, 2010).

The concept of a “microbiome”, a term originally coined by Whipps and colleagues in 1988, describes a characteristic microbial community residing within a reasonably well-defined habitat with distinct physio-chemical properties, which serves as the stage for their activities (Whipps *et al.*, 1988). The term microbiome extends to encompass various elements, including structural components (such as nucleic acids, proteins, lipids, and polysaccharides) and metabolites (including signalling molecules, toxins, and organic and inorganic compounds (Berg, 1999). Conversely, microbial communities or “microbiota” are commonly defined as collections of (living) microorganisms coexisting within a shared environment (Konopka, 2009). In the context of the gastrointestinal tract, the assemblage of bacteria, archaea, eukarya and viruses has been termed the "gut microbiota". This complex ecosystem has

evolved in tandem with its host over millennia, establishing an intricate, mutually beneficial relationship that plays a pivotal role in maintaining host health and function (Bäckhed *et al.*, 2005). The normal gut microbiota plays a pivotal role in maintaining host immunologic, hormonal, and metabolic homeostasis, while also providing protection against pathogens (Thursby *et al.*, 2017). Disruptions to the gut microbiota composition and function and their interactions with the host have been associated with a wide range of diseases, including immune-related diseases (Forbes *et al.*, 2016), chronic metabolic diseases (Hernández-Ceballos *et al.*, 2021), and neurodevelopmental diseases (Griffiths *et al.*, 2018).

The composition of the gut microbiota displays spatial variations depending on the location and depth within the GI tract. Notably, the abundance of microbial cells increases progressively from the stomach to the colon. This divergence in microbial communities is a consequence of their adaptation to specific niches, influenced by the host's physiological factors, dietary preferences, and growth environment. As a result, distinct microbial communities develop in different regions of the GI tract, leading to the formation of regionally specific microbial populations (Thursby *et al.*, 2017). Furthermore, these communities can undergo dynamic changes over time in response to factors such as disease or shifts in environmental conditions (Flores *et al.*, 2014, Johnson *et al.*, 2019). Those exposures, in conjunction with host genetic variations, lead to a high degree of interpersonal variation in gut microbiota composition (Eckburg *et al.*, 2005).

The dominant phyla within the gut microbiota comprise of *Firmicutes, Bacteroidetes, Actinobacteria*, *Proteobacteria, Fusobacteria*, and *Verrucomicrobia* (Lozupone *et al.*, 2012), where *Firmicutes* and *Bacteroidetes* collectively account for 90% of the gut microbiota (Mariat *et al.*, 2009). The *Firmicutes* phylum includes over 200 genera, including *Lactobacillus, Bacillus, Clostridium, Enterococcus*, and *Ruminococcus*, with *Clostridium* making up 95% of this phylum. *Bacteroidetes* is most prominently represented by *Bacteroides* and *Prevotella*, and the phylum *Actinobacteria* primarily by the *Bifidobacterium* genus (Rinninella *et al.*, 2019). Due to their widespread prevalence and the relative ease of accessibility through the highly conserved 16S rRNA gene, the bacterial microbiota has been extensively studied (Clarridge, 2004).

In contrast, the non-bacterial populations, including archaea, fungi, viruses, and protists, have only recently garnered attention in microbiome research. Emerging evidence suggests that these populations also play critical roles in maintaining human homeostasis and contributing to disease aetiology. However, their role in the pathogenesis of IBD has been reviewed elsewhere (Ungaro *et al.*, 2019, Matijašić *et al.*, 2020, Hsu *et al.*, 2023, Le *et al.*, 2022) and falls outside the scope of this thesis, as the methodologies employed here primarily capture bacteria.

3.1. Microbial composition in IBD

Evidence from both animal models and human subjects provides substantial support for the involvement of the gut microbiome in the pathogenesis of IBD. In animal models, the absence of gut microbes in germ-free mice prevented colitis development in genetically susceptible mice (Veltkamp *et al.*, 2001), and transferring pro-inflammatory bacteria from diseased mice to healthy ones induced inflammation (Schaubeck *et al.*, 2016, Du *et al.*, 2015). Additionally, the transfer of naïve CD4+ lymphocytes from healthy mice into those lacking T and B lymphocytes was shown to induce colitis (Powrie, 1995), and this outcome was dependent on differences in their gut microbiota (Reinoso Webb *et al.*, 2018). Observations in humans align with these findings, with IBD activity being particularly pronounced in regions characterized by elevated bacterial populations situated in proximity to the affected tissues. Notably, these regions include the terminal ileum and the colon, where faecal material tends to exhibit relative stagnation (Martinez-Medina *et al.*, 2014). For patients with CD, modifying the continuity of the faecal stream through surgical intervention has been demonstrated to improve inflammation in the excluded portion, despite minimal changes in histopathology (Janowitz *et al.*, 1998). Antibiotic therapy has proven effective in specific cases of IBD (Thia *et al.*, 2009, Perencevich *et al.*, 2006, Turner *et al.*, 2014), and certain microbes have been identified as either driving or suppressing the inflammation (Henke *et al.*, 2019, Chassaing *et al.*, 2013, Zhou *et al.*, 2020).

A substantial body of evidence suggests that the gut microbiome is altered in patients with IBD compared to healthy controls. The most consistent findings associated with

IBD, which were confirmed in various systematic reviews and meta-analyses, were a reduced microbial diversity, higher temporal instability of the community structure and an imbalanced microbial composition (*Figure 5*). Latter is characterized by a depletion of beneficial microorganisms, including *Clostridium IV* and *XIVa*, *Bacteroides, Sutterella, Roseburia, Bifidobacterium*, and *Faecalibacterium prausnitzii* in conjunction with an increase in certain potential pathogens, such as members of the *Proteobacteria* (including invasive and adherent *Escherichia coli, Salmonella, Yersinia and Helicobacter*), *Veillonellaceae, Pasteurellaceae, Fusobacterium*, and the mucin degrader *Ruminococcus gnavus* (Abdel-Rahman *et al.*, 2023, Pittayanon *et al.*, 2020, Aldars-Garcia *et al.*, 2021, Prosberg *et al.*, 2016). Several studies with treatment naïve patients suggested that those alterations are, at least to some extent, already present at disease onset. (Gevers *et al.*, 2014, Rausch *et al.*, 2023, Olbjørn *et al.*, 2019) Microbiome studies often vary, at times substantially, in their reported differentially abundant taxa and disease-dependent effect sizes, especially between the IBD subtypes and disease activity states (Abdel-Rahman *et al.*, 2023, Pittayanon *et al.*, 2020, Aldars-Garcia *et al.*, 2021). The microbiome structure of patients with UC is often reported to be closer to that of controls than that of CD, with higher microbe diversity, less severe net loss of beneficial symbionts such as *F. prausnitzii* or *Blautia* as well as less increased *Proteobacteria* populations (Sankarasubramanian *et al.*, 2020, Yilmaz *et al.*, 2019, Vich Vila *et al.*, 2018). However, despite the proposal of a microbiome signature to differentiate between CD and non-CD samples consisting of *Faecalibacterium, Peptostreptococcaceae, Anaerostipes, Methanobrevibacter, Christensenellaceae, Collinsella, Fusobacterium,* and *Escherichia (Pascal et al., 2017)*, other studies have identified *Eubacteriaceae*, *Bacteroidaceae*, and *Verrucomicrobiaceae* as most discriminating taxa (Papa *et al.*, 2012). In yet another model designed to differentiate between IBD subtypes, key features encompassed several members of *Lachnospiraceae* and *Ruminococcaceae*, as well as *Gammaproteobacteria (*Halfvarson *et al., 2017)*.

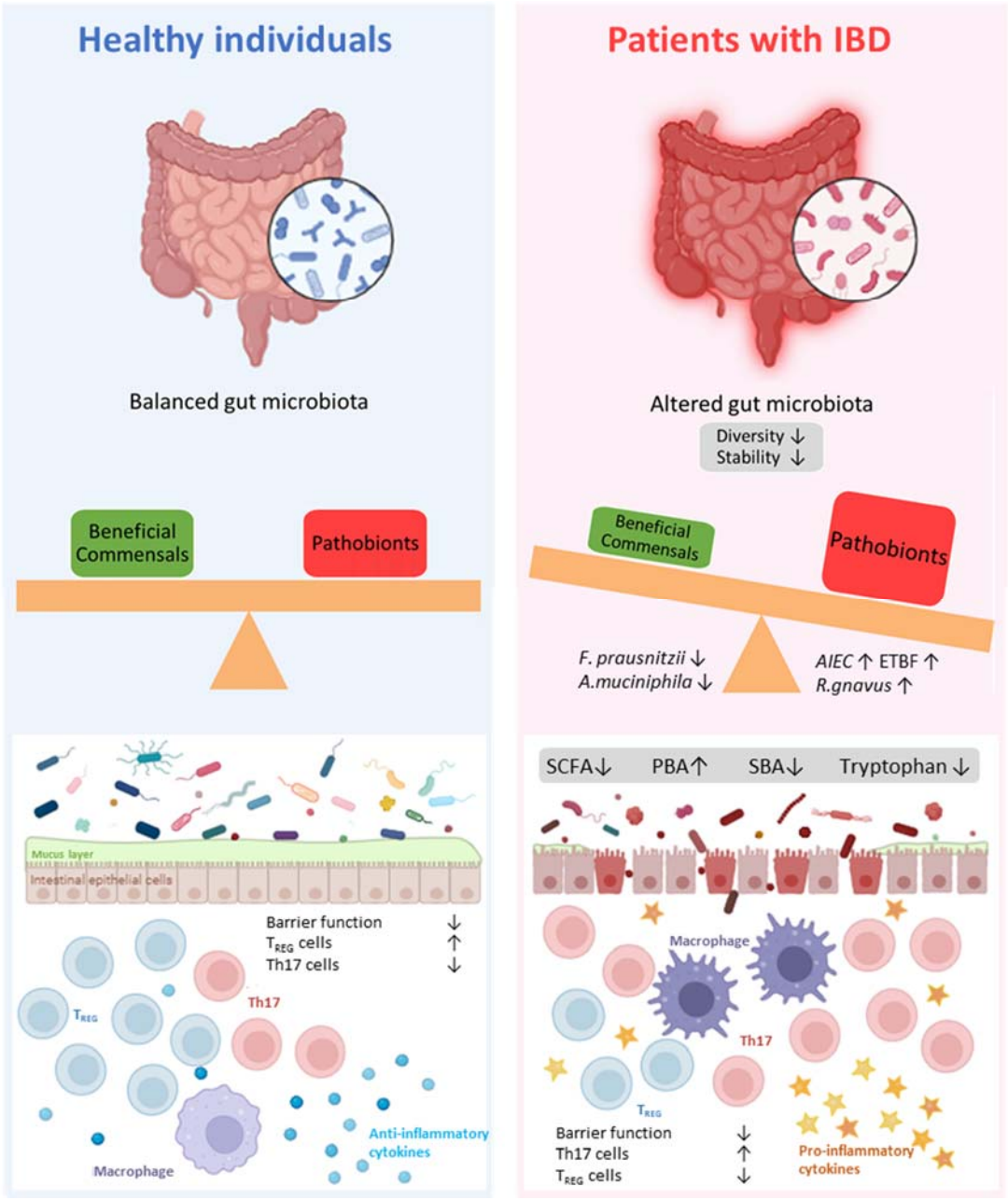


Figure 5: The role of gut microbiota in the pathogenesis of IBD. In healthy individuals, a diverse gut microbiome serves beneficial functions for the host and contributes to maintaining intestinal homeostasis (left). However, in individuals with IBD, there is an alteration of the gut microbial composition, characterized by an increase in 'pathobionts' and a decrease in commensal microbes and microbial metabolites (right). This imbalance leads to impaired barrier function, pathogen translocation, and an imbalance in TREG and Th17 cells, ultimately resulting in elevated levels of pro-inflammatory cytokines (adapted from Cabezas-Cruz et al., 2023, Round et al., 2009 and Hou et al., 2022).

For the distinction of disease activity states there is a similar picture. While certain studies reported decreased bacterial richness alongside an increase in *F. prausnitzii* and *Clostridiales* among patients in remission from IBD, with a simultaneous rise in *Proteobacteria* in those experiencing relapses (Zhang *et al.*, 2007, Tedjo *et al.*, 2016, Fukuda *et al.*, 2014, Nishikawa *et al.*, 2009, Zhu *et al.*, 2022) other studies have failed

to replicate these findings (Galazzo *et al.*, 2019, Bibiloni *et al.*, 2006, Nemoto *et al.*, 2012). Naftlali and co-authors proposed that a potential explanation for this variation could be the distinct microbial profiles observed in patients with CD involving the ileum versus the colon, with discernible differences in relation to disease activity evident in the former but not in the latter (Naftali *et al.*, 2016). While stool samples are readily accessible, they do not entirely capture the microbial communities adherent to the mucosa, which directly interact with the epithelial barrier. These mucosal communities can, however, only be obtained through an invasive biopsy procedure through endoscopy, which is especially problematic with regards to paediatric and longitudinal studies in humans. Although studies in macaques found a high correlation of taxa and their relative abundances between stool and mucosa (Yasuda *et al.*, 2015), most human IBD studies found significant differences between faecal and mucosal samples (Nowicki *et al.*, 2023, Sartor, 2015, Hou *et al.*, 2022). It is, however, important to note that not all variation arises solely from biological factors or as reaction to the aforementioned environmental influences. Technical variations and the diversity in analysis methods and experimental designs also contribute significantly to the observed variability (Ni *et al.*, 2017, Aldars-Garcia *et al.*, 2021). Microbiome studies encompass numerous experiments and analytical procedures, each of which is capable of introducing biases that can potentially undermine the study's biological findings and reproducibility. In the absence of one-size-fits-all approaches and standardized methods for microbiome studies this can cause significant variability in study outcomes (Schloss, 2018, Poussin *et al.*, 2018). Discrepancies on the microbiome composition outcomes were shown to arise from factors such as sample storage (Choo *et al.*, 2015), DNA extraction and purification (Costea *et al.*, 2018), amplification of 16S rDNA via PCR (Hiergeist *et al.*, 2016), sequencing methodologies and platforms (Clooney *et al.*, 2016b), as well as bioinformatics analysis (Tanca *et al.*, 2016, Goodrich *et al.*, 2014, Nearing *et al.*, 2022).

Another issue, unrelated to variation, is that despite the abundant evidence pointing to differences in the microbiome between health and disease, and the essential role of the microbiome observed in disease progression, researchers encounter

difficulties in providing a coherent argument whether the observed changes in gut microbiota are causal or consequential. An argument for the latter would be the observations from a recent meta-analysis of microbiome data from 28 studies spanning ten different diseases. It revealed that fifty percent of the taxa displaying differential abundance between cases and controls were common across all investigated diseases (Duvallet *et al.*, 2017). A substantial number of these taxonomic changes are also linked to the aging process (Ghosh *et al.*, 2020). Significantly, many of the beneficial commensal microorganisms that exhibit depletion in disease states belong to the category of strict anaerobes, such as *F. prausnitzii*, which are sensitive to oxygen exposure (Zou *et al.*, 2021). In contrast, *Proteobacteria* encompass obligate or facultative anaerobes, rendering them more tolerant to the presence of oxygen in their surroundings (Moon *et al.*, 2018). While the oxygen concentration in the healthy large intestine typically remains exceedingly low, it can escalate during episodes of inflammation, either due to an influx of blood into the intestine or as a consequence of immune activity by phagocytes, which have the capacity to generate substantial amounts of ROS (Rigottier-Gois, 2013). Henson *et al.* were able to in silico replicate changes akin to those observed in IBD by augmenting the oxygen levels in their multispecies biofilm models (Henson *et al.*, 2017). Ultimately, in instances of chronic diseases that were initially linked to microbiota, subsequent research has demonstrated that the observed effects were primarily attributable to treatment (Forslund *et al.*, 2015) or dietary preferences (Yap *et al.*, 2021). There are several species that are consistently reported altered in IBD across studies and for which potential molecular mechanisms have been described which can be roughly separated into beneficial or potentially harmful.

Faecalibacterium prausnitzii

Faecalibacterium prausnitzii is frequently found decreased in CD, but less so in UC (Lavelle *et al.*, 2015), where its decreased abundance was found to be predictive of relapse (Rajca *et al.*, 2014) and CD recurrence after surgery (Quevrain *et al.*, 2016, Sokol *et al.*, 2008). This species is an important producer of butyrate from fermentation of complex polysaccharides in the diet (Lavelle *et al.*, 2015). Butyrate is an important energy source for IECs and also acts as signalling molecule for the

regulation of cell proliferation and differentiation. It also strengthens the intestinal mucosal barrier function and exhibits immunomodulatory effects (Plöger *et al.*, 2012, Hamer *et al.*, 2008). In addition to these indirect beneficial effects, *F. prausnitzii* has been shown to directly interact with the immune system by inducing a tolerogenic cytokine profile characterized by the induction of very low levels of IL-12 and IFN-γ, along with high levels of IL-10, in peripheral blood mononuclear cells *in vitro* (Sokol *et al.*, 2008). Additionally, oral administration of *F. prausnitzii* or its supernatant to mice alleviated the severity of trinitrobenzene sulfonic acid induced colitis while also improving the disrupted gut microbiota (Sokol *et al.*, 2008). Quévrain and colleagues found a 15kDa protein of unknown function named microbial anti-inflammatory molecule (MAM). MAM-derived peptides were shown to suppress the r NF-κB pathway in epithelial cell lines and to effectively prevent dinitrobenzene sulfonic acid induced colitis in a mouse model (Quevrain *et al.*, 2016). Rossi *et al.* conducted an investigation into the immunomodulatory properties of a previously reported anti-inflammatory strain, *F. prausnitzii* A2-165 and discovered its capability to modify the phenotype of mucosal DCs, suppressing IFN-γ differentiation while inducing IL-10 producing T cells which might contributes to overall immune homeostasis in the gut (Rossi *et al.*, 2016, Alameddine *et al.*, 2019).

Akkermansia muciniphila

Akkermansia muciniphila is another bacterium that is regularly observed to have reduced abundances in patients with IBD (Png *et al.*, 2010, Dunn *et al.*, 2016, Wang *et al.*, 2020, Rajilić-Stojanović *et al.*, 2013). It is recognized for its ability to inhabit the mucosal layer of the human intestinal tract, primarily by utilizing mucin glycoproteins as its primary source of carbon and nitrogen (Derrien *et al.*, 2004). This mucin-degrading, strictly anaerobe bacterium has been shown to contribute to the stability of the intestinal barrier, likely due to nutrient exchange, immune system maturation and defence against invading pathogens (Derrien *et al.*, 2004, Everard *et al.*, 2013, van der Lugt *et al.*, 2019). Extracellular vesicles derived from *A. muciniphila* reduced the production of the pro-inflammatory cytokine IL-6 in IECs *in vitro* and when orally administered exhibited protective effects against dextran sodium sulfate induced colitis (Kang *et al.*, 2013, Bian *et al.*, 2019). Through the activation of Toll-like

receptors TLR2 and TLR4, *A. muciniphila* induces elevated levels of IL-10, akin to the response elicited by *F. prausnitzii* A2-165. This is facilitated by the presence of an abundant pili-like protein on their outer membrane, called Amuc_1100 (Plovier *et al.*, 2017). Conversely, the presence of *A. muciniphila* in an *in vitro* culture of peripheral blood mononuclear cells induced a pro-inflammatory profile characterized by TNF-α, IFN-γ secretion, indicating this species cannot be categorised as solely anti-inflammatory (Cekanaviciute *et al.*, 2017). Indeed, Ansaldo *et al.* reported that *A. muciniphila* induced homeostatic IgG production and antigen-specific T cell responses in gnotobiotic mice, but could also elicit different T cell responses in conventional mice (Ansaldo *et al.*, 2019).

Bacteroides fragilis

Bacteroides fragilis is an opportunistic pathogen that occupies different roles in the gut. While the non-toxigenic strain NCTC 9343 of *B. fragili*s (NTBF) is known to enhance mucosal immune development and mitigate colitis, the enterotoxigenic strain of *B. fragilis* (ETBF) is strongly linked to IBD and colorectal cancer (CRC) (Chan *et al.*, 2019, Prindiville *et al.*, 2000, Basset *et al.*, 2004). A key feature of *B. fragilis* is that a substantial part of its genetic makeup is reserved to carbohydrate metabolism (Xu *et al.*, 2003), encoding functions such as degradation of dietary polysaccharides and the synthesis of various surface capsular polysaccharides (Krinos *et al.*, 2001, Coyne *et al.*, 2005). One of them, Polysaccharide A (PSA) stands out among other zwitterionic polysaccharides, as it can trigger T cell-dependent immune responses. These responses play a role in influencing the development of CD4+ T cells mediating a correct Th-1/Th-2 balance for the host (Mazmanian *et al.*, 2005) and also promote immunologic tolerance through TLR2 signalling on Foxp3(+) T_{REG} cells (Round *et al.*, 2011). PSA was found to protect mice from *Helicobacter hepaticus* induced colitis (Mazmanian *et al.*, 2008) and its promotor showed an inversed orientation, effectively downregulating the expression of PSA in patients with IBD (Blandford *et al.*, 2019). Conversely, ETBF secrete *B. fragilis* toxin (BFT), a 20-kDa zinc-dependent metalloprotease with three isotypes (BFT-1, BFT-2, and BFT-3) (Kang *et al.*, 2014). The strain also promotes the breakdown of E-cadherin, which leads to heightened permeability of the gut barrier and triggers the activation of the NF-κB signalling

pathway, causing colonic epithelial cells to release pro-inflammatory cytokines (Wick *et al.*, 2010). ETBF is linked to inflammatory diarrhoea in humans and animals (Sears *et al.*, 2008, Sack *et al.*, 1992), and is also associated with clinically active IBD (Prindiville *et al.*, 2000).

Escherichia coli

Similarly, oral application of the probiotic strain *Escherichia coli* Nissle 1917, which is commercially available as a probiotic, can improve a number of GI issues such as diarrhoea and symptomatic uncomplicated diverticular disease (SUDD) as well as maintain remission in patients with UC (Scaldaferri *et al.*, 2016, Losurdo *et al.*, 2015), while a genetically, phylogenetically, and functionally variable group of *E. coli* with adherent-invasive phenotype (AIEC) have been shown to elicit colitis in mouse models and to potentially contribute to IBD (Chassaing *et al.*, 2013, Benoit *et al.*, 2014, Small *et al.*, 2013, Kittana *et al.*, 2023, Nadalian *et al.*, 2021). AIEC strains were shown to have a higher prevalence in ileal CD compared to colonic CD (Darfeuille-Michaud *et al.*, 2004, Martinez-Medina *et al.*, 2009a), and also have been associated with both disease activity (Sasaki *et al.*, 2007, Lopez-Siles *et al.*, 2014) and the severity of inflammation (Elliott *et al.*, 2013, Baumgart *et al.*, 2007). AIEC strains do not possess the known virulence factors typically associated with other pathotypes of *E. coli* (Boudeau *et al.*, 1999, Nash *et al.*, 2010). Additionally, no other virulence factors have been identified that are either common to all AIEC strains or unique to them (Conte *et al.*, 2014, Martinez-Medina *et al.*, 2009b, O'Brien *et al.*, 2017). Instead, AIEC strains are distinguished by their *in vitro* characteristics. This phenotype is characterized by its ability to adhere to and invade human epithelial cells (Boudeau *et al.*, 1999), its capacity for survival and replication within macrophages, and its ability to induce the production of the pro-inflammatory cytokine TNF-α (Glasser *et al.*, 2001). AIEC have the ability to breach the mucous layer through the facilitation of mucin degrading proteases (McGuckin *et al.*, 2011). Additionally, they have been demonstrated to withstand the action of antimicrobial peptides through two genes encoding a Mig-14 family protein which is linked to defensin resistance, and an OmpT family outer membrane protease (McPhee *et al.*, 2014). The attachment to the intestinal epithelium is then facilitated by type 1 pili (FimH) found on the surface of

AIEC. These pili interact with cell adhesion molecule 6 (CEACAM6) receptors present on the ileal enterocytes. Notably, in patients with CD, CEACAM6 host receptors are overexpressed, thereby promoting the colonization of the ileal mucosa by AIEC (Barnich *et al.*, 2007). This colonization induces the upregulation of the pore-forming tight junction protein claudin-2 within IECs (Denizot *et al.*, 2012), resulting in the rearrangement of tight junctions from the exterior, causing a reduction and ultimately the loss of barrier function (Shawki *et al.*, 2017). The presence of flagella on AIEC surfaces, which play a role in biofilm formation (Martinez-Medina *et al.*, 2009b), has the capacity to activate and upregulate TLR5. When flagella bind to TLR5 located on the surface of IECs, it stimulates the production of hypoxia-inducible factor-1a and triggers the NF-κB pathway, which results in the secretion of pro-inflammatory IL-8. This cascade leads to an escalated inflammatory response and inhibits autophagy in host cells (Eaves-Pyles *et al.*, 2008, Subramanian *et al.*, 2008). Via FimH AIEC are also able to bind onto glycoprotein 2 on the surface of microfold cells, which are distributed among enterocytes, and through a mechanism involving long polar fimbriae they are also able to enter and translocate through them (Roberts *et al.*, 2010). In patients with CD, aphthoid ulcers develop at the site of microfold cells, which constitute around 5% of the follicle-associated epithelium in Peyer's patches, suggesting a role for AIEC in their formation (Fujimura *et al.*, 1996). Two AIEC strains, LF82 and 083:H1, were observed to stimulate the expression of TNF-α, IFN-γ, and IL-8 transcripts in colon biopsies from individuals with CD (Ambrose *et al.*, 1984). The release of the pro-inflammatory cytokine IL-8 by IECs results in the recruitment of macrophages and DCs to the infection site. Meanwhile, the secretion of IFN-γ and TNF-α by macrophages leads to the upregulation of CEACAM6 expression, thereby promoting further AIEC colonization (Palmela *et al.*, 2018).

Ruminococcus gnavus

Ruminococcus gnavus is a prevalent member of the healthy gut microbiome with a median abundance of around 0.1% across 90% of the population (Qin *et al.*, 2010). In IBD, its numbers have been shown to temporarily increase to up to 69% to the total microbial composition (Hall *et al.*, 2017) and the increased abundance has been linked to disease severity in patients with CD (Png *et al.*, 2010), as well as the risk of

pouchitis in UC patients who have undergone a total colectomy (Machiels *et al.*, 2017). Adaptation of *R. gnavus* to the gut environment varies depending on the specific strain, and there is a significant degree of variation among strains in terms of their functional characteristics. A study by Hall *et al.* identified 199 strain specific genes that were associated with various functions, including adhesion, responses to oxidative stress, iron acquisition, and the breakdown and utilization of mucus (Hall *et al.*, 2017). The *Ruminococcus* species *R. gnavus* and *R. torques* were shown to produce extracellular α- and β-glycosidases that have the capability to cleave the terminal sugars from MUC2 O-glycans (Hoskins *et al.*, 1985, Beaud *et al.*, 2005). These O-glycans compose a combination of sialic acid, sulfate esters, blood group antigens, and Lewis antigens, which collectively render the mucin macromolecule resistant to microbial degradation (Van Klinken *et al.*, 1995). Once these terminal sugars are enzymatically cleaved, revealing oligosaccharides and the protein core, the mucin matrix's structural integrity becomes compromised. It becomes vulnerable to degradation by other colonic bacteria, potentially resulting in a compromised gut barrier stability (Tailford *et al.*, 2015). Additionally, *R. gnavus* strain ATCC 29149, which is able to grow on mucin as a sole carbon source (Crost *et al.*, 2016), was observed to be produce an intramolecular trans-sialidase that selectively converts α2–3-linked sialic acid substrates into 2,7-anhydro-Neu5Ac. This unique enzymatic activity provides *R. gnavus* with a nutritional advantage, as 2,7-anhydro-Neu5Ac is deemed less accessible to other gut microbiota members (Tailford *et al.*, 2015). Another strategy of *R. gnavus* to gain competitive advantage is the production of bacteriocins. Ruminococcin A (RumA) is a small peptide composed of fewer than 40 amino acids with trypsin-dependent lantibiotic activity. Genes similar to *rum*A have been observed to be widespread among different strains of *R. gnavus* (Donia *et al.*, 2015, Marcille *et al.*, 2002). Furthermore, it has been demonstrated *in vivo* that the *R. gnavus* strain E1 can synthesize the anti-clostridial sactipeptide Ruminococcin C during its colonization of the digestive tract of rats (Crost *et al.*, 2011).

In addition to influencing the microbial composition via secretion of antimicrobial peptides, *R. gnavus* strains can also interact with the host immune system. On the one hand, they produced a complex glucose-rhamnose cell wall polysaccharide

(glucorhamnan), which potently induces the pro-inflammatory cytokine TNFα secretion by DCs in a TLR4-dependent manner (Henke *et al.*, 2019). On the other hand, some strains of *R. gnavus* carry a large capsular polysaccharide (CPS) on their cell surface, which has immunomodulatory properties, as CPS-positive strains did not induce an immune response (Henke *et al.*, 2021). Thus, *R. gnavus* strains can elicit both pro- and anti-inflammatory responses.

These examples of mechanistic processes that are implicated to either contribute to or protect against IBD highlight the significant disparity in behaviour among different strains of the same species within the gut environment. This variability is influenced by their genetic composition and the specific selective pressures they encounter. Consequently, it becomes crucial to assess the gut microbial community not merely by the presence of particular taxa but also by the functions they perform.

3.2. Microbial metabolites in IBD

Metabolic pathways from microbes produce a number of small molecules, either as intermediate or end product of dietary substrates, modified host molecules or directly microbiota derived metabolites. These metabolites serve as the primary means through which microbes engage with their host, impacting immune development and balance, the host's energy metabolism, and the preservation of mucosal integrity (Postler *et al.*, 2017, Lavelle *et al.*, 2020). The influence of microbial metabolites on their host has been demonstrated in germfree-mice which exhibit drastically altered metabolic profiles than that of conventional mice (Wikoff *et al.*, 2009), with differences observed in plasma, urine, liver and gut tissues (Claus *et al.*, 2008). Multiple studies investigated how the metabolic profiles, i.e. a snapshot of small molecules in a sampled environment, of patients with IBD differed from that of healthy controls (Morgan *et al.*, 2012b, Schirmer *et al.*, 2018, Pascal *et al.*, 2017). Most notably, Franzosa et al combined metagenomics and metabolomics to analyse stool samples from 155 IBD patients of the PRISM cohort. The team used 4 separate liquid chromatography mass spectroscopy methods to measure polar metabolites, lipids, free fatty acids and bile acids and found more than 8000 molecules, of which only a fraction could be annotated. Similar to the taxonomic profiles, the metabolic

profiles of patients with CD showed greater dissimilarity from healthy controls compared to patients with UC. The authors argued that the variation between patients with UC, with some resembling those with CD and others resembling healthy controls, could be attributed to inflammation. Furthermore, they found that 31% of metabolic features were altered between IBD and healthy controls, with the majority of features shared between the IBD subtypes. The changes in metabolic profiles primarily included enrichments in sphingolipids and bile acids, as well as depletions in triacylglycerols and tetrapyrroles. By integrating the metagenomics and metabolomics data they were able to find strong associations between differentially abundant metabolites and taxa (Franzosa *et al.*, 2019). Overall, bile acids, short-chain fatty acids (SCFAs) as well as tryptophan metabolites have emerged as key classes of metabolites that are altered and have been implicated in the pathogenesis of IBD (Lavelle *et al.*, 2020).

Short-chain fatty acids

SCFAs, mainly acetate, propionate and butyrate, are produced by microbes via fermentation of otherwise indigestible complex carbohydrates (Sonnenburg *et al.*, 2014). Acetate, which accounts for about 50-70% of SCFAs is produced by a wide range of bacteria, while propionate, accounts for 10-20% of the SCFA pool is mainly produced by *Bacteroidetes* and a small number of *Firmicutes*. Butyrate, which is only produced by a small number of *Clostridia* accounts for most of the rest (Rowland *et al.*, 2018). The ratios of SCFAs change depending on the intake of dietary fibre (Velázquez *et al.*, 2000). SFCAs have been shown to influence a large number of host processes, including but not limited to central appetite regulation (Frost *et al.*, 2014), energy metabolism (Tolhurst *et al.*, 2012), regulation of intestinal macrophages (Singh *et al.*, 2014, Chang *et al.*, 2014), T_{REG} cell homeostasis (Furusawa *et al.*, 2013, Smith *et al.*, 2013), gut epithelial homeostasis (Macia *et al.*, 2015) and B cell differentiation and responses (Kim, 2016). In patients with IBD faecal SCFA levels are typically reduced (Marchesi *et al.*, 2016, Takaishi *et al.*, 2008, Vernia *et al.*, 1988). On the microbial side this reduction of SCFAs is associated with a loss of butyrate producers like *F. prausnitzii* and *Roseburia hominis* (Sokol *et al.*, 2008, Machiels *et al.*, 2014) and with a reduced genetic capacity for butyrate synthesis by the gut

microbiota (Laserna-Mendieta *et al.*, 2018). On the host side, genes encoding for enzymes involved in butyrate metabolism are downregulated in inflamed tissues (De Preter *et al.*, 2012, Palmieri *et al.*, 2015).

Tryptophan

Tryptophan is an essential aromatic amino acid that plays a crucial role in many metabolic functions is a precursor of molecules such as serotonin, melatonin and vitamin B3 (Kałużna-Czaplińska *et al.*, 2019). Dietary tryptophan from food like poultry, fish, and dairy can be metabolised from the host via the kynurine (KYN) and serotonin pathways or by the gut microbiota which produce a range of indole metabolites, some of which can interact with the aryl hydrocarbon receptor that has been implicated in the pathogenesis of IBD (Lavelle *et al.*, 2020). The rate limiting step of the KYN pathway, which accounts for 90% of the tryptophan metabolism, is facilitated by indoleamine 2,3-dioxygenase 1 (IDO1). IDO1, which is expressed in various immune cells and epithelial cells, plays a role in regulating both systemic and local immune responses (Richard *et al.*, 2009). In conditions marked by chronic inflammation, such as IBD, the expression of IDO1 is induced by pro-inflammatory cytokines (Ciorba, 2013). Nikolaus *et al.* reported elevated tryptophan degradation through the KYN pathway and identified a negative association between serum tryptophan levels and disease activity in IBD patients (Nikolaus *et al.*, 2017). This observation suggests that tryptophan deficiency may contribute to IBD progression or worsen disease activity, akin to what has been observed in mouse models where tryptophan deficiency exacerbated colitis (Hashimoto *et al.*, 2012). Moreover, the aryl hydrocarbon receptor (AhR), which is an important transcription factor that mainly exerts protective and anti-inflammatory effects (Zenewicz *et al.*, 2008, Qiu *et al.*, 2012), can be activated by certain indole derivatives, produced by a limited number of bacteria, including *Peptostreptococcus russellii* and some *Lactobacilli* (Zelante *et al.*, 2013). AhR signalling, which was shown to inhibit inflammation and colitis in mouse models, is reduced in patients with IBD as AhR is downregulated in inflamed mucosa (Monteleone *et al.*, 2011) and also Indoleacrylic acid (AhR agonist) and indole-3-pyruvic acid (AhR agonist precursor) were found reduced in patients with IBD (Wlodarska *et al.*, 2017, Alexeev *et al.*, 2018).

Bile acids

Bile acids (BA) are small amphipathic molecules with a steroid backbone that are synthesized from cholesterol, accounting for around half of the daily turnover of cholesterol. The process of converting cholesterol into BAs in the liver can be initiated by two enzymes: cholesterol 7α-hydroxylase (CYP7A1) in the classic (neutral) pathway, or mitochondrial sterol 27-hydroxylase (CYP27A1) in the alternative (acidic) pathway (Staels *et al.*, 2009). The products generated by these pathways, namely cholic acid (CA) and chenodeoxycholic acid (CDCA), are referred to as primary BAs. To increase the solubility for secretion into the bile and storage in the gallbladder, the primary BAs are conjugated with either glycine or taurine which form Na^+ salts under physiological conditions (Chiang, 2004). Following the consumption of a meal, the gallbladder contracts, releasing micellar BAs into the intestine to facilitate digestion. Through enterohepatic circulation, approximately 95% of these BAs are reabsorbed in the distal ileum and then transported back to the liver through the portal circulation (Hofmann, 1999). The replacement of the remaining 5% of primary BAs that are excreted in the faeces is regulated by fibroblast growth factor 19 (FGF19). This regulation occurs when BAs bind to the farnesoid X receptor (FXR), creating a feedback mechanism where BAs effectively control their own synthesis (Kim *et al.*, 2007). In addition to their function in lipid digestion and the absorption of fat-soluble vitamins in the small intestine, BAs also function as signalling molecules that exert metabolic and immune regulatory effects. These effects are mediated by their binding to various receptors, including FXR, the transmembrane G protein-coupled receptor 5 (TGR5), the pregnane X receptor (PXR), the vitamin D receptor (VDR), and the constitutive androstane receptor (CAR) (Schaap *et al.*, 2014).

The gut microbiota further increases the chemical diversity of the BA pool by modifying the host derived BAs. Microbial bile salt hydrolases deconjugate amino acid residues from primary BAs, leading to the formation of free primary BAs (Jones *et al.*, 2008). These unconjugated primary BAs can then undergo further modifications facilitated by a rather small fraction of the resident gut microbiota, including dehydroxylation, dehydrogenation, and sulfation, giving rise to secondary and tertiary BA forms. In humans, deoxycholic acid (DCA) and lithocholic acid (LCA) are the major secondary BAs produced from CA and CDCA, respectively (Ridlon *et al.*, 2006). A normal microbial BA metabolism was shown to play a

crucial role in regulating the immune response by modulating colonic RORγ^{+} T_{REG} cell differentiation thus ameliorating host susceptibility to inflammatory colitis (Song *et al.*, 2020, Hang *et al.*, 2019). However, while BAs can be modified by microbiota, they in turn exert an significant effect on microbiota, either directly by inhibiting the growth of certain taxa (Kurdi *et al.*, 2006) or indirectly by stimulating the expression of antimicrobial peptides such as cathelicidin (D'Aldebert *et al.*, 2009), angiogenin (Inagaki *et al.*, 2006, Hooper *et al.*, 2003), and inducible nitric oxide synthase (Wallace *et al.*, 2000, Inagaki *et al.*, 2006), while FXR activation was shown to prevent bacterial overgrowth and translocation (Lorenzo-Zúñiga *et al.*, 2003). Faecal BA profiles in individuals with IBD have been repeatedly reported altered compared to healthy controls, characterized by reduced secondary BAs and elevated primary BAs (Franzosa *et al.*, 2019, Jansson *et al.*, 2009, Jacobs *et al.*, 2016, Vantrappen *et al.*, 1977, Kruis *et al.*, 1986). Furthermore, it was shown that not only the faecal concentrations of LCA and DCA were decreased in the ileal pouches of patients with UC, but also the biosynthetic capabilities to convert primary BAs into secondary BAs was reduced. The same study also showed that secondary BA supplementation was able to reduce intestinal inflammation in mouse models of colitis (Sinha *et al.*, 2020). Finally, several taxa have been associated with the reduction in secondary BAs in patients with IBD, namely *Bifidobacterium* and *Clostridium clusters IV* and *XIV*b (Wang *et al.*, 2021), *Roseburia* and unclassified *Subdoligranulum* species (Lloyd-Price *et al.*, 2019) as well as Ruminococcaceae (particularly *Clostridium leptum*) (Yilmaz *et al.*, 2019, Vich Vila *et al.*, 2018, Sinha *et al.*, 2020).

We know that the microbiota are central to the pathogenesis of IBD and to a certain degree malleable, which makes it an attractive target for new treatment strategies. Due to the high inter-personal heterogeneity and the lack of consistency between studies we have problems to fully exploit the opportunities microbiome research represents for personalised medicines or even to decide what came first: the chicken or the egg. To address this question, influences of IBD on the gut microbiota must be thoroughly characterised, with a particular focus on the multitude of factors, both external and internal, that can potentially obscure our picture of the interactions between the gut microbiota and its host.

Section II: Bioinformatic Platforms for Metagenomics

1. Glossary 51
2. Introduction 53
   2.1. Choice of Metagenomic Approach 54
   2.2. Overview of Sequencing Technologies 57
   2.3. Quality Filtering of Metagenomics Data 59
3. Amplicon Sequencing 60
   3.1. Controlling for Biases and Contamination 61
   3.2. Bioinformatic Analysis 62
   3.3. Pre-processing 64
   3.4. OTU Clustering 65
   3.5. Sub-OTU Methods 67
   3.6. Taxonomic Assignment 68
   3.7. Copy Number Variation 71
4. Whole Genome Shotgun Sequence Analysis 72
   4.1. Translating Sequence Data into Genomes 73
   4.2. Genomic Assembly 74
   4.3. Metagenomic Assembly 75
   4.4. Analysis of Assembled Genomes 78
   4.5. Variant Calling 80
   4.6. Analysis of the Functional Capacity of a Metagenome 81
5. Further Analysis 84
6. Reproducibility and Benchmarking 86
7. Data Deposition 87
8. Concluding Remarks 88

1. Glossary

16S Ribosomal RNA gene Transcribed form of the small subunit gene located in the 30S subunit of a prokaryotic ribosome. Contains 9 variable regions that can be targeted for amplification and used for profiling of microbial communities.

Amplicon Naturally or artificially amplified target gene sequence.

Amplicon sequence variant (ASV) Refers to individual DNA sequences recovered from a high-throughput marker gene analysis after the removal of spurious sequences generated during PCR amplification and sequencing.

Annotation Assigning a taxonomic lineage to a marker gene or genome or a function to a gene.

Archaeome Collective coding capacity of archaea within a defined environment.

Assembly Aligning and merging fragments from a longer DNA sequence in order to reconstruct the original sequence.

Binning Grouping reads or contigs based on particular attributes and assigning them to operational taxonomic units.

Bioinformatics Use of computer science and statistics to analyze and interpret biological data.

Biomarker Measurable indicator of some biological state or condition.

Centroid The most representative point/sequence/ASV within in a cluster.

Chimeric sequences An artificial DNA sequence generated during amplification, consisting of a combination of two (or more) parental sequences.

Clonal culture A culture of cells produced in such a way that individual clones, i.e. identical cells that are derived from the same cell, can be selected.

Contig Short for contiguous, assembly of short overlapping reads that together build a consensus region of DNA.

Diversity A measure of the taxonomic distribution within a community, either in terms of distinct taxa or in terms of their evolutionary/phylogenetic distance.

Functional expression screening Method of screening prepared genomic or cDNA libraries for genes of interest.

Hamming distance A metric for comparing two data strings. It is defined as the number of positions in which the corresponding symbols are different when comparing two binary strings of equal length.

Heuristic Practical method to solve a problem/task, speeding up the process of finding a satisfactory solution while not guaranteeing optimal results.

Internal transcribed spacer (ITS) Highly polymorphic non-coding spacer DNA situated between the small-subunit rRNA and large-subunit rRNA genes which length ranges from 450 to 750 bp and is utilized as universal fungal barcode sequence.

***k*-mer** Sub-sequences of length k contained within a biological sequence.

Lowest common ancestor (LCA) In graph theory the lowest common ancestor of two nodes x and y in a tree is the lowest (i.e. deepest) node that has both x and y as descendants.

Marker gene sequencing Also called amplicon sequencing, after extracting DNA from a sample, certain marker genes are amplified and successively sequenced and serve as proxy for the whole genomes in the further analysis.

Metagenome The experimentally determined data set of genomic sequences in a given environment.

Metagenomics The study of microbial communities, typically relying on high-throughput experimental data and bioinformatic techniques.

Microbiome Collective coding capacity of microorganisms within a defined environment.

Microbiota Total collection of microbial organisms within a community or sample.

Mock community Defined mixture of microbial cells or nucleic acid molecules created in vitro to simulate the composition of a microbiome sample or the nucleic acid isolated therefrom.

Mycobiome Collective coding capacity of fungi within a defined environment.

Operational Taxonomic Unit (OTU) A cluster of sequence which are similar beyond some threshold (e.g. 97%) used in place of species, genus, etc.

Orthologues Functionally equivalent genes in different species that have been derived from a common ancestral gene.

Paralog Homologous genes within a genome that are the result of gene duplication events.

Phageome Collective coding capacity of bacteriophages within a defined environment.

Phylogenetically independent contrasts Generalization of the paired comparison method in comparative biology, which maps data for phenotypic traits onto an existing phylogenetic tree.

Phylogenetics The study of the evolutionary history and relationships among individuals or groups of organisms.

Pipeline Series of tools or scripts for the analysis of particular data where the output of the previous step is the input for the next step.

Reads DNA code of segments obtained from a sequencing experiment.

Shotgun sequencing Extracted DNA from a sample is sheered into fragments and randomly sequenced.

Taxa Populations of phylogenetically related organisms.

Taxonomy Science of naming, defining and classifying groups of biological organisms on the basis of shared characteristics.

Variant calling Process of identifying variants between closely related sets of sequence data, which typically take the form of single-nucleotide polymorphisms.

Virome Collective coding capacity of viruses within a defined environment.

Whole Genome Shotgun (WGS) sequencing Shotgun sequencing of individual organisms or microbial communities.

2. Introduction

Metagenomics is the process of recovering genetic material of all microbes present in an environmental sample with the aim of studying the population dynamics and functional potential of these microorganisms. Although the term microbiome refers to all bacteria, fungi, archaea, viruses and protozoa, bacteria and archaea have been the focus of the majority of research to this point. Studies of other phyla are increasing, for instance of the phageome as it has a major impact on dairy production (Pujato *et al.*, 2019).

When the concept of metagenomics was first described in 1998 it involved the cultivation of clonal cultures followed by a functional expression screening, which for the first time, at least theoretically, allowed access to the collective metagenome of a given sample and a route to harvest its biosynthetic capabilities (Handelsman *et al.*, 1998). The advent of next generation sequencing (NGS) techniques after 2005 resulted in rapidly decreasing costs for sequence-based metagenomics analysis and provided researchers with the capability to capture complete microbial profiles without prior knowledge of composition and independent of the ability to culture the organisms concerned.

Metagenomics today serves as an umbrella term for two different approaches, amplicon or marker gene sequencing, and true whole genome shotgun (WGS) sequencing. Marker gene studies typically involve the use of generic primers to amplify conserved regions between species and their proximal hyper-variable regions. Whole genome shotgun metagenomics, on the other hand, entails non-targeted sequencing of all genetic material in a microbiome. The key difference between the two is that, while marker gene studies can detail what organisms at what relative abundance are present in a microbial sample, the presence of entire genomes yielded by WGS also gives detailed information about metabolic potential, evolutionary relationships, and the structure and organization of microbial genomes.

The rapid development of high throughput sequencing (HTS) technologies has granted unprecedented insights into microbial populations within different environments across disciplines including medicine, biology, environmental and food

sciences. While traditional labour intensive “wet lab” experiments have been, by and large, replaced, new challenges have emerged in dealing with the analysis of huge and noisy NGS data. Noise in this context refers to a large amount of additional information and random errors introduced by preparation of the genetic material, the sequencing process, and the downstream processing, obscuring the view to the data signal of interest. The analysis of metagenomics sequences is further complicated by the substantial complexity of these data, differing in read length, sequencing depth, base quality and error profiles. An increasing number of bioinformatics pipelines, platforms and tools are being developed and updated to accurately reveal the underlying microbial community profiles with an increasing resolution, while dealing with the disparities in the produced data.

Studies have shown that not only do the wet lab stages of an experiment (sample storage, DNA extraction, primer type, amplification, and NGS platform) influence the outcome of a metagenomics experiment but also the choice of bioinformatics pipeline used to treat the sequence data. Although numerous benchmarking studies attempt to find the optimal workflow to process and analyse raw data from sequencing experiments, there is currently no agreed optimal pipeline. Instead, the approach taken is based on the environment and research question to be asked (Xue *et al.*, 2018, Golob *et al.*, 2017, Walsh *et al.*, 2018).

The principal steps of a metagenomics survey are (i) sampling, microbial DNA extraction and sequencing, followed by (ii) bioinformatics sequence processing of the raw sequence data and finally (iii) statistical analysis and visualization. In this chapter we will concentrate mostly on the second step and give an overview of the general workflow from the raw data delivered by sequencers to a human interpretable state, suitable for further statistical analysis.

2.1. Choice of Metagenomic Approach

Both sequencing strategies (amplification and whole genome shotgun) possess their own advantages and disadvantages which need to be considered before choosing an approach. In addition to the considerable economic considerations, the environment

being sampled can also dictate this decision, as well as the availability of computational resources and bioinformatics expertise.

WGS metagenomics sequences fragments of the total microbial DNA of a sample and thus not only allows the inference of the relative abundance of every member of the community at high resolution (species and even strain level), but it also enables quantification of metabolic pathways to give a picture of the metabolic potential of the studied microbial community. As random fragments of the complete length of the genomes are sequenced, a much greater depth of sequencing is required for this approach compared to amplicon sequencing. This is not only slower and more costly, but also results in a much larger amount of data whose analysis is computationally expensive due to the need to assemble the fragments, either against a reference database or using de novo assembling techniques. Another problem with reference-based assembly is that not all genomes are yet represented. More recently characterized species are better represented in 16S rRNA databases, meaning that for some environments, amplicon sequencing is more effective than WGS sequencing despite the lower information content (Tessler *et al.*, 2017).

Marker gene studies involve the sequencing of well-defined and relatively small phylogenetic markers following amplification via polymerase chain reaction (PCR). This results in a smaller amount of data as each organism is represented by its phylogenetic marker. As only a defined region of each genome in the sample is sequenced, namely the marker gene, more samples can be processed in every instrument run. This does not only decrease the required minimum sequencing depth but also the costs. While the amplification step is known to create biases, it also allows the analysis of samples with a low microbial load or with a high amount of host DNA, as it specifically amplifies the DNA of interest. On the downside, marker gene studies usually deliver only a genus level resolution of the microbial composition, even though bioinformatics tools exist that try to infer species (Spingo (Allard *et al.*, 2015)) and even strain level (Himap (Segota *et al.*, 2019)) resolution of the microbiome under study. Although there have been efforts to estimate functional profiles from amplicons using a variety of tools (PICRUSt (Douglas *et al.*, 2018), Tax4Fun (Asshauer *et al.*, 2015), Piphillin (Iwai *et al.*, 2016)), these tools are generally

applied in an exploratory sense to narrow the field for future more in-depth studies with shotgun sequencing.

These two approaches are not mutually exclusive; often an effective trade-off is to carry out a broader analysis with an amplicon sequencing method with a larger sample size, before proceeding to WGS metagenomics with a subset of samples. *Table 1* compares amplicon sequencing and WGS metagenomics.

Table 1: Comparing sequencing strategies: Advantages and disadvantages

	Amplicon sequencing	**WGS metagenomics**
Cost	• Minimum 10x cheaper per sample than WGS metagenomics sequencing • Analysis more accessible as less computational power needed	• Prohibitively expensive for smaller labs • Requires considerable processing power for most analysis platforms
Information content	• Provides data specific to amplicon used, i.e. Bacteria/ archaea with 16S rRNA • Genus level resolution with some species level resolution possible • Taxonomy only, with some inferred metagenomics information using tools such as PICRUST2	• Provides data on all microorganisms present in a sample • Species and strain level resolution possible • Detailed taxonomic and functional information available as full genomes generated with sufficient sequencing coverage
Suitability	• More sensitive due to the enrichment of the target sequence but also prone to bias due to amplification step • Databases more comprehensive when characterizing novel environments, as they are easier to populate	• Less susceptible to bias as no amplification is required. • No amplification means that e.g. in biopsies where host DNA can make up 99% of all DNA, WGS is unsuitable without microbial enrichment.
Reference databases	• Extensive databases include also less well studied species • Different databases are known to contradict each other in the assignment of certain sequences	• More exotic species might not be represented and can get overlooked

As the WGS method sequences most DNA in a sample, regardless of what the user intended target may be, the sample type and preparation must be taken into consideration. In environmental samples this may not be a concern, but when trying to sequence microbial populations within host tissues, for example contaminated meat products, the DNA sequenced will be overwhelmingly host DNA, making it difficult to gain sufficient sequencing depth of the microbial content. In instances such as this, amplicon sequencing strategies may be more suitable, at least for initial analysis. The next issue relates to the nature of the sequence data required. There are a variety of sequencing solutions available, discussed in detail in the next section,

but the correct choice of technology, or combinations thereof, has an enormous impact on any downstream research outputs. Other important aspects of experimental design include the following:

1. The use of both positive and negative controls to ensure the reproducibility and comparability of results.
2. The recruitment of an adequate sample size to draw statistically significant conclusions, which can be aided by carrying out a pilot study to measure the effect size.
3. The collection and (manual) curation of comprehensive meaningful metadata for all samples to enable a robust statistical analysis of the generated sequence data.

Figure 6 highlights the key differences in terms of required inputs and the resulting output data between amplicon sequences and WGS.

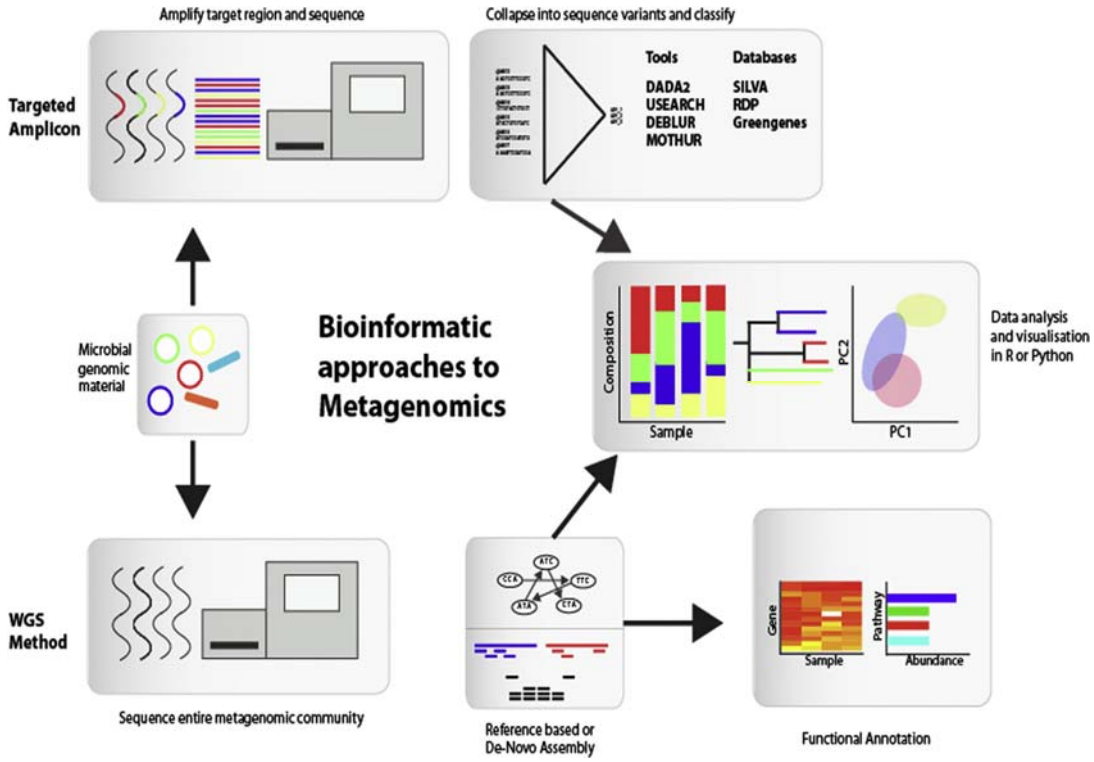


Figure 6: Bioinformatic approaches to metagenomics

2.2. Overview of Sequencing Technologies

Moore's Law states that the number of transistors on a microchip doubles every two years, resulting a halving in the overall cost (Muir *et al.*, 2016). Sequencing technology

increases in accuracy and decreases in cost that even supersede Moore's Law (Wetterstrand, 2022), making it a rapidly-evolving field that is difficult to precisely define. To assist in this, sequencing strategies are grouped together in different generations of the technology, a current snapshot of the state of the art is as follows.

Sanger sequencing, developed in 1977, is referred to as the first generation of sequencing. While still widely used for some projects due to its relatively long read length (on average 650 base pairs), and high accuracy, it is not an appropriate tool for metagenomics studies due to its low throughput and high cost (Escobar-Zepeda *et al.*, 2015). Next generation sequencing methods are massively parallel and can often produce millions of reads during a typical run, where the genomic DNA present is sequenced repeatedly in small random fragments. The two predominant NGS methods are Ion Torrent and Illumina but differences between them in terms of cost, underlying chemistry, output and accuracy mean they are not always suited to the same tasks. The Illumina sequencing platforms provide the most popular sequencing solutions owing to their low cost, high accuracy, and high output (Utturkar *et al.*, 2014). The Illumina MiSeq platform offers low sample output of up to 15 Gb but relatively long reads at an affordable price. The paired-end functionality (i.e. DNA fragments are sequenced from both sides, 5' and 3') offers overlapping reads of up to 300 bp each, making this the technology of choice for amplification-based sequencing experiments such as 16S rRNA gene sequencing. Other Illumina platforms such as the HiSeq, NextSeq and NovaSeq offer much higher output, with the NovaSeq generating up to 6000 Gb, but with shorter read lengths of up to 150 bp (Ambardar *et al.*, 2016). Ion Torrent sequencing yields reads up to 400 bp in length. However only single-end reads (i.e. DNA fragments are only sequenced from one end) are available, and the low total output of up to 15 Gb makes this technology impractical for anything other than amplicon sequencing. In fact, this method has a higher error rate when long repeat regions are present in the sample (Quince *et al.*, 2017).

In the context of WGS, the short reads generated by second generation technologies often yield incomplete genome assemblies. In more complex cases such as genomes with long repeat regions or paralogs, or when sequencing viruses such as bacteriophages, longer reads are often required to close the genomes (Pollard *et al.*,

2018). Third generation sequencing platforms targeting individual DNA molecules, have been developed to meet this demand. Two platforms are currently available, the Oxford Nanopore sequencing methods, and the Pacific Biosciences PacBio platform. A variety of Nanopore solutions are available based on the number of flow cells they contain, ranging from the portable MinION, designed for use in the field to the scaled up PromethION with on board data processing. These Nanopore methods produce reads of up to 100 kb in length (Bowden *et al.*, 2019). The Sequel2 is the current state of the art PacBio technology and can sequence reads up to 60 kb in length. As these third generation strategies begin to reliably upscale their output, their use in metagenomics studies will become more widespread (Quince *et al.*, 2017). A summary of the sequencing options available is found in *Table 2*.

Table 2: Comparison of sequencing solutions across 3 generations of sequencing

Generation	Method	Output type	Max read length	Throughput	Error rate	Runtime
First	Sanger	SE	650–900 bp	62.4 kb	0.0001%	1–3 hours
Second	Illumina	SE and PE	75–300 bp	13.2–6000 Gb	0.26%–0.8%	Up to 6 days
Second	Ion torrent	SE	150–400 bp	10 Gb	1.785%	7.3 hours
Third	Nanopore	SE	Up to 100 kb	0.1–1 Gb	12%–38%	Real time data analysis
Third	PacBio	SE	Up to 60 kb	1 Gb	11%–15%	2 hours

SE: single-end sequencing; PE paired-end sequencing

2.3. Quality Filtering of Metagenomics Data

As the quality of sequencing data returned can be variable, quality filtering is a crucial first step in any analysis. This is particularly important when working with second generation sequence data. Although the error rate with third generation technologies is higher, it remains stable across the length of the sequence, whereas with second generation sequencing technologies, Illumina in particular, a quality profile is generated that depreciates with sequence length, particularly in the reverse read of a pair (Tan *et al.*, 2019). This makes quality filtering prior to downstream analysis a critical step in any analysis pipeline.

The typical output of a sequencing experiment (Illumina and Ion Torrent) is paired or unpaired reads in the fastq format (*Figure 7*). It contains sequence data in conjunction with machine readable error annotations for each sequenced base, the so called a Phred score. This per base quality score ranges from 1 to 40 and is

translated as 10 log10(P), with P being the probability of a base being called erroneously. For example, a Phred score of 10 means there is a 1:10 chance of the base being called incorrectly, whereas the maximum score of 40 means a 1:10,000 chance of error. The fastq format serves as input for most programmes which are utilized to trim away lower quality reads. The trimming itself can either be based on manual visual analysis and cutting a given sequence to a specified length or setting a minimum quality cut off, or a combination of both. Freely accessible technologies for quality filtering sequence data include Trimmomatic (Bolger *et al.*, 2014) and the FASTX-Toolkit (Hannon, 2010).

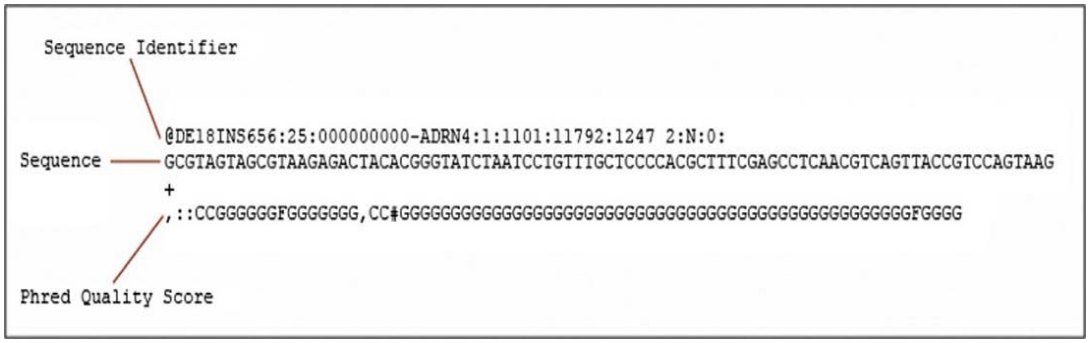


Figure 7: The Fastq format

3. Amplicon Sequencing

Amplification-based sequencing is a fast and cost-effective way to gain insight into the taxonomic composition of a microbial community and the phylogenetic relationship between their members in an environment of interest. As mentioned, amplicon sequencing involves amplification of marker genes via PCR from DNA, extracted from the microbial community of a given sample. Amplicons are subsequently sequenced and aligned against a reference database to gain information about the host organism of that sequence. Such a molecular marker is ideally present in every member of a population, and differs between any individuals with distinct genomes in a manner correlating with the evolutionary distance between individuals of the population (Morgan *et al.*, 2012a).

The 16S small subunit ribosomal RNA (rRNA) gene is commonly used as marker gene as it combines a number of necessary traits. The rRNA gene, which is ubiquitous in bacteria and most archaea, is an integral part of the ribosome and its function has not changed over time, implying that random sequence variations are a precise

measure of evolutionary time. The gene consists of both highly conserved regions, which allow targeted amplification with generic primers, and nine hypervariable regions (HVR),which serve as phylogenetic clock (Tringe *et al.*, 2008).

The shorter read length of second generation HTS platforms does not allow the sequencing of the full 16S rRNA gene, which is 1.5 kb in size, and so one or more adjacent variable regions are usually amplified and subsequently sequenced. It has been shown that those sub-sequences allow for accurate taxonomic classification (Liu *et al.*, 2007), but naturally, this restriction to HVRs comes with a lower taxonomic resolution compared to the full-length gene (Schloss, 2010). Furthermore, different HVRs of the same 16S rRNA gene may result in different taxonomic classification (Youssef *et al.*, 2009). A popular choice is the V4 region. It not only allow the complete overlap of the paired reads in the Illumina MiSeq platform, improving the quality of the generated reads (Kozich *et al.*, 2013), but also has been shown to have the highest correlation to the phylogenetic signal of the complete 16S gene (Yang *et al.*, 2016). Ultimately, the choice of target region should be determined by its discriminatory power between different taxa in the environment of interest, which ideally should be tested with a small subset of samples.

16S rRNA gene sequencing is primarily a method for characterizing the bacterial members of a microbial community. Although some resolution of the archaeome is also provided, it does not include the whole spectrum of archaea, thus knowledge about the archaeome is more a by-product (Koskinen *et al.*, 2017). More recently, the mycobiome has been studied via the use of Internal Transcribed Spacers (ITS) located within the ribosomal gene operon, but the specific databases which are used for taxonomic assignments are still poorly filled in comparison with the 16S specific databases (Xu, 2016, Nilsson *et al.*, 2019). Viruses completely lack a universal gene marker and the analysis of the virome relies completely on shotgun sequencing (Cantalupo *et al.*, 2019).

3.1. Controlling for Biases and Contamination

Regardless of the study goal, it is highly recommended to take measures to control for contamination and bias at each stage of the process, from sampling to data

analysis. Such bias can introduce changes in the observed relative abundance and may have an impact on the community diversity (Pollock *et al.*, 2018). It can be controlled by the introduction of positive controls. For example, synthetic mock communities, composed of predetermined ratios of DNA originating from a mixture of selected bacteria, can be sequenced and analysed along with the experimental samples. This allows the quantification of sequencing errors but also provides information on which taxa are likely to be over or underrepresented (Schloss *et al.*, 2011). Another positive control are spike-in standards, which are added directly to the samples and enable a per sample basis quality control of the used methodology. Then again, adding genetic material directly to the sample bears the risk of crossover between the 16S rRNA gene sequences contained in the standards and those found in samples. Hence, it is important to select bacteria which are highly unlikely to occur in the samples of interest (Tourlousse *et al.*, 2017). Furthermore, it is crucial to include a negative control along with the samples to identify contaminants arising from DNA extraction kits, reagents and the lab environment. It has been shown that not only contaminants between different DNA extraction kits are distinguishable, but also between batches of the same DNA extraction kits (Salter *et al.*, 2014). In high biomass environments contaminants might play a less important role, as their relative abundance in the sample is negligible, but in low biomass samples contaminants can make up a relative high proportion of the sample. This can lead to erroneous conclusions in the downstream analysis if not handled carefully (Karstens *et al.*, 2019). While it is not appropriate to simply remove reads from taxa that have been identified as contaminants since some of these reads might originate from the tested sample, there are tools available (e.g. the R package Decontam (Davis *et al.*, 2018) or SourceTracker (Knights *et al.*, 2011)), that attempt to statistically identify and remove reads from contaminants in metagenomics sequencing data.

3.2. Bioinformatic Analysis

Bioinformatic analysis of amplicon sequences generally consists of the pre-treatment of the raw sequence data followed by microbial diversity analysis and advanced data analysis. Several comprehensive pipelines are available which bundle different tools into one platform, most notably Quantitative Insights Into Microbial Ecology

(QIIME2) (Bolyen *et al.*, 2019), mothur (Schloss et al., 2009), UPARSE (Edgar, 2013), the Divisive Amplicon Denoising Algorithm (DADA2) (Callahan *et al.*, 2016), and the web based MG-Rast (Meyer *et al.*, 2008), which are summarized in *Table 3*. These and other existing platforms not only differ in the underlying algorithms used, but also in available documentation (i.e. extend of publicly available tutorials, manuals and support) as well as in required computational power and runtime. While all these packages provide effective pipelines for 16S rRNA gene analysis, the choice usually depends on the quality of the raw data, the available bioinformatics skills, and computational resources. The mothur and QIIME platforms have been deemed superior to their competitors due to the availability of extensive documentation, the flexibility of the provided processing and analysing methods, accepted input and output formats, and supported sequencing technologies (D'Argenio *et al.*, 2014, Nilakanta *et al.*, 2014, Plummer *et al.*, 2015).

Table 3: Premier 16S rRNA sequence analysis pipelines

	Web	Description	Open source	Access	Reference
QIIME2	https://qiime2.org/	• Rewrite of the QIIME 1 microbiome analysis pipeline • End to end microbiome analysis pipeline • Visualisation	yes	local	(Bolyen *et al.*, 2019)
mothur	https://www.mothur.org	• Incorporated the functionality of DOTUR, SONS, TreeClimber, s-LIBSHUFF, and UNIFRAC • End to end microbiome analysis pipeline • Visualisation	yes	local	(Schloss *et al.*, 2009)
UPARSE	https://www.drive5.com/uparse/	• End to end microbiome analysis pipeline • No visualisation	• 32 bit version free for personal/ academic use • 64 bit version 1485 USD	local	(Edgar, 2013)
DADA2	https://benjjneb.github.io/dada2/index.html	• Amplicon pipeline from raw data to assigning taxonomy for amplicon sequence variants • No visualisation	yes	local	(Callahan *et al.*, 2016)
MG Rast	www.mg-rast.org	• End to end microbiome analysis pipeline • No visualisation • Data repository	yes	web-based	(Meyer *et al.*, 2008)

3.3. Pre-processing

Bioinformatics analysis usually starts with the pre-processing of the raw sequence data, followed by clustering and taxonomic annotation of the quality filtered reads and results in abundance tables which serve as input for the downstream statistical analysis.

The in silico treatment of the raw amplicons begins with quality filtering. As low-quality reads, artefacts and contamination can compromise any downstream analysis, and taxonomic annotation in particular, it is crucial to apply a stringent sequence curation. Pre-processing starts with the removal of sequences with ambiguous bases in the primer and the barcode region, as well as reads with overall low quality scores, homopolymers longer than eight nucleotides and sequences well above or below the expected lengths for a given NGS platform (Kunin *et al.*, 2010, Schloss *et al.*, 2011). The quality filtered reads are subsequently demultiplexed, i.e. the reads are sorted into their respective sample bins according to the unique barcode sequences which have been added to the primer sequences during the amplification. After the removal of any sequencing adapters, amplification primers, and multiplex identifiers, it is necessary to identify and remove chimeric sequences. Chimeric sequences are a known artefact from PCR amplification, resulting from cross-hybridization of unrelated DNA fragments which can inflate the diversity metrics if left untreated. Chimera removal is integrated in the most commonly used classifiers and can either be performed by comparing sequences to a known chimera free data base (Chimera Slayer (Haas *et al.*, 2011) and DECIPHER (Wright *et al.*, 2012)) or de novo by detecting low abundance sequences that could be formed by a combination of two sequences of higher abundance within the dataset (Perseus (Quince *et al.*, 2011) and UCHIME (Edgar *et al.*, 2011)). The database-independent approaches have the advantage of being able to detect chimeric sequences of unknown organisms which is particular useful if one expects a high degree of uncultured bacteria in their sample. Finally, one optional step to accelerate processing times is read de-replication. Here identical reads are compressed into one representative sequence with adjoined abundance count, which considerably reduces the amount of data for later analysis.

Provided their overlapping regions are sufficiently long, paired-end reads can be merged, generating consensus sequences by aligning the forward and reverse reads and resolving any mismatches found in the alignment. This enables calculation of an improved estimate of the number of errors in the read. Depending on the quality of the data, this can be done either at the beginning of the quality filtering or at subsequent steps (Hugerth *et al.*, 2017).

3.4. OTU Clustering

Most of the 16S rRNA gene sequence is conserved, but in the HVRs small numbers of base pairs can change in a very short evolutionary time. Thus, a difference of few base pairs can differentiate between different species. The amplification step can introduce substitution and deletions due to incorrect base pairing and polymerase slipping (Turnbaugh *et al.*, 2007), while during sequencing point errors occur due to incorrect base calls and omitted or spurious base calls. Error rates are often thought to be around 0.01%, hence, in a typical study spanning millions of sequences, it is expected that many sequences have at least one sequencing error. This creates the bioinformatical challenge of distinguishing between PCR or sequencing artefacts and true biological variation. A similarity cut-off of 100% would separate biologically clonal sequences which only differ due to errors in the procedure of generating amplicon reads, obscuring correct taxon identification and inflating diversity measures. This problem is usually solved by clustering sequences at a convenient similarity threshold, typically 97% identity, into so called operational taxonomic units (OTUs) which are used as the equivalent of species in the downstream analysis (Morgan *et al.*, 2012a). OTU clustering is performed by one of three methods, namely closed reference, open reference and de novo clustering. The closed reference approach clusters the pre-processed reads exclusively by aligning them against a sequence collection in a reference database, and any unaligned sequences are discarded. With the open reference approach, the OTU picking is firstly reference-based but instead of discarding any unaligned reads, they will be subjected to de novo clustering (Tatusov *et al.*, 2003, Navas-Molina *et al.*, 2013, Tatusov *et al.*, 2000).

In contrast, de novo clustering groups the sequences without the use of a reference database, which is computationally more expensive but said to be more accurate

(Westcott *et al.*, 2015). This approach is also less stable, meaning that the same sequence might be assigned to a different OTU depending on the community context (He *et al.*, 2015). Hierarchical clustering, which is implemented for instance by mothur (Schloss *et al.*, 2009) and Ortho-MCL (Li *et al.*, 2003), involves gathering sequences based on a distance matrix containing all pairwise differences and can be broadly divided into single-, complete- and average-linkage. For single-linkage, a sequence is placed into a certain cluster if it has a similarity above a threshold to at least one other sequence in the cluster. This approach forms very large, heterogeneous clusters and is rarely used except for clustering sequences of the rapidly evolving fungal ITS region (Lindahl *et al.*, 2013). For complete linkage, on the other hand, sequences with a similarity above a threshold to all other members of the clusters are grouped together, which therefore generates much more and smaller OTUs and has the tendency to overestimate the community richness. Finally, average linkage clusters are constructed when the average similarity between a sequence and all others in the same cluster is above a threshold. It is computationally very demanding to run all against all comparisons on data sets with millions of reads so heuristic techniques like USEARCH (Edgar, 2010) , SEED (Bao *et al.*, 2011), and VSEARCH (Westcott *et al.*, 2015) have been developed. These approximate average linkage clustering by only comparing each sequence to a "centroid" sequence within each cluster, grouping a candidate into an existing cluster if a preselected distance cut-off is met. Centroids are chosen from the input sequences in decreasing abundance, under the premises that high-abundance reads are more likely to be correct amplicon sequences corresponding to true biological sequences.

The selected OTU clustering method can result in very dissimilar outputs for identical datasets with the outputs dependent on the chosen identity threshold, the complexity of the data and the abundance ratios of the reads (Schmidt *et al.*, 2015). Benchmarking of OTU clustering methods found that average-linkage clustering produced the most stable and meaningful OTUs but at the expense of very high computational costs. Thus, heuristic methods are usually preferred, but extra care has to be taken with the quality of the data, the clustering threshold and the actual clustering method (Chen *et al.*, 2013).

3.5. Sub-OTU Methods

Many scientific questions can be addressed by broad taxonomic profiles as provided by OTU clustering, but some applications require higher specificity. Additionally, the biological relevance of the 97% identity threshold has been questioned as it has been shown that 16S rRNA gene sequences with 99% identity represent functionally distinct microorganisms (Patin *et al.*, 2013). Furthermore, the OTU picking strategy has a major effect on the data interpretation and reproducibility and thus tools have been developed that skip the OTU picking, so-called Sub-OTU methods. These methods attempt to correct sequencing errors to determine the true underlying sequence and provide the maximum possible biological resolution for a given dataset by collapsing the reads into so called sequence variants after error correction. The first amplicon sequencing error correction tools were designed to remove noise from 454 sequencing data, but now a collection of Illumina sequencing denoisers (i.e. algorithms that attempt to rectify sequencing errors) are available that differ in the way they attempt to correct sequencing errors (Quince *et al.*, 2011).

The Divisive Amplicon Denoising Algorithm (DADA2) collapses Illumina reads into amplicon sequence variants (ASVs). Its denoising algorithm is based on the Illumina error model which is trained on each sequencing run separately. For each sequencing run a unique mathematical error model is build which considers the properties of each particular run. DADA2 clusters amplicons by considering their abundance distribution as well as the distance from other reads. It uses the above mentioned values in combination with the quality scores provided by the sequencing platform for the given run to calculate a substitution error model. Finally, this model is used to assign low frequency reads to high abundance reads from which they are likely to be derived from by substitution (Callahan *et al.*, 2016).

UNOISE, which is part of the UPARSE pipeline, also denoises reads produced by Illumina sequencers and is optimized for speed. It uses a one-pass clustering strategy that does not use quality scores, but instead uses two parameters with pre-set values, that are believed to work well on a broad range of datasets, to cluster reads into so-called zero radius OTUs, which are an attempt to report all correct biological sequences in a dataset. UNOISE also uses all reads of a dataset simultaneously and

starts by eliminating low abundance OTUs. It then fits an error model based on the premise that the larger the distance between a low-abundance unique sequence and a higher abundance unique sequence, the larger the difference in their relative abundance should be (Edgar, 2013).

In contrast, DEBLUR (Amir *et al.*, 2017) operates on each sample separately and thus substantially reduces the computational demands relative to similar Sub-OTU methods. After sequences are sorted by abundance, the number of predicted error-derived reads is subtracted from neighbouring reads based on their Hamming distance (see Glossary), using an upper bound error probability profile based on a parameterized maximal probability for indels and a parameterized mean read error rate for normalization. Any sequence whose abundance drops to zero after subtraction is considered to be noise and is removed (Amir *et al.*, 2017).

A recent independent benchmarking paper that compared open reference OTU clustering at 97% and the three above mentioned denoising algorithms found that the four methods produce vastly different numbers of OTUs or ASVs, respectively, with open reference clustering finding most OTUs followed by DADA2. But while OTU picking clustered reads on average in the 95%–97% identity range, the sub-OTU methods achieved 99%–100%. Unfortunately, this ability to detect low abundance species seems to come with the trade-off of a higher false positive rate. Overall, all three denoisers resulted in a comparable community structure, suggesting that the choice of denoising tool should not affect biological conclusions that are based on relative abundance data. In terms of computational run time DADA2 was significantly slower than the other two, followed by DEBLUR, and UNOISE3 being the fastest (Nearing *et al.*, 2018).

3.6. Taxonomic Assignment

The next crucial steps in analysis is the inference of community composition via taxonomic classification of the previously determined OTUs. Broadly, there are three strategies for the taxonomic assignment of sequences: word match, best hit and Lowest Common Ancestor (LCA).

1. Word match approaches compare k-mer frequencies between sequences, where the higher the degree of shared k-mer nucleotide frequencies, the more similar the queried sequences. This approach is computationally less expensive than the other two approaches, which require the alignment of the full query sequence against reference databases.
2. In the best hit approach, a query sequence is aligned to every sequence in a reference database and is assigned to the sequence that obtains the highest alignment score.
3. In the LCA method, on the other hand, the query sequence is assigned to the lowest common ancestor if the sequence obtains multiple hits against the reference database.

The Ribosomal Database Project (RDP) classifier (Wang *et al.*, 2007) uses a naïve Bayesian approach to classify sequences on exact matches of k-mers of length 8 in conjunction with bootstrapping to provide a measure of confidence for the given assignment. Initially, SILVA Incremental Aligner (SINA) (Pruesse *et al.*, 2012) also implements a k-mer based search, but then uses the subset of the best matched sequences to build a phylogenetic tree representing the selected sequences. Finally, exact alignments between the query and the reference sequence are calculated and the taxonomy is assigned as the lowest common ancestor of the top-scoring alignments. Another tool supporting taxonomic assignment by the LCA algorithm is MEGAN (Huson *et al.*, 2007), which employs two different LCA approaches, naïve and weighted. In the naïve LCA approach, each read is directly assigned to some taxon matching the LCA criteria, while the weighted approach is a two-step computation. In the first step, all reads get assigned a weight based on the number of reads that uniquely align to a given reference sequence. In the second step, all reads are assigned to the taxon that is bigger than 75% of the total weight of all references for which the read has a significant alignment. This approach leads to a more sensitive taxonomic assignment but at the cost of longer runtimes.

The most cited tool for pairwise alignment is probably BLAST (Altschul *et al.*, 1990) which yields high accuracy at reasonable computational costs, while BLAT (Kent,

2002) and USEARCH (Edgar, 2010) are developed for increased speed, but with the cost of decreased sensitivity.

Regardless of the method used for taxonomic assignment, each tool compares a set of queried sequences against a defined 16S rRNA reference database. Ultimately, the success of this process relies not only on the accuracy and speed of the used classification algorithms, but also on the range of annotated sequences available in public repositories. For bacteria in particular, there are very large reference databases of the 16S rRNA gene, which have been sequenced to a much greater extent than whole genomes. However, on the downside, this means that most information about included organisms are only derived from environmental sequencing rather than single strain sequencing, leading to conflicting annotations between the different databases (Edgar, 2018).

The Ribosomal Database Project (RDP) (Maidak *et al.*, 2001) is maintained by the Center for Molecular Ecology at the University of Michigan and is updated regularly. Its latest release in 2016 contained more than 3 million 16S rRNA gene sequences and approximately 125,000 fungal 28S rRNA sequences. SILVA (Yilmaz *et al.*, 2014) was last updated in 2017 and currently contains over 6 million small subunit sequences and 900,000 large subunit sequences. The SILVA database is maintained by the Microbial Genomics and Bioinformatics Group at the Max Planck Institute for Marine Microbiology in Bremen. Greengenes (DeSantis *et al.*, 2006) is a database of full length 16S rRNA genes that combines the NCBI taxonomy with the cyanoDB (Hauer, 2019) and allows taxonomic identification based on de novo tree inferences. However, it has not been updated since 2013 and hence misses all organisms described since then. *Table 4* shows the aforementioned reference databases for 16S rRNA genes as well as the UNITE reference database for the molecular identification of fungi containing 1.75 million public fungal ITS sequences (Nilsson *et al.*, 2018).

Table 4: Predominant marker gene sequence databases

Database	Size	Latest revision	Target organism	References
SILVA	695,171 non-redundant, 6,073,181 total	December 2017	85% bacteria, some archaea and fungi	(Yilmaz *et al.*, 2014)
RDP	3,356,809 total	September 2016	Comprehensive bacteria and fungi, with some archaea	(Cole *et al.*, 2014)
Greengenes	92,684 non-redundant, 1,012,863 total	May 2013	Bacteria and some archaea	(DeSantis *et al.*, 2006)
UNITE	98,183 non-redundant, 1,750,261 total	September 2019	Fungi	(Nilsson *et al.*, 2018)

Errors in databases due to sequencing or amplification errors as well as incorrect labelling of sequences can lead to misidentification of sequences. It is recommended to use the most diverse database available, or preferably to create personalized reference databases containing only organisms known to inhabit the environment under study allowing an improved assignment of lower taxa level (Werner *et al.*, 2012, Ritari *et al.*, 2015). Any choice of algorithm and database to profile a microbial community has their own inevitable built-in biases. Which method ultimately reveals most of the underlying community is difficult to assess and depends on the specific environment of interest (Golob *et al.*, 2017, De Filippis *et al.*, 2018).

3.7. Copy Number Variation

Another difficulty that arises when studying the microbiome of a given sample with amplicon sequencing is the varying number of 16S rRNA gene copies among different species, the so-called copy number variation (CNV). The copy number per genome varies from 1 to 15 copies between species (Klappenbach *et al.*, 2000), and to a lesser extent, there is even variation among strains of the same species (Acinas *et al.*, 2004). Both the multiplicity and the variability in 16S rRNA gene copy numbers can lead to problems in the assessment of bacterial community structure if not accounted for in the analysis. The multiplicity will skew the abundance estimates of individual taxa along with the relative abundance, while the variability biases the diversity measures. Differential abundance between different samples or treatments groups on the other hand is unlikely to be biased by CNV as different taxa will have the same copy numbers in all groups. In cases where the copy number of the 16S rRNA gene of a species is known, it is recommended to correct abundance estimates with the inverse of the gene copy numbers, i.e. if the gene copy number is 5 for a particular gene, the

counts for that particular gene would be divided by 5. Obviously, this is not quite as straight forward when the copy number of a species is not known. In those cases, possible solutions include using the gene copy number of a closely related species or to estimate them based on phylogenetically independent contrasts (Vetrovsky *et al.*, 2013, Felsenstein, 1985). While tools to correct for copy number variation like CopyRighter (Angly *et al.*, 2014), rrnDB (Stoddard *et al.*, 2015) and the R-package picante (Kembel *et al.*, 2010) have been shown to increase the accuracy of diversity measures, their reliance on databases makes the correction of rare and poorly studied organisms challenging, as they are less likely to be represented.

4. Whole Genome Shotgun Sequence Analysis

As mentioned earlier, whole genome shotgun sequencing is the non-targeted sequencing of all genetic material present in a sample. Since the seminal work by Craig Venter in 2004 (Venter *et al.*, 2004), sequencing microbial populations present in the Sargasso Sea, this sequencing strategy has contributed to breakthroughs in a range of research areas. Most notably, the Human Microbiome Project. This project had the stated aims (i) of understanding the full impact of microbial components on their human host by studying samples from multiple body sites, (ii) investigating the role of the microbiome in health and disease and (iii) providing a comprehensive standardized repository of data to support future research (Peterson *et al.*, 2009). This was expanded upon in the Integrative Human Microbiome Project which focused on microbiome-host interactions (Proctor *et al.*, 2019, Butte *et al.*, 2014). Further work ranges from bacterial associations in Inflammatory Bowel Disease (Gevers *et al.*, 2014), and describing the effect of the human microbiota on host directed drugs (Peppercorn *et al.*, 1972), to tracking the outbreak of human and foodborne pathogens (Loman *et al.*, 2013).

The expansion in applicability of whole genome shotgun sequencing has been mirrored only by the rapid advances in the number of bioinformatics platforms available for analysis of the ensuing data. Whether for aligning reads to reference genomes, assembling reads into contiguous segments of overlapping DNA (contigs), or functionally annotating a metagenome, there is a myriad of possible tools, none of which outperform all others in every circumstance. A comprehensive review of the

relevant literature is outside the scope of this chapter, but is highly recommended before undertaking any analysis of shotgun sequence data.

4.1. Translating Sequence Data into Genomes

The most obvious and important questions in any metagenomics survey is what microorganisms inhabit the ecological niche being characterized, in what proportions, and what do they do. These questions have predominantly been posed of the bacterial constituents of a given microbial community but have been expanded to include viruses and fungi. These questions can be answered in two different ways. While many modern bioinformatics platforms combine metagenomics assembly and downstream analysis into one pipeline, the principles remain the same as for individual genomes.

Whole genome sequencing strategies fall under two umbrella terms, based on whether the target DNA is well characterized, or if novel DNA is expected. Simply put, if the sample is expected to contain previously characterized bacteria, taxonomic and functional annotation can be performed by referring to available databases. If these are not available, sequencing reads must be manually assembled into contigs using a variety of methods. There are advantages and disadvantages to both approaches that must be considered before proceeding beyond this point as outlined below.

The depth of read coverage of the genomes present can dictate which method to use, as 20x coverage is generally recommended as a minimum for genome assembly (Quince *et al.*, 2017). This limits the effectiveness of assembly-based methods in analysing complex microbial communities given the current sequencing technologies available. Genome assembly, and the associated downstream tasks are computationally intensive which can often limit the size of the study, whereas most reference-based methods emphasize efficiency allowing for large scale metagenomic analyses (Vollmers *et al.*, 2017). As is indicated by the name, reference-based methods require a database which contains at a minimum closely related microbes to those found in the samples, whereas assembly-based strategies are able to resolve genomes of novel organisms (Lischer *et al.*, 2017). While some degree of manual supervision is required for reference-based assembly, the tools involved require

minimal intervention when compared to the degree of curation required for accurate contig assembly and taxonomic binning (Quince *et al.*, 2017). The theory of metagenomics assembly is the same in principle as genomic assembly, and the same underlying principles are observed.

4.2. Genomic Assembly

4.2.1. Reference-Based Genome Assembly

Also called assembly-free profiling, this approach makes use of already sequenced reference genomes. The quality filtered short reads generated by a sequencing platform are aligned to a reference genome using an alignment tool. Commonly used alignment tools for single genome methods include Bowtie (Langmead, 2010) and Burrows-Wheeler aligner (Li *et al.*, 2009).

4.2.2. De novo Genome Assembly

This refers to sequencing DNA when no reference genome is available for alignment. Instead, genomes are assembled from the DNA segments yielded from a sequencing experiment, these segments are referred to as reads (Ghurye *et al.*, 2016).

Genome assembly occurs in one of three ways. Greedy assembly is the simplest method, whereby reads with overlapping regions are combined into contigs in an iterative manner starting with the longest overlaps and ending once no further reads can be merged. This method is easy to implement but has considerable drawbacks that limit its use in modern sequencing experiments. Among these, the most glaring is that only local alignments are considered.

The Overlap-Layout-Consensus method assembles by calculating pairwise overlaps between all paired reads, which are computed with an alignment algorithm which enables the method to allow for read errors. Finally, an overlap graph is constructed, where each node is a read and all vertices are overlapped between them (Ghurye *et al.*, 2016). De Bruijn Graphs are constructed by breaking up each sequencing read generated into overlapping fragments of a fixed size. These become the edges and vertices of the graph. The genome is assembled by finding a Eulerian path through this graph, as first described by Leonhard Euler to solve the Bridges of Königsberg problem some 300 years ago (Compeau *et al.*, 2011, Quince *et al.*, 2017). De novo

sequence assembly is commonly assessed using a metric called N50, which is defined as the minimum contig length needed to cover 50% of the genome. However, it is not simply the median over all contigs lengths, but weighted to emphasize the importance of longer contigs (Castro *et al.*, 2017).

These methods were originally developed for single genome assemblies. The fact that metagenomics data contains a variety of different organisms, with varying abundances and degrees of relatedness makes metagenomics assembly considerably more complicated (Ghurye *et al.*, 2016). Metagenomics is firmly established as the go-to technology for analysis of bacterial populations. Unfortunately, assembly methods for these data remain inadequate, which is stifling research, particularly into novel environments with minimal reference material available (Nurk *et al.*, 2017).

4.3. Metagenomic Assembly

As mentioned before, the principles of assembly remain the same for metagenomes, but the process becomes more complicated. Metagenomic resequencing-based methods skip the assembly stage as they rely on existing sequence data for reference and proceed directly to the alignment of reads to taxonomic and functional databases. Many tools exist for the de novo assembly of metagenomes, and while they all incorporate at least one of the three genome assembly methods outlined previously, they differ in how they are applied to the significantly more complicated task of assembling an entire metagenome.

4.3.1. De novo Metagenome Assembly

Most metagenomics assemblers use a modification of the De Bruijn graph approach. However, as sequencing errors dramatically increase the size of the graph, and therefore the processing power required, overlap-layout-consensus methods have returned to the fore for the assembly of data from single molecule technologies, such as PACBIO and Nanopore (Ghurye *et al.*, 2016).

Metagenome assembly is still undoubtedly an imperfect science, and no one assembler can be relied upon to outperform all others in every situation. As such, a variety of assemblers have been designed to supplement the existing established

assemblers IDBA and SPAdes, which have added metagenome specific extensions. IDBA-UD (Peng *et al.*, 2012) is built as an extension on to IDBA, and was specifically designed to take into account the Uneven read Depth of typical metagenomics sequence data. A drawback in terms of implementing this tool is that it was originally designed for read lengths up to 100 bp, and a k-mer size of 12 bp. Modification of this, to suit a longer fragment length requires the modification of some accessory scripts when compiling the tool, undefined for the casual user. MetaSPAdes (Quince *et al.*, 2017, Nurk *et al.*, 2017) first constructs a De Bruijn graph of all reads available, using SPAdes. This is followed by a variety of graph simplification strategies to transform this into an assembly graph. This step constructs the paths corresponding with fragments within the genomes sequenced. Variations between highly similar contigs, potentially due to strain level variation are not considered by MetaSPAdes. The tool instead aims to assemble reliable consensus sequences giving the most accurate representation of species present, without accounting for strain level variation. Unlike the two previously mentioned tools, BBAP (Lin *et al.*, 2017) is an overlap-consensus based genome assembler, making it more suited to sequence data with high error rates, such as those from SMRT technologies or highly polymorphic DNA.

Considering no tool consistently outperforms all others, multiple assembly methods should be attempted and compared before proceeding with this type of analysis. A summary of other commonly used metagenomics assemblers is shown in *Table 5*.

Table 5: Additional options for metagenome alignment

Platform	Description	Multithreading	References
METAVELVET	De Bruijn graph based metagenomics assembler. Initial graphs divided into sub-graphs treated as species from which scaffolds are constructed.	Yes	(Namiki *et al.*, 2012)
MEGAHIT	Assembler optimized for metagenomes by using variety of k-mer sizes, uses De Bruijn graphs	Yes, + GPU options	(Li *et al.*, 2015)
SOAPdenovo2	De Bruijn graph metagenome assembler optimized to reduce memory consumption	Yes	(Luo *et al.*, 2012)
Omega	Overlap consensus based metagenome assembler, suitable for third generation sequencing data	Not possible	(Haider *et al.*, 2014)
Ray meta	De Bruijn graph metagenome assembler	Yes, with run agent.	(Boisvert *et al.*, 2012)

In addition to the tools outlined above, the efficiency of metagenomics assembly can be further improved by either reference assisted de novo assembly or hybrid assembly using multiple sequencing technologies.

4.3.2. Reference Assisted De novo Assembly

Binning of contigs by their species can be done prior to or during assembly (Lischer *et al.*, 2017). Clustering of sequencing reads by taxonomy, referred to as binning, can help to reduce the assembly of mixed genomes. In theory, this will produce one bin per unique genome present in the sample, and is useful when a metagenome contains, or is suspected to contain many closely related species (Roumpeka *et al.*, 2017). A variety of programmes exists to facilitate this, approaching the problem from a number of different angles and accounting for variable levels of complexity. Lightweight tools include LikelyBin (Kislyuk *et al.*, 2009)(, which is based on the underlying assumption that oligonucleotide frequency remains constant within a given bacterial genome, and is therefore only suitable for low complexity communities. Other popular tools for phylogenetic binning include PHYSCIMM (Kelley *et al.*, 2010) and MetaWatt (Strous *et al.*, 2012).

Once assembly has been completed, it is important to assess the quality of that assembly and therefore the performance of the assembly tools. There are a number of statistical methods available for this task, which measure some combination of number of scaffolds assembled, length of scaffolds, coverage rate, N50, and in rarer cases, completeness of gene prediction. Scaffold length can have downstream effects on many components of metagenomics analysis, particularly if the N50 is shorter than gene length. While the N50 is a purely statistical measure that can be manually calculated, MetaQUAST (Mikheenko *et al.*, 2016) is a dedicated platform for the assessment of metagenome assembly quality. This tool is a modified implementation of the original QUAST (Gurevich *et al.*, 2013) algorithm designed for individual assemblies, and evaluates assembly quality based on alignments to closely related reference genomes.

4.3.3. Hybrid Assembly Using Multiple Sequencing Technologies

As mentioned earlier, despite the many advantages of Illumina sequencing, the short read length means the data yielded by this approach is not always suited to de novo sequencing, particularly in genomes with long repeat regions (Utturkar *et al.*, 2014), such as ribosomal DNA operons. As third generation sequencing platforms continue to improve in accuracy, hybrid assemblies combining the high volume of reads yielded by second generation technologies, and the length of reads provided by third generation technologies will become the norm.

4.4. Analysis of Assembled Genomes

Many tools combine both assembly and classification/function analysis, particularly in the case of reference-based methods. Profiling of microbial species present, and their abundance, in an environmental sample can also be carried out without prior genome assembly. Assembly-free profiling is similar in principle to the tools used for single genome alignments to a reference, but adapted for metagenomics use. This approach has a number of advantages, as it is less computationally demanding and can provide information on low abundance organisms that would not be sequenced to sufficient depths for assembly (Quince *et al.*, 2017). Also, reference-based approaches generally require less manual intervention than assembly-based methods. The principal limitation is that even with improvements to the algorithms used, if the biological material present in the sample has not, at a minimum, had a close ancestor sequenced, they are impossible to identify (Quince *et al.*, 2017). Despite this, the complex community structure of common sample sites such as the human gut, make assembly free profiling the more suitable method.

At a fundamental level, assembly free profiling means comparing sequence data yielded from a sequencing experiment to existing databases of microorganism genomes in a way that is accurate and computationally efficient. In practice, this is carried out in four different ways. Sequences can be classified by sequence similarity to a reference genome, similarities in composition such as codon usage, hybrid methods that combine elements of the first two approaches, or finally, marker-based methods. These classify sequences based on the presence of specific marker sequences such as the 16S rRNA gene fragment in bacteria, or the ITS region in fungi

(Peabody *et al.*, 2015). These tools all rely on models derived from reference sequences of existing sequenced genomes.

Simple “brute force” mapping of reads to sequenced genomes in a database leads to spurious false positives in terms of taxonomic classification. A selection of more reliable approaches are outlined below. Kraken (Wood *et al.*, 2014) which uses as default the REFseq database hosted by NCBI extracts k-mers of default length 31 from sequence data and finds the lowest common ancestor in which this is present in the database. It is therefore still a similarity-based method, but considerably faster than BLAST based methods. Relative synonymous codon usage (RCSU), based on the established fact that codons are differentially favoured in different organisms, is a computationally efficient way to discriminate between microbes, particularly at higher levels of taxonomic resolution (Novoa *et al.*, 2019).

MetaPhlAn (Segata *et al.*, 2012), and the more recent extension MetaPhlAn2 (Truong *et al.*, 2015) use clade specific marker genes to characterize microbial communities. As of the most recent release, this classification program contains over 1 million clade specific markers, which equates to approximately 145 markers per bacterial species for over 7000 commonly identified species. Additional functionality for classification of viral and eukaryotic components of metagenomics samples has also been added.

When classifying microbial reads, a balance must be found between accuracy and speed. Similarity based classification methods based on BLAST are often the slowest methods, but modifications of similarity-based classification using short regions, considerably improves on this speed without sacrificing accuracy. A summary of more commonly used metagenomics classification tools is shown in *Table 6* below, with the suggestion that a review of the relevant literature is performed before selection of one.

Table 6: Additional tools for taxonomic profiling

Tool	Organisms detected	Function	References
MiCoP	All microbes including viruses and eukaryotes	Mapping based alignment	(LaPierre *et al.*, 2019)
CLARK	Bacteria and archaea	k-mer based whole genome matching	(Ounit *et al.*, 2015)
MALT	All microorganism if recorded in RefSeq	k-mer based matching and alignment	(Herbig *et al.*, 2016)
Mothur	Bacteria and fungi	Marker gene recognition	(Schloss *et al.*, 2009)

4.5. Variant Calling

Reference based sequence assembly also allows for variant calling analysis. This is the process of identifying variants between closely related sets of sequence data, which typically take the form of single-nucleotide polymorphisms (SNPs). Variant calling analysis based on WGS data has superseded more traditional methods such as PFGE or MLST, as the level of sensitivity to small, localized changes is much higher (Yoshimura *et al.*, 2019).

A typical workflow for variant calling involves aligning WGS sequence data to a reference genome, creating BAM files. This is done using a genome aligner such as Bowtie (Langmead, 2010), or Burrows-Wheeler aligner (Li *et al.*, 2009). Following this, differences between the aligned reads and the reference genomes are identified and written to a variant call file (VCF). This can be done using tools within the SAMtools package (Li, 2011). Lastly, this VCF file must be filtered to ensure results are significant, and not resulting from artefacts of the sequencing process. This can also be done within the SAMtools package.

Variant calling allows for the differentiation of microorganisms at the strain level (Segata, 2018), which can be of crucial significance in metagenomics. For example, when tracking the evolution of antibiotic resistance in methicillin-resistant *Staphylococcus aureus* (Harris *et al.*, 2010). Variant calling can be scaled upwards to process entire metagenomics datasets, and tools such as StrainPhlan (Truong *et al.*, 2017) apply the same principle, but are tuned for the complexities of large mixed samples.

4.6. Analysis of the Functional Capacity of a Metagenome

Once the primary question of what microorganisms inhabit the environment or environments sampled has been answered, the focus usually shifts to what it is that they can do, i.e. the functional potential of a microbial community.

Following an assembly-based approach, methods for characterizing the metabolic potential of a microbial community can again be adapted from methods designed for single genome characterization. There are several gene prediction tools available ranging from basic tools which predict genes based on presence of start and stop codons and GC content, to more complex tools such as GeneMarkHMM (Lukashin *et al.*, 1998) and Glimmer (Kelley *et al.*, 2012). The former is an extension of the original GeneMark (Besemer *et al.*, 2005), which was the first widespread methodology for gene finding in genome projects. It was the tool used for annotations of many of the earliest full genomes to be sequenced such as *Methanococcus jannaschii*. The output is a probability for each fragment of DNA to carry a coding sequence in one of the six possible reading frames, or to be designated as non-coding DNA (Lukashin *et al.*, 1998). Glimmer is more recent, and similarly to GeneMark uses interpolated Markov Models to identify coding regions and distinguish them from non-coding DNA (Kelley *et al.*, 2012).

The functional analysis tools mentioned previously, require that the sequences to be annotated have been assembled to a degree in advance. Two of the biggest challenges faced in gene prediction are short read lengths of sequence data, and the increased error rate that occurs in longer read length sequencing technologies (Mantere *et al.*, 2019). Average gene length in microorganisms is 950 base pairs (Rho *et al.*, 2010), which is considerably longer than the reads available using the more accurate second generation sequencing technology. In these situations, FragGeneScan (Rho *et al.*, 2010) can be used, which is designed specifically with this situation in mind.

As with taxonomic classification, functional annotation is at its most fundamental, the process of identifying coding regions within sequenced genomes, and aligning these to a translated protein database. There are several databases containing

functional information relating to genes and genomes. The Kyoto Encyclopaedia of Genes and Genomes (KEGG) (Kanehisa *et al.*, 2017) is an online database of genomes and genes with the primary aim of assigning functional meaning to both. The information is stored in a hierarchy of different levels as Kegg Orthology (KO) containing molecular level functional annotations, with each annotated KO being homologous to a gene or protein. Higher level functional information for a gene or protein is kept in BRITE hierarchies and KEGG pathway maps (Kanehisa *et al.*, 2017). This database can be queried with amino acid sequence data using the BlastKOALA tool (Kanehisa *et al.*, 2016). The Clusters of Orthologous Groups (COGs) of proteins database is curated by clustering together orthologues from different genomes, with the hypothesis that orthologous genes can be expected to have a conserved function. Functional prediction of proteins is performed by querying which cluster the protein falls into, using the COGNITOR program (Tatusov *et al.*, 2000, Tatusov *et al.*, 2003). UniProt (UniProt, 2008) is a vast database containing both a manually annotated and curated database of protein sequences and functional information in UniProtKB/Swiss-Prot, and a much larger database of automatically annotated records in UniProtKB/TrEMBL (Junker *et al.*, 2000). This database is provided by the UniProt Consortium, which is comprised of the subsidiaries Swiss Institute of Bioinformatics, the Protein Information Resource, and the European Bioinformatics Institute (UniProt, 2014).

Without refinement, analysis of this kind leads to considerable computational bottlenecks, particularly in a large scale metagenomics dataset. It is impossible to manually cross-reference all available data with databases of this type, intermediary programmes such as HuManN2 (Franzosa *et al.*, 2018) and MEGAN (Huson *et al.*, 2007) are therefore required to bridge the gap. MEGAN, currently in its sixth iteration, has an advantage over many other metagenomics platforms in that it is compatible with all Windows, Mac and Linux operating systems. MEGAN first uses the DIAMOND alignment tool to align all reads to a database, typically the NCBI-nr database. MEGAN then takes this alignment file as a reference for binning the reads, functionally and taxonomically. The lowest common ancestor (LCA) algorithm assigns each sequenced read to the lowest taxonomic rank of common ancestor of all

organisms the read in question aligns to. This is repeated for all reads in the dataset. Functional annotation is carried out by searching for the best alignment between a sequenced read, and a functionally annotated DNA sequence from one of the following databases, SEED (Bao *et al.*, 2011), KEGG (Kanehisa *et al.*, 2000), InterPro2Go (Burge *et al.*, 2012) or eggNOG (Huerta-Cepas *et al.*, 2019).

HuManN2 (Franzosa *et al.*, 2018) performs species level functional annotation of metagenomes and metatranscriptomes. Unlike taxonomic profiling, functional profiling quantifies the metabolic potential of a microbial community. HuManN2 uses a "tiered search" strategy to rapidly profile the functional composition of a metagenome. Initially, MetaPhIAn2 is used to identify previously characterized microbes in the sample, and constructs a database per sample, merging existing data with pan genomes of identified species. Following this, reads in a given sample are mapped against the samples pan genome database at the nucleotide level. The reads that do not align are then translated and used to query a protein database which has by default either UniRef90 or UniRef50. The alignments created by this tiered search strategy, once weighted by sequence length and quality of alignment, are used to generate per-organism and total community, gene family abundance (Franzosa *et al.*, 2018).

As with any other aspect of metagenomics research, a thorough understanding of the tools available and their strengths and weaknesses is recommended. A non-exhaustive summary of these tools is shown in ***Error! Reference source not found.***.

Table 7: Commonly used tools for functional annotation of metagenomes

Tool	Gene prediction	Functional annotation of sequences	Functional profiling	Interface	Databases used	References
HuManN2	No	No	Yes	Local	MetaCyc	(Franzosa *et al.*, 2018)
MEGAN	No	Yes	Yes	Local	COG, KEG, SEED	(Huson *et al.*, 2016)
COGNIZER	No	No	Yes	Local	COG, KEGG, Pfam, SEED, GO, FIG	(Bose *et al.*, 2015)
metaCLADE	No	Yes	Yes	Local	Pfam	(Ugarte *et al.*, 2018)
MOCAT2	Yes	Yes	Yes	Local	KEGG, MetaCyc, Pfam, SEED, ARDB, and others.	(Kultima *et al.*, 2016)
MG-RAST	No	Yes	Yes	Web based	GenBank, SEED, RefSeq, KEGG, UniProt and others.	(Keegan *et al.*, 2016)

For more specific investigations such as antibiotic genes, databases such as the Comprehensive Antibiotic Resistance Database (CARD) (McArthur *et al.*, 2013), the Antibiotic Resistance Genes Database (ARDB) (Liu *et al.*, 2009) and resfams (Gibson *et al.*, 2015) can be used to complement functional annotation. CARD is an intensively curated database consisting of data on antimicrobial resistance at the genetic level. Another database of interest is the Virulence Factor Database (VFDB) (Chen *et al.*, 2005), a curated database of microbial pathogens including in particular foodborne pathogens. Unlike CARD, VFD has an inbuilt search tool called VFanalyzer, but in the cases of both databases and the numerous others that exist, search tools such as BLAST (Altschul *et al.*, 1990) and alignment tools such as BOWTIE (Langmead, 2010) or BWA (Li *et al.*, 2009) can be used to query sequencing results against these databases. *Table 8* shows a non-exhaustive list of bioinformatics tools and platforms for metagenomics analysis.

5. Further Analysis

Metagenomics data is usually summarized as a table of read counts per OTU/ASV or gene/genome per sample. Those tables tend to be very sparse, where counts of zeroes may mean the true absence of an OTU/ASV or gene/genome or that its presence is below the detection limit. This detection limit can vary between samples due to differences between sequencing runs and an unequal representation of samples in pooled sequencing libraries. One way of bioinformatically dealing with this problem is to discard instances which are observed in less than a certain percent of samples, proportion of reads or given number of independent samples (de la Cuesta-Zuluaga *et al.*, 2016).

Once a metagenomics dataset has been characterized taxonomically and functionally, statistical comparisons between groups, experimental conditions or time series experiments among others can be carried out using regular parametric and non-parametric methods within traditional multivariate statistical approaches. Beyond this, several traditional ecological methods can also be applied to microbial ecology, and packages such as vegan (Oksanen J, 2022) and phyloseq (McMurdie *et al.*, 2013, Oksanen J, 2022) exist within the R environment to facilitate their use.

Table 8: Non-exhaustive list of bioinformatics platforms for metagenomics analysis

Category	Details	Platform and references
Quality Filtering Tools	Tools for read assessment, read filtering and trimming	FASTQC (Andrews, 2010), FASTX-Toolkit (Hannon, 2010), Trimmomatic (Bolger *et al.*, 2014), PRINTSEQ (Schmieder *et al.*, 2011), NGS QC Toolkit (Patel *et al.*, 2012), meta-QC-chain (Zhou *et al.*, 2014b)
Metagenomic assemblers	Tools for alignment and merging of reads generated by sequencing into longer contiguous regions or scaffolds to reconstruct the original sequence found in the sample.	IDBA-UD (Peng *et al.*, 2012), metaSPAdes (Nurk *et al.*, 2017), Meta-IDBA (Peng *et al.*, 2011), MetaVelvet (Namiki *et al.*, 2012), MetaVelvet-SL (Afiahayati *et al.*, 2015), (Lin *et al.*, 2017)BBAP(Lin *et al.*, 2017), ray meta (Boisvert *et al.*, 2012), MEGAHIT (Li *et al.*, 2015), MetaBAT (Kang *et al.*, 2019), Omega (Haider *et al.*, 2014), SOAPdenovo2 (Luo *et al.*, 2012), Minimus2 (Sommer *et al.*, 2007)
Shotgun read binning	Tools to group contigs or reads based on specific features to classify taxonomically or aid assembly	MetaBin (Sharma *et al.*, 2012), LikelyBin (Kislyuk *et al.*, 2009), PHYSCIMM (Kelley *et al.*, 2010), MetaWatt (Strous *et al.*, 2012), MaxBin (Wu *et al.*, 2016), SortITEMS (Monzoorul Haque *et al.*, 2009)
Reference based classification methods for shotgun data	Approaches are based on aligning to reference genome, similarities in codon	Kraken (Wood *et al.*, 2014), MetaPhlAn (Truong *et al.*, 2015), MiCoP (LaPierre *et al.*, 2019), CLARK (Ounit *et al.*, 2015), MALT (Warinner *et al.*, 2017), mothur (Schloss *et al.*, 2009), Kaiju (Menzel *et al.*, 2016), DeepMicrobes (Liang *et al.*, 2020), MSC, k-SLAM (Ainsworth *et al.*, 2017), MegaBLAST (Zhang *et al.*, 2000)
Functional annotation of shotgun data	Tools to predict genes in shotgun data Tools to carry out translated search of functional databases	GeneMarkHMM (Lukashin *et al.*, 1998), Glimmer (Kelley *et al.*, 2012), FragGeneScan (Rho *et al.*, 2010), MetaGeneMark (Zhu *et al.*, 2010), MetaPath (Liu *et al.*, 2011), HuManN2 (Franzosa *et al.*, 2018), MEGAN (Huson *et al.*, 2016), COGNIZER (Bose *et al.*, 2015), metaCLADE (Ugarte *et al.*, 2018), MOCAT2 (Kultima *et al.*, 2016), MG-RAST (Keegan *et al.*, 2016)
Functional databases	Databases containing translated protein sequences and their assigned function. Can be manually or computationally annotated.	MetaCyc (Caspi *et al.*, 2020), COG (Tatusov *et al.*, 2003), KEGG (Kanehisa *et al.*, 2017), SEED (Bao *et al.*, 2011), pfam (Finn *et al.*, 2014), GO (Ashburner *et al.*, 2000), ARDB (Liu *et al.*, 2009), GenBank (Sayers *et al.*, 2019), RefSeq (Rajput *et al.*, 2019), eggNOGv5 (Huerta-Cepas *et al.*, 2019)
OTU-picking	Tools to perform OTU clustering with open reference, closed reference or *de novo* methods	Mothur (Schloss *et al.*, 2009), OrthoMCL (Li *et al.*, 2003), DOTUR (Schloss *et al.*, 2005), CD-HIT (Li *et al.*, 2006), ESPRiT-Tree (Cai *et al.*, 2011), DNACLUST (Ghodsi *et al.*, 2011), HmmUFOtu (Zheng *et al.*, 2018), SEED (Bao *et al.*, 2011), VSEARCH (Westcott *et al.*, 2015)
Sub-OUT methods	Tools to generate amplicon sequence variants	DADA2 (Callahan *et al.*, 2016), UNOISE2 (Edgar, 2016), Deblur (Amir *et al.*, 2017), AmpliconNoise (Quince *et al.*, 2011), CFF (Tikhonov *et al.*, 2015)
Chimaera removal	Tools for removing chimeric sequences	UCHIME (Edgar *et al.*, 2011b), ChimeraSlayer (Haas *et al.*, 2011), perseus (Quince *et al.*, 2011), CATCh (Mysara *et al.*, 2015), DECIPHER (Wright *et al.*, 2012)
16S rRNA gene amplicon classifiers	Software to assign taxonomy to 16S marker gene OTUs	Mothur (Schloss *et al.*, 2009b), RDP Classifier (Wang *et al.*, 2007), UTAX (Edgar, 2013), 16S Classifier (Chaudhary *et al.*, 2015, Edgar, 2018), SPINGO (Allard *et al.*, 2015a), SINA (Pruesse *et al.*, 2012)
Amplicon databases	Reference databases for marker genes	SILVA (Pruesse *et al.*, 2007), Greengenes (DeSantis *et al.*, 2006), RDP (Maidak *et al.*, 2001), rrnDB (Stoddard *et al.*, 2015), PhylOPDb (Jaziri *et al.*, 2014), UNITE (Nilsson *et al.*, 2018)
16S rRNA gene pipelines	Packages of tools and commands to analyse 16S marker gene data from raw data to visualization	QIIME2 (Bolyen *et al.*, 2019), mothur (Schloss *et al.*, 2009b), SILVA (Yilmaz *et al.*, 2014), MEGAN (Huson *et al.*, 2016), FASTGroup (Seguritan *et al.*, 2001), PANGEA (Giongo *et al.*, 2010), CLOTU (Kumar *et al.*, 2011), JAguc (Nebel *et al.*, 2011), DADA2 (Callahan *et al.*, 2016), MICCA (Albanese *et al.*, 2015), FunFrame (Weisman *et al.*, 2013), UPARSE (Edgar, 2013), MG-Rast (Keegan *et al.*, 2016)

These can be complemented by multidimensional scaling tools which are extremely important for the visual representation of high dimensional data and are facilitated by the ape package (Paradis *et al.*, 2004) within the R environment.

Specific tools tailored to metagenomics data also exist, which have been modified to be more sensitive to the specific nature of the data. For instance, the problem of calculating differential abundance across experimental conditions, be it of taxa or gene expression, is one that has been significantly improved on with the development of bespoke bioinformatics tools. The DeSeq algorithm was developed for RNAseq data, with the aim of finding genes that are differentially expressed across treatment groups, samples or time points based on the negative binomial distribution. The second iteration of DeSeq2 has been extended to other types of HTS data and can now be used as a tool for detecting gradients of bacterial colonization between experimental conditions and biomarker discovery (Love *et al.*, 2014). MetagenomeSeq was specifically designed for marker gene surveys such as 16S rRNA gene sequencing and the detection of differentially enriched bacterial taxa but can equally be used for count tables generated by whole genome shotgun sequencing experiments. It addresses the effects of both normalization and under-sampling of microbial communities and also incorporates the testing of feature correlations (Paulson *et al.*, 2013). The Anova-Like differential expression tool for high through put sequencing data (ALDEx2) uses underlying assumptions of compositionality (Fernandes *et al.*, 2013).

6. Reproducibility and Benchmarking

There are a multitude of bioinformatics platforms available for metagenomics analysis, and not only the tools of choice but also how a specific pipeline is used, can have an effect on the conclusions drawn by the resulting data. As every step of a metagenomics study can bias the result and change our perception of the underlying microbial community, it is vital to keep all variables consistent throughout a study and include them in the method section. Apart from the DNA extraction method and choice of sequencing technologies, and in case of amplicon sequencing studies which 16S rRNA region was targeted with which exact primers, this also must include a

description of how DNA contamination was controlled for, how the sequencing error rate was assessed and importantly an in depth description of the in silico analysis. It is not sufficient to only report the used tools but also the versions and specific parameters (if divergent from the defaults) have to be indicated. To allow reproducibility and comparison of studies it is recommend to not only make the generated data of study available but also the implemented code used to analyse the data (De Filippis *et al.*, 2018, Golob *et al.*, 2017, Ye *et al.*, 2019).

When developing a suite of tools or pipeline for analysis, the benchmarking of different potential platforms is an effective way to ensure the developed pipeline suits the experimental needs. Tools such as MetaSim (Richter *et al.*, 2008) provide the raw materials for such a benchmarking project. This allows the user to define and simulate a sequencing dataset of known microbial composition, and consequently to assess the accuracy and speed of potential metagenomics platforms (Richter *et al.*, 2008). In terms of reproducibility, an important initial step in promoting this is to deposit the results of any sequencing experiment into one of the online sequence data repositories available. Two of these are the Sequence Read Archive (SRA) (Leinonen *et al.*, 2011) and Metagenomic Rapid Annotations using Subsystems Technology (MG-RAST) (Wilke *et al.*, 2016).

7. Data Deposition

The sequence read archive, or SRA, is one of the most important platforms involved in metagenomics research. It is the primary repository of high throughput sequencing data hosted by the National Institute of Health in the United States, and part of the International Nucleotide Sequence Database Collaboration. A wide range of sequencing data is accepted, such as Roche454, Illumina and Pacific Biosystems data. All data submitted to this portal is publicly available, and serves the purpose of aiding new discoveries by increasing access to data, and promoting the reproducibility of the field of metagenomics, which is at present one of the key weaknesses of the field (Leinonen *et al.*, 2011). MG-RAST is another repository for sequence data. It currently stores over 150,000 datasets with over 23,000 of them in the public domain. As the name suggests, it also provides some limited functionality in metagenomics analysis. Raw reads can be uploaded in fastq format, which are then taxonomically and

functionally annotated with minimal user input. Further analysis and visualization is then possible through an interactive web server (Wilke *et al.*, 2016).

8. Concluding Remarks

The sheer number of different bioinformatics tools and pipelines available can seem daunting when planning a metagenomics project. The first step should always be a thorough review of the literature to ensure that the optimal strategy is used based on the research question, the computational power, and bioinformatics experience available. Benchmarking papers are particularly useful as they can illustrate the variable output given by many of these different pipelines. As computing power increases, big data analysis of metagenomic data and the storage of said data will become more accessible as less specialized equipment is required. However, reproducibility of results, already a point of contention, is a key challenge for the future. Diligent storage in public repositories of sequence data and robust annotated code is the responsibility of all metagenomics researchers.

Aim of this thesis

Clinical and experimental research consistently highlights alterations in gut microbial composition among individuals with IBD, underscoring the pivotal role of the gut microbiota in disease pathogenesis. However, the presence of high inter-personal heterogeneity, inconsistency across studies, and numerous confounding factors have hindered the establishment of a definitive microbial signature for IBD. Consequently, the full potential of microbiome research for personalized medicine remains untapped.

Therefore, this thesis aims to conduct a comprehensive investigation into the multifaceted influences of lifestyle and environmental factors, treatments for IBD and comorbidities, as well as microbial metabolites of the gut microbiota in individuals with IBD. Additionally, it seeks to explore novel technologies, particularly machine learning approaches, for navigating the complexities of such convoluted data.

Chapter II of this thesis focuses on quantifying the relative contributions of various lifestyle and environmental factors to the compositional variability of the gut microbiota within a large longitudinally sampled IBD cohort. Whereas Chapter III aims to untangle microbiota variance, shedding light on the extent to which differences in treatments may contribute to apparent geographic influences on gut microbial composition.

In Chapter IV, the attention is directed towards the microbes-metabolic axis in individuals with Crohn's disease, aiming to elucidate how the bidirectional interactions between bile acids and microbiota in the gut influence the microbial composition.

Finally, Chapter V investigates the utility of machine learning approaches as tools for querying microbiome data. Specifically, it examines whether certain pre-processing techniques or algorithms exhibit superior performance in an endeavour to provide insights into the strengths and weaknesses of different commonly used ML algorithms and to help guide future application of ML in microbiome research.

Chapter II

Ranking microbiome variance in inflammatory bowel disease: A large longitudinal intercontinental study

Adam G Clooney[1,2], Julia Eckenberger[1,2*], Emilio Laserna-Mendieta[1,2], Kathryn A Sexton[3], Matthew T Bernstein[3], Kathy Vagianos[3], Michael Sargent[3,4], Feargal J Ryan[1,2], Carthage Moran[1,5], Donal Sheehan[1,5], Roy D Sleator[1,6], Laura E Targownik[3,4], Charles N Bernstein[3,4], Fergus Shanahan[1,5], Marcus J Claesson[1,2]*

1. APC Microbiome Ireland, University College Cork, Cork, Ireland
2. School of Microbiology, University College Cork, Cork, Ireland
3. The University of Manitoba Inflammatory Bowel Disease Clinical and Research Centre, Winnipeg, Manitoba, Canada
4. Section of Gastroenterology, Department of Internal Medicine, University of Manitoba, Winnipeg, Manitoba, Canada
5. Department of Medicine, University College, Cork, Ireland
6. Department of Biological Sciences, Cork Institute of Technology, Cork, Ireland

Published in: Gut, 2021. 70(3): p. 499-510, DOI: 10.1136/gutjnl-2020-321106

Author contributions: *MJC, FS, LT* and *CNB* designed and managed the project; *AGC, JE* and *EL- M* performed the analyses; *FS, CNB, CM* and *DS* provided clinical samples and expertise; *AGC, JE, MJC* and *FS* wrote the manuscript; *MJC, RDS* and *FS* secured funding. *KAS* and *MTB* created and managed the database for Canadian subjects, *KV* administered the FFQ to Canadian subjects, *MS* handled the stool samples and measured fCAL in Canadian subjects. *AGC* and *JE* contributed equally.

Supplementary material:
https://gut.bmj.com/highwire/filestream/239187/field_highwire_adjunct_files/1/gutjnl-2020-321106supp002_data_supplement.xlsx

Table of Contents

Abstract 92
1. Introduction 93
2. Methods 93
2.1. Study Populations, Sample and Data Collection 93
2.2. Sample Processing and Sequencing 94
2.3. Bioinformatic analysis and statistics 95
3. Results 97
3.1. Microbial composition in IBD differs from that of controls 97
3.2. Species associated with disease 101
3.3. Hierarchical clustering reveals groups dominated by prevalent species 104
3.4. Temporal stability 106
3.5. Age, diet and geography affect the microbiota 106
3.6. Medications and bowel resection associated with a changing microbiota 108
4. Discussion 111
5. Acknowledgements 112
6. Appendix 113

Abstract

Objective: The microbiome contributes to the pathogenesis of inflammatory bowel disease (IBD) but the relative contribution of different lifestyle and environmental factors to the compositional variability of the gut microbiota is unclear.

Design: Here, we rank the size effect of disease activity, medications, diet and geographic location of the faecal microbiota composition (16S rRNA gene sequencing) in patients with Crohn's disease (CD; n=303), ulcerative colitis (UC; n = 228) and controls (n=161), followed longitudinally (at three time-points with 16 weeks intervals)

Results: Reduced microbiota diversity but increased variability was confirmed in CD and UC compared with controls. Significant compositional differences between diseases, particularly CD, and controls were evident. Longitudinal analyses revealed reduced temporal microbiota stability in IBD, particularly in patients with changes in disease activity. Machine learning separated disease from controls, and active from inactive disease, when consecutive time points were modelled. Geographic location accounted for most of the microbiota variance, second to the presence or absence of CD, followed by history of surgical resection, alcohol consumption and UC diagnosis, medications and diet with most (90.3%) of the compositional variance stochastic or unexplained.

Conclusion: The popular concept of precision medicine and rational design of any therapeutic manipulation of the microbiota will have to contend not only the heterogeneity of the host response, but also with widely differing lifestyles and with much variance still unaccounted for.

1. Introduction

Clinical and experimental research has linked the gut microbiota with the pathogenesis of Crohn's disease (CD) and ulcerative colitis (UC) (Bernstein *et al.*, 2017, Nishida *et al.*, 2018, Sheehan *et al.*, 2017, Zuo *et al.*, 2018, Ryan *et al.*, 2020). Reduced microbial diversity and other non-disease-specific changes in the microbiome, which probably reflect changes secondary to inflammation, have been reported by several investigators. In addition, some have recently reported apparent disease-specific changes (Franzosa *et al.*, 2019, Pascal *et al.*, 2017, Yilmaz *et al.*, 2019, Halfvarson *et al.*, 2017). However, diet, medications, ethnicity, geography, and a multitude of lifestyle or environmental variables may confound the interpretation and replication of microbiome studies. This is particularly problematic with small studies. Few large-scale longitudinal studies across stages of disease activity have been conducted (Halfvarson *et al.*, 2017, Rampton *et al.*, 2016, Lin *et al.*, 2014) particularly with intercontinental comparisons. Therefore, we undertook a study over different time-points of a large population of patients with CD and UC from two socioeconomically developed but geographically separated regions (Ireland and Canada), during periods of active (relapse) and inactive (remission) disease and ranked the contribution of variables to microbiota composition.

The results confirm microbiota compositional instability in patients with fluctuating disease activity, but show that geographic location (which may reflect ethnic and lifestyle differences) ranks higher in its contribution to microbiota variance than common variables such as prior surgical resection, age, gender, and diet. However, the majority of the compositional variance remains unaccounted. The findings have implications for the prospect of personalized therapeutics based on microbial manipulation.

2. Methods

2.1. Study Populations, Sample and Data Collection

This study was approved by the Cork hospitals' research ethics committee and the University of Manitoba Health Research Board. All patients had well-established diagnoses by conventional and investigative criteria (Rampton *et al.*, 2016). Patients were not invited to comment on the study design and were not consulted to develop

patient relevant outcomes or interpret the results. Patients were not invited to contribute to the writing or editing of this document for readability or accuracy. An active state of IBD was defined as a faecal calprotectin measurement of ≥250µg/g (Lin *et al.*, 2014). Patients from Cork, Ireland (n=283) and Manitoba, Canada (n=409), were sampled at three separate time points approximately 16 weeks apart (Table 1). All subjects completed a Food Frequency Questionnaire (FFQ) to capture long-term dietary habits through frequencies of medium food servings of 157 items (Claesson *et al.*, 2012, Harrington *et al.*, 2011), The FFQ used in Ireland was adapted for the Canadian participants to reflect common food items consumed by the Canadian population (Clooney *et al.*, 2016a). These frequencies were normalized to per month counts and were grouped into 28 broader food categories.

2.2. Sample Processing and Sequencing

All samples were processed in a single laboratory in Cork using the same protocols. Irish subjects brought samples to a morning-time gastroenterology clinic which were promptly delivered to the microbiology laboratory and frozen at -80°C. Canadian samples were stored at -80°C, before shipment to Cork on dry ice. There were no inadvertent episodes of freeze-thawing. An experiment was conducted to investigate the time delay in transporting the Canadian samples (see Results).

Aliquots of approximately 0.2 g were transferred into tubes with one 3.5 mm glass bead, 0.1 mL of 1.0 mm zirconia/silica beads and 0.1 mL of 0.1 mm glass beads (Biospec, Bartlesville, USA). QIAamp Fast DNA stool kit (Qiagen GmbH, Germany) was used for DNA extraction. First, 1 mL of InhibitEX buffer was added to the faecal samples, which were disrupted by bead-beating in a Mini-Beadbeater-24 (Biospec) for 30 s, thrice at maximum speed (3450 strokes/min). Samples were heated at 95°C for 5 mins and subsequently processed according to the kit instructions resulting in genomic DNA eluted in 200 µL ATE buffer. DNA concentrations were measured using a NanoDrop 2000 Spectrophotometer (Thermo Scientific, USA) and stored at -80°C until 16S rRNA gene amplicon sequencing library preparation. Library preparation was performed following Illumina (San Diego, USA) recommendations. Aliquots of 15 ng of extracted DNA were subjected to PCR amplification of the 16S rRNA V3-V4 hypervariable region in a total volume of 30 µL. Primers (Forward: 5'-

CCTACGGGNGGCWGCAG-3', Reverse: 5'-GACTACHVGGGTATCTAATCC-3') were selected from Klindworth *et al.* and Illumina adapters, containing overhang nucleotide sequences, were added to the gene-specific sequences (Klindworth *et al.*, 2013). The PCR primers (Eurofins Genomics, Germany) at a concentration of 0.2 µM were used with the Phusion High-Fidelity DNA polymerase (Thermo Scientific, USA). PCR amplification was performed on a 2720 Thermal Cycler (Applied Biosystems, USA) at 98°C for 30 s, followed by 25 cycles of 98°C for 10 s, 55°C for 15 s, 72°C for 20 s and 72°C for 5 min. Post-PCR products of the amplified 16S rRNA gene band was verified in an agarose gel and purified using Agencourt AMPure XP magnetic beads (Beckman-Coulter, USA) and eluted in 50 µL of EB Buffer (Qiagen). After purification, 5 µL of DNA was amplified in a second PCR employing Nextera XT Index primer (Illumina) run at 98°C for 30 s, and followed by eight cycles of 98°C for 10 s, 55°C for 15 s, 72°C for 20 s and 72°C for 5 min, followed by a second purification step with Agencourt AMPure XP magnetic beads. The amplicons containing the Nextera indexes were finally eluted in 25 µL of EB Buffer and DNA concentration quantified using Qubit high sensitivity dsDNA assay kit (Thermo Scientific). A randomized pooled library was created by adding 40 ng of each sample. Finally, a diluted sample of this library with a 30 nM concentration was sent for MiSeq sequencing (Illumina) to Eurofins Genomics.

An additional -80°C frozen aliquot containing approximately 0.5 g of faecal material was used to measure calprotectin concentration in an ImmunoCap 250 autoanalyser employing EliA^{TM} Calprotectin Immunoassay v2 (Phadia-Thermo Scientific, Sweden). The BÜHLMANN fCAL ELISA kit was deployed for faecal calprotectin measurement in the Canadian cohort. Results were expressed as µg/g feces for both assays.

2.3. Bioinformatic analysis and statistics

First, "*cutadapt*" (Martin, 2011) was used to remove the adapter sequences from the V3-V4 region of the 16S rRNA amplicon reads allowing for an error rate of 0.2. USEARCH (v.8.1.1861) script "*fastq_merge*" was employed to merge forward and reverse reads. The QIIME script "*split_libraries*" was used to demultiplex retaining reads with a minimum average PHRED quality of Q25 and a length of 390-465bp. *De novo* clustering of Operational Taxonomic Units (OTUs) was carried out using

"*cluster_otus*" in USEARCH, while chimeras were removed using "*uchime_ref*" with the ChimeraSlayer GOLD database(Edgar *et al.*, 2011). The Mothur implementation of the Ribosomal Database Project (RDP) classifier (v.11.4) was used to classify OTU taxonomy (phylum and genus) with a bootstrap cut-off of 80%, and all others assigned as unclassified at that particular rank (Schloss *et al.*, 2009). For species classification and clostridium clusters SPINGO (v.1.3) (Allard *et al.*, 2015) was performed against the RDP database (v.11.4) with similarity score of 0.5 and bootstrap cut-off of 0.8. Sequence data are available at NCBI SRA PRJNA414072.

All downstream analysis was performed in R v3.6.0. Alpha and beta diversity was calculated using R package "*phyloseq*" while plots were constructed in "*ggplot2*". Differential taxa abundance was done using "*MetagenomeSeq*" (v.1.26.3), while differences in diet consumption were calculated using Wilcoxon tests. Heatplots were constructed using the "*made4*" library with Spearman correlations combined with hierarchical Ward-clustering. Hierarchical clusters were cut using "*dynamicTreeCut"* with a minimum cluster size of 75 samples and tested with Fisher's tests for the categorical data and the Wilcoxon tests for quantitative data. Spearman correlations were used to assess correlations between food groups/taxa/metadata and PCoA/PCA axes. Adonis from the "*vegan*" library based on Bray-Curtis dissimilarity was performed to investigate and rank the effect of 14 environmental factors including medications on overall microbial composition. The FFQ data were condensed into one factor, the Healthy Food Diversity (HFD) Index, as previously described (Drescher *et al.*, 2007), using individual food items. For the different therapies and disease activities, the variance was adjusted for IBD group to quantify effects between active/inactive and treated/untreated, rather than the effect of healthy/IBD sufferer. Cumulative explained variance was calculated in a greedy stepwise approach, whereby environmental factors contributing with the highest R^2 value (P-value<0.1) were iteratively added to the model in the previous iteration. To distinguish between health/disease and active/inactive disease, gradient boosted trees models were generated via "*xgboost*" based on either proportional normalized OTUs present in at least 5% of the samples, or ratios between these OTUs of two time points from patients with non-changing disease states. Optimal model parameters

were determined with bootstrapping of 1000 iterations and 5-fold cross-validations. For each binary classification *n* leave-one-out models were generated to predict the class of the n^{th} sample. OTU importance was determined by the gain an OTU added to a model and the frequency with which each OTU was used for a model. Performance of each classification was measured by Area Under the Receiver Operating Characteristics (AUROC). P-values were adjusted for multiple testing where appropriate, using the Benjamini and Hochberg method (Benjamini *et al.*, 2001).

3. Results

3.1. Microbial composition in IBD differs from that of controls

We collected 1,816 stool samples from 303 patients with CD, 228 patients with UC and 161 healthy controls from the regions of Manitoba, Canada (59% of subjects) and Cork, Ireland. The samples were longitudinally collected from subjects who were either in relapse (active; 34% of 1,515 IBD samples) or in remission (inactive) with time intervals of on average 16 weeks (95% CI: 15.6-17.0). See *Table 9* for subject characteristics.

The 16S rRNA V3-V4 region of all samples were amplified, sequenced and subjected to quality and chimera filtering in Cork, resulting in a mean of 21,647 (95% CI: 21,298-21,996) usable reads per sample. These were clustered into 3,148 OTUs with ≥97% identity and analyzed further. To ensure that the time lag in mailing the Canadian samples to Cork did not confound microbiota results, we analyzed four fresh samples stored in duplicates for up to 7 days at room temperature. There was no significant difference between time points in terms of microbiota alpha diversity, nor was any OTU differentially abundant between samples processed at different time points (*Supplementary Figure 1*).

Beta diversity analysis based on Bray-Curtis distances (*Figure 8a*) showed a significant disease-associated shift along Principal Components (PCs) 1 and 2, with CD samples located furthest from healthy controls, followed by UC. Samples from patients with active disease were further from controls than those from patients in remission (inactive) for both UC along PC1 (p-value<0.0.13) and CD along PC2 (p-value<0.012).

Table 9: Subject characteristics and sample sizes for the a) Canadian and b) Irish cohorts.

a)	MANITOBA				
Cohort	**CD**		**UC**		**non-IBD Control**
Number of subjects	186		144		79
Gender (%female)	72		58.3		36.7
Age (mean +/- SD)	54.4 ±12.6		56.9±12.6		57.6±15.9
Age of disease onset (mean =/- SD)	30.5±11.9		36.1±13.2		NA
Smoking Status					
Non (%)	40.1		41.1		35.4
Ex (%)	44.3		44.2		24
Current (%)	9.5		6.1		2.1
Alcohol Consumption					
> 2 units per day (%)	3.3		4.9		11.5
< 2 units per day (%)	47.9		49.1		38.5
Non (%)	42.9		37.5		11.5
Sample type	**Active**	**Inactive**	**Active**	**Inactive**	**non-IBD Control**
Number of samples	238	340	167	324	96
Prior resection (%)	47.5	41.8	91.6	91.7	NA
Medication					
Biologics (%)	17.6	8	4.2	2.8	0
5-ASA (%)	26.9	25	50.9	49.4	0
Mercaptopurine (%)	13.9	24.1	11.4	11.4	0
Corticosteroids (%)	5.9	3.5	9.6	3.7	1
Proton pump inhibitors (%)	26.1	20.9	23.4	11.1	31.3
Other GIT drugs (%)	19.3	22.1	13.8	7.1	2.1
Antibiotics (%)	2.9	4.4	4.2	0.9	0
Supplements (%)	60.9	66.8	47.9	45.4	21.9
NSAI Drugs (%)	28.2	28.5	31.1	23.1	18.8
Diabetes drugs (%)	1.7	1.7	6.6	4.9	0

b)	CORK				
Cohort	**CD**		**UC**		**non-IBD Control**
Number of subjects	117		84		82
Gender (%female)	48.7		36.9		50
Age (mean +/- SD)	41.0 ±11.7		45.6 ±12.6		40.88 ±16.5
Age of disease onset (mean =/- SD)	26.4 ±10.6		30.8 ±12.1		NA
Smoking Status					
Non (%)	39.8		51.9		55.3
Ex (%)	48.1		41.5		14.5
Current (%)	12		6.6		6.1
Alcohol Consumption					
> 2 units per day (%)	12		9.9		5
< 2 units per day (%)	42.3		47.2		31.8
Non (%)	45.6		41.5		14
Sample type	**Active**	**Inactive**	**Active**	**Inactive**	**non-IBD Control**
Number of samples	55	180	61	150	179
Prior resection (%)	65.5	52.2	93.4	98	NA
Medication					
Biologics (%)	27.3	20.6	11.5	10.7	0
5-ASA (%)	40	32.8	57.4	57.3	0
Mercaptopurine (%)	14.5	22.2	18	24.7	0
Corticosteroids (%)	25.5	10.6	24.6	7.3	1.1
Proton pump inhibitors (%)	18.2	18.3	4.9	17.3	2.2
Other GIT drugs (%)	14.5	10	3.3	6.7	1.1
Antibiotics (%)	0	0	1.6	0.7	0
Supplements (%)	12.7	20.6	14.8	18	10.1
NSAI Drugs (%)	14.5	4.4	0	8	1.7
Diabetes drugs (%)	3.6	3.9	4.9	4	0

Several food groups and other metadata (discussed later) were significantly correlated to PC1 and/or PC2 axes and, thus, associated with the overall variation in microbiota composition (*Figure 8b*; *Supplementary Table 1*). Among the strongest correlates were geographic location, resection, age of onset, and food supplements. However, the overall strongest correlation was for alpha diversity, which was lower in patients with IBD. Irrespective of correlations with beta-diversities, alpha diversity was significantly lower in patients with CD and UC compared to controls (*Figure 8c*; *Supplementary Table 2*), and also in active vs inactive UC. For individual subjects (pair-wise version of Figure 8c), active UC showed significantly lower alpha diversity than the corresponding inactive sample from the same subject (*Supplementary Figure 2*).

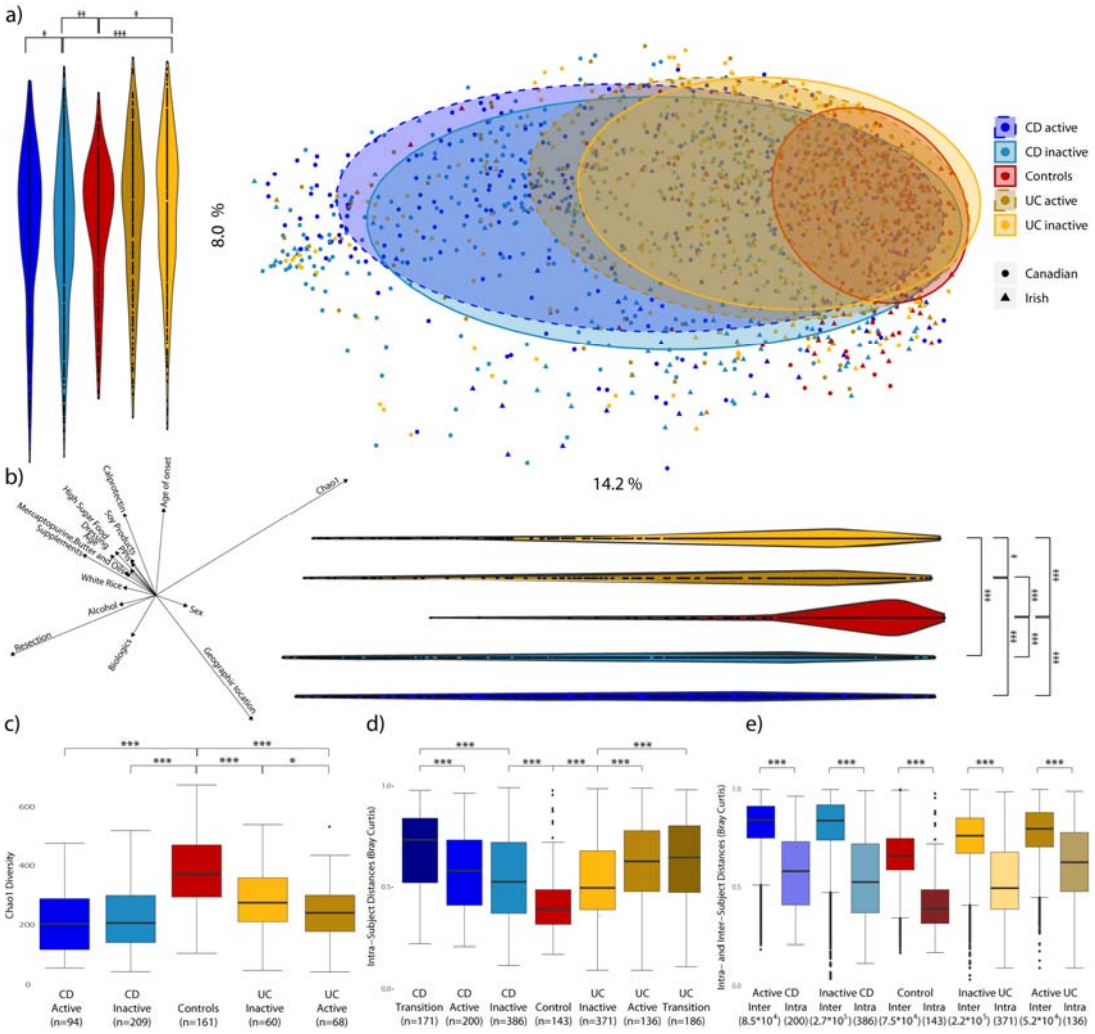


*Figure 8: Microbiota composition and diversity in CD, UC and control subjects. a) Principle Coordinate Analysis based on Bray-Curtis distances with proportional normalization on all OTUs present in more than 5% of samples, with samples grouped by disease and activity state. Violin plots show projections of PCoA points into PC1 and PC2. b) Spearman correlations between PC axes and food groups/metadata and alpha diversity. Only categories with significant correlations are represented. The direction and length of the arrows indicate the direction and strength of the correlation. c) Comparison of distances between time points from the same subject based on Bray-Curtis distances. d) Comparison of Chao1 diversity for disease type and status of patients with IBD versus healthy controls. e) Comparison of intra- and inter-individual Bray-Curtis distances. P-values: *< 0.05; ** < 0.01; *** <0.001.*

3.2. Species associated with disease

The OTUs were classified to species level and filtered down to 200 species present in at least 5% of the samples in order to remove outlier taxa with low prevalence. MetagenomeSeq analysis showed a greater number of species with significantly higher abundance in disease compared to controls. Of these, 68 were significantly decreased and 35 significantly increased in CD relative to controls. Similarly, 60 species were significantly decreased, and 27 species were increased in UC compared to controls (*Figure 9; Supplementary Tables 3-5*). The most significantly increased species in CD compared to controls were *Ruminococcus gnavus* and *Eggerthella lenta*. In contrast, several *Eubacteria (E. eligens* and *E. rectale)* and *Faecalibacterium prausnitzii species* were decreased in CD when compared to controls. *Eggerthella lenta* was also increased in UC versus controls along with *Holdemania filiformis and Clostridium innocuum*, whereas *Eubacterium eligens* and *Clostridium aminobutyricum* were decreased in UC relative to controls. There were no significant differences between active and inactive CD, or active and inactive UC.

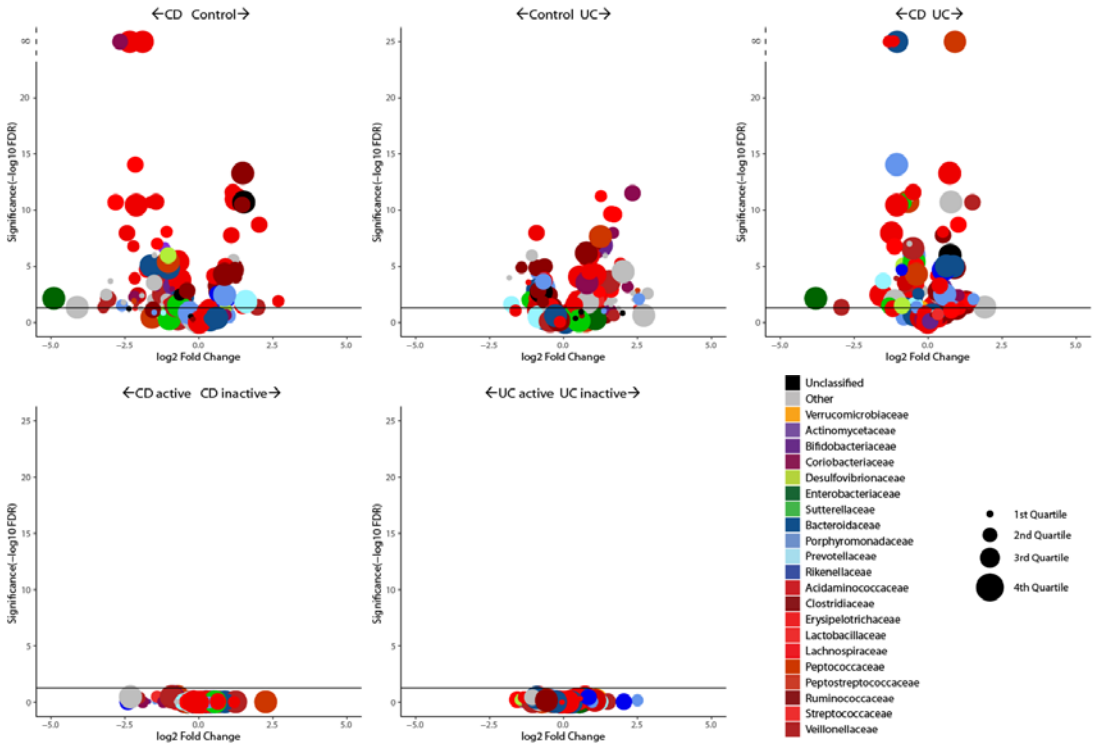


Figure 9: Differential species abundances (volcano plots) between disease groups and disease state. Points above the horizontal line are significant while the x-axis position of each point indicates the direction of fold change. The size of each point refers to the abundance of the species across the cohort while the color indicates the family rank.

A microbial signature for CD has been suggested using a Boolean algorithm on eight differentially abundant taxa resulting in accuracies between 64-82% and 77-85% to distinguish patients against UC and healthy controls, respectively (Pascal *et al.*, 2017). Applying the same algorithm and genera to our dataset generated lower accuracies of 61% (UC vs CD) and 68% (Control vs CD). A more powerful approach, particularly for heterogeneous datasets (Chen *et al.*, 2016), is the Machine Learning technique Extreme Gradient Boosting. With this method, we observed an Area Under the Curve (AUC) of 0.88 (84% accuracy) for CD vs control, and 0.88 (83% accuracy) for UC vs control, with *Eubacterium rectale* and *Clostridium* cluster XIVa being the most important discriminatory OTUs in both diseases, respectively (*Figure 10*; *Supplementary Table 6*). The AUC for CD vs UC was, however, much lower (0.67; 64% accuracy). Repeating the classification within each location only marginally changed the AUCs (*Figure 10*), whereas intercontinental cross-validation reduced AUCs somewhat with an average of 0.071 (*Supplementary Figure 3; Supplementary Table 7*). Curiously, classifying disease activity was only possible by using longitudinal within-subject change between two consecutive time points as ratios of each OTU, as opposed to OTUs from a single time point. With the former method, we obtained higher AUCs of 0.81 (81% Accuracy; *H. saccharovorans* most important) for CD, 0.73 (85%; *Bifidobacterium*) for UC and 0.91 (89% Accuracy; *A. hadrus*) for patients combined (*Figure 11; Supplementary Table 8*).

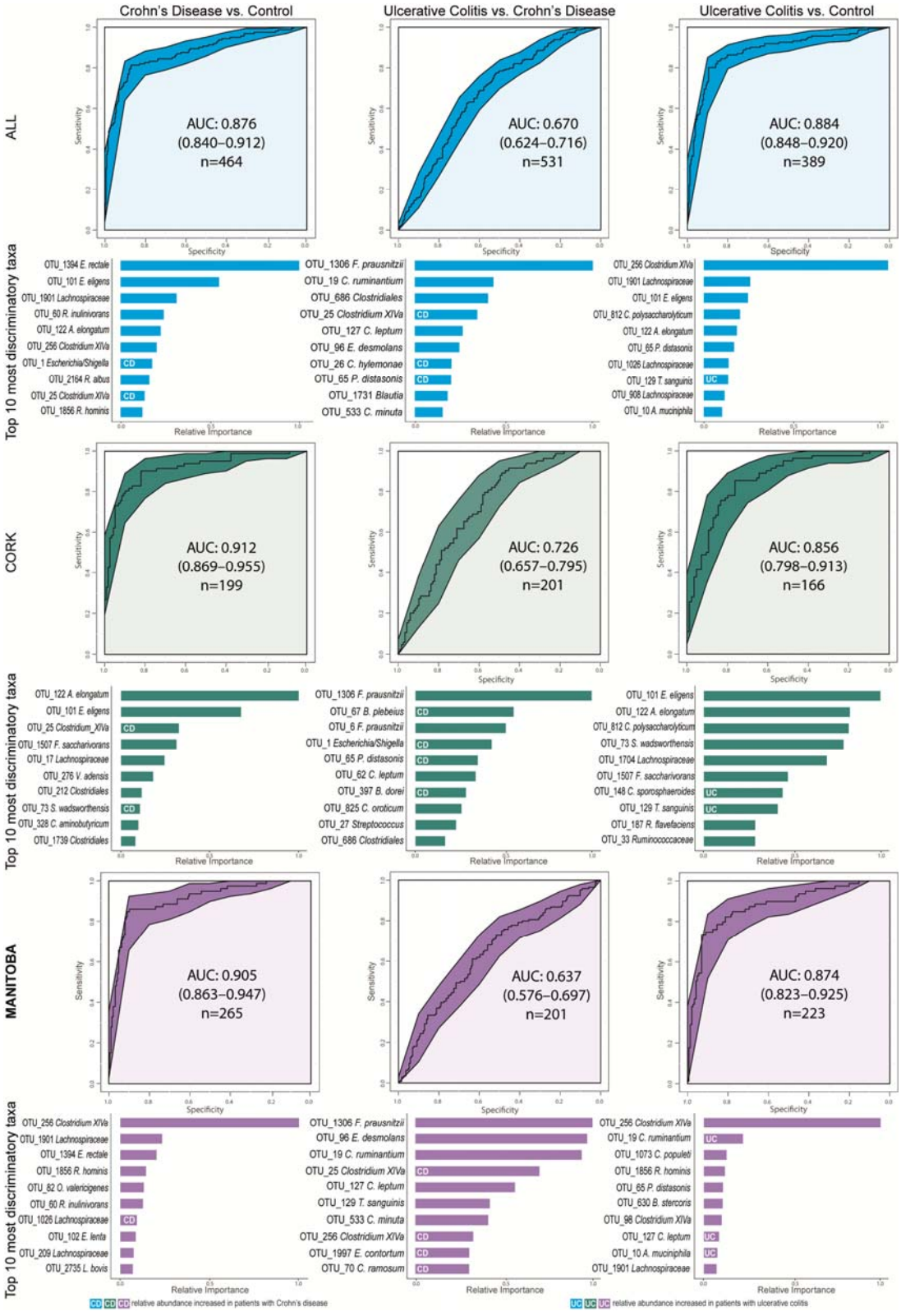


Figure 10: Machine learning classification of the subject cohorts, first combined and then separated by geographic location. Receiver operating characteristic curves (ROC) for the boosted tree models on the bases of proportional normalized OTUs present in more than 5% of samples. Below each ROC-curve variable importance plots show the relative importance for the 10 OTUs with the highest gain for each comparison alongside their highest known classification. The white UCs, CDs within the bars indicate which OTUs are increased in their respective patient group, bars without mark indicate that their respective taxon is increased in the other class of the model. A model with an AUC of 0.5 has no discriminatory capacity, whereas an AUC of 1 indicates perfect separation of the response variables.

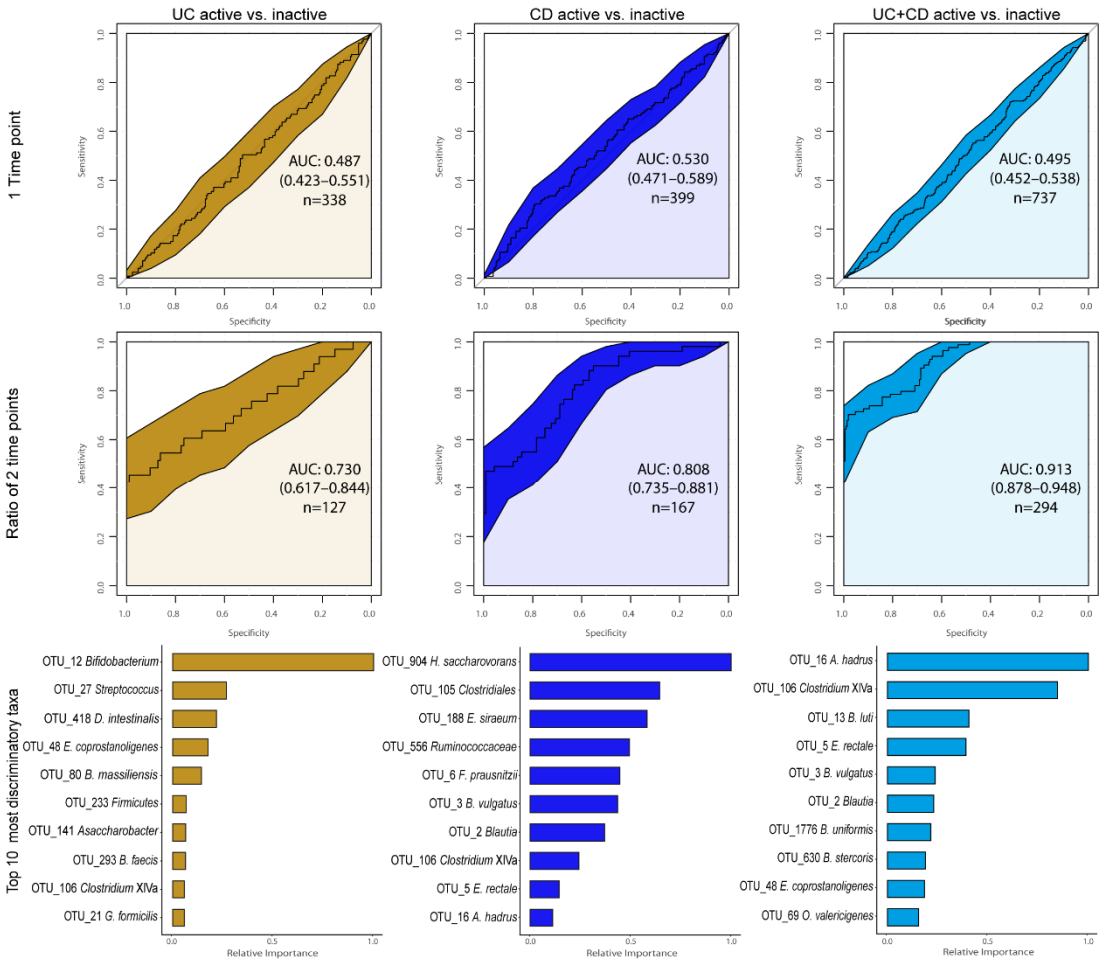


Figure 11: Single and dual time-point Machine Learning classification of disease activity for patient with CD and UC separately and combined cohorts based on proportional normalized OTUs present in more than 5% of samples. The ROCs for the boosted tree models on the second row were generated based on the ratio of each OTU between two consecutive time-points. Only subjects that did not transition between disease states were included, Canadian and Irish combined. Below each ROC-curve are variable importance displayed indicating the relative importance for the 10 OTUs with the highest gain for each model.

3.3. Hierarchical clustering reveals groups dominated by prevalent species

In view of the heterogeneity of UC and CD, we assessed potential stratification based on microbiota composition. Hierarchical clustering revealed 10 sub-groups (clusters) based on their proportional species composition, with several clusters dominated by particular species (*Figure 12*). Clusters ranging in size between 79 and 328 samples were often distinguished by varying relative family abundances. Among these, clusters 7-10 had significantly more (mainly Canadian) patients with IBD than the other clusters. Cluster 1 (*Bacteroides vulgatus*), cluster 7 (*Escherichia/Shigella*) and cluster 8 (*Akkermansia muciniphila* for 1/3 of subjects) contained significantly more subjects post resection, while clusters 1, 2 (*Prevotella copri*) and 10 had more subjects taking biologics (anti-TNF) (*Supplementary Figure 4*; *Supplementary Table 9*).

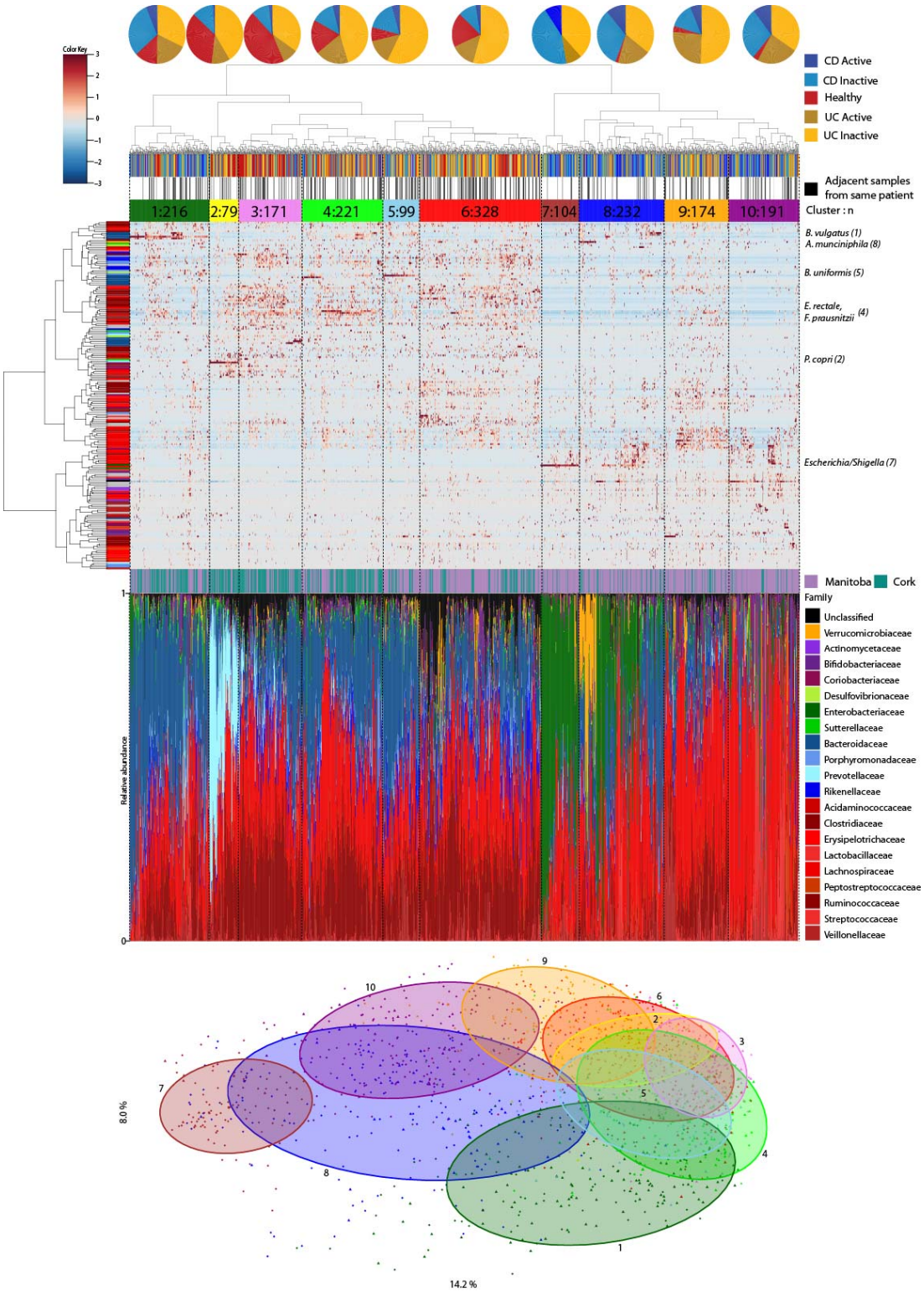


Figure 12: Hierarchical clustering of stool microbiota. Heatplot of OTUs classified at species level with spearman correlations and ward-linkage clustering. The vertical color bar represents the family of each species. The samples were divided into 10 clusters (tertiary horizontal color bar) using dynamicTreeCut. The first horizontal color bar indicates the disease and status of each sample, the second horizontal bar shows samples of the same patient that clustered adjacent to each other. The species names on the right of the plot, along with the cluster number, are the drivers of the clustering. The bar chart shows the taxonomic composition at family level of each sample ordered by phylum. The PCoA at the bottom is the same as Figure 8 but labelled by the 10 clusters.

3.4. Temporal stability

Temporal stability was assessed by comparing within-subject Bray Curtis distances across multiple time points. Patients with CD or UC had significantly greater within-subject differences (reduced microbiota stability) compared to controls (*Figure 8d; Supplementary Table 10*). In addition, we observed significantly higher within-subject transitional microbiota differences in samples taken across different activity stages for both CD and UC, than in samples from patients whose disease activity did not change. Inter-individual differences were also substantially larger than intra-individual differences for all cohorts (*Figure 8e*).

We further investigated temporal stability by testing whether samples from the same subject had a higher tendency to be adjacent to each other in the hierarchical tree clustering (*Figure 12*). Of all samples in this tree, 36% were closest to samples from the same subject. In support of the temporal observations above (*Figure 8d; Supplementary Table 11*), samples from the same (either CD or UC) subject, but with transitioning activity states, were less likely to be next to each other than with a sample with non-changing activity state. Samples from the same control subjects were more likely to be next to each other than samples for patients with CD or UC.

3.5. Age, diet and geography affect the microbiota

Separating the beta diversity PCoA into the two geographical locations (*Supplementary Figure 5*) showed that CD microbiota from Manitoba subjects were to a higher degree responsible for the horizontal shift observed in *Figure 8a* and *Supplementary Figure 6*, primarily attributed to resection (see next section). There were otherwise no major differences between disease, activity and control cohorts across the two locations. The alpha diversity for Cork controls and UC inactive subjects was significantly increased compared to the Manitoba cohort (*Supplementary Figure 7*). As the latter cohort were older than the Cork cohort (*Table 9*), we adjusted for geographical location, but still found the same two Cork cohorts having significantly higher alpha diversities than the corresponding Manitoba cohorts. Age was also negatively correlated with alpha diversity for Cork inactive and Manitoba active CD cohorts (*Supplementary Figure 8a*). We observed significant shifts in microbiota composition between the two locations along both principal

components also for separate cohorts (*Supplementary Figure 6*). Thus, a number of species were significantly different between the two locations with 68 species for CD (13 active, 51 inactive), 57 for UC (13 active, 27 inactive), and 20 different between the control groups (*Supplementary Figure 9*; *Supplementary Tables 12-18*). The carbohydrate fermenting *Clostridium leptum* was the most significantly abundant species in Manitoba subjects for both CD and UC compared to Cork, overall and for inactive states.

Several food groups (high sugar foods, brown pasta, poultry, red meats, alcohol and brown bread) were significantly correlated with overall microbiota composition, as illustrated by correlations to the two PC axes in *Figure 8b* (*Supplementary Table 1*). Across both geographical locations, consumption of high sugar foods, nuts, butter and oils were increased in subjects with IBD, while controls consumed more fruit, vegetables, high-fiber and breakfast cereals, sauces and alcohol (*Figure 8b; Supplementary Table 19*). Additionally, 20 food groups were consumed differently between the population cohorts (*Supplementary Figure 10*) with potatoes, beans, white bread, processed meats, sauces and breakfast cereals significantly increased in the Cork cohort, while among others vegetables, high-sugar foods, nuts, pasta and rice were increased in the Manitoba group.

We found microbiota diversity to be positively correlated with the HFD index for all cohorts combined and for Manitoba UC inactive (*Supplementary Figure 8*). Unexpectedly, microbiota diversity in patients with inactive UC in Cork was negatively correlated with the HFD index, in contrast to their Canadian counterparts. A microbiota-independent PCA based on the 157 food items showed that the long-term dietary habits of patients with CD and UC were somewhat different from controls with a significant shift away from controls along PC2 (*Figure 13*). The main drivers of this shift were brown pasta/rice and dressing (higher in controls) and butter and oils (higher in IBD) (*Figure 13; Supplementary Table 20*).

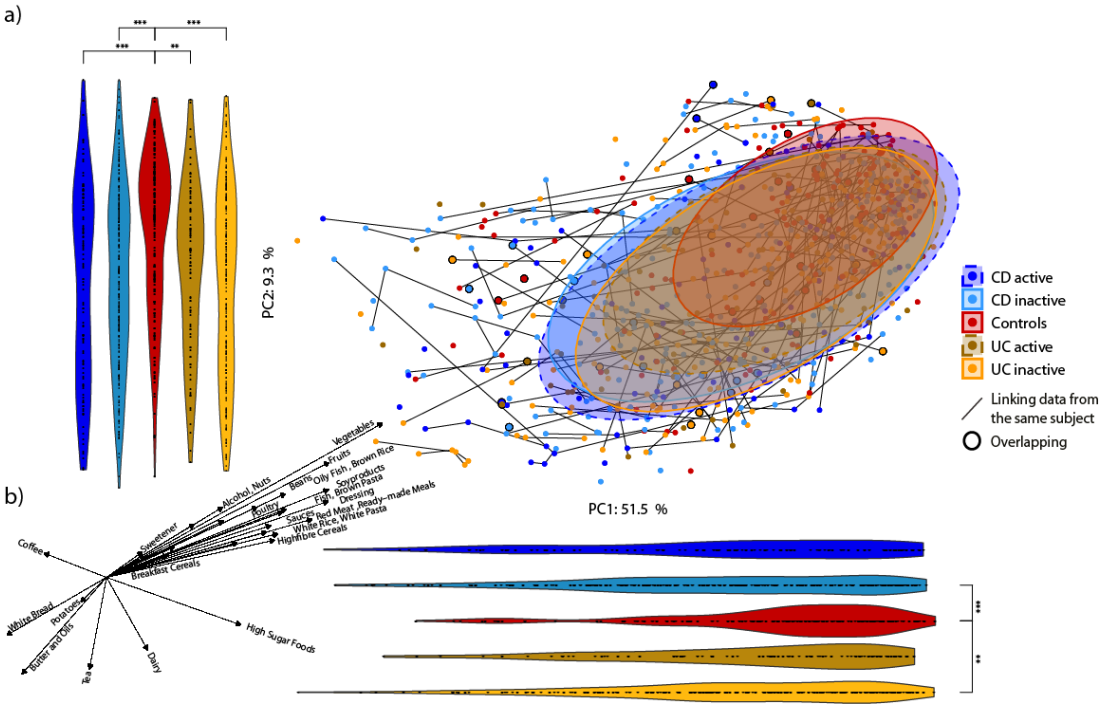


Figure 13: Food composition in CD, UC and control subjects. a) Principle Component Analysis (PCA) of the food groups colored by disease and disease status. Violin plots represent the points in the PCA projected to the PC1 and PC2 axes to assess shifts in the groups. Patients with greater than 1 identical questionnaire are indicated by black circles. b) Spearman correlations between PC axis of PCA and food groups/metadata. Only features with significant correlations are represented. The direction and length of the arrows indicate the direction and strength of the correlation.

3.6. Medications and bowel resection associated with a changing microbiota

Both microbiota diversity and composition were significantly associated with bowel resection and the use of some medications. Resected subjects had lower alpha diversity than non-resected subjects and controls (*Supplementary Figure 11*), and also showed a significant shift in beta diversity away from non-resected subjects and from healthy controls (*Supplementary Figure 12*). Resection was a treatment primarily within our CD cohorts; thus, when comparing these subjects it became evident that resection in Manitoba had a particularly marked effect on the microbiota composition (*Supplementary Figure 13*). The significant shifts along PC2 away from controls and non-resected subjects had the same directions for both countries, but was much more pronounced for Manitoba.

There were up to 63 species differentially abundant (*Supplementary Table 21-25*) between any of these resection/non-resection groups, including *Faecalibacterium prausnitzii* and *Clostridium leptum* being increased in non-resected subjects, and

Blautia producta, *Bacteroides fragilis* and *Ruminococcus gnavus* being increased in resected subjects.

Subjects on proton pump inhibitor (PPIs) medication were located further away from controls than non-users and had also significantly lower alpha diversity (*Supplementary Figure 12; Supplementary Table 26*). Two out of 10 significantly increased species in PPI-users belonged to the *Streptococcus* genus (*S. agalactiae* and *S. mutans*). The only medicated subjects with microbiota composition closer to controls (higher PC1 values) were those taking 5-aminosalicylates (*Supplementary Figure 12*). Similarly, alpha diversity (*Supplementary Figure 11*) was significantly higher for patients on 5-aminosalicylates, and controls, compared to those who were not. However, these changes were not attributable to any particular species.

Finally, we tested how much microbiota variance (beta-diversity) was explained by the 25 environmental factors for 650 subjects at one randomly selected time point (*Figure 14*). Presence or absence of a CD diagnosis had the greatest impact on the microbiota, followed by geographic location, previous surgical resection, alcohol consumption and UC diagnosis. Diet (summarized in the HFD Index), anthropometrics and medications also explained some variation in the microbial composition. The reason why smoking and disease activity showed non-significance may be due to the relatively small and uneven sizes of these groups (*Table 9*). It should, however, be noted that when combined, these factors explain less than 9.7% of the total microbiota variance.

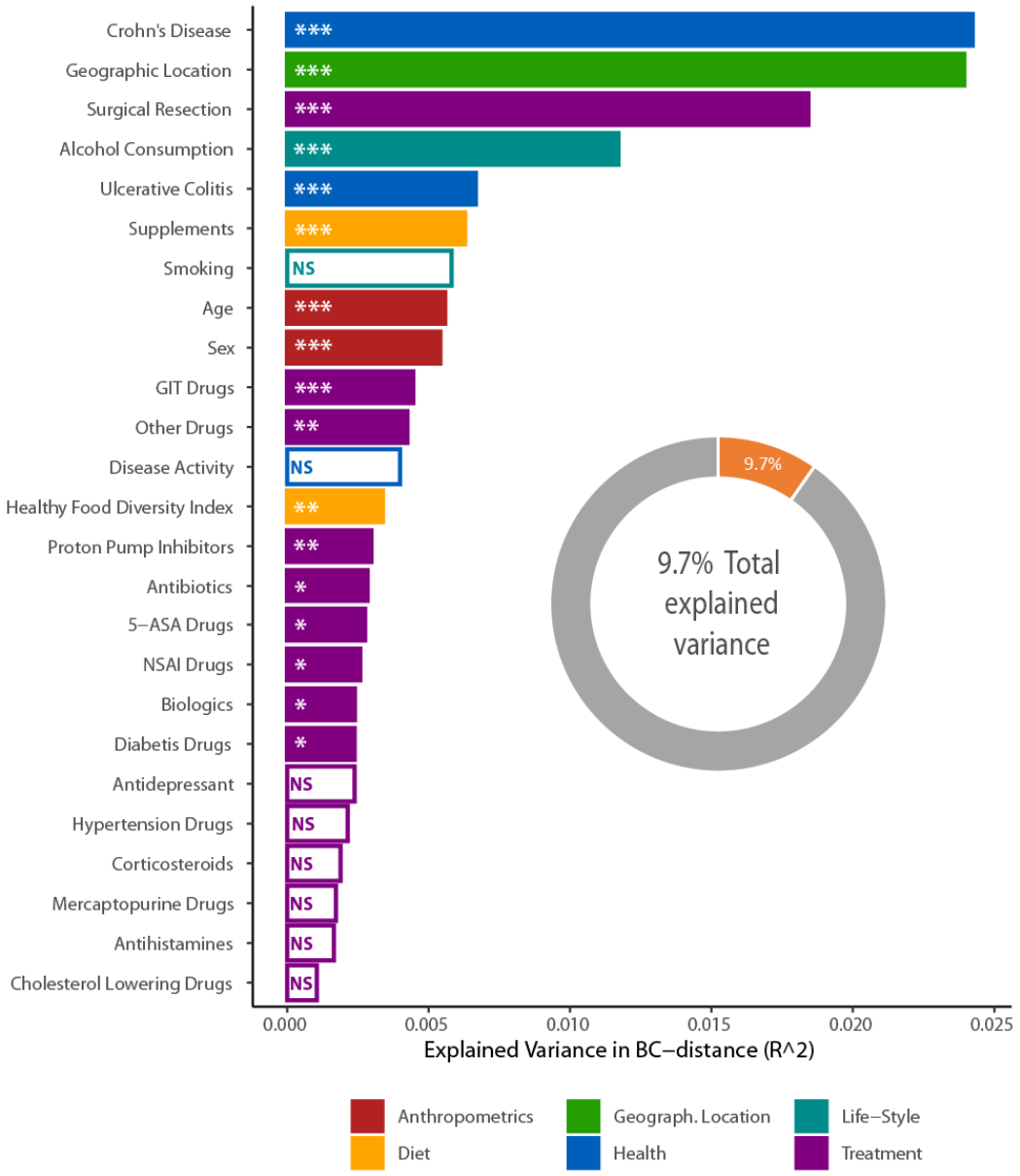


*Figure 14: Factors explaining microbiota variance. Association between 25 tested environmental factors and the microbiota beta diversity of n=650 individuals (291 CD, 236 UC, 120 controls; one randomly chosen time-point per subject) at OTU-level in terms of explained fraction of the variance in Bray-Curtis dissimilarity. Of available 692 subjects, 42 were excluded due to missing values in the metadata. Adjusted PERMANOVA P-values: *< 0.05; ** < 0.01; *** < 0.001.*

4. Discussion

The relationship between microbiota and pathogenesis of IBD is well established, but there is limited information to rank the influence of known variables on microbiota composition, including longitudinal follow up at different phases of disease activity. Our results show that disturbances in faecal microbial composition are most marked in active disease, particularly in CD. The longitudinal study design allowed for microbiota dynamics to be investigated, showing that inter-individual variance was greater than intra-individual, and that the microbiota was more unstable in CD and UC compared to controls. This is consistent with earlier reports of smaller cohorts (Scanlan *et al.*, 2006, Halfvarson *et al.*, 2017, Lloyd-Price *et al.*, 2019). The greatest changes in microbiota composition were linked with transitions across active and inactive phases of disease. The importance of longitudinal sampling became even more apparent with machine learning models capable of separating active and inactive disease, even though specific taxa were not linked with different activity states. This was only possible when using intra-individual ratios of each OTU between two consecutive time-points, but not when using OTUs from single time-points. Biologically, this is intuitively obvious since inflammation and other disease-related factors (such as medications to treat inflammation, antibiotic use, surgery, etc.) are more likely to lead to microbiota variability in patients with IBD than in controls. Thus, frequent and long-term sampling might permit an improved classification, and potentially relapse-predicting, models (Vazquez-Baeza *et al.*, 2018).

As expected, faecal microbiota alpha diversity was reduced in both CD and UC, but in contrast to microbiota composition, diversity did not vary significantly with disease activity. Of the microbial species found to be significantly increased in CD compared to controls, we concur with previous reports for *Ruminococcus gnavus* (Willing *et al.*, 2010, Joossens *et al.*, 2011) and *Fusobacterium nucleatum* (Gevers *et al.*, 2014), although the opposite has also been reported for the former species (Gevers *et al.*, 2014). We found decreased abundance in CD for *Ruminococcus albus, Eubacterium rectale and Faecalibacterium prausnitzii* compared to controls, in agreement with previous studies (Kang *et al.*, 2010, Gevers *et al.*, 2014) and meta-analysis (Cao *et al.*, 2014). *Eubacterium* and *Roseburia* species were among the most important taxa for classifying CD and UC compared to controls.

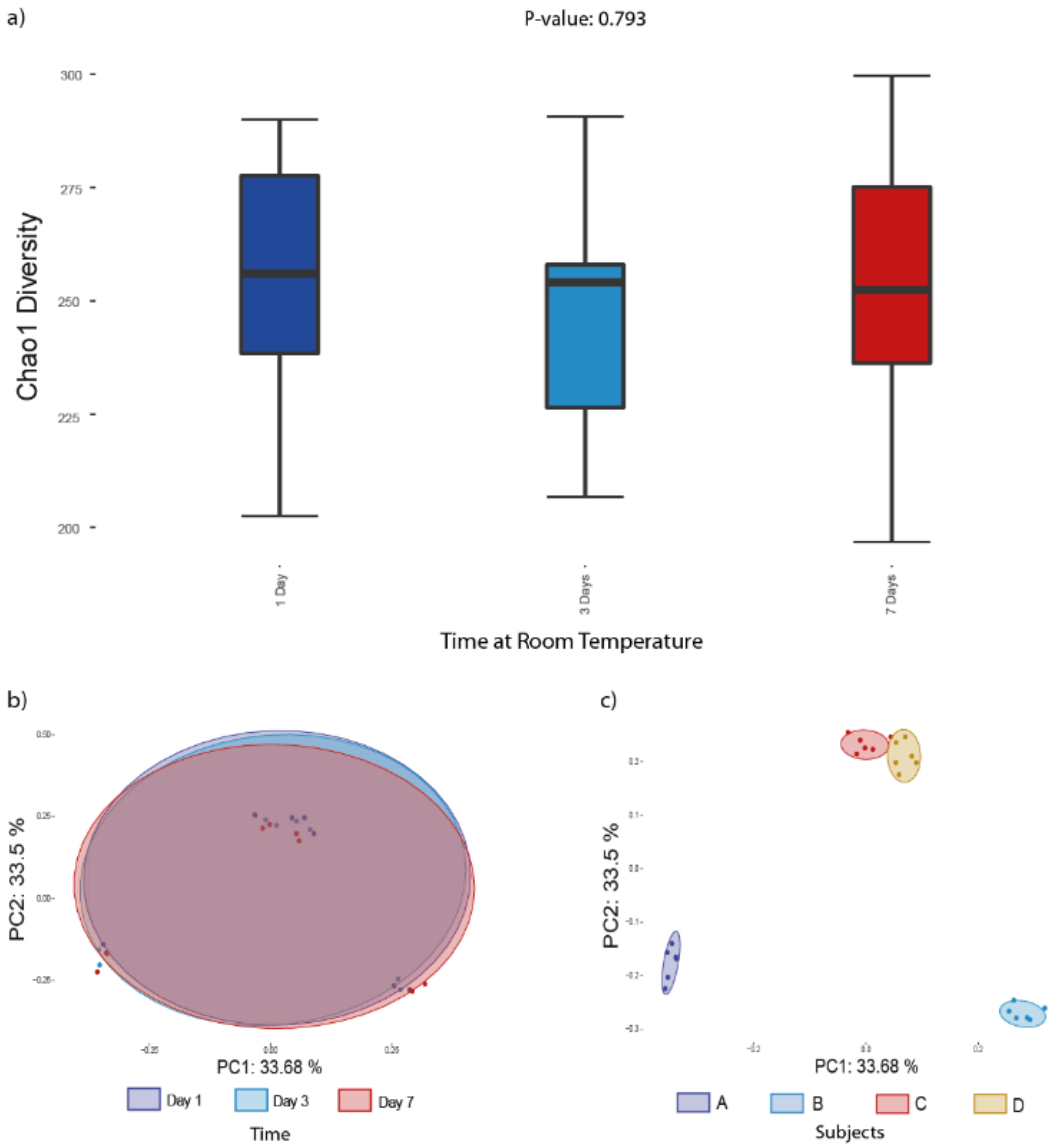
Some species were particularly prominent in certain subgroups of patients. For example, clusters associated with *B. vulgatus*, *A. muciniphila* and *Escherichia/Shigella* were prominent in patients with prior surgical resections, which has also been noted by others (Neut *et al.*, 2002). It has been proposed that reduced stool pH with 5-aminosalicylate might account for blooms of *Bifidobacteria* and *Lactobacilli*, theoretically reducing mucosal inflammation and minimizing alterations in the microbiota (Xue *et al.*, 2012). Patients receiving 5-ASA drugs may also have a microbiota closer to that of controls because of milder disease than those requiring biologics.

Significant associations between compositional changes and habitual diet were observed and which are consistent with reports of dietary constituents as potential risk factors for IBD (reviewed (Lewis *et al.*, 2017). Remarkably, about 90% of the microbiota variance in IBD remains unaccounted for, either not measured or due to stochastic factors. This is higher than what was previously observed (84%) in a large Belgian-Dutch cohort (Falony *et al.*, 2016), where UC had a relatively small effect on overall variance. Geographic location (which may reflect variance in lifestyle, diet and ethnicity) had the second greatest non-disease-related influence in our cohort and was recently shown to have significant impact among infants (Stewart *et al.*, 2018). Although similar trends were evident in both geographic locations studied by us, the work of others has highlighted ethnicity (Deschasaux *et al.*, 2018) and geographic location (He *et al.*, 2018) as important considerations for studies of the microbiota. This might also contribute to some of the inconsistency in microbiota studies of IBD. Regardless, both microbiota and host heterogeneity, which may be inter-dependent, will represent a challenge for the prospect of personalised prognostics or therapeutics based on microbiota manipulation.

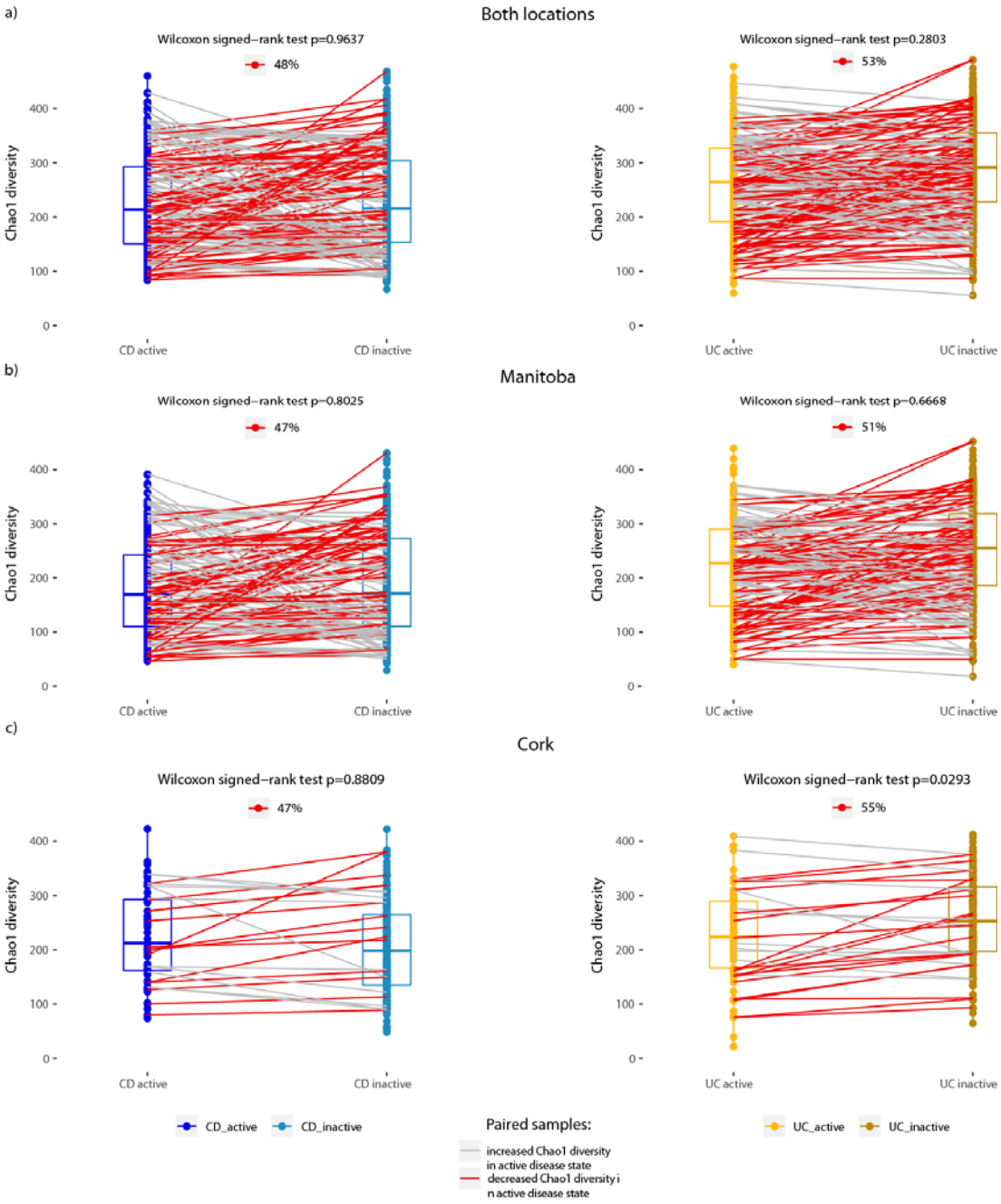
5. Acknowledgements

This research was supported in part by Science Foundation Ireland (grant numbers SFI/12/RC/2273, 11/SIRG/B2162 and 17/CDA/4765) and European Crohn's and Colitis Organization (year 2014). We also thank research nurses Margot Hurley and Catherine O'Riordan for their assistance.

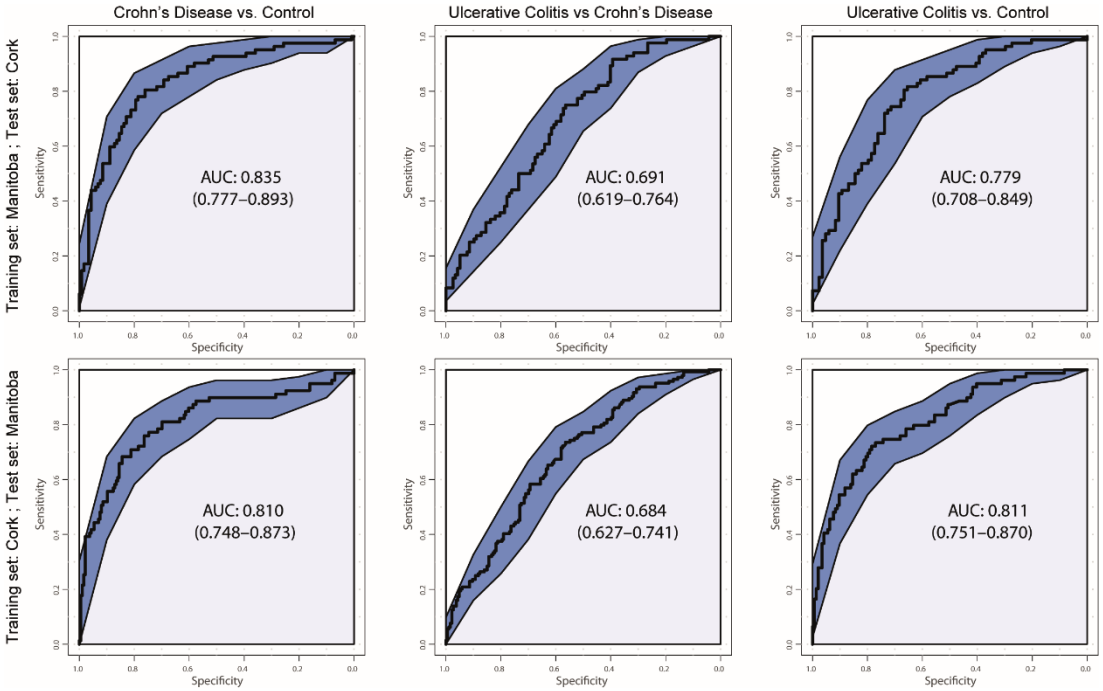
6. Appendix



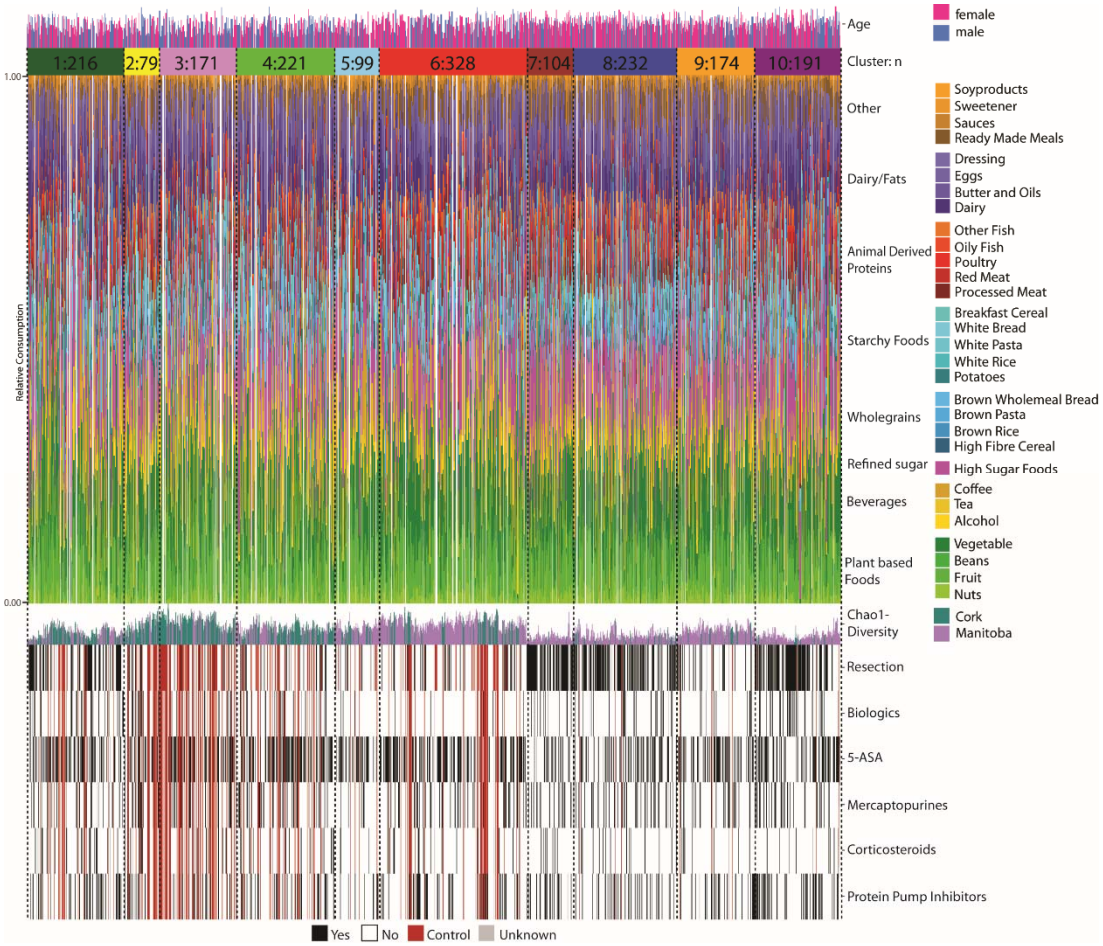
Supplementary Figure 1: Comparison of samples stored for various periods of time at room temperature. a) Chao1 diversity, b) samples clustered by time at RT, c) samples clustered by subjects.



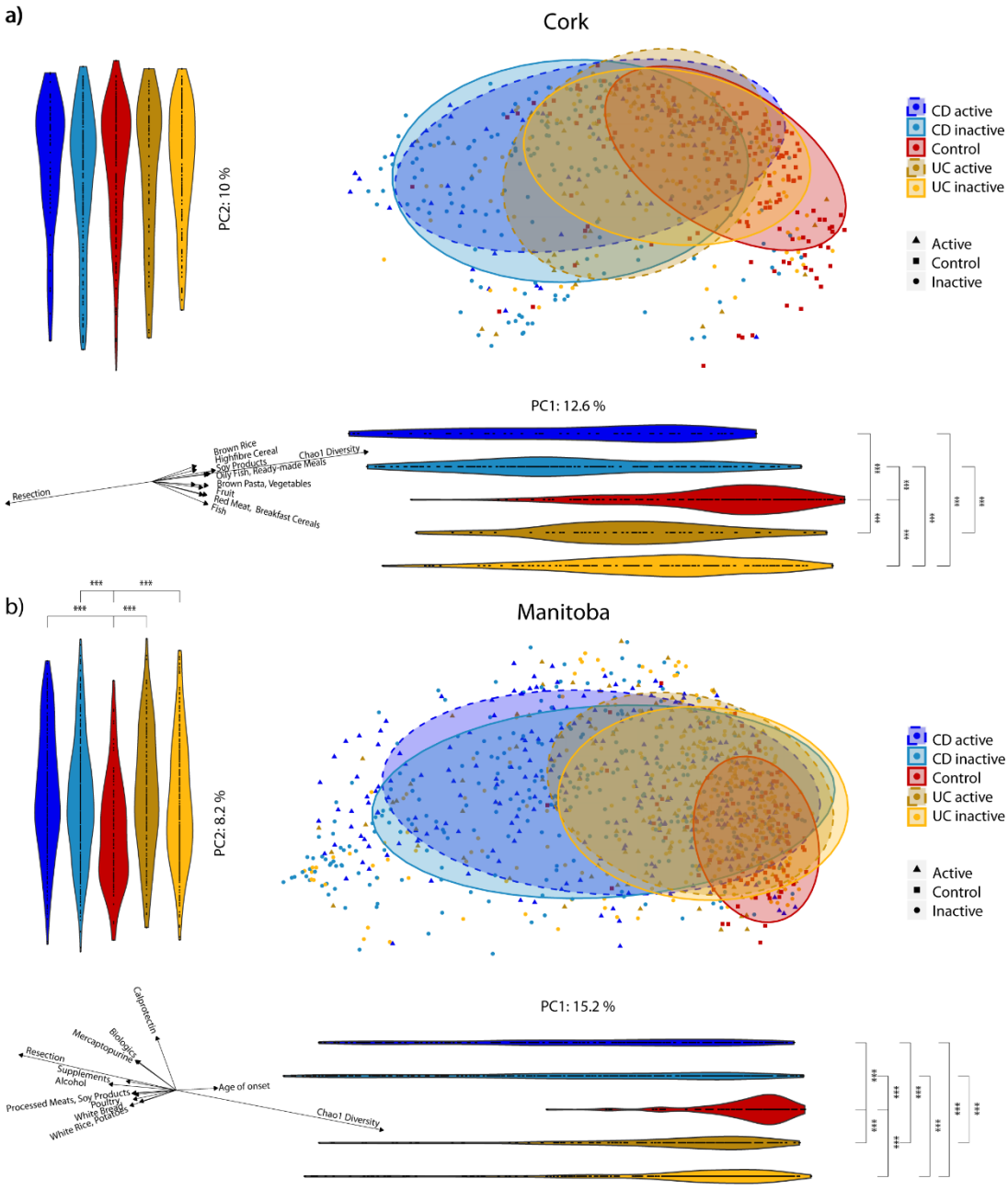
Supplementary Figure 2: Changes in alpha diversity (Chao1) between active and inactive disease state, lines connect samples of patients whose activity status change between time-points. a) Across both geographic locations separately for b) Manitoba and c) Cork. Red lines represent cases where alpha diversity was lower in active than in inactive disease, and grey lines where the opposite was true.



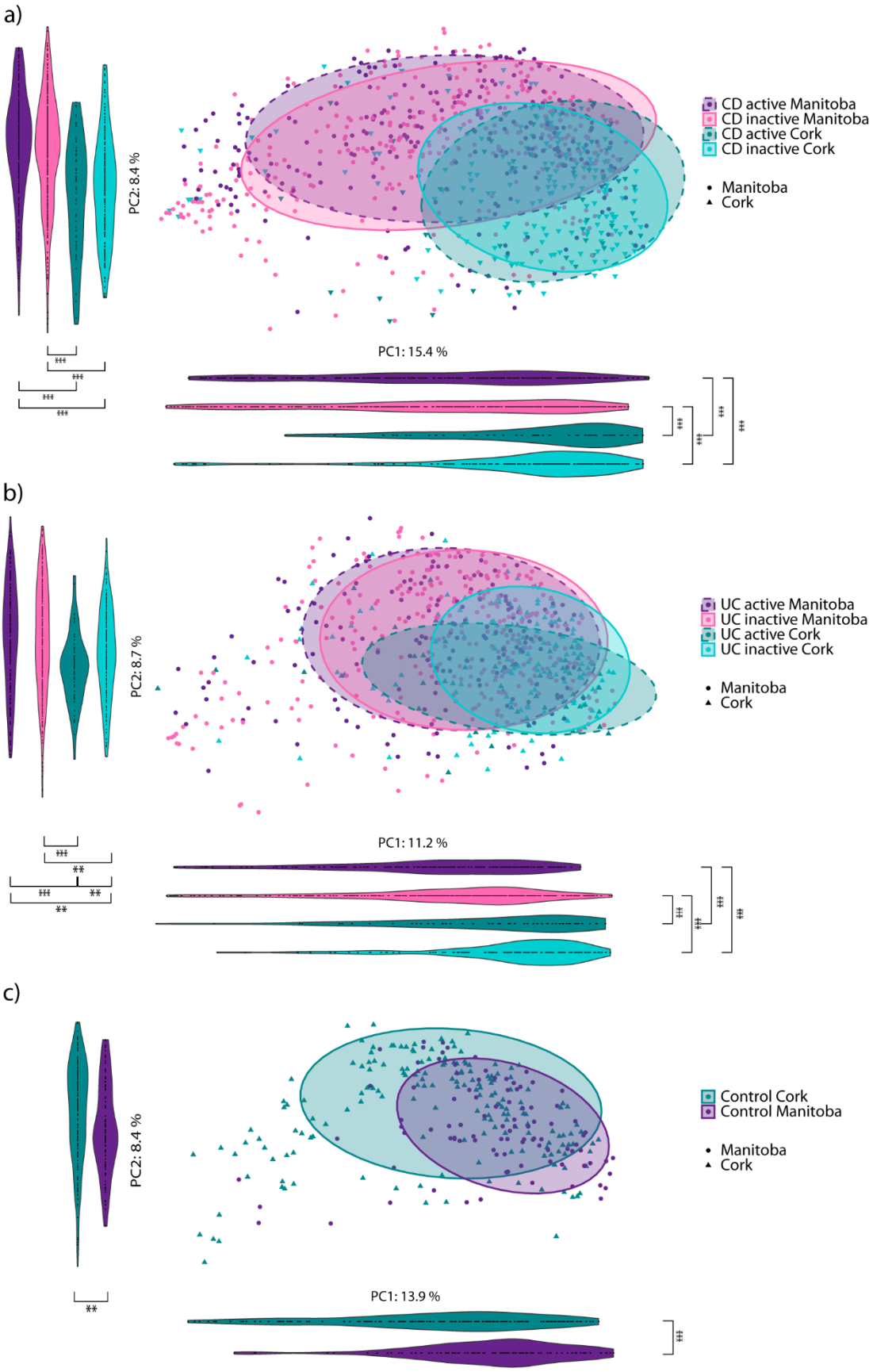
Supplementary Figure 3: Cross-geographic Machine Learning classification of the subject cohorts based on proportional normalized OTUs present in more than 5% of samples. Receiver operating characteristic curves (ROC) for the intercontinental validation of boosted tree classifiers between the same groups as in Figure 10. Here, the first row of ROC-curves used Manitoba subjects as training set and Cork subjects as test set, whereas the second row shows the results of the opposite case (Cork subjects as training model and Manitoba as test set). A model with an AUC of 0.5 has no discriminatory capacity, whereas an AUC of 1 indicates perfect separation of the response variables.



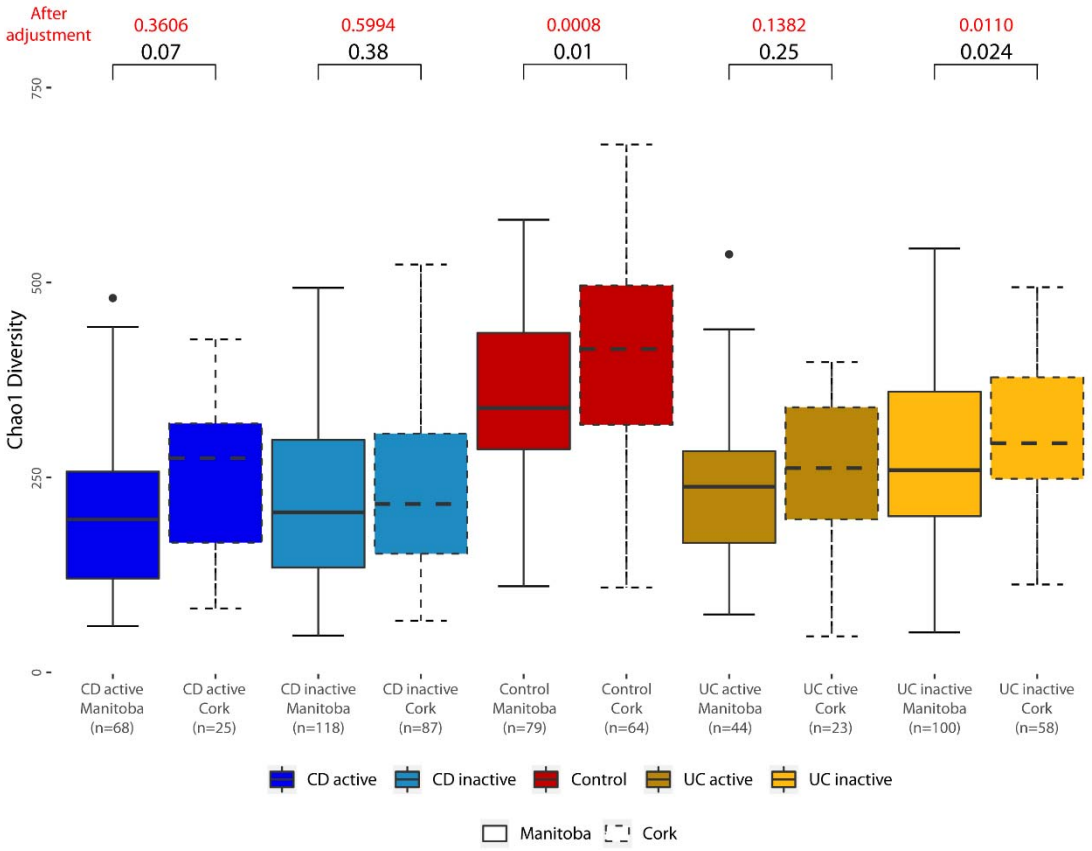
Supplementary Figure 4: Gender, medication status, resection status, dietary categories and Chao1 diversity with samples displayed in the same order as Figure 12.



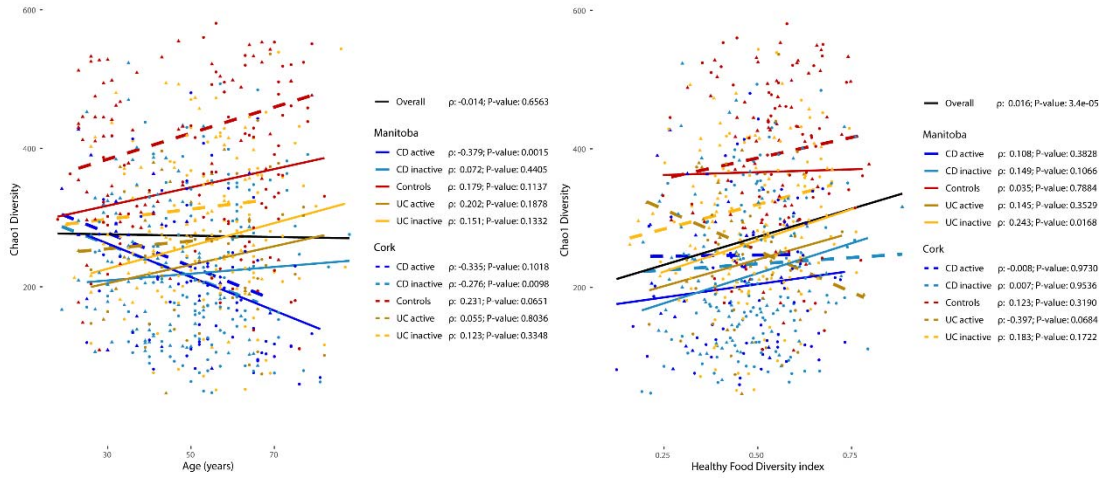
Supplementary Figure 5: Principle Coordinate Analysis based on Bray-Curtis distances with proportional normalization on all OTUs present in more than 5% of samples, with samples grouped by disease and activity state for a) Cork and b) Manitoba separately. Violin plots show projections of PCoA points onto PC1 and PC2. The arrows represent Spearman correlations between PC axes and food groups/metadata and alpha diversity. Only categories with significant correlations are represented.



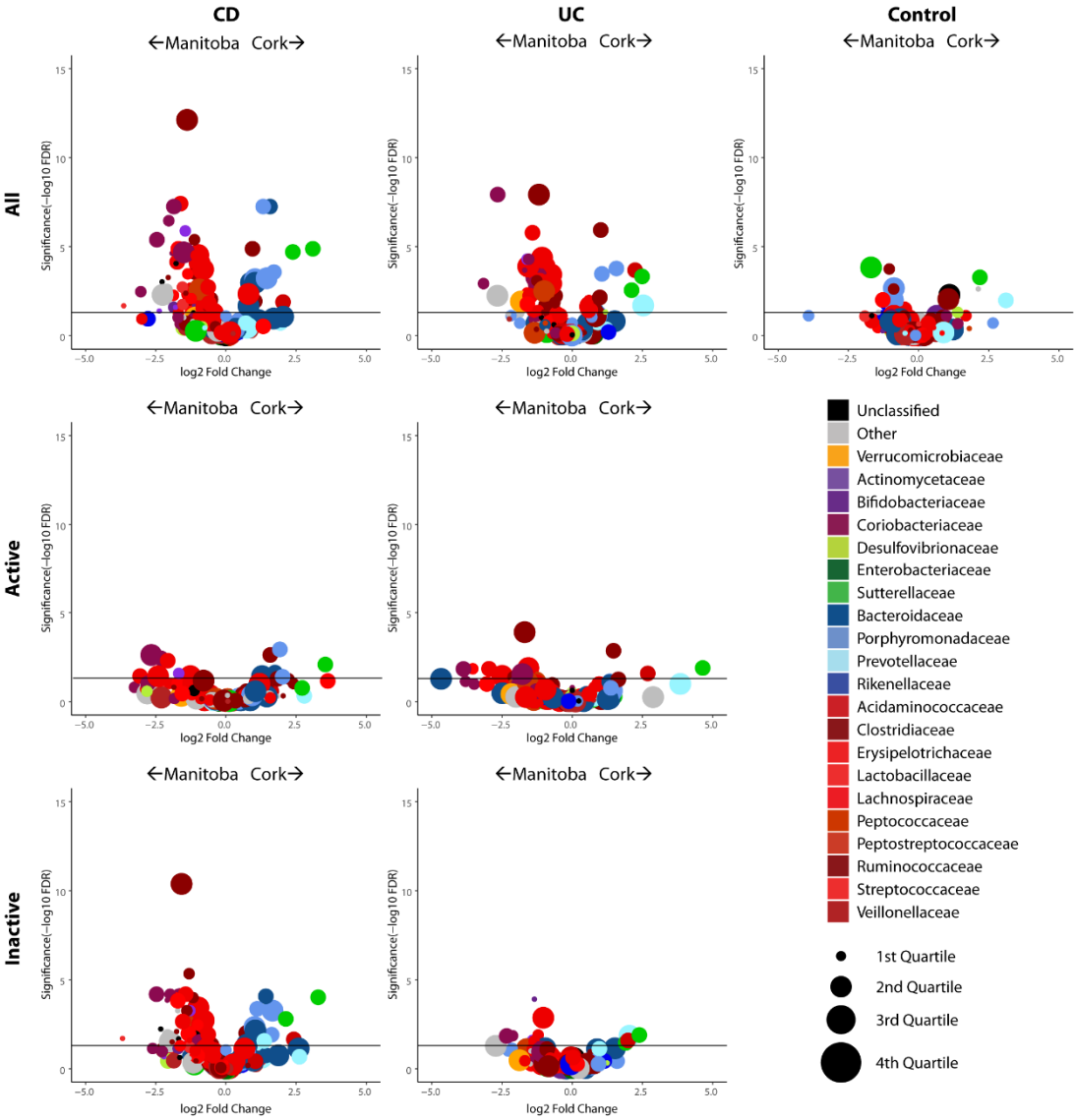
Supplementary Figure 6: Principle Coordinate Analysis based on Bray-Curtis distances with proportional normalization on all OTUs present in more than 5% of samples, grouped by disease state and geographic location for a) CD patients, b) UC patients, and c) controls. Violin plots show projections of PCoA points onto PC1 and PC2.



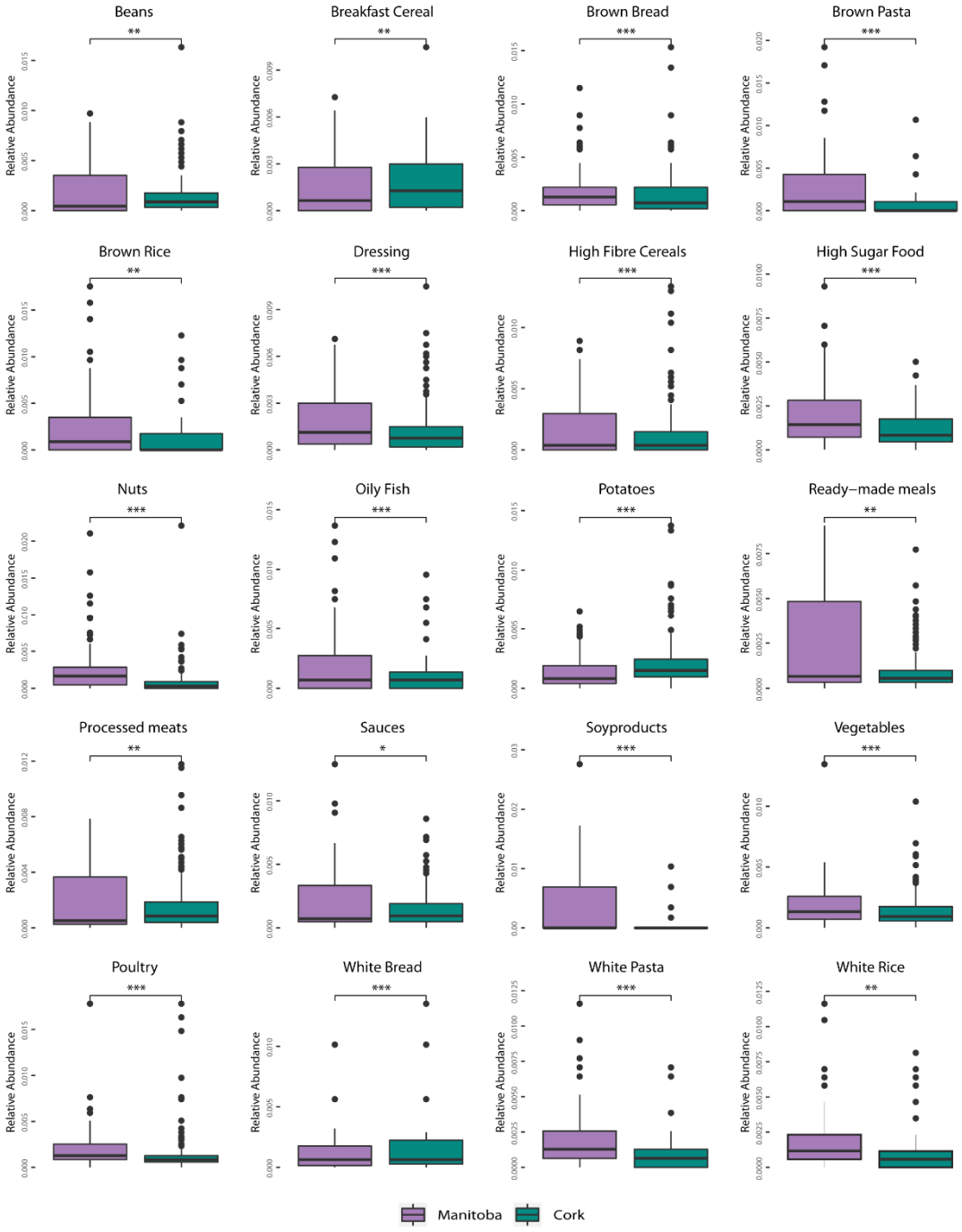
Supplementary Figure 7: Comparison of Alpha diversity across disease groups, disease activity and geographic location. P-values in black represent differences before adjustment for age while the red values show P-values after adjustment.



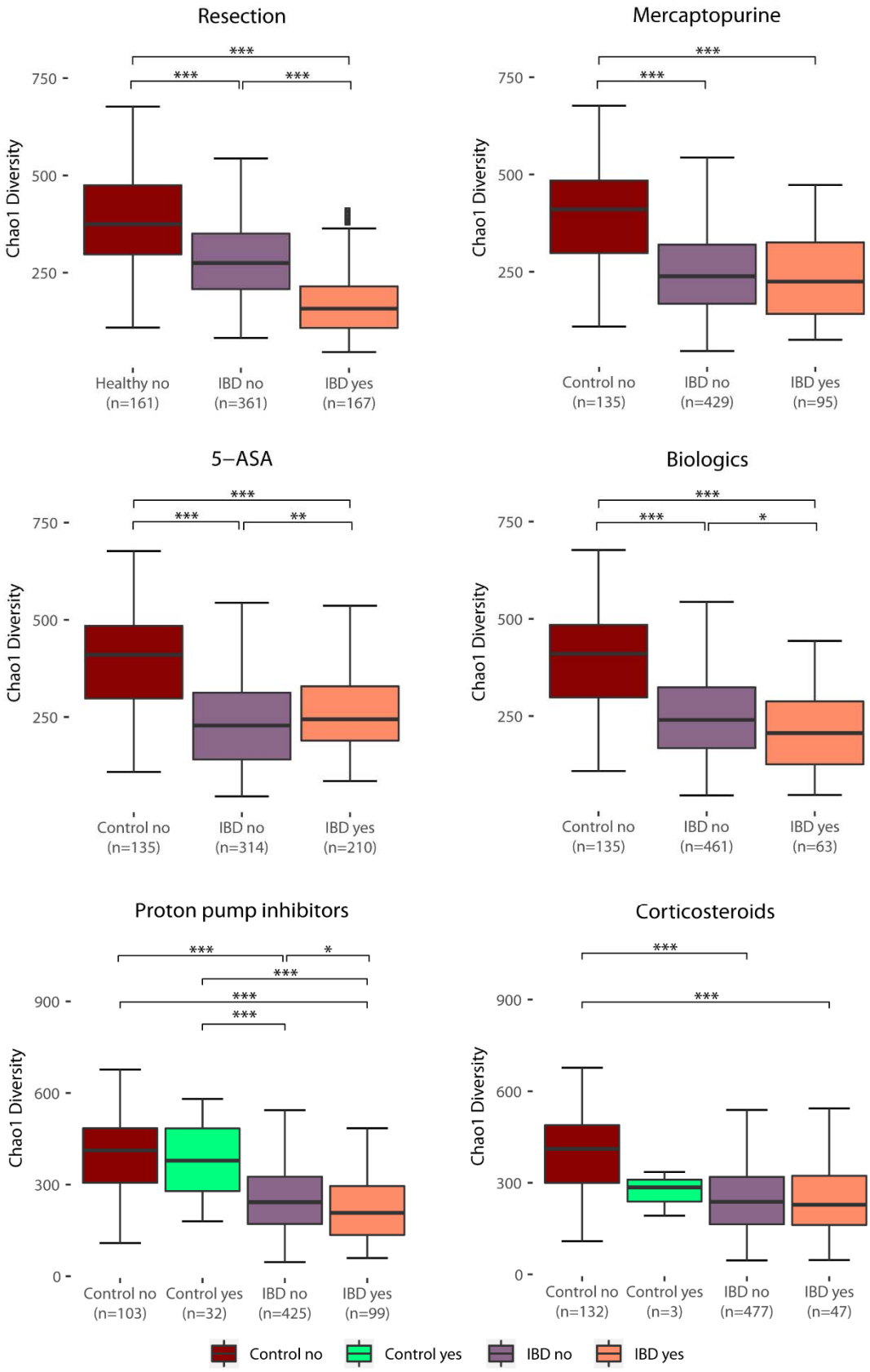
Supplementary Figure 8: Spearman correlations between alpha diversity (chao1) and a) age and b) Healthy Food Diversity (HFD) index.



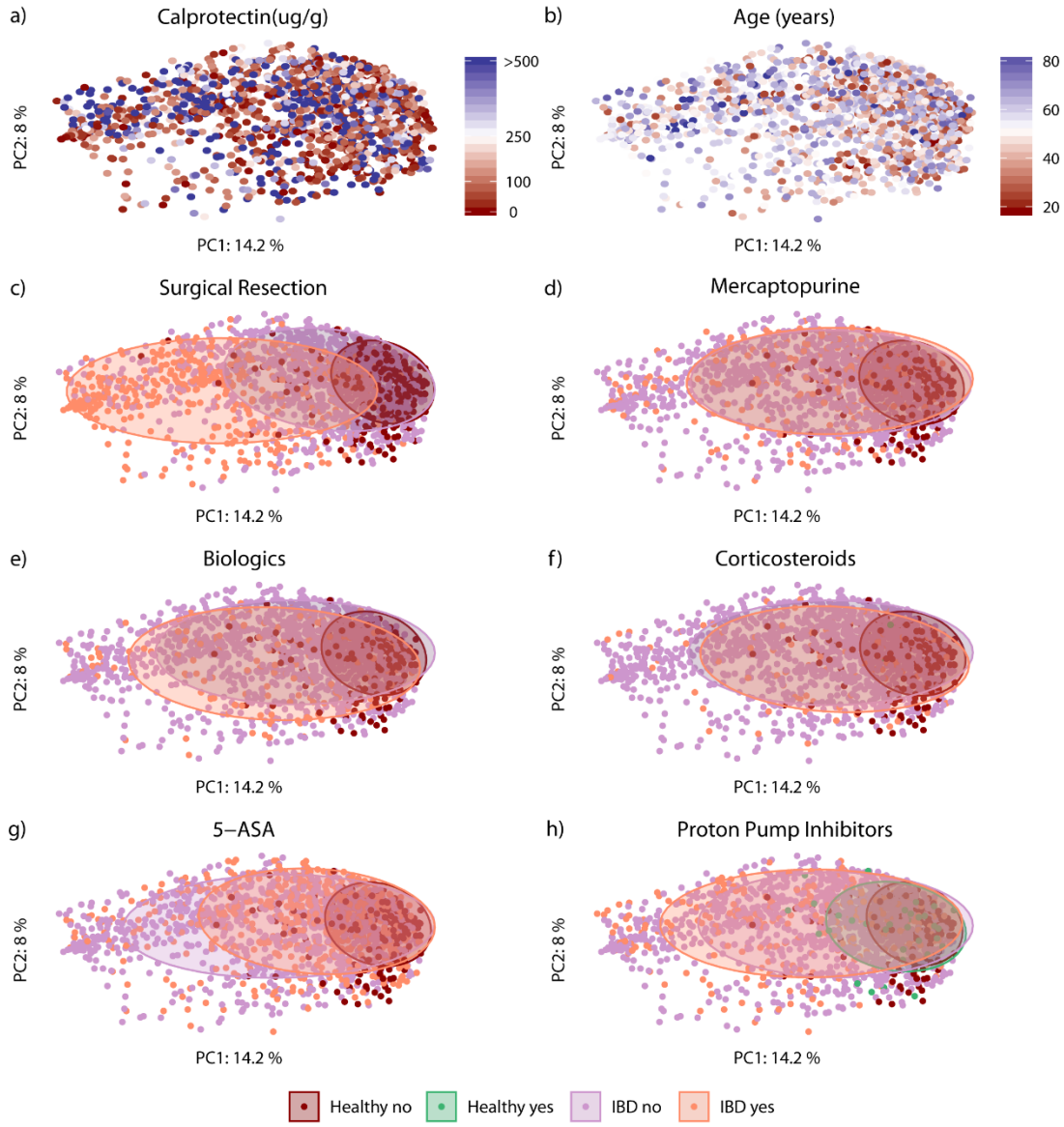
Supplementary Figure 9: Volcano plots showing differential abundant species between the Cork and Manitoba cohorts across disease and disease status. Points above the horizontal line are significant while the position on the x-axis of each point indicates the direction of fold change. The size of each point refers to the abundance of the species across the cohort while the color indicates the family rank.



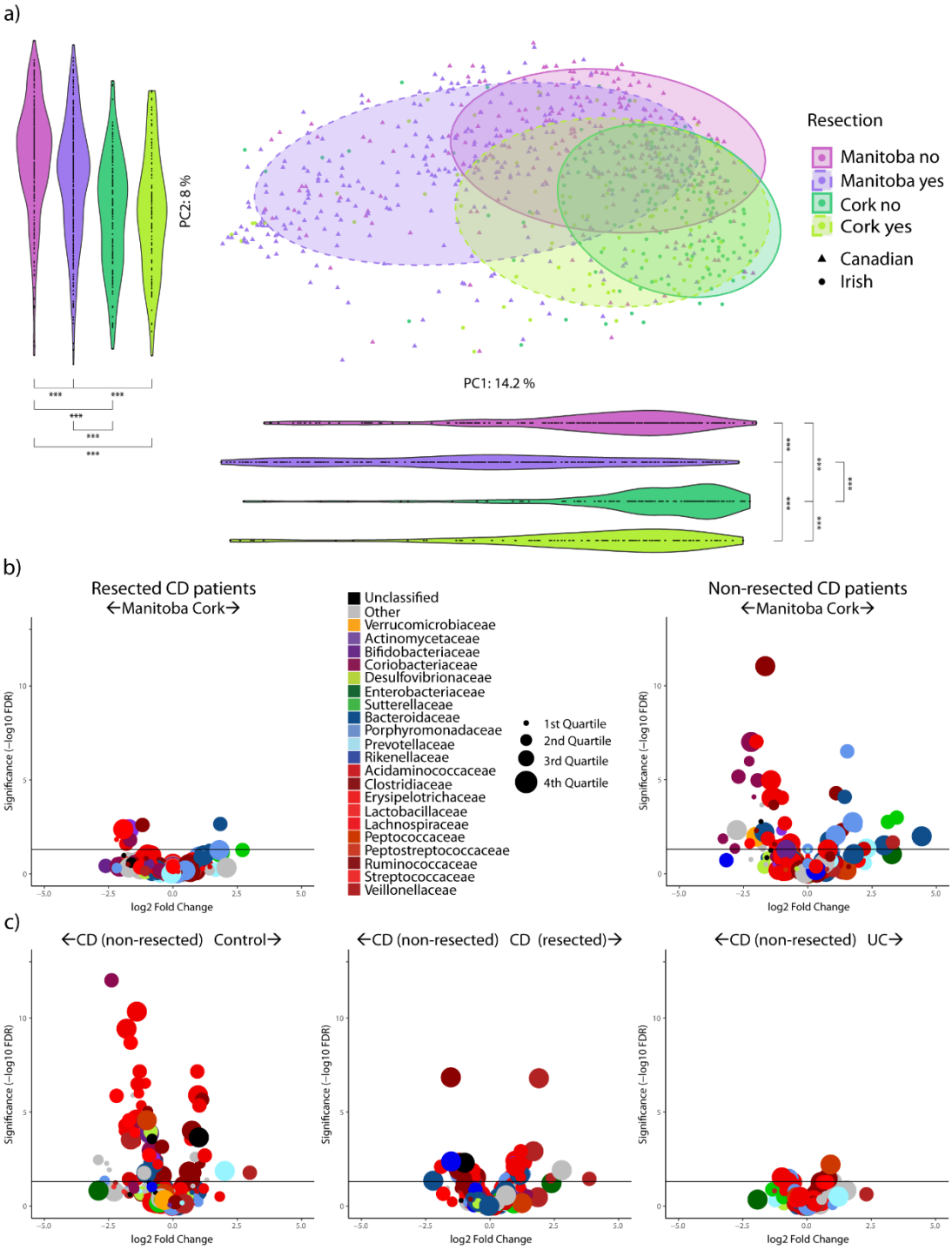
Supplementary Figure 10: Boxplots representing differential consumption of food groups between the Cork and Manitoba cohorts. Only food groups with a significant difference were plotted (20).



Supplementary Figure 11: Alpha diversity (Chao1) categorized by medication information and resection status.



Supplementary Figure 12: Principle Coordinate Analysis based on Bray-Curtis metric with proportional normalization on all OTUs present in greater than 5% of samples with samples grouped by a) faecal calprotectin, b) age, c) resection status, and d)-h) medication information.



Supplementary Figure 13: a) Principle Coordinate Analysis based on Bray-Curtis distances with proportional normalization on all OTUs present in more than 5% of samples, grouped by grouped resection status and geographic location. Volcano plots showing differential abundant species between b) Cork and Manitoba cohorts across resected and non-resected CD patients and c) between non-resected CD patient and Controls, resected CD patients, and UC patients, respectively. Points above the horizontal line are significant while the position on the x-axis of each point indicates the direction of fold change. The size of each point refers to the abundance of the species across the cohort while the color indicates the family rank.

Chapter III

Interactions between medications and the gut microbiome in inflammatory bowel disease

Eckenberger, J.[1, 2], Butler, J.C.[1, 2, 3], Bernstein,C.N.[4, 5], Shanahan,F.[1, 6], Claesson, M.J.[1, 2]

1. APC Microbiome Ireland, University College Cork, Cork, Ireland
2. School of Microbiology, University College Cork, Cork, Ireland
3. The SFI Centre for Research Training in Genomics Data Science, Ireland
4. The University of Manitoba Inflammatory Bowel Disease Clinical and Research Centre, Winnipeg,
   Manitoba, Canada
5. Section of Gastroenterology, Department of Internal Medicine, University of Manitoba, Winnipeg, Manitoba, Canada
6. Department of Medicine, University College, Cork, Ireland

Published in: *Microorganisms* **2022**, *10*(10).
https://doi.org/10.3390/microorganisms10101963

Author contributions: Conceptualization, J.E.; C.N.B.; F.S. and M.J.C.; methodology, J.E.; formal analysis, J.E. and J.C.B.; investigation, J.E. and M.J.C.; resources, C.N.B.; F.S. and M.J.C.; writing—original draft preparation, J.E. and M.J.C.; writing—review and editing, C.N.B. and F.S.; visualization, J.E.; supervision, M.J.C.; funding acquisition, F.S. and M.J.C. All authors have read and agreed to the published version of the manuscript.

Supplementary material:

https://www.mdpi.com/article/10.3390/microorganisms10101963/s1

Table of Contents

Abstract 127
1. Introduction 128
2. Methods 129
3. Results 132
3.1. Microbiota composition is different both between IBD subtypes and geographic locations ... 133
3.2. Higher usage of resection, medications, and supplements among Canadian patients 134
3.3. Polypharmacy associated with the disease-associated microbiota shift in Crohn's disease 138
3.4. Treatment explains more variation in microbial composition than IBD-subtype 140
3.5. Disease-associated bacteria are less affected by medications than beneficial bacteria 141
3.6. Additive effects of multiple medications amplify the effect of IBD-subtype on gut microbiota... 144
4. Discussion 146

Abstract

In view of the increasing evidence that commonly prescribed, non-antibiotic drugs interact with the gut microbiome, we re-examined the microbiota variance in inflammatory bowel disease (IBD) to determine the degree to which medication and supplement intake might account for compositional differences between disease-subtypes and geographic location. We assessed the confounding effects of various treatments on the faecal microbiota composition (16S rRNA gene sequencing) in persons with Crohn's disease (CD; n=188) or ulcerative colitis (UC; n = 161) from either Cork (Ireland) or Manitoba (Canada) sampled at 3 time points. The medication profiles between persons with UC and CD and from different countries varied in number and type of drugs taken. Among Canadian participants with CD, surgical resection, and overall medication and supplement usage is significantly more common than for their Irish counterparts. Treatments explained more microbiota variance (3.5%) than all other factors combined (2.4%) and 40 of the 78 tested medications and supplements showed significant associations with at least one taxon in the gut microbiota. But while treatments accounted for a relatively small proportion of the geographic contribution to microbiome variance between Irish and Canadian patients, additive effects from multiple medications contributed significantly to microbiome differences between UC and CD.

1. Introduction

Inflammatory bowel disease (IBD) which comprises Crohn's disease (CD) and ulcerative colitis (UC) have been linked with changes in the gut microbiome (Ryan *et al.*, 2020, Alam *et al.*, 2020), although cause and consequence have not been disentangled (Khan *et al.*, 2019, Aldars-Garcia *et al.*, 2021). Moreover, no uniform microbiome pattern or signature has been identified consistently across multiple studies (Pittayanon *et al.*, 2020). This may be due to the heterogeneity of IBD and/or variations in study design and patient populations (Khan *et al.*, 2019). Also medications have been shown to alter the gut microbiome and add an additional layer of confounding factors. Not only antibiotics have a well-described, cumulative long-lasting effect on the intestinal microbiome (Elvers *et al.*, 2020), but also commonly prescribed non-antibiotic medications have been shown to impact the gut microbiome (Oh *et al.*, 2021, Maier *et al.*, 2018). Standard therapies for IBD include anti-inflammatory drugs like 5-aminosalicylic acids (5-ASA) and corticosteroids, immunosuppressants such as thiopurines, methotrexate and biologics (TNF-α inhibitors and integrin or interleukin receptor antagonists). Several studies have shown that these drugs can impact the microbial composition (Sargautiene *et al.*, 2020, Huang *et al.*, 2015, Swidsinski *et al.*, 2007, Letertre *et al.*, 2020, Magnusson *et al.*, 2016), but vice versa can gut microbiota modulate their pharmacological activity via drug (in)activation and biotransformation (Franzin *et al.*, 2021). For instance, bacterial azoreductases release 5-ASA from its prodrugs, while bacterial N-acetyltransferases are responsible for the inactivation of 5-ASA (Sargautiene *et al.*, 2020). Aminosalicylates in turn can increase the levels of some Firmicutes while reducing the levels of Bacteroidetes and Proteobacteria (Xu *et al.*, 2018).

Moreover commonly used non-IBD drugs can also affect the composition and metabolic function of gut microbiota (Vich Vila *et al.*, 2020). In a major in vitro study, Maier and colleagues (Maier *et al.*, 2018) tested the effect of more than 1000 drugs against 40 representative gut bacterial strains and found that approximately a quarter of human targeted drugs inhibited the growth of at least one test strain, with butyrate and propionate producing bacteria being more sensitive, and γ-Proteobacteria being more drug-resistant. Pathogens and commensals alike have

been shown to be able to metabolize and/or bio-transform a variety of drugs leading to an altered bacterial metabolism and changes in microbial composition due to metabolic cross-feeding and changes in the intestinal microenvironment (Zimmermann *et al.*, 2019, Klunemann *et al.*, 2021). To this end, a variety of studies have elucidated the microbiota-altering effects of commonly prescribed medications (proton pump inhibitors (Clooney *et al.*, 2016a, Weersma *et al.*, 2020), lipid lowering statins (Vieira-Silva *et al.*, 2020), laxatives (Tropini *et al.*, 2018), metformin (Forslund *et al.*, 2015), beta blockers (Lin *et al.*, 2021) and ACE inhibitors (Tuteja *et al.*, 2019), SSRI antidepressants (Munoz-Bellido *et al.*, 2000). In some cases, the drugs explained more of the microbiota variability than the disease itself (Falony *et al.*, 2016, Forslund *et al.*, 2021b), which bears the question as to what degree this variability is secondary to the disease, or to the drugs treating it.

We recently showed in a large intercontinental twin city study of the microbiome in IBD that geographic location (Ireland vs. Canada) had a major influence on microbiota variance almost equivalent to that of a diagnosis of Crohn's disease itself (Clooney *et al.*, 2021). While the influence of geography may, in part, be due to cultural and ethnic influences, differences in treatment on either side of the Atlantic may also have a contribution. Therefore, the purpose of the present study was to re-examine and disentangle microbiota variance in IBD to determine the degree to which differences in treatment at different locations might account for the apparent geographic influence. The results confirm that the overall trends of microbiota composition and diversity as previously reported by us remain different across IBD subtypes and geographic location. Only a small part of the effect of geographic location is explained by the differences in medication and supplement intake. However, a large proportion of the disease-associated shift in microbial composition between patients with UC and CD can be explained by additive interaction effects from multiple medications.

2. Methods

The V3-V4 16S rRNA amplicon sequences from our previous study, which had been processed in a single laboratory in Cork utilizing the same protocols, were downloaded and pre-processed as previously described (Clooney *et al.*, 2021), with

the exemption of the taxonomic classification, which was performed within the mothur suite (v1.39.3)(Schloss *et al.*, 2009) via the classify.seqs function against the SILVA database (v132) (Quast *et al.*, 2013) with a bootstrap cut-off of 80%. OTUs that fell below that cut-off were assigned as unclassified at that particular rank. Species-level resolution was provided by SPINGO (Allard *et al.*, 2015) using a similarity score of 0.5 and bootstrap cut-off of 0.8 against the SILVA database (v132). Conflicts between the two methods were resolved by means of BLASTn (v2.8.1; May 10 2021) (Altschul *et al.*, 1990). The raw medication data was classified based on the anatomical therapeutic chemical (ATC) classification system, which hierarchically classifies medications and dietary supplements based on the therapeutic use of their main active ingredient (1st level: anatomical main group; 2nd level: therapeutic subgroup; 3rd level: pharmacological subgroup; 4th level: chemical subgroup; 5th level: chemical substance) (WHO Collaborating Centre for Drug Statistics Methodology, 2021), and recorded as a qualitative variable. While surgical resection is neither medication or supplement, it is a common treatment for IBD as 80% of patients with CD require surgery during their lifetime (Sica *et al.*, 2013). It is also has been shown to significantly alter the microbiome of patients (Fang *et al.*, 2021) and was therefore included in the analysis as treatment. Long-term dietary habits were captured through frequencies of medium food servings of 157 items via Food Frequency Questionnaires (FFQ) and were summarized into one factor, the Healthy Food Diversity (HFD) Index, as previously described (Clooney *et al.*, 2021, Drescher *et al.*, 2007). An active state of IBD was defined as a faecal calprotectin measurement of ≥ 250 µg/g (Lin *et al.*, 2014) .

All statistical analyses were performed in the R environment version 4.1.0 and visualizations were produced with the ggplot2_3.3.5 package (Wickham, 2016). In order to alleviate the constant sum constraint of compositional data (Gloor *et al.*, 2017), zeros were removed from the raw counts via the count zero multiplicative method within the zCompositions_1.3.4 package (Palarea-Albaladejo *et al.*, 2015) and subsequently subjected to a centred log ratio (CLR) transform, i.e. the data were expressed as logarithms of ratios with the geometric mean as denominator using the propr_4.2.6 package (Quinn *et al.*, 2019). Beta diversity (i.e. between sample

diversity) of the microbiome was evaluated via principal component analysis on Aitchison distances (Aitchison *et al.*, 2000) within phyloseq_1.36.0 (McMurdie *et al.*, 2013), while vegan_2.5.7 (Oksanen *et al.*, 2013) was used for permutational multivariate analysis of variance between groups. Alpha diversity (i.e. within sample diversity) was calculated with iNEXT_2.0.20 (Chao *et al.*, 2014). Differential taxa abundance and effect sizes without adjusting for confounding factors was computed via ALDEx2_1.24.0 (Fernandes *et al.*, 2013) using 1000 Monte Carlo samples. Differences in the usage of medications and supplements between the different IBD-subtypes and patients from the two different geographic locations were tested for significance with Fischer's Exact tests. Distance based redundancy analysis (dbRDA) was performed with the capscale function in vegan_2.5.7 (Oksanen *et al.*, 2013) to evaluate the effect of environmental factors on the medication profiles (Jaccard distances) and gut microbial composition (Aitchison distances). Here, medication profiles and microbial abundances constitute the set of response variables, while the environmental factors represent the predictive variables. The proportion of explained, compared to the total fitted variance indicates how much of the variation between samples is due to differences in environmental factors. Selection of the most relevant species and features in the dbRDA was implemented with the ordiselect function in goeveg_0.5.1 (von Lampe, 2021). Differences in numbers of used medications as well as the number of changed medications and intra-personal differences in alpha and beta diversity measures were assessed via Wilcoxon rank sum tests.

Explained variance of single covariates in a multivariate data set was computed with the VpThemAll_0.0.0.9 wrapper (Moitinho-Silva *et al.*, 2021), which chooses a model via the ordistep function that explains most variance in microbiota composition and then looks at each included metadata variable separately, conditioning out the effect of all other included metadata variables using the varpart function. The difference between the naïve (shared) effect and the unique effects amounts to the interaction effect. The variation in community data with respect to explanatory tables was calculated with the varpart function in vegan_2.5-7 (Oksanen *et al.*, 2013) allowing only permutations within the samples of the same patient to adjust for multiple

measurements. Significant associations and effect sizes of covariates with single taxa were computed with the metadeconfoundR_0.2.8 package (Forslund *et al.*, 2021a), specifying “patient” as random variable. P-values were adjusted for multiple testing where appropriate, using the Benjamini and Hochberg method (Benjamini *et al.*, 2001).

3. Results

After extracting from the Clooney et al. (Clooney *et al.*, 2021) data set samples of persons with IBD with all 3 time-points and no missing data, 349 persons were included for further analysis, of which 188 were diagnosed with CD and 161 with UC. The sampling time points were on average 15.18±0.42 weeks apart. Usage information on medication and dietary supplements had been collected for all of the 1,047 samples. Of the originally 3,148 operational taxonomic units (OTUs) that were clustered at ≥97% identity, 2,409 remained after filtering, leaving on average 22,024 ±492 usable reads per sample. Applying the ATC classification system (WHO Collaborating Centre for Drug Statistics Methodology, 2021) to the raw medication data yielded 302 different chemical substances (5^{th} ATC level). A number of dietary supplements (n=74) had no ATC classifier and were aggregated under “Other supplements”. For the analysis the medications and supplements were then further combined into 120 pharmacological subgroups (3^{rd} ATC level) and drug usage was recorded as a qualitative yes-no variable. Long-term dietary habits were summarized into one factor, the Healthy Food Diversity (HFD) Index, as previously described (Clooney *et al.*, 2021, Drescher *et al.*, 2007). An active state of IBD was defined as a faecal calprotectin measurement of ≥250 µg/g (Lin *et al.*, 2014). (*Figure 15, Supplementary Table 1 and 2*).



	Cork		Manitoba	
	CD	UC	CD	UC
IBD-subtype	44.3%	55.7%	57.1%	42.9%
Disease activity	29.1%	22.4%	39.6%	31.8%
Years since diagnosis	18.36±2.28	16.46±1.8	24.45±0.95	20.17±1.23
Sex (female)	40.2%	36.7%	71.8%	59.8%
Age	44.21±2.4	47.25±2.1	54.77±1.13	57.47±1.39
HFD-Index	0.43±0.03	0.47±0.02	0.5±0.01	0.51±0.01
Smoker	15.4%	4.8%	10.7%	5.7%
Alcohol	49.6%	55.1%	39.6%	59.2%

Figure 15: Subject characteristics and sample sizes of the study cohort. The ATC classification system classifies medications hierarchically based on the therapeutic use of the main active ingredient: 1st level: anatomical main group; 2nd level: therapeutic subgroup; 3rd level: pharmacological subgroup; 4th level: chemical subgroup; 5th level: chemical substance.

3.1. Microbiota composition is different both between IBD subtypes and geographic locations

In contrast to our original study, here we used a 'compositionally-aware' analysis approach, which more accurately quantifies taxa without the confounding relative effect of total read count per sample (Gloor *et al.*, 2017). As previously reported (Clooney *et al.*, 2021), alpha (i.e. within sample) diversities were lower for all patients with CD compared to UC, and also for Canadian participants in general (Wilcox $p < 0.05$; *Figure 16a and b, Supplementary Table 3*). Abundances of OTUs were CLR-transformed and beta (i.e. between sample) diversity analysis was calculated from Aitchison distances of all OTUs present in at least 10% of the samples. The first two principal component axes captured a much higher proportion of the microbiota variation in the data set (71.6%) whilst showing the same significance (PERMANOVA $p < 0.05$) for disease and location-associated shifts (*Figure 16c and d*). Differential taxa abundance analysis was carried out using ALDEx2, initially without accounting for any confounders. Here, the OTUs were aggregated based on their highest known

taxonomic classification and filtered for an abundance in at least 10% of samples, resulting in 233 tested taxa. Of these, 108 OTUs were significantly different (Wilcox/Welch $p < 0.05$) between persons with UC and CD (*Figure 16e*), but displayed only weak to moderate effect sizes, ranging from -0.38 to 0.41. *Faecalibacterium prausnitzii*, Ruminococcaceae UCG-002 and *Eubacterium* were the most increased taxa in patients with UC, while *Lachnoclostridium*, *Erysipelatoclostridium* and *Escherichia/Shigella* had the highest abundance in the microbiome of patients with CD (*Supplementary Table 4*). The effect sizes of the 100 taxa that were significantly different (Wilcox/Welch $p < 0.05$) between subjects from Manitoba and Cork were slightly higher than between the IBD subtypes (*Figure 16f*), ranging from -0.33 to 0.59. Members of Saccharimonadaceae, *Streptococcus*, and *Ruthenibacterium lactatiformans* were the most enriched taxa in the Manitoba cohort, while Lachnospiraceae_UCG-004, *Odoribacter splanchnicus* and Bacteroidales were the most increased taxa in the Irish cohort (*Supplementary Table 5*). Thus, with updated analysis, compared to our previous work (Clooney *et al.*, 2021), the overall trends of microbiota composition and diversity remain different across IBD subtypes and geographic location.

3.2. Higher usage of resection, medications, and supplements among Canadian patients

To disentangle the effect of medication on microbiota we first quantified any significant differences in IBD treatments, both dependently and independently of geography. Of the 120 different pharmacological subgroups found in this study, 78 were recorded at least 5 times as “yes” in total and were included in the downstream analysis (default MetadeconfoundR requirement). We also included surgical resection as a treatment. Of these, 28 medications and supplements showed significantly (Fischer $p < 0.05$) different usage between the Irish and the Canadian cohort (*Figure 17a*). Between patients with either CD or UC, 18 pharmacological subgroups were differentially (Fischer $p < 0.05$) used. Surgical resection was significantly more present in the Manitoba cohort and in persons with CD. Intestinal anti-inflammatory agents, a subgroup which contains ‘locally acting corticosteroids’, and ‘aminosalicylic acid and similar agents’ were the most commonly used

treatments, which were over-represented in persons with UC and Irish participants. The most taken supplements were vitamin A and D, which were over-represented in Canadians with CD (*Figure 17a, Supplementary Table 6*).

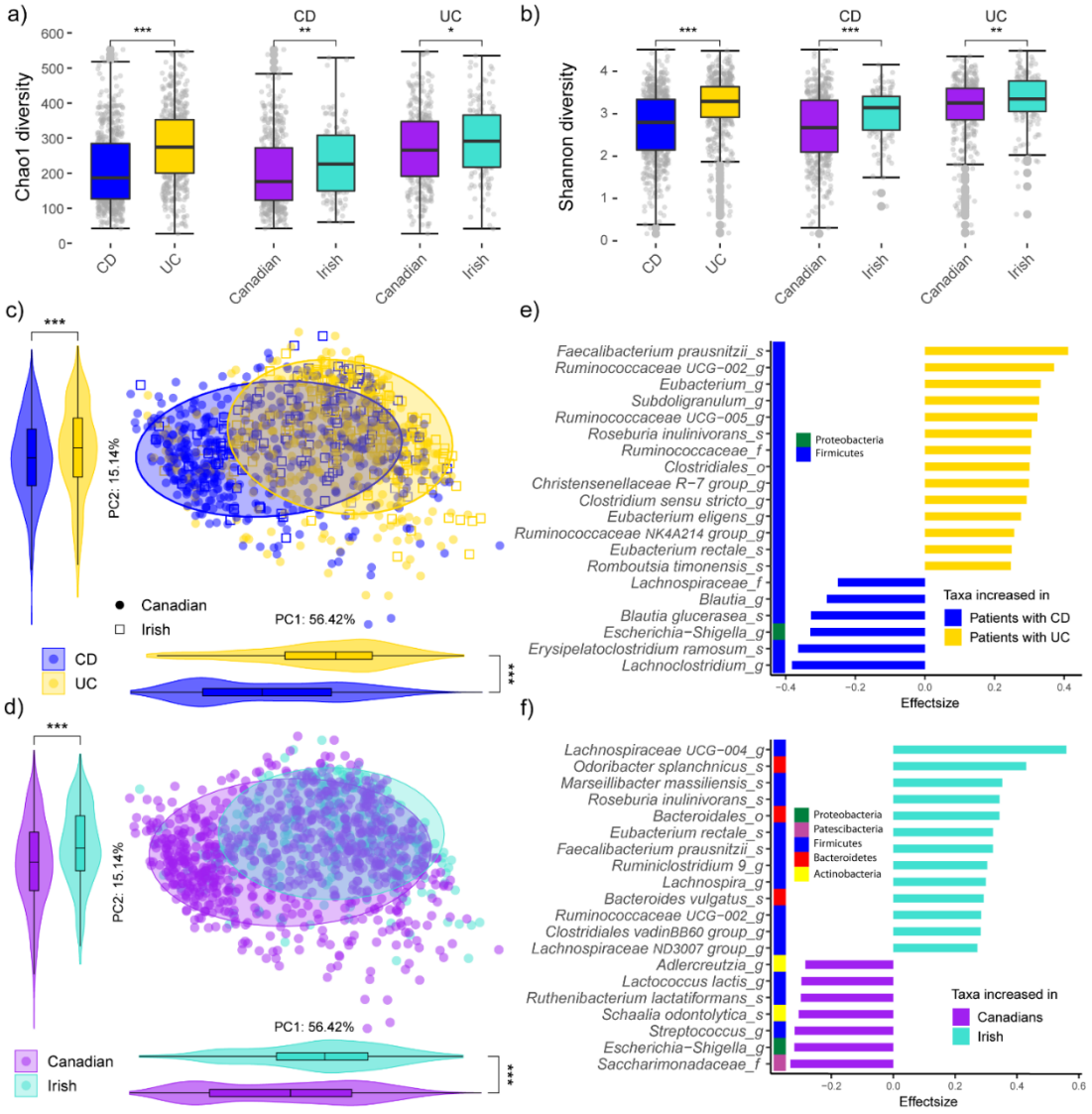


*Figure 16: Comparison of (a) Chao1 (species richness) and (b) Shannon diversity (species richness and evenness) between different IBD-subtypes and geographic location. Principal component analysis (PCA) based on Aitchison distances on all operational taxonomic units (OTUs) present in >10% of samples, with samples grouped by (c) IBD subtype and (d) geographic location. Violin plots show projections of the PCA points onto PC1 and PC2. Stars show significant differences between the groups as determined by Wilcoxon test. Top 20 most differential OTU abundances between (e) IBD subtypes and (f) geographic location calculated with ALDEx2; *p < 0.05; **p < 0.01; *** p < 0.001.*

Jaccard distance-based redundancy analysis (dbRDA), which is more suitable for quantifying distances between binary data (e.g. medication A: TRUE vs. FALSE), was used to evaluate the effect of IBD subtype and geographic location on medication usage patterns. Both the IBD subtype ($F=33.71$; $p < 0.001$) and geographic location

(F=25.96; p < 0.001) showed a significant effect, with a CD/UC diagnosis accounting for 26.6% and location for 19.5% of the explained variation in treatments (*Figure 17b and c*). Surgical resection and vitamin A and D supplementation were most prevalent in Canadian participants, while intestinal anti-inflammatories best distinguished Irish participants with UC from the other cohorts.

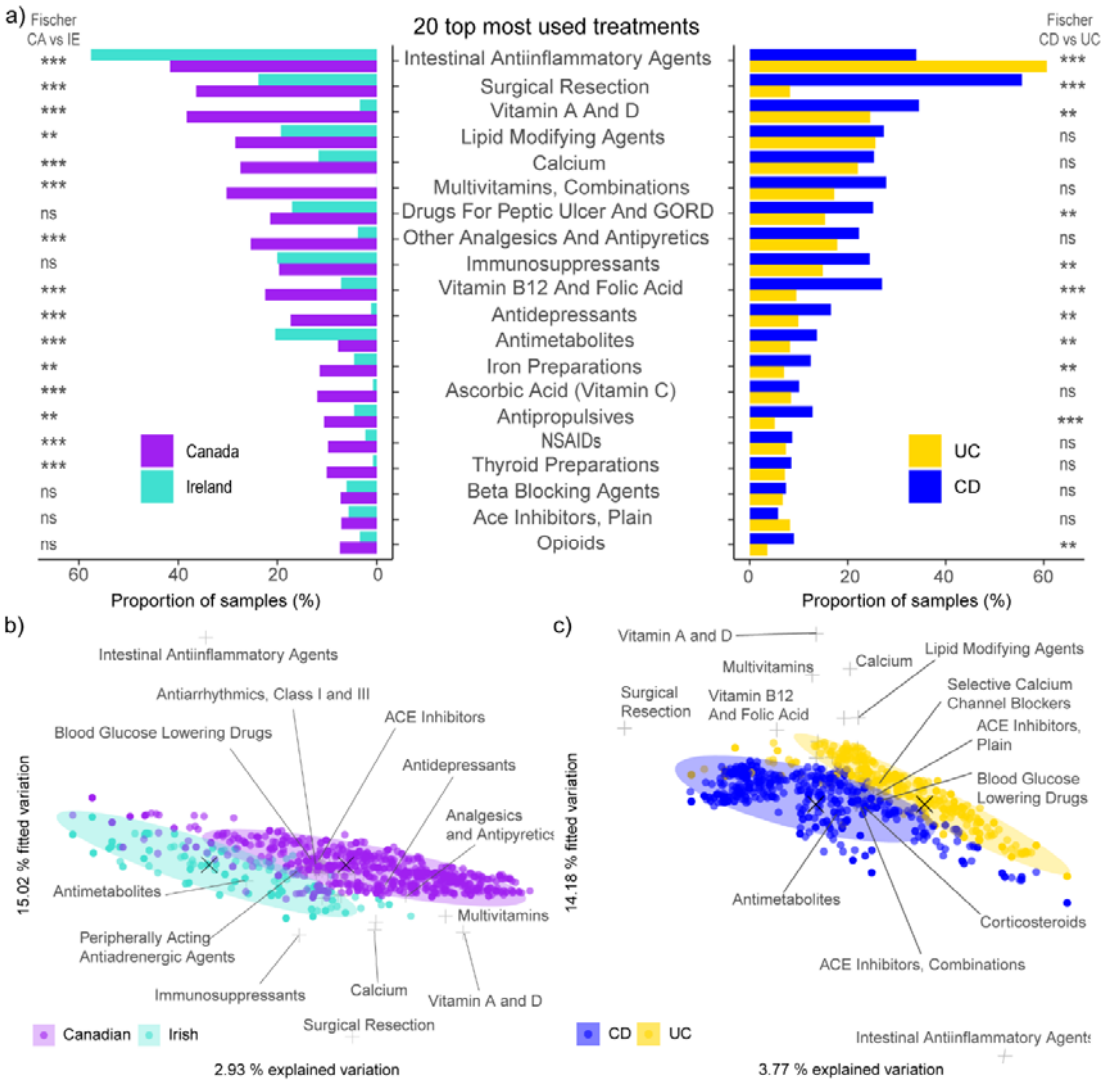


*Figure 17: Comparison of the 20 most common medications and supplements taken in the study cohort separated by IBD subtype and geographic location (a). Ordination plot of Jaccard distance-based redundancy analysis (dbRDA) of used medications and supplements constrained by (b) geographic location and (c) IBD subtype. The points represent samples, crosses represent medications and X represent the centroids of the depicted groups. The 5% of medications with the best axis fit are labelled; Fishers test: *p < 0.05; **p < 0.01; ***p < 0.001.*

The participant groups from varying locations and with a different disease type not only diverged in their drug usage pattern, but also in the amount of ingested pharmaceutical compounds. Overall, persons with CD took a significantly higher number of supplements and medications (Wilcox p < 0.05) than persons with UC.

Similarly, Canadian participants used a significantly greater number of supplements and medications than their Irish counterparts (*Figure 18 a and b*). Another source of variation between the groups was the frequency with which specific medications were changed over the course of the study. Only 73 of 349 participants did not change their medication regime during the 6 months of study, while another 93 persons changed 5 or more medications over the three time points (*Figure 18c*). Unsurprisingly, due to the higher usage of pharmaceutical products in the Canadian cohort, there were also significantly (Wilcox $p < 0.05$) more changes of medications in this group, while there were no significant disparities between persons with UC and CD within their cohort (*Figure 18d*).

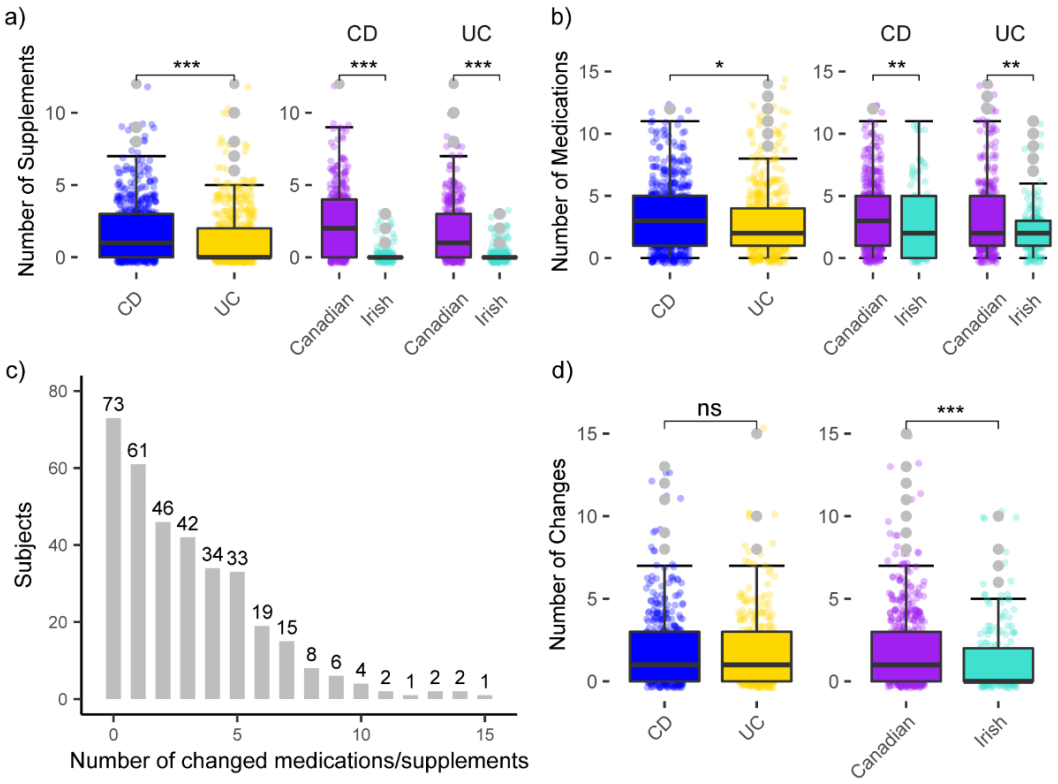


*Figure 18: Comparison of the amount of (a) supplements and (b) medication taken per person and time point. (c) Histogram of changes in medications and supplements with ATC classification per subject over the course of the study. (d) Comparison of changes in medications and supplements between patients with differing IBD subtype and from different geographic locations. Wilcoxon test: *p < 0.05; **p < 0.01; ***p < 0.001.*

3.3. Polypharmacy associated with the disease-associated microbiota shift in Crohn's disease

Polypharmacy (PP) is defined as the concurrent and regular use of at least 5 medications, and excessive polypharmacy (EPP) as use of 10 or more different drugs. In our study, 120 persons at 233 time points fit in the former category, while 22 persons (42) fit the latter group (*Figure 19b*). Polypharmacy expectedly increased with age as persons in the EPP and PP group were significantly older than those that took less than 5 medications on a regular basis (*Figure 19c*). The microbial composition of persons with EPP, PP and no polypharmacy (NP) showed a gradual shift along the first PC axis (*Figure 19a*). This corresponded with the shift seen when grouping by IBD-subtype (*Figure 16c*), which is not surprising as persons with CD used significantly more medications than persons with UC (*Figure 18b*). Taking multiple medications at the same time can lead to increased side effects and drug interactions that amplify adverse complications (Ticinesi *et al.*, 2017).

Since the intake of these pharmaceuticals might influence disease-associated microbiome disturbances, we investigated the influence of changes in the treatment regime on the microbiota. We compared intra-individual alpha and beta diversity differences between the first and second, and second and third, time points of patients who did change their medications and those who did not (*Supplementary Table 7*). Persons with CD who started new medications between time points displayed significantly higher Shannon diversity differences (*Figure 19e*) and intra-individual Jaccard distances (*Figure 19g*). This finding was not reproduced in persons with UC, which might be due to the higher microbial richness compared to persons with CD. Neither disease group showed significant differences in intra-individual species richness (*Figure 19d*) nor in Aitchison distances (*Figure 19f*) after starting a new medication. Thus, while presence or absence of certain taxa can be affected by a change in medication, this does not necessarily translate to overall differences in species richness. Additionally, it may have also been important to know when exactly a change in the medication regime has occurred. Unfortunately, this information was not available to us.

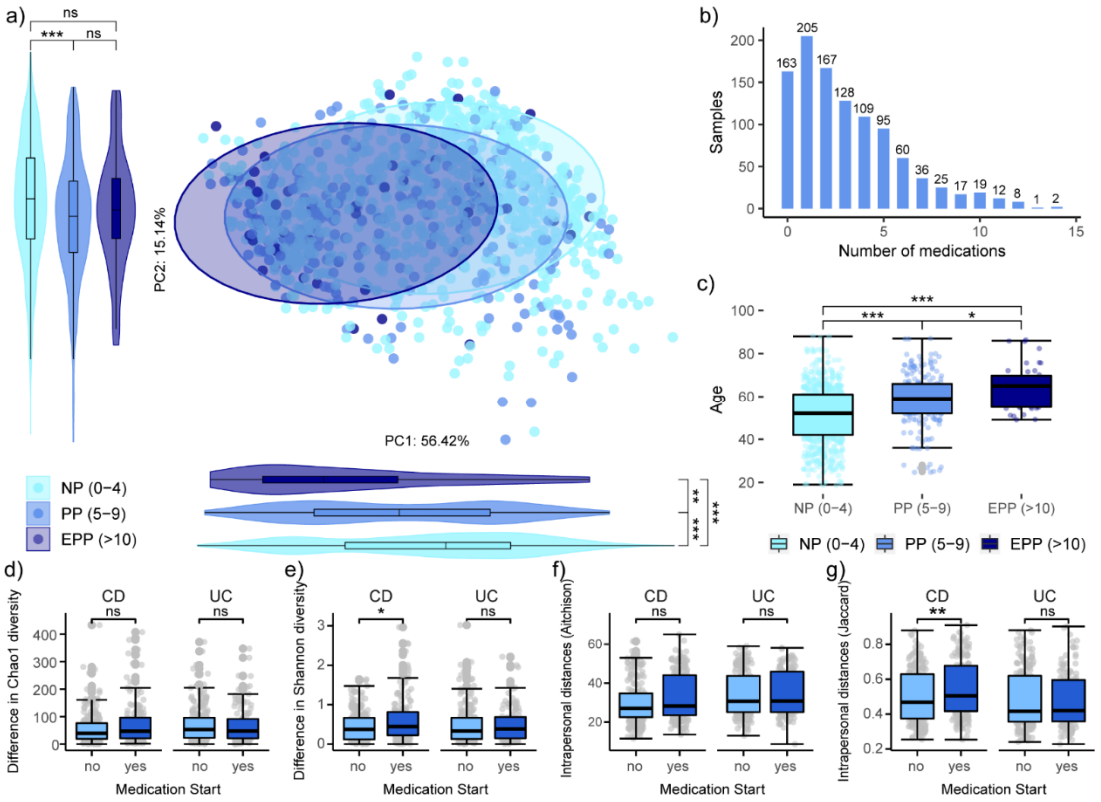


*Figure 19: (a) PCA based on Aitchison distances with CLR transformed OTUS with a prevalence of at least 10% grouped by polypharmacy. Violin plots show projections of the PCA points onto PC1 and PC2 (b) Histogram of the number of medications used per sample. (c) Comparison of the age of patients between the polypharmacy groups. Comparison of intra-personal (d) Chao1 diversity and (e) Shannon diversity differences, as well as (f) Aitchison and (g) Jaccard distances. Stars show significant differences between the groups as determined by Wilcoxon test; *p < 0.05; **p < 0.01; *** p < 0.001.*

3.4. Treatment explains more variation in microbial composition than IBD-subtype

Having established differences in both microbiota and treatment between locations, as well as microbiota-polypharmacy associations, we further assessed the overall effect of these treatments on the gut microbiota in more detail. Thus, we performed multivariate regression analysis of the explained microbiota variance (Aitchison distances) of different variables such as medications, patient demographics, environmental factors, and disease status and activity. The full explanatory regression models contained 16 of those variables, which together explained 7.88% of the total variation in microbial composition (R^2: 0.788; *Figure 20a, Supplementary Table 8*). This is slightly less than in our original study (9.7%) (Clooney *et al.*, 2021), and is likely due to the fact that here we considered a slightly different number of participants and no control subjects. A conservative estimate of the unique effect of each variable present in the full model was obtained by evaluating all the variables separately, while at the same time partitioning out the effects of all other included covariates. Thus, the difference in explained variation between the naive effect and the unique effect are the interaction effects between the variables. When taken together, all medications explained more microbiota variation than IBD-subtype alone (1.47% versus 0.25%), but less than surgical resection (1.68%) or geographic location (1.23%). It was also notable that the interaction effects associated with the proportion of variance explained by IBD-subtype were much higher (1.57%) than the interaction effects explained by geographic location (0.50%) (*Figure 20a, Supplementary Table 8*). Treatment (surgical resection and medications combined) explained more variation (3.48%) than all other clinical and environmental factors together (2.40%) *(Figure 20b*). The effects of disease remission and intestinal anti-inflammatory agents were positively correlated with UC, while surgical resection, drugs for peptic ulcers and gastro-oesophageal reflux disease (GORD), and antacids as well as EPP and Insulins and analogues were positively associated with the effect of CD on the gut microbiome (*Figure 20c*). In concordance to the previous differential taxa abundance analysis (*Figure 16e*), *Escherichia/Shigella*, *Klebsiella* and *Blautia* were most associated with CD, whilst a high abundance of several *F. prausnitzii*

species characterized persons with UC. Irish participants exhibited a higher abundance of Lachnospiraceae UCG-004 than Canadians, who in turn showed a higher proportion of *Escherichia/Shigella* and *Ruthenibacterium lactatiformans* (*Figure 20d*).

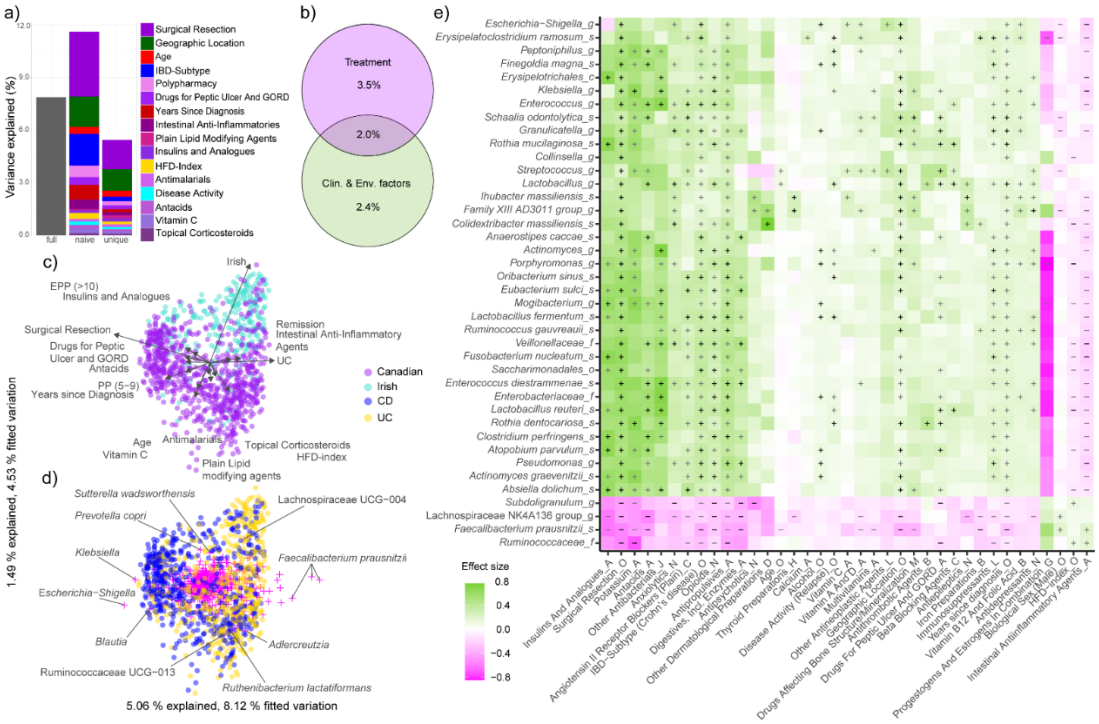


Figure 20: Variance explained (%) by (a) individual variables and (b) variable groups. "full" denotes the explained variance for all 16 factors combined, "naïve" the explained variance for each variables separately, and "unique" the explained variance for each variable after removing the interaction effects with all other variables. (c) Biplot of the dbRDA based on Aitchison distances on all operational taxonomic units (OTUs) present in >10% of samples, colored by geographic location. The arrows represent the effects of the constraining variables. (d) Same dbRDA but colored by IBD-subtype. Crosses denote species sites, and the labels represent the top 10% of OTUs with the best axis fit. (e) The top 40 taxa with the highest number of significant associations to metadata. Plus denotes positive associations, minus negative associations. Black signs show strictly deconfounded associations (i.e. the association can be reduced to this metadata variable/taxonomic feature pair even if another metadata variable is associated to the same feature), grey signs confounded ones (association is distorted by one or more other variables). Taxonomic rank: s: species, g: genus, f: family, c: class, o: order; ATC-levels: A: Alimentary tract and metabolism, B: Blood and blood forming organs, C: Cardiovascular system, D: Dermatologicals, G: Genito and urinary system and sex hormones, H: Systemic hormonal, excl. sex hormones and insulins L: Antineoplastic and immunomodulating agents, M: Musculo-skeletal system, N: Nervous system; O: Other metadata

3.5. Disease-associated bacteria are less affected by medications than beneficial bacteria

To further quantify individual drug-microbiota effects, we implemented a univariate statistical approach that identifies naïve associations between taxa and metadata for confounding effects, i.e. distortions of the association between taxonomic feature and metadata variable which are caused by additional variables. We included 79 treatments, which were recorded at least 5 times as "yes" within the study, as well

as 8 additional metadata variables (geographic location, IBD-subtype, disease activity, smoking status, alcohol consumption, HFD-index, years since diagnosis and biological sex). Of these, 49 variables had at least one significant association (FDR<0.05, Cliff's Delta >0.1) with a taxonomic feature in the gut microbiome (*Supplementary Table 9*).

When clustering the taxa and meta-data associations by their effect sizes (*Figure 20e*), we noticed that the majority of treatments seem to amplify the effect of IBD on the microbiota in so far that many of the associations between treatments and microbiota had the same direction (i.e. positive or negative) as the associations between IBD and microbiota. As detailed further below, many of the taxa that we found to be increased or decreased in patients with IBD were also correlated with commonly used treatments of these patients. The health-associated taxa *F. prausnitzii*, Lachnospiraceae NK4AI36-group, and *Subdoligranulum* (Aldars-Garcia *et al.*, 2021, Pittayanon *et al.*, 2020, Shanahan *et al.*, 2021, Deleu *et al.*, 2021) were not only depleted in CD compared to UC, but also had the highest number of (negative) correlations with other covariates (21, 20, 17 and 17 associations, respectively). Equally, many taxa that are positively correlated with IBD, such as *Escherichia/Shigella* (n=18), *Streptococcus* (n=22), *Klebsiella* (n=19) and Veillonellaceae (n=19), as well as bacteria from the oral cavity such as *Rothia dentocariosa* (n=22), *Fusobacterium nucleatum* (n=21) or *Oribacterium sinus* (n=19), were among those that showed the highest number of positive correlations with medication and supplements (*Figure 20e*). At higher taxonomic ranks, Verrucomicrobiae, Bacteroidia and γ-Proteobacteria were significantly (Fischer $p < 0.05$) less associated with medications than Actinobacteria, Bacilli and Clostridia, among others (*Figure 21a, Supplementary Table 10*). It was also notable that the Clostridia taxa Ruminococcaceae, Clostridiales vadinBB60 group and unclassified Clostridiales had the highest proportion of negative associations. While Lachnospiraceae were relatively unaffected by medications (*Figure 21c*), they also showed a relatively high proportion of negative associations (*Figure 21d, Supplementary Table 11*).

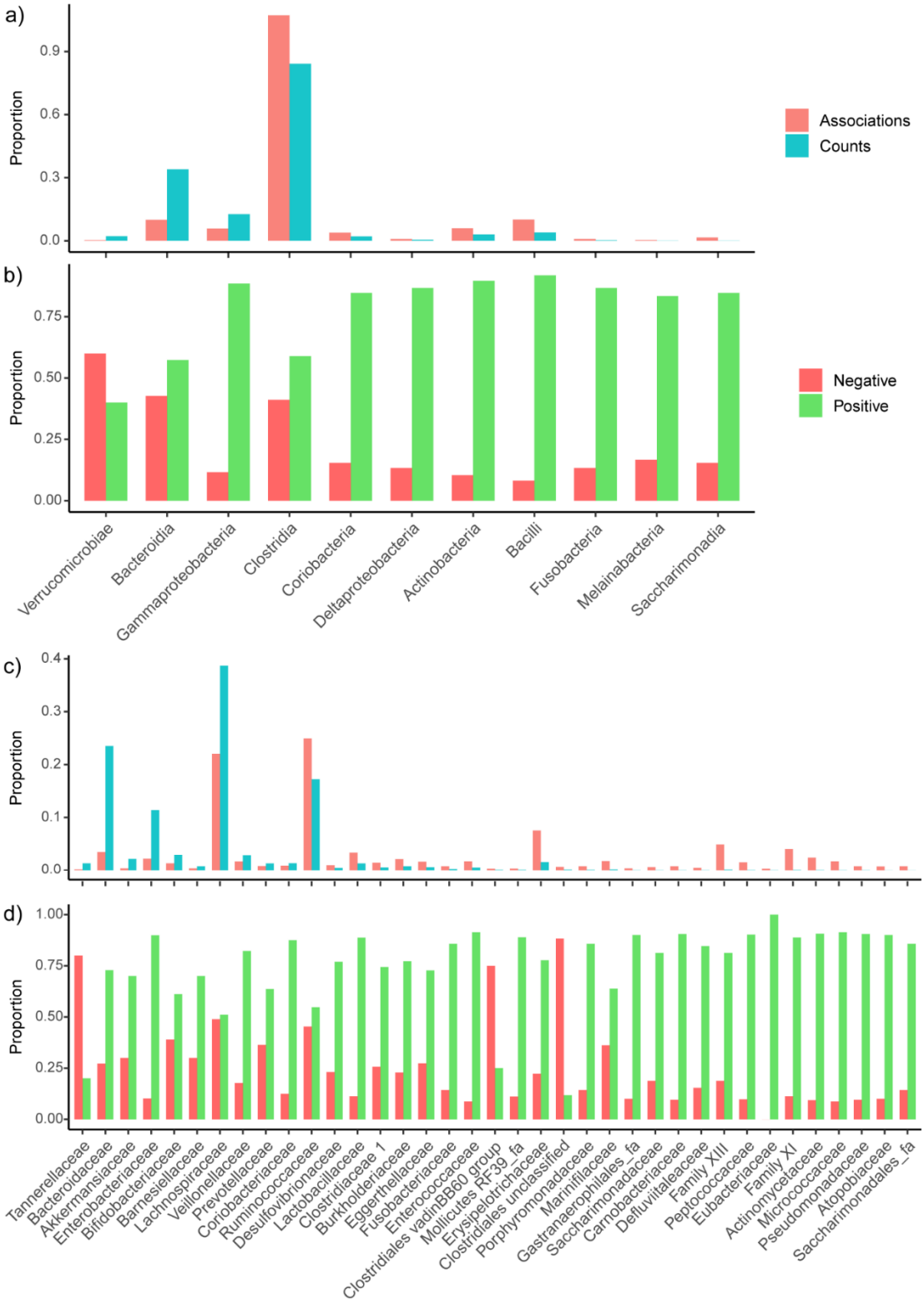


Figure 21: Taxa whose proportions of associations were significantly over- or under-represented compared to the overall distribution of OTU- counts at (a) class and (c) family level as determined by Fisher exact test. Proportion of positive and negative associations for (b) class level and (d) family level. Bars are ordered by the ratio of associations to counts and negative to positive associations, respectively.

3.6. Additive effects of multiple medications amplify the effect of IBD-subtype on gut microbiota

In terms of specific treatment-association to taxa, surgical resection (198 associations), antipropulsives (167), and intestinal anti-inflammatory agents (140) showed the highest number of significant associations. A combination of oestrogens and progesterons, chemotherapeutics for topical use, insulins and peripherally acting antiandrenergic agents, showed the strongest associations with single taxa (*Figure 22a*). Of the overall 2,754 significant associations between covariates and the taxonomic features, 35% were confounded i.e. the effect size value for a metadata/ taxonomic feature pair could be reduced to at least one other covariate while retaining its own significance (*Figure 22b*). Markedly, the proportion of confounded associations was lower for geographic location (16.8%) but much higher for IBD-subtype (60.4%). IBD-subtype showed interaction effects with either years since diagnosis, geographic location, biological sex or surgical resection, as well as 5 other treatments (A07E, B03A, A07D, N02A, and A12B in *Supplementary Table 2*). Moreover, 2,547 of all associations (92.5%) involve taxa that were also significant for the IBD-subtype, whereof 80% were in the same +/- direction and thus potentially amplifying the separation of the gut microbiome composition between persons with CD and persons with UC. For example, the immunosuppressant subgroup (L04A), which contains selective immunosuppressants such as mycophenolic acid but also tumour necrosis factor alpha (TNFα) inhibitors and other immunosuppressants like methotrexate and azathioprine showed significant associations with 124 taxonomic features. Of these, 120 were not only also significant for IBD-subtype but also co-directional. A notable exemption to this exacerbating effect were intestinal anti-inflammatories, a pharmacological subgroup that contains locally acting corticosteroids, and aminosalicylic acid and similar agents and shared 129 of its 140 significant associations with IBD-subtype, but none of those were co-directional, suggesting that they counteracted the effect of IBD-subtype towards a healthier microbiome. Other variables that reduced the effect of IBD-subtype were a combination of progesterone and oestrogen (54 shared, whereof only 4 were co-directional), HFD-index (106 shared, 6 co-directional), and biological sex (26 shared,

3 co-directional; *Figure 22c*). The amplifying effect for geographic location was much less pronounced, as 2,037 associations (80%) were shared with other variables and only 60% of those were co-directional. Surgical resection, Vitamin A/D supplementation and antipropulsives had the most co-directional associations with geographic location (111, 92 and 85 respectively). Years since diagnosis and IBD-subtype also showed a high overlap with the effect of geographic location (107 and 109 co-directional; *Figure 22d*). While most of these effects were weak or moderate (*Figure 22a*), the summation of the many co-directional effects from different treatments might obscure which taxa were most depleted or increased due to a variable of interest in this case IBD-subtype and to a lesser degree geographic location.

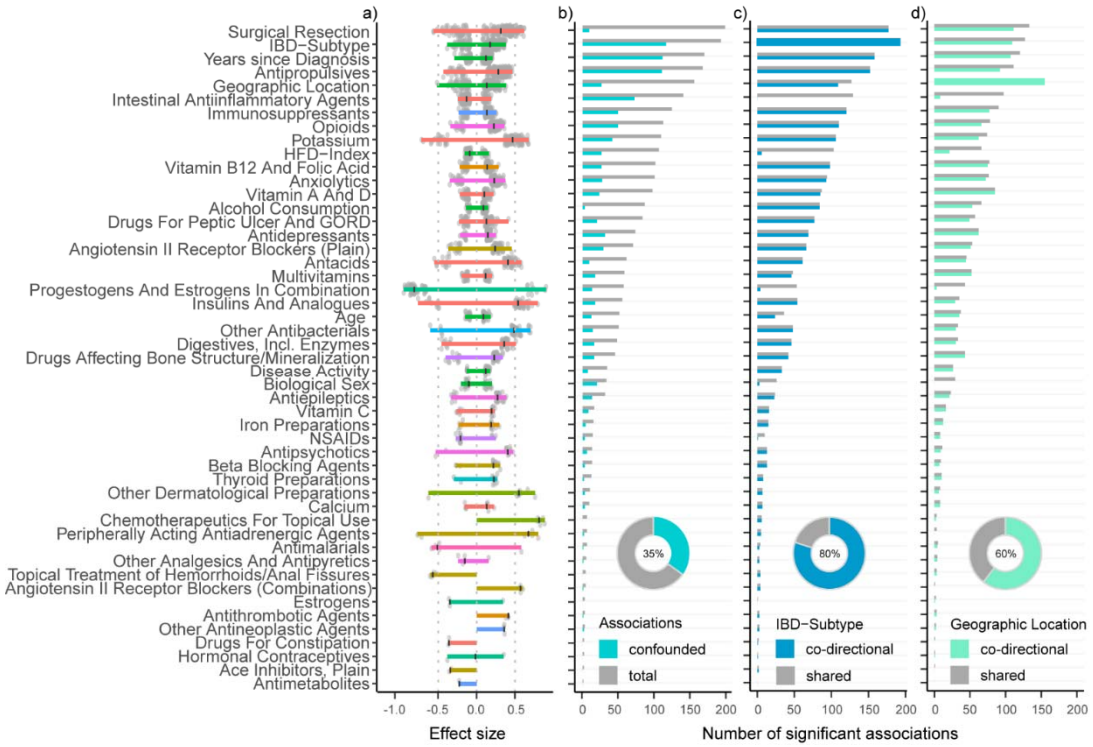


Figure 22: Effect sizes (a) and number of significant associations by metadata variable that are (b) confounded, (c) shared/co-directional with disease and (d) shared/co-directional with geographic location.

4. Discussion

In our previous study we showed that geographic location accounted for the second highest explained variance in gut microbial composition after a diagnosis with CD (Clooney *et al.*, 2021). The present study extends the earlier observations by addressing differences in medication profiles in greater detail. We discovered that multiple medications and supplements were differentially used between persons with UC and CD as well as between Canadian and Irish participants. Indeed, Canadians were found to take significantly more medications and supplements than their Irish counterparts and IBD-subtype accounted for only slightly more variation in drug usage patterns than geographic location. Despite this, when assessing the confounding effects of treatments on the microbiota, only a small part of the variation in microbial composition between patients from the different geographic locations was explained by the differences in medication and supplement intake.

In contrast, a major part of the disparity between the gut microbiomes of persons with UC versus CD seems to be due to, or amplified by, interaction effects with treatment. About half of the tested medications and supplements showed significant associations with at least one taxon from the gut microbiota and together treatments, including surgical resection and, medications and supplements, explained more variation in gut microbial composition than all other tested environmental factors.

Several taxa whose increase are generally reported with a shift away from healthy gut microbiome composition to an inflamed state, including *Escherichia/Shigella*, *Streptococcus*, *Klebsiella* and *Veillonellaceae* (Khan *et al.*, 2019, Aldars-Garcia *et al.*, 2021, Buttó *et al.*, 2016, Mirsepasi-Lauridsen *et al.*, 2019) were found here to be increased in the microbiome of persons with CD. These taxa also notably ranked amongst the highest number of positive associations with the tested covariates. Bacteria from the oral cavity such as *Rothia dentocariosa* and *Fusobacterium nucleatum*, which have been reported to be increased in PPI users (Weersma *et al.*, 2020, Clooney *et al.*, 2016a) showed not only significant positive associations with drugs for peptic ulcers and GORD in the present study but also belong to the top 40 most affected bacteria. A depletion of *Faecalibacterium prausnitzii*, was here and in

other studies associated with CD (Pittayanon *et al.*, 2020, Clooney *et al.*, 2021, Sokol *et al.*, 2008, Melgar *et al.*, 2010) and was also described in persons treated with the immunosuppressant azathioprine (Effenberger *et al.*, 2021). And whilst we were unable to confirm a negative association of *F. prausnitzii* with the intake of immunosuppressants, it was among the taxa which had the most negative associations with the tested drugs. The lack of diminution of this taxon by immunosuppressants here, might be explainable due to the fact that the ATC subgroup L04A not only includes thiopurines but also TNFα inhibitors, latter of which have been shown to increase SCFA producing bacteria (Estevinho *et al.*, 2020). It is noticeable though that immunosuppressants share nearly all their significant associations to taxonomic features with IBD subtype and all of those shared associations showed the same directionality and thus increases the disparity between the gut microbiota of persons with UC and CD. In contrast to that, intestinal anti-inflammatory agents are among the few drugs that did not follow this exacerbating trend. While this medication subgroup also shared most of its significant associations to taxa with IBD-subtype, none of them were co-directional. This observation is in agreement with earlier reports that 5-ASA drugs can partially recover the gut microbiome to a healthy status (Clooney *et al.*, 2021, Olaisen *et al.*, 2019).

Comparing the distribution of the number of reads and significant associations to medications for each taxon showed that some taxa were more resistant towards the effect of human targeted drugs, e.g. Verrucomicrobiae, Bacteroidia and γ-Proteobacteria, while others, such as Actinobacteria, Bacilli and Clostridia, were more sensitive. These results concurred with a study from Maier *et al.* which found γ-Proteobacteria to be more drug resistant than the highly abundant commensals like *Roseburia intestinalis*, *Eubacterium rectale*, and *Blautia obeum* (Maier *et al.*, 2018).

While this study couldn't confirm treatment as a major factor explaining dissimilarities in gut microbiota of persons from different geographic locations, it shows that the highly variable medications profiles of persons with IBD and their effect on the faecal gut microbiota likely impede the discovery of a universally valid

microbial signature distinguishing the IBD-subtypes and at least in part, explain the high disparity between different IBD microbiome studies. Furthermore, it highlights the need to include an exhaustive list of medication intake (and ideally dosages) of study participants in the analysis that go beyond the common IBD therapeutics to improve reproducibility between IBD-studies.

There are several limitations to our study. Due to the highly variable drug usage patterns not all medications could be assessed for univariate analysis. The evaluation of the confounding effects of medication on gut microbiota was further hampered by extreme polypharmacy and a multitude of changes in medication regimes over the course of the study. It is therefore possible that some of the reported effects of particular medications were under-estimated. However, while the exact effect sizes of single medications may not be robust, taken together they reveal a trend that treatments can exacerbate disease-associated shifts by additive effects of multiple medications on the microbiome as well as by affecting some groups of microbiota more than others. It is noteworthy though, that all described effects between medication and microbiome in this study are associative rather than causative and thus intervention studies with treatment naïve persons with IBD and animal models will be needed to comprehensively disentangle the role of treatment and disease on the variation the gut microbiome of IBD patients.

University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

CORA
Cork Open Research Archive
Cartlann Taighde Oscailte Chorcaí

Eckenberger, J. 2023. The gut microbiome in inflammatory bowel disease and its confounders. PhD Thesis, University College Cork.

Please note that Chapters IV & V (pp. 149-245) are unavailable due to a request by the author.

CORA Cork Open Research Archive http://cora.ucc.ie

General Discussion

Despite substantial efforts in the research field, the precise cause of IBD remains elusive. The prevailing theory for IBD pathogenesis is that a combination of genetic susceptibility and environmental factors trigger an abnormal immune response against the gut microbiota (Kostic *et al.*, 2014, Matsuoka *et al.*, 2015). Evidence from both animal models and human subjects strengthens the case for the microbiome's involvement in IBD. The absence of gut microbes in germ–free mice prevented the development of colitis in genetically susceptible mice (Veltkamp *et al.*, 2001), and transferring pro-inflammatory bacteria from diseased to healthy mice induced inflammation (Schaubeck *et al.*, 2016, Du *et al.*, 2015). Moreover, the transfer of naive CD4+ lymphocytes from healthy mice into those lacking T and B lymphocytes was shown to induce colitis (Powrie, 1995), contingent on the differences in their gut microbiota (Reinoso Webb *et al.*, 2018).

Observations in humans also align with these findings, as disease activity in IBD is most pronounced in areas with high bacterial populations in close proximity to the affected tissues, such as the terminal ileum and the colon, where faecal material remains relatively stagnant (Martinez-Medina *et al.*, 2014). In patients with Crohn's disease (CD), surgical manipulation of the continuity of the faecal stream has been shown to improve inflammation in the excluded part, despite little change in histopathology (Janowitz *et al.*, 1998). Antibiotic therapy has proven effective in certain IBD cases (Thia *et al.*, 2009, Perencevich *et al.*, 2006, Turner *et al.*, 2014), and specific microbes have been identified as drivers or suppressors of inflammation (Henke *et al.*, 2019, Chassaing *et al.*, 2013, Zhou *et al.*, 2020). There is also growing interest in the ameliorative potential of prebiotics, probiotics, and even faecal microbiota transplantation for some IBD patients (Gionchetti *et al.*, 2003, Zocco *et al.*, 2006, Zhang *et al.*, 2022b). Genome-wide association studies focusing on IBD have identified over 230 independent risk single nucleotide polymorphisms (de Lange *et al.*, 2017, Turpin *et al.*, 2018). A substantial portion of these genetic risk factors revealed strong associations with immune function, and specifically with the immune system's interaction with the microbiota (Aschard *et al.*, 2019, Jostins *et al.*, 2012, Liu *et al.*, 2015, Lavoie *et al.*, 2019).

Thus, it is now widely accepted that the microbiome plays a key role in IBD pathogenesis, albeit the exact mechanisms, causality and microbial agents are not equivocally understood (Pittayanon *et al.*, 2020, Aldars-Garcia *et al.*, 2021, Khan *et al.*, 2019). A multitude of studies, including our own (Clooney *et al.*, 2021), consistently reported differences between patients with IBD and controls including lower microbial diversity, decreases in commensal bacteria, increases in pathobionts and oral microbiota, as well as variations in bile acid profiles (Upadhyay *et al.*, 2023). However, inconsistencies in methodologies, cohorts and study design, as well as variations in reported ecological impacts and their magnitudes across IBD studies hamper the identification of clinical and biological factors that can best predict disease progression and treatment outcomes over time (Aldars-Garcia *et al.*, 2021, Abdel-Rahman *et al.*, 2023). Importantly, these factors are a key requirement for precision medicine, needed to tailor treatments to individual patients based on their clinical and biological characteristics, and thereby optimizing patient care (Borg-Bartolo *et al.*, 2020, Kashyap *et al.*, 2017). Yet, even after extensive work in the field a specific microbial signature for IBD or other diseases remains elusive (Pizarro *et al.*, 2019, Khan *et al.*, 2019). In this thesis we identified multiple sources of variation between subjects and studies contributing to this challenge, emphasizing the need for continued research.

Chapter II of this thesis examined the faecal gut microbiota of a large population of patients with CD and UC relative to healthy controls from two geographically separated regions but with similar socioeconomic background over the course of several months. The longitudinal design allowed for the observation of active and inactive disease states, while the collection of anthropometrics as well as medication intake, dietary information and lifestyle variables enabled the identification of potentially confounding effects. Similar to other studies (Pittayanon *et al.*, 2020, Abdel-Rahman *et al.*, 2023, Franzosa *et al.*, 2019) we observed a decreased microbial diversity in patients with IBD and altered faecal microbiota communities in patients with UC or CD compared to healthy controls. Furthermore, we confirmed an elevated compositional instability of gut microbiota in patients with fluctuating disease activity (Scanlan *et al.*, 2006, Halfvarson *et al.*, 2017, Lloyd-Price *et al.*, 2019). Although patients with active disease, especially those with UC, showed only a modest shift further away from

healthy controls and a slight decline in microbial diversity compared to their remittent counterparts, we did not observe a difference in relative species abundances between patients in relapse or remission in this study. While a ML approach was able to successfully discriminate between patients with UC or CD and healthy controls, and to a lesser extent, between the IBD subtypes based on relative abundances of taxa at a given point in time, separating patients with active and inactive disease state was only possible when considering temporal changes, indicating that cross-sectional study design does not sufficiently capture all necessary information. Moreover, an increased sampling frequency in future studies might not only allow to seize the exact moment of a relapse, but such information could potentially enable relapse prediction based on gut microbial changes, allowing a timely medical intervention for the benefit of the patient.

However, it is not only disease related factors that influence the gut microbial community. A multitude of factors encompassing almost all facets of human existence have been shown to shape the gut microbiome, including but not limited to ethnicity (Deschasaux *et al.*, 2018) birth mode (Dominguez-Bello *et al.*, 2010), breastfeeding (Penders *et al.*, 2006), ageing (Claesson *et al.*, 2012), cohabitation (Rothschild *et al.*, 2018), diet (Johnson *et al.*, 2019, Asnicar *et al.*, 2021), physical activity (Clarke *et al.*, 2014), smoking (Biedermann, 2013) and alcohol consumption (Falony *et al.*, 2016). Notably, in our study all 25 included factors explained only 10% of the microbiota variance. While this is consistent with other studies (Falony *et al.*, 2016, Zhernakova *et al.*, 2016) it shows that most of the variability in gut microbial communities is yet unaccounted for. Likely contributors to this unexplained variance include uncollected patient information, inter-individual variability and stochastic noise. Hence any biological conclusions researchers draw should be considered in respect to the great unknown.

In terms of assessing and ranking the collected sources of microbiota variation, we discovered that geographic location exerted the most significant non-disease-related influence. This influence surpassed the impact of a diagnosis of UC and even previous surgical resection, coming second only to the impact of a diagnosis of CD. Geographic location may reflect variations in a wide range of environmental factors, such as

ethnicity, lifestyle, diet, food additives and pollutants, healthcare differences, or social networks. This does, however, not only underscore the critical importance of considering environmental factors when designing studies, but also provides a potential explanation for inconsistencies among different microbiota studies.

Moreover, the increasing evidence of not only antibiotics but also commonly prescribed non-antibiotic drugs interacting with the gut microbiome (Maier *et al.*, 2018, Weersma *et al.*, 2020, Radhakrishnan *et al.*, 2022), raises concerns about potential distortions in microbial community. Inadequate control for drug treatments can introduce noise into data, hindering our ability to accurately attribute microbiota variations to their true causes. This issue is exemplified in studies on type 2 diabetes mellitus and gut microbiota, where metformin treatment obscured microbial signals, leading to conflicting conclusions (Zhang *et al.*, 2013, Qin *et al.*, 2012, Karlsson *et al.*, 2013, Forslund *et al.*, 2015).

In **Chapter III**, we therefore undertook a thorough re-evaluation of microbiota variance among patients with IBD, with a specific focus on investigating the microbiota impact of potential confounding factors related to treatment and supplement usage. Our study revealed notable disparities in treatment regimens, not only between subtypes of IBD but also across different countries of origin. However, a multivariate regression analysis assessing the confounding effects found that the combined impact of medication and supplement intake, while significant, accounted for only a modest portion of the observed variation in gut microbiota composition attributed to geographic location. Interestingly, the additive effects stemming from the usage of multiple medications not only explained a greater degree of variation in gut microbiota profiles than the specific IBD subtype but they also exacerbated the differences in patients diagnosed with CD and UC. In our study patients with CD or UC ingested 302 different chemical substances, which were subsequently combined to 120 pharmacological subgroups. About half of those displayed significant associations between at least taxon from the gut microbiota. Notably, several taxa typically associated with a microbiota shift towards an inflamed state, including *Escherichia/Shigella, Streptococcus, Klebsiella, Veillonellaceae,* and *Fusobacterium nucleatum*, not only exhibited increased presence in the microbiomes of CD patients

but also had the strongest positive correlations with medication and supplement usage. Conversely, generally health-associated taxa, such as *F. prausnitzii*, *Lachnospiraceae NK4AI36-group* and *Subdoligranulum*, were not only depleted in CD when compared to UC but also displayed the most prominent negative correlations with the tested pharmacological subgroups. Comparing the distribution of taxa and their associations to medications showed that *Verrucomicrobiae, Bacteroidia,* and *γ-Proteobacteria,* exhibited greater resilience to the influence of human-targeted drugs. In contrast, *Actinobacteria, Bacilli*, and *Clostridia* were more susceptible to medication effects. These findings align with a study by Maier *et al.*, which identified *γ-Proteobacteria* as more resistant to drugs compared to highly abundant commensals like *Roseburia intestinalis, Eubacterium rectale*, and *Blautia obeum* (Maier *et al.*, 2018). Consequently, it becomes evident that at least some of these microbial variations can be attributed to treatment and supplement usage among the groups under investigation.

When we compare the substantial diversity of medications and supplements ingested in our study to the typically reported medication profiles, a noticeable knowledge gap emerges. Given that the prevalence of polypharmacy is high in IBD patient cohorts (Cross *et al.*, 2005, Mesonero *et al.*, 2022), this discrepancy likely results from underreporting rather than our study being an outlier. Assuming that participants in other IBD microbiome studies consume similar quantities of pharmaceutical substances, it is plausible that their medication profiles are equally diverse and potentially distinct across studies. This variability in medication profiles across studies likely contributes to confounding effects on faecal gut microbiota, potentially hindering the discovery of a universally applicable microbial signature for IBD. It also underscores the importance of incorporating comprehensive documentation of medication intake, extending beyond the common treatments of the disease under study, and ideally including dosage information in order to enhance the reproducibility between IBD studies and separate disease-associated effects from treatment-associated effects.

Variability in gut composition is not only influenced by external factors such as diet, medication and lifestyle choices, but also by host-derived factors, such as BAs, as well

as the products of the gut microbial metabolism, which add another dimension to the observed variability in gut structure. Primary bile acids (PBAs), originating from the host, possess bacteriostatic properties and can impact microbial composition. Gut microbes detoxify BAs, altering the pool of available BA in the gut. Microbiota-modified secondary bile acids (SBAs) further influence microbiota composition and concurrently modulate host immunity (Thomas *et al.*, 2022). BA malabsorption (BAM) is a condition in which the digestive system fails to reabsorb BAs properly in the small intestine, resulting in an excess of BA in the colon and often increased intestinal motility and so called BA diarrhoea (BAD). It is a common manifestation in IBD, particularly affecting patients with ileal inflammation or resection (Vítek, 2015).

In **Chapter IV**, we assessed faecal and circulating levels of metabolites associated with the BA-sensing farnesoid X receptor (FXR) in conjunction with faecal microbiota composition of patients with CD, both in those who have previously undergone surgical resection and those who have not. Our findings indicated that the normal signalling from the gut to the liver is disrupted in patients with CD relative to healthy controls, resulting in excessive hepatic BA synthesis, reminiscent of BAM/BAD, in a subset of patients. We observed signs of BAM/BAD in approximately one-third of patients with CD, with a notably higher prevalence among those who had undergone resection surgery; however, fewer than half of these individuals received treatment for this condition. This places the BAM incidence in our study cohort toward the lower end of the reported prevalence range, with previous studies indicating higher rates in both non-resected and resected patients (Barkun *et al.*, 2013, Bromke *et al.*, 2021). Furthermore, among CD patients, differences in surgical resection history were associated with variations in hepatic BA synthesis (directly measured through 7α-hydroxy-4-cholesten-3-one levels) and BA reabsorption (indirectly assessed by fibroblast growth factor 19 concentrations). Within the CD cohort, the rate of hepatic BA synthesis and the resulting increased levels of PBAs accounted for more variation in the gut microbial composition than the patient's resection status. They also showed a similar association pattern with taxonomic features of the gut microbiome as the disease-phenotype, suggesting that changes in BA metabolism likely contribute to the observed disease- and resection-associated shifts in the gut microbiota. The group of

patients with elevated BA production exhibited significantly higher relative abundances of the bile-resistant genus *Enterocloster* (Vich Vila *et al.*, 2023) as well as a depletion of health associated butyrate producing bacteria like *Faecalibacterium*, *Lachnospiraceae, Ruminococcaceae, Christensenellaceae* and *Subdoligranulum* (Aldars-Garcia *et al.*, 2021). Recent studies have demonstrated the ability of *Enterocloster bolteae*, formerly known as *Clostridium bolteae*, to engage in the re-conjugation of BAs with amino acids, resulting in the formation of compounds known as microbial conjugated BAs (MCBAs) (Quinn *et al.*, 2020). The amide conjugations occurring at the C24 acyl site cause notable alterations in their chemical properties, primarily an increase in hydrophobicity, which has been proposed as a protective mechanism employed by bacteria to counteract potential toxic effects (Garcia *et al.*, 2022). On the other hand, the depletion of beneficial species may be linked to their relative sensitivity to bile (Lopez-Siles *et al.*, 2012, Foditsch *et al.*, 2015) as well as the pathology of BAM, which can manifest in watery diarrhoea. Research has consistently demonstrated that diarrhoea can lead to significant alterations in the composition of the gut microbiota. These changes are characterized by an increase in *Proteobacteria* and a reduction in butyrate-producing bacteria, including members of the *Clostridia* families *Lachnospiraceae* and *Ruminococcus* (Duvallet *et al.*, 2017). The observed shifts in microbial composition and their associations with FXR-associated metabolites strongly suggest that changes in BA metabolites significantly contribute to the observed differences in microbial composition between health and disease. This study further illustrates that both CD and a patient's resection status significantly affect BA signalling, leading to altered BA profiles and therefore increased variability among patients with CD.

Combined, these results demonstrate that microbiomes are dynamic ecosystems which not only have to adapt to shifting environmental conditions but are also under active regulation by the host (Staley *et al.*, 2017, Benson *et al.*, 2010). This added complexity, as well as the increasing data volumes, necessitate innovative analytical approaches to capture all available information, especially considering the unique characteristics of microbiome data. These data are not only compositional (Gloor *et al.*, 2017) but also exhibit profound heterogeneity and high dimensionality, yet they

are often sparse in nature. The capacity of ML to harness computer algorithms to uncover intricate patterns within extensive datasets, without the need for prior knowledge of the underlying data to be explicitly coded into the model, renders it exceptionally well-suited for the demands of microbiome research.

In **Chapter V**, a comparative methods analysis was therefore conducted to assess various non-linear ML algorithms (Support Vector Machine, Random Forest, and Extreme Gradient Boosting) along with different filtering and normalization techniques for the prediction of host-phenotypes, specifically distinguishing between CD from control groups, as well as differentiating IBD subtypes based on taxonomic features from the gut microbial community. The internal validation, via a nested for loop approach and external validation via publicly available independent datasets, showed a comparable discriminatory performance, indicating a good generalizability of the models despite differences in treatment status, geographic location and study cohort. The models from all three algorithms were well able to distinguish between health and disease. Notably, in a meta-analysis of 28 microbiome studies of 10 different diseases, Duvallet *et al*. observed that many genera associated with individual studies respond to more than one disease, suggesting that many association are shared non-IBD-specific responses (Duvallet *et al.*, 2017), while others reported many of those taxa also being associated with unhealthy aging (Ghosh *et al.*, 2022) or non-antibiotic medications (Maier *et al.*, 2018). Thus the signal between Crohn's patients and healthy controls likely consists of disease specific and non-disease specific alterations of taxa. This also aligned with our observation that patients with several confounding patient characteristics, such as higher age, longer disease duration, higher medication intake or previous surgical resection, were less likely to be misclassified. Conversely, the classification performance for the distinction of IBD subtype was a lot less profound, even though all algorithms were able to build classifiers that were better than random for the distinction of IBD subtype. This drop in performance might on the one hand reflect the fact that only disease-specific signals are useful in this case, which are a lot less pronounced (Clooney *et al.*, 2021), and on the other hand showcases the heterogeneity of patients with IBD (Verstockt *et al.*, 2022).

However, in microbiome research the goal is not solely to achieve precise classification between groups of interest, but also identifying specific microorganisms or their metabolic by products that directly exert an influence on the host phenotype. The efficacy of a discriminatory model is therefore just as important as the taxa that were successfully used to train such a model. In fact, interpretability is imperative for any prospective application of ML systems in medicine and healthcare (Vellido, 2020, Yoon *et al.*, 2021). Notably, contemporary complex (non-linear) ML methodologies are renowned for their inherent opaqueness, and increasing interpretability and explainability of ML models is the focus of ongoing research (Tjoa *et al.*, 2021).

An analysis of feature importance, which measures the individual contribution of specific predictors to the model performance, found several butyrate producing commensals including several members of the *Clostridium cluster XIVa* as well as potential pathogens such as *Eggerthella lenta* to be most important for the classification of health versus disease within our data. For the distinction of IBD subtypes, a depletion of *Faecalibacterium prausnitzii, C. leptum* and *C. ruminantium* separated patients with CD from patients with UC. Yet, when comparing the contributions across different models for the same classification task, there was only minimal overlap in the set of important features between models of different algorithms. Only three OTUs were found in top 20 most important features in all models for the classification of Crohn's disease, while the classification of the IBD subtype gained four shared OTUs. Part of the disparity in selected features between the models certainly arises due to the different underlying mathematical modelling approaches of the classifiers, as well as differences in methods to obtain feature importance and lastly the metric of importance itself (Saarela *et al.*, 2021), but some variation will also stem from redundancy in the feature space. If we consider multiple predictors that each have the same correlation with the response variable, but are not correlated among each other, then adding any of the predictors to a model will reduce the variance of the response variable to a certain, equal amount. However, if the predictors themselves are correlated with each other, then the earlier predictors will reduce the variance in the response to a higher degree than the later added predictors because much of the predictive power has already been accounted for, although the

predictors in reality may have the same correlation to the response variable (Hand, 2006) This is problematic, as we know that microbiome data are highly collinear. But while we can computationally circumvent misinterpretation of feature importance by grouping correlated features together or removing them altogether (Molnar *et al.*, 2022) this further complicates the identification of specific candidates that directly affect the host phenotype.

Not accounting for collinearity when applying or evaluating ML based approaches is not the only pitfall of ML based science. A recent study by Quinn evaluated 102 ML-based microbiome papers and found that the vast majority of them had methodological problems, first and foremost of these being test set omission and test set data leakage (Quinn, 2021). But microbiome research is not the only field that struggles with methodological problems when applying ML, a study from Kapoor and Narayanan surveyed ML reviews across a broad range of scientific fields for issues and likewise reported a multitude of methodological problems, such as lack of clean separation of training and test set, use of illegitimate features, non-independence of training and test samples and selection bias. Along with limited data availability and the use of inappropriate methods or performance metrics this is widely causing irreproducible results and overoptimistic performance estimates (Kapoor *et al.*, 2023, Greener *et al.*, 2022). While the application of ML to biological data has its own challenges, these methodological inconsistencies only add to the so called "reproducibility crisis" in microbiome research. In additional to inter-lab differences in sample processing, i.e. sample storage, DNA extraction method, primer choice, amplification, or choice of sequencing platform, which can all have an substantial effect on the resulting data, pre-processing of the raw sequences data and taxonomic assignment can influence the outcome of a metagenomics experiment (Salonen *et al.*, 2010, Kuczynski *et al.*, 2011, Clooney *et al.*, 2016b, Golob *et al.*, 2017, Cardona *et al.*, 2012). A lack of universally accepted standards for data processing and analysis of microbiome data can cause technical batch effects that surpass biological effects (Goodrich *et al.*, 2014) and impede comparability between studies. Moreover, high inter- and intra-personal variability, the influence of environmental factors (Rothschild *et al.*, 2018), as well as difficulties in actually defining a healthy microbiome (Shanahan

et al., 2021) for case-control studies making it a challenge to establish consistent patterns between different study cohorts. Incomplete reporting of study details, limited data sharing and incomplete metadata can also hinder reproducibility (Huang *et al.*, 2021).

Future perspectives

Projects like International Human Microbiome standards (IHMS) or more recently the Human Microbiome Action (HMA) aim to improve microbiome research methods and develop a set of minimum standards in microbiome research to increase international comparability and aid the translation of findings into practical clinical applications and public health guidance. In particular the use of ML for human microbiome research has been further highlighted and optimised in the recently completed EU COST Action ML4Microbiome (D'Elia *et al.*, 2023, Papoutsoglou *et al.*, 2023). Increasing reproducibility is just one aspect in these goals, in addition to harmonizing the collection, pre-processing, and analysis of microbial samples for effective data sharing and the integration of multiple study cohorts and data types. In order to elucidate the complex interplay between genetic, environmental, and microbial factors in the context of IBD, robust research will require very large, well-defined study cohorts encompassing various geographic locations and diverse treatment patterns. Study populations that accurately represent the entire IBD population without biases toward, for example, specific ethnicities, socioeconomic backgrounds, or exposomes are essential to establish causal effects for certain environmental factors on IBD and which biological signatures can be used as indicators of specific clinical outcomes (Scott *et al.*, 2019). Given the substantial cost and time investment required to generate such comprehensive, deeply phenotyped data, achieving this goal is likely beyond the capacity of a single research group. Thus, it calls for extensive collaboration and community-wide efforts, and ultimately underscoring the critical need for FAIR (Findable, Accessible, Interoperable and Reusable) data sharing, the implementation of which is currently still found to be deficient (Hughes *et al.*, 2023, Wilkinson *et al.*, 2016).

Furthermore, longitudinal studies with frequent sampling time-points are needed to allow for the investigation of the natural progression of IBD pathology and intra-individual changes of the microbiome in respect to environmental triggers or

treatments. Following patients longitudinally would also eliminate inter-personal variation among patients, in so far as patients can be used as their own control, thereby facilitating the identification and validation of robust unbiased biomarkers for disease susceptibility, prognosis, and treatment response, which are essential to the concept of precision medicine (Lloyd-Price *et al.*, 2019).

Contemporary microbiome studies demand a shift beyond the conventional practice of community profiling towards a more nuanced exploration of microbial functionality. While researchers have only recently begun to direct their attention to the non-bacterial subpopulations of the gut microbiome and begin to understand microbial interactions among themselves and with the host organism, this is not sufficient to answer which species are present in a certain environment (Laforest-Lapointe *et al.*, 2018, Zhang *et al.*, 2022a, Du Toit, 2022). Given the functional redundancy between different microbial species and functional dissimilarity within the same species, host-microbiota interactions should extend to the metabolic environment and functional capacity (Schirmer *et al.*, 2018). For instance, at the species level gut microbiomes are not readily discernible between patients with active and inactive diseases, but they differ within their functional repertoire (Lloyd-Price *et al.*, 2019). Nuanced changes in function might allow potential therapeutic targets, diagnostic markers, and disease mechanisms to be elucidated.

Genomics data are anticipated to become one of the most prominent examples of big data in the coming decades, potentially surpassing domains like astronomy and the internet (Stephens *et al.*, 2015). This seems to present an ideal opportunity for the application of ML to address problems where traditional data analysis approaches may encounter limitations. However, caution is warranted when drawing conclusions from models trained on specific datasets, as these may not accurately reflect the characteristics of the underlying study population (Clemmensen *et al.*, 2022). Even with the emergence of model-agnostic approaches to feature importance, interpretability, and explainability, especially for complex non-linear ML methods, challenges persist. Given the high inter-correlatedness among members of the microbiome, our ability to definitively identify the contributions of various features to a model is limited (Molnar *et al.*, 2022). Consequently, the primary focus should be on

employing models to accurately predict risk factors, such as disease progression or treatment response, rather than relying on models to explain disease phenotypes (Greener *et al.*, 2022). Moreover, it is imperative to exercise diligence to ensure that ML models are used correctly and only where they are well-suited. It is therefore recommended that scientific claims made through predictive modelling are accompanied by so-called model cards. These cards provide crucial standardized information about a particular machine learning models and are designed to enhance transparency, accountability, and responsible AI practices (Mitchell *et al.*, 2019, Kapoor *et al.*, 2023).

Final Conclusion

The research conducted in this thesis quantified both intrinsic and extrinsic factors that influence the composition of the human gut microbiome, which can confound shifts associated with disease. Notably, in the context of medications and polypharmacy, these factors explained a greater degree of variation in the structure of the gut microbial community than the disease phenotype itself. This underscores the imperative need for studies with well-described metadata that enable us to adjust for confounding factors in microbiome analysis and sheds light on the challenges of reproducibility in IBD studies. Moreover, the research demonstrated that a cross-sectional study design alone is insufficient to distinguish between patients in relapse and those in remission. This distinction became possible only when incorporating temporal shifts in taxonomic features, indicating that intra-personal variability can be leveraged to increase the diagnostic and prognostic power of microbiome data. Lastly, a comparative analysis of various ML algorithms aimed at classifying IBD disease phenotypes based on taxonomic features in the gut microbiome revealed similar effectiveness across algorithms. However, these algorithms placed importance on different members of the microbial community when training models, underscoring the need for caution when employing machine learning for tasks that aim to comprehend underlying biological aspects rather than simply accurately predicting specific phenotypes.

Bibliography

ABBAS-EGBARIYA, H., HABERMAN, Y., BRAUN, T., HADAR, R., DENSON, L., GAL-MOR, O. & AMIR, A. 2022. Meta-analysis defines predominant shared microbial responses in various diseases and a specific inflammatory bowel disease signal. *Genome Biology,* 23**,** 61.

ABDEL-RAHMAN, L. I. H. & MORGAN, X. C. 2023. Searching for a Consensus Among Inflammatory Bowel Disease Studies: A Systematic Meta-Analysis. *Inflamm Bowel Dis,* 29**,** 125-139.

ABOU-ISMAIL, M. Y., CITLA SRIDHAR, D. & NAYAK, L. 2020. Estrogen and thrombosis: A bench to bedside review. *Thromb Res,* 192**,** 40-51.

ABRAHAM, C. & CHO, J. H. 2009. Inflammatory bowel disease. *New England Journal of Medicine,* 361**,** 2066-2078.

ACHESON, E. D. & TRUE LOVE, S. C. 1961. Early weaning in the aetiology of ulcerative colitis. A study of feeding in infancy in cases and controls. *Br Med J,* 2**,** 929-33.

ACINAS, S. G., MARCELINO, L. A., KLEPAC-CERAJ, V. & POLZ, M. F. 2004. Divergence and redundancy of 16S rRNA sequences in genomes with multiple rrn operons. *J Bacteriol,* 186**,** 2629-35.

AFIAHAYATI, SATO, K. & SAKAKIBARA, Y. 2015. MetaVelvet-SL: an extension of the Velvet assembler to a de novo metagenomic assembler utilizing supervised learning. *DNA Res,* 22**,** 69-77.

AINSWORTH, D., STERNBERG, M. J. E., RACZY, C. & BUTCHER, S. A. 2017. k-SLAM: accurate and ultra-fast taxonomic classification and gene identification for large metagenomic data sets. *Nucleic Acids Res,* 45**,** 1649-1656.

AITCHISON, J., BARCELÓ-VIDAL, C., MARTÍN-FERNÁNDEZ, J. A. & PAWLOWSKY-GLAHN, V. 2000. Logratio Analysis and Compositional Distance. *Mathematical Geology,* 32**,** 271-275.

ALAM, M. T., AMOS, G. C. A., MURPHY, A. R. J., MURCH, S., WELLINGTON, E. M. H. & ARASARADNAM, R. P. 2020. Microbial imbalance in inflammatory bowel disease patients at different taxonomic levels. *Gut Pathog,* 12**,** 1.

ALAMEDDINE, J., GODEFROY, E., PAPARGYRIS, L., SARRABAYROUSE, G., TABIASCO, J., BRIDONNEAU, C., YAZDANBAKHSH, K., et al. 2019. Faecalibacterium prausnitzii Skews Human DC to Prime IL10-Producing T Cells Through TLR2/6/JNK Signaling and IL-10, IL-27, CD39, and IDO-1 Induction. *Frontiers in Immunology,* 10.

ALBANESE, D., FONTANA, P., DE FILIPPO, C., CAVALIERI, D. & DONATI, C. 2015. MICCA: a complete and accurate software for taxonomic profiling of metagenomic data. *Scientific Reports,* 5**,** 9743.

ALBENBERG, L. G. & WU, G. D. 2014. Diet and the Intestinal Microbiome: Associations, Functions, and Implications for Health and Disease. *Gastroenterology,* 146**,** 1564-1572.

ALDARS-GARCIA, L., CHAPARRO, M. & GISBERT, J. P. 2021. Systematic Review: The Gut Microbiome and Its Potential Clinical Application in Inflammatory Bowel Disease. *Microorganisms,* 9.

ALEXANDER, M., ANG, Q. Y., NAYAK, R. R., BUSTION, A. E., SANDY, M., ZHANG, B., UPADHYAY, V., et al. 2022. Human gut bacterial metabolism drives Th17 activation and colitis. *Cell Host Microbe,* 30**,** 17-30.e9.

ALEXEEV, E. E., LANIS, J. M., KAO, D. J., CAMPBELL, E. L., KELLY, C. J., BATTISTA, K. D., GERICH, M. E., et al. 2018. Microbiota-Derived Indole Metabolites Promote Human and Murine Intestinal Homeostasis through Regulation of Interleukin-10 Receptor. *Am J Pathol,* 188**,** 1183-1194.

ALLAIS, L., KERCKHOF, F.-M., VERSCHUERE, S., BRACKE, K. R., DE SMET, R., LAUKENS, D., VAN DEN ABBEELE, P., et al. 2016. Chronic cigarette smoke exposure induces microbial and inflammatory shifts and mucin changes in the murine gut. *Environmental Microbiology,* 18**,** 1352-1363.

ALLARD, G., RYAN, F. J., JEFFERY, I. B. & CLAESSON, M. J. 2015. SPINGO: a rapid species-classifier for microbial amplicon sequences. *BMC Bioinformatics,* 16**,** 324.

ALTSCHUL, S. F., GISH, W., MILLER, W., MYERS, E. W. & LIPMAN, D. J. 1990. Basic local alignment search tool. *J Mol Biol,* 215**,** 403-10.

AMBARDAR, S., GUPTA, R., TRAKROO, D., LAL, R. & VAKHLU, J. 2016. High Throughput Sequencing: An Overview of Sequencing Chemistry. *Indian J Microbiol,* 56**,** 394-404.

AMBROSE, N. S., JOHNSON, M., BURDON, D. W. & KEIGHLEY, M. R. 1984. Incidence of pathogenic bacteria from mesenteric lymph nodes and ileal serosa during Crohn's disease surgery. *Br J Surg,* 71**,** 623-5.

AMIR, A., MCDONALD, D., NAVAS-MOLINA, J. A., KOPYLOVA, E., MORTON, J. T., ZECH XU, Z., KIGHTLEY, E. P., et al. 2017. Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns. *mSystems,* 2.

AMRE, D. K., D'SOUZA, S., MORGAN, K., SEIDMAN, G., LAMBRETTE, P., GRIMARD, G., ISRAEL, D., et al. 2007. Imbalances in dietary consumption of fatty acids, vegetables, and fruits are associated with risk for Crohn's disease in children. *The American journal of gastroenterology,* 102**,** 2016-2025.

ANANTHAKRISHNAN, A. N. 2015. Epidemiology and risk factors for IBD. *Nature Reviews Gastroenterology & Hepatology,* 12**,** 205-217.

ANANTHAKRISHNAN, A. N., CAGAN, A., GAINER, V. S., CAI, T., CHENG, S. C., SAVOVA, G., CHEN, P., et al. 2013a. Normalization of plasma 25-hydroxy vitamin D is associated with reduced risk of surgery in Crohn's disease. *Inflamm Bowel Dis,* 19**,** 1921-7.

ANANTHAKRISHNAN, A. N., HIGUCHI, L. M., HUANG, E. S., KHALILI, H., RICHTER, J. M., FUCHS, C. S. & CHAN, A. T. 2012a. Aspirin, nonsteroidal anti-inflammatory drug use, and risk for Crohn disease and ulcerative colitis: a cohort study. *Ann Intern Med,* 156**,** 350-9.

ANANTHAKRISHNAN, A. N., KHALILI, H., HIGUCHI, L. M., BAO, Y., KORZENIK, J. R., GIOVANNUCCI, E. L., RICHTER, J. M., et al. 2012b. Higher predicted vitamin D status is associated with reduced risk of Crohn's disease. *Gastroenterology,* 142**,** 482-9.

ANANTHAKRISHNAN, A. N., KHALILI, H., KONIJETI, G. G., HIGUCHI, L. M., DE SILVA, P., FUCHS, C. S., WILLETT, W. C., et al. 2014. Long-term intake of dietary fat and risk of ulcerative colitis and Crohn's disease. *Gut,* 63**,** 776-84.

ANANTHAKRISHNAN, A. N., KHALILI, H., KONIJETI, G. G., HIGUCHI, L. M., DE SILVA, P., KORZENIK, J. R., FUCHS, C. S., et al. 2013b. A prospective study of long-term intake of dietary fiber and risk of Crohn's disease and ulcerative colitis. *Gastroenterology,* 145**,** 970-7.

ANANTHAKRISHNAN, A. N., KHALILI, H., SONG, M., HIGUCHI, L. M., RICHTER, J. M. & CHAN, A. T. 2015. Zinc intake and risk of Crohn's disease and ulcerative colitis: a prospective cohort study. *Int J Epidemiol,* 44**,** 1995-2005.

ANDERSSON, P. & SÖDERHOLM, J. D. Surgery in ulcerative colitis: Indication and timing. Digestive Diseases, 2009. 335-340.

ANDREWS, S. 2010. FastQC: A Quality Control Tool for High Throughput Sequence Data.

ANGLY, F. E., DENNIS, P. G., SKARSHEWSKI, A., VANWONTERGHEM, I., HUGENHOLTZ, P. & TYSON, G. W. 2014. CopyRighter: a rapid tool for improving the accuracy of microbial community profiles through lineage-specific gene copy number correction. *Microbiome,* 2**,** 11.

ANIWAN, S., HARMSEN, W. S., TREMAINE, W. J., KANE, S. V. & LOFTUS, E. V., JR. 2018. Overall and Cause-Specific Mortality of Inflammatory Bowel Disease in Olmsted County, Minnesota, From 1970 Through 2016. *Mayo Clin Proc,* 93**,** 1415-1422.

ANSALDO, E., SLAYDEN, L. C., CHING, K. L., KOCH, M. A., WOLF, N. K., PLICHTA, D. R., BROWN, E. M., et al. 2019. Akkermansia muciniphila induces intestinal adaptive immune responses during homeostasis. *Science,* 364**,** 1179-1184.

ARAB, J. P., KARPEN, S. J., DAWSON, P. A., ARRESE, M. & TRAUNER, M. 2017. Bile acids and nonalcoholic fatty liver disease: Molecular insights and therapeutic perspectives. *Hepatology,* 65**,** 350-362.

ASCHARD, H., LAVILLE, V., TCHETGEN, E. T., KNIGHTS, D., IMHANN, F., SEKSIK, P., ZAITLEN, N., et al. 2019. Genetic effects on the commensal microbiota in inflammatory bowel disease patients. *PLoS Genet,* 15**,** e1008018.

ASHBURNER, M., BALL, C. A., BLAKE, J. A., BOTSTEIN, D., BUTLER, H., CHERRY, J. M., DAVIS, A. P., et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet,* 25**,** 25-9.

ASNICAR, F., BERRY, S. E., VALDES, A. M., NGUYEN, L. H., PICCINNO, G., DREW, D. A., LEEMING, E., et al. 2021. Microbiome connections with host metabolism and habitual diet from 1,098 deeply phenotyped individuals. *Nat Med,* 27**,** 321-332.

ASSHAUER, K. P., WEMHEUER, B., DANIEL, R. & MEINICKE, P. 2015. Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data. *Bioinformatics,* 31**,** 2882-4.

ATTINKARA, R., MWINYI, J., TRUNINGER, K., REGULA, J., GAJ, P., ROGLER, G., KULLAK-UBLICK, G. A., et al. 2012. Association of genetic variation in the NR1H4 gene, encoding the nuclear bile acid receptor FXR, with inflammatory bowel disease. *BMC Res Notes,* 5**,** 461.

BÄCKHED, F., LEY RUTH, E., SONNENBURG JUSTIN, L., PETERSON DANIEL, A. & GORDON JEFFREY, I. 2005. Host-Bacterial Mutualism in the Human Intestine. *Science,* 307**,** 1915-1920.

BAJER, L., WOHL, P. & DRASTICH, P. 2018. PSC-IBD: specific phenotype of inflammatory bowel disease associated with primary sclerosing cholangitis. *Vnitr Lek,* 64**,** 659-664.

BAKDASH, G., VOGELPOEL, L. T., VAN CAPEL, T. M., KAPSENBERG, M. L. & DE JONG, E. C. 2015. Retinoic acid primes human dendritic cells to induce gut-homing, IL-10-producing regulatory T cells. *Mucosal Immunol,* 8**,** 265-78.

BANERJEE, S., TIAN, T., WEI, Z., SHIH, N., FELDMAN, M. D., PECK, K. N., DEMICHELE, A. M., et al. 2018. Distinct Microbial Signatures Associated With Different Breast Cancer Types. *Frontiers in Microbiology,* 9.

BANKEVICH, A., NURK, S., ANTIPOV, D., GUREVICH, A. A., DVORKIN, M., KULIKOV, A. S., LESIN, V. M., et al. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol,* 19**,** 455-77.

BAO, E., JIANG, T., KALOSHIAN, I. & GIRKE, T. 2011. SEED: efficient clustering of next-generation sequences. *Bioinformatics,* 27**,** 2502-9.

BARBALHO, S. M., GOULART, R. D. A. & BATISTA, G. L. D. S. A. 2019. Vitamin A and inflammatory bowel diseases: from cellular studies and animal models to human disease. *Expert Review of Gastroenterology & Hepatology,* 13**,** 25-35.

BARCLAY, A. R., RUSSELL, R. K., WILSON, M. L., GILMOUR, W. H., SATSANGI, J. & WILSON, D. C. 2009. Systematic review: the role of breastfeeding in the development of pediatric inflammatory bowel disease. *J Pediatr,* 155**,** 421-6.

BARKUN, A. N., LOVE, J., GOULD, M., PLUTA, H. & STEINHART, H. 2013. Bile acid malabsorption in chronic diarrhea: pathophysiology and treatment. *Can J Gastroenterol,* 27**,** 653-9.

BARNES, E. L., LOFTUS, E. V. & KAPPELMAN, M. D. 2021. Effects of Race and Ethnicity on Diagnosis and Management of Inflammatory Bowel Diseases. *Gastroenterology,* 160**,** 677-689.

BARNICH, N., CARVALHO, F. A., GLASSER, A. L., DARCHA, C., JANTSCHEFF, P., ALLEZ, M., PEETERS, H., et al. 2007. CEACAM6 acts as a receptor for adherent-invasive E. coli, supporting ileal mucosa colonization in Crohn disease. *J Clin Invest,* 117**,** 1566-74.

BASSET, C., HOLTON, J., BAZEOS, A., VAIRA, D. & BLOOM, S. 2004. Are Helicobacter species and enterotoxigenic Bacteroides fragilis involved in inflammatory bowel disease? *Dig Dis Sci,* 49**,** 1425-32.

BASSO, P. J., FONSECA, M. T., BONFÁ, G., ALVES, V. B. F., SALES-CAMPOS, H., NARDINI, V. & CARDOSO, C. R. D. B. 2014. Association Among Genetic Predisposition, Gut Microbiota, and Host Immune Response in the Etiopathogenesis of Inflammatory Bowel Disease. *Brazilian Journal of Medical and Biological Research*.

BATTAT, R., DUIJVESTEIN, M., VANDE CASTEELE, N., SINGH, S., DULAI, P. S., VALASEK, M. A., MIMMS, L., et al. 2019. Serum Concentrations of 7α-hydroxy-4-cholesten-3-one Are Associated With Bile Acid Diarrhea in Patients With Crohn's Disease. *Clin Gastroenterol Hepatol,* 17**,** 2722-2730.e4.

BATTAT, R., SCHERL, E. J., LUKIN, D., CHARILAOU, P., MAHTANI, P., GERBER, J., GANDARA, J. A., et al. 2023. Increased Primary Bile Acids with Ileocolonic Resection Impact Ileal Inflammation and Gut Microbiota in Inflammatory Bowel Disease. *J Crohns Colitis,* 17**,** 795-803.

BAUMGART, M., DOGAN, B., RISHNIW, M., WEITZMAN, G., BOSWORTH, B., YANTISS, R., ORSI, R. H., et al. 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**,** 403-18.

BEAUD, D., TAILLIEZ, P. & ANBA-MONDOLONI, J. 2005. Genetic characterization of the beta-glucuronidase enzyme from a human intestinal bacterium, Ruminococcus gnavus. *Microbiology (Reading),* 151**,** 2323-2330.

BEGLEY, M., GAHAN, C. G. & HILL, C. 2005. The interaction between bacteria and bile. *FEMS microbiology reviews,* 29**,** 625-651.

BENCHIMOL, E. I., MACK, D. R., NGUYEN, G. C., SNAPPER, S. B., LI, W., MOJAVERIAN, N., QUACH, P., et al. 2014. Incidence, Outcomes, and Health Services Burden of Very Early Onset Inflammatory Bowel Disease. *Gastroenterology,* 147**,** 803-813.e7.

BENJAMINI, Y., DRAI, D., ELMER, G., KAFKAFI, N. & GOLANI, I. 2001. Controlling the false discovery rate in behavior genetics research. *Behavioural brain research,* 125**,** 279-284.

BENOIT, C., OMRY, K., FREDERIC, A. C., RUTH, E. L. & ANDREW, T. G. 2014. AIEC pathobiont instigates chronic colitis in susceptible hosts by altering microbiota composition. *Gut,* 63**,** 1069.

BENSON, A. K., KELLY, S. A., LEGGE, R., MA, F., LOW, S. J., KIM, J., ZHANG, M., et al. 2010. Individuality in gut microbiota composition is a complex polygenic trait shaped by multiple environmental and host genetic factors. *Proc Natl Acad Sci U S A,* 107**,** 18933-8.

BERG, R. D. 1999. Bacterial translocation from the gastrointestinal tract. *Adv Exp Med Biol,* 473**,** 11-30.

BERKOWITZ, L., SCHULTZ, B. M., SALAZAR, G. A., PARDO-ROA, C., SEBASTIÁN, V. P., ÁLVAREZ-LOBOS, M. M. & BUENO, S. M. 2018. Impact of Cigarette Smoking on the Gastrointestinal Tract Inflammation: Opposing Effects in Crohn's Disease and Ulcerative Colitis. *Frontiers in Immunology,* 9.

BERNELL, O., LAPIDUS, A. & HELLERS, G. 2000. Risk Factors for Surgery and Postoperative Recurrence in Crohn's Disease. *Annals of Surgery,* 231.

BERNSTEIN, C. N., ELIAKIM, A., FEDAIL, S., FRIED, M., GEARRY, R., GOH, K. L., HAMID, S., et al. 2016. World Gastroenterology Organisation Global Guidelines Inflammatory Bowel Disease: Update August 2015. *J Clin Gastroenterol,* 50**,** 803-818.

BERNSTEIN, C. N. & FORBES, J. D. 2017. Gut Microbiome in Inflammatory Bowel Disease and Other Chronic Immune-Mediated Inflammatory Diseases. *Inflamm Intest Dis,* 2**,** 116-123.

BERNSTEIN, C. N., FRIED, M., KRABSHUIS, J. H., COHEN, H., ELIAKIM, R., FEDAIL, S., GEARRY, R., et al. 2010. World Gastroenterology Organization Practice Guidelines for the Diagnosis and Management of IBD in 2010. *Inflammatory Bowel Diseases,* 16**,** 112-124.

BERNSTEIN, C. N., RAWSTHORNE, P., CHEANG, M. & BLANCHARD, J. F. 2006. A population-based case control study of potential risk factors for IBD. *Am J Gastroenterol,* 101**,** 993-1002.

BESEMER, J. & BORODOVSKY, M. 2005. GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic Acids Res,* 33**,** W451-4.

BIAN, X., WU, W., YANG, L., LV, L., WANG, Q., LI, Y., YE, J., et al. 2019. Administration of Akkermansia muciniphila Ameliorates Dextran Sulfate Sodium-Induced Ulcerative Colitis in Mice. *Front Microbiol,* 10**,** 2259.

BIBILONI, R., MANGOLD, M., MADSEN, K. L., FEDORAK, R. N. & TANNOCK, G. W. 2006. The bacteriology of biopsies differs between newly diagnosed, untreated, Crohn's disease and ulcerative colitis patients. *J Med Microbiol,* 55**,** 1141-1149.

BIEDERMANN, L., BRÜLISAUER, K., ZEITZ, J., FREI, P., SCHARL, M., VAVRICKA, S. R., FRIED, M., et al. 2014. Smoking Cessation Alters Intestinal Microbiota: Insights from Quantitative Investigations on Human Fecal Samples Using FISH. *Inflammatory Bowel Diseases,* 20**,** 1496-1501.

BITTON, A., VUTCOVICI, M., SEWITCH, M., SUISSA, S. & BRASSARD, P. 2016. Mortality Trends in Crohn's Disease and Ulcerative Colitis: A Population-based Study in Québec, Canada. *Inflamm Bowel Dis,* 22**,** 416-23.

BLACKWELL, J., SAXENA, S., JAYASOORIYA, N., BOTTLE, A., PETERSEN, I., HOTOPF, M., ALEXAKIS, C., et al. 2020. Prevalence and duration of gastrointestinal symptoms before diagnosis of Inflammatory Bowel Disease and predictors of timely specialist review: a population-based study. *J Crohns Colitis*.

BLANDFORD, L. E., JOHNSTON, E. L., SANDERSON, J. D., WADE, W. G. & LAX, A. J. 2019. Promoter orientation of the immunomodulatory Bacteroides fragilis capsular polysaccharide A (PSA) is off in individuals with inflammatory bowel disease (IBD). *Gut Microbes,* 10**,** 569-577.

BOISVERT, S., RAYMOND, F., GODZARIDIS, E., LAVIOLETTE, F. & CORBEIL, J. 2012. Ray Meta: scalable de novo metagenome assembly and profiling. *Genome Biol,* 13**,** R122.

BOLGER, A. M., LOHSE, M. & USADEL, B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics,* 30**,** 2114-20.

BOLYEN, E., RIDEOUT, J. R., DILLON, M. R., BOKULICH, N. A., ABNET, C. C., AL-GHALITH, G. A., ALEXANDER, H., et al. 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat Biotechnol,* 37**,** 852-857.

BONNER, G. F., FAKHRI, A. & VENNAMANENI, S. R. 2004. A long-term cohort study of nonsteroidal anti-inflammatory drug use and disease activity in outpatients with inflammatory bowel disease. *Inflamm Bowel Dis,* 10**,** 751-7.

BORG-BARTOLO, S. P., BOYAPATI, R. K., SATSANGI, J. & KALLA, R. 2020. Precision medicine in inflammatory bowel disease: concept, progress and challenges. *F1000Res,* 9.

BORGHEDE, M. K., SCHLÜTTER, J. M., AGNHOLT, J. S., CHRISTENSEN, L. A., GORMSEN, L. C. & DAHLERUP, J. F. 2011. Bile acid malabsorption investigated by selenium-75-homocholic acid taurine ((75)SeHCAT) scans: causes and treatment responses to cholestyramine in 298 patients with chronic watery diarrhoea. *Eur J Intern Med,* 22**,** e137-40.

BORUP, C., WILDT, S., RUMESSEN, J. J., BOUCHELOUCHE, P. N., GRAFF, J., DAMGAARD, M., MCQUITTY, C., et al. 2017. Chenodeoxycholic acid stimulated fibroblast growth factor 19 response - a potential biochemical test for bile acid diarrhoea. *Aliment Pharmacol Ther,* 45**,** 1433-1442.

BOSCH, A., GKOGKA, E., LE GUYADER, F. S., LOISY-HAMON, F., LEE, A., VAN LIESHOUT, L., MARTHI, B., et al. 2018. Foodborne viruses: Detection, risk assessment, and control options in food processing. *International Journal of Food Microbiology,* 285**,** 110-128.

BOSE, T., HAQUE, M. M., REDDY, C. & MANDE, S. S. 2015. COGNIZER: A Framework for Functional Annotation of Metagenomic Datasets. *PLoS One,* 10**,** e0142102.

BOUDEAU, J., GLASSER, A. L., MASSERET, E., JOLY, B. & DARFEUILLE-MICHAUD, A. 1999. Invasive ability of an Escherichia coli strain isolated from the ileal mucosa of a patient with Crohn's disease. *Infect Immun,* 67**,** 4499-509.

BOURGONJE, A. R., ROO-BRAND, G., LISOTTO, P., SADAGHIAN SADABAD, M., REITSEMA, R. D., DE GOFFAU, M. C., FABER, K. N., et al. 2022. Patients With Inflammatory Bowel Disease Show IgG Immune Responses Towards Specific Intestinal Bacterial Genera. *Front Immunol,* 13**,** 842911.

BOWDEN, R., DAVIES, R. W., HEGER, A., PAGNAMENTA, A. T., DE CESARE, M., OIKKONEN, L. E., PARKES, D., et al. 2019. Sequencing of human genomes with nanopore technology. *Nature Communications,* 10**,** 1869.

BRANT, S. R. 2013. Promises, delivery, and challenges of inflammatory bowel disease risk gene discovery. *Clin Gastroenterol Hepatol,* 11**,** 22-6.

BREIMAN, L. 2001. Random Forests. *Machine Learning,* 45**,** 5-32.

BROMKE, M. A. & KRZYSTEK-KORPACKA, M. 2021. Bile Acid Signaling in Inflammatory Bowel Disease. *Int J Mol Sci,* 22.

BROWN, C., GIBSON, P. R., HART, A., KAPLAN, G. G., KACHROO, S., DING, Q., HAUTAMAKI, E., et al. 2015a. Long-term outcomes of colectomy surgery among patients with ulcerative colitis. *Springerplus,* 4**,** 573.

BROWN, C. C., ESTERHAZY, D., SARDE, A., LONDON, M., PULLABHATLA, V., OSMA-GARCIA, I., AL-BADER, R., et al. 2015b. Retinoic acid is essential for Th1 cell lineage stability and prevents transition to a Th17 cell program. *Immunity,* 42**,** 499-511.

BRYDON, W. G., NYHLIN, H., EASTWOOD, M. A. & MERRICK, M. V. 1996. Serum 7 alpha-hydroxy-4-cholesten-3-one and selenohomocholyltaurine (SeHCAT) whole body retention in the assessment of bile acid induced diarrhoea. *Eur J Gastroenterol Hepatol,* 8**,** 117-23.

BURGE, S., KELLY, E., LONSDALE, D., MUTOWO-MUELLENET, P., MCANULLA, C., MITCHELL, A., SANGRADOR-VEGAS, A., et al. 2012. Manual GO annotation of predictive protein signatures: the InterPro approach to GO curation. *Database (Oxford),* 2012**,** bar068.

BURISCH, J., JESS, T., MARTINATO, M., LAKATOS, P. L. & ON BEHALF OF, E. E. 2013. The burden of inflammatory bowel disease in Europe. *Journal of Crohn's and Colitis,* 7**,** 322-337.

BUTTE, A. J. & NULL, N. 2014. The Integrative Human Microbiome Project: Dynamic Analysis of Microbiome-Host Omics Profiles during Periods of Human Health and Disease. *Cell host & microbe,* 16**,** 276 - 289.

BUTTÓ, L. F. & HALLER, D. 2016. Dysbiosis in intestinal inflammation: Cause or consequence. *Int J Med Microbiol,* 306**,** 302-309.

CABEZAS-CRUZ A, BERMÚDEZ-HUMARÁN LG. Exploring the relationship between Faecalibacterium duncaniae and Escherichia coli in inflammatory bowel disease (IBD): Insights and implications. Comput Struct Biotechnol J. 2023 Nov 19;23:1-9. doi: 10.1016/j.csbj.2023.11.027.

CAI, Y. & SUN, Y. 2011. ESPRIT-Tree: hierarchical clustering analysis of millions of 16S rRNA pyrosequences in quasilinear computational time. *Nucleic Acids Research,* 39**,** e95-e95.

CALKINS, B. M. 1989. A meta-analysis of the role of smoking in inflammatory bowel disease. *Dig Dis Sci,* 34**,** 1841-54.

CALLAHAN, B. J., MCMURDIE, P. J., ROSEN, M. J., HAN, A. W., JOHNSON, A. J. & HOLMES, S. P. 2016. DADA2: High-resolution sample inference from Illumina amplicon data. *Nat Methods,* 13**,** 581-3.

CAMILLERI, M. 2016. Dissecting Molecular Mechanisms in Bile Acid Diarrhea. *Am J Gastroenterol,* 111**,** 433-5.

CAMILLERI, M., CARLSON, P., CHEDID, V., VIJAYVARGIYA, P., BURTON, D. & BUSCIGLIO, I. 2019. Aquaporin Expression in Colonic Mucosal Biopsies From Irritable Bowel Syndrome With Diarrhea. *Clin Transl Gastroenterol,* 10**,** e00019.

CAMILLERI, M., NADEAU, A., TREMAINE, W. J., LAMSAM, J., BURTON, D., ODUNSI, S., SWEETSER, S., et al. 2009. Measurement of serum 7alpha-hydroxy-4-cholesten-3-one (or 7alphaC4), a surrogate test for bile acid malabsorption in health, ileal disease and irritable bowel syndrome using liquid chromatography-tandem mass spectrometry. *Neurogastroenterol Motil,* 21**,** 734-e43.

CANTALUPO, P. G. & PIPAS, J. M. 2019. Detecting viral sequences in NGS data. *Current Opinion in Virology,* 39**,** 41-48.

CANTORNA, M. T. 2006. Vitamin D and its role in immunology: Multiple sclerosis, and inflammatory bowel disease. *Progress in Biophysics and Molecular Biology,* 92**,** 60-64.

CANTORNA, M. T., MUNSICK, C., BEMISS, C. & MAHON, B. D. 2000. 1,25-Dihydroxycholecalciferol prevents and ameliorates symptoms of experimental murine inflammatory bowel disease. *J Nutr,* 130**,** 2648-52.

CAO, Y., SHEN, J. & RAN, Z. H. 2014. Association between Faecalibacterium prausnitzii reduction and inflammatory bowel disease: a meta-analysis and systematic review of the literature. *Gastroenterology research and practice,* 2014.

CARDONA, S., ECK, A., CASSELLAS, M., GALLART, M., ALASTRUE, C., DORE, J., AZPIROZ, F., et al. 2012. Storage conditions of intestinal microbiota matter in metagenomic analysis. *BMC Microbiology,* 12**,** 158.

CARVALHO, D. V., PEREIRA, E. M. & CARDOSO, J. S. 2019. Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics* [Online], 8.

CASPI, R., BILLINGTON, R., KESELER, I. M., KOTHARI, A., KRUMMENACKER, M., MIDFORD, P. E., ONG, W. K., et al. 2020. The MetaCyc database of metabolic pathways and enzymes - a 2019 update. *Nucleic Acids Res,* 48**,** D445-d453.

CASTRO, C. J. & NG, T. F. F. 2017. U50: A New Metric for Measuring Assembly Output Based on Non-Overlapping, Target-Specific Contigs. *Journal of Computational Biology,* 24**,** 1071-1080.

CEKANAVICIUTE, E., YOO, B. B., RUNIA, T. F., DEBELIUS, J. W., SINGH, S., NELSON, C. A., KANNER, R., et al. 2017. Gut bacteria from multiple sclerosis patients modulate human T cells and exacerbate symptoms in mouse models. *Proc Natl Acad Sci U S A,* 114**,** 10713-10718.

CHAN, J. L., WU, S., GEIS, A. L., CHAN, G. V., GOMES, T. A. M., BECK, S. E., WU, X., et al. 2019. Non-toxigenic Bacteroides fragilis (NTBF) administration reduces bacteria-driven chronic colitis and tumor development independent of polysaccharide A. *Mucosal Immunology,* 12**,** 164-177.

CHAN, S. S., LUBEN, R., OLSEN, A., TJONNELAND, A., KAAKS, R., LINDGREN, S., GRIP, O., et al. 2014a. Association between high dietary intake of the n-3 polyunsaturated fatty acid docosahexaenoic acid and reduced risk of Crohn's disease. *Aliment Pharmacol Ther,* 39**,** 834-42.

CHAN, S. S., LUBEN, R., VAN SCHAIK, F., OLDENBURG, B., BUENO-DE-MESQUITA, H. B., HALLMANS, G., KARLING, P., et al. 2014b. Carbohydrate intake in the etiology of Crohn's disease and ulcerative colitis. *Inflamm Bowel Dis,* 20**,** 2013-21.

CHAN, S. S. M. 2019. Editorial: differences in sex-based incidence of inflammatory bowel diseases between Asia-Pacific and Western countries – are these due to differences in environment, access to healthcare, or pathophysiology? *Alimentary Pharmacology & Therapeutics,* 49**,** 1098-1099.

CHANG, P. V., HAO, L., OFFERMANNS, S. & MEDZHITOV, R. 2014. The microbial metabolite butyrate regulates intestinal macrophage function via histone deacetylase inhibition. *Proc Natl Acad Sci U S A,* 111**,** 2247-52.

CHAO, A., GOTELLI, N. J., HSIEH, T. C., SANDER, E. L., MA, K. H., COLWELL, R. K. & ELLISON, A. M. 2014. Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. *Ecological Monographs,* 84**,** 45-67.

CHAPKIN, R. S., DAVIDSON, L. A., LY, L., WEEKS, B. R., LUPTON, J. R. & MCMURRAY, D. N. 2007. Immunomodulatory Effects of (n-3) Fatty Acids: Putative Link to Inflammation and Colon Cancer1,23. *The Journal of Nutrition,* 137**,** 200S-204S.

CHASSAING, B., KOREN, O., CARVALHO, F. A., LEY, R. E. & GEWIRTZ, A. T. 2013. AIEC Pathobiont Instigates Chronic Colitis in Susceptible Hosts by Altering Microbiota Composition. *Gut*.

CHAUDHARY, N., SHARMA, A. K., AGARWAL, P., GUPTA, A. & SHARMA, V. K. 2015. 16S Classifier: A Tool for Fast and Accurate Taxonomic Classification of 16S rRNA Hypervariable Regions in Metagenomic Datasets. *PLOS ONE,* 10**,** e0116106.

CHEIFETZ, A. S., GIANOTTI, R., LUBER, R. & GIBSON, P. R. 2017. Complementary and Alternative Medicines Used by Patients With Inflammatory Bowel Diseases. *Gastroenterology,* 152**,** 415-429.e15.

CHEN, L., YANG, J., YU, J., YAO, Z., SUN, L., SHEN, Y. & JIN, Q. 2005. VFDB: a reference database for bacterial virulence factors. *Nucleic Acids Res,* 33**,** D325-8.

CHEN, P., ZHOU, G., LIN, J., LI, L., ZENG, Z., CHEN, M. & ZHANG, S. 2020. Serum Biomarkers for Inflammatory Bowel Disease. *Front Med (Lausanne),* 7**,** 123.

CHEN, T. & GUESTRIN, C. 2016. XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining***,** 785-794.

CHEN, W., ZHANG, C. K., CHENG, Y., ZHANG, S. & ZHAO, H. 2013. A comparison of methods for clustering 16S rRNA sequences into OTUs. *PLoS One,* 8**,** e70837.

CHIANG, J. Y. 2009. Bile acids: regulation of synthesis: thematic review series: bile acids. *Journal of lipid research,* 50**,** 1955-1966.

CHIANG, J. Y. L. 2004. Regulation of bile acid synthesis: pathways, nuclear receptors, and mechanisms. *Journal of Hepatology,* 40**,** 539-551.

CHOO, J. M., LEONG, L. E. & ROGERS, G. B. 2015. Sample storage conditions significantly influence faecal microbiome profiles. *Sci Rep,* 5**,** 16350.

CHOY, M. C., VISVANATHAN, K. & DE CRUZ, P. 2017. An Overview of the Innate and Adaptive Immune System in Inflammatory Bowel Disease. *Inflammatory Bowel Diseases,* 23**,** 2-13.

CIORBA, M. A. 2013. Indoleamine 2,3 dioxygenase in intestinal disease. *Current Opinion in Gastroenterology,* 29**,** 146-152.

CLAESSON, M. J., JEFFERY, I. B., CONDE, S., POWER, S. E., O'CONNOR, E. M., CUSACK, S., HARRIS, H. M., et al. 2012. Gut microbiota composition correlates with diet and health in the elderly. *Nature,* 488**,** 178-84.

CLARKE, S. F., MURPHY, E. F., O'SULLIVAN, O., LUCEY, A. J., HUMPHREYS, M., HOGAN, A., HAYES, P., et al. 2014. Exercise and associated dietary extremes impact on gut microbial diversity. *Gut,* 63**,** 1913-20.

CLARRIDGE, J. E., 3RD 2004. Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. *Clin Microbiol Rev,* 17**,** 840-62, table of contents.

CLAUS, S. P., TSANG, T. M., WANG, Y., CLOAREC, O., SKORDI, E., MARTIN, F. P., REZZI, S., et al. 2008. Systemic multicompartmental effects of the gut microbiome on mouse metabolic phenotypes. *Mol Syst Biol,* 4**,** 219.

CLEMMENSEN, L. K. H. & KJÆRSGAARD, R. D. 2022. Data Representativity for Machine Learning and AI Systems. *ArXiv,* abs/2203.04706.

CLICK, B., BARNES, E. L., COHEN, B. L., SANDS, B. E., HANSON, J. S., RUBIN, D. T., DUBINSKY, M. C., et al. 2022. Objective disease activity assessment and therapeutic drug monitoring prior to biologic therapy changes in routine inflammatory bowel disease clinical practice: TARGET-IBD. *BMC Gastroenterol,* 22**,** 71.

CLOONEY, A. G., BERNSTEIN, C. N., LESLIE, W. D., VAGIANOS, K., SARGENT, M., LASERNA-MENDIETA, E. J., CLAESSON, M. J., et al. 2016a. A comparison of the gut microbiome between long-term users and non-users of proton pump inhibitors. *Aliment Pharmacol Ther,* 43**,** 974-84.

CLOONEY, A. G., ECKENBERGER, J., LASERNA-MENDIETA, E., SEXTON, K. A., BERNSTEIN, M. T., VAGIANOS, K., SARGENT, M., et al. 2021. Ranking microbiome variance in inflammatory bowel disease: a large longitudinal intercontinental study. *Gut,* 70**,** 499-510.

CLOONEY, A. G., FOUHY, F., SLEATOR, R. D., O' DRISCOLL, A., STANTON, C., COTTER, P. D. & CLAESSON, M. J. 2016b. Comparing Apples and Oranges?: Next Generation Sequencing and Its Impact on Microbiome Analysis. *PLOS ONE,* 11**,** e0148028.

COHEN, L. J., CHO, J. H., GEVERS, D. & CHU, H. 2019. Genetic Factors and the Intestinal Microbiome Guide Development of Microbe-Based Therapies for Inflammatory Bowel Diseases. *Gastroenterology,* 156**,** 2174-2189.

COHEN, R. D. 2001. Efficacy and safety of repeated infliximab infusions for Crohn's disease: 1-year clinical experience. *Inflammatory Bowel Diseases,* 7**,** S17-S22.

COLE, J. R., WANG, Q., FISH, J. A., CHAI, B., MCGARRELL, D. M., SUN, Y., BROWN, C. T., et al. 2014. Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res,* 42**,** D633-42.

COMPEAU, P. E. C., PEVZNER, P. A. & TESLER, G. 2011. How to apply de Bruijn graphs to genome assembly. *Nature biotechnology,* 29**,** 987-991.

CONTE, M. P., LONGHI, C., MARAZZATO, M., CONTE, A. L., ALEANDRI, M., LEPANTO, M. S., ZAGAGLIA, C., et al. 2014. Adherent-invasive Escherichia coli (AIEC) in pediatric Crohn's disease patients: phenotypic and genetic pathogenic features. *BMC Res Notes,* 7**,** 748.

CORNISH, J. A., TAN, E., SIMILLIS, C., CLARK, S. K., TEARE, J. & TEKKIS, P. P. 2008. The Risk of Oral Contraceptives in the Etiology of Inflammatory Bowel Disease: A Meta-Analysis. *Official journal of the American College of Gastroenterology | ACG,* 103.

COSNES, J. 2008. What is the link between the use of tobacco and IBD? *Inflamm Bowel Dis,* 14 Suppl 2**,** S14-5.

COSTEA, P. I., HILDEBRAND, F., ARUMUGAM, M., BÄCKHED, F., BLASER, M. J., BUSHMAN, F. D., DE VOS, W. M., et al. 2018. Enterotypes in the landscape of gut microbial community composition. *Nature Microbiology,* 3**,** 8-16.

COTILLARD, A., KENNEDY, S. P., KONG, L. C., PRIFTI, E., PONS, N., LE CHATELIER, E., ALMEIDA, M., et al. 2013. Dietary intervention impact on gut microbial gene richness. *Nature,* 500**,** 585-8.

COX, K. D., BLACK, M. J., FILIP, N., MILLER, M. R., MOHNS, K., MORTIMOR, J., FREITAS, T. R., et al. 2017. Community assessment techniques and the implications for rarefaction and extrapolation with Hill numbers. *Ecol Evol,* 7**,** 11213-11226.

COYNE, M. J., REINAP, B., LEE, M. M. & COMSTOCK, L. E. 2005. Human symbionts use a host-like pathway for surface fucosylation. *Science,* 307**,** 1778-1781.

CROSS, R. K., WILSON, K. T. & BINION, D. G. 2005. Polypharmacy and Crohn's disease. *Aliment Pharmacol Ther,* 21**,** 1211-6.

CROST, E. H., AJANDOUZ, E. H., VILLARD, C., GERAERT, P. A., PUIGSERVER, A. & FONS, M. 2011. Ruminococcin C, a new anti-Clostridium perfringens bacteriocin produced in the gut by the commensal bacterium Ruminococcus gnavus E1. *Biochimie,* 93**,** 1487-1494.

CROST, E. H., TAILFORD, L. E., MONESTIER, M., SWARBRECK, D., HENRISSAT, B., CROSSMAN, L. C. & JUGE, N. 2016. The mucin-degradation strategy of Ruminococcus gnavus: The importance of intramolecular trans-sialidases. *Gut Microbes,* 7**,** 302-312.

CRYER, B. 2004. COX-2–specific inhibitor or proton pump inhibitor plus traditional NSAID: Is either approach sufficient for patients at highest risk of NSAID-induced ulcers? *Gastroenterology,* 127**,** 1256-1258.

CURRY, K. D., NUTE, M. G. & TREANGEN, T. J. 2021. It takes guts to learn: machine learning techniques for disease detection from the gut microbiome. *Emerg Top Life Sci,* 5**,** 815-827.

CUTOLO, M., CAPELLINO, S. & STRAUB, R. H. 2008. Oestrogens in rheumatic diseases: friend or foe? *Rheumatology (Oxford),* 47 Suppl 3**,** iii2-5.

D'ALDEBERT, E., BIYEYEME BI MVE, M. J., MERGEY, M., WENDUM, D., FIRRINCIELI, D., COILLY, A., FOUASSIER, L., et al. 2009. Bile salts control the antimicrobial peptide cathelicidin through nuclear receptors in the human biliary epithelium. *Gastroenterology,* 136**,** 1435-43.

D'ARGENIO, V., CASABURI, G., PRECONE, V. & SALVATORE, F. 2014. Comparative metagenomic analysis of human gut microbiome composition using two different bioinformatic pipelines. *Biomed Res Int,* 2014**,** 325340.

D'AFFLITTO, M., UPADHYAYA, A., GREEN, A. & PEIRIS, M. 2022. Association Between Sex Hormone Levels and Gut Microbiota Composition and Diversity—A Systematic Review. *Journal of Clinical Gastroenterology,* 56.

D'AMBROSIO, D. N., CLUGSTON, R. D. & BLANER, W. S. 2011. Vitamin A Metabolism: An Update. *Nutrients* [Online], 3.

D'ELIA, D., TRUU, J., LAHTI, L., BERLAND, M., PAPOUTSOGLOU, G., CECI, M., ZOMER, A., et al. 2023. Advancing microbiome research with machine learning: key findings from the ML4Microbiome COST action. *Frontiers in Microbiology,* 14.

DA RIO, L., SPADACCINI, M., PARIGI, T. L., GABBIADINI, R., DAL BUONO, A., BUSACCA, A., MASELLI, R., et al. 2023. Artificial intelligence and inflammatory bowel disease: Where are we going? *World J Gastroenterol,* 29**,** 508-520.

DANESE, S., COLOMBEL, J.-F., LUKAS, M., GISBERT, J. P., D'HAENS, G., HAYEE, B. H., PANACCIONE, R., et al. 2022. Etrolizumab versus infliximab for the treatment of moderately to severely active ulcerative colitis (GARDENIA): a randomised, double-blind, double-dummy, phase 3 study. *The lancet Gastroenterology & hepatology,* 7**,** 118-127.

DARFEUILLE-MICHAUD, A., BOUDEAU, J., BULOIS, P., NEUT, C., GLASSER, A. L., BARNICH, N., BRINGER, M. A., et al. 2004. High prevalence of adherent-invasive Escherichia coli associated with ileal mucosa in Crohn's disease. *Gastroenterology,* 127**,** 412-21.

DAVID, L. A., MAURICE, C. F., CARMODY, R. N., GOOTENBERG, D. B., BUTTON, J. E., WOLFE, B. E., LING, A. V., et al. 2014. Diet rapidly and reproducibly alters the human gut microbiome. *Nature,* 505**,** 559-63.

DAVIS, N. M., PROCTOR, D. M., HOLMES, S. P., RELMAN, D. A. & CALLAHAN, B. J. 2018. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome,* 6**,** 226.

DE FILIPPIS, F., PARENTE, E., ZOTTA, T. & ERCOLINI, D. 2018. A comparison of bioinformatic approaches for 16S rRNA gene profiling of food bacterial microbiota. *Int J Food Microbiol,* 265**,** 9-17.

DE FILIPPIS, F., PELLEGRINI, N., VANNINI, L., JEFFERY, I. B., LA STORIA, A., LAGHI, L., SERRAZANETTI, D. I., et al. 2016. High-level adherence to a Mediterranean diet beneficially impacts the gut microbiota and associated metabolome. *Gut,* 65**,** 1812-1821.

DE LA CUESTA-ZULUAGA, J. & ESCOBAR, J. S. 2016. Considerations For Optimizing Microbiome Analysis Using a Marker Gene. *Front Nutr,* 3**,** 26.

DE LANGE, K. M., MOUTSIANAS, L., LEE, J. C., LAMB, C. A., LUO, Y., KENNEDY, N. A., JOSTINS, L., et al. 2017. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nat Genet,* 49**,** 256-261.

DE PRETER, V., ARIJS, I., WINDEY, K., VANHOVE, W., VERMEIRE, S., SCHUIT, F., RUTGEERTS, P., et al. 2012. Impaired butyrate oxidation in ulcerative colitis is due to decreased butyrate uptake and a defect in the oxidation pathway. *Inflamm Bowel Dis,* 18**,** 1127-36.

DE SOUZA, H. S. & FIOCCHI, C. 2016. Immunopathogenesis of IBD: current state of the art. *Nat Rev Gastroenterol Hepatol,* 13**,** 13-27.

DEGIROLAMO, C., RAINALDI, S., BOVENGA, F., MURZILLI, S. & MOSCHETTA, A. 2014. Microbiota modification with probiotics induces hepatic bile acid synthesis via downregulation of the Fxr-Fgf15 axis in mice. *Cell Rep,* 7**,** 12-8.

DEGLI ESPOSTI, L., DAPERNO, M., DOVIZIO, M., FRANCHI, A., SANGIORGI, D., SAVARINO, E. V., SCALDAFERRI, F., et al. 2023. A retrospective analysis of treatment patterns, drug discontinuation and healthcare costs in Crohn's disease patients treated with biologics. *Digestive and Liver Disease,* 55**,** 1214-1220.

DELEON, M. & STOCCHI, L. 2022. Elective and Emergent Surgery in the Ulcerative Colitis Patient. *Clinics in Colon and Rectal Surgery*.

DELEU, S., MACHIELS, K., RAES, J., VERBEKE, K. & VERMEIRE, S. 2021. Short chain fatty acids and its producing organisms: An overlooked therapy for IBD? *eBioMedicine,* 66.

DENIZOT, J., SIVIGNON, A., BARREAU, F., DARCHA, C., CHAN, H. F., STANNERS, C. P., HOFMAN, P., et al. 2012. Adherent-invasive Escherichia coli induce claudin-2 expression and barrier defect in CEABAC10 mice and Crohn's disease patients. *Inflamm Bowel Dis,* 18**,** 294-304.

DERRIEN, M., ALVAREZ, A.-S. & VOS, W. M. D. 2019. The Gut Microbiota in the First Decade of Life. *Trends in Microbiology*.

DERRIEN, M., VAUGHAN, E. E., PLUGGE, C. M. & DE VOS, W. M. 2004. Akkermansia muciniphila gen. nov., sp. nov., a human intestinal mucin-degrading bacterium. *International Journal of Systematic and Evolutionary Microbiology,* 54**,** 1469-1476.

DESANTIS, T. Z., HUGENHOLTZ, P., LARSEN, N., ROJAS, M., BRODIE, E. L., KELLER, K., HUBER, T., et al. 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol,* 72**,** 5069-72.

DESCHASAUX, M., BOUTER, K. E., PRODAN, A., LEVIN, E., GROEN, A. K., HERREMA, H., TREMAROLI, V., et al. 2018. Depicting the composition of gut microbiota in a population with varied ethnic origins but shared geography. *Nature Medicine,* 24**,** 1526-1531.

DEVKOTA, S. & CHANG, E. B. 2015. Interactions between Diet, Bile Acid Metabolism, Gut Microbiota, and Inflammatory Bowel Diseases. *Dig Dis,* 33**,** 351-6.

DODEN, H. L. & RIDLON, J. M. 2021. Microbial Hydroxysteroid Dehydrogenases: From Alpha to Omega. *Microorganisms,* 9.

DOMINGUEZ-BELLO, M. G., COSTELLO, E. K., CONTRERAS, M., MAGRIS, M., HIDALGO, G., FIERER, N. & KNIGHT, R. 2010. Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proc Natl Acad Sci U S A,* 107**,** 11971-5.

DONIA, M. S. & FISCHBACH, M. A. 2015. HUMAN MICROBIOTA. Small molecules from the human microbiota. *Science,* 349**,** 1254766.

DOUGLAS, G. M., BEIKO, R. G. & LANGILLE, M. G. I. 2018. Predicting the Functional Potential of the Microbiome from Marker Genes Using PICRUSt. *Methods Mol Biol,* 1849**,** 169-177.

DRESCHER, L. S., THIELE, S. & MENSINK, G. B. 2007. A new index to measure healthy food diversity better reflects a healthy diet than traditional measures. *J Nutr,* 137**,** 647-51.

DU TOIT, A. 2022. Remember the phages. *Nature Reviews Microbiology,* 20**,** 254-254.

DU, Z., HUDCOVIC, T., MRAZEK, J., KOZAKOVA, H., SRUTKOVA, D., SCHWARZER, M., TLASKALOVA-HOGENOVA, H., et al. 2015. Development of gut inflammation in mice colonized with mucosa-associated bacteria from patients with ulcerative colitis. *Gut Pathogens,* 7**,** 32.

DUBOC, H., RAJCA, S., RAINTEAU, D., BENAROUS, D., MAUBERT, M. A., QUERVAIN, E., THOMAS, G., et al. 2013. Connecting dysbiosis, bile-acid dysmetabolism and gut inflammation in inflammatory bowel diseases. *Gut,* 62**,** 531-9.

DULAI, P. S., SANDBORN, W. J. & GUPTA, S. 2016a. Colorectal Cancer and Dysplasia in Inflammatory Bowel Disease: A Review of Disease Epidemiology, Pathophysiology, and Management. *Cancer Prev Res (Phila),* 9**,** 887-894.

DULAI, P. S., SINGH, S., CASTEELE, N. V., BOLAND, B. S. & SANDBORN, W. J. 2016b. How Will Evolving Future Therapies and Strategies Change How We Position the Use of Biologics in Moderate to Severely Active Inflammatory Bowel Disease. *Inflamm Bowel Dis,* 22**,** 998-1009.

DUNN, K. A., MOORE-CONNORS, J., MACINTYRE, B., STADNYK, A. W., THOMAS, N. A., NOBLE, A., MAHDI, G., et al. 2016. Early Changes in Microbial Community Structure Are Associated with Sustained Remission After Nutritional Treatment of Pediatric Crohn's Disease. *Inflamm Bowel Dis,* 22**,** 2853-2862.

DUVALLET, C., GIBBONS, S. M., GURRY, T., IRIZARRY, R. A. & ALM, E. J. 2017. Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. *Nature Communications,* 8**,** 1784.

EAVES-PYLES, T., ALLEN, C. A., TAORMINA, J., SWIDSINSKI, A., TUTT, C. B., JEZEK, G. E., ISLAS-ISLAS, M., et al. 2008. Escherichia coli isolated from a Crohn's disease patient adheres, invades, and induces inflammatory responses in polarized intestinal epithelial cells. *Int J Med Microbiol,* 298**,** 397-409.

ECKBURG, P. B., BIK, E. M., BERNSTEIN, C. N., PURDOM, E., DETHLEFSEN, L., SARGENT, M., GILL, S. R., et al. 2005. Diversity of the human intestinal microbial flora. *Science,* 308**,** 1635-8.

ECKENBERGER, J., BUTLER, J. C., BERNSTEIN, C. N., SHANAHAN, F. & CLAESSON, M. J. 2022. Interactions between Medications and the Gut Microbiome in Inflammatory Bowel Disease. *Microorganisms,* 10.

ECKMANN, L. & KARIN, M. 2005. NOD2 and Crohn's disease: loss or gain of function? *Immunity,* 22**,** 661-7.

EDGAR, R. 2018. Taxonomy annotation and guide tree errors in 16S rRNA databases. *PeerJ,* 6**,** e5030.

EDGAR, R. C. 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics,* 26**,** 2460-2461.

EDGAR, R. C. 2013. UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nature Methods,* 10**,** 996.

EDGAR, R. C. 2016. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. *bioRxiv***,** 081257.

EDGAR, R. C., HAAS, B. J., CLEMENTE, J. C., QUINCE, C. & KNIGHT, R. 2011. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics,* 27**,** 2194-2200.

EFFENBERGER, M., REIDER, S., WASCHINA, S., BRONOWSKI, C., ENRICH, B., ADOLPH, T. E., KOCH, R., et al. 2021. Microbial Butyrate Synthesis Indicates Therapeutic Efficacy of Azathioprine in IBD Patients. *Journal of Crohn's and Colitis,* 15**,** 88-98.

ELLIOTT, T. R., HUDSPITH, B. N., WU, G., COOLEY, M., PARKES, G., QUIÑONES, B., RANDALL, L., et al. 2013. Quantification and characterization of mucosa-associated and intracellular Escherichia coli in inflammatory bowel disease. *Inflamm Bowel Dis,* 19**,** 2326-38.

ELVERS, K. T., WILSON, V. J., HAMMOND, A., DUNCAN, L., HUNTLEY, A. L., HAY, A. D. & VAN DER WERF, E. T. 2020. Antibiotic-induced changes in the human gut microbiota for the most commonly prescribed antibiotics in primary care in the UK: a systematic review. *BMJ Open,* 10**,** e035677.

EMA, E. M. A. 1995. *ICH Q2(R2) Validation of analytical procedures - Scientific guideline* [Online]. Available: https://www.ema.europa.eu/en/ich-q2r2-validation-analytical-procedures-scientific-guideline.

ESCOBAR-ZEPEDA, A., VERA-PONCE DE LEÓN, A. & SANCHEZ-FLORES, A. 2015. The Road to Metagenomics: From Microbiology to DNA Sequencing Technologies and Bioinformatics. *Frontiers in Genetics,* 6.

ESTEVINHO, M. M., ROCHA, C., CORREIA, L., LAGO, P., MINISTRO, P., PORTELA, F., TRINDADE, E., et al. 2020. Features of Fecal and Colon Microbiomes Associate With Responses to Biologic Therapies for Inflammatory Bowel Diseases: A Systematic Review. *Clinical Gastroenterology and Hepatology,* 18**,** 1054-1069.

EUSUFZAI, S., AXELSON, M., ANGELIN, B. & EINARSSON, K. 1993. Serum 7 alpha-hydroxy-4-cholesten-3-one concentrations in the evaluation of bile acid malabsorption in patients with diarrhoea: correlation to SeHCAT test. *Gut,* 34**,** 698-701.

EVERARD, A., BELZER, C., GEURTS, L., OUWERKERK, J. P., DRUART, C., BINDELS, L. B., GUIOT, Y., et al. 2013. Cross-talk between Akkermansia muciniphila and intestinal epithelium controls diet-induced obesity. *Proc Natl Acad Sci U S A,* 110**,** 9066-71.

FABISIAK, N., FABISIAK, A., WATALA, C. & FICHNA, J. 2017. Fat-soluble Vitamin Deficiencies and Inflammatory Bowel Disease: Systematic Review and Meta-Analysis. *J Clin Gastroenterol,* 51**,** 878-889.

FALONY, G., JOOSSENS, M., VIEIRA-SILVA, S., WANG, J., DARZI, Y., FAUST, K., KURILSHIKOV, A., et al. 2016. Population-level analysis of gut microbiome variation. *Science,* 352**,** 560-4.

FANG, X., VAZQUEZ-BAEZA, Y., ELIJAH, E., VARGAS, F., ACKERMANN, G., HUMPHREY, G., LAU, R., et al. 2021. Gastrointestinal Surgery for Inflammatory Bowel Disease Persistently Lowers Microbiome and Metabolome Diversity (vol 27, pg 1368, 2021). *Inflammatory Bowel Diseases,* 27**,** 1368-1368.

FELSENSTEIN, J. 1985. Phylogenies and the comparative method. *The American Naturalist,* 125**,** 1-15.
FERNANDES, A. D., MACKLAIM, J. M., LINN, T. G., REID, G. & GLOOR, G. B. 2013. ANOVA-Like Differential Expression (ALDEx) Analysis for Mixed Population RNA-Seq. *PLOS ONE,* 8**,** e67019.
FINN, R. D., BATEMAN, A., CLEMENTS, J., COGGILL, P., EBERHARDT, R. Y., EDDY, S. R., HEGER, A., et al. 2014. Pfam: the protein families database. *Nucleic Acids Res,* 42**,** D222-30.
FLAK, M. B., NEVES, J. F. & BLUMBERG, R. S. 2013. Immunology. Welcome to the microgenderome. *Science,* 339**,** 1044-5.
FLETCHER, J., COOPER, S. C., GHOSH, S. & HEWISON, M. 2019. The Role of Vitamin D in Inflammatory Bowel Disease: Mechanism to Management. *Nutrients,* 11.
FLORES, G. E., CAPORASO, J. G., HENLEY, J. B., RIDEOUT, J. R., DOMOGALA, D., CHASE, J., LEFF, J. W., et al. 2014. Temporal variability is a personalized feature of the human microbiome. *Genome Biology,* 15**,** 531.
FLORES, R., SHI, J., FUHRMAN, B., XU, X., VEENSTRA, T. D., GAIL, M. H., GAJER, P., et al. 2012. Fecal microbial determinants of fecal and systemic estrogens and estrogen metabolites: a cross-sectional study. *Journal of Translational Medicine,* 10**,** 253.
FODITSCH, C., SANTOS, T. M. A., TEIXEIRA, A. G. V., PEREIRA, R. V. V., DIAS, J. M., GAETA, N. & BICALHO, R. C. 2015. Isolation and Characterization of Faecalibacterium prausnitzii from Calves and Piglets. *PLOS ONE,* 9**,** e116465.
FORBES, J. D., VAN DOMSELAAR, G. & BERNSTEIN, C. N. 2016. The Gut Microbiota in Immune-Mediated Inflammatory Diseases. *Frontiers in Microbiology,* 7.
FORSLUND, K., HILDEBRAND, F., NIELSEN, T., FALONY, G., LE CHATELIER, E., SUNAGAWA, S., PRIFTI, E., et al. 2015. Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota. *Nature,* 528**,** 262-266.
FORSLUND S.K., C. R., ZIMMERMANN-KOGADEEVA M., MARKÓ L., ARON-WISNEWSKY J., NIELSEN T., BIRKNER T. 2022. Data Analysis Pipeline for Investigating Drug-Host-Microbiome Relationships in Cardiometabolic Disease. R package version 0.2.8 ed.
FORSLUND, S. K., CHAKAROUN, R., ZIMMERMANN-KOGADEEVA, M., MARKÓ, L., ARON-WISNEWSKY, J., NIELSEN, T. & BIRKNER, T. 2021a. Data analysis pipeline for investigating drug-host-microbiome relationships in cardiometabolic disease (MetaCardis cohort).
FORSLUND, S. K., CHAKAROUN, R., ZIMMERMANN-KOGADEEVA, M., MARKO, L., ARON-WISNEWSKY, J., NIELSEN, T., MOITINHO-SILVA, L., et al. 2021b. Combinatorial, additive and dose-dependent drug-microbiome associations. *Nature,* 600**,** 500-505.
FOURNIER, B. M. & PARKOS, C. A. 2012. The role of neutrophils during intestinal inflammation. *Mucosal Immunol,* 5**,** 354-66.
FRADET, C., KERN, J., ATANASOV, P., WIRTH, D. & BORSI, A. 2020. Impact of surgery and its complications in ulcerative colitis patients in clinical practice: A systematic literature review of real-world evidence in Europe. *International Journal of Surgery Open,* 22**,** 22-32.
FRANZIN, M., STEFANCIC, K., LUCAFO, M., DECORTI, G. & STOCCO, G. 2021. Microbiota and Drug Response in Inflammatory Bowel Disease. *Pathogens,* 10.
FRANZOSA, E. A., MCIVER, L. J. & RAHNAVARD, G. 2018. Species-level functional profiling of metagenomes and metatranscriptomes. 15**,** 962-968.
FRANZOSA, E. A., SIROTA-MADI, A., AVILA-PACHECO, J., FORNELOS, N., HAISER, H. J., REINKER, S., VATANEN, T., et al. 2019. Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nat Microbiol,* 4**,** 293-305.
FROLKIS, A. D., DYKEMAN, J., NEGRÓN, M. E., DEBRUYN, J., JETTE, N., FIEST, K. M., FROLKIS, T., et al. 2013. Risk of Surgery for Inflammatory Bowel Diseases Has Decreased Over Time: A Systematic Review and Meta-analysis of Population-Based Studies. *Gastroenterology,* 145**,** 996-1006.
FROMM, H. & MALAVOLTI, M. 1986. Bile acid-induced diarrhoea. *Clin Gastroenterol,* 15**,** 567-82.
FROST, G., SLEETH, M. L., SAHURI-ARISOYLU, M., LIZARBE, B., CERDAN, S., BRODY, L., ANASTASOVSKA, J., et al. 2014. The short-chain fatty acid acetate reduces appetite via a central homeostatic mechanism. *Nature Communications,* 5**,** 3611.
FUJIMURA, Y., KAMOI, R. & IIDA, M. 1996. Pathogenesis of aphthoid ulcers in Crohn's disease: correlative findings by magnifying colonoscopy, electron microscopy, and immunohistochemistry. *Gut,* 38**,** 724.
FUKUDA, K. & FUJITA, Y. 2014. Determination of the discriminant score of intestinal microbiota as a biomarker of disease activity in patients with ulcerative colitis. *BMC Gastroenterol,* 14**,** 49.

FURUSAWA, Y., OBATA, Y., FUKUDA, S., ENDO, T. A., NAKATO, G., TAKAHASHI, D., NAKANISHI, Y., et al. 2013. Commensal microbe-derived butyrate induces the differentiation of colonic regulatory T cells. *Nature,* 504**,** 446-450.

GALAZZO, G., TEDJO, D. I., WINTJENS, D. S. J., SAVELKOUL, P. H. M., MASCLEE, A. A. M., BODELIER, A. G. L., PIERIK, M. J., et al. 2019. Faecal Microbiota Dynamics and their Relation to Disease Course in Crohn's Disease. *J Crohns Colitis,* 13**,** 1273-1282.

GAO, C.-H., YU, G. & CAI, P. 2021. ggVennDiagram: An Intuitive, Easy-to-Use, and Highly Customizable R Package to Generate Venn Diagram. *Frontiers in Genetics,* 12.

GARCIA, C. J., KOSEK, V., BELTRÁN, D., TOMÁS-BARBERÁN, F. A. & HAJSLOVA, J. 2022. Production of New Microbially Conjugated Bile Acids by Human Gut Microbiota. *Biomolecules,* 12.

GASSLER, N., ROHR, C., SCHNEIDER, A., KARTENBECK, J., BACH, A., OBERMÜLLER, N., OTTO, H. F., et al. 2001. Inflammatory bowel disease is associated with changes of enterocytic junctions. *Am J Physiol Gastrointest Liver Physiol,* 281**,** G216-28.

GEBOES, K., COLOMBEL, J. F., GREENSTEIN, A., JEWELL, D. P., SANDBORN, W. J., VATN, M. H., WARREN, B., et al. 2008. Indeterminate colitis: a review of the concept--what's in a name? *Inflamm Bowel Dis,* 14**,** 850-7.

GENSOLLEN, T., IYER, S. S., KASPER, D. L. & BLUMBERG, R. S. 2016. How colonization by microbiota in early life shapes the immune system. *Science,* 352**,** 539-544.

GÉRARD, P. 2013. Metabolism of cholesterol and bile acids by the gut microbiota. *Pathogens,* 3**,** 14-24.

GEREMIA, A., BIANCHERI, P., ALLAN, P., CORAZZA, G. R. & DI SABATINO, A. 2014. Innate and adaptive immunity in inflammatory bowel disease. *Autoimmun Rev,* 13**,** 3-10.

GEVERS, D., KUGATHASAN, S., DENSON, L. A., VÁZQUEZ-BAEZA, Y., VAN TREUREN, W., REN, B., SCHWAGER, E., et al. 2014. The treatment-naive microbiome in new-onset Crohn's disease. *Cell host & microbe,* 15**,** 382-392.

GHODSI, M., LIU, B. & POP, M. 2011. DNACLUST: accurate and efficient clustering of phylogenetic marker genes. *BMC Bioinformatics,* 12**,** 271.

GHOSH, S. & MITCHELL, R. 2007. Impact of inflammatory bowel disease on quality of life: Results of the European Federation of Crohn's and Ulcerative Colitis Associations (EFCCA) patient survey. *J Crohns Colitis,* 1**,** 10-20.

GHOSH, T. S., DAS, M., JEFFERY, I. B. & O'TOOLE, P. W. 2020. Adjusting for age improves identification of gut microbiome alterations in multiple diseases. *eLife,* 9**,** e50240.

GHOSH, T. S., SHANAHAN, F. & O'TOOLE, P. W. 2022. Toward an improved definition of a healthy microbiome for healthy aging. *Nature Aging,* 2**,** 1054-1069.

GHURYE, J. S., CEPEDA-ESPINOZA, V. & POP, M. 2016. Metagenomic Assembly: Overview, Challenges and Applications. *The Yale journal of biology and medicine,* 89**,** 353-362.

GIBSON, M. K., FORSBERG, K. J. & DANTAS, G. 2015. Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *ISME J,* 9**,** 207-16.

GIONCHETTI, P., RIZZELLO, F., HELWIG, U., VENTURI, A., LAMMERS, K. M., BRIGIDI, P., VITALI, B., et al. 2003. Prophylaxis of pouchitis onset with probiotic therapy: a double-blind, placebo-controlled trial. *Gastroenterology,* 124**,** 1202-9.

GIONGO, A., CRABB, D. B., DAVIS-RICHARDSON, A. G., CHAULIAC, D., MOBBERLEY, J. M., GANO, K. A., MUKHERJEE, N., et al. 2010. PANGEA: pipeline for analysis of next generation amplicons. *The ISME Journal,* 4**,** 852-861.

GKLAVAS, A., DELLAPORTAS, D. & PAPACONSTANTINOU, I. 2017. Risk factors for postoperative recurrence of Crohn's disease with emphasis on surgical predictors. *Ann Gastroenterol,* 30**,** 598-612.

GLASSER, A. L., BOUDEAU, J., BARNICH, N., PERRUCHOT, M. H., COLOMBEL, J. F. & DARFEUILLE-MICHAUD, A. 2001. Adherent invasive Escherichia coli strains from patients with Crohn's disease survive and replicate within macrophages without inducing host cell death. *Infect Immun,* 69**,** 5529-37.

GLEESON, M. H. & DAVIS, A. J. 2003. Non-steroidal anti-inflammatory drugs, aspirin and newly diagnosed colitis: a case-control study. *Aliment Pharmacol Ther,* 17**,** 817-25.

GLOOR, G. B., MACKLAIM, J. M., PAWLOWSKY-GLAHN, V. & EGOZCUE, J. J. 2017. Microbiome Datasets Are Compositional: And This Is Not Optional. *Front Microbiol,* 8**,** 2224.

GODET, P. G., MAY, G. R. & SUTHERLAND, L. R. 1995. Meta-analysis of the role of oral contraceptive agents in inflammatory bowel disease. *Gut,* 37**,** 668-73.

GOLOB, J. L., MARGOLIS, E., HOFFMAN, N. G. & FREDRICKS, D. N. 2017. Evaluating the accuracy of amplicon-based microbiome computational pipelines on simulated human gut microbial communities. *BMC Bioinformatics,* 18**,** 283.

GOMOLLÓN, F., DIGNASS, A., ANNESE, V., TILG, H., VAN ASSCHE, G., LINDSAY, J. O., PEYRIN-BIROULET, L., et al. 2017. 3rd European Evidence-based Consensus on the Diagnosis and Management of Crohn's Disease 2016: Part 1: Diagnosis and Medical Management. *Journal of Crohn's and Colitis,* 11**,** 3-25.

GONZALEZ, C. G., MILLS, R. H., ZHU, Q., SAUCEDA, C., KNIGHT, R., DULAI, P. S. & GONZALEZ, D. J. 2022. Location-specific signatures of Crohn's disease at a multi-omics scale. *Microbiome,* 10**,** 133.

GOODRICH, J. K., DI RIENZI, S. C., POOLE, A. C., KOREN, O., WALTERS, W. A., CAPORASO, J. G., KNIGHT, R., et al. 2014. Conducting a microbiome study. *Cell,* 158**,** 250-262.

GREENER, J. G., KANDATHIL, S. M., MOFFAT, L. & JONES, D. T. 2022. A guide to machine learning for biologists. *Nature Reviews Molecular Cell Biology,* 23**,** 40-55.

GRENINGER, A. L. 2018. A decade of RNA virus metagenomics is (not) enough. *Virus Research,* 244**,** 218-229.

GRIFFITHS, J. A. & MAZMANIAN, S. K. 2018. Emerging evidence linking the gut microbiome to neurologic disorders. *Genome Med,* 10**,** 98.

GUO, X., OKPARA, E. S., HU, W., YAN, C., WANG, Y., LIANG, Q., CHIANG, J. Y. L., et al. 2022. Interactive Relationships between Intestinal Flora and Bile Acids. *International Journal of Molecular Sciences* [Online], 23.

GUPTA, A., DHAKAN, D. B., MAJI, A., SAXENA, R., P, K. V., MAHAJAN, S., PULIKKAN, J., et al. 2019. Association of Flavonifractor plautii, a Flavonoid-Degrading Bacterium, with the Gut Microbiome of Colorectal Cancer Patients in India. *mSystems,* 4.

GÜRBÜZ, M. & AKTAÇ, Ş. 2022. Understanding the role of vitamin A and its precursors in the immune system. *Nutrition Clinique et Métabolisme,* 36**,** 89-98.

GUREVICH, A., SAVELIEV, V., VYAHHI, N. & TESLER, G. 2013. QUAST: quality assessment tool for genome assemblies. *Bioinformatics,* 29**,** 1072-5.

GUZIOR, D. V. & QUINN, R. A. 2021. Review: microbial transformations of human bile acids. *Microbiome,* 9**,** 140.

HAAS, B. J., GEVERS, D., EARL, A. M., FELDGARDEN, M., WARD, D. V., GIANNOUKOS, G., CIULLA, D., et al. 2011. Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Res,* 21**,** 494-504.

HADLEY, W. 2016. *Ggplot2,* New York, NY, Springer Science+Business Media, LLC.

HAEUSLER, R. A., ASTIARRAGA, B., CAMASTRA, S., ACCILI, D. & FERRANNINI, E. 2013. Human insulin resistance is associated with increased plasma levels of 12α-hydroxylated bile acids. *Diabetes,* 62**,** 4184-91.

HAIDER, B., AHN, T. H., BUSHNELL, B., CHAI, J., COPELAND, A. & PAN, C. 2014. Omega: an overlap-graph de novo assembler for metagenomics. *Bioinformatics,* 30**,** 2717-22.

HALFVARSON, J., BRISLAWN, C. J., LAMENDELLA, R., VAZQUEZ-BAEZA, Y., WALTERS, W. A., BRAMER, L. M., D'AMATO, M., et al. 2017. Dynamics of the human gut microbiome in inflammatory bowel disease. *Nat Microbiol,* 2**,** 17004.

HALL, A. B., YASSOUR, M., SAUK, J., GARNER, A., JIANG, X., ARTHUR, T., LAGOUDAS, G. K., et al. 2017. A novel Ruminococcus gnavus clade enriched in inflammatory bowel disease patients. *Genome Medicine,* 9**,** 103.

HAMER, H. M., JONKERS, D., VENEMA, K., VANHOUTVIN, S., TROOST, F. J. & BRUMMER, R. J. 2008. Review article: the role of butyrate on colonic function. *Aliment Pharmacol Ther,* 27**,** 104-19.

HAN, X., DING, S., JIANG, H. & LIU, G. 2021. Roles of Macrophages in the Development and Treatment of Gut Inflammation. *Front Cell Dev Biol,* 9**,** 625423.

HANAUER, S. B., SANDBORN, W. J. & LICHTENSTEIN, G. R. 2019. Evolving Considerations for Thiopurine Therapy for Inflammatory Bowel Diseases-A Clinical Practice Update: Commentary. *Gastroenterology,* 156**,** 36-42.

HAND, D. J. 2006. Classifier Technology and the Illusion of Progress. *Statistical Science,* 21**,** 1-14.

HANDELSMAN, J., RONDON, M. R., BRADY, S. F., CLARDY, J. & GOODMAN, R. M. 1998. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem Biol,* 5**,** R245-9.

HANG, S., PAIK, D., YAO, L., KIM, E., TRINATH, J., LU, J., HA, S., et al. 2019. Bile acid metabolites control TH17 and Treg cell differentiation. *Nature,* 576**,** 143-148.

HANNON, G. J. 2010. FASTX-Toolkit. http://hannonlab.cshl.edu/fastx_toolkit

HANSEN, T. & DUERKSEN, D. R. 2018. Enteral Nutrition in the Management of Pediatric and Adult Crohn's Disease. *Nutrients,* 10.

HARRINGTON, J., FITZGERALD, A. P., LAYTE, R., LUTOMSKI, J., MOLCHO, M. & PERRY, I. J. 2011. Sociodemographic, health and lifestyle predictors of poor diets. *Public Health Nutr,* 14**,** 2166-75.

HARRIS, S. R., FEIL, E. J., HOLDEN, M. T., QUAIL, M. A., NICKERSON, E. K., CHANTRATITA, N., GARDETE, S., et al. 2010. Evolution of MRSA during hospital transmission and intercontinental spread. *Science,* 327**,** 469-74.

HASHIMOTO, T., PERLOT, T., REHMAN, A., TRICHEREAU, J., ISHIGURO, H., PAOLINO, M., SIGL, V., et al. 2012. ACE2 links amino acid malnutrition to microbial ecology and intestinal inflammation. *Nature,* 487**,** 477-481.

HAUER, T. K., J. 2019. CyanoDB.cz 2.0 - On-line database of cyanobacterial genera. Univ. of South Bohemia & Inst. of Botany AS CR: World-wide electronic publication.

HAWWA, A. F., MILLERSHIP, J. S., COLLIER, P. S., VANDENBROECK, K., MCCARTHY, A., DEMPSEY, S., CAIRNS, C., et al. 2008. Pharmacogenomic studies of the anticancer and immunosuppressive thiopurines mercaptopurine and azathioprine. *British Journal of Clinical Pharmacology,* 66**,** 517-528.

HE, Y., CAPORASO, J. G., JIANG, X. T., SHENG, H. F., HUSE, S. M., RIDEOUT, J. R., EDGAR, R. C., et al. 2015. Stability of operational taxonomic units: an important but neglected property for analyzing microbial diversity. *Microbiome,* 3**,** 20.

HE, Y., LIU, S., KLING, D. E., LEONE, S., LAWLOR, N. T., HUANG, Y., FEINBERG, S. B., et al. 2016. The human milk oligosaccharide 2'-fucosyllactose modulates CD14 expression in human enterocytes, thereby attenuating LPS-induced inflammation. *Gut,* 65**,** 33-46.

HE, Y., WU, W., ZHENG, H. M., LI, P., MCDONALD, D., SHENG, H. F., CHEN, M. X., et al. 2018. Regional variation limits applications of healthy gut microbiome reference ranges and disease models. *Nat Med,* 24**,** 1532-1535.

HENKE, M. T., BROWN, E. D., CASSILLY, C. D., VLAMAKIS, H., XAVIER, R. J. & CLARDY, J. 2021. Capsular Polysaccharide Correlates With Immune Response to the Human Gut Microbe *Ruminococcus Gnavus*. *Proceedings of the National Academy of Sciences*.

HENKE, M. T., KENNY, D. J., CASSILLY, C. D., VLAMAKIS, H., XAVIER, R. J. & CLARDY, J. 2019. Ruminococcus gnavus, a member of the human gut microbiome associated with Crohn's disease, produces an inflammatory polysaccharide. *Proc Natl Acad Sci U S A,* 116**,** 12672-12677.

HENSON, M. A. & PHALAK, P. 2017. Microbiota dysbiosis in inflammatory bowel diseases: in silico investigation of the oxygen hypothesis. *BMC Syst Biol,* 11**,** 145.

HERBIG, A., MAIXNER, F., BOS, K. I., ZINK, A., KRAUSE, J. & HUSON, D. H. 2016. MALT: Fast alignment and analysis of metagenomic DNA sequence data applied to the Tyrolean Iceman. *bioRxiv***,** 050559.

HERNÁNDEZ-CEBALLOS, W., CORDOVA-GALLARDO, J. & MENDEZ-SANCHEZ, N. 2021. Gut Microbiota in Metabolic-associated Fatty Liver Disease and in Other Chronic Metabolic Diseases. *J Clin Transl Hepatol,* 9**,** 227-238.

HIBBERD, M. C., WU, M., RODIONOV, D. A., LI, X., CHENG, J., GRIFFIN, N. W., BARRATT, M. J., et al. 2017. The effects of micronutrient deficiencies on bacterial species from the human gut microbiota. *Sci Transl Med,* 9.

HIERGEIST, A., REISCHL, U. & GESSNER, A. 2016. Multicenter quality assessment of 16S ribosomal DNA-sequencing for microbiome analyses reveals high inter-center variability. *Int J Med Microbiol,* 306**,** 334-342.

HIGUCHI, L. M., KHALILI, H., CHAN, A. T., RICHTER, J. M., BOUSVAROS, A. & FUCHS, C. S. 2012. A prospective study of cigarette smoking and the risk of inflammatory bowel disease in women. *Am J Gastroenterol,* 107**,** 1399-406.

HOFMANN, A. F. 1999. The continuing importance of bile acids in liver and intestinal disease. *Archives of Internal Medicine,* 159**,** 2647-2658.

HOOPER, L. V., STAPPENBECK, T. S., HONG, C. V. & GORDON, J. I. 2003. Angiogenins: a new class of microbicidal proteins involved in innate immunity. *Nat Immunol,* 4**,** 269-73.

HOROWITZ, J. E., WARNER, N., STAPLES, J., CROWLEY, E., GOSALIA, N., MURCHIE, R., VAN HOUT, C., et al. 2021. Mutation spectrum of NOD2 reveals recessive inheritance as a main driver of Early Onset Crohn's Disease. *Scientific Reports,* 11.

HOSKINS, L. C., AGUSTINES, M., MCKEE, W. B., BOULDING, E. T., KRIARIS, M. & NIEDERMEYER, G. 1985. Mucin degradation in human colon ecosystems. Isolation and properties of fecal strains that degrade ABH blood group antigens and oligosaccharides from mucin glycoproteins. *J Clin Invest,* 75**,** 944-53.

HOU, J. K., ABRAHAM, B. & EL-SERAG, H. 2011. Dietary intake and risk of developing inflammatory bowel disease: a systematic review of the literature. *Am J Gastroenterol,* 106**,** 563-73.

HOU, K., WU, ZX., CHEN, XY., WANG, JQ., ZHANG, D., XIAO, C., ZHU, D. *et al.* Microbiota in health and diseases. Sig Transduct Target Ther 7, 135 (2022). https://doi.org/10.1038/s41392-022-00974-4

HOU, Y., DONG, L., LU, X., SHI, H., XU, B., ZHONG, W., MA, L., et al. 2022. Distinctions Between Fecal and Intestinal Mucosal Microbiota in Subgroups of Irritable Bowel Syndrome. *Dig Dis Sci,* 67**,** 5580-5592.
HSU, C., GHANNOUM, M., COMINELLI, F. & MARTINO, L. D. 2023. Mycobiome and Inflammatory Bowel Disease: Role in Disease Pathogenesis, Current Approaches and Novel Nutritional-based Therapies. *Inflamm Bowel Dis,* 29**,** 470-479.
HU, X., MADDEN, L. V., EDWARDS, S. & XU, X. 2015. Combining Models is More Likely to Give Better Predictions than Single Models. *Phytopathology®,* 105**,** 1174-1182.
HUANG, E. Y., INOUE, T., LEONE, V. A., DALAL, S., TOUW, K., WANG, Y. W., MUSCH, M. W., et al. 2015. Using Corticosteroids to Reshape the Gut Microbiome: Implications for Inflammatory Bowel Diseases. *Inflammatory Bowel Diseases,* 21**,** 963-972.
HUANG, Y., RAJESH, A., AYYALA, R., SARKAR, A. P., GUO, R., LING, E., NAKASHIDZE, I., et al. 2021. The Systematic Assessment of Completeness of Public Metadata Accompanying Omics Studies.
HUERTA-CEPAS, J., SZKLARCZYK, D., HELLER, D., HERNÁNDEZ-PLAZA, A., FORSLUND, S. K., COOK, H., MENDE, D. R., et al. 2019. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research,* 47**,** D309-D314.
HUGERTH, L. W. & ANDERSSON, A. F. 2017. Analysing Microbial Community Composition through Amplicon Sequencing: From Sampling to Hypothesis Testing. *Front Microbiol,* 8**,** 1561.
HUGHES, L. D., TSUENG, G., DIGIOVANNA, J., HORVATH, T. D., RASMUSSEN, L. V., SAVIDGE, T. C., STOEGER, T., et al. 2023. Addressing barriers in FAIR data practices for biomedical data. *Scientific Data,* 10**,** 98.
HUI, A. J., LEONG, R. W.-L. & CHAN, F. K.-L. 2004. Gastric Ulcer. *In:* JOHNSON, L. R. (ed.) *Encyclopedia of Gastroenterology.* New York: Elsevier.
HUSON, D. H., AUCH, A. F., QI, J. & SCHUSTER, S. C. 2007. MEGAN analysis of metagenomic data. *Genome research,* 17**,** 377-386.
HUSON, D. H., BEIER, S., FLADE, I., GORSKA, A., EL-HADIDI, M., MITRA, S., RUSCHEWEYH, H. J., et al. 2016. MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. *PLoS Comput Biol,* 12**,** e1004957.
HUSSENBUX, A. & DE SILVA, A. 2021. Steroids in inflammatory bowel disease: a clinical review. *Journal of Prescribing Practice,* 3**,** 107-111.
HUTKA, B., LÁZÁR, B., TÓTH, A. S., ÁGG, B., LÁSZLÓ, S. B., MAKRA, N., LIGETI, B., et al. 2021. The Nonsteroidal Anti-Inflammatory Drug Ketorolac Alters the Small Intestinal Microbiota and Bile Acids Without Inducing Intestinal Damage or Delaying Peristalsis in the Rat. *Frontiers in Pharmacology,* 12.
INAGAKI, T., CHOI, M., MOSCHETTA, A., PENG, L., CUMMINS, C. L., MCDONALD, J. G., LUO, G., et al. 2005. Fibroblast growth factor 15 functions as an enterohepatic signal to regulate bile acid homeostasis. *Cell Metabolism,* 2**,** 217-225.
INAGAKI, T., MOSCHETTA, A., LEE, Y. K., PENG, L., ZHAO, G., DOWNES, M., YU, R. T., et al. 2006. Regulation of antibacterial defense in the small intestine by the nuclear bile acid receptor. *Proc Natl Acad Sci U S A,* 103**,** 3920-5.
ISIDRO, R. A. & APPLEYARD, C. B. 2016. Colonic macrophage polarization in homeostasis, inflammation, and cancer. *Am J Physiol Gastrointest Liver Physiol,* 311**,** G59-73.
ITO, S., FUJIMORI, T., FURUYA, A., SATOH, J., NABESHIMA, Y. & NABESHIMA, Y. I. 2005. Impaired negative feedback suppression of bile acid synthesis in mice lacking βKlotho. *Journal of Clinical Investigation,* 115**,** 2202-2208.
IWAI, S., WEINMAIER, T., SCHMIDT, B. L., ALBERTSON, D. G., POLOSO, N. J., DABBAGH, K. & DESANTIS, T. Z. 2016. Piphillin: Improved Prediction of Metagenomic Content by Direct Inference from Human Microbiomes. *PLoS One,* 11**,** e0166104.
JACOBS, J. P., GOUDARZI, M., SINGH, N., TONG, M., MCHARDY, I. H., RUEGGER, P., ASADOURIAN, M., et al. 2016. A Disease-Associated Microbial and Metabolomics State in Relatives of Pediatric Inflammatory Bowel Disease Patients. *Cell Mol Gastroenterol Hepatol,* 2**,** 750-766.
JANOWITZ, H. D., CROEN, E. C. & SACHAR, D. B. 1998. The role of the fecal stream in Crohn's disease: an historical and analytic review. *Inflamm Bowel Dis,* 4**,** 29-39.
JANSSON, J., WILLING, B., LUCIO, M., FEKETE, A., DICKSVED, J., HALFVARSON, J., TYSK, C., et al. 2009. Metabolomics Reveals Metabolic Biomarkers of Crohn's Disease. *PLOS ONE,* 4**,** e6386.
JANTCHOU, P., MOROIS, S., CLAVEL-CHAPELON, F., BOUTRON-RUAULT, M. C. & CARBONNEL, F. 2010. Animal protein intake and risk of inflammatory bowel disease: The E3N prospective study. *Am J Gastroenterol,* 105**,** 2195-201.
JAZIRI, F., PARISOT, N., ABID, A., DENONFOUX, J., RIBIÈRE, C., GASC, C., BOUCHER, D., et al. 2014. PhylOPDb: a 16S rRNA oligonucleotide probe database for prokaryotic identification. *Database,* 2014**,** bau036.

JIANG, Y., LUO, J., HUANG, D., LIU, Y. & LI, D. D. Machine Learning Advances in Microbiology: A Review of Methods and Applications.

JOHNE, B., FAGERHOL, M. K., LYBERG, T., PRYDZ, H., BRANDTZÆG, P., NÆSS-ANDRESEN, C. F. & DALE, I. 1997. Functional and Clinical Aspects of the Myelomonocyte Protein Calprotectin. *Molecular Pathology*.

JOHNSON, A. J., VANGAY, P., AL-GHALITH, G. A., HILLMANN, B. M., WARD, T. L., SHIELDS-CUTLER, R. R., KIM, A. D., et al. 2019. Daily Sampling Reveals Personalized Diet-Microbiome Associations in Humans. *Cell Host Microbe,* 25**,** 789-802 e5.

JOHNSTON, I. M., NOLAN, J. D., PATTNI, S. S., APPLEBY, R. N., ZHANG, J. H., KENNIE, S. L., MADHAN, G. K., et al. 2016. Characterizing Factors Associated With Differences in FGF19 Blood Levels and Synthesis in Patients With Primary Bile Acid Diarrhea. *Am J Gastroenterol,* 111**,** 423-32.

JONES, B. V., BEGLEY, M., HILL, C., GAHAN, C. G. & MARCHESI, J. R. 2008. Functional and comparative metagenomic analysis of bile salt hydrolase activity in the human gut microbiome. *Proc Natl Acad Sci U S A,* 105**,** 13580-5.

JONES, S. 2013. Trends in microbiome research. *Nature Biotechnology,* 31**,** 277-277.

JOOSSENS, M., HUYS, G., CNOCKAERT, M., DE PRETER, V., VERBEKE, K., RUTGEERTS, P., VANDAMME, P., et al. 2011. Dysbiosis of the faecal microbiota in patients with Crohn's disease and their unaffected relatives. *Gut***,** gut. 2010.223263.

JORDAN, M. I. & MITCHELL, T. M. 2015. Machine learning: Trends, perspectives, and prospects. *Science,* 349**,** 255-260.

JOSTINS, L., RIPKE, S., WEERSMA, R. K., DUERR, R. H., MCGOVERN, D. P., HUI, K. Y., LEE, J. C., et al. 2012. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature,* 491**,** 119-24.

JOWETT, S. L., SEAL, C. J., PEARCE, M. S., PHILLIPS, E., GREGORY, W., BARTON, J. R. & WELFARE, M. R. 2004. Influence of dietary factors on the clinical course of ulcerative colitis: a prospective cohort study. *Gut,* 53**,** 1479.

JOYCE, S. A., MACSHARRY, J., CASEY, P. G., KINSELLA, M., MURPHY, E. F., SHANAHAN, F., HILL, C., et al. 2014. Regulation of host weight gain and lipid metabolism by bacterial bile acid modification in the gut. *Proc Natl Acad Sci U S A,* 111**,** 7421-6.

JUILLERAT, P., GRUEBER, M. M., RUETSCH, R., SANTI, G., VUILLÈMOZ, M. & MICHETTI, P. 2022. Positioning biologics in the treatment of IBD: A practical guide – Which mechanism of action for whom? *Current Research in Pharmacology and Drug Discovery,* 3**,** 100104.

JUNKER, V., CONTRINO, S., FLEISCHMANN, W., HERMJAKOB, H., LANG, F., MAGRANE, M., MARTIN, M. J., et al. 2000. The role SWISS-PROT and TrEMBL play in the genome research environment. *J Biotechnol,* 78**,** 221-34.

KAISTHA, A. & LEVINE, J. 2014. Inflammatory bowel disease: the classic gastrointestinal autoimmune disease. *Curr Probl Pediatr Adolesc Health Care,* 44**,** 328-34.

KALMAN, T. D. 2021. Surgery and Stomas in Crohn's Disease.

KAŁUŻNA-CZAPLIŃSKA, J., GĄTAREK, P., CHIRUMBOLO, S., CHARTRAND, M. S. & BJØRKLUND, G. 2019. How important is tryptophan in human health? *Critical Reviews in Food Science and Nutrition,* 59**,** 72-88.

KANEHISA, M., FURUMICHI, M., TANABE, M., SATO, Y. & MORISHIMA, K. 2017. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic acids research,* 45**,** D353-D361.

KANEHISA, M. & GOTO, S. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res,* 28**,** 27-30.

KANEHISA, M., SATO, Y. & MORISHIMA, K. 2016. BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. *Journal of Molecular Biology,* 428**,** 726-731.

KANG, C.-S., BAN, M., CHOI, E.-J., MOON, H.-G., JEON, J.-S., KIM, D.-K., PARK, S.-K., et al. 2013. Extracellular Vesicles Derived from Gut Microbiota, Especially Akkermansia muciniphila, Protect the Progression of Dextran Sulfate Sodium-Induced Colitis. *PLOS ONE,* 8**,** e76520.

KANG, D. D., LI, F., KIRTON, E., THOMAS, A., EGAN, R., AN, H. & WANG, Z. 2019. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ,* 7**,** e7359.

KANG, G., HART, C. A. & SHEARS, P. 2014. 24 - Bacterial Enteropathogens. *In:* FARRAR, J., HOTEZ, P. J., JUNGHANSS, T., KANG, G., LALLOO, D. & WHITE, N. J. (eds.) *Manson's Tropical Infectious Diseases (Twenty-third Edition).* London: W.B. Saunders.

KANG, S., DENMAN, S. E., MORRISON, M., YU, Z., DORE, J., LECLERC, M. & MCSWEENEY, C. S. 2010. Dysbiosis of fecal microbiota in Crohn's disease patients as revealed by a custom phylogenetic microarray. *Inflammatory bowel diseases,* 16**,** 2034-2042.

KAPLAN, G. G. 2015. The global burden of IBD: from 2015 to 2025. *Nature Reviews Gastroenterology & Hepatology,* 12**,** 720-727.

KAPOOR, S. & NARAYANAN, A. 2023. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns***,** 100804.

KARATZOGLOU, A., MEYER, D. & HORNIK, K. 2006. Support Vector Machines in R. *Journal of Statistical Software,* 15**,** 1 - 28.

KARLSSON, F. H., TREMAROLI, V., NOOKAEW, I., BERGSTROM, G., BEHRE, C. J., FAGERBERG, B., NIELSEN, J., et al. 2013. Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature,* 498**,** 99-103.

KARSTENS, L., ASQUITH, M., DAVIN, S., FAIR, D., GREGORY, W. T., WOLFE, A. J., BRAUN, J., et al. 2019. Controlling for Contaminants in Low-Biomass 16S rRNA Gene Sequencing Experiments. *mSystems,* 4.

KASER, A. & BLUMBERG, R. S. 2011. Autophagy, microbial sensing, endoplasmic reticulum stress, and epithelial function in inflammatory bowel disease. *Gastroenterology,* 140**,** 1738-47.

KASHYAP, P. C., CHIA, N., NELSON, H., SEGAL, E. & ELINAV, E. 2017. Microbiome at the Frontier of Personalized Medicine. *Mayo Clin Proc,* 92**,** 1855-1864.

KASSAMBARA, A. 2022. ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.5.0 ed.

KASSAMBARA, A. 2023. rstatix: Pipe-Friendly Framework for Basic Statistical Tests. R package version 0.7.2 ed.

KEEGAN, K. P., GLASS, E. M. & MEYER, F. 2016. MG-RAST, a Metagenomics Service for Analysis of Microbial Community Structure and Function. *Methods Mol Biol,* 1399**,** 207-33.

KELISHADI, R. & FARAJIAN, S. 2014. The protective effects of breastfeeding on chronic non-communicable diseases in adulthood: A review of evidence. *Adv Biomed Res,* 3**,** 3.

KELLEY, D. R., LIU, B., DELCHER, A. L., POP, M. & SALZBERG, S. L. 2012. Gene prediction with Glimmer for metagenomic sequences augmented by classification and clustering. *Nucleic acids research,* 40**,** e9-e9.

KELLEY, D. R. & SALZBERG, S. L. 2010. Clustering metagenomic sequences with interpolated Markov models. *BMC Bioinformatics,* 11**,** 544.

KEMBEL, S. W., COWAN, P. D., HELMUS, M. R., CORNWELL, W. K., MORLON, H., ACKERLY, D. D., BLOMBERG, S. P., et al. 2010. Picante: R tools for integrating phylogenies and ecology. *Bioinformatics,* 26**,** 1463-4.

KENT, T. H., CARDELLI, R. M. & STAMLER, F. W. 1969. Small intestinal ulcers and intestinal flora in rats given indomethacin. *Am J Pathol,* 54**,** 237-49.

KENT, W. J. 2002. BLAT--the BLAST-like alignment tool. *Genome Res,* 12**,** 656-64.

KHALILI, H., HIGUCHI, L. M., ANANTHAKRISHNAN, A. N., MANSON, J. E., FESKANICH, D., RICHTER, J. M., FUCHS, C. S., et al. 2012. Hormone therapy increases risk of ulcerative colitis but not Crohn's disease. *Gastroenterology,* 143**,** 1199-1206.

KHAN, I., ULLAH, N., ZHA, L., BAI, Y., KHAN, A., ZHAO, T., CHE, T., et al. 2019. Alteration of Gut Microbiota in Inflammatory Bowel Disease (IBD): Cause or Consequence? IBD Treatment Targeting the Gut Microbiome. *Pathogens,* 8.

KHOR, B., GARDET, A. & XAVIER, R. J. 2011. Genetics and Pathogenesis of Inflammatory Bowel Disease. *Nature*.

KIM, I., AHN, S. H., INAGAKI, T., CHOI, M., ITO, S., GUO, G. L., KLIEWER, S. A., et al. 2007. Differential regulation of bile acid homeostasis by the farnesoid X receptor in liver and intestine. *J Lipid Res,* 48**,** 2664-72.

KISLYUK, A., BHATNAGAR, S., DUSHOFF, J. & WEITZ, J. S. 2009. Unsupervised statistical clustering of environmental shotgun sequences. *BMC Bioinformatics,* 10**,** 316.

KITTANA, H., GOMES, J. C., HECK, K., JURITSCH, A. F., SUGHROUE, J., XIAN, Y., MANTZ, S., et al. 2023. Evidence for a Causal Role for Escherichia Coli Strains Identified as Adherent-Invasive (AIEC) in Intestinal Inflammation. *Msphere*.

KLAPPENBACH, J. A., DUNBAR, J. M. & SCHMIDT, T. M. 2000. rRNA operon copy number reflects ecological strategies of bacteria. *Appl Environ Microbiol,* 66**,** 1328-33.

KLEMENT, E., COHEN, R. V., BOXMAN, J., JOSEPH, A. & REIF, S. 2004. Breastfeeding and risk of inflammatory bowel disease: a systematic review with meta-analysis. *Am J Clin Nutr,* 80**,** 1342-52.

KLINDWORTH, A., PRUESSE, E., SCHWEER, T., PEPLIES, J., QUAST, C., HORN, M. & GLOCKNER, F. O. 2013. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res,* 41**,** e1.

KLUNEMANN, M., ANDREJEV, S., BLASCHE, S., MATEUS, A., PHAPALE, P., DEVENDRAN, S., VAPPIANI, J., et al. 2021. Bioaccumulation of therapeutic drugs by human gut bacteria. *Nature,* 597**,** 533-538.

KNIGHTS, D., KUCZYNSKI, J., CHARLSON, E. S., ZANEVELD, J., MOZER, M. C., COLLMAN, R. G., BUSHMAN, F. D., et al. 2011. Bayesian community-wide culture-independent microbial source tracking. *Nature Methods,* 8**,** 761.

KNOWLES, S. R., GRAFF, L. A., WILDING, H., HEWITT, C., KEEFER, L. & MIKOCKA-WALUS, A. 2018. Quality of Life in Inflammatory Bowel Disease: A Systematic Review and Meta-analyses-Part I. *Inflamm Bowel Dis,* 24**,** 742-751.

KONOPKA, A. 2009. What is microbial community ecology? *The ISME Journal,* 3**,** 1223-1230.

KOSKINEN, K., PAUSAN, M. R., PERRAS, A. K., BECK, M., BANG, C., MORA, M., SCHILHABEL, A., et al. 2017. First Insights into the Diverse Human Archaeome: Specific Detection of Archaea in the Gastrointestinal Tract, Lung, and Nose and on Skin. *MBio,* 8.

KOSTIC, A. D., XAVIER, R. J. & GEVERS, D. 2014. The microbiome in inflammatory bowel disease: current status and the future ahead. *Gastroenterology,* 146**,** 1489-99.

KOZICH, J. J., WESTCOTT, S. L., BAXTER, N. T., HIGHLANDER, S. K. & SCHLOSS, P. D. 2013. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Appl Environ Microbiol,* 79**,** 5112-20.

KRINOS, C. M., COYNE, M. J., WEINACHT, K. G., TZIANABOS, A. O., KASPER, D. L. & COMSTOCK, L. E. 2001. Extensive surface diversity of a commensal microorganism by multiple DNA inversions. *Nature,* 414**,** 555-558.

KRONMAN, M. P., ZAOUTIS, T. E., HAYNES, K., FENG, R. & COFFIN, S. E. 2012. Antibiotic exposure and IBD development among children: a population-based cohort study. *Pediatrics,* 130**,** e794-803.

KRUIS, W., KALEK, H. D., STELLAARD, F. & PAUMGARTNER, G. 1986. Altered fecal bile acid pattern in patients with inflammatory bowel disease. *Digestion,* 35**,** 189-98.

KUBINSKI, R., DJAMEN-KEPAOU, J. Y., ZHANABAEV, T., HERNANDEZ-GARCIA, A., BAUER, S., HILDEBRAND, F., KORCSMAROS, T., et al. 2022. Benchmark of Data Processing Methods and Machine Learning Models for Gut Microbiome-Based Diagnosis of Inflammatory Bowel Disease.

KUCZYNSKI, J., LAUBER, C. L., WALTERS, W. A., PARFREY, L. W., CLEMENTE, J. C., GEVERS, D. & KNIGHT, R. 2011. Experimental and analytical tools for studying the human microbiome. *Nat Rev Genet,* 13**,** 47-58.

KUHN, M. 2008. Building Predictive Models in R Using the caret Package. *Journal of Statistical Software,* 28**,** 1 - 26.

KUHN, M. & WICKAM, H. 2020. Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles.

KULTIMA, J. R., COELHO, L. P., FORSLUND, K., HUERTA-CEPAS, J., LI, S. S., DRIESSEN, M., VOIGT, A. Y., et al. 2016. MOCAT2: a metagenomic assembly, annotation and profiling framework. *Bioinformatics (Oxford, England),* 32**,** 2520-2523.

KUMAR, S., CARLSEN, T., MEVIK, B. H., ENGER, P., BLAALID, R., SHALCHIAN-TABRIZI, K. & KAUSERUD, H. 2011. CLOTU: an online pipeline for processing and clustering of 454 amplicon reads into OTUs followed by taxonomic annotation. *BMC Bioinformatics,* 12**,** 182.

KUNIN, V., ENGELBREKTSON, A., OCHMAN, H. & HUGENHOLTZ, P. 2010. Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environ Microbiol,* 12**,** 118-23.

KURAKEVICH, E., HENNET, T., HAUSMANN, M., ROGLER, G. & BORSIG, L. 2013. Milk oligosaccharide sialyl(α2,3)lactose activates intestinal CD11c+ cells through TLR4. *Proc Natl Acad Sci U S A,* 110**,** 17444-9.

KURDI, P., KAWANISHI, K., MIZUTANI, K. & YOKOTA, A. 2006. Mechanism of growth inhibition by free bile acids in lactobacilli and bifidobacteria. *J Bacteriol,* 188**,** 1979-86.

LAFOREST-LAPOINTE, I. & ARRIETA, M. C. 2018. Microbial Eukaryotes: a Missing Link in Gut Microbiome Studies. *mSystems,* 3.

LALA, S., OGURA, Y., OSBORNE, C., HOR, S. Y., BROMFIELD, A., DAVIES, S., OGUNBIYI, O., et al. 2003. Crohn's disease and the NOD2 gene: a role for paneth cells. *Gastroenterology,* 125**,** 47-57.

LANGMEAD, B. 2010. Aligning short sequencing reads with Bowtie. *Current protocols in bioinformatics,* Chapter 11**,** Unit-11.7.

LAPIERRE, N., MANGUL, S., ALSER, M., MANDRIC, I., WU, N. C., KOSLICKI, D. & ESKIN, E. 2019. MiCoP: microbial community profiling method for detecting viral and fungal organisms in metagenomic samples. *BMC Genomics,* 20**,** 423.

LASERNA-MENDIETA, E. J., CLOONEY, A. G., CARRETERO-GOMEZ, J. F., MORAN, C., SHEEHAN, D., NOLAN, J. A., HILL, C., et al. 2018. Determinants of Reduced Genetic Capacity for Butyrate Synthesis by the Gut Microbiome in Crohn's Disease and Ulcerative Colitis. *J Crohns Colitis,* 12**,** 204-216.

LAVELLE, A., LENNON, G., O'SULLIVAN, O., DOCHERTY, N., BALFE, A., MAGUIRE, A., MULCAHY, H. E., et al. 2015. Spatial variation of the colonic microbiota in patients with ulcerative colitis and control volunteers. *Gut,* 64**,** 1553-61.

LAVELLE, A. & SOKOL, H. 2020. Gut microbiota-derived metabolites as key actors in inflammatory bowel disease. *Nat Rev Gastroenterol Hepatol,* 17**,** 223-237.

LAVOIE, S., CONWAY, K. L., LASSEN, K. G., JIJON, H. B., PAN, H., CHUN, E., MICHAUD, M., et al. 2019. The Crohn's disease polymorphism, ATG16L1 T300A, alters the gut microbiota and enhances the local Th1/Th17 response. *Elife,* 8.

LE GOALLEC, A., TIERNEY, B. T., LUBER, J. M., COFER, E. M., KOSTIC, A. D. & PATEL, C. J. 2020. A systematic machine learning and data type comparison yields metagenomic predictors of infant age, sex, breastfeeding, antibiotic usage, country of origin, and delivery type. *PLoS Comput Biol,* 16**,** e1007895.

LE, P.-H., CHIU, C.-T., YEH, P.-J., PAN, Y.-B. & CHIU, C.-H. 2022. Clostridium innocuum infection in hospitalised patients with inflammatory bowel disease. *Journal of Infection,* 84**,** 337-342.

LEE, S. H., YUN, Y., KIM, S. J., LEE, E.-J., CHANG, Y., RYU, S., SHIN, H., et al. 2018. Association between Cigarette Smoking Status and Composition of Gut Microbiota: Population-Based Cross-Sectional Study. *Journal of Clinical Medicine* [Online], 7.

LEFEBVRE, P., CARIOU, B., LIEN, F., KUIPERS, F. & STAELS, B. 2009. Role of Bile Acids and Bile Acid Receptors in Metabolic Regulation. *Physiological Reviews,* 89**,** 147-191.

LEINONEN, R., SUGAWARA, H., SHUMWAY, M. & INTERNATIONAL NUCLEOTIDE SEQUENCE DATABASE, C. 2011. The sequence read archive. *Nucleic acids research,* 39**,** D19-D21.

LESAGE, S., ZOUALI, H., CÉZARD, J. P., COLOMBEL, J. F., BELAICHE, J., ALMER, S., TYSK, C., et al. 2002. CARD15/NOD2 mutational analysis and genotype-phenotype correlation in 612 patients with inflammatory bowel disease. *Am J Hum Genet,* 70**,** 845-57.

LETERTRE, M. P. M., MUNJOMA, N., WOLFER, K., PECHLIVANIS, A., MCDONALD, J. A. K., HARDWICK, R. N., CHERRINGTON, N. J., et al. 2020. A Two-Way Interaction between Methotrexate and the Gut Microbiota of Male Sprague-Dawley Rats. *Journal of Proteome Research,* 19**,** 3326-3339.

LEWIS, J. D. & ABREU, M. T. 2017. Diet as a Trigger or Therapy for Inflammatory Bowel Diseases. *Gastroenterology,* 152**,** 398-414 e6.

LI, D., LIU, C. M., LUO, R., SADAKANE, K. & LAM, T. W. 2015. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics,* 31**,** 1674-6.

LI, H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics (Oxford, England),* 27**,** 2987-2993.

LI, H. & DURBIN, R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England),* 25**,** 1754-1760.

LI, J., CHEN, N., WANG, D., ZHANG, J. & GONG, X. 2018. Efficacy of vitamin D in treatment of inflammatory bowel disease: A meta-analysis. *Medicine (Baltimore),* 97**,** e12662.

LI, L., STOECKERT, C. J., JR. & ROOS, D. S. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res,* 13**,** 2178-89.

LI, T. & CHIANG, J. Y. 2014. Bile acid signaling in metabolic disease and drug therapy. *Pharmacol Rev,* 66**,** 948-83.

LI, W. & GODZIK, A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics,* 22**,** 1658-1659.

LIANG, Q., BIBLE, P. W., LIU, Y., ZOU, B. & WEI, L. 2020. DeepMicrobes: taxonomic classification for metagenomics with deep learning. *NAR Genomics and Bioinformatics,* 2**,** lqaa009.

LIMPIAS KAMIYA, K. J. L., HOSOE, N., HAYASHI, Y., KAWAGUCHI, T., TAKABAYASHI, K., OGATA, H. & KANAI, T. 2022. Video capsule endoscopy in inflammatory bowel disease. *DEN Open,* 2**,** e26.

LIN, J. F., CHEN, J. M., ZUO, J. H., YU, A., XIAO, Z. J., DENG, F. H., NIE, B., et al. 2014. Meta-analysis: fecal calprotectin for assessment of inflammatory bowel disease activity. *Inflamm Bowel Dis,* 20**,** 1407-15.

LIN, S. C. & CHEIFETZ, A. S. 2018. The Use of Complementary and Alternative Medicine in Patients With Inflammatory Bowel Disease. *Gastroenterol Hepatol (N Y),* 14**,** 415-425.

LIN, Y.-T., LIN, T.-Y., HUNG, S.-C., LIU, P.-Y., HUNG, W.-C., TSAI, W.-C., TSAI, Y.-C., et al. 2021. Differences in the Microbial Composition of Hemodialysis Patients Treated with and without β-Blockers. *Journal of Personalized Medicine,* 11.

LIN, Y.-Y., HSIEH, C.-H., CHEN, J.-H., LU, X., KAO, J.-H., CHEN, P.-J., CHEN, D.-S., et al. 2017. De novo assembly of highly polymorphic metagenomic data using in situ generated reference sequences and a novel BLAST-based assembly pipeline. *BMC bioinformatics,* 18**,** 223-223.

LINDAHL, B. D., NILSSON, R. H., TEDERSOO, L., ABARENKOV, K., CARLSEN, T., KJOLLER, R., KOLJALG, U., et al. 2013. Fungal community analysis by high-throughput sequencing of amplified markers--a user's guide. *New Phytol,* 199**,** 288-99.

LINDELL, G., FARNEBO, L. O., CHEN, D., NEXØ, E., MADSEN, J. R., BUKHAVE, K. & GRAFFNER, H. 1993. Acute Effects of Smoking during Modified Sham Feeding in Duodenal Ulcer Patients: An Analysis of Nicotine, Acid Secretion, Gastrin, Catecholamines, Epidermal Growth Factor, Prostaglandin E2, and Bile Acids. *Scandinavian Journal of Gastroenterology,* 28**,** 487-494.

LISCHER, H. E. L. & SHIMIZU, K. K. 2017. Reference-guided de novo assembly approach improves genome reconstruction for related species. *BMC bioinformatics,* 18**,** 474-474.

LIU, B. & POP, M. 2009. ARDB--Antibiotic Resistance Genes Database. *Nucleic Acids Res,* 37**,** D443-7.

LIU, B. & POP, M. 2011. MetaPath: identifying differentially abundant metabolic pathways in metagenomic datasets. *BMC Proceedings,* 5**,** S9.

LIU, J. Z., VAN SOMMEREN, S., HUANG, H., NG, S. C., ALBERTS, R., TAKAHASHI, A., RIPKE, S., et al. 2015. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat Genet,* 47**,** 979-986.

LIU, X. Y., TANG, H., ZHOU, Q. Y., ZENG, Y. L., CHEN, D., XU, H., LI, Y., et al. 2023. Advancing the precision management of inflammatory bowel disease in the era of omics approaches and new technology. *World J Gastroenterol,* 29**,** 272-285.

LIU, Z., LOZUPONE, C., HAMADY, M., BUSHMAN, F. D. & KNIGHT, R. 2007. Short pyrosequencing reads suffice for accurate microbial community analysis. *Nucleic Acids Res,* 35**,** e120.

LIU, Z., ZHANG, Y., JIN, T., YI, C., OCANSEY, D. K. W. & MAO, F. 2022. The role of NOD2 in intestinal immune response and microbiota modulation: A therapeutic target in inflammatory bowel disease. *International Immunopharmacology,* 113**,** 109466.

LLOYD-PRICE, J., ARZE, C., ANANTHAKRISHNAN, A. N., SCHIRMER, M., AVILA-PACHECO, J., POON, T. W., ANDREWS, E., et al. 2019. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature,* 569**,** 655-662.

LODDO, I. & ROMANO, C. 2015. Inflammatory Bowel Disease: Genetics, Epigenetics, and Pathogenesis. *Front Immunol,* 6**,** 551.

LODES, M. J., CONG, Y., ELSON, C. O., MOHAMATH, R., LANDERS, C. J., TARGAN, S. R., FORT, M., et al. 2004. Bacterial flagellin is a dominant antigen in Crohn disease. *J Clin Invest,* 113**,** 1296-306.

LOK, K.-H., HUNG, H.-G., NG, C.-H., KWONG, K. C., YIP, W.-M., LAU, S.-F., LI, K.-K., et al. 2008. Epidemiology and clinical characteristics of ulcerative colitis in Chinese population: Experience from a single center in Hong Kong. *Journal of Gastroenterology and Hepatology,* 23**,** 406-410.

LOMAN, N. J., CONSTANTINIDOU, C., CHRISTNER, M., ROHDE, H., CHAN, J. Z., QUICK, J., WEIR, J. C., et al. 2013. A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of Shiga-toxigenic Escherichia coli O104:H4. *Jama,* 309**,** 1502-10.

LONG, M. D., KAPPELMAN, M. D., MARTIN, C. F., CHEN, W., ANTON, K. & SANDLER, R. S. 2016. Role of Nonsteroidal Anti-Inflammatory Drugs in Exacerbations of Inflammatory Bowel Disease. *J Clin Gastroenterol,* 50**,** 152-6.

LOOIJER-VAN LANGEN, M., HOTTE, N., DIELEMAN, L. A., ALBERT, E., MULDER, C. & MADSEN, K. L. 2011. Estrogen receptor-β signaling modulates epithelial barrier function. *Am J Physiol Gastrointest Liver Physiol,* 300**,** G621-6.

LOPEZ-SILES, M., KHAN, T. M., DUNCAN, S. H., HARMSEN, H. J., GARCIA-GIL, L. J. & FLINT, H. J. 2012. Cultured representatives of two major phylogroups of human colonic Faecalibacterium prausnitzii can utilize pectin, uronic acids, and host-derived substrates for growth. *Appl Environ Microbiol,* 78**,** 420-8.

LOPEZ-SILES, M., MARTINEZ-MEDINA, M., BUSQUETS, D., SABAT-MIR, M., DUNCAN, S. H., FLINT, H. J., ALDEGUER, X., et al. 2014. Mucosa-associated Faecalibacterium prausnitzii and Escherichia coli co-abundance can distinguish Irritable Bowel Syndrome and Inflammatory Bowel Disease phenotypes. *Int J Med Microbiol,* 304**,** 464-75.

LOPEZ-SILES, M., MARTINEZ-MEDINA, M., SURÍS-VALLS, R., ALDEGUER, X., SABAT-MIR, M., DUNCAN, S. H., FLINT, H. J., et al. 2016. Changes in the Abundance of Faecalibacterium prausnitzii Phylogroups I and II in the Intestinal Mucosa of Inflammatory Bowel Disease and Patients with Colorectal Cancer. *Inflamm Bowel Dis,* 22**,** 28-41.

LORENZO-ZÚÑIGA, V., BARTOLÍ, R., PLANAS, R., HOFMANN, A. F., VIÑADO, B., HAGEY, L. R., HERNÁNDEZ, J. M., et al. 2003. Oral bile acids reduce bacterial overgrowth, bacterial translocation, and endotoxemia in cirrhotic rats. *Hepatology,* 37**,** 551-7.

LOSURDO, G., IANNONE, A., CONTALDO, A., IERARDI, E., DI LEO, A. & PRINCIPI, M. 2015. Escherichia coli Nissle 1917 in Ulcerative Colitis Treatment: Systematic Review and Meta-analysis. *J Gastrointestin Liver Dis,* 24**,** 499-505.

LOVE, M. I., HUBER, W. & ANDERS, S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology,* 15**,** 550-550.

LOZUPONE, C. A., STOMBAUGH, J. I., GORDON, J. I., JANSSON, J. K. & KNIGHT, R. 2012. Diversity, stability and resilience of the human gut microbiota. *Nature,* 489**,** 220-30.

LU, S. C., SWISHER, C. L., CHUNG, C., JAFFRAY, D. & SIDEY-GIBBONS, C. 2023. On the importance of interpretable machine learning predictions to inform clinical decision making in oncology. *Front Oncol,* 13**,** 1129380.

LUCAS, L. N., BARRETT, K., KERBY, R. L., ZHANG, Q., CATTANEO, L. E., STEVENSON, D., REY, F. E., et al. 2021. Dominant Bacterial Phyla from the Human Gut Show Widespread Ability To Transform and Conjugate Bile Acids. *mSystems,* 6**,** 10.1128/msystems.00805-21.

LUKASHIN, A. V. & BORODOVSKY, M. 1998. GeneMark.hmm: new solutions for gene finding. *Nucleic acids research,* 26**,** 1107-1115.

LUO, R., LIU, B., XIE, Y., LI, Z., HUANG, W., YUAN, J., HE, G., et al. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience,* 1**,** 18-18.

LYUTAKOV, I., NAKOV, R., VALKOV, H., VATCHEVA-DOBREVSKA, R., VLADIMIROV, B. & PENCHEV, P. 2021. Serum Levels of Fibroblast Growth Factor 19 Correlate with the Severity of Diarrhea and Independently from Intestinal Inflammation in Patients with Inflammatory Bowel Disease or Microscopic Colitis. *Turk J Gastroenterol,* 32**,** 374-381.

M'KOMA, A. E. 2022. Inflammatory Bowel Disease: Clinical Diagnosis and Surgical Treatment-Overview. *Medicina* [Online], 58.

MA, C., BATTAT, R., PARKER, C. E., KHANNA, R., JAIRATH, V. & FEAGAN, B. G. 2019. Update on C-reactive protein and fecal calprotectin: are they accurate measures of disease activity in Crohn's disease? *Expert Review of Gastroenterology & Hepatology,* 13**,** 319-330.

MA, H. & PATTI, M. E. 2014. Bile acids, obesity, and the metabolic syndrome. *Best Pract Res Clin Gastroenterol,* 28**,** 573-83.

MACH, N. & CLARK, A. 2017. Micronutrient Deficiencies and the Human Gut Microbiota. *Trends in Microbiology,* 25**,** 607-610.

MACHIELS, K., JOOSSENS, M., SABINO, J., DE PRETER, V., ARIJS, I., EECKHAUT, V., BALLET, V., et al. 2014. A decrease of the butyrate-producing species Roseburia hominis and Faecalibacterium prausnitzii defines dysbiosis in patients with ulcerative colitis. *Gut,* 63**,** 1275-83.

MACHIELS, K., SABINO, J., VANDERMOSTEN, L., JOOSSENS, M., ARIJS, I., DE BRUYN, M., EECKHAUT, V., et al. 2017. Specific members of the predominant gut microbiota predict pouchitis following colectomy and IPAA in UC. *Gut,* 66**,** 79-88.

MACIA, L., TAN, J., VIEIRA, A. T., LEACH, K., STANLEY, D., LUONG, S., MARUYA, M., et al. 2015. Metabolite-sensing receptors GPR43 and GPR109A facilitate dietary fibre-induced gut homeostasis through regulation of the inflammasome. *Nature Communications,* 6**,** 6734.

MAGNUSSON, M. K., STRID, H., SAPNARA, M., LASSON, A., BAJOR, A., UNG, K. A. & OHMAN, L. 2016. Anti-TNF Therapy Response in Patients with Ulcerative Colitis Is Associated with Colonic Antimicrobial Peptide Expression and Microbiota Composition. *Journal of Crohns & Colitis,* 10**,** 943-952.

MAHID, S. S., MINOR, K. S., SOTO, R. E., HORNUNG, C. A. & GALANDIUK, S. 2006. Smoking and inflammatory bowel disease: a meta-analysis. *Mayo Clin Proc,* 81**,** 1462-71.

MAIDAK, B. L., COLE, J. R., LILBURN, T. G., PARKER, C. T., JR., SAXMAN, P. R., FARRIS, R. J., GARRITY, G. M., et al. 2001. The RDP-II (Ribosomal Database Project). *Nucleic Acids Res,* 29**,** 173-4.

MAIER, L., PRUTEANU, M., KUHN, M., ZELLER, G., TELZEROW, A., ANDERSON, E. E., BROCHADO, A. R., et al. 2018. Extensive impact of non-antibiotic drugs on human gut bacteria. *Nature,* 555**,** 623-628.

MAIER, T. V., LUCIO, M., LEE, L. H., VERBERKMOES, N. C., BRISLAWN, C. J., BERNHARDT, J., LAMENDELLA, R., et al. 2017. Impact of Dietary Resistant Starch on the Human Gut Microbiome, Metaproteome, and Metabolome. *mBio,* 8.

MANTERE, T., KERSTEN, S. & HOISCHEN, A. 2019. Long-Read Sequencing Emerging in Medical Genetics. *Frontiers in Genetics,* 10.

MARASCO, G., CREMON, C., BARBARO, M. R., FALANGONE, F., MONTANARI, D., CAPUANI, F., MASTEL, G., et al. 2022. Pathophysiology and Clinical Management of Bile Acid Diarrhea. *J Clin Med,* 11.

MARCHESI, J. R., ADAMS, D. H., FAVA, F., HERMES, G. D., HIRSCHFIELD, G. M., HOLD, G., QURAISHI, M. N., et al. 2016. The gut microbiota and host health: a new clinical frontier. *Gut,* 65**,** 330-9.

MARCILLE, F., GOMEZ, A., JOUBERT, P., LADIRÉ, M., VEAU, G., CLARA, A., GAVINI, F., et al. 2002. Distribution of genes encoding the trypsin-dependent lantibiotic ruminococcin A among bacteria isolated from human fecal microbiota. *Applied and Environmental Microbiology,* 68**,** 3424-3431.

MARCÓ, M. B., MOINEAU, S. & QUIBERONI, A. 2012. Bacteriophages and dairy fermentations. *Bacteriophage,* 2**,** 149-158.

MARCOS-ZAMBRANO, L. J., KARADUZOVIC-HADZIABDIC, K., LONCAR TURUKALO, T., PRZYMUS, P., TRAJKOVIK, V., AASMETS, O., BERLAND, M., et al. 2021. Applications of Machine Learning in Human Microbiome Studies: A Review on Feature Selection, Biomarker Identification, Disease Prediction and Treatment. *Front Microbiol,* 12**,** 634511.

MARIAT, D., FIRMESSE, O., LEVENEZ, F., GUIMARĂES, V. D., SOKOL, H., DORÉ, J., CORTHIER, G., et al. 2009. The Firmicutes/Bacteroidetes ratio of the human microbiota changes with age. *BMC Microbiology,* 9**,** 123.

MARTIN, M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal,* 17**,** pp. 10-12.

MARTINEZ-MEDINA, M., ALDEGUER, X., LOPEZ-SILES, M., GONZÁLEZ-HUIX, F., LÓPEZ-OLIU, C., DAHBI, G., BLANCO, J. E., et al. 2009a. 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,* 15**,** 872-82.

MARTINEZ-MEDINA, M. & GARCIA-GIL, L. J. 2014. Escherichia coli in chronic inflammatory bowel diseases: An update on adherent invasive Escherichia coli pathogenicity. *World J Gastrointest Pathophysiol,* 5**,** 213-27.

MARTINEZ-MEDINA, M., NAVES, P., BLANCO, J., ALDEGUER, X., BLANCO, J. E., BLANCO, M., PONTE, C., et al. 2009b. Biofilm formation as a novel phenotypic feature of adherent-invasive Escherichia coli (AIEC). *BMC Microbiol,* 9**,** 202.

MARTÍNEZ, I., LATTIMER, J. M., HUBACH, K. L., CASE, J. A., YANG, J., WEBER, C. G., LOUK, J. A., et al. 2013. Gut microbiome composition is linked to whole grain-induced immunological improvements. *Isme j,* 7**,** 269-80.

MARZOLINI, C., TIRONA, R. G., GERVASINI, G., POONKUZHALI, B., ASSEM, M., LEE, W., LEAKE, B. F., et al. 2007. A common polymorphism in the bile acid receptor farnesoid X receptor is associated with decreased hepatic target gene expression. *Mol Endocrinol,* 21**,** 1769-80.

MASNADI SHIRAZI, K., NIKNIAZ, Z., MASNADI SHIRAZI, A. & ROHANI, M. 2018. Vitamin A supplementation decreases disease activity index in patients with ulcerative colitis: A randomized controlled clinical trial. *Complementary Therapies in Medicine,* 41**,** 215-219.

MATIJAŠIĆ, M., MEŠTROVIĆ, T., PALJETAK, H., PERIĆ, M., BAREŠIĆ, A. & VERBANAC, D. 2020. Gut Microbiota beyond Bacteria-Mycobiome, Virome, Archaeome, and Eukaryotic Parasites in IBD. *Int J Mol Sci,* 21.

MATSUOKA, K. & KANAI, T. 2015. The gut microbiota and inflammatory bowel disease. *Semin Immunopathol,* 37**,** 47-55.

MATSUOKA, K., KOBAYASHI, T., UENO, F., MATSUI, T., HIRAI, F., INOUE, N., KATO, J., et al. 2018. Evidence-based clinical practice guidelines for inflammatory bowel disease. *J Gastroenterol,* 53**,** 305-353.

MATTAR, M. C., LOUGH, D., PISHVAIAN, M. J. & CHARABATY, A. 2011. Current management of inflammatory bowel disease and colorectal cancer. *Gastrointest Cancer Res,* 4**,** 53-61.

MAX KUHN, S. J., JORGE CIMENTADA 2022. corrr: Correlations in R. R package version 0.4.4 ed.

MAZMANIAN, S. K., LIU, C. H., TZIANABOS, A. O. & KASPER, D. L. 2005. An immunomodulatory molecule of symbiotic bacteria directs maturation of the host immune system. *Cell,* 122**,** 107-18.

MAZMANIAN, S. K., ROUND, J. L. & KASPER, D. L. 2008. A microbial symbiosis factor prevents intestinal inflammatory disease. *Nature,* 453**,** 620-625.
MCARTHUR, A. G., WAGLECHNER, N., NIZAM, F., YAN, A., AZAD, M. A., BAYLAY, A. J., BHULLAR, K., et al. 2013. The comprehensive antibiotic resistance database. *Antimicrobial agents and chemotherapy,* 57**,** 3348-3357.
MCGUCKIN, M. A., LINDEN, S. K., SUTTON, P. & FLORIN, T. H. 2011. Mucin dynamics and enteric pathogens. *Nat Rev Microbiol,* 9**,** 265-78.
MCMURDIE, P. J. & HOLMES, S. 2013. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLOS ONE,* 8**,** e61217.
MCPHEE, J. B., SMALL, C. L., REID-YU, S. A., BRANNON, J. R., LE MOUAL, H. & COOMBES, B. K. 2014. Host defense peptide resistance contributes to colonization and maximal intestinal pathology by Crohn's disease-associated adherent-invasive Escherichia coli. *Infection and immunity,* 82**,** 3383-3393.
MELGAR, S. & SHANAHAN, F. 2010. Inflammatory bowel disease-from mechanisms to treatment strategies. *Autoimmunity,* 43**,** 463-77.
MENZEL, P., NG, K. L. & KROGH, A. 2016. Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nature Communications,* 7**,** 11257.
MESONERO, F., FERNÁNDEZ, C., SÁNCHEZ-RODRÍGUEZ, E., GARCÍA-GARCÍA PAREDES, A., SENOSIAIN, C., ALBILLOS, A. & LÓPEZ-SANROMÁN, A. 2022. Polypharmacy in Patients With Inflammatory Bowel Disease: Prevalence and Outcomes in a Single-center Series. *J Clin Gastroenterol,* 56**,** e189-e195.
METWALY, A., DUNKEL, A., WALDSCHMITT, N., RAJ, A. C. D., LAGKOUVARDOS, I., CORRALIZA, A. M., MAYORGAS, A., et al. 2020. Integrated microbiota and metabolite profiles link Crohn's disease to sulfur metabolism. *Nat Commun,* 11**,** 4322.
MEYER, F., PAARMANN, D., D'SOUZA, M., OLSON, R., GLASS, E. M., KUBAL, M., PACZIAN, T., et al. 2008. The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics,* 9**,** 386.
MICHIELAN, A. & D'INCÀ, R. 2015. Intestinal Permeability in Inflammatory Bowel Disease: Pathogenesis, Clinical Evaluation, and Therapy of Leaky Gut. *Mediators Inflamm,* 2015**,** 628157.
MIKHEENKO, A., SAVELIEV, V. & GUREVICH, A. 2016. MetaQUAST: evaluation of metagenome assemblies. *Bioinformatics,* 32**,** 1088-90.
MILAJERDI, A., EBRAHIMI-DARYANI, N., DIELEMAN, L. A., LARIJANI, B. & ESMAILLZADEH, A. 2021. Association of Dietary Fiber, Fruit, and Vegetable Consumption with Risk of Inflammatory Bowel Disease: A Systematic Review and Meta-Analysis. *Adv Nutr,* 12**,** 735-743.
MIRKOV, M. U., VERSTOCKT, B. & CLEYNEN, I. 2017. Genetics of inflammatory bowel disease: beyond NOD2. *Lancet Gastroenterol Hepatol,* 2**,** 224-234.
MIRSEPASI-LAURIDSEN, H. C., VALLANCE, B. A., KROGFELT, K. A. & PETERSEN, A. M. 2019. Escherichia coli Pathobionts Associated with Inflammatory Bowel Disease. *Clin Microbiol Rev,* 32.
MITCHELL, M., WU, S., ZALDIVAR, A., BARNES, P., VASSERMAN, L., HUTCHINSON, B., SPITZER, E., et al. 2019. Model Cards for Model Reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency.* Atlanta, GA, USA: Association for Computing Machinery.
MODI, S. K., COLLINS, J. J. & RELMAN, D. A. 2014. Antibiotics and the Gut Microbiota. *Journal of Clinical Investigation*.
MOITINHO-SILVA, L., FORSLUND, S. K., CHAKAROUN, R., ZIMMERMANN-KOGADEEVA, M., MARKÓ, L., ARON-WISNEWSKY, J., NIELSEN, T., et al. 2021. grp-bork/vpthemall: Release of vpthemall - assistant functions for variation partition with dbRDA method. Zenodo.
MOLLER, F. T., ANDERSEN, V., WOHLFAHRT, J. & JESS, T. 2015. Familial Risk of Inflammatory Bowel Disease: A Population-Based Cohort Study 1977–2011. *Official journal of the American College of Gastroenterology | ACG,* 110.
MOLNAR, C., CASALICCHIO, G. & BISCHL, B. 2018. iml: An R package for Interpretable Machine Learning. *Journal of Open Source Software,* 3(26).
MOLNAR, C., KÖNIG, G., HERBINGER, J., FREIESLEBEN, T., DANDL, S., SCHOLBECK, C. A., CASALICCHIO, G., et al. 2022. General Pitfalls of Model-Agnostic Interpretation Methods for Machine Learning Models. *In:* HOLZINGER, A., GOEBEL, R., FONG, R., MOON, T., MÜLLER, K.-R. & SAMEK, W. (eds.) *xxAI - Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers.* Cham: Springer International Publishing.
MOLODECKY, N. A., SOON, I. S., RABI, D. M., GHALI, W. A., FERRIS, M., CHERNOFF, G., BENCHIMOL, E. I., et al. 2012. Increasing Incidence and Prevalence of the Inflammatory Bowel Diseases With Time, Based on Systematic Review. *Gastroenterology,* 142**,** 46-54.e42.

MONTELEONE, I., RIZZO, A., SARRA, M., SICA, G., SILERI, P., BIANCONE, L., MACDONALD, T. T., et al. 2011. Aryl hydrocarbon receptor-induced signals up-regulate IL-22 production and inhibit inflammation in the gastrointestinal tract. *Gastroenterology,* 141**,** 237-48, 248.e1.

MONZOORUL HAQUE, M., GHOSH, T. S., KOMANDURI, D. & MANDE, S. S. 2009. SOrt-ITEMS: Sequence orthology based approach for improved taxonomic estimation of metagenomic sequences. *Bioinformatics,* 25**,** 1722-1730.

MOON, C. D., YOUNG, W., MACLEAN, P. H., COOKSON, A. L. & BERMINGHAM, E. N. 2018. Metagenomic insights into the roles of Proteobacteria in the gastrointestinal microbiomes of healthy dogs and cats. *Microbiologyopen,* 7**,** e00677.

MOOSSAVI, S., MILIKU, K., SEPEHRI, S., KHAFIPOUR, E. & AZAD, M. B. 2018. The Prebiotic and Probiotic Properties of Human Milk: Implications for Infant Immune Development and Pediatric Asthma. *Front Pediatr,* 6**,** 197.

MORAN, C. P., NEARY, B. & DOHERTY, G. A. 2016. Endoscopic evaluation in diagnosis and management of inflammatory bowel disease. *World J Gastrointest Endosc,* 8**,** 723-732.

MORENO-INDIAS, I., LAHTI, L., NEDYALKOVA, M., ELBERE, I., ROSHCHUPKIN, G., ADILOVIC, M., AYDEMIR, O., et al. 2021. Statistical and Machine Learning Techniques in Human Microbiome Studies: Contemporary Challenges and Solutions. *Frontiers in Microbiology,* 12.

MORGAN, X. C. & HUTTENHOWER, C. 2012a. Chapter 12: Human microbiome analysis. *PLoS Comput Biol,* 8**,** e1002808.

MORGAN, X. C., TICKLE, T. L., SOKOL, H., GEVERS, D., DEVANEY, K. L., WARD, D. V., REYES, J. A., et al. 2012b. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biology,* 13**,** R79.

MOWAT, C., DIGBY, J., STRACHAN, J. A., WILSON, R., CAREY, F. A., FRASER, C. G. & STEELE, R. J. 2016. Faecal haemoglobin and faecal calprotectin as indicators of bowel disease in patients presenting to primary care with bowel symptoms. *Gut,* 65**,** 1463-9.

MUELLER, M., THORELL, A., CLAUDEL, T., JHA, P., KOEFELER, H., LACKNER, C., HOESEL, B., et al. 2015. Ursodeoxycholic acid exerts farnesoid X receptor-antagonistic effects on bile acid and lipid metabolism in morbid obesity. *J Hepatol,* 62**,** 1398-404.

MUIR, P., LI, S., LOU, S., WANG, D., SPAKOWICZ, D. J., SALICHOS, L., ZHANG, J., et al. 2016. The real cost of sequencing: scaling computation to keep pace with data generation. *Genome biology,* 17**,** 53-53.

MULLISH, B. H., MCDONALD, J. A. K., PECHLIVANIS, A., ALLEGRETTI, J. R., KAO, D., BARKER, G. F., KAPILA, D., et al. 2019. Microbial bile salt hydrolases mediate the efficacy of faecal microbiota transplant in the treatment of recurrent Clostridioides difficile infection. *Gut,* 68**,** 1791-1800.

MUNOZ-BELLIDO, J. L., MUNOZ-CRIADO, S. & GARCìA-RODRìGUEZ, J. A. 2000. Antimicrobial activity of psychotropic drugs: Selective serotonin reuptake inhibitors. *International Journal of Antimicrobial Agents,* 14**,** 177-180.

MYSARA, M., SAEYS, Y., LEYS, N., RAES, J. & MONSIEURS, P. 2015. CATCh, an ensemble classifier for chimera detection in 16S rRNA sequencing studies. *Appl Environ Microbiol,* 81**,** 1573-84.

NADALIAN, B., YADEGAR, A., HOURI, H., OLFATIFAR, M., SHAHROKH, S., ASADZADEH AGHDAEI, H., SUZUKI, H., et al. 2021. Prevalence of the pathobiont adherent-invasive Escherichia coli and inflammatory bowel disease: a systematic review and meta-analysis. *Journal of Gastroenterology and Hepatology,* 36**,** 852-863.

NAFTALI, T., RESHEF, L., KOVACS, A., PORAT, R., AMIR, I., KONIKOFF, F. M. & GOPHNA, U. 2016. Distinct Microbiotas are Associated with Ileum-Restricted and Colon-Involving Crohn's Disease. *Inflammatory Bowel Diseases,* 22**,** 293-302.

NAMIKI, T., HACHIYA, T., TANAKA, H. & SAKAKIBARA, Y. 2012. MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Res,* 40**,** e155.

NASH, J. H., VILLEGAS, A., KROPINSKI, A. M., AGUILAR-VALENZUELA, R., KONCZY, P., MASCARENHAS, M., ZIEBELL, K., et al. 2010. Genome sequence of adherent-invasive Escherichia coli and comparative genomic analysis with other E. coli pathotypes. *BMC Genomics,* 11**,** 667.

NAVAS-MOLINA, J. A., PERALTA-SÁNCHEZ, J. M., GONZÁLEZ, A., MCMURDIE, P. J., VÁZQUEZ-BAEZA, Y., XU, Z., URSELL, L. K., et al. 2013. Advancing our understanding of the human microbiome using QIIME. *Methods Enzymol,* 531**,** 371-444.

NEARING, J. T., DOUGLAS, G. M., COMEAU, A. M. & LANGILLE, M. G. I. 2018. Denoising the Denoisers: an independent evaluation of microbiome sequence error-correction approaches. *PeerJ,* 6**,** e5364.

NEARING, J. T., DOUGLAS, G. M., HAYES, M. G., MACDONALD, J., DESAI, D. K., ALLWARD, N., JONES, C. M. A., et al. 2022. Microbiome differential abundance methods produce different results across 38 datasets. *Nature Communications,* 13**,** 342.

NEBEL, M. E., WILD, S., HOLZHAUSER, M., HÜTTENBERGER, L., REITZIG, R., SPERBER, M. & STOECK, T. 2011. JAGUC--a software package for environmental diversity analyses. *J Bioinform Comput Biol,* 9**,** 749-73.

NEMOTO, H., KATAOKA, K., ISHIKAWA, H., IKATA, K., ARIMOCHI, H., IWASAKI, T., OHNISHI, Y., et al. 2012. Reduced Diversity and Imbalance of Fecal Microbiota in Patients with Ulcerative Colitis. *Digestive Diseases and Sciences,* 57**,** 2955-2964.

NEUT, C., BULOIS, P., DESREUMAUX, P., MEMBREÉ, J.-M., LEDERMAN, E., GAMBIEZ, L., CORTOT, A., et al. 2002. Changes in the bacterial flora of the neoterminal ileum after ileocolonic resection for Crohn's disease. *The American journal of gastroenterology,* 97**,** 939.

NG, S. C., SHI, H. Y., HAMIDI, N., UNDERWOOD, F. E., TANG, W., BENCHIMOL, E. I., PANACCIONE, R., et al. 2017. Worldwide incidence and prevalence of inflammatory bowel disease in the 21st century: a systematic review of population-based studies. *Lancet,* 390**,** 2769-2778.

NGUYEN, N. H., PICETTI, D., DULAI, P. S., JAIRATH, V. A.-O. X., SANDBORN, W. J., OHNO-MACHADO, L., CHEN, P. L., et al. Machine Learning-based Prediction Models for Diagnosis and Prognosis in Inflammatory Bowel Diseases: A Systematic Review.

NI, J., WU, G. D., ALBENBERG, L. & TOMOV, V. T. 2017. Gut microbiota and IBD: causation or correlation? *Nat Rev Gastroenterol Hepatol,* 14**,** 573-584.

NIKOLAUS, S., SCHULTE, B., AL-MASSAD, N., THIEME, F., SCHULTE, D. M., BETHGE, J., REHMAN, A., et al. 2017. Increased Tryptophan Metabolism Is Associated With Activity of Inflammatory Bowel Diseases. *Gastroenterology,* 153**,** 1504-1516 e2.

NILAKANTA, H., DREWS, K. L., FIRRELL, S., FOULKES, M. A. & JABLONSKI, K. A. 2014. A review of software for analyzing molecular sequences. *BMC Res Notes,* 7**,** 830.

NILSSON, R. H., ANSLAN, S., BAHRAM, M., WURZBACHER, C., BALDRIAN, P. & TEDERSOO, L. 2019. Mycobiome diversity: high-throughput sequencing and identification of fungi. *Nat Rev Microbiol,* 17**,** 95-109.

NILSSON, R. H., LARSSON, K.-H., TAYLOR, A. F S., BENGTSSON-PALME, J., JEPPESEN, T. S., SCHIGEL, D., KENNEDY, P., et al. 2018. The UNITE database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications. *Nucleic Acids Research,* 47**,** D259-D264.

NISHIDA, A., INOUE, R., INATOMI, O., BAMBA, S., NAITO, Y. & ANDOH, A. 2018. Gut microbiota in the pathogenesis of inflammatory bowel disease. *Clin J Gastroenterol,* 11**,** 1-10.

NISHIKAWA, J., KUDO, T., SAKATA, S., BENNO, Y. & SUGIYAMA, T. 2009. Diversity of mucosa-associated microbiota in active and inactive ulcerative colitis. *Scand J Gastroenterol,* 44**,** 180-6.

NITZAN, O., ELIAS, M., PERETZ, A. & SALIBA, W. 2016. Role of antibiotics for treatment of inflammatory bowel disease. *World J Gastroenterol,* 22**,** 1078-87.

NOVOA, E. M., JUNGREIS, I., JAILLON, O. & KELLIS, M. 2019. Elucidation of Codon Usage Signatures across the Domains of Life. *Molecular Biology and Evolution*.

NOWICKI, C., RAY, L., ENGEN, P., MADRIGRANO, A., WITT, T., LAD, T., COBLEIGH, M., et al. 2023. Comparison of gut microbiome composition in colonic biopsies, endoscopically-collected and at-home-collected stool samples. *Frontiers in Microbiology,* 14.

NURK, S., MELESHKO, D. & KOROBEYNIKOV, A. 2017. metaSPAdes: a new versatile metagenomic assembler. 27**,** 824-834.

O'BRIEN, C. L., BRINGER, M. A., HOLT, K. E., GORDON, D. M., DUBOIS, A. L., BARNICH, N., DARFEUILLE-MICHAUD, A., et al. 2017. Comparative genomics of Crohn's disease-associated adherent-invasive Escherichia coli. *Gut,* 66**,** 1382-1389.

O'LEARY, S. M., COLEMAN, M. M., CHEW, W. M., MORROW, C., MCLAUGHLIN, A. M., GLEESON, L. E., O'SULLIVAN, M. P., et al. 2014. Cigarette Smoking Impairs Human Pulmonary Immunity to Mycobacterium tuberculosis. *American Journal of Respiratory and Critical Care Medicine,* 190**,** 1430-1436.

ODES, H. S., FICH, A., REIF, S., HALAK, A., LAVY, A., KETER, D., ELIAKIM, R., et al. 2001. Effects of current cigarette smoking on clinical course of Crohn's disease and ulcerative colitis. *Dig Dis Sci,* 46**,** 1717-21.

OH, B., BOYLE, F., PAVLAKIS, N., CLARKE, S., GUMINSKI, A., EADE, T., LAMOURY, G., et al. 2021. Emerging Evidence of the Gut Microbiome in Chemotherapy: A Clinical Review. *Frontiers in Oncology,* 11.

OHLAND, C. L. & JOBIN, C. 2015. Microbial Activities and Intestinal Homeostasis: A Delicate Balance Between Health and Disease. *Cellular and Molecular Gastroenterology and Hepatology,* 1**,** 28-40.

OKSANEN, J., BLANCHET, F. G., KINDT, R., LEGENDRE, P., MINCHIN, P., O'HARA, R., SIMPSON, G., et al. 2013. Community ecology package. *R package version,* 2.

OKSANEN J, S. G., BLANCHET F, KINDT R, LEGENDRE P, MINCHIN P, O'HARA R, SOLYMOS P, STEVENS M, SZOECS E, WAGNER H, BARBOUR M, BEDWARD M, BOLKER B, BORCARD D, CARVALHO G, CHIRICO M, DE CACERES M,DURAND S, EVANGELISTA H, FITZJOHN R, FRIENDLY M, FURNEAUX B, HANNIGAN G, HILL M, LAHTI L, MCGLINN D, OUELLETTE M, RIBEIRO CUNHA E, SMITH T, STIER A, TER BRAAK C, WEEDON J 2022. vegan: Community Ecology Package. R package version 2.6-4 ed.

OLAISEN, M., SPIGSET, O., FLATBERG, A., GRANLUND, A. V. B., BREDE, W. R., ALBREKTSEN, G., ROYSET, E. S., et al. 2019. Mucosal 5-aminosalicylic acid concentration, drug formulation and mucosal microbiome in patients with quiescent ulcerative colitis. *Aliment Pharmacol Ther,* 49**,** 1301-1313.

OLBJØRN, C., CVANCAROVA SMÅSTUEN, M., THIIS-EVENSEN, E., NAKSTAD, B., VATN, M. H., JAHNSEN, J., RICANEK, P., et al. 2019. Fecal microbiota profiles in treatment-naïve pediatric inflammatory bowel disease - associations with disease phenotype, treatment, and outcome. *Clin Exp Gastroenterol,* 12**,** 37-49.

ORTIZO, R., LEE, S. Y., NGUYEN, E. T., JAMAL, M. M., BECHTOLD, M. M. & NGUYEN, D. L. 2017. Exposure to oral contraceptives increases the risk for development of inflammatory bowel disease: a meta-analysis of case-controlled and cohort studies. *European Journal of Gastroenterology & Hepatology,* 29.

OUNIT, R., WANAMAKER, S., CLOSE, T. J. & LONARDI, S. 2015. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC Genomics,* 16**,** 236.

OVADIA, C., PERDONES-MONTERO, A., SPAGOU, K., SMITH, A., SARAFIAN, M. H., GOMEZ-ROMERO, M., BELLAFANTE, E., et al. 2019. Enhanced Microbial Bile Acid Deconjugation and Impaired Ileal Uptake in Pregnancy Repress Intestinal Regulation of Bile Acid Synthesis. *Hepatology,* 70**,** 276-293.

PACHIKIAN, B. D., NEYRINCK, A. M., DELDICQUE, L., DE BACKER, F. C., CATRY, E., DEWULF, E. M., SOHET, F. M., et al. 2010. Changes in intestinal bifidobacteria levels are associated with the inflammatory response in magnesium-deficient mice. *J Nutr,* 140**,** 509-14.

PALAREA-ALBALADEJO, J. & MARTÍN-FERNÁNDEZ, J. A. 2015. zCompositions — R package for multivariate imputation of left-censored data under a compositional approach. *Chemometrics and Intelligent Laboratory Systems,* 143**,** 85-96.

PALJETAK, C. H., BAREŠIĆ, A., PANEK, M., PERIĆ, M., MATIJAŠIĆ, M., LOJKIĆ, I., BARIŠIĆ, A., et al. 2022. Gut microbiota in mucosa and feces of newly diagnosed, treatment-naïve adult inflammatory bowel disease and irritable bowel syndrome patients. *Gut Microbes,* 14**,** 2083419.

PALMELA, C., CHEVARIN, C., XU, Z., TORRES, J., SEVRIN, G., HIRTEN, R., BARNICH, N., et al. 2018. Adherent-invasive Escherichia coli in inflammatory bowel disease. *Gut,* 67**,** 574-587.

PALMIERI, O., CREANZA, T. M., BOSSA, F., PALUMBO, O., MAGLIETTA, R., ANCONA, N., CORRITORE, G., et al. 2015. Genome-wide Pathway Analysis Using Gene Expression Data of Colonic Mucosa in Patients with Inflammatory Bowel Disease. *Inflamm Bowel Dis,* 21**,** 1260-8.

PAN, X., ZHU, Q., PAN, L. L. & SUN, J. 2022. Macrophage immunometabolism in inflammatory bowel diseases: From pathogenesis to therapy. *Pharmacol Ther,* 238**,** 108176.

PANZITT, K., ZOLLNER, G., MARSCHALL, H.-U. & WAGNER, M. 2022. Recent advances on FXR-targeting therapeutics. *Molecular and Cellular Endocrinology,* 552**,** 111678.

PAPA, E., DOCKTOR, M., SMILLIE, C., WEBER, S., PREHEIM, S. P., GEVERS, D., GIANNOUKOS, G., et al. 2012. Non-Invasive Mapping of the Gastrointestinal Microbiota Identifies Children with Inflammatory Bowel Disease. *PLOS ONE,* 7**,** e39242.

PAPOUTSOGLOU, G., TARAZONA, S., LOPES, M. B., KLAMMSTEINER, T., IBRAHIMI, E., ECKENBERGER, J., NOVIELLI, P., et al. 2023. Machine learning approaches in microbiome research: challenges and best practices. *Frontiers in Microbiology,* 14.

PARADIS, E., CLAUDE, J. & STRIMMER, K. 2004. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics,* 20**,** 289-90.

PARIGI, T. L., MASTROROCCO, E., DA RIO, L., ALLOCCA, M., D'AMICO, F., ZILLI, A., FIORINO, G., et al. 2022. Evolution and New Horizons of Endoscopy in Inflammatory Bowel Diseases. *J Clin Med,* 11.

PARK, J. H., PEYRIN-BIROULET, L., EISENHUT, M. & SHIN, J. I. 2017. IBD immunopathogenesis: A comprehensive review of inflammatory molecules. *Autoimmun Rev,* 16**,** 416-426.

PASCAL, V., POZUELO, M., BORRUEL, N., CASELLAS, F., CAMPOS, D., SANTIAGO, A., MARTINEZ, X., et al. 2017. A microbial signature for Crohn's disease. *Gut,* 66**,** 813-822.

PATANGIA, D. V., ANTHONY RYAN, C., DEMPSEY, E., PAUL ROSS, R. & STANTON, C. 2022. Impact of antibiotics on the human microbiome and consequences for host health. *Microbiologyopen,* 11**,** e1260.
PATEL, R. K. & JAIN, M. 2012. NGS QC Toolkit: A Toolkit for Quality Control of Next Generation Sequencing Data. *PLOS ONE,* 7**,** e30619.
PATHIRANA, W. G. W., CHUBB, S. P., GILLETT, M. J. & VASIKARAN, S. D. 2018. Faecal Calprotectin. *Clin Biochem Rev,* 39**,** 77-90.
PATIN, N. V., KUNIN, V., LIDSTROM, U. & ASHBY, M. N. 2013. Effects of OTU clustering and PCR artifacts on microbial diversity estimates. *Microb Ecol,* 65**,** 709-19.
PATTNI, S. S., BRYDON, W. G., DEW, T., JOHNSTON, I. M., NOLAN, J. D., SRINIVAS, M., BASUMANI, P., et al. 2013. Fibroblast growth factor 19 in patients with bile acid diarrhoea: a prospective comparison of FGF19 serum assay and SeHCAT retention. *Aliment Pharmacol Ther,* 38**,** 967-76.
PAULSON, J. N., STINE, O. C., BRAVO, H. C. & POP, M. 2013. Differential abundance analysis for microbial marker-gene surveys. *Nature Methods,* 10**,** 1200.
PEABODY, M. A., VAN ROSSUM, T., LO, R. & BRINKMAN, F. S. L. 2015. Evaluation of shotgun metagenomics sequence classification methods using in silico and in vitro simulated communities. *BMC Bioinformatics,* 16**,** 362.
PEDERSEN, T. L. 2022. patchwork: The Composer of Plots. R package version 1.1.2 ed.
PENDERS, J., THIJS, C., VINK, C., STELMA, F. F., SNIJDERS, B., KUMMELING, I., VAN DEN BRANDT, P. A., et al. 2006. Factors influencing the composition of the intestinal microbiota in early infancy. *Pediatrics,* 118**,** 511-21.
PENG, Y., LEUNG, H. C., YIU, S. M. & CHIN, F. Y. 2011. Meta-IDBA: a de Novo assembler for metagenomic data. *Bioinformatics,* 27**,** i94-101.
PENG, Y., LEUNG, H. C., YIU, S. M. & CHIN, F. Y. 2012. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics,* 28**,** 1420-8.
PEPPERCORN, M. A. & GOLDMAN, P. 1972. The role of intestinal bacteria in the metabolism of salicylazosulfapyridine. *J Pharmacol Exp Ther,* 181**,** 555-62.
PEREIRA-FANTINI, P. M., LAPTHORNE, S., JOYCE, S. A., DELLIOS, N. L., WILSON, G., FOUHY, F., THOMAS, S. L., et al. 2014. Altered FXR signalling is associated with bile acid dysmetabolism in short bowel syndrome-associated liver disease. *J Hepatol,* 61**,** 1115-25.
PERENCEVICH, M. & BURAKOFF, R. 2006. Use of antibiotics in the treatment of inflammatory bowel disease. *Inflamm Bowel Dis,* 12**,** 651-64.
PETERSON, J., GARGES, S., GIOVANNI, M., MCINNES, P., WANG, L., SCHLOSS, J. A., BONAZZI, V., et al. 2009. The NIH Human Microbiome Project. *Genome Res,* 19**,** 2317-23.
PETERSON, L. W. & ARTIS, D. 2014. Intestinal epithelial cells: regulators of barrier function and immune homeostasis. *Nat Rev Immunol,* 14**,** 141-53.
PEYRIN-BIROULET, L., GERMAIN, A., PATEL, A. S. & LINDSAY, J. O. 2016. Systematic review: outcomes and post-operative complications following colectomy for ulcerative colitis. *Aliment Pharmacol Ther,* 44**,** 807-16.
PEYRIN-BIROULET, L., SANDBORN, W., SANDS, B. E., REINISCH, W., BEMELMAN, W., BRYANT, R. V., D'HAENS, G., et al. 2015. Selecting Therapeutic Targets in Inflammatory Bowel Disease (STRIDE): Determining Therapeutic Goals for Treat-to-Target. *Official journal of the American College of Gastroenterology | ACG,* 110.
PIGHTLING, A. W., PETTENGILL, J. B., LUO, Y., BAUGHER, J. D., RAND, H. & STRAIN, E. 2018. Interpreting Whole-Genome Sequence Analyses of Foodborne Bacteria for Regulatory Applications and Outbreak Investigations. *Frontiers in microbiology,* 9**,** 1482-1482.
PITTAYANON, R., LAU, J. T., LEONTIADIS, G. I., TSE, F., YUAN, Y., SURETTE, M. & MOAYYEDI, P. 2020. Differences in Gut Microbiota in Patients With vs Without Inflammatory Bowel Diseases: A Systematic Review. *Gastroenterology,* 158**,** 930-946 e1.
PIZARRO, T. T., STAPPENBECK, T. S., RIEDER, F., ROSEN, M. J., COLOMBEL, J.-F., DONOWITZ, M., TOWNE, J., et al. 2019. Challenges in IBD Research: Preclinical Human IBD Mechanisms. *Inflammatory Bowel Diseases,* 25**,** S5-S12.
PLÖGER, S., STUMPFF, F., PENNER, G. B., SCHULZKE, J. D., GÄBEL, G., MARTENS, H., SHEN, Z., et al. 2012. Microbial butyrate and its role for barrier function in the gastrointestinal tract. *Ann N Y Acad Sci,* 1258**,** 52-9.
PLOVIER, H., EVERARD, A., DRUART, C., DEPOMMIER, C., VAN HUL, M., GEURTS, L., CHILLOUX, J., et al. 2017. A purified membrane protein from Akkermansia muciniphila or the pasteurized bacterium improves metabolism in obese and diabetic mice. *Nature Medicine,* 23**,** 107-113.

PLUMMER, E. & TWIN, J. 2015. A Comparison of Three Bioinformatics Pipelines for the Analysis of Preterm Gut Microbiota using 16S rRNA Gene Sequencing Data. *Journal of Proteomics & Bioinformatics,* 8.

PNG, C. W., LINDÉN, S. K., GILSHENAN, K. S., ZOETENDAL, E. G., MCSWEENEY, C. S., SLY, L. I., MCGUCKIN, M. A., et al. 2010. Mucolytic bacteria with increased prevalence in IBD mucosa augment in vitro utilization of mucin by other bacteria. *Am J Gastroenterol,* 105**,** 2420-8.

POLLARD, M. O., GURDASANI, D., MENTZER, A. J., PORTER, T. & SANDHU, M. S. 2018. Long reads: their purpose and place. *Human Molecular Genetics,* 27**,** R234-R241.

POLLOCK, J., GLENDINNING, L., WISEDCHANWET, T. & WATSON, M. 2018. The Madness of Microbiome: Attempting To Find Consensus "Best Practice" for 16S Microbiome Studies. *Appl Environ Microbiol,* 84.

PORTER, R. J., ARENDS, M. J., CHURCHHOUSE, A. M. D. & DIN, S. 2021. Inflammatory Bowel Disease-Associated Colorectal Cancer: Translational Risks from Mechanisms to Medicines. *Journal of Crohn's and Colitis,* 15**,** 2131-2141.

POSTLER, T. S. & GHOSH, S. 2017. Understanding the Holobiont: How Microbial Metabolites Affect Human Health and Shape the Immune System. *Cell Metab,* 26**,** 110-130.

POUNCEY, A., SCOTT, A., ALEXANDER, J., MARCHESI, J. R. & KINROSS, J. 2018. Gut Microbiota, Chemotherapy and the Host: The Influence of the Gut Microbiota on Cancer Treatment. *Ecancermedicalscience*.

POUSSIN, C., SIERRO, N., BOUÉ, S., BATTEY, J., SCOTTI, E., BELCASTRO, V., PEITSCH, M. C., et al. 2018. Interrogating the microbiome: experimental and computational considerations in support of study reproducibility. *Drug Discov Today,* 23**,** 1644-1657.

POWRIE, F. 1995. T cells in inflammatory bowel disease: protective and pathogenic roles. *Immunity,* 3**,** 171-4.

PRINDIVILLE, T. P., SHEIKH, R. A., COHEN, S. H., TANG, Y. J., CANTRELL, M. C. & SILVA, J., JR. 2000. Bacteroides fragilis enterotoxin gene sequences in patients with inflammatory bowel disease. *Emerg Infect Dis,* 6**,** 171-4.

PROCTOR, L. M., CREASY, H. H., FETTWEIS, J. M., LLOYD-PRICE, J., MAHURKAR, A., ZHOU, W., BUCK, G. A., et al. 2019. The Integrative Human Microbiome Project. *Nature,* 569**,** 641-648.

PROSBERG, M., BENDTSEN, F., VIND, I., PETERSEN, A. M. & GLUUD, L. L. 2016. The association between the gut microbiota and the inflammatory bowel disease activity: a systematic review and meta-analysis. *Scandinavian Journal of Gastroenterology,* 51**,** 1407-1415.

PRUESSE, E., PEPLIES, J. & GLOCKNER, F. O. 2012. SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics,* 28**,** 1823-9.

PRUESSE, E., QUAST, C., KNITTEL, K., FUCHS, B. M., LUDWIG, W., PEPLIES, J. & GLÖCKNER, F. O. 2007. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res,* 35**,** 7188-96.

PUDJIHARTONO, N., FADASON, T., KEMPA-LIEHR, A. W. & O'SULLIVAN, J. M. 2022. A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction. *Frontiers in Bioinformatics,* 2.

PUJATO, S. A., QUIBERONI, A. & MERCANTI, D. J. 2019. Bacteriophages on dairy foods. *J Appl Microbiol,* 126**,** 14-30.

QIE, D., ZHANG, Y., GONG, X., HE, Y., QIAO, L., LU, G. & LI, Y. 2021. SLC10A2 deficiency-induced congenital chronic bile acid diarrhea and stunting. *Mol Genet Genomic Med,* 9**,** e1740.

QIN, J., LI, R., RAES, J., ARUMUGAM, M., BURGDORF, K. S., MANICHANH, C., NIELSEN, T., et al. 2010. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature,* 464**,** 59-65.

QIN, J., LI, Y., CAI, Z., LI, S., ZHU, J., ZHANG, F., LIANG, S., et al. 2012. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature,* 490**,** 55-60.

QIU, J., HELLER, J. J., GUO, X., CHEN, Z. M., FISH, K., FU, Y. X. & ZHOU, L. 2012. The aryl hydrocarbon receptor regulates gut immunity through modulation of innate lymphoid cells. *Immunity,* 36**,** 92-104.

QUAST, C., PRUESSE, E., YILMAZ, P., GERKEN, J., SCHWEER, T., YARZA, P., PEPLIES, J., et al. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res,* 41**,** D590-6.

QUEVRAIN, E., MAUBERT, M. A., MICHON, C., CHAIN, F., MARQUANT, R., TAILHADES, J., MIQUEL, S., et al. 2016. Identification of an anti-inflammatory protein from Faecalibacterium prausnitzii, a commensal bacterium deficient in Crohn's disease. *Gut,* 65**,** 415-425.

QUEZADA, S. M., MCLEAN, L. P. & CROSS, R. K. 2018. Adverse events in IBD therapy: the 2018 update. *Expert Rev Gastroenterol Hepatol,* 12**,** 1183-1191.

QUINCE, C., LANZEN, A., DAVENPORT, R. J. & TURNBAUGH, P. J. 2011. Removing noise from pyrosequenced amplicons. *BMC Bioinformatics,* 12**,** 38.
QUINCE, C., WALKER, A. W., SIMPSON, J. T., LOMAN, N. J. & SEGATA, N. 2017. Shotgun metagenomics, from sampling to analysis. *Nature Biotechnology,* 35**,** 833.
QUINN, R. A., MELNIK, A. V., VRBANAC, A., FU, T., PATRAS, K. A., CHRISTY, M. P., BODAI, Z., et al. 2020. Global chemical effects of the microbiome include new bile-acid conjugations. *Nature,* 579**,** 123-129.
QUINN, T. P. 2021. Stool Studies Don't Pass the Sniff Test: A Systematic Review of Human Gut Microbiome Research Suggests Widespread Misuse of Machine Learning. Available: https://ui.adsabs.harvard.edu/abs/2021arXiv210703611Q [Accessed July 01, 2021].
QUINN, T. P., ERB, I., GLOOR, G., NOTREDAME, C., RICHARDSON, M. F. & CROWLEY, T. M. 2019. A field guide for the compositional analysis of any-omics data. *GigaScience,* 8**,** giz107.
RADHAKRISHNAN, S. T., ALEXANDER, J. L., MULLISH, B. H., GALLAGHER, K. I., POWELL, N., HICKS, L. C., HART, A. L., et al. 2022. Systematic review: the association between the gut microbiota and medical therapies in inflammatory bowel disease. *Aliment Pharmacol Ther,* 55**,** 26-48.
RAJCA, S., GRONDIN, V., LOUIS, E., VERNIER-MASSOUILLE, G., GRIMAUD, J. C., BOUHNIK, Y., LAHARIE, D., et al. 2014. Alterations in the intestinal microbiome (dysbiosis) as a predictor of relapse after infliximab withdrawal in Crohn's disease. *Inflamm Bowel Dis,* 20**,** 978-86.
RAJILIĆ-STOJANOVIĆ, M., SHANAHAN, F., GUARNER, F. & DE VOS, W. M. 2013. Phylogenetic analysis of dysbiosis in ulcerative colitis during remission. *Inflamm Bowel Dis,* 19**,** 481-8.
RAJPUT, B., PRUITT, K. D. & MURPHY, T. D. 2019. RefSeq curation and annotation of stop codon recoding in vertebrates. *Nucleic Acids Research,* 47**,** 594-606.
RAMAKRISHNA, B. S., VARGHESE, R., JAYAKUMAR, S., MATHAN, M. & BALASUBRAMANIAN, K. A. 1997. Circulating antioxidants in ulcerative colitis and their relationship to disease severity and activity. *J Gastroenterol Hepatol,* 12**,** 490-4.
RAMPTON, D. & SHANAHAN, F. 2016. *Fast Facts: Inflammatory Bowel Disease*.
RAMPTON, D. S. & SHANAHAN, F. 2020. Fast Facts: Inflammatory Bowel Disease: Translating the science into compassionate IBD care. S.Karger AG.
RASHIDI, A., EBADI, M., REHMAN, T. U., ELHUSSEINI, H., NALLURI, H., KAISER, T., HOLTAN, S. G., et al. 2021. Gut Microbiota Response to Antibiotics Is Personalized and Depends on Baseline Microbiota. *Microbiome*.
RAUSCH, P., ELLUL, S., PISANI, A., BANG, C., TABONE, T., MARANTIDIS CORDINA, C., ZAHRA, G., et al. 2023. Microbial Dynamics in Newly Diagnosed and Treatment Naïve IBD Patients in the Mediterranean. *Inflammatory Bowel Diseases,* 29**,** 1118-1132.
REESE, G. E., NANIDIS, T., BORYSIEWICZ, C., YAMAMOTO, T., ORCHARD, T. & TEKKIS, P. P. 2008. The effect of smoking after surgery for Crohn's disease: a meta-analysis of observational studies. *Int J Colorectal Dis,* 23**,** 1213-21.
REINOSO WEBB, C., DEN BAKKER, H., KOBOZIEV, I., JONES-HALL, Y., RAO KOTTAPALLI, K., OSTANIN, D., FURR, K. L., et al. 2018. Differential Susceptibility to T Cell-Induced Colitis in Mice: Role of the Intestinal Microbiota. *Inflamm Bowel Dis,* 24**,** 361-379.
REN, J., AHLGREN, N. A., LU, Y. Y., FUHRMAN, J. A. & SUN, F. 2017. VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome,* 5**,** 69.
RESCIGNO, M. & DI SABATINO, A. 2009. Dendritic cells in intestinal homeostasis and disease. *J Clin Invest,* 119**,** 2441-50.
REZAEI, S. R. & AHMADI, A. 2023. A hierarchical GAN method with ensemble CNN for accurate nodule detection. *Int J Comput Assist Radiol Surg,* 18**,** 695-705.
REŽEN, T., ROZMAN, D., KOVÁCS, T., KOVÁCS, P., SIPOS, A., BAI, P. & MIKÓ, E. 2022. The role of bile acids in carcinogenesis. *Cell Mol Life Sci,* 79**,** 243.
RHO, M., TANG, H. & YE, Y. 2010. FragGeneScan: predicting genes in short and error-prone reads. *Nucleic acids research,* 38**,** e191-e191.
RICHARD, D. M., DAWES, M. A., MATHIAS, C. W., ACHESON, A., HILL-KAPTURCZAK, N. & DOUGHERTY, D. M. 2009. L-tryptophan: Basic metabolic functions, behavioral research and therapeutic indications. *International Journal of Tryptophan Research,* 2**,** 45-60.
RICHTER, D. C., OTT, F., AUCH, A. F., SCHMID, R. & HUSON, D. H. 2008. MetaSim: a sequencing simulator for genomics and metagenomics. *PLoS One,* 3**,** e3373.
RIDLON, J. M. & BAJAJ, J. S. 2015. The human gut sterolbiome: bile acid-microbiome endocrine aspects and therapeutics. *Acta Pharm Sin B,* 5**,** 99-105.

RIDLON, J. M., KANG, D. J. & HYLEMON, P. B. 2006. Bile salt biotransformations by human intestinal bacteria. *J Lipid Res,* 47**,** 241-59.

RIGOTTIER-GOIS, L. 2013. Dysbiosis in inflammatory bowel diseases: the oxygen hypothesis. *Isme j,* 7**,** 1256-61.

RIMOLDI, M., CHIEPPA, M., SALUCCI, V., AVOGADRI, F., SONZOGNI, A., SAMPIETRO, G. M., NESPOLI, A., et al. 2005. Intestinal immune homeostasis is regulated by the crosstalk between epithelial cells and dendritic cells. *Nat Immunol,* 6**,** 507-14.

RINNINELLA, E., RAOUL, P., CINTONI, M., FRANCESCHI, F., MIGGIANO, G. A. D., GASBARRINI, A. & MELE, M. C. 2019. What is the Healthy Gut Microbiota Composition? A Changing Ecosystem across Age, Environment, Diet, and Diseases. *Microorganisms,* 7.

RITARI, J., SALOJARVI, J., LAHTI, L. & DE VOS, W. M. 2015. Improved taxonomic assignment of human intestinal 16S rRNA sequences by a dedicated reference database. *BMC Genomics,* 16**,** 1056.

ROBERTS, C. L., KEITA, Å. V., DUNCAN, S. H., O'KENNEDY, N., SÖDERHOLM, J. D., RHODES, J. M. & CAMPBELL, B. J. 2010. Translocation of Crohn's disease Escherichia coli across M-cells: contrasting effects of soluble plant fibres and emulsifiers. *Gut,* 59**,** 1331.

ROBIN, X., TURCK, N., HAINARD, A., TIBERTI, N., LISACEK, F., SANCHEZ, J.-C. & MÜLLER, M. 2011. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics,* 12**,** 77.

ROCCHI, A., BENCHIMOL, E. I., BERNSTEIN, C. N., BITTON, A., FEAGAN, B., PANACCIONE, R., GLASGOW, K. W., et al. 2012. Inflammatory Bowel Disease: A Canadian Burden of Illness Review. *Canadian Journal of Gastroenterology,* 26**,** 984575.

ROSSI, O., VAN BERKEL, L. A., CHAIN, F., TANWEER KHAN, M., TAVERNE, N., SOKOL, H., DUNCAN, S. H., et al. 2016. Faecalibacterium prausnitzii A2-165 has a high capacity to induce IL-10 in human and murine dendritic cells and modulates T cell responses. *Sci Rep,* 6**,** 18507.

ROTHMAN, K. J. & GREENLAND, S. 2005. Validity and Generalizability in Epidemiologic Studies. *Encyclopedia of Biostatistics.*

ROTHSCHILD, D., WEISSBROD, O., BARKAN, E., KURILSHIKOV, A., KOREM, T., ZEEVI, D., COSTEA, P. I., et al. 2018. Environment dominates over host genetics in shaping human gut microbiota. *Nature,* 555**,** 210-215.

ROUMPEKA, D. D., WALLACE, R. J., ESCALETTES, F., FOTHERINGHAM, I. & WATSON, M. 2017. A Review of Bioinformatics Tools for Bio-Prospecting from Metagenomic Sequence Data. *Front Genet,* 8**,** 23.

ROUND, J. L., LEE, S. M., LI, J., TRAN, G., JABRI, B., CHATILA, T. A. & MAZMANIAN, S. K. 2011. The Toll-like receptor 2 pathway establishes colonization by a commensal of the human microbiota. *Science,* 332**,** 974-7.

ROUND, J., MAZMANIAN, S. 2009. The gut microbiota shapes intestinal immune responses during health and disease. Nat Rev Immunol 9, 313–323. https://doi.org/10.1038/nri2515

ROUX, S., ENAULT, F., HURWITZ, B. L. & SULLIVAN, M. B. 2015. VirSorter: mining viral signal from microbial genomic data. *PeerJ,* 3**,** e985.

ROUX, S., TOURNAYRE, J., MAHUL, A., DEBROAS, D. & ENAULT, F. 2014. Metavir 2: new tools for viral metagenome comparison and assembled virome analysis. *BMC bioinformatics,* 15**,** 76-76.

RUBIO, A., PIGNEUR, B., GARNIER-LENGLINÉ, H., TALBOTEC, C., SCHMITZ, J., CANIONI, D., GOULET, O., et al. 2011. The efficacy of exclusive nutritional therapy in paediatric Crohn's disease, comparing fractionated oral vs. continuous enteral feeding. *Aliment Pharmacol Ther,* 33**,** 1332-9.

RYAN, F. J., AHERN, A. M., FITZGERALD, R. S., LASERNA-MENDIETA, E. J., POWER, E. M., CLOONEY, A. G., O'DONOGHUE, K. W., et al. 2020. Colonic microbiota is associated with inflammation and host epigenomic alterations in inflammatory bowel disease. *Nature Communications,* 11**,** 1512.

SAARELA, M. & JAUHIAINEN, S. 2021. Comparison of feature importance measures as explanations for classification models. *SN Applied Sciences,* 3**,** 272.

SACK, R. B., MYERS, L. L., ALMEIDO-HILL, J., SHOOP, D. S., BRADBURY, W. C., REID, R. & SANTOSHAM, M. 1992. Enterotoxigenic Bacteroides fragilis: epidemiologic studies of its role as a human diarrhoeal pathogen. *J Diarrhoeal Dis Res,* 10**,** 4-9.

SAIRENJI, T., COLLINS, K. L. & EVANS, D. V. 2017. An Update on Inflammatory Bowel Disease. *Prim Care,* 44**,** 673-692.

SAKAMOTO, N., KONO, S., WAKAI, K., FUKUDA, Y., SATOMI, M., SHIMOYAMA, T., INABA, Y., et al. 2005. Dietary risk factors for inflammatory bowel disease: a multicenter case-control study in Japan. *Inflamm Bowel Dis,* 11**,** 154-63.

SALONEN, A., NIKKILÄ, J., JALANKA-TUOVINEN, J., IMMONEN, O., RAJILIĆ-STOJANOVIĆ, M., KEKKONEN, R. A., PALVA, A., et al. 2010. Comparative analysis of fecal DNA extraction methods with phylogenetic microarray: effective recovery of bacterial and archaeal DNA using mechanical cell lysis. *J Microbiol Methods,* 81**,** 127-34.

SALTER, S. J., COX, M. J., TUREK, E. M., CALUS, S. T., COOKSON, W. O., MOFFATT, M. F., TURNER, P., et al. 2014. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biology,* 12**,** 87.

SANDBORN, W. J., D'HAENS, G. R., REINISCH, W., PANÉS, J., CHAN, D., GONZALEZ, S., WEISEL, K., et al. 2022. Guselkumab for the Treatment of Crohn's Disease: Induction Results From the Phase 2 GALAXI-1 Study. *Gastroenterology,* 162**,** 1650-1664.e8.

SANDBORN, W. J., FEAGAN, B. G., RUTGEERTS, P., HANAUER, S., COLOMBEL, J. F., SANDS, B. E., LUKAS, M., et al. 2013. Vedolizumab as induction and maintenance therapy for Crohn's disease. *N Engl J Med,* 369**,** 711-21.

SANDBORN, W. J., SU, C., SANDS, B. E., D'HAENS, G. R., VERMEIRE, S., SCHREIBER, S., DANESE, S., et al. 2017. Tofacitinib as Induction and Maintenance Therapy for Ulcerative Colitis. *N Engl J Med,* 376**,** 1723-1736.

SANKARASUBRAMANIAN, J., AHMAD, R., AVUTHU, N., SINGH, A. B. & GUDA, C. 2020. Gut Microbiota and Metabolic Specificity in Ulcerative Colitis and Crohn's Disease. *Front Med (Lausanne),* 7**,** 606298.

SAR, J. V. D., DONKERVLIET, J. & IOSUP, A. 2019. Yardstick: A Benchmark for Minecraft-like Services. *Proceedings of the 2019 ACM/SPEC International Conference on Performance Engineering.* Mumbai, India: Association for Computing Machinery.

SARGAUTIENE, V., LIGERE, R., KALNIŅA, I., JĀKOBSONE, I., NIKOLAJEVA, V. & DEROVS, A. 2020. The Effect of 5-Aminosalicylic Acid on Intestinal Microbiota. *Proceedings of the Latvian Academy of Sciences. Section B. Natural, Exact, and Applied Sciences.,* 74**,** 53-57.

SARTOR, R. B. 2015. Optimal sampling of the intestinal microbiota for research. *Nature Reviews Gastroenterology & Hepatology,* 12**,** 253-254.

SASAKI, M., SITARAMAN, S. V., BABBIN, B. A., GERNER-SMIDT, P., RIBOT, E. M., GARRETT, N., ALPERN, J. A., et al. 2007. Invasive Escherichia coli are a feature of Crohn's disease. *Lab Invest,* 87**,** 1042-54.

SAUTER, G. H., MUNZING, W., RITTER, C. V. & PAUMGARTNER, G. 1999. Bile Acid Malabsorption as a Cause of Chronic Diarrhea Diagnostic Value of 7α-Hydroxy-4-Cholesten-3-One in Serum. *Digestive Diseases and Sciences,* 44**,** 14-19.

SAYERS, E. W., AGARWALA, R., BOLTON, E. E., BRISTER, J. R., CANESE, K., CLARK, K., CONNOR, R., et al. 2019. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res,* 47**,** D23-d28.

SAYIN, S. I., WAHLSTROM, A., FELIN, J., JANTTI, S., MARSCHALL, H. U., BAMBERG, K., ANGELIN, B., et al. 2013. Gut microbiota regulates bile acid metabolism by reducing the levels of tauro-beta-muricholic acid, a naturally occurring FXR antagonist. *Cell Metab,* 17**,** 225-35.

SAZONOVS, A., KENNEDY, N. A., MOUTSIANAS, L., HEAP, G. A., RICE, D. L., REPPELL, M., BEWSHEA, C. M., et al. 2020. HLA-DQA1*05 Carriage Associated With Development of Anti-Drug Antibodies to Infliximab and Adalimumab in Patients With Crohn's Disease. *Gastroenterology,* 158**,** 189-199.

SCALDAFERRI, F., GERARDI, V., MANGIOLA, F., LOPETUSO, L. R., PIZZOFERRATO, M., PETITO, V., PAPA, A., et al. 2016. Role and mechanisms of action of Escherichia coli Nissle 1917 in the maintenance of remission in ulcerative colitis patients: An update. *World J Gastroenterol,* 22**,** 5505-11.

SCANLAN, P. D., SHANAHAN, F., O'MAHONY, C. & MARCHESI, J. R. 2006. Culture-independent analyses of temporal variation of the dominant fecal microbiota and targeted bacterial subgroups in Crohn's disease. *Journal of clinical microbiology,* 44**,** 3980-3988.

SCHAAP, F. G., TRAUNER, M. & JANSEN, P. L. 2014. Bile acid receptors as targets for drug development. *Nat Rev Gastroenterol Hepatol,* 11**,** 55-67.

SCHAUBECK, M., CLAVEL, T., CALASAN, J., LAGKOUVARDOS, I., HAANGE, S. B., JEHMLICH, N., BASIC, M., et al. 2016. Dysbiotic gut microbiota causes transmissible Crohn's disease-like ileitis independent of failure in antimicrobial defence. *Gut,* 65**,** 225-37.

SCHIRMER, M., FRANZOSA, E. A., LLOYD-PRICE, J., MCIVER, L. J., SCHWAGER, R., POON, T. W., ANANTHAKRISHNAN, A. N., et al. 2018. Dynamics of metatranscription in the inflammatory bowel disease gut microbiome. *Nature Microbiology,* 3**,** 337-346.

SCHLOSS, P. D. 2010. The effects of alignment quality, distance calculation method, sequence filtering, and region on the analysis of 16S rRNA gene-based studies. *PLoS Comput Biol,* 6**,** e1000844.

SCHLOSS, P. D. 2018. Identifying and Overcoming Threats to Reproducibility, Replicability, Robustness, and Generalizability in Microbiome Research. *mBio,* 9.

SCHLOSS, P. D., GEVERS, D. & WESTCOTT, S. L. 2011. Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PLoS One,* 6**,** e27310.

SCHLOSS, P. D. & HANDELSMAN, J. 2005. Introducing DOTUR, a Computer Program for Defining Operational Taxonomic Units and Estimating Species Richness. *Applied and Environmental Microbiology,* 71**,** 1501-1506.

SCHLOSS, P. D., WESTCOTT, S. L., RYABIN, T., HALL, J. R., HARTMANN, M., HOLLISTER, E. B., LESNIEWSKI, R. A., et al. 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol,* 75**,** 7537-41.

SCHMIDT, T. S., MATIAS RODRIGUES, J. F. & VON MERING, C. 2015. Limits to robustness and reproducibility in the demarcation of operational taxonomic units. *Environ Microbiol,* 17**,** 1689-706.

SCHMIEDER, R. & EDWARDS, R. 2011. Quality control and preprocessing of metagenomic datasets. *Bioinformatics (Oxford, England),* 27**,** 863-864.

SCHOENFELD, KIMMEY, SCHEIMAN, BJORKMAN & LAINE 1999. Review article: nonsteroidal anti-inflammatory drug-associated gastrointestinal complications—guidelines for prevention and treatment. *Alimentary Pharmacology & Therapeutics,* 13**,** 1273-1285.

SCHOEPFER, A. M., DEHLAVI, M. A., FOURNIER, N., SAFRONEEVA, E., STRAUMANN, A., PITTET, V., PEYRIN-BIROULET, L., et al. 2013. Diagnostic delay in Crohn's disease is associated with a complicated disease course and increased operation rate. *Am J Gastroenterol,* 108**,** 1744-53; quiz 1754.

SCIBERRAS, M., KARMIRIS, K., NASCIMENTO, C., TABONE, T., NIKOLAOU, P., THEODOROPOULOU, A., MULA, A., et al. 2022. Mental Health, Work Presenteeism, and Exercise in Inflammatory Bowel Disease. *Journal of Crohn's and Colitis,* 16**,** 1197-1201.

SCOTT, F. I., RUBIN, D. T., KUGATHASAN, S., BOUSVAROS, A., ELSON, C. O., NEWBERRY, R. D., MELMED, G. Y., et al. 2019. Challenges in IBD Research: Pragmatic Clinical Research. *Inflammatory Bowel Diseases,* 25**,** S40-S47.

SEARS, C. L., ISLAM, S., SAHA, A., ARJUMAND, M., ALAM, N. H., FARUQUE, A. S. G., SALAM, M. A., et al. 2008. Association of Enterotoxigenic Bacteroides fragilis Infection with Inflammatory Diarrhea. *Clinical Infectious Diseases,* 47**,** 797-803.

SEGAL, J. P., LEBLANC, J. F. & HART, A. L. 2021. Ulcerative colitis: an update. *Clin Med (Lond),* 21**,** 135-139.

SEGATA, N. 2018. On the Road to Strain-Resolved Comparative Metagenomics. *mSystems,* 3**,** e00190-17.

SEGATA, N., WALDRON, L., BALLARINI, A., NARASIMHAN, V., JOUSSON, O. & HUTTENHOWER, C. 2012. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat Methods,* 9**,** 811-4.

SEGOTA, I. & LONG, T. 2019. A high-resolution pipeline for 16S-sequencing identifies bacterial strains in human microbiome. *bioRxiv***,** 565572.

SEGURITAN, V. & ROHWER, F. 2001. FastGroup: A program to dereplicate libraries of 16S rDNA sequences. *BMC Bioinformatics,* 2**,** 9.

SENDER, R., FUCHS, S. & MILO, R. 2016. Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLoS Biol,* 14**,** e1002533.

SHABAN, N., HOAD, C. L., NAIM, I., ALSHAMMARI, M., RADFORD, S. J., CLARKE, C., MARCIANI, L., et al. 2022. Imaging in inflammatory bowel disease: current and future perspectives. *Frontline Gastroenterol,* 13**,** e28-e34.

SHAH, S. C., KHALILI, H., CHEN, C. Y., AHN, H. S., NG, S. C., BURISCH, J. & COLOMBEL, J. F. 2019. Sex-based differences in the incidence of inflammatory bowel diseases-pooled analysis of population-based studies from the Asia-Pacific region. *Aliment Pharmacol Ther,* 49**,** 904-911.

SHAH, S. C., KHALILI, H., GOWER-ROUSSEAU, C., OLEN, O., BENCHIMOL, E. I., LYNGE, E., NIELSEN, K. R., et al. 2018. Sex-Based Differences in Incidence of Inflammatory Bowel Diseases-Pooled Analysis of Population-Based Studies From Western Countries. *Gastroenterology,* 155**,** 1079-1089.e3.

SHANAHAN, F., GHOSH, T. S. & O'TOOLE, P. W. 2021. The Healthy Microbiome—What Is the Definition of a Healthy Gut Microbiome? *Gastroenterology,* 160**,** 483-494.

SHARMA, V. K., KUMAR, N., PRAKASH, T. & TAYLOR, T. D. 2012. Fast and Accurate Taxonomic Assignments of Metagenomic Sequences Using MetaBin. *PLOS ONE,* 7**,** e34030.

SHAW, S. Y., BLANCHARD, J. F. & BERNSTEIN, C. N. 2010. Association between the use of antibiotics in the first year of life and pediatric inflammatory bowel disease. *Am J Gastroenterol,* 105**,** 2687-92.

SHAW, S. Y., BLANCHARD, J. F. & BERNSTEIN, C. N. 2011. Association between the use of antibiotics and new diagnoses of Crohn's disease and ulcerative colitis. *Am J Gastroenterol,* 106**,** 2133-42.

SHAW, S. Y., BLANCHARD, J. F. & BERNSTEIN, C. N. 2013. Association between early childhood otitis media and pediatric inflammatory bowel disease: an exploratory population-based analysis. *J Pediatr,* 162**,** 510-4.

SHAWKI, A. & MCCOLE, D. F. 2017. Mechanisms of intestinal epithelial barrier dysfunction by adherent-invasive Escherichia coli. *Cellular and molecular gastroenterology and hepatology,* 3**,** 41-50.

SHEEHAN, D. & SHANAHAN, F. 2017. The Gut Microbiota in Inflammatory Bowel Disease. *Gastroenterol Clin North Am,* 46**,** 143-154.

SHEIKH, H., PRINS, C. & SCHRIJVERS, E. 2023. Artificial Intelligence: Definition and Background. *In:* SHEIKH, H., PRINS, C. & SCHRIJVERS, E. (eds.) *Mission AI: The New System Technology.* Cham: Springer International Publishing.

SHI, A. X., ZHOU, Y., ZHANG, X.-Y., ZHAO, Y.-S., QIN, H.-Y., WANG, Y.-P. & WU, X.-A. 2017. Irinotecan-induced bile acid malabsorption is associated with down-regulation of ileal Asbt (Slc10a2) in mice. *European Journal of Pharmaceutical Sciences,* 102**,** 220-229.

SICA, G. S. & BIANCONE, L. 2013. Surgery for inflammatory bowel disease in the era of laparoscopy. *World J Gastroenterol,* 19**,** 2445-8.

SINGH, A., MAHAJAN, R., KEDIA, S., DUTTA, A. K., ANAND, A., BERNSTEIN, C. N., DESAI, D., et al. 2022. Use of thiopurines in inflammatory bowel disease: an update. *Intest Res,* 20**,** 11-30.

SINGH, G. 2000. Gastrointestinal complications of prescription and over-the-counter nonsteroidal anti-inflammatory drugs: A view from the ARAMIS database. *American Journal of Therapeutics,* 7**,** 115-121.

SINGH, N., GURAV, A., SIVAPRAKASAM, S., BRADY, E., PADIA, R., SHI, H., THANGARAJU, M., et al. 2014. Activation of Gpr109a, receptor for niacin and the commensal metabolite butyrate, suppresses colonic inflammation and carcinogenesis. *Immunity,* 40**,** 128-39.

SINHA, S. R., HAILESELASSIE, Y., NGUYEN, L. P., TROPINI, C., WANG, M., BECKER, L. S., SIM, D., et al. 2020. Dysbiosis-Induced Secondary Bile Acid Deficiency Promotes Intestinal Inflammation. *Cell Host Microbe,* 27**,** 659-670.e5.

SIVA, S., RUBIN, D. T., GULOTTA, G., WROBLEWSKI, K. & PEKOW, J. 2017. Zinc Deficiency is Associated with Poor Clinical Outcomes in Patients with Inflammatory Bowel Disease. *Inflamm Bowel Dis,* 23**,** 152-157.

SLATTERY, S. A., NIAZ, O., AZIZ, Q., FORD, A. C. & FARMER, A. D. 2015. Systematic review with meta-analysis: the prevalence of bile acid malabsorption in the irritable bowel syndrome with diarrhoea. *Aliment Pharmacol Ther,* 42**,** 3-11.

SMALL, C. L., REID-YU, S. A., MCPHEE, J. B. & COOMBES, B. K. 2013. Persistent infection with Crohn's disease-associated adherent-invasive Escherichia coli leads to chronic inflammation and intestinal fibrosis. *Nat Commun,* 4**,** 1957.

SMITH, P. M., HOWITT, M. R., PANIKOV, N., MICHAUD, M., GALLINI, C. A., BOHLOOLY, Y. M., GLICKMAN, J. N., et al. 2013. The microbial metabolites, short-chain fatty acids, regulate colonic Treg cell homeostasis. *Science,* 341**,** 569-73.

SOKOL, H., PIGNEUR, B., WATTERLOT, L., LAKHDARI, O., BERMUDEZ-HUMARAN, L. G., GRATADOUX, J. J., BLUGEON, S., et al. 2008. 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**,** 16731-6.

SOMMER, D. D., DELCHER, A. L., SALZBERG, S. L. & POP, M. 2007. Minimus: a fast, lightweight genome assembler. *BMC Bioinformatics,* 8**,** 64.

SONG, G., FIOCCHI, C. & ACHKAR, J.-P. 2019a. Acupuncture in Inflammatory Bowel Disease. *Inflammatory Bowel Diseases,* 25**,** 1129-1139.

SONG, X., SUN, X., OH, S. F., WU, M., ZHANG, Y., ZHENG, W., GEVA-ZATORSKY, N., et al. 2020. Microbial bile acid metabolites modulate gut RORγ+ regulatory T cell homeostasis. *Nature,* 577**,** 410-415.

SONG, Z., CAI, Y., LAO, X., WANG, X., LIN, X., CUI, Y., KALAVAGUNTA, P. K., et al. 2019b. Taxonomic profiling and populational patterns of bacterial bile salt hydrolase (BSH) genes based on worldwide human gut microbiome. *Microbiome,* 7**,** 9.

SONNENBURG, ERICA D. & SONNENBURG, JUSTIN L. 2014. Starving our Microbial Self: The Deleterious Consequences of a Diet Deficient in Microbiota-Accessible Carbohydrates. *Cell Metabolism,* 20**,** 779-786.

SONNENBURG, J. L., XU, J., LEIP, D. D., CHEN, C.-H., WESTOVER, B. P., WEATHERFORD, J., BUHLER, J. D., et al. 2005. Glycan Foraging in Vivo by an Intestine-Adapted Bacterial Symbiont. *Science,* 307**,** 1955-1959.

STAELS, B. & FONSECA, V. A. 2009. Bile acids and metabolic regulation: mechanisms and clinical responses to bile acid sequestration. *Diabetes Care,* 32 Suppl 2**,** S237-45.

STAFFORD, I. A.-O., GOSINK, M. M., MOSSOTTO, E., ENNIS, S. & HAUBEN, M. 2022. A Systematic Review of Artificial Intelligence and Machine Learning Applications to Inflammatory Bowel Disease, with Practical Guidelines for Interpretation.

STALEY, C., WEINGARDEN, A. R., KHORUTS, A. & SADOWSKY, M. J. 2017. Interaction of gut microbiota with bile acid metabolism and its influence on disease states. *Appl Microbiol Biotechnol,* 101**,** 47-64.

STALLMACH, A., HAGEL, S. & BRUNS, T. 2010. Adverse effects of biologics used for treating IBD. *Best Practice & Research Clinical Gastroenterology,* 24**,** 167-182.

STEPHENS, Z. D., LEE, S. Y., FAGHRI, F., CAMPBELL, R. H., ZHAI, C., EFRON, M. J., IYER, R., et al. 2015. Big data: astronomical or genomical? *PLoS biology,* 13**,** e1002195.

STEWART, C. J., AJAMI, N. J., O'BRIEN, J. L., HUTCHINSON, D. S., SMITH, D. P., WONG, M. C., ROSS, M. C., et al. 2018. Temporal development of the gut microbiome in early childhood from the TEDDY study. *Nature,* 562**,** 583-588.

STODDARD, S. F., SMITH, B. J., HEIN, R., ROLLER, B. R. & SCHMIDT, T. M. 2015. rrnDB: improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development. *Nucleic Acids Res,* 43**,** D593-8.

STROBER, W., MURRAY, P. J., KITANI, A. & WATANABE, T. 2006. Signalling pathways and molecular interactions of NOD1 and NOD2. *Nature Reviews Immunology,* 6**,** 9-20.

STROBER, W. & WATANABE, T. 2011. NOD2, an intracellular innate immune sensor involved in host defense and Crohn's disease. *Mucosal Immunology,* 4**,** 484-495.

STRONG, S., STEELE, S. R., BOUTROUS, M., BORDINEAU, L., CHUN, J., STEWART, D. B., VOGEL, J., et al. 2015. Clinical Practice Guideline for the Surgical Management of Crohn's Disease. *Diseases of the Colon & Rectum,* 58.

STROUS, M., KRAFT, B., BISDORF, R. & TEGETMEYER, H. E. 2012. The binning of metagenomic contigs for microbial physiology of mixed cultures. *Front Microbiol,* 3**,** 410.

STURNIOLO, G. C., DI LEO, V., FERRONATO, A., D'ODORICO, A. & D'INCÀ, R. 2001. Zinc supplementation tightens "leaky gut" in Crohn's disease. *Inflamm Bowel Dis,* 7**,** 94-8.

SU, W., CHEN, Y., CAO, P., CHEN, Y., GUO, Y., WANG, S. & DONG, W. 2020. Fusobacterium nucleatum Promotes the Development of Ulcerative Colitis by Inducing the Autophagic Cell Death of Intestinal Epithelial. *Front Cell Infect Microbiol,* 10**,** 594806.

SUBRAMANIAN, S., RHODES, J. M., HART, C. A., TAM, B., ROBERTS, C. L., SMITH, S. L., CORKILL, J. E., et al. 2008. Characterization of epithelial IL-8 response to inflammatory bowel disease mucosal E. coli and its inhibition by mesalamine. *Inflamm Bowel Dis,* 14**,** 162-75.

SULTAN, M., RAO, A., ELPELEG, O., VAZ, F. M., ABU-LIBDEH, B., KARPEN, S. J. & DAWSON, P. A. 2018. Organic solute transporter-β (SLC51B) deficiency in two brothers with congenital diarrhea and features of cholestasis. *Hepatology,* 68**,** 590-598.

SUN, Y., YUAN, S., CHEN, X., SUN, J., KALLA, R., YU, L., WANG, L., et al. 2023. The Contribution of Genetic Risk and Lifestyle Factors in the Development of Adult-Onset Inflammatory Bowel Disease: A Prospective Cohort Study. *Am J Gastroenterol,* 118**,** 511-522.

SWIDSINSKI, A., LOENING-BAUCKE, V., BENGMARK, S., LOCHS, H. & DORFFEL, Y. 2007. Azathioprine and mesalazine-induced effects on the mucosal flora in patients with IBD colitis. *Inflammatory Bowel Diseases,* 13**,** 51-56.

TAILFORD, L. E., OWEN, C. D., WALSHAW, J., CROST, E. H., HARDY-GODDARD, J., LE GALL, G., DE VOS, W. M., et al. 2015. Discovery of intramolecular trans-sialidases in human gut microbiota suggests novel mechanisms of mucosal adaptation. *Nat Commun,* 6**,** 7624.

TAKAISHI, H., MATSUKI, T., NAKAZAWA, A., TAKADA, T., KADO, S., ASAHARA, T., KAMADA, N., et al. 2008. Imbalance in intestinal microflora constitution could be involved in the pathogenesis of inflammatory bowel disease. *Int J Med Microbiol,* 298**,** 463-72.

TAKENAKA, K., FUJII, T., KAWAMOTO, A., SUZUKI, K., SHIMIZU, H., MAEYASHIKI, C., YAMAJI, O., et al. 2022. Deep neural network for video colonoscopy of ulcerative colitis: a cross-sectional study. *Lancet Gastroenterol Hepatol,* 7**,** 230-237.

TAKEUCHI, K., SMALE, S., PREMCHAND, P., MAIDEN, L., SHERWOOD, R., THJODLEIFSSON, B., BJORNSSON, E., et al. 2006. Prevalence and mechanism of nonsteroidal anti-inflammatory drug-induced clinical relapse in patients with inflammatory bowel disease. *Clin Gastroenterol Hepatol,* 4**,** 196-202.

TAMBURINI, S., SHEN, N., WU, H. C. & CLEMENTE, J. C. 2016. The microbiome in early life: implications for health outcomes. *Nat Med,* 22**,** 713-22.

TAN, G., OPITZ, L., SCHLAPBACH, R. & REHRAUER, H. 2019. Long fragments achieve lower base quality in Illumina paired-end sequencing. *Scientific reports,* 9**,** 2856-2856.

TANCA, A., PALOMBA, A., FRAUMENE, C., PAGNOZZI, D., MANGHINA, V., DELIGIOS, M., MUTH, T., et al. 2016. The impact of sequence database choice on metaproteomic results in gut microbiota studies. *Microbiome,* 4**,** 51.

TARCA, A. L., CAREY, V. J., CHEN, X.-W., ROMERO, R. & DRĂGHICI, S. 2007. Machine Learning and Its Applications to Biology. *PLOS Computational Biology,* 3**,** e116.

TATUSOV, R. L., FEDOROVA, N. D., JACKSON, J. D., JACOBS, A. R., KIRYUTIN, B., KOONIN, E. V., KRYLOV, D. M., et al. 2003. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics,* 4**,** 41.

TATUSOV, R. L., GALPERIN, M. Y., NATALE, D. A. & KOONIN, E. V. 2000. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic acids research,* 28**,** 33-36.

TEDJO, D. I., SMOLINSKA, A., SAVELKOUL, P. H., MASCLEE, A. A., VAN SCHOOTEN, F. J., PIERIK, M. J., PENDERS, J., et al. 2016. The fecal microbiota as a biomarker for disease activity in Crohn's disease. *Sci Rep,* 6**,** 35216.

TESSLER, M., NEUMANN, J. S., AFSHINNEKOO, E., PINEDA, M., HERSCH, R., VELHO, L. F. M., SEGOVIA, B. T., et al. 2017. Large-scale differences in microbial biodiversity discovery between 16S amplicon and shotgun sequencing. *Sci Rep,* 7**,** 6589.

THIA, K. T., LOFTUS, E. V., SANDBORN, W. J. & YANG, S.-K. 2008. An Update on the Epidemiology of Inflammatory Bowel Disease in Asia. *The American Journal of Gastroenterology*.

THIA, K. T., MAHADEVAN, U., FEAGAN, B. G., WONG, C., COCKERAM, A., BITTON, A., BERNSTEIN, C. N., et al. 2009. Ciprofloxacin or metronidazole for the treatment of perianal fistulas in patients with Crohn's disease: a randomized, double-blind, placebo-controlled pilot study. *Inflamm Bowel Dis,* 15**,** 17-24.

THIA, K. T., SANDBORN, W. J., HARMSEN, W. S., ZINSMEISTER, A. R. & LOFTUS, E. V., JR. 2010. Risk factors associated with progression to intestinal complications of Crohn's disease in a population-based cohort. *Gastroenterology,* 139**,** 1147-55.

THIELE, C. & HIRSCHFELD, G. 2021. cutpointr: Improved Estimation and Validation of Optimal Cutpoints in R. *Journal of Statistical Software,* 98**,** 1 - 27.

THOMAS, C., PELLICCIARI, R., PRUZANSKI, M., AUWERX, J. & SCHOONJANS, K. 2008. Targeting bile-acid signalling for metabolic diseases. *Nat Rev Drug Discov,* 7**,** 678-93.

THOMAS, J. P., MODOS, D., RUSHBROOK, S. M., POWELL, N. & KORCSMAROS, T. 2022. The Emerging Role of Bile Acids in the Pathogenesis of Inflammatory Bowel Disease. *Front Immunol,* 13**,** 829525.

THRIENE, K. & MICHELS, K. B. 2023. Human Gut Microbiota Plasticity throughout the Life Course. *Int J Environ Res Public Health,* 20.

THURSBY, E. & JUGE, N. 2017. Introduction to the human gut microbiota. *Biochem J,* 474**,** 1823-1836.

TICINESI, A., MILANI, C., LAURETANI, F., NOUVENNE, A., MANCABELLI, L., LUGLI, G. A., TURRONI, F., et al. 2017. Gut microbiota composition is associated with polypharmacy in elderly hospitalized patients. *Sci Rep,* 7**,** 11102.

TIKHONOV, M., LEACH, R. W. & WINGREEN, N. S. 2015. Interpreting 16S metagenomic data without clustering to achieve sub-OTU resolution. *The ISME Journal,* 9**,** 68-80.

TJOA, E. & GUAN, C. 2021. A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. *IEEE Transactions on Neural Networks and Learning Systems,* 32**,** 4793-4813.

TOLHURST, G., HEFFRON, H., LAM, Y. S., PARKER, H. E., HABIB, A. M., DIAKOGIANNAKI, E., CAMERON, J., et al. 2012. Short-chain fatty acids stimulate glucagon-like peptide-1 secretion via the G-protein-coupled receptor FFAR2. *Diabetes,* 61**,** 364-71.

TOURLOUSSE, D. M., YOSHIIKE, S., OHASHI, A., MATSUKURA, S., NODA, N. & SEKIGUCHI, Y. 2017. Synthetic spike-in standards for high-throughput 16S rRNA gene amplicon sequencing. *Nucleic Acids Res,* 45**,** e23.

TRIANTAFYLLIDI, A., XANTHOS, T., PAPALOIS, A. & TRIANTAFILLIDIS, J. K. 2015. Herbal and plant therapy in patients with inflammatory bowel disease. *Ann Gastroenterol,* 28**,** 210-220.

TRINGE, S. G. & HUGENHOLTZ, P. 2008. A renaissance for the pioneering 16S rRNA gene. *Curr Opin Microbiol,* 11**,** 442-6.

TROPINI, C., MOSS, E. L., MERRILL, B. D., NG, K. M., HIGGINBOTTOM, S. K., CASAVANT, E. P., GONZALEZ, C. G., et al. 2018. Transient Osmotic Perturbation Causes Long-Term Alteration to the Gut Microbiota. *Cell,* 173**,** 1742-1754.e17.

TRUONG, D. T., FRANZOSA, E. A., TICKLE, T. L., SCHOLZ, M., WEINGART, G., PASOLLI, E., TETT, A., et al. 2015. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat Methods,* 12**,** 902-3.

TRUONG, D. T., TETT, A. & PASOLLI, E. 2017. Microbial strain-level population structure and genetic diversity from metagenomes. 27**,** 626-638.

TURNBAUGH, P. J., LEY, R. E., HAMADY, M., FRASER-LIGGETT, C. M., KNIGHT, R. & GORDON, J. I. 2007. The human microbiome project. *Nature,* 449**,** 804-810.

TURNER, D., LEVINE, A., KOLHO, K. L., SHAOUL, R. & LEDDER, O. 2014. Combination of oral antibiotics may be effective in severe pediatric ulcerative colitis: a preliminary report. *J Crohns Colitis,* 8**,** 1464-70.

TURPIN, W., GOETHEL, A., BEDRANI, L. & CROITORU MDCM, K. 2018. Determinants of IBD Heritability: Genes, Bugs, and More. *Inflamm Bowel Dis,* 24**,** 1133-1148.

TUTEJA, S. & FERGUSON, J. F. 2019. Gut Microbiome and Response to Cardiovascular Drugs. *Circulation: Genomic and Precision Medicine,* 12**,** e002314.

UGARTE, A., VICEDOMINI, R., BERNARDES, J. & CARBONE, A. 2018. A multi-source domain annotation pipeline for quantitative metagenomic and metatranscriptomic functional profiling. *Microbiome,* 6**,** 149.

UNGAR, B. & KOPYLOV, U. 2016. Advances in the development of new biologics in inflammatory bowel disease. *Ann Gastroenterol,* 29**,** 243-8.

UNGARO, F., MASSIMINO, L., D'ALESSIO, S. & DANESE, S. 2019. The gut virome in inflammatory bowel disease pathogenesis: From metagenomics to novel therapeutic approaches. *United European Gastroenterol J,* 7**,** 999-1007.

UNGARO, R., BERNSTEIN, C. N., GEARRY, R., HVIID, A., KOLHO, K. L., KRONMAN, M. P., SHAW, S., et al. 2014. Antibiotics associated with increased risk of new-onset Crohn's disease but not ulcerative colitis: a meta-analysis. *Am J Gastroenterol,* 109**,** 1728-38.

UNIPROT, C. 2008. The universal protein resource (UniProt). *Nucleic acids research,* 36**,** D190-D195.

UNIPROT, C. 2014. Activities at the Universal Protein Resource (UniProt). *Nucleic acids research,* 42**,** D191-D198.

UPADHYAY, K. G., DESAI, D. C., ASHAVAID, T. F. & DHERAI, A. J. 2023. Microbiome and metabolome in inflammatory bowel disease. *J Gastroenterol Hepatol,* 38**,** 34-43.

UTTURKAR, S. M., KLINGEMAN, D. M., LAND, M. L., SCHADT, C. W., DOKTYCZ, M. J., PELLETIER, D. A. & BROWN, S. D. 2014. Evaluation and validation of de novo and hybrid assembly techniques to derive high-quality genome sequences. *Bioinformatics,* 30**,** 2709-16.

VAN DER LUGT, B., VAN BEEK, A. A., AALVINK, S., MEIJER, B., SOVRAN, B., VERMEIJ, W. P., BRANDT, R. M. C., et al. 2019. Akkermansia muciniphila ameliorates the age-related decline in colonic mucus thickness and attenuates immune activation in accelerated aging Ercc1 (-/Delta7) mice. *Immun Ageing,* 16**,** 6.

VAN GENNEP, S., EVERS, S. W., RIETDIJK, S. T., GIELEN, M. E., DE BOER, N. K. H., GECSE, K. B., PONSIOEN, C. I. J., et al. 2021. High Disease Burden Drives Indirect Costs in Employed Inflammatory Bowel Disease Patients: The WORK-IBD Study. *Inflamm Bowel Dis,* 27**,** 352-363.

VAN KLINKEN, B. J., DEKKER, J., BÜLLER, H. A. & EINERHAND, A. W. 1995. Mucin gene structure and expression: protection vs. adhesion. *Am J Physiol,* 269**,** G613-27.

VANGAY, P., JOHNSON, A. J., WARD, T. L., AL-GHALITH, G. A., SHIELDS-CUTLER, R. R., HILLMANN, B. M., LUCAS, S. K., et al. 2018. US Immigration Westernizes the Human Gut Microbiome. *Cell,* 175**,** 962-972.e10.

VANTRAPPEN, G., GHOOS, Y., RUTGEERTS, P. & JANSSENS, J. 1977. Bile acid studies in uncomplicated Crohn's disease. *Gut,* 18**,** 730-5.

VAVASSORI, P., MENCARELLI, A., RENGA, B., DISTRUTTI, E. & FIORUCCI, S. 2009. The Bile Acid Receptor FXR Is a Modulator of Intestinal Innate Immunity1. *The Journal of Immunology,* 183**,** 6251-6261.

VAVRICKA, S. R., SPIGAGLIA, S. M., ROGLER, G., PITTET, V., MICHETTI, P., FELLEY, C., MOTTET, C., et al. 2012. Systematic Evaluation of Risk Factors for Diagnostic Delay in Inflammatory Bowel Disease. *Inflammatory Bowel Diseases,* 18**,** 496-505.

VAZEILLE, E., BUISSON, A., BRINGER, M. A., GOUTTE, M., OUCHCHANE, L., HUGOT, J. P., DE VALLEE, A., et al. 2015. Monocyte-derived macrophages from Crohn's disease patients are impaired in the ability to control intracellular adherent-invasive Escherichia coli and exhibit disordered cytokine secretion profile. *J Crohns Colitis,* 9**,** 410-20.

VAZQUEZ-BAEZA, Y., GONZALEZ, A., XU, Z. Z., WASHBURNE, A., HERFARTH, H. H., SARTOR, R. B. & KNIGHT, R. 2018. Guiding longitudinal sampling in IBD cohorts. *Gut,* 67**,** 1743-1745.

VELÁZQUEZ, M., DAVIES, C., MARETT, R., SLAVIN, J. L. & FEIRTAG, J. M. 2000. Effect of Oligosaccharides and Fibre Substitutes on Short-chain Fatty Acid Production by Human Faecal Microflora. *Anaerobe,* 6**,** 87-92.

VELLIDO, A. 2020. The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural Computing and Applications,* 32**,** 18069-18083.

VELTKAMP, C., TONKONOGY, S. L., DE JONG, Y. P., ALBRIGHT, C., GRENTHER, W. B., BALISH, E., TERHORST, C., et al. 2001. Continuous stimulation by normal luminal bacteria is essential for the development and perpetuation of colitis in Tg(epsilon26) mice. *Gastroenterology,* 120**,** 900-13.

VENTER, J. C., REMINGTON, K., HEIDELBERG, J. F., HALPERN, A. L., RUSCH, D., EISEN, J. A., WU, D., et al. 2004. Environmental Genome Shotgun Sequencing of the Sargasso Sea. *Science,* 304**,** 66.

VERMEIRE, S., VAN ASSCHE, G. & RUTGEERTS, P. 2004. C-reactive protein as a marker for inflammatory bowel disease. *Inflamm Bowel Dis,* 10**,** 661-5.

VERNIA, P., CAPRILLI, R., LATELLA, G., BARBETTI, F., MAGLIOCCA, F. M. & CITTADINI, M. 1988. Fecal lactate and ulcerative colitis. *Gastroenterology,* 95**,** 1564-8.

VERSTOCKT, B., BRESSLER, B., MARTINEZ-LOZANO, H., MCGOVERN, D. & SILVERBERG, M. S. 2022. Time to Revisit Disease Classification in Inflammatory Bowel Disease: Is the Current Classification of Inflammatory Bowel Disease Good Enough for Optimal Clinical Management? *Gastroenterology,* 162**,** 1370-1382.

VETROVSKY, T. & BALDRIAN, P. 2013. The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses. *PLoS One,* 8**,** e57923.

VICH VILA, A., COLLIJ, V., SANNA, S., SINHA, T., IMHANN, F., BOURGONJE, A. R., MUJAGIC, Z., et al. 2020. Impact of commonly used drugs on the composition and metabolic function of the gut microbiota. *Nature Communications,* 11**,** 362.

VICH VILA, A., HU, S., ANDREU-SÁNCHEZ, S., COLLIJ, V., JANSEN, B. H., AUGUSTIJN, H. E., BOLTE, L. A., et al. 2023. Faecal Metabolome and Its Determinants in Inflammatory Bowel Disease. *Gut*.

VICH VILA, A., IMHANN, F., COLLIJ, V., JANKIPERSADSING, S. A., GURRY, T., MUJAGIC, Z., KURILSHIKOV, A., et al. 2018. Gut microbiota composition and functional changes in inflammatory bowel disease and irritable bowel syndrome. *Science Translational Medicine,* 10**,** eaap8914.

VIEIRA-SILVA, S., FALONY, G., BELDA, E., NIELSEN, T., ARON-WISNEWSKY, J., CHAKAROUN, R., FORSLUND, S. K., et al. 2020. Statin therapy is associated with lower prevalence of gut microbiota dysbiosis. *Nature,* 581**,** 310-315.

VIJAYVARGIYA, P. & CAMILLERI, M. 2019. Current Practice in the Diagnosis of Bile Acid Diarrhea. *Gastroenterology,* 156**,** 1233-1238.

VÍTEK, L. 2015. Bile acid malabsorption in inflammatory bowel disease. *Inflamm Bowel Dis,* 21**,** 476-83.

VOLLMERS, J., WIEGAND, S. & KASTER, A.-K. 2017. Comparing and Evaluating Metagenome Assembly Tools from a Microbiologist's Perspective - Not Only Size Matters! *PLOS ONE,* 12**,** e0169662.

VON LAMPE, F. 2021. *goeveg R-package* [Online]. github. Available: https://github.com/fvlampe/goeveg.

WALKER, A. W., INCE, J., DUNCAN, S. H., WEBSTER, L. M., HOLTROP, G., ZE, X., BROWN, D., et al. 2011. Dominant and diet-responsive groups of bacteria within the human colonic microbiota. *ISME Journal,* 5**,** 220-230.

WALLACE, J. L. & MILLER, M. J. 2000. Nitric oxide in mucosal defense: a little goes a long way. *Gastroenterology,* 119**,** 512-20.

WALSH, A. M., CRISPIE, F., O'SULLIVAN, O., FINNEGAN, L., CLAESSON, M. J. & COTTER, P. D. 2018. Species classifier choice is a key consideration when analysing low-complexity food microbiome data. *Microbiome,* 6**,** 50.

WALTERS, J. R. F., TASLEEM, A. M., OMER, O. S., BRYDON, W. G., DEW, T. & LE ROUX, C. W. 2009. A New Mechanism for Bile Acid Diarrhea: Defective Feedback Inhibition of Bile Acid Biosynthesis. *Clinical Gastroenterology and Hepatology,* 7**,** 1189-1194.

WANG, D., HARTMANN, K., SEWERYN, M. & SADEE, W. 2018. Interactions Between Regulatory Variants in CYP7A1 (Cholesterol 7α-Hydroxylase) Promoter and Enhancer Regions Regulate CYP7A1 Expression. *Circ Genom Precis Med,* 11**,** e002082.

WANG, J. & XIE, G. 2017. Bile Acid–microbiota Crosstalk in Gastrointestinal Inflammation and Carcinogenesis. *Nature Reviews Gastroenterology & Hepatology*.

WANG, Q., GARRITY, G. M., TIEDJE, J. M. & COLE, J. R. 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol,* 73**,** 5261-7.

WANG, R., LI, Z., LIU, S. & ZHANG, D. 2023. Global, regional and national burden of inflammatory bowel disease in 204 countries and territories from 1990 to 2019: a systematic analysis based on the Global Burden of Disease Study 2019. *BMJ Open,* 13**,** e065186.

WANG, X.-W. & LIU, Y.-Y. 2020. Comparative study of classifiers for human microbiome data. *Medicine in Microecology,* 4**,** 100013.

WANG, Y., GAO, X., ZHANG, X., XIAO, F., HU, H., LI, X., DONG, F., et al. 2021. Microbial and metabolic features associated with outcome of infliximab therapy in pediatric Crohn's disease. *Gut Microbes,* 13**,** 1-18.

WARINNER, C., HERBIG, A., MANN, A., FELLOWS YATES, J. A., WEIß, C. L., BURBANO, H. A., ORLANDO, L., et al. 2017. A Robust Framework for Microbial Archaeology. *Annual Review of Genomics and Human Genetics,* 18**,** 321-356.

WARNER, B., JOHNSTON, E., ARENAS-HERNANDEZ, M., MARINAKI, A., IRVING, P. & SANDERSON, J. 2018. A practical guide to thiopurine prescribing and monitoring in IBD. *Frontline Gastroenterol,* 9**,** 10-15.

WEDLAKE, L., A'HERN, R., RUSSELL, D., THOMAS, K., WALTERS, J. R. & ANDREYEV, H. J. 2009. Systematic review: the prevalence of idiopathic bile acid malabsorption as diagnosed by SeHCAT scanning in patients with diarrhoea-predominant irritable bowel syndrome. *Aliment Pharmacol Ther,* 30**,** 707-17.

WEERSMA, R. K., ZHERNAKOVA, A. & FU, J. 2020. Interaction between drugs and the gut microbiome. *Gut,* 69**,** 1510-1519.

WEHKAMP, J., HARDER, J., WEICHENTHAL, M., MUELLER, O., HERRLINGER, K. R., FELLERMANN, K., SCHROEDER, J. M., et al. 2003. Inducible and constitutive beta-defensins are differentially expressed in Crohn's disease and ulcerative colitis. *Inflamm Bowel Dis,* 9**,** 215-23.

WEHKAMP, J., HARDER, J., WEICHENTHAL, M., SCHWAB, M., SCHÄFFELER, E., SCHLEE, M., HERRLINGER, K. R., et al. 2004. NOD2 (CARD15) mutations in Crohn's disease are associated with diminished mucosal alpha-defensin expression. *Gut,* 53**,** 1658-64.

WEISMAN, D., YASUDA, M. & BOWEN, J. L. 2013. FunFrame: functional gene ecological analysis pipeline. *Bioinformatics,* 29**,** 1212-1214.

WEISSHOF, R. & CHERMESH, I. 2015. Micronutrient deficiencies in inflammatory bowel disease. *Curr Opin Clin Nutr Metab Care,* 18**,** 576-81.

WÉRA, O., LANCELLOTTI, P. & OURY, C. 2016. The Dual Role of Neutrophils in Inflammatory Bowel Diseases. *J Clin Med,* 5.

WERNER, J. J., KOREN, O., HUGENHOLTZ, P., DESANTIS, T. Z., WALTERS, W. A., CAPORASO, J. G., ANGENENT, L. T., et al. 2012. Impact of training sets on classification of high-throughput bacterial 16s rRNA gene surveys. *ISME J,* 6**,** 94-103.

WESTCOTT, S. L. & SCHLOSS, P. D. 2015. De novo clustering methods outperform reference-based methods for assigning 16S rRNA gene sequences to operational taxonomic units. *PeerJ,* 3**,** e1487.

WETTERSTRAND, K. A. 2022. *The Cost of Sequencing a Human Genome* [Online]. National Human Genome Research Institute. Available: https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost [Accessed].

WHIPPS, J. M., LEWIS, K. & COOKE, R. 1988. Mycoparasitism and plant disease control. Manchester University Press Manchester.

WHO Collaborating Centre for Drug Statistics Methodology 2021. Guidelines for ATC classification and DDD assignment. Norwegian Institute of Public Health.

WICK, E. C. & SEARS, C. L. 2010. Bacteroides spp. and diarrhea. *Curr Opin Infect Dis,* 23**,** 470-4.

WICKHAM, H. 2016. *ggplot2: Elegant Graphics for Data Analysis*, Springer-Verlag New York.

WIKOFF, W. R., ANFORA, A. T., LIU, J., SCHULTZ, P. G., LESLEY, S. A., PETERS, E. C. & SIUZDAK, G. 2009. Metabolomics analysis reveals large effects of gut microflora on mammalian blood metabolites. *Proceedings of the National Academy of Sciences,* 106**,** 3698-3703.

WILKE, A., BISCHOF, J., GERLACH, W., GLASS, E., HARRISON, T., KEEGAN, K. P., PACZIAN, T., et al. 2016. The MG-RAST metagenomics database and portal in 2015. *Nucleic acids research,* 44**,** D590-D594.

WILKINS, T., JARVIS, K. & PATEL, J. 2011. Diagnosis and management of Crohn's disease. *Am Fam Physician,* 84**,** 1365-75.

WILKINSON, M. D., DUMONTIER, M., AALBERSBERG, I. J., APPLETON, G., AXTON, M., BAAK, A., BLOMBERG, N., et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data,* 3**,** 160018.

WILLING, B. P., DICKSVED, J., HALFVARSON, J., ANDERSSON, A. F., LUCIO, M., ZHENG, Z., JÄRNEROT, G., et al. 2010. A pyrosequencing study in twins shows that gastrointestinal microbial profiles vary with inflammatory bowel disease phenotypes. *Gastroenterology,* 139**,** 1844-1854. e1.

WILSON, A., WANG, Q., ALMOUSA, A. A., JANSEN, L. E., CHOI, Y.-H., SCHWARZ, U. I. & KIM, R. B. 2020. Genetic variation in the farnesoid X-receptor predicts Crohn's disease severity in female patients. *Scientific Reports,* 10**,** 11725.

WIRBEL, J., ZYCH, K., ESSEX, M., KARCHER, N., KARTAL, E., SALAZAR, G., BORK, P., et al. 2021. Microbiome meta-analysis and cross-disease comparison enabled by the SIAMCAT machine learning toolbox. *Genome Biol,* 22**,** 93.

WLODARSKA, M., LUO, C., KOLDE, R., D'HENNEZEL, E., ANNAND, J. W., HEIM, C. E., KRASTEL, P., et al. 2017. Indoleacrylic Acid Produced by Commensal Peptostreptococcus Species Suppresses Inflammation. *Cell Host Microbe,* 22**,** 25-37.e6.

WOLPERT, D. H. 2002. The Supervised Learning No-Free-Lunch Theorems. *In:* ROY, R., KÖPPEN, M., OVASKA, S., FURUHASHI, T. & HOFFMANN, F. (eds.) *Soft Computing and Industry: Recent Applications.* London: Springer London.

WONG, C. P., RINALDI, N. A. & HO, E. 2015. Zinc deficiency enhanced inflammatory response by increasing immune cell activation and inducing IL6 promoter demethylation. *Mol Nutr Food Res,* 59**,** 991-9.

WOOD, D. E. & SALZBERG, S. L. 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol,* 15**,** R46.

WRIGHT, E. S., YILMAZ, L. S. & NOGUERA, D. R. 2012. DECIPHER, a search-based approach to chimera identification for 16S rRNA sequences. *Appl Environ Microbiol,* 78**,** 717-25.

WU, G. D., CHEN, J., HOFFMANN, C., BITTINGER, K., CHEN, Y. Y., KEILBAUGH, S. A., BEWTRA, M., et al. 2011. Linking long-term dietary patterns with gut microbial enterotypes. *Science,* 334**,** 105-8.

WU, Y. W., SIMMONS, B. A. & SINGER, S. W. 2016. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics,* 32**,** 605-7.

XAVIER, R. J. & PODOLSKY, D. K. 2007. Unravelling the pathogenesis of inflammatory bowel disease. *Nature,* 448**,** 427-434.

XU, J. 2016. Fungal DNA barcoding. *Genome,* 59**,** 913-932.

XU, J., BJURSELL, M. K., HIMROD, J., DENG, S., CARMICHAEL, L. K., CHIANG, H. C., HOOPER, L. V., et al. 2003. A genomic view of the human-Bacteroides thetaiotaomicron symbiosis. *Science,* 299**,** 2074-6.

XU, J., CHEN, N., WU, Z., SONG, Y., ZHANG, Y., WU, N., ZHANG, F., et al. 2018. 5-Aminosalicylic Acid Alters the Gut Bacterial Microbiota in Patients With Ulcerative Colitis. *Frontiers in Microbiology,* 9.

XU, L., LOCHHEAD, P., KO, Y., CLAGGETT, B., LEONG, R. W. & ANANTHAKRISHNAN, A. N. 2017. Systematic review with meta-analysis: breastfeeding and the risk of Crohn's disease and ulcerative colitis. *Aliment Pharmacol Ther,* 46**,** 780-789.

XUE, L., HUANG, Z., ZHOU, X. & CHEN, W. 2012. The possible effects of mesalazine on the intestinal microbiota. *Alimentary pharmacology & therapeutics,* 36**,** 813-814.

XUE, L. N., XU, K. Q., ZHANG, W., WANG, Q., WU, J. & WANG, X. Y. 2013. Associations between vitamin D receptor polymorphisms and susceptibility to ulcerative colitis and Crohn's disease: a meta-analysis. *Inflamm Bowel Dis,* 19**,** 54-60.

XUE, Z., KABLE, M. E. & MARCO, M. L. 2018. Impact of DNA Sequencing and Analysis Methods on 16S rRNA Gene Bacterial Community Analysis of Dairy Products. *mSphere,* 3.

YANG, B., WANG, Y. & QIAN, P. Y. 2016. Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis. *BMC Bioinformatics,* 17**,** 135.

YANG, L., LUO, H., TAN, D., ZHANG, S., ZHONG, Z., WANG, S., VONG, C. T., et al. 2021. A recent update on the use of Chinese medicine in the treatment of inflammatory bowel disease. *Phytomedicine,* 92**,** 153709.

YANG, X., CHARLEBOIS, P., GNERRE, S., COOLE, M. G., LENNON, N. J., LEVIN, J. Z., QU, J., et al. 2012. De novo assembly of highly diverse viral populations. *BMC genomics,* 13**,** 475-475.

YAP, C. X., HENDERS, A. K., ALVARES, G. A., WOOD, D. L. A., KRAUSE, L., TYSON, G. W., RESTUADI, R., et al. 2021. Autism-related dietary preferences mediate autism-gut microbiome associations. *Cell,* 184**,** 5916-5931 e17.

YASUDA, K., OH, K., REN, B., TICKLE, T. L., FRANZOSA, E. A., WACHTMAN, L. M., MILLER, A. D., et al. 2015. Biogeography of the intestinal mucosal and lumenal microbiome in the rhesus macaque. *Cell Host Microbe,* 17**,** 385-391.

YE, S. H., SIDDLE, K. J., PARK, D. J. & SABETI, P. C. 2019. Benchmarking Metagenomics Tools for Taxonomic Classification. *Cell,* 178**,** 779-794.

YILMAZ, B., JUILLERAT, P., OYAS, O., RAMON, C., BRAVO, F. D., FRANC, Y., FOURNIER, N., et al. 2019b. Microbial network disturbances in relapsing refractory Crohn's disease. *Nat Med,* 25**,** 323-336.

YILMAZ, P., PARFREY, L. W., YARZA, P., GERKEN, J., PRUESSE, E., QUAST, C., SCHWEER, T., et al. 2014. The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. *Nucleic Acids Res,* 42**,** D643-8.

YIP, A. A.-O., SMITH-ROBERGE, J., KHORASANI, S. A.-O., AUCOIN, M. A.-O. & INGALLS, B. A.-O. 2022 Calibrating spatiotemporal models of microbial communities to microscopy data: A review.

YONGHOON, L. & RINA FOYGEL, B. 2022. Binary classification with corrupted labels. *Electronic Journal of Statistics,* 16**,** 1367-1392.

YOON, C. W., TORRANCE, R. & SCHEINERMAN, N. 2021. Machine Learning in Medicine: Should the Pursuit of Enhanced Interpretability Be Abandoned? *Journal of Medical Ethics*.

YOSHIMURA, D., KAJITANI, R., GOTOH, Y., KATAHIRA, K., OKUNO, M., OGURA, Y., HAYASHI, T., et al. 2019. Evaluation of SNP calling methods for closely related bacterial isolates and a novel high-accuracy pipeline: BactSNP. *Microbial Genomics,* 5.

YOUSSEF, N., SHEIK, C. S., KRUMHOLZ, L. R., NAJAR, F. Z., ROE, B. A. & ELSHAHED, M. S. 2009. Comparison of species richness estimates obtained using nearly complete fragments and simulated pyrosequencing-generated fragments in 16S rRNA gene-based environmental surveys. *Appl Environ Microbiol,* 75**,** 5227-36.

YU, C., WANG, F., KAN, M., JIN, C., JONES, R. B., WEINSTEIN, M., DENG, C. X., et al. 2000. Elevated cholesterol metabolism and bile acid synthesis in mice lacking membrane tyrosine kinase receptor FGFR4. *Journal of Biological Chemistry,* 275**,** 15482-15489.

YU, Z. T., NANTHAKUMAR, N. N. & NEWBURG, D. S. 2016. The Human Milk Oligosaccharide 2'-Fucosyllactose Quenches Campylobacter jejuni-Induced Inflammation in Human Epithelial Cells HEp-2 and HT-29 and in Mouse Intestinal Mucosa. *J Nutr,* 146**,** 1980-1990.

ZELANTE, T., IANNITTI, R. G., CUNHA, C., DE LUCA, A., GIOVANNINI, G., PIERACCINI, G., ZECCHI, R., et al. 2013. Tryptophan catabolites from microbiota engage aryl hydrocarbon receptor and balance mucosal reactivity via interleukin-22. *Immunity,* 39**,** 372-85.

ZENEWICZ, L. A., YANCOPOULOS, G. D., VALENZUELA, D. M., MURPHY, A. J., STEVENS, S. & FLAVELL, R. A. 2008. Innate and adaptive interleukin-22 protects mice from inflammatory bowel disease. *Immunity,* 29**,** 947-57.

ZENG, M. Y., INOHARA, N. & NUÑEZ, G. 2017. Mechanisms of inflammation-driven bacterial dysbiosis in the gut. *Mucosal Immunology,* 10**,** 18-26.

ZHANG, C., ZHANG, J., ZHANG, Y., SONG, Z., BIAN, J., YI, H. & MA, Z. 2023a. Identifying neutrophil-associated subtypes in ulcerative colitis and confirming neutrophils promote colitis-associated colorectal cancer. *Front Immunol,* 14**,** 1095098.

ZHANG, F., ASCHENBRENNER, D., YOO, J. Y. & ZUO, T. 2022a. The gut mycobiome in health, disease, and clinical applications in association with the gut bacterial microbiome assembly. *The Lancet Microbe,* 3**,** e969-e983.

ZHANG, J., GUO, Y. & DUAN, L. 2022b. Features of Gut Microbiome Associated With Responses to Fecal Microbiota Transplantation for Inflammatory Bowel Disease: A Systematic Review. *Front Med (Lausanne),* 9**,** 773105.

ZHANG, L., REN, J. W., WONG, C. C., WU, W. K., REN, S. X., SHEN, J., CHAN, R. L., et al. 2012. Effects of cigarette smoke and its active components on ulcer formation and healing in the gastrointestinal mucosa. *Curr Med Chem,* 19**,** 63-9.

ZHANG, M., LIU, B., ZHANG, Y., WEI, H., LEI, Y. & ZHAO, L. 2007. Structural shifts of mucosa-associated lactobacilli and Clostridium leptum subgroup in patients with ulcerative colitis. *J Clin Microbiol,* 45**,** 496-500.

ZHANG, Q., SU, X., ZHANG, C., CHEN, W., WANG, Y., YANG, X., LIU, D., et al. 2023b. Klebsiella pneumoniae Induces Inflammatory Bowel Disease Through Caspase-11-Mediated IL18 in the Gut Epithelial Cells. *Cell Mol Gastroenterol Hepatol,* 15**,** 613-632.

ZHANG, X., SHEN, D., FANG, Z., JIE, Z., QIU, X., ZHANG, C., CHEN, Y., et al. 2013. Human gut microbiota changes reveal the progression of glucose intolerance. *PLoS One,* 8**,** e71108.

ZHANG, X., ZHANG, D., JIA, H., FENG, Q., WANG, D., LIANG, D., WU, X., et al. 2015. The oral and gut microbiomes are perturbed in rheumatoid arthritis and partly normalized after treatment. *Nature Medicine,* 21**,** 895-905.

ZHANG, Z., SCHWARTZ, S., WAGNER, L. & MILLER, W. 2000. A greedy algorithm for aligning DNA sequences. *J Comput Biol,* 7**,** 203-14.

ZHENG, Q., BARTOW-MCKENNEY, C., MEISEL, J. S. & GRICE, E. A. 2018. HmmUFOtu: An HMM and phylogenetic placement based ultra-fast taxonomic assignment and OTU picking tool for microbiome amplicon sequencing studies. *Genome Biology,* 19**,** 82.

ZHENG, X., HUANG, F., ZHAO, A., LEI, S., ZHANG, Y., XIE, G., CHEN, T., et al. 2017. Bile acid is a significant host factor shaping the gut microbiome of diet-induced obese mice. *BMC Biology,* 15**,** 120.

ZHERNAKOVA, A., KURILSHIKOV, A., BONDER, M. J., TIGCHELAAR, E. F., SCHIRMER, M., VATANEN, T., MUJAGIC, Z., et al. 2016. Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science,* 352**,** 565-9.

ZHOU, H. & HYLEMON, P. B. 2014a. Bile acids are nutrient signaling hormones. *Steroids,* 86**,** 62-68.

ZHOU, Q., SU, X., JING, G. & NING, K. 2014b. Meta-QC-Chain: Comprehensive and Fast Quality Control Method for Metagenomic Data. *Genomics, Proteomics & Bioinformatics,* 12**,** 52-56.

ZHOU, Y., XU, H., XU, J., ZHAO, H., CHEN, Y. & NIE, Y. 2020. F. Prausnitzii and Its Supernatant Increase SCFAs-producing Bacteria to Restore Gut Dysbiosis in TNBS-Induced Colitis.

ZHOU, Z.-H. 2012. *Ensemble Methods: Foundations and Algorithms*, Chapman & Hall/CRC.

ZHU, S., HAN, M., LIU, S., FAN, L., SHI, H. & LI, P. 2022. Composition and diverse differences of intestinal microbiota in ulcerative colitis patients. *Front Cell Infect Microbiol,* 12**,** 953962.

ZHU, W., LOMSADZE, A. & BORODOVSKY, M. 2010. Ab initio gene identification in metagenomic sequences. *Nucleic Acids Res,* 38**,** e132.

ZIMMERMANN, M., ZIMMERMANN-KOGADEEVA, M., WEGMANN, R. & GOODMAN, A. L. 2019. Mapping human microbiome drug metabolism by gut bacteria and their genes. *Nature,* 570**,** 462-467.

ZOCCO, M. A., DAL VERME, L. Z., CREMONINI, F., PISCAGLIA, A. C., NISTA, E. C., CANDELLI, M., NOVI, M., et al. 2006. Efficacy of Lactobacillus GG in maintaining remission of ulcerative colitis. *Aliment Pharmacol Ther,* 23**,** 1567-74.

ZOU, Y., LIN, X., XUE, W., TUO, L., CHEN, M.-S., CHEN, X.-H., SUN, C.-H., et al. 2021. Characterization and description of Faecalibacterium butyricigenerans sp. nov. and F. longum sp. nov., isolated from human faeces. *Scientific Reports,* 11**,** 11340.

ZUO, L., LI, Y., WANG, H., WU, R., ZHU, W., ZHANG, W., CAO, L., et al. 2014. Cigarette smoking is associated with intestinal barrier dysfunction in the small intestine but not in the large intestine of mice. *Journal of Crohn's and Colitis,* 8**,** 1710-1722.

ZUO, T., KAMM, M. A., COLOMBEL, J. F. & NG, S. C. 2018. Urbanization and the gut microbiota in health and inflammatory bowel disease. *Nat Rev Gastroenterol Hepatol,* 15**,** 440-452.

ZUPO, R., SILA, A., CASTELLANA, F., BRINGIOTTI, R., CURLO, M., DE PERGOLA, G., DE NUCCI, S., et al. 2022. Prevalence of Zinc Deficiency in Inflammatory Bowel Disease: A Systematic Review and Meta-Analysis. *Nutrients,* 14.

ZUURVELD, M., VAN WITZENBURG, N. P., GARSSEN, J., FOLKERTS, G., STAHL, B., VAN'T LAND, B. & WILLEMSEN, L. E. M. 2020. Immunomodulation by Human Milk Oligosaccharides: The Potential Role in Prevention of Allergic Diseases. *Frontiers in Immunology,* 11.

Acknowledgements

Finishing up this document, that is supposed to be the accumulation of what I have done in the last five (ish) years, I come to think what, or better who, brought me here. While this document might be my work and sweat and tears, I did not embark on this journey alone and there are many people that contributed and helped me along the way.

First and foremost I would like to thank the volunteers whose samples are the foundation of this PhD, without them, this work wouldn't exist. It is cumbersome at the best of times, to give your blood and (let's call it) sweat to science, but doing so while experiencing a full flare up from a at times debilitating disease, takes a lot more out of a person and I am truly grateful to all of you. I hope I did this offering justice.

Surely, the next one in line has to be my supervisor Marcus Claesson, for offering me the opportunity to pursue this PhD and keeping me funded the whole time, COVID extensions and all. He has given me the freedom to develop my understanding of the field, and also gave me the initial nudge towards machine learning for which I am very thankful, as well as for setting me up with the right people to learn from and to collaborate with. Here, I have to give a shout out to Andriy Temko, who allowed me to tag along to meetings with his actual PhD student, so I can learn more about the intricacies of machine learning and its many, many pitfalls. Your insights were invaluable. But also the people, Emilio, Ana, Alvaro and Chloe, who did the actual lab work and provided me with the sweet data to play with, can't stay unmentioned. Nor can James "Paddy" Mullan, who kept our servers running and our data safe. I appreciate you even more since I have to deal with people's unfortunate tendency to hog cores and create zombies.

And of course I have to thank all of you members of the 4.11 office and the remaining Claesson lab, present and past. It was, to say it with Jill's words, "truly delightful" working with you guys (while I also thoroughly enjoyed the *not* working part). Who else would have stopped what they were doing and go on a virtual hunt for external datasets with me? Jamie, Jill, Maria, Sydney, Rachel, Thomas, Janiki, Shriram, Kardokh and James (I'm sure I'm forgetting someone here, sorry) the interactions and discussions with you over the years taught me a lot and made this PhD journey worthwhile. It also let me hang onto my sanity in the last weeks.

Last but not least, I want to thank my family and friends, without whom I wouldn't be here, nor what I am today. I want to thank my Mum for pushing me to get my PhD and my Dad for being proud of me whatever I'm doing with my live as long as I'm happy. I would like to thank my friend Kasia for forming a pack of weirdos with me and for keeping me entertained and (borderline) sane. I also have to thank my partner, James for being my stable support throughout the years and for making me laugh, every freaking day. Thanks for running through the emotion with me in the last weeks, I couldn't have done it without you.