Data integration on inflammatory bowel disease

Lluís Revilla Sancho

2022-04-13

Contents

List of Figures

List of Tables

Preface

The main topic of the thesis is data integration applied in the inflammatory bowel disease (IBD) research. This disease is complex, for instance it is not known if the cause behind Crohn's disease and ulcerative colitis is the same. There are hypothesis pointing that the microbiome is a major factor in the disease, which together with an aberrant immune response is the dominant theory. In order to find robust relationships between the microbiome and the immune system it is important to consider all the relevant variables that influence a disease. On this thesis we seek these relationships using data from different sequencing technologies and the observed or reported phenotype of the patients.

The thesis was carried out on the Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS) research institute and funded by Centro de Investigación Biomédica en Red (CIBER). The thesis was done on the IBD unit which is a translational team of biologist, microbiologists, veterinaries, bioinformaticians, doctors and nurses (at Hospital Clínic) in a multidisciplinary team. The leading doctor of the unit was Julian Panés whose interest on the disease made possible this thesis.

The thesis is on the doctoral programme in biomedicine of the University of Barcelona (UB). My thesis directors' were Juanjo Lozano and Azucena Salas, my boss, who helped as bioinformatician and disease expert respectively. They provided advice and guidance on how to analyse the data and where to focus on the different experiments/analysis.

Abstracts

English

Introduction: Inflammatory bowel disease is a complex intestinal disease with several genetic and environmental factors that can influence its course. The ethiology and pathophysiology of the disease is not fully understood. There is some evidence that microbiome can play a role. Finding relationships between microbiome and host's mucosa could help advance prevention, diagnosis or treatment.

Methods: We based our analysis on intestinal bacterial 16S rRNA and human transcriptome data from biopsies from multiple timepoints and intestine segments. We extended regularized generalized canonical correlation analysis to find models that are coherent with previous knowledge on the disease taking into account the samples' information. Multiple inflammatory bowel disease datasets on different treatments and conditions were analysed and the models defining those dataset were compared. The results were compared with multiple co-inertia analysis.

Results: Splitting sample variables into different blocks results in models of these relationships that show differences on the genes and microorganisms selected. The models generated using our new method inteRmodel outperformed multiple co-inertia analysis to classify the samples according to their location. Despite being used on datasets of different sources the resulting models show similar relationships between variables.

Discussion: Comparing multiple models helps find out the relationships within datasets. Our method finds how strong are the relationships between the microbiome, transcriptome and environmental variables. On different datasets genes selected were common. This approach is robust and flexible to different datasets and settings.

Conclusion: With inteRmodel we found that the microbiome relates more closely to the sample location than to disease, but the transcriptome is highly related to the location of the sample on the intestine. There is a common transcriptome between datasets while microorganisms depend of the dataset. We can improve sample classification by taking into account both bacterial 16S rRNA and host transcriptome.

Spanish

Introducción: La enfermedad inflamatoria intestinal es una enfermedad intestinal compleja con factores genéticos y ambientales que pueden influir en su curso. La etiología y la fisiopatología de la enfermedad no se conocen por completo. Existen evidencias que el microbioma puede desempeñar un papel relevante. Encontrar relaciones entre el microbioma y la mucosa del huésped podría ayudar a avanzar en la prevención, el diagnóstico o el tratamiento.

Métodos: Basamos nuestro análisis en el ARNr 16S bacteriano intestinal y en datos de transcriptomas humanos de biopsias de múltiples puntos temporales y segmentos

intestinales. Extendimos el análisis de correlación canónica generalizada regularizado para encontrar modelos coherentes con el conocimiento previo sobre la enfermedad teniendo en cuenta la información de las muestras. Se analizaron múltiples conjuntos de datos de enfermedad inflamatoria intestinal en diferentes tratamientos y condiciones y se compararon los modelos que definen esos conjuntos de datos. Los resultados se compararon con análisis de coinercia múltiple.

Resultados: Dividir las variables de la muestra en diferentes bloques resulta en modelos de estas relaciones que muestran diferencias en los genes y microorganismos seleccionados. Los modelos generados con nuestro nuevo método, interRmodel, superaron el análisis de múltiples coinercias para clasificar las muestras según su ubicación. A pesar de ser utilizados en conjuntos de datos de diferentes fuentes, los modelos resultantes muestran unas relaciones similares entre las variables.

Discusión: La comparación de varios modelos ayuda a descubrir las relaciones dentro de los conjuntos de datos. Nuestro método encuentra cuán fuertes son las relaciones entre el microbioma, el transcriptoma y las variables ambientales. En diferentes conjuntos de datos, los genes seleccionados eran comunes. Este enfoque es robusto y flexible para diferentes conjuntos de datos y configuraciones.

Conclusión: Con inteRmodel descubrimos que el microbioma se relaciona más estrechamente con la ubicación de la muestra que con la enfermedad, pero el transcriptoma está muy relacionado con la ubicación de la muestra en el intestino. Hay un transcriptoma común entre los conjuntos de datos, mientras que los microorganismos dependen del conjunto de datos. Podemos mejorar la clasificación de las muestras teniendo en cuenta tanto el ARNr 16S bacteriano como el transcriptoma del huésped.

Catalan

Introducció: La malaltia inflamatòria intestinal és una malaltia intestinal complexa amb diversos factors genètics i ambientals que poden influir en el seu curs. L'etiologia i fisiopatologia de la malaltia no es coneix del tot. Hi ha evidències que el microbioma pot tenir un paper rellevant. Trobar relacions entre el microbioma i la mucosa de l'hoste podria ajudar a avançar en la prevenció, el diagnòstic o el tractament.

Mètodes: Vam basar la nostra anàlisi en dades d'ARNr 16S bacteriana intestinal i de transcriptoma humà de biòpsies de múltiples punts de temps i segments intestinals. Hem ampliat l'anàlisi de correlació canònica generalitzada regularitzada per trobar models coherents amb el coneixement previ sobre la malaltia tenint en compte la informació de les mostres. Es van analitzar diversos conjunts de dades de malaltia inflamatòria intestinal sobre diferents tractaments i condicions i es van comparar els models que defineixen aquest conjunt de dades. Els resultats es van comparar amb l'anàlisi de coinèrcia múltiple.

Resultats: Dividir les variables de la mostra en diferents blocs dóna com a resultat models d'aquestes relacions que mostren diferències en els gens i els microorganismes seleccionats. Els models generats mitjançant el nostre nou mètode intermodel van superar l'anàlisi de coinèrcia múltiple per classificar les mostres segons la seva ubicació. Tot i utilitzar-se en conjunts de dades de diferents fonts, els models resultants mostren relacions similars entre variables.

Discussió: La comparació de diversos models ajuda a esbrinar les relacions dins dels conjunts de dades. El nostre mètode troba com de fortes són les relacions entre el microbioma, el transcriptoma i les variables ambientals. En diferents conjunts de dades, els gens seleccionats eren comuns. Aquest enfocament és robust i flexible per a diferents conjunts de dades i configuracions.

Conclusió: Amb inteRmodel vam trobar que el microbioma es relaciona més estretament amb la ubicació de la mostra que amb la malaltia, però el transcriptoma està molt relacionat amb la ubicació de la mostra a l'intestí. Hi ha un transcriptoma comú entre conjunts de dades, mentre que els microorganismes depenen del conjunt de dades. Podem millorar la classificació de les mostres tenint en compte tant l'ARNr 16S bacterià com el transcriptoma hoste.

Introduction

Inflammatory bowel disease (IBD) involves Crohn's disease (CD) and ulcerative colitis (UC). It generally affects the terminal ileum and the colon but it can involve any segment of the gastrointestinal tract. UC is a recurrent, chronic and continuous inflammation of the colon and rectum while the CD is not a continuous inflammation and affects the whole gastrointestinal tract causing transmural inflammation.

IBD etiology is unknown. However, once it has initiated the most prevalent hypothesis of its chronicity suggests an aberrant immunological response to antigens of the commensal microbiome.

To diagnose IBD doctors use endoscopy and/or magnetic resonance imaging and histologies.

Treatments provided for IBD include, noninflammatory drugs, suppressors and biologics, i.e, anti TNF- anti-IL2, 23, anti-integrin $\alpha 4 \beta 7$. The therapeutic options can induce remission in some patients, but they often need continuous treatment to avoid recurrence. Nevertheless, many patients are refractory or intolerant to those therapies and need to undergo surgery or other strategies like dietetic and psychological support [1] .

1.1 Inflammatory bowel disease

IBD includes the CD and UC which are characterized by alternating periods of remission and clinical relapse that mainly affect the gastrointestinal tract. CD is a progressive relapsing disease that can affect all the gastrointestinal tract but shows mostly on both terminal ileum and colon with a discontinuous inflammation. UC is a colonic relapsing disease characterized by continuous inflammation of the colon. Both of them have different risk factors, clinical, endoscopic an histological characteristics (see sections 1.1.2 and 1.1.3).

Around 3.5 million individuals have IBD in Europe and North America combined [2]. IBD is more commonly found in industrialized and developed regions, suggesting that environmental factors might greatly influence IBD occurrence. In addition, the incidence of IBD is increasing in areas, such as Asia or Eastern Europe, where the number of cases was relatively low hitherto [3].

The dysregulation of the inflammatory response observed in IBD requires interplay between host genetic factors and the intestinal microbiome. Several studies support the concept that IBD arise from an exacerbate immune response against commensal gut microorganisms. Nonetheless, the disease could result from an imbalanced microbial composition leading to generalized or localized dysbiosis[1].

[1] A signature is usually a group of features that describe/are representative of a cell line or a

The role of the gut microbiome in IBD is an active ongoing field of research. Several authors are currently studying the alterations reported in IBD of the intestinal microbiome. However, it is still unclear the cause-effect relation between dysbiosis and IBD. Partly due to the multiple variables already identified that have been linked to IBD; for instance, age, diet, usage of antibiotic, tobacco, and socioeconomic status [4, 5].

The relationship between host and microbiome has been proposed to play a fundamental role in maintaining disease. For instance, some *Proteobacteria* species which have adherent and invasive properties might exploit host defenses and promote a proinflammatory environment, altering the intestinal microbiota in favor of dysbiosis [6].

The epithelium is often damaged and might present ulcers or other inflammation symptoms. A segment of the gastrointestinal tract might recover if the patient receives treatment or due the natural cycles of the disease. But once a segment is affected by the disease it can be considered as involved, as some damage remains even if the tissues is no longer inflammed.

1.1.1 Etiology and pathogenesis

Several mechanisms have been proposed to drive IBD pathogenesis [7, 8]. Some of them are based on a relationship between the immune system and the microbiome [9, 10]. It is also unclear if CD and the UC share the same origin considering their different symptoms.

There is also evidence of some genetic component on the onset of the disease, specially if the disease appears very early (less than 2 years old patients) [11, 12]. Disease can be classified based on age at onset as very early, early or adult on-set disease [12]. Genome-wide association studies (GWAS) have linked IBD to over 100 genetic loci, including a *NOD2* gene, but so far there is not any known mechanism how polymorfism on this genes are driving the disease [13]. On early pedriatic and adult disease the genetic component is lower than on very early on set and it is thought that the environmental factors are the main cause of the disease at those ages.

On the following sections we will explore the role of several of the possible factors involved on the pahtogenesis, starting with the genetics.

1.1.1.1 Genetics

IBD is not an heritable disease, except for very early onset IBD, but it has some genetic influence that predisposes people to have it.

This has lead to look for genetic factors on IBD both on general population and on the early cases. Genome wide association studies (GWAS) are one of the most common genetic studies performed, together with methylation studies. To discover through linkage desequilibrium genetic variations linked to phenotypes and regulatory transcription changes, respectively.

process or a stage.

With GWAS several allels on protein coding loci have been found, rising to around 300 genetic variants [14]. Particularly, the *NOD2* gene is highly relevant for the disease on European patients, as it is a risk alleles for CD loci but show significant protective effects in UC [15, 16]. The mechanism of how this gene protects from UC has not been confirmed yet [13].

Many of the relevant genetic loci related to IBD are not on protein coding fragments of the genome. Recently expression quantitative trait loci (eQTL) particularly showed [11] that locis are on enhancers or promoters like e.g. H3K27Ac or promoter e.g. H3K4me1 marks as found by chromatin immunoprecipitation sequencing (ChIP-Seq).

1.1.1.2 Microbiome

The human intestine is a large reservoir of co-existing microorganisms (bacteria, fungi, viruses, and unicellular eukaryotes) . This microbiome community exerts different functions in the human body influencing nutrients' metabolism, the maturation of the immune system while suppressing the growth of harmful microorganisms' [17].

The role of the gut microbiota has been proposed to play a role in IBD pathogenesis. IBD has been characterized by a breakdown in the balance between beneficial and harmful bacteria that are present in the human gut compared to healthy individuals [18, 19].

Indeed, many studies show that patients with IBD have less biodiversity. Biodiversity is measured on α (alfa) and β (beta) diversity. α-diversity is a measure of the species present on a single sample and its abundance while β-diversity compares the diversity between samples. There are some reports of taxonomic changes and increase on *Enterobacteriaceae sp*, *Escherichia coli* (specially the invasive strain) at the mucosal layer of IBD patients [20]. At the same time there is often a reduction of protective species like *Bifidobacterium*, *Lactobacillus* and *Faecalibacterium* which might be able to protect individuals from mucosal inflammation via several mechanism such as a downregulation of proinflammatory cytokines or the stimulation of IL-10 and antiinflammatory cytokines [20]. Specially *Faecalibacterium prausnitzii* is one microorganism of interest [21, 22].

In fact, it has been recently proposed that several unique microbial species can distinguish healthy controls from UC and CD patients [23, 24].

One of the proposed mechanism of crosstalk between bacteria and host is through bacterial metabolites. They interact with the cells and modulate the state of the intestine. One example of such metabolite is butyrate which has been linked to microorganisms presents on healthy intestines and shown to interact with intestine cells and help regulate some genes [25].

As previously mentioned, adherent invasive *Escherichia coli*, a proteobacteria specie, has been associated with IBD [26]. Adherent invasive strains are mainly found in ileal and colonic samples of CD patients and their presence in UC is less clear. These adherent invasive cells enter through the epithelium of the more permeable cells and live on their cytosol.

The metabolic cocktail composed of soluble factors secreted by life probiotic bacteria, living microorganisms which, when administered in adequate amounts, confer health

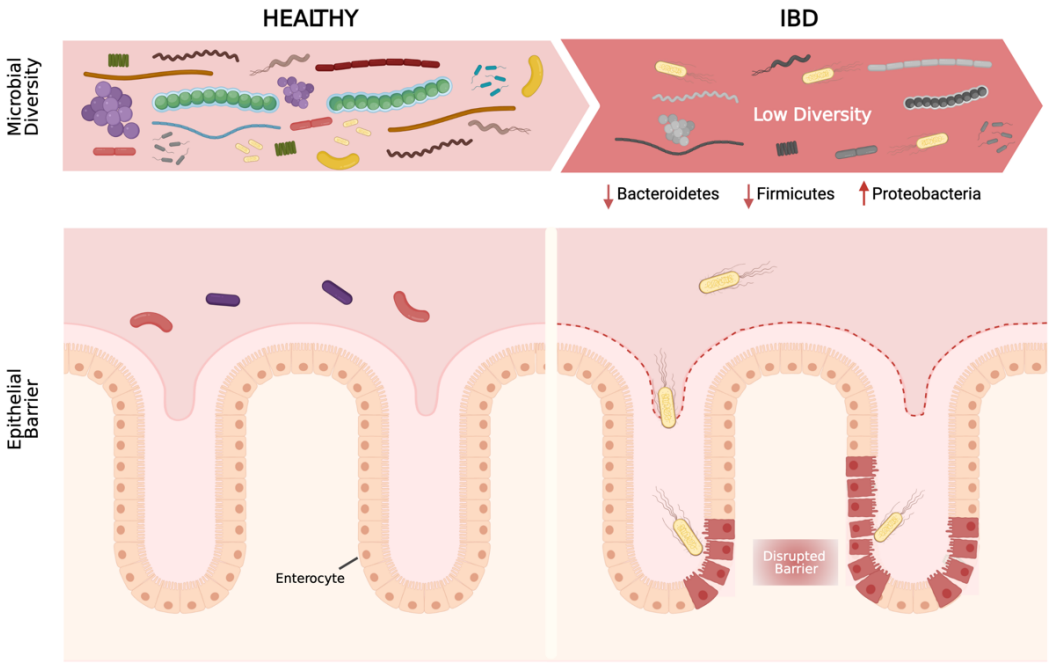

Figure 1.1: The microbial composition in the gut. On the left healthy gut is represented as having a high microbiome diversity and no damage on the epithelial barrier. On the right the IBD gut were microbiome diversity is lower and some bacteria is in physical contact with the damaged epithelium

benefits on the host [27–31] or any bacterial-released molecule capable of providing health benefits through a direct or indirect mechanism, has been collectively known as postbiotics since 2012 [27].

1.1.1.3 Immune response

As explained previously the immune system plays a role in IBD pathogenesis and pathophysiology.

Loss of tolerance to commensal bacteria has been suggested as the underlying mechanism triggering the inflammation on the intestine. The immune response involves many different cells lines and regions, which are important to know how they organize for a better understanding of the disease.

From the luminal side of the intestine, the first layer is the mucosa (See sections 1.2 or 1.1). In the colon the mucus is organized in two layers: the inner layer, a firm mucus layer; and the outer, loose mucus layer [32]. The intestinal epithelium is a single layer of cells organized into crypts and villi (and circular folds on the large intestine) that carries out a diverse array of functions besides digestion performed by specialized cell lineages.

Immune response in the intestinal mucous is mainly excreted by the gut associated

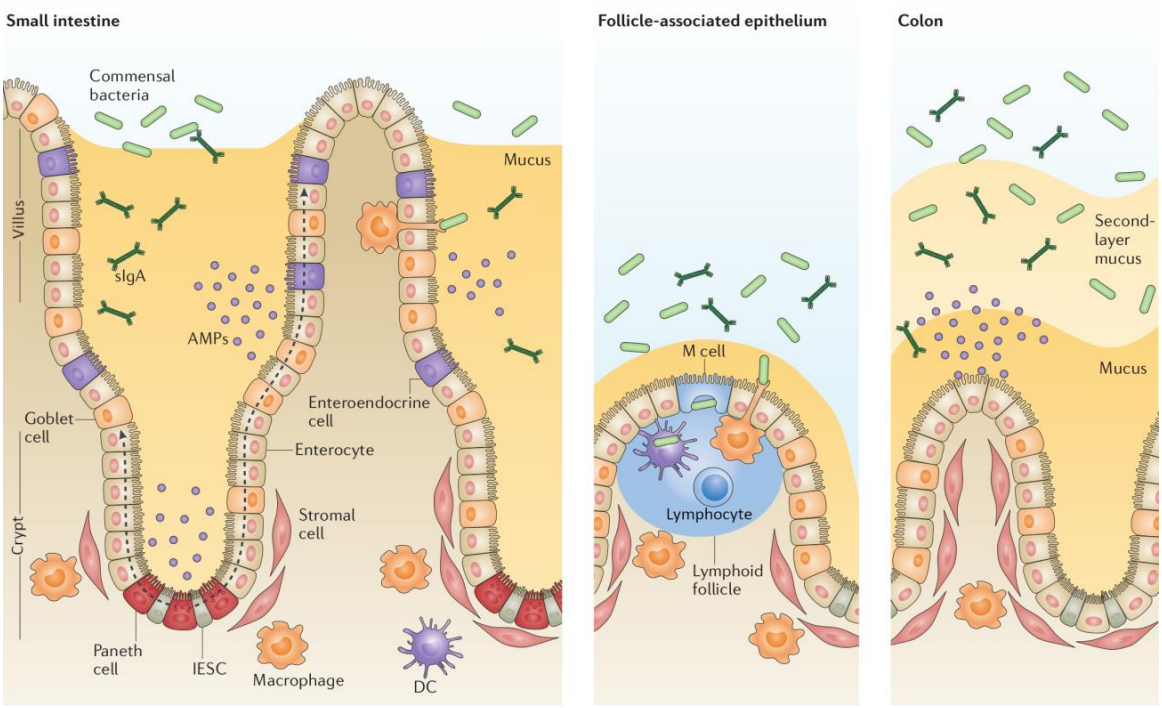

Figure 1.2: The intestinal epithelial barrier.

lymphoid tissue [33]. Genetically predisposed patients when exposed to certain environmental factors activate immune responses against microbials or self-antigens which in turn, may impair the mucosal barrier of the intestinal mucosa, the first physical barrier on the mucosal surface.

Both the adaptive and the innate immune cells are present in the intestine, right below the epithelium. On IBD due to antigen translocation into the lamina propria, the immune response leads the adaptive cells to generate immune response to harmless components of the intestinal microbiota. This initial response induces a local increase in the production of pro-inflammatory cytokines and mediators which damages the mucosa. Therefore, the loss of integrity on this barrier enables the intestinal luminal bacteria to access the intestinal epithelium and to interact with the immune system underneath it more directly [34].

The intestinal epithelium is another line of defense against bacterial invasion. Intestinal epithelial cells play a key role in controlling the integrity of the physical barrier to the intestinal microorganisms [34] not only physically but also secreting antimicrobial peptides and defensins, both of which are altered in IBD patients [35]. The intestinal epithelium also plays a key role on the intake and diffusion of metabolites from the intestinal lumen to the lamina propia.

Damaging or increasing the permeability of the intestinal epithelium results on a response from the immune system. Detecting signals of any foreign particle can also trigger the immune system. On the intestine this starts with the identification of these signals by intestinal epithelial cells have pattern recognition receptors. There are two main pattern recognition receptors: toll-like receptors (TLR), which are present on the surface, and nucleotide-binding oligomerization domain-like receptors (NOD-like receptors), present on the cytoplasm of the cells. These receptors upon recognition

of pathogen associated molecular patterns (PAMPs) start an amplifying signaling producing chemokines and cytokines which activates the transcription and translation of pro-inflammatory mediators to ensure an effective immune response. Initially the innate response is triggered but the cells also increase the antigen presentation to T cells and thus activate the immune adaptive response [36].

Other cell types, such as monocytes, macrophages and dendritic cells also present the pattern recognition receptors. From those, macrophages and dendritic cells are antigen presenting cells too and secrete several cytokines to activate other immune cells. Usually CD patients express higher amounts of TLR than healthy individuals, which might trigger a stronger immune response. This response is driven by $CD4^+$ T cells proliferation in secondary lymphoid tissues to T helpers in the presence of the antigens and cytokines nearby.

T helpers (T_h) differentiate depending of the cytokines at which they are exposed. Th type 1 are driven by exposure to IL-12 secreted primarily by dentritic cells. T_h2 are driven by cytokines secreted by macrophages. The imbalance between T_h1 and T_h2-promoting cytokines determines the intensity and duration of the inflammatory response in experimental colitis [37].

T_h17-promoting cytokines are less well characterized in human. T_{reg} cells differentiate after exposure to cytokines IL-10, IFN-γ and TGF-β. Overall the presence of certain cytokines and the response to self-antigens are factors leads to an inflammation and damage that is related to the onset and establishment of IBD

On these kind of diseases autologous hematopoietic stem cell transplantation has shown some benefits on IBD [38]. The benefit of hematopoietic stem cell transplant (HSCT) in autoimmunity is thought to originate from the depletion of auto-reactive cells regardless of their specificity. However, due to its associated risk this therapy is only given when patients are refractory to all available therapeutic options.

1.1.1.4 Environmental

Chronic inflammatory disorders and neoplasms have become the main cause of morbidity and mortality in the countries with high standards of personal cleanliness. A decrease in human exposure to microbes or hygene which might affect the proper maturation of the immune system so that it provides less immune response or exacerbated response towards “friendly” microbes [39, 40].

From other environmental factors related to IBD such as tobacco, diet, certain drugs and stress; tobacco is the most influential environmental factor. Surprisingly, it has an opposite effect on UC and CD: in CD tobacco is a risk factor that increases the risk of relapse and/or surgical intervention. In UC, it has been observed that smoking cessation worsens the disease [41].

Pharmacological treatments such as oral contraceptives, non steroidal anti-inflammatory drugs are also related to develop or relapse the disease [42, 43].

The psychological welfare of people also plays an important role in the disease progression, stress, anxiety and depression might be important in relapse and deterioration of the disease [44]. Other environmental factors have been linked to IBD but without enough evidence to support a causative effect in the development of the disease.

1.1.2 CD physiology

As previously introduced, CD is a chronic inflammatory disorder characterized by a discontinuous inflammation of the gastrointestinal tract. Inflammation on the gastrointestinal tract is transmural and can affect from the mouth to the anus, but mainly it manifests on the ileum and colon [10]. It is frequently associated with extraintestinal manifestation and/or concomitant immune-mediated diseases.

The disease itself manifest an heterogeneous symptoms that can involve, diarrhea, weight loss, abdominal pain, fever, anorexia and malaise. Other less frequent co-occurring manifestations are arthritis, primary sclerosing cholangitis, skin disorders venous or arterial thromboembolism and/or pulmonary involvement [45] . These symptoms make it hard to correctly diagnose the disease by non-specialists, in addition there is not a non invasive easy procedure to diagnose it. All these can lead to delays on correct diagnosis of the disease.

The detection of parasites or bacteria, such as *Clostridium difficile*, have been associated with CD [46]. The detection of fecal calprotectin, is generally a good marker of endoscopic activity with sensitivity above 70% and specificity above 80% [47, 48].

Usually the best diagnosis method is to perform a colonoscopy, whether there is inflammation on the gastrointestinal tract on discontinuous regions then is CD. This inflammation could also present ulceration with rectal sparing and histological lesions which also help to diagnose the patients [49].

Usually on CD a granulome, that is a region with big multinucleous cells, can appear on any intestinal layer. In addition to the inflamed location(s), mosaic zones (patches of inflamed and non-inflamed areas) are more characteristic of CD [50].

The Montreal classification aims to classify patients according to their age of disease onset, standardized anatomical disease location an disease behavior. This classification assumes that the location of CD remains stable over time after diagnosis but behavioral phenotypes change. Other scores consider area affected:

- Montreal classification allows for early onset of disease to be categorized separately those with age of diagnosis at 16 years or younger, diagnosis at 17–40 years and >40 years, respectively [12].
- SES-CD: simple endoscopic score for CD [51]. Score based on size of ulcers, ulcerate surface percentage, affected surface and presence of narrowing on the bowel.
- CDAI: CD Activity Index takes into account weight, ideal body weight, sex, and events on the last week such as liquid stools, abdominal pain, general well-being and if anti-diarrhea drug usage, as well as knowing if there are fistulas, fever, and other complications [52]

To some extent, there is a disassociation between clinical symptoms and the endoscopy finding. Often patients report feeling better despite lack of muscular healing [53]. To overcome this disassociation and be able to compare the well-being of patients several scores and thresholds are used on research and by physicians that will be described later.

In the early stages of the disease the relapsing and remitting course is more frequent. Often relapses are accompanied by clinical symptoms, and few have prolonged clinical remission (without treatment) [54]. When there is clinical remission, there can still remain some other lesions and often subclinical inflammation persists. Frequently the damage caused by the disease evolves to fibrostenotic stricture or penetrating lesions (fistula and abscess).

Damage of the disease might not be apparent to patients and might be only seen several years later than the first detection [53]. Mucosal healing is a first step towards the healing of deeper layers of the inflamed bowel wall on the CD.

Patients might progress from an inflammatory phenotype to a stricturing or penetrating one [12]. Stricturing is a narrowing of a part of the intestine often because of scar tissue and fibrosis in its wall. Penetrating is when the epithelium has some holes or tubes. If these tubes result in an abnormal connection between two body parts it is a fistula, it might also result in an abcess, a collections of pus, often developed in the abdomen, pelvis, or around the anal area.

1.1.3 UC physiology

As previously introduced, UC is a chronic inflammatory disorder characterized by a continuous inflammation of the colon. Depending on the inflamed segments of the intestine it is classified in several phenotypes.

Around a third of the patients with UC suffer proctitis, the inflammation of the rectum. If the segments from rectum to the sigmoid colon are affected is a distal colitis, if it affects the left colon then it becomes a left colitis. If the inflammation continues to the descending colon it is then an extensive colitis until it affects the whole colon when it becomes a pancolitis. The extension of UC is inversely related to the frequency. However, the extension and severity of the disease correlates: the prognosis is worse the more extended it is [55]. In addition, the damage usually consists in many neutrophil in the lumen crypt [50].

The goal of the clinical care is to recover. As a first step, the symptoms of IBD have to lessen to the point that they are mostly absent, gone, or barely noticeable, this is known as clinical remission. However, this is not enough as the mucosa might be still inflamed and thus the reconstitution of the structure and function of the intestinal mucosa is not complete. Other lesions, might aid to the progression to other phenotypes such as fibrostenotic stricture or penetrating lesions or primary sclerosing cholangitis [56].

To prevent and avoid further damage several procedures are followed:

When there is dysplasia, an abnormal development of cells within tissues or organs (which is considered a precedence before colorectal cancer growth [57]), or the damage on the colon has been too big a surgical procedure to remove part or all of the colon must be done.
Patients that undergo a colectomy need to have their bowel reconnected with a procedure called ileoanal anastomosis (also know as J-pouch by the shape it takes) surgery. Often the lining of the pouch created during surgery becomes inflamed on what is known as pouchitis [58].

Many scores have been proposed for several purposes, from quality of life to disease severity or patient status. Among the scores most used are the following:

- Mayo: A score designed to be simple to calculate based on stool frequency, bleeding, mucosal appearance at endoscopy and physicians assessment [59].
- IBDQ: A 32 questionnaire used to assess the quality of life grouped into four categories: bowel, systemic, social and emotional [60].
- UCEDIS: An endoscopic score based on vascular pattern, internal bleeding and erosion and ulcers [61].

Other measured parameters include, weight, effective weight, fecal calprotectin, C reactive protein and hemoglobine.

1.1.4 Treatment

Current treatment attempt to induce and keep the remission of patients and reduce secondary effects of the disease instead of revert the pathogenic mechanisms. As standard of care corticosteroids, aminosalicylates and immunosuppressor and some other drugs like antibiotics or metronidazol are util in some cases.

Acid 5-aminosalicylic (5-ASA or mesalazina, pentasa) can be given in a topic way (either liquid, enemas, or suppository) or in oral form (pills or dilutions). In CU it helps in the clinic remission but it does not always mean that there is remission (is twice much likely than placebo to reach remission) [62]. On CD the effects are not so stark and generally it does not produce changes on the disease [63].

Antibiotics, such as metronidazol and ciprofloxacina, are effective to deal with secondary effects of IBD such as abscess and bacterian overgrowth in CD [64, 65], but they do not seem effective on UC [65].

Corticosteroids can be taken orally, such as prednisolona, prednisona and Budesonide; intravenous, hidrocortisona, metilprednisolona; or via enemas and suppositories. Budesonide is not absorved well and has a limited biodistribution but it has good therapeutic benefits with a reduced systemic toxicity in IBD [54, 66]. These drugs work very well as antiinflammatory for mild or severe IBD but do not work well as maintenance drug [67, 68].

Thiopurines (Azathioprine, mercaptopurina) are immunosuppressants drugs that deactivate key process of limphocytes T that might trigger the inflammation. As a side effect they are toxic due to their interaction with nucleic acids [69]. on CD they are useful to induce and keep remission [70], while on UC they are used to keep the remission [71].

In the last two decades IBD treatment has moved from aminosalicylates, corticosteroids and immunomodulators to anti-TNFα. Anti-TNFα drugs has changed IBD treatment as it reduced the hospitalization associated with previous treatments, reducing medical costs and risk of surgery as well as induce a better mucosal healing and quality of life for patients [72]. However, 20-30% of patients have no response to this treatments and another 30-40% lose response in a year [73].

Recently a new wave of drugs has been developed targeting different molecules such as vedolizumab, targeting anti-integrin$\alpha 4\beta 7$, ustekinumab, targeting both IL-12 and IL-23, risankizumab an anti IL-23, tofacitinib an inhibitor of JAKs, infliximab an anti-TNFα.

Patients might become refractory to drug. Thus, drugs do not have the same effect as previously and the dose might need to be increased with the risk of more secondary effects [1]. Surgery resection might be needed on these patients.

Close to 35% of patients with UC will need to have a surgery resection, either due to complications or because the inflammation can not be controlled. Surgery usually removes the inflamed segment of the colon. The most common procedure used is a colectomy (whole colon removal), with ileostomy [74]. CD patients usually require surgery associated to complications like stenosis, abscess, and fistulas) between 70% and 90% at some point of their lives [75]. Usually the surgery is limited to removing the inflamed segment but occasionally an ileostomy is required [76].

If the drugs fail to contain the inflammation and heal the mucosa doctors might recommend a different procedure. In some cases HSCT is recommended which have shown to improve the life of the patients [77]. This is a new procedure given only to the most extreme cases to reset the immunological state of the patient.

To reset or hugely modify the microbiota fecal microbiota transplantation between different people is currently being explored [78].

1.1.5 Summary

IBD is a complex disease that impacts the health of many people for long time and with lasting impact on their quality of life.

Current clinical care in some cases is enough to have a sustained clinical and endoscopic remission but most often is not enough and relapse is expected. Several factors, such as becoming refractory to drugs, intermittent course of the disease and lack of validated predictors of disease course or response to therapy make the treatment complex.

Lack of knowledge of what are the factors cause of the disease make those treatments and drugs to be addressed to block further inflammation and damage, but cannot prevent it and often they do not stop it completely.

1.2 Integration studies on IBD

Many studies have looked up to the origin of the disease. As seen, one of the hypothesis behind the maintenance of the inflammation involves the microbiome and the host epithelium. This has been studied using several data sources, mostly from sequencing data. The technical methods used to obtain the data of the inflamed tissue differ between extracted from biopsied samples at colonoscopy or from surgical samples.

Those samples are usually used later on to diagnose or for research purpose. To obtain research-quality data it usually imply using techniques such as immunohistochemistry, histopathology, immunohistochemistry, fluorescence in situ hybridization

and polymerase chain reaction. These techniques allow to measure or visualize where are the cells expressing certain proteins or genes, thus helping with the analysis validation.

Furthermore several studies have been carried out to discover links between microbiome and the inflammation, followed by those looking for some relationship between genetics and the disease and more recently the metabolome. These studies, known as integration, multi-omic or interaction studies, usually use multiple sequencing assays as the bases of the analysis [79].

However, confirming causal interactions of the variables of each essay is difficult. To find relationships some articles use correlation [80], there are others that use a combination of methods from correlations, partial correlations to integrative methods [81–83] and network integrations [82].

Very rarely there is an experimental confirmation of the relationships between variables of the different assays because it is complicated to test an interaction and to set up the right conditions for the many variables that are accounted for on the integrations. One of the few methods published that shows an interaction between genes and microorganisms on IBD is to expose the *ex-vivo* sample or cell lines with microbiomes or supernatant of at their culture [84].

1.2.1 Type of data used for integration analysis

According to the data used, we can classify the studies:

1.2.1.1 Transcriptome

Most of the integrations refer some other source of data to the transcriptomics of the patient. The transcriptome of patients derived samples has been extensively studied since the existence of microarrays. There are known marker gens of inflammation and many research focus on identifying prognosis predictors and treatment response prediction based on gene expression [85–87]. Recently single-cell RNA-seq technology has enabled to estimate cell populations of the samples with better degree of success than bulk RNA-sequencing. Single cell technologies are starting to be used for integration.

1.2.1.2 Microbiome

Many of the integration analysis on IBD are done between host transcriptome and the microbiome. These studies use datasets from IBD patients usually stratified by disease activity or severity of inflammation or location of the disease. Most of them are based on correlation analysis between the microbiome and RNA-seq [80]. Conclusions of these integrations range from finding differences on the correlation depending on the type of disease ([80]) to finding relationships with inflammatory genes [81].

1.2.1.3 Genetics

Genetics is the next most common data source used to integrate data on IBD. Most studies on genetics and IBD are genome-wide association studies. The genetic component is specially important on IBD that starts on children [14, 83, 88].

When using genetic data to integrate it with transcriptomics it is usually to understand how a genetic variant is affecting a gene expression. This has lead to the identification of expression quantitative trait loci (eQTL) [89–92].

1.2.1.4 Metabolome

More recently there have been an increased interest on the study of the metabolic stat of IBD patients, given that microorganisms interact with the host also via their products and metabolites. There is evidence some of these metabolomic products come from the microbiome [25, 84]. Some studies have integrated the metabolome with the RNAseq and state of the epithelium [93, 94].

1.3 Integration

The term "data integration" is widely used with varying meanings. According to the dictionary integration is defined as:

> "the process of combining two or more things into one" — Cambridge Dictionary

Other words used are integration, and if specific to data from sequencing technologies, multi-omics or pluri-omics. Here integration will be used as it is the more general one and not restricted to omics or sequencing technologies.

Since the beginning of the integration methods there have been many methods proposed [95]. Some of the early methods were initially used for surveying the agreement of different evaluating systems [96], others were developed for agricultural sciences [97] or food industry [98]. Some of these methods are specific for one application or data type while others are more generally applicable. Lately, the access to bigger datasets with more variables and often from the same samples has increased the focus of the research community on the methodologies available on several disciplines but mainly on biological sciences. The explosion of data on biological sciences has been driven by the new sequencing technologies that allow to measure thousand of variables of many samples at the same time. If done with multiple sequencing technologies it is usually referred as multi-omics methods, which usually only uses omic data.

It is crucial to classify, review and compare tools available, as well as, to benchmark these tools against the same dataset as a way to provide clear recommendations to anyone wishing to use them [99]. Part of these efforts use the methods' strategies to classify them [100, 101]. Following this view we will review the available integration methods according to several axes: type of data used, aim of the method, relationships between variables, relationships between samples, relationship between variables and samples, input data, mathematical framework and results of the method.

1.3.1 Classification of integration methods

Integration methods have very different properties that allow them to be classified and compared [102]. Here we classify those meant to be used with omics datasets, with references to concrete methodology and in occasionally to articles using them.

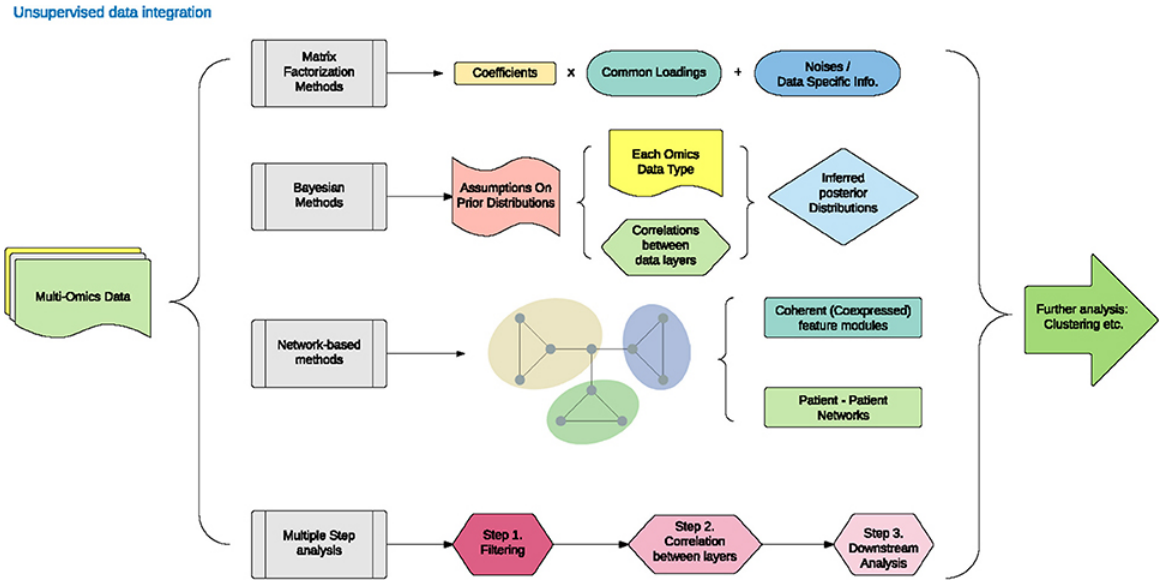

Figure 1.3: Unsupervised data integration methodology. Figure 1 from Huang 2017.

1.3.1.1 Data type: numeric or categorical

The most important distinction in integration methods is what kind of data are combined. In general data can be divided between categorical and numeric variables, which are usually found in several fields. Sometimes clinicians want to understand the relationship between a phenotype they observe and the underlying mechanism. Usually this involves looking how metabolites, gene expression, methylation, number of variants a gene has, and other numeric variables are related to the observed phenotype (i.e. pain).

Depending on the method's aim it handles numeric and categorical data types or just one. Often they are used differently. The most common way to handle different types of data is converting the categorical values to a mock or dummy variable. For each categorical factor there is a new variable whose value is 1 if that sample had this factor and 0 otherwise. For instance, if the categorical variable has three values (A, B, C) it would be converted to A (1, 0, 0) B (0, 1, 0) and C (0, 0, 1). Often the number of variables created is one less than the number of factors that existed, on this example only A and B would be kept. This transformation allows to use categorical values as numeric variables.

If the method only accepts categorical data but you want to provide numeric values usually those values are categorized. For example if a variable is (0.123, 0.25, 0.56, 0.78) one could make to categorical values like ("<0.5", "<0.5", ">0.5", ">0.5"). The number of categories to use and how the numeric value splits depends on each case.

Very rarely methods allow to use both data types as they are. If they allow so, it is usually for classification purposes only.

1.3.1.2 Objective

The objective of any method is one of its most important defining properties. Data integration can be classified according to the (biological) question they try to answer. In general all of them aim to provide a better understanding of the relationships between the 'omics types. Often a single method is not enough and several methods are used on the same dataset. This is specially relevant when a potentially relationship between omics is discovered. For instance, checking that in a particular case or condition a given relationship is present, might require experimental confirmation.

Most of the times one (or several) of the following results are expected from integration methods:

- Classification of the patients or samples

One of the purposes of integration might be to use multiple sources of data to accurately describe how do samples or patients fit on a predefined possibles states. The objective here might be to accurately describe whether a patient has one or other related disease of a possible subset of diseases or phenotype [103].

- An overview of the role of each individual omic in a biological system

Sometimes the question is which omic method is the best describing a disease. This knowledge could prevent performing expensive tests and replace them by more affordable or easier technique that has enough predictive power or is sensitive enough and specific for the task. An example of this is the search of markers on blood to identify links between different cell populations [81, 104].

- Finding a molecular signature

A signature is usually a group of features that describe/are representative of a cell type, a process or a stage. Identifying a subset of variables from the omics that are related is often a desired goal because it reduces the amount of variables allowing to perform experiments on the bench on just those that might be important. In other fields, such as machine learning, selecting the important variables is known as feature selection. There are several methods that are used to do this [100]. An example of this is when performing eQTL analysis, where a locus is related to a change in gene expression [89–92].

- A predictive model

Predictive models usually require a very good understanding of the current and/or past relationships, as well as, a good feature selection procedure. If a good model on previous data exists it might be used to predict future events. Sometimes, models are only built to predict events without being able to accurately understand the underlying mechanism. This kind of methods are used to improve treatment selection, diagnosis and prediction of prognosis [105].

- Impute values

Some methods aim to accurately guess the values of missing data given some other information. Missing values can happen for a variety of reasons from practical ones, like a sample not being available, to technical ones, such as laboratory mistake [106]. However, this is often a intermediate step to other goals.

To complete these goals it is important to have enough statistical power to determine the significance of tests performed (if any) and to understand how complete are the data sources used on the integration [107]. Having more statistical power helps identifying the relationships one seeks when using this methods.

1.3.1.3 Relationship between variables and samples

Depending on the amount of variables and samples used in studies can be classified in two types. Traditionally for each sample few variables are measured, for instance on a biopsy with RT-PCR only a few genes are measured, however with the new omics techniques (transcriptomics, metabolomics, methylomics, genomics), thousands of variables are measured for the same sample. This has lead to the following situation:

- More variables than samples

For a single sample of RNA around 50k genome identifiers (genes, long non coding RNAs, iRNA, pseudogenes,...) can be measured. Which leads to the case where there are many more variables than samples. Thus high-throughput data analysis typically falls into the category of $p \gg n$ problems (big p, little n), where the number of genes or proteins, p, is considerably larger than the number of samples, n. With such high number of variables the identification of the relevant variables is hard because variables will co-variate. When many variables are tightly correlated, discovering which one is important using just numerical methods is challenging. This is even more difficult when looking for causal relationships.

- More samples than variables

This was the usual case when for instance, from a cohort of patients the temperature is measured along the stage of a disease: two variables, time and temperature for each sample. If there are more than 2 patients, then the number of samples is greater than the number of variables studied. This is described in the literature as $n \gg p$ (or big n, little p). Nowadays this is less frequent on the bioscience world, and does not

causes trouble analyzing it because the high number of samples allows to accurately estimate the dispersion of variables.

There are several methods available to estimate the number of samples required [108]. Having just the enough samples for the desired statistical power, however, might not be enough in case some samples are not correctly processed.

In addition, variables might be separated on different blocks of variables. These blocks might be just of the same source or from multiple methods. Depending on the method this blocks might have special meaning: when all the data is joined in a single block it is known as superblock.

1.3.1.4 Relationship between samples

Depending on the relationship between the samples, the questions answerable and the methods that work on them differ.

A sample can have multiple or one data source, for instance we could have RNA-seq and 16S data from the same sample. In a study if all the samples have all the data from multiple data sources it is a complete case. If some samples have data from some data sources but not from others the study is not a with a complete case.

Sometimes because the sample is not enough, or there are some technical or organizational problems a source of data for a sample (which is known as an incomplete case) might be lost. This results in a new source of variation that has to be dealt with, which complicates the conclusion one can draw from the studies of these kind of data.

Even when all the cases of a patient are complete the samples can come from several sites of the same individual or with different combinations of variables, which makes it relevant to understand the relationships between the different samples.

There is no easy classification of this as each experiment might be designed differently. In general, experiments are designed to be as consistent as possible but in face of adverse events that become a variation of the design the analysis complicates. Either some data is imputed or some samples are omitted for the analysis. This can happen with samples taken at different timepoints as patients for instance if they miss a follow up visit.

Time As mentioned above, time is one of the factors that sometimes cannot be controlled, despite having programmed visits every two weeks, for instance, some patients might come early or later due to multiple reasons.

Sometimes, the objective of the study is precisely to analyze the relationships at different time, or to asses how the relationships change with time. To discover causality between two variables the cause must precede the consequence, which highlights the importance of time. Being aware of the time differences and time scales is crucial in most cases.

During *in vitro* experiments, conditions can be reproduced even if they are at different time. However, when using patients material replicates can not usually be performed like *in vitro* experiments. This makes it harder to study time-related change on patients.

Lastly, time between the collection of a fresh samples and its processing also influences the readings of the samples of the omics technology, specially RNA-seq [109, 110]. Some genes are more influenced by time than others but as they are measured at the same time in all samples this might distort the data. Keeping track of the time that it takes to process samples is also hard to due and requires a highly coordinated effort [111].

1.3.1.5 Relationship between variables

Once data is collected, the next step is to understand the relationship between the variables present. As mentioned earlier, some variables influence others which can affect the outcome in complex ways. With many variables present in a dataset it is important to be aware of known relationships between variables. Even in a simple dataset, like an RNA-seq dataset, it is important to be aware of the relationships between variables.

Since the discovery of the lactose operon it is known how some genes regulate each other [112]. However, it is not know how other variables are related between them. For instance, how does the increase in expression of a gene affects the growth of a microorganism? Usually the relationships between variables are mediated by many factors or interactions.

One of the best examples of such interactions is when some variables correlate. Their correlation can be used to reduce the number of variables being analyzed by ignoring the relationships between them and using the most representative variable (less widely correlated and with more variation). This step is usually done by dimension reduction methods. However, sometimes this is not desirable or feasible as the correlation does not explain the direction of the causality of the interaction between the variables (if there is any).

Network approaches relate the variables to each other [113]. These approaches are fairly new and growing in popularity partly because they can address the direction of the interaction.

In partial correlations the effect of other variables on the two being under study are taken into account [114]. They assumes a linear relationship between the co-occurring variables and those of interest. However, it is computationally expensive when there are thousands of variables.

1.3.1.6 Input data

We have classified the studies according to the data they use (as seen previously). But, some methods to account for relationships of variables only work when a dataset is complete while other do not:

- Data from the same samples:

These methods do not handle well or at all missing data. They need complete cases/data of the samples in order to be able to integrate the results. These methods include Regularized Generalized Canonical Correlation Analysis (RGCCA) [115,

116], Multiple co-inertia analysis (MCIA) [117], Multi-Study Factor Analysis (MSFA) [118], Multi-Omics Factor Analysis (MOFA) [119] and STATegRa [120].

- Data from different samples:

These methods do not need data from the same sample. They draw their conclusions generalizing from the data available. Some of them handle missing data, while others do use the data at face value. These method includes MetaPhlAn2 [121], HUMAnN,[122] and LEfSe [123].

Furthermore, some methods are designed to integrate specific types of datasets, (usually because they make some assumptions that are only met on that kind of data). For instance, HCG, 16S rRNA-seq, RNA-seq and metabolomics do not share the same data distribution, and are different between them. Also even with the same data depending on the processing of the data they can have very different properties: OTUs (operational taxonomic unit) properties are not the same as ASV (amplicon sequence variants) when analyzing 16S rRNAseq data.

1.3.1.7 Mathematical framework

Methods use different mathematical frameworks to process the data. Here we briefly describe some common mathematical frameworks, some of which have previously appeared:

- Networks

  Networks methods were mentioned because they use and find information about the interaction of variables. Multilayer networks, including the multiplex, Molti-C-DREAM [124], Random Walk with Restart on Multiplex and Heterogeneous Biological Networks RWR-MH, Random walk with Restart on Multiplex RWR-M [125]. Network embedding MultiVERSE are some of the methods using networks [126].

  Bayesian approaches are also quite frequent, these methods use the Bayes' theorem to see the relationships between variables. The Bayes' theorem explains that the conditional probability of a variable is related to the prior knowledge of conditions that might be related to the event [127]. Some methods that use these approaches are Reconstructing Integrative Molecular Bayesian Network (RIMBANET) [128] and Bayesian Consensus Clustering (BCC) [129].

- Dimensional Reduction

  These methods focus on finding just a few variables and summarizing them using a function that has some desired property such as the correlation between thre transformed variables is maximal while the components are orthogonal.s The selection of variables is usually done with L1 or Lasso Regression regularization technique or L2 also known as Ridge Regression. L1-regularization adds a penalty equal to the absolute value of the magnitude of coefficients which might leads to some coefficients becoming zero and the variable eliminated from the model. On the other hand, L2-regularization does not result in elimination of

coefficients or sparse models and can only be used when there is multicollinearity as it works well to avoid over-fitting.

Several tools use this approach, Momix [130], regularized canonical correlation analysis (RGCCA) [131], mixOmics [132] and STATegRa [120]. Other methods use bayesian approches like the Bayesian Group Factor Analysis [133].

- Active module identification

  Multi-omic objective genetic algorithm (scores based in two metrics; node score and density of interactions score). An example of a method using this approach is Multi-Objective Genetic Algorithm to Find Active Modules in Multiplex Biological Networks (MOGAMUN) [134]

Usually depending on the mathematical framework used, these methods return similar outputs.

1.3.1.8 Output results

According to the output the integration methods can be classified in several groups:

For the network methods the following output is usually returned: Connections between the variables/nodes, a measure of how strong is the connection (or simply if there is a connection or not).

For dimensional reduction methods there are three possible outputs: Shared factor across the data, specific factors for each data or mixed factors.

- Shared factors:

Integration results in a vector of the samples in a lower dimensional space that is shared by all the data set used to integrate. Such methods include iCluster, Multi-Omics Factor Analysis (MOFA) [119].

- Specific factors:

Integration results in several vectors of the samples in a lower dimensional space of each data set used to integrate. Such methods include Regularized Generalized Canonical Correlation Analysis (RGCCA) [115, 116], Multiple co-inertia analysis (MCIA) [117] and Multi-Study Factor Analysis (MSFA) [118].

- Mixed factors:

Integration results in both shared and specific factors, to each dataset and common to all them. Such methods include Joint and Individual Variation Explained (JIVE) [135] and integrative Non-negative Matrix Factorization (iNMF) [136].

1.3.2 Interpretation

How to interpret the results of applying the different methods is highly linked to understanding the method and its output. On a correlation between two variables, the interpretation of the analysis is clear, if one variable increase, the other one too. The implications of this correlation can be far reaching but the principles to understand them are simple.

However, on more complex methods the interpretation becomes less clear. The interpretation of a canonical correlation analysis is much harder [137]. Also on more complex methods the number of parameters required increases so the time and intellectual effort to understand the relationships between the parameters is also higher.

The interpretation also helps to discuss the results and relate it to other previously know information.

- Individually:

Here we study how each variable relates to another. In the correlation analysis, the relationship between two variables under study. Or if looking by patient: how do interpret that in these patient variable A and B is X and Y?

- Globally:

In a principal component analysis for instance how do we interpret that some variables have the same loading? What happens in a more complex method like canonical correlation analysis?

There are some articles about how to interpret those methods on real datasets [138]. Others, to benchmark and to learn how to interpret propose analyzing a simulated dataset [139, 140]. Which is used to compare the results of the integration with the dataset of interest and to compare different tools. These datasets are created with some relationships that the tools are expected to find.

There exists several methods to create synthetic datasets like MOSim [140], metaSPARSim [141], CAMISIM [142], ballgown [143], polyester [144] and even edgeR [145]. These methods are useful to compare different setup but they can miss some subtle not previously reported relations on real data.

1.3.3 Reviews

The comparison and review of methods independently from original authors have become a crucial step for selecting the right tool to apply a given dataset and research question [130].

Some of these reviews focus on a specific type of data integration: metabolomics [100], genomics [11], microbiomics... Others focus on the disease and the challenges of each omics and the need of an integrative approach to provide better therapies [107, 146, 147]. On this regard there are several efforts to integrate data on IBD but no comprehensive review to date is known to the author. The most comprehensive article to date is a very recent review identifying problems and providing recommendations for future work [148].

1.3.4 Summary

The field of integration is large and complex, with increasing interest over the last few years, specially in the psychology and omics fields. As a methodology they are quite complex and diverse but there is a growing interest on them to help answer complex questions without using other complex tools like deep neuronal networks or other machine learning approaches (despite not being incompatible).

Methods to integrate have many characteristics, depending on the objectives and data that available. Regardless of the method used, interpretation and reporting is usually the main challenge.

1.4 Summary

IBD is a gastrointestinal disease that includes two different diseases UC and CD. It affects preferentially the lower intestinal tract causing lesions on the epithelial barrier. Bacteria causes or use these lesions to further damage the patients. But this interaction is also influenced by many other variables.

Data integration are methods to analyze datasets with data from multiple sources. They include a variety of methods and there are methods for several purposes. On IBD it has been used mostly sequencing data, disregarding the other variables known to be relevant.

Data integration on IBD has been underexplored despite being known that many factors interact on the pathogenesis and maintenance of the disease.

Hypothesis and objectives

2.1 Hypothesis

IBD is a gastrointestinal disease affecting preferentially the lower intestinal tract causing lesions on the epithelium. Bacteria interact with host's epithelium and with many other variables. There are data integration methods for a variety of purposes, including finding relationships between variables. On IBD these methods have not been used much to find which bacteria is related to the host's genes.

We hypothesize that the expression state of the epithelium and the **microorganisms present might identify whether patients are suffering IBD if environmental factors are taken into account**.

2.2 Objectives

The main objective of this thesis is finding the relationships between microbiome and gene expression in the intestinal mucosa. For this reason we identify microorganisms and genes related to IBD.

- Identify relationships between intestinal mucosal gene expression and microbiome presence on the intestine.
- Determine the influence in the microbiota and mucosa relationship of clinical variables such as location (colon and ileum), disease status, sex, age, treatment received, etc. in the microbiome and mucosa relationship.
- Identify groups of possible underlying mechanism interactions between the microbiome and mucosa on IBD.

Materials and methods

This chapter contains a brief description of the main characteristics of the different datasets used on this thesis . The complete processing protocol before obtaining the data of those dataset that were not generated at Hosptial Clínic can be found on their respective reference.

Samples of the different cohorts collected in Hospital Clínic were collected similarly and described only once. Differences between protocols are noted on the respective dataset's section.

Here we describe all the methods used to analyze data from the multiple cohorts included in this thesis. The code used can be found on the links provided in the appendix.

3.1 Datasets

3.1.1 Puget's dataset

The glioma dataset is the data provided as an example of biological data by the authors of RGCCA from a previous publication [149]. The data came from diffuse intrinsic pontine glioma patients whose transcriptome was analyzed with Agilent 44K Whole Human Genome Array G4410B and G4112F. The copy number variation of the samples was processed with the ADM-2 algorithm, and data from comparative genomic hybridization (CGH) analyzed using Mutation Surveyor software. In addition, this dataset contained information on age, localization of the tumor, sex and a numerical grading of the severity of the tumor [149].

Table 3.1: Characteristics of samples from the Puget's dataset.

Characteristic	Puget's
Samples	53
Sex (female/male)	28/25
Location (cort/dipg/midl)	20/22/11

3.1.2 HSCT dataset

Samples from the HSCT dataset used in this thesis were from a cohort of patients with severe refractory CD undergoing HSCT. Patients were treated in the Department of Gastroenterology (Hospital Clínic de Barcelona –Spain–). The protocol was approved by the Catalan Transplantation Organization and by the Institutional Ethics

Committee of the Hospital Clinic de Barcelona (Study Number HCB/2012/7244). All patients provided written consent following extensive counselling about being included on the study and using their data on publications.

Colonic and ileal biopsies were obtained at several time points during ileo-colonoscopy, at inclusion and every 6 or 12 months after HSCT up to 4 years after the start of the treatment. Samples were obtained when possible from both uninvolved and involved areas. In addition, biopsies were taken from the ileum and colon regions of 19 non-IBD controls consisting of individuals with no history of IBD and who presented no significant pathological findings following endoscopic examination for colon cancer surveillance (Hospital Univesitari Mútua de Terrassa –Spain–). The protocol was approved by the Institutional Ethics Committee of the Hospital Univesitari Mútua de Terrassa (Study Number NA1651).

At least one biopsy was collected and fresh-frozen at -80°C for microbial DNA extraction. The remaining biopsies were placed in RNAlater RNA Stabilization Reagent (Qiagen, Hilde, Germany) and stored at -80°C until total RNA extraction. In total 158 samples with both RNA and DNA extraction of the same segment and time were available 3.2:

Table 3.2: Characteristics of samples from the HSCT dataset.

Characteristic	HSCT
Sex (female/male)	22/15
Age at diagnostic (<17/<40/>40 years)	7/11/0
Duration (in years): mean (min-max)	14 (8-28)
Age: mean (min-max)	44 (23-70)
Disease status (non-IBD/CD)	51/107
Sample location (ileum/colon/unknown)	48/108/2
Local simple endoscopic score for CD: mean (min-max)	2.15 (0-12)
CDAI: mean (min-max)	120 (0-450

3.1.3 Häsler's dataset

An IBD-related dataset was obtained by Prof. Dr. Rosentiel and Prof. Dr. Robert Häsler [80]. Biopsies were obtained endoscopically during routine diagnosis RNA and DNA were extracted using standard procedures. DNA from the 16S rRNA gene was amplified with primers 319F and 806R. Both RNA and DNA was then sequenced on HiSeq 2000 and MiSeq respectively. These biopsies included samples from the terminal ileum and sigma from CD, UC, infectious disease-controls and healthy non-IBD. The dataset included information about gender, location, age, and the status (inflamed or non-inflamed) of the region from which the biopsy was taken.

Table 3.3: Characteristics of samples from the Häsler's dataset.

Characteristic	Häsler's
Disease status (non-IBD/IBD)	33/26
Sex (female/male)	42/17

Characteristic	Häsler's
Sample location (ileum/colon)	30/29

3.1.4 Morgan's dataset

A previously published dataset from a pouchitis study was analyzed [150]. In this study patients having undergone proctocolectomy with ileal pouch-anal anastomosis for treatment of UC or familial adenomatous polyposis at least 1 year prior to enrollment were recruited at Mount Sinai Hospital (Toronto, Canada) excluding individuals with a diagnosis of CD. The dataset has a total of 255 samples from 203 patients, containing data for both host transcriptome and intestinal microbiome. On some cases several biopsyes were collected from the same patients. This dataset included anonymous identifiers for patients, whether the sample was from the pre-pouch ileum (PPI) or from the pouch, the sex, the outcome of the procedure and an inflammation score. The pouch ileum might be inflamed or not.

Table 3.4: Characteristics of samples from the Morgan's dataset.

Characteristic	Morgan's
Samples (n)	255
Sex patients (female/male)	101/102
Sample location (Pouch/PPI)	59/196

3.1.5 Howell's dataset

This dataset included a cohort of 66 treatment-naïve children at diagnosis of their IBD, along with 30 age- and sex-matched non-inflammatory control children, recruited at the Paediatric Gastroenterology unit at Addenbrooke's Hospital (England) [151].

Data from 77 samples that had both RNAseq and 16S data was used. There are 10 non-IBD samples, 11 with CD and 11 with UC. Data has the following characteristics: disease, age at diagnostic, age at time of study, sex, sample location, and disease activity:

Table 3.5: Characteristics of samples from the Howell's dataset.

Characteristic	Howell's
Disease (non-IBD/CD/UC)	11/10/11
Age at diagnostic (<17/<40/>40 years)	32/0/0
Age: mean (min-max)	12 (6-15)
Sex (female/male)	10/22
Segment (ileum/colon)	31/46
Clinical history (inflammation/no inflammation)	24/53

3.2 Sample processing

3.2.1 RNA sequencing

Total RNA from mucosal samples (HSCT cohort) was isolated using the RNAeasy kit (Qiagen, Hilde, Germany). RNA sequencing libraries were prepared for paired-end sequencing using HighSeq-4000 platform. Samples with good enough quality as recommended by FastQC were processed with cutadapt (version 1.7.1 [152]) for quality filtering. Later, the libraries were mapped against the human reference genome using the STAR aligner (2.5.2a) with Ensembl annotation (release 26 of GENCODE, GRCh38.p10 or superior) [153].

Read counts per gene were obtained with RSEM (version 1.2.31) [154] as previously described [155].

3.2.2 Microbial DNA sequencing

Biopsies from the HSCT CD cohort were resuspended in 180 μl TET (TrisHCl 0.02M, EDTA 0.002M, Triton 1X) buffer and 20mg/ml lysozyme (Carl Roth, Quimivita, S.A.). Samples were incubated for 1h at 37°C and vortexed with 25 μl Proteinase K before incubating at 56°C for 3h. Buffer B3 (NucleoSpin Tissue Kit–Macherey-Nagel) was added followed by a heat treatment for 10 min at 70°C. After adding 100% ethanol, samples were centrifuged at 11000 x g for 1 min. Two washing steps were performed before eluting DNA. Concentrations and purity were checked using NanoDrop One (Thermo Fisher Scientific). Samples were immediately used or placed at -20°C for long-term storage until DNA sequencing.

3.2.2.1 DNA sequencing

Microbial cells were disrupted by mechanical lysis using FastPrep-24. Heat treatment and centrifugation were conducted after adding a cooling adaptor. Supernatants were treated with RNase to eliminate RNA. Total DNA was purified using gDNA columns as described in detail previously [156]. Briefly, the V3-V4 regions of 16S rRNA gene were amplified (15x15 cycles) following a previously described two-step protocol [157] using forward and reverse primers 341F-ovh/785R-ovh [158]. Purification of amplicons was performed by using the AMPure XP system (Beckmann). Next, sequencing was performed with pooled samples in paired-end modus (PE275) using an MiSeq system (Illumina, Inc.) according to the manufacturer's instructions and 25% (v/v) PhiX standard library.

Library preparation and sequencing of the HSCT dataset were performed at the Technische Universität München. Briefly, volumes of 600μL DNA stabilization solution (STRATEC biomedical) and 400μL Phenol:choloform:isoamyl alcohol (25:24:1, Sigma-Aldrich) were added to the aliquots.

3.2.2.2 Microbial profiling

For the HSCT dataset the processing of raw-reads was performed by using the IMNGS (version 1.0 Build 2007) [159] pipeline based on the UPARSE approach [158]. Sequences were demultiplexed, trimmed to the first base with a quality score <3 and then paired. Sequences with less than 300 and more than 600 nucleotides and paired reads with an expected error >3 were excluded from the analysis. The 5 nucleotides from each end of the remaining reads were trimmed to avoid GC bias and non-random base composition. Operational taxonomic units (OTUs) were clustered at 97% sequence similarity. Taxonomy assignment was performed at 80% confidence level using the RDP classifier and the SILVA ribosomal RNA gene database project. Later the data was normalized using the same method as for RNA-seq described above. The microbiome was visually inspected for batch effects in PCA; none were found. The resulting OTUs table was normalized using edgeR (Version 3.28 or later) [145].

For all the other datasets dada2 [160] (Version 1.14 or later) was used to analyze microbiome data. It creates amplicon sequencing variants from the 16S sequencing data, without merging similar sequences at any threshold. It is an alternative to the use of OTUs which allows to compare results between studies and provides more resolution to identify differences on the fragment of 16S amplified.

We used Silva v138.1 to annotate the 16S fragments whenever possible [161]. If we did not have access to the direct ASV we used the annotation provided by the original authors of the dataset.

3.3 Integration methods

The main method used on this thesis has been regularized generalized canonical correlation analysis (RGCCA) a method derived from the canonical correlation. Canonical correlation is a method that uses data about the same unit but from different origins to find how much the different sources agree. Regularized generalized canonical correlation analysis is implemented on the homonymous package RGCCA [131] which was used here. The method and implementation will be explained in detail in the next section.

3.3.1 Regularized generalized canonical correlation analysis

To understand the regularized canonical correlation analysis we first provide a brief description of principal component analysis.

Principal component analysis defines a new orthogonal coordinate system that optimally describes variance in a single dataset. It does so by decomposing the numerical matrix into the eigenvalues and eigenvectors with decreasing variance. Its results include new variables (the eigenvectors) and a vector of loadings or weight indicating the importance of the original variables for these new variables. Typically, data is normalized and standardized to avoid that a variable with higher variance and different scale dominate the results.

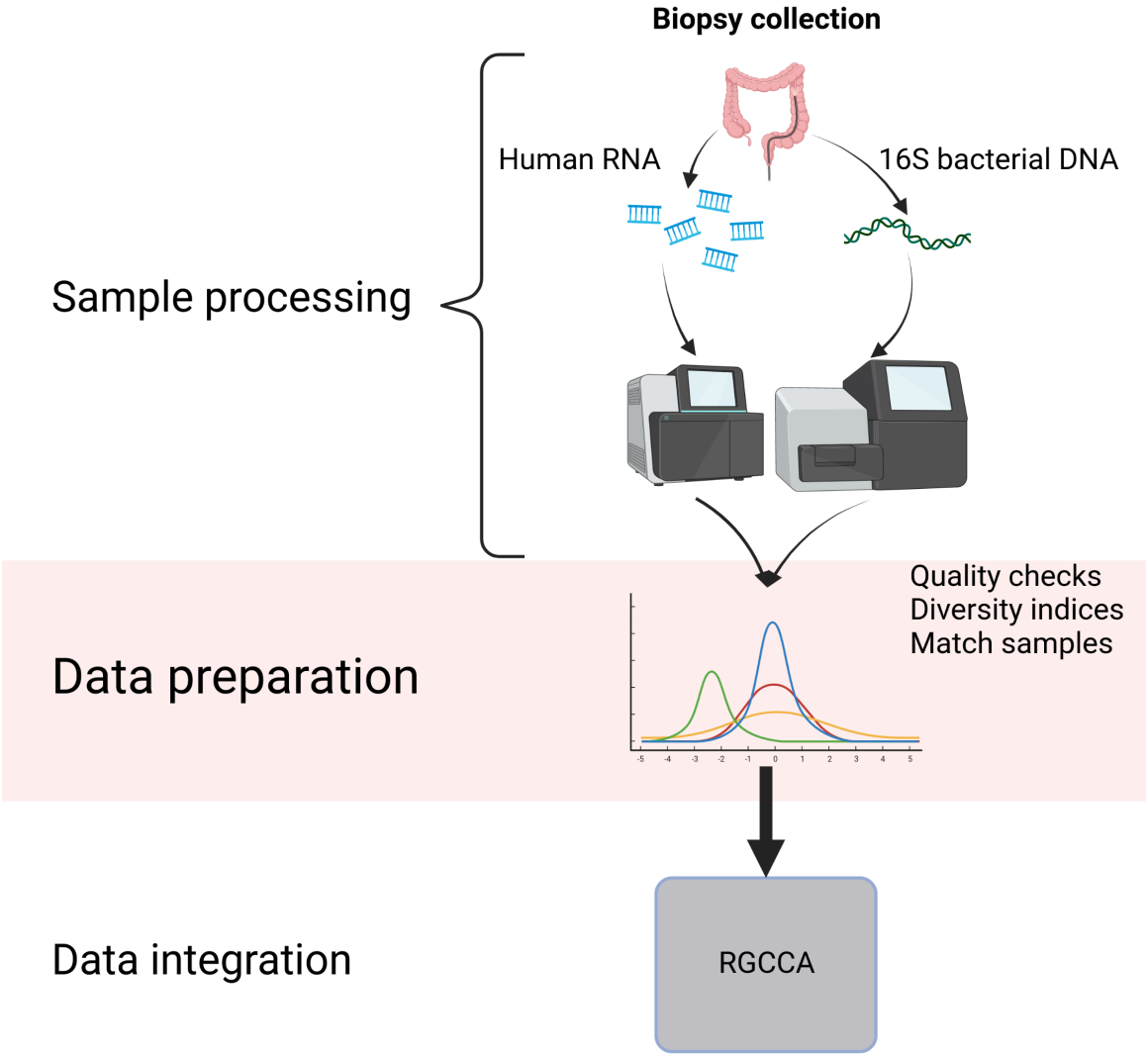

Figure 3.1: Workflow of the main analysis process of the thesis. Created with BioRender.com

The canonical correlation analysis [97, 162] is a method to find agreement between two, or more, scorers (as it was first introduced on the literature) or sources. It extends the PCA to a two sources of data. It provides a similar output, new variables and weights of the existing variables indicating their importance. The function it maximizes is:

$$\rho = \underset{a,b}{\operatorname{argmax}}(\operatorname{cor}(a^T X, b^T Y))$$

On this equation a and b are random variables that given X and Y, the two data sources, maximize the correlation between $a^T X$ and $b^T Y$, the first canonical variables, the new variables. a and b are also known as weights or loadings of the variables.

Over several years of progress on the field of canonical correlations [115, 116, 163–167], regularized generalized canonical correlation analysis (RGCCA) emerged with a generalization from canonical correlations extending the procedure to more than two sources of data [**R-RGCCA?**] and being made more flexible generalizing from other

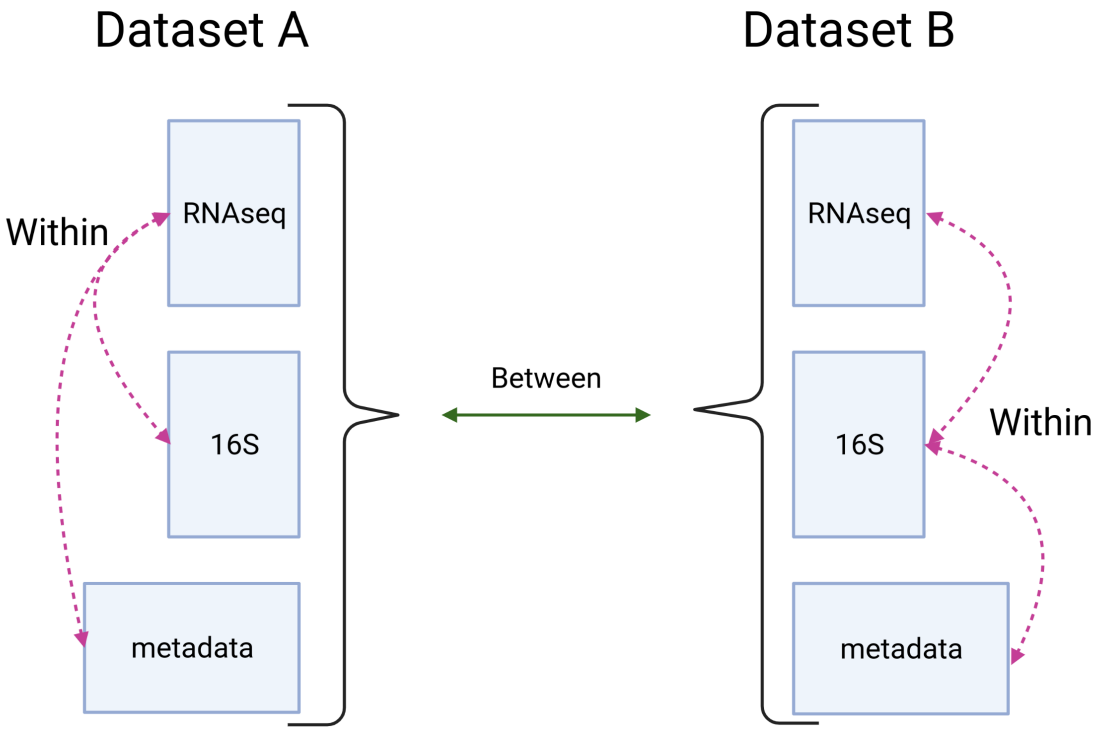

Figure 3.2: Multi-omic relationships on different datasets. Integration methods focus on relationships within datasets. Common relationships between datasets are used as confirmation/validation. Created with BioRender.com

proposed methods.

3.3.1.1 Description

RGCCA works with numeric matrices that can be as big as needed, as it is designed for datasets with more variables than samples ($p \gg n$). However, for each sample it needs to have a complete case with no missing values (at the time of writing this thesis there is an in-development version, not released on CRAN[1] yet, that replaces any missing value by a 0) and in which time is not considered as a special variable. It uses a dimensional reduction approach to relate the different blocks of data between them and produce specific factors for each dataset. The objective function is:

$$\underset{a_1,a_2,\ldots,a_J}{\text{maximize}} \sum_{j,k=1}^{J} c_{jk} g(\text{cov}(X_j a_j, X_k a_k)) \text{ s.t. } (1-\tau_j) var(X_j a_j) + \tau_j \|a_j\|^2 = 1, j = 1, \ldots, J$$

Being X_j the values from sample j, the weights of the variables of said sample are represented by a. While g is a function that can take the form of x, also known as Horst method, $|x|$ known as centroid method, x^2 known as factorial method, or any user-supplied function. C is a symmetric matrix describing the network between blocks.

[1]CRAN is The Comprehensive R Archive Network available at https://cran.r-project.org/

The shrinkage parameter is defined as:

$$\widehat{\lambda}^{\star} = \frac{\sum_{i \neq j} \widehat{Var}(r_{ij})}{\sum_{i \neq j} r_{ij}^2}$$

Where the r_{ij} are the correlation coefficients of the matrix between variables i and j . Where the variance is defined as:

$$\widehat{Var}(S_{ij}) = \frac{n^2}{(n-1)\check{s}} \widehat{Var}(w_{ij}) = \frac{n}{(n-1)^3} \sum_{k=1}^{n} (w_{kij} - \overline{w_{ij}})^2$$

And its components are: $w_{kij} = (x_{ki} - \overline{x}_i)(x_{kj} - \overline{x}_j)$ and $\overline{w}_{ij} = \frac{1}{n} \sum_{k=1}^{n} w_{kij}$ representing x_{ij} the values of a sample j on a variable i.

The authors realized that there is a special problem due to sparsity on biological data which could be handled using first another normalization to improve the stability and success of the canonical correlation methodology. The method to perform the dimensional reduction using the sparse method consists on maximizing the same equation but with a different constraint: Specifically, RGCCA with all $\tau_j = 1$, combined with an L1-penalty that gives rise to SGCCA:

$$\underset{a_1, a_2, \ldots, a_J}{\text{maximize}} \sum_{j,k=1}^{J} c_{jk} g(\text{cov}(X_j a_j, X_k a_k)) \quad \text{s.t.} \quad \|a_j\|_2 = 1 \text{ and } \|a_j\|_1 \leq s_j, j = 1, \ldots, J$$

The s_j controls the sparsity estimated of the data, the smaller it is, the higher the sparsity of X_j is. As s_j is closer to 0, more features are selected as it looks to optimize covariance; while if it is closer to 1, less features are selected and the function resembles the correlation. The values of s_j were estimated using the Schäfer method [168] when the block included 16S data or RNAseq, otherwise 1 was used.

As previously mentioned, there are different g functions that could be used; but the centroid method was chosen to detect both positive and negative relationships.

Categorical data was encoded as binary (dummy) variables for each factor except one to keep degrees of freedom, where 0 indicates not present and 1 indicates present. One level was omitted to avoid overfitting the data. Each block, regardless if it had continuous numeric variables or dummy variables was standardized to zero mean and unit variance. Later, it was divided by the square root of the number of variables of the block for an unbiased estimation.

3.3.1.2 Output

RGCCA, as other dimensional reduction techniques, provides specific weights for each variable on each dimension and a sample score on each dimension, together with quality scores. The most important output are the canonical components of each block. For each block there can be as many canonical components as the number of variables of a block minus 1. These canonical components represent a block by the largest source of variation on that block that satisfies the constrains explained below.

To measure the quality of the model, the implementation provides indicators based on the Average Variance Explained (AVE). RGCCA returns an AVE score for each block, which measures how the variables of the block correlate with the dimension component of the block. It also provides two AVE scores of the whole model: the inner, which measures how each dimension accounts for the variance, and the outer, which measures how variables correlate with the dimension components. The closer the inner AVE is to 1, the better the model adjusts to the data. However, that mathematically fits better to the data does not guarantee that the model provides more insights into the biology.

It can also generate results as other related methods based on the maximization of a function of correlations: SUMCOR (sum of correlations method) [169], SSQCOR (sum of squared correlations method) [170], SABSCOR (sum of absolute values of the correlations method) [171]. Others methods are based on the maximization of a function of covariances: SUMCOV (sum of covariances method), SSQCOV (sum of squared covariances method), SABSCOV (sum of absolute values of the covariances method). The following table summarizes the equivalent parameters needed on RGCCA to work:

Table 3.6: Equivalences of RGCCA to other methods

Method	Scheme	Normalization	Shrinkage
SUMCOR	Horst	$Var(X_j a_j) = 1$	0
SSQCOR	Factorial	$Var(X_j a_j) = 1$	0
SABSCOR	Centroid	$Var(X_j a_j) = 1$	0
SUMCOV	Horst	$\|a_j\| = 1$	1
SSQCOV	Factorial	$\|a_j\| = 1$	1
SABSCOV	Centroid	$\|a_j\| = 1$	1

There are other methods that can be performed with RGCCA, for example: The classical canonical correlation analysis would be equivalent to data variance equal to 1 and shrinkage of 0. Partial least squares (PLS) regression, which maximizes covariance, would be equal to data variance of 1 and shrinkage of 1 in RGCCA. Finally, redundancy analysis could be performed where one block's weight normalized value is 1 and the variance of the other equals to 1.

3.3.1.3 Models

There is no formal definition of what constitutes a block of data on multi-omics tools. Most multi-omics and integration tools assume one block for each type of data, such as an essay a survey or an experiment. We decided to split the block with the variables about the samples to separate independent variables. The hypothesis we made was that more blocks with highly related variables but independent from the other blocks would fit better the data and thus help to identify causal or dependent variables.

To model what might be the relationships within datasets current practices include using a pre-selected model of relations between blocks (See figure 3.2). However, this model might not be an accurate representation of the relations between blocks and

several models might need to be fitted. To help find the fitting model for the data we created an R package, named inteRmodel, which helps finding the right model for the dataset via a bootstrapping procedure.

This method was applied to the previously described datasets to find the relationship between microorganisms and the disease. Following this method; to provide a ground truth, a model with only the relationships between the two experimental obtained data is analyzed, on what it is called the model 0.
The next models analyzed consisted on relationships between the two experimental blocks and a block with all the metadata of the samples. These models are denoted by 1.Y, where 1 denotes the family 1 and Y is used to label some of the models of this family.

Later instead of a big metadata block, following our theory we split this metadata block on several ones, having a block for time related variables, another one for location and the other about the people on the study. This allows to design a model with an expected relationships between these blocks and makes more interpretable the relationships. These models are from the family 2 denoted by the name 2.Y, where 2 denotes the family of the model and Y change for particular models with different relationships between the blocks change.

For each family of models we tested all possible relationships with weights between 0 and 1 by 0.1 intervals to find the best model on each datasets according to the AVE score.

The final models were further validated using a bootstrap approach to measure their accuracy and likelihood on the data available.

3.3.2 Other

3.3.2.1 MCIA

Multiple co-inertia analysis, also known as MCIA, is a method to examine covariant gene expression patterns between two blocks [172]. It is implemented on the package omicade4. On its core MCIA maximizes the following formula:

$$\sum_{k=1}^{K} w_k cov^2(X_k Q_k u_k, v)$$

where K is the total number of matrices, X_k the transformed matrices and Q_k is a square matrix with r_{ij} in diagonal elements indicating the hyperspace of features metrics, u_k are auxiliary axes, v the reference structure and w the weights of the matrices. This can be used to obtain a dimension of $P_k^d = u_k^d(u_k^d Q_k u_k^{dT})^{-1} u_k^d Q_k$ given that for each dimension the residuals are obtained following $X_k^{d-1} = X_1^d - X_1^d P_1^{d-1}$ where d are the dimensions needed.

MCIA was used as a baseline method to compare the RGCCA integration.

3.3.2.2 STATegRa

We used STATegRa To explore how much do different blocks of a dataset have in common [173]. It is a framework for integrating datasets with two data types using parametric and non-parametric methods. The methods used are omics component analysis based on singular value decomposition (SVD) of the data matrix. There are three different methods provided to this end: *DISCO-SCA*, *JIVE* and *O2PLS*.

DISCO-SCA uses:

$$X_k = TP_k^T + E_k$$

Where T is the $I \times R$ matrix of components scores that is shared between all blocks and P_k the $J_k \times R$ matrix of components loadings for block k.

Let $X_1, X_2, \ldots X_i$ be blocks of data and $X = [X_1, X_2, \ldots X_i]$ represent the joint data, then the JIVE decomposition is defined as:

$$X_i = J_i + A_i + \epsilon_i,\ i = 1, 2, \ldots$$

where $J = [J_1, J_2, \ldots J_i]$ is the $p \times n$ matrix of rank $r < rank(X)$ representing the joint structure, A_i is the $p_i \times n$ matrix of rank $r_i < rank(X_i)$ representing the individual structure of X_i and ϵ_i are $p_i \times n$ error matrices of independent entries.

Finally, the O2PLS approach uses multiple linear regression to estimate the pure constituent profiles and divides the systematic part into two, one common to both blocks and one not. The O2PLS model can be written as a factor analysis where some factors are common between both blocks.

$$\begin{aligned}
&\text{X model : } X = TW^T + T_{\text{Y-ortho}}P_{\text{Y-ortho}}^T + E \\
&\text{Y model : } Y = UC^T + U_{\text{X-ortho}}P_{\text{X-ortho}}^T + F \\
&\text{Inner relation : } U = T + H
\end{aligned}$$

Each model is built similarly by adding the subtraction of the projected values of the other component keeping the relationship between them as stated on the third line.

3.4 Functional enrichment methods

3.4.1 Over representation analysis

Functional enrichment methods are those methods that aim to provide with more information about the variables besides their numerical value measured. They can be very different in nature but they all use the numeric values of the variables and other information, being it from the same experiment data collection or from external data sources.

Many functional enrichment methods are based on an over representation analysis, where a group of elements is tested for their mesure in other groups. This can be

done with `clusterProfiler` which tests genes enrichment for functionality based on information on pathway databases [174] that it is used in several publications [175]. `clusterProfiler` checks the enrichment of features of a given group on the (background) list provided.

$$H_0 : P_{subset} \leq P_{overall}$$
$$H_1 : P_{subset} > P_{overall}$$

The statistical test used is usually the fisher test, the hypergeometric test or the proportion test. We describe the hypergeometric test and the proportion test below.

3.4.1.1 Fisher test

The fisher test calculates how independently are two categories. Say if genes in category X are independent of category Y, if we use the contingency table

Table 3.7: Fisher contingency table

	In category X	Not in category X	Row total
In category Y	a	b	$a+b$
Not in category Y	c	d	$c+d$
Column total	$a+c$	$b+d$	$a+b+c+d=n$

Given this contingency table 3.7, the probability of X and Y being independent can be calculated with:

$$\begin{aligned} P &= \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} \\ &= \frac{(a+b)!\ (c+d)!\ (a+c)!\ (b+d)!}{a!\ \ b!\ \ c!\ \ d!\ \ n!} \end{aligned}$$

3.4.1.2 Hypergeometric test

The hypergeometric distribution describes the probability of k successes (when the object drawn has a specified feature) in n draws[2], from a finite population of size N that contains exactly K objects with that feature, wherein each draw is either a success or a failure.

In this context N is the number of genes being used and n the number of genes on a pathway. It can be used to compare the genes found on a pathway (k genes) compared to the expected K numbers of the distribution using the following equation:

$$P_X(k) = P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

[2] without replacement

This looks like the fisher test, because hypergeometric test assesses the extremeness of observing k or more of that overlap (a on) and thus it is the same as a one-sided Fisher's exact test.

3.4.1.3 Proportion test

The overrepresentation of a given group of elements can also be tested with the proportion test, which is sometimes also used on clusterProfiler. The proportion test uses the χ^2 distribution to test if the observed frequency (O_i) is close to the expected frequency (E_i):

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \simeq \chi^2_{n-p}$$

As this is usually done on a 2x2 contingency table it is equivalent to the Z-test of proportion. Sometimes, the expected frequency is so low that a correction must be done to the estimation:

$$\chi^2_{\text{Yates}} = \sum_{i=1}^{n} \frac{(|O_i - E_i| - 0.5)^2}{E_i}$$

This increases the p-value as it raises the Chi-square statistic.

3.4.2 Gene Set Enrichment Analysis

There are other methods that test if some variables show an unexpected importance according to a statistic like fold change or value, such as gene set enrichment analysis (GSEA) [176]. GSEA is a computational method originally developed to determine whether a priori defined set of genes shows statistically significant and concordant differences between two biological states. This methods check if a group of variables present in an ordered list shows a skewed distribution and it compares it against a random group of similar size. It has been widely used since its original publication, also on IBD [177].

This method calculates the rank of genes $rank(g_j) = r_j$ where each g is a gene, and then it calculates the following functions:

$$P_{hit}(S, i) = \sum_{g_j \in S, j \leq i} \frac{|r_j|^p}{N_R}, \text{ where } N_R = \sum_{g_j \in S} |r_j|^p$$

$$P_{miss}(S, i) = \sum_{g_j \notin S, j \leq i} \frac{1}{N - N_H}$$

With these values the enrichment score (ES) defined as: $ES = max(|P_{hit}(S, i) - P_{miss}(S, i)|)$ is calculated from the walk. At least 1000 permutations are usually used but a high number of permutations are required for an accurate estimation of the enrichment score. However, when more than one pathway (S) is evaluated in order to compare between their enrichment scores, they must be normalized by dividing the

scores by the mean of all the ES. When power p is 0 it is equivalent to the standard Kolmogorov–Smirnov statistic, though it is usually set to 1.

For testing GSEA we used `fgsea` [178] implementation for its speed and integration with other methods used in this thesis. Gene pathways from the REACTOME database were tested on the weight of different models or on the comparisons performed [179].

3.4.2.1 GSVA

To estimate the expression of the pathways and compare their expression levels between conditions gene set variation analysis as implemented on GSVA was used [180]. It is a method that summarizes the variables' numerical value changing the space of variable x sample to group x sample . This enables other methods to use this new space instead of the original variables, which provides a successful way to look into data [181]. GSVA was used (again from the REACTOME database) to find the relationships between the pathways and the microbiome at different taxonomic levels.

This is done via an estimation and a comparison with a discrete Poisson kernel: i indicates the gene from a total of k genes, and samples are indicated by j from n number of samples.

$$z_{ij} = \hat{F}_r(x_{ij}) = \frac{1}{n}\sum_{k=1}^{n}\sum_{y=0}^{x_{ij}} \frac{e^{-(x_{ik}+r)}(x_{ik}+r)^y}{y!}$$

$r = 0.5$ is used to set the mode of the Poisson kernel at each x_{ik}, that is, similar to the expression of a gene for a given sample.

Later this is converted to ranks $z_{(i)j}$ for each sample and normalized: $r_{ij} = |\frac{p}{2} - z_{(i)j}|$ to make the distribution of ranks symmetric around zero to later compare with a normal distribution using a Kolmogorov-Smirnov-like random walk statistic:

$$v_{jk}(l) = \frac{\sum_{i=1}^{l} |r_{ij}|^\tau I(g_{(i)} \in \gamma_k)}{\sum_{i=1}^{p} |r_{ij}|^\tau I(g_{(i)} \in \gamma_k)} - \frac{\sum_{i=1}^{l} I(g_{(i)} \notin \gamma_k)}{p - |\gamma_k|}$$

Here τ describes the weight of the tail in the random walk (default is set to 1). γ_k is the k-th gene set and $I(g_{(i)} \in \gamma_k)$ is the indicator function whether the gene ranked i-th belongs to the gene set γ_k . $|\gamma_k|$ indicates the ordination of the gene set, the number of genes of the gene set and p the number of genes in the dataset.

This difference is later converted to enrichment score for each gene set for each sample, similar to GSEA. This score can be calculated as a difference of hits and misses or the maximum deviation from zero of the random walk (which allows to detect gene sets that have genes with different opposing expression patterns).

3.5 Variance and diversity methods

3.5.1 PERMANOVA

The PERMANOVA method [182, 183], provided by the vegan package on the `adonis` function, was used to test if microbiome data variance is due to other variables when using distances metrics. It uses the residual sum of squares such as:

$$SS_W = \frac{1}{n} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} d_{ij}^2 \epsilon ij$$

When using euclidian distances (d) it is equivalent to MANOVA. Here ϵ_{ij} takes the value of 1 if the observation i and the observation j are in the same group, otherwise it takes the value of 0. This can be later used to test which variance is bigger, inter-groups or intra-groups by using the following formula:

$$F = \frac{SS_A/(\alpha - 1)}{SS_W/(N - \alpha)}$$

Where SS_A is the among group sum of squares, representing the intra-group variance. N is the number of samples and α the number of different groups.

This allows to test if the variables are related to the variance of the data as it can be compared with the F statistic after a high number of permutations.

3.5.2 globaltest

It is a method for testing complex hypothesis and calculate the influence of each variable on a given outcome [184]. We tested which variables, (sex, age, location, time since diagnosis, treatment) are important on the datasets with globaltest (Version 5.40 or later). This method provides a general statistic to test hypothesis against a high dimensional dataset.

$$S = \sum_{i=1}^{p} x_i' x_i g(t_i^2)$$

The global test performs a test statistic on the transformed t-test, where if p, the number of variables, is large the test is more powerful on average over all possible sparse alternatives of general functions g.

It was performed with variables individually and also with interactions between the different variables.

3.5.3 Diversity indices

Microbiome diversity was measured using vegan and phyloseq methods [185]. α-diversity is a measure of how much a given microbiome at a taxonomic level is present

on a sample. Several measures exists, on the thesis we used the effective Simpson or effective Shannon diversity index to compare diversity between samples and conditions.

The effective Simpson (also known as inverse Simpson) and the effective Shannon are:

$$D_{\text{effective Simpson}} = \frac{1}{\sum_{1=1}^{S} p_i^2}$$
$$D_{\text{effective Shannon}} = \frac{1}{-\sum_{i=1}^{S} p_i \log_e p_i}$$

Where p_i is the proportion of species i and S is the number of species.

β-diversity is a diversity index that compares how similar are two samples. It was calculated using the phyloseq package for exploratory analysis.

3.6 Other methods

3.6.1 Statistics

Differential expression analysis was performed with the limma-trend method [186, 187] and edgeR [145] (Bioconductor version 3.10 or superior) packages. Data was normalized using the trimmed mean of M-values and log-2 transformed into counts per millions following the workflow previously described using voom [188]. This approach assumes tests for each gene that:

$$H_0 : \mu_X = \mu_Y H_1 : \mu_X \neq \mu_Y$$

Limma assumes that $E(y_g) = X\alpha_g$, and $var(y_g) = W_g\sigma_g^2$ where X is a design matrix and α_g is an unknown coefficient vector and σ_g^2 is the gene-variance and W_g is a known non-negative definite weight matrix. The design matrix of limma can be combined with contrast estimators defined such that $\beta_g = C^T\alpha_g$ where C^T is a constant vector typically defining how a given variable is related to a sample.

These statistics are later estimated via bayesian method whose prior is defined as:

$$\frac{1}{\sigma_g^2} \sim \frac{1}{d_0 S_0^2}\chi_{d_0}^2$$

This results on a moderated t-test:

$$\tilde{t}_{gj} = \frac{\hat{\beta}_{gj}}{\tilde{S}_g\sqrt{v_{gj}}} \sim t_{d_g+d_0}$$

Were $\hat{\beta}_g$ can be interpreted as the difference of the effect of two factors.

As gene comparisons are done along many genes there is an increase probability to have a type I error. Multiple testing correction method have been designed to

correct this kind of family-wise comparisons. To correct for multiple testing, the false discovery rate was estimated using the method of Benjamini and Hochberg [189]. A gene was considered differentially expressed when it was significant at 5% FDR. Special attention was given to those genes that showed a fold-change higher than |1.5|.

3.6.2 WGCNA

To look for relationships between the microbiome and the RNAseq we used weighted gene co-expression network analysis. We used weighted gene co-expression network analysis as implemented on WGCNA [190] as well as correlations. The Spearman rank correlation coefficient is:

$$R_s(X,Y) = \frac{\sum_{i=1}^{n}(R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_{i=1}^{n}(R_i - \bar{R})^2}\sqrt{\sum_{i=1}^{n}(S_i - \bar{S})^2}}$$

Being $(X_1, Y_1), \ldots, (X_n, Y_n)$, assign a rank where $(R_1, S_1), \ldots, (R_n, S_n)$ for n being all the variables where $\bar{R} = \frac{1}{n}\sum_{i=1}^{n} R_i$ and likewise $\bar{S} = \frac{1}{n}\sum_{i=1}^{n} S_i$. The distribution of the Spearman correlation coefficient is symmetric around 0 and can be approximated to a normal distribution as $\sqrt{n-1}R_{s(X,Y)} \sim N(0,1)$ which can be used to calculate the p-value of a given estimation.

3.6.3 BaseSet

BaseSet was developed and used to find which variables are really involved on the interaction and how likely they were to be together. It is a package that uses fuzzy set logic to calculate the probability to belong to a group, in this case, those genes and bacteria selected by the model that interact with the other.

Under the standard fuzzy set logic a set S is a group of elements for which each element e has an α membership to that set [191, 192]. α is usually bounded between 0 and 1: $\alpha \in [0,1]$. A given element e can belong to more than one set. Assimilating the membership function to probability we can calculate the probability of a given element e to belong to a set S and not any other set: $P(e \in S | e \notin S^c)$. Which applied to the data and the case at hand, it is the probability that a given variable is associated with a given outcome and not with any other outcome.

The membership function was derived from the bootstraps used for each model on the thousand iterations of the integrative method applied to give an estimation of how probable is a given gene and bacteria to be selected as relevant for the model. The bootstraps of the models where used to calculate the probability of a variable to be selected by RGCCA. This probability was used to calculate (via `set_size`) the genes and bacteria that are specific of the model that allows to separate the transcriptome by its location and the microbiome by the disease status.

3.6.4 experDesign

experDesign was developed [193] to prevent and quantify if a given experiment has

batch effect due to the batches used to measure the values or other known variable. It might help to detect a bad design of the experiment.

On pseudo code the core of the program can be described as:

```
for each index:

    for each batch

        pick size(batch) samples

        if samples are in another batch

            pick other samples

    for each batch

        calculate some summary statistics

        compare with the summary statistics of all the samples

keep the index with less differences between the index and all the samples
```

Summary statistics taken into account are the median, the variance, the range, the number of missing values, and the entropy of the categorical variables. It can take into account spatial distribution and, given the number of samples that fit on a batch, provide which technical replicates[3] are best to use.

3.6.5 ROC- AUC

To estimate if the selected features (genes or microorganism) by the integration methods have some biological meaningful contribution we measured if they can classify features, such as, which gastrointestinal segment is each sample from, or which type of disease does each patient have.

To compare between different models the area under the curve (AUC) of the receiver operating characteristic (ROC) was calculated with the pROC package [194]. It is based on the following formulas, where FP is false positive, N is a negative, P is positive, TP is true positive and FN is false negative:

$$FPR = \frac{FP}{N} = \frac{FP}{FP + TN}$$
$$TPR = \frac{TP}{P} = \frac{FP}{FP + FN}$$

[3]Technical replicates are samples that are measured multiple times to ensure the accuracy of the measurement. In contrast biological replicates are sample with the same conditions but different individuals or process before the measurement.

The ROC curve is that where the true positive rate (TPR) or sensitivity, recall or hit rate is represented against the false positive rate (FPR) on the abscissa. The area under this curve is a measure of how good such classifications performs overall, being 0.5 as good as a random selection. The closer it is to 1 the better as it classifies incorrectly less samples and accurately classify more.

Results

4.1 Packages/methods

4.1.1 experDesign

`experDesign` package built in R was released for the first time on CRAN on 2020-09-08 after nearly a year after the initial release made on github. After peer review it was published on a journal on 2021-11 [193].

The package uses functional programming to create and modify objects and the features used. The package bases its performance on the large body of work made by the R core team. It adds the information to the introduced `data.frame` or returns an vector with the appropriate information.

`experDesign` functions are divided into several categories:

- Helper functions to aid on deciding how many batches are or how many samples per batch. There are some also that report how good a given distribution of the samples felt for a given dataset.
- Functions generating indexes.
- Functions distributing the samples on indexes

Regarding time related variables `experDesign` will use them as factors, while issuing a warning to the user.

Since its development it has been used on a couple of RNA sequencing experiments that required a batch design, one of organoids bulk RNA-seq (data not related to this thesis) and another one of biopsies bulk RNA-seq from the BARCELONA cohort (See appendix section D.1). It was also used to check if there is any observable batch effect on the datasets analyzed.

On the designed datasets `experDesign` avoided batch effects from the sequencing process. However, on the organoids dataset, a change on the matrigel used to produce them introduced a batch effect that made it impossible to compare samples before and after that change (there were not any shared sample before and after the change of matrigel). On the BARCELONA cohort there were other problems described on the appendices (section D.1).

Since its release on CRAN it has had a median of ~400 downloads each month from RStudio repository mirror, showing the interest the community have on solutions like this.

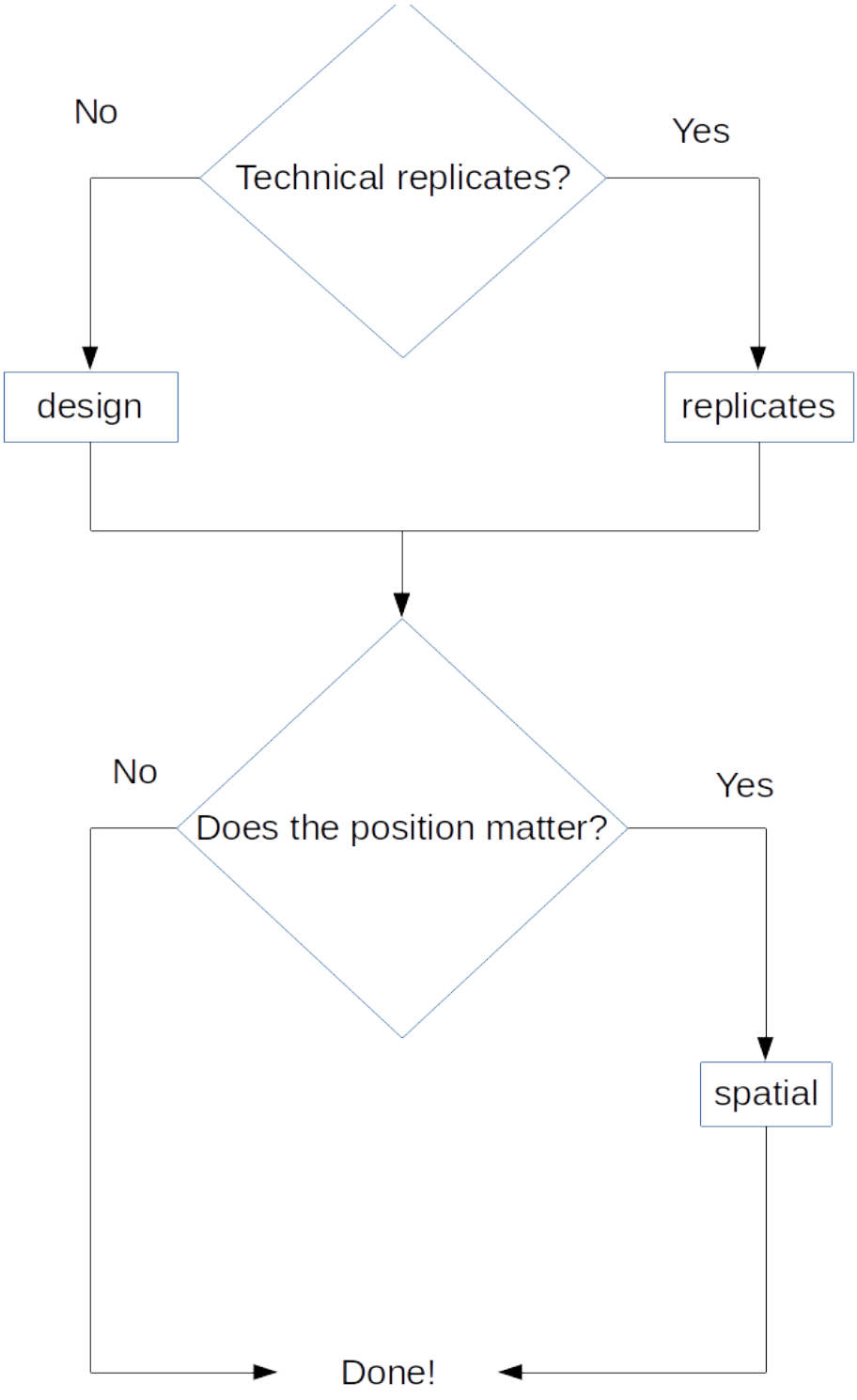

Figure 4.1: ‘experDesign‘ functions and workflow. Workflow for users of the package showing which functions can be used depending on the experiment design they have.

4.1.2 BaseSet

`BaseSet` package, built in R, was released for the first time on CRAN on 2020-11-11, nearly two years after the initial work started on github.

The package uses both functional programming and object oriented program to create and modify the TidySet S4 object defined[1]. Mixing it with S3 generic functions it provides a powerful interface compatible with the tidyverse principles, a group of packages following the same design. The package provides a new class to handle fuzzy sets and the associate information.

`BaseSet` methods are divided into several categories:

- General functions to create sets of the TidySet class or convert from it to a list or about the package.
- Set operations like adjacency cartesian product, cardinality, complement, incidence, independence, intersection, union, subtract, power set or size.
- Functions to work with TidySets to add relationships, sets, elements or some complimentary data about them. Remove the same or simply move around data or calculate the number of elements, relations and sets.
- Functions to read files from formats where sets are usually stored in the bioinformatician field: GAF, GMT and OBO formats.
- Last, some utility functions to use set name conventions and other auxiliary functions.

The package had a long development process with initial iterations based on GSEABase package which was later abandoned (GSEAdv) to also include some uncertainty on the relationship of a gene with a given gene set.

The package formed part of an exploration of the Bioconductor community (project to develop, support, and disseminate free open source software that facilitates rigorous and reproducible analysis of data from current and emerging biological assay) for more modern and faster handling of sets than GSEABase. There were three different packages created as part of this process, `BaseSet`, `BiocSet` released on Bioconductor and `unisets`, available on github. The three different approaches were presented at a birds of feather on BioC2019, the annual conference of Bioconductor on 2019.

The package passed the review on the rOpenSci organization (See review) and is now part of the packages hosted there too.

Since its release on CRAN it has had a median close to ~400 downloads each month from RStudio package manager.

[1] S4 is one of the object programming paradigms on R. For a more complete overview and differences see Advanced R [195].

4.1.3 inteRmodel

The package was build once the method used to find accurate models of the relationships of the data available of a dataset using RGCCA was established. Using the package on github simplifies the process and makes easier to redo the model optimization used on this thesis.

The package has functions that can be grouped in three categories:

- Look for models and evaluate them: To search for a model given some conditions, such as that all the blocks are connected, and check the models via bootstrapping or leave one out procedures.
- Reporting: To make better reports by improving handling of names or simplifying the objects or how to calculate scores.
- Building: To easier build correct models on RGCCA, simplifying the process to create a symmetric matrix.

Currently it is only available on github, so the number of downloads and usage is unknown but since its release a user has contacted to keep it up to date with development versions of RGCCA. Currently, it is compatible with the next release of RGCCA being prepared[2].

The functions `analyze` helps to analyze code of a single integration, providing the results on a tidy format for further processing. To create the connections between blocks the function `weight_design` is available. It creates all the possible matrices given a number of blocks and a number of weights. Optionally it can create just a subset of those based on a numeric vector. However, it does not provide a way to have the design named.
If the user wants to create their own design matrices, they can use `symm` and modify the design of the model with `subSymm`. `symm`, takes an initial matrix to pick up the row and column names. It is recommended that the user checks the design matrix is fully connected, which the package facilitates with the function `correct`. This is also recommended even if the design matrices are created with `weight_design`.

To search models `search_models` starts with a initial connectivity of the blocks and creates all the combinations of connections given.

For the bootstrap procedure there is the function `boot_index` to create the bootstrapped index of samples to be used by `boot_index_sgcca`. `boot_index` randomly selects as much samples as specified by the arguments to create as much indices as the required by the second argument. If the bootstrapped samples used is not important one can use `boot_samples_sgcca`. If the users want to perform a leave-one-out procedure they can use `looIndex`.

For more information, you can access the manual online or once it is installed.

[2]We also contributed with some comments and feedback to the package to make it easier to read the source and check the inputs and improve the documentation so that it is coherent with the code and previous results of the functions.

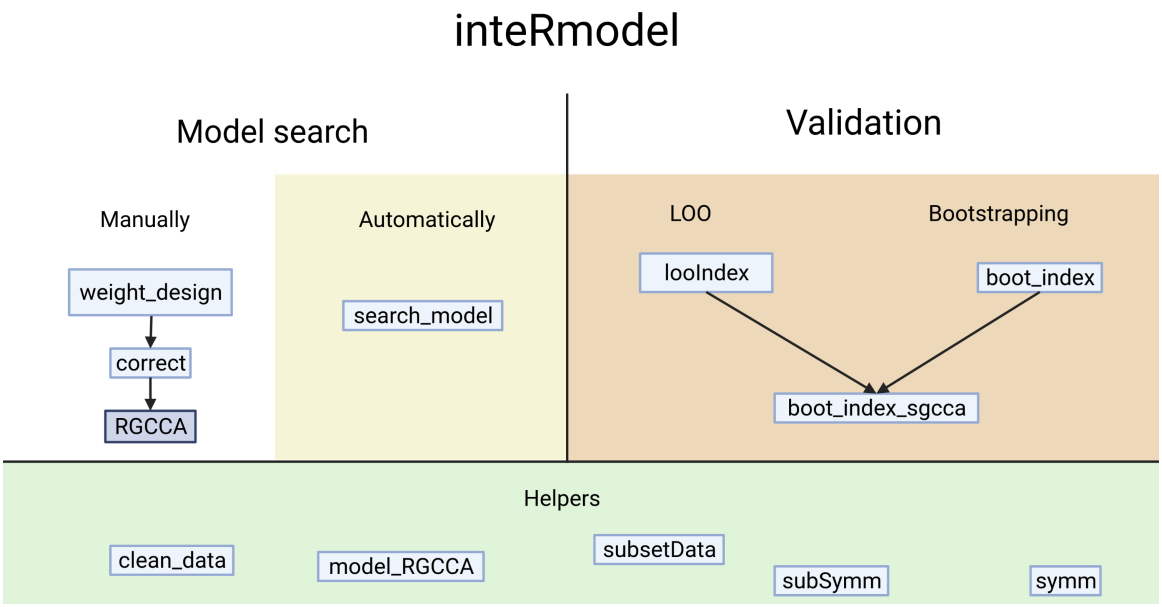

Figure 4.2: 'inteRmodel' functions and workflow. Functions provided by the inteRmodel package to search and validate models of relationships using RGCCA. Created with BioRender.com

The package cannot choose which variables use from the block with information to split into several blocks. However, it provides the `model_RGCCA` function to make it easy to prepare such variables for RGCCA input. The inteRmodel procedure will only be useful if the user should decide which variables are independent from others and split them into different blocks. To asses this the user can use the methods we used, as described on the above section 3.5. It is important to keep in mind the possible causal relationships on user's data [196].

4.2 Analysis

On the following sections the main results of analyzing each dataset are presented.

4.2.1 Puget's dataset

On this dataset the different parameters and capabilities of RGCCA were tested.

The three different methods, centroid, factorial or horst were tested and compared. The main result of this comparison was that the differences of the selection of the variables mattered more than the number of variables selected with each method. The models were tested with different weights on all three schemes: horst, centroid and factorial. The horst and the centroid scheme were similar while the factorial resulted in the most different AVE values (see S1 Data of [197]). The centroid scheme was selected because it takes into account all the relationship regardless of the sign of the canonical correlation between the blocks. It is similarity to horst scheme.

The effect of the sparsity value was measured by its effect on the inner AVE scores and the combination of the different values for each block as can be seen on Figure 4.3.

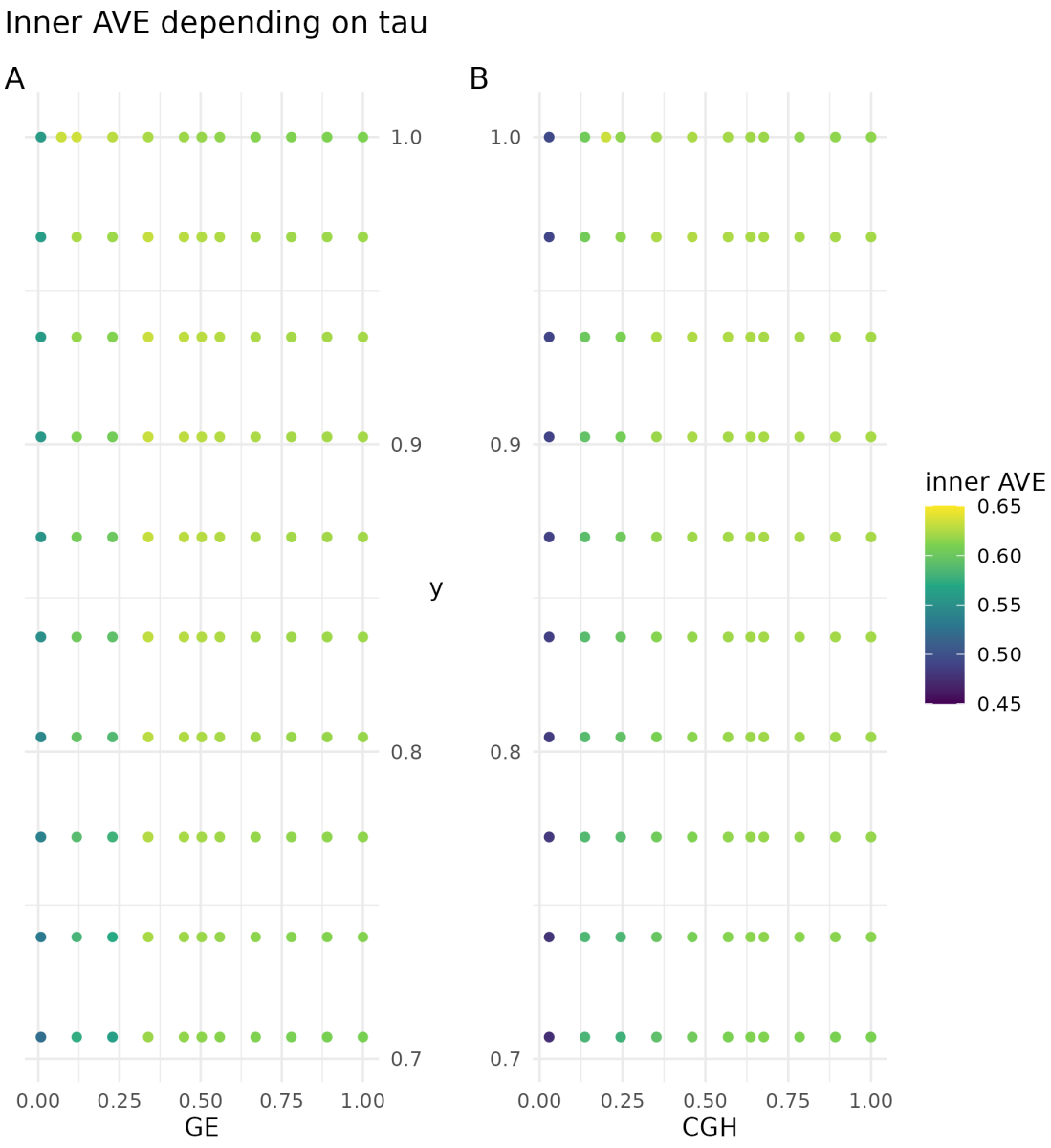

Figure 4.3: Effect of tau on the inner AVE on Puget's dataset. The suggested tau value is the column between the regular grid, on the ordinate axis the y's tau values and on the abscissa the gene expression (GE) on the left and the comparative genomic hybridization (CGH) on the right. The highest inner AVE is with high tau values for y and middle to upper values for GE and CGH.

The first model of the family of models 1 can be seen on Table 4.1:

Table 4.1: Model 1 for Puget's dataset. Relationships between the different blocks in the Puget dataset for model 1. 0 indicates no relationship and 1 indicates a strong relationship.

Model 1	***GE***	***CGH***	***Localization***
GE	0	0	1
CGH	0	0	1
Localization	1	1	0

When looking for the model that adjust better following this structure we arrived to model 1.2, described below (Table 4.2) :

Table 4.2: Model 1.2 for Puget's dataset. Relationships between the different blocks in the Puget dataset for model 1. 0 indicates no relationship and 1 indicates a strong relationship.

Model.1.2	***GE***	***CGH***	***Localization***
GE	0	0.0	1.0
CGH	0	0.0	0.1
Localization	1	0.1	0.0

On model 2 we split the invariable variables from those related to the location (Table 4.3):

Table 4.3: Model 2 for Puget's dataset. Relationships between the different blocks in the Puget dataset for model 1. 0 indicates no relationship and 1 indicates a strong relationship.

Model.2	***GE***	***CGH***	***Invariable***	***Localization***
GE	0	1	1	1
CGH	1	0	1	1
Invariable	1	1	0	0
Localization	1	1	0	0

Following this split, the model that has higher inner AVE for these blocks is the following (Table 4.4):

Table 4.4: Model 2.2 for Puget's dataset. Relationships between the different blocks in the Puget dataset for model 1. 0 indicates no relationship and 1 indicates a strong relationship.

Model.2.2	***GE***	***CGH***	***Invariable***	***Localization***
GE	1	1/3	0	1
CGH	1/3	0	1/3	0
Invariable	0	1/3	0	0
Localization	1	0	0	0

If we added a superblock with all the data of the different blocks from model 1 we started with the standard relationship between blocks (Table 4.5):

Table 4.5: Relationships between the different blocks in the Puget's dataset for model superblock. 0 indicates no relationship and 1 indicates a strong relationship.

Model.superblock	*GE*	*CGH*	*Superblock*	*Localization*
GE	0	0	1	0
CGH	0	0	1	0
Superblock	1	1	0	1
Localization	0	0	1	0

But when the best model with the superblock that had highest inner AVE is quite different (Table 4.6):

Table 4.6: Relationships between the different blocks in the Puget's dataset for model superblock.2. 0 indicates no relationship and 1 indicates a strong relationship.

Model.superblock	*GE*	*CGH*	*Superblock*	*Localization*
GE	1	1/3	0	1
CGH	1/3	0	1	0
Superblock	0	1	0	0
Localization	1	0	1	0

Exploratory analysis with the superblock model was done. The first two components of the superblock did not help to explain the biology or classify the tumors (See 4.4):

The same data was used to look for a good model from the data itself including a model with a superblock but looking at the first component of the CGH and transcriptome block. This allowed to visually inspect if each model's components helped to classify the samples (Figure 4.5):

The first components of the CGH and the transcriptomics blocks of the superblock.2 model show better classification than that of the superblock. However, the other models show a better classification of the samples with much simpler models.

To find these models the three blocks with the best tau and the centroid scheme were analyzed by changing the weights between 0 and 1 by 0.1 intervals. According to the inner AVE, the best model was the one in which the weights (1) between the host transcriptome and location, (2) the host transcriptome and the CGH, and (3) the CGH block were linked to variables related to the location with weights of 1, 0.1 and 0.1, respectively.

When we added a superblock to the data, there was a slight increase of 0.01 on the inner AVE of the model (See Table 4.7). The model with the superblock that explained most of the variance was that in which the weights of the interaction within (1) the host transcriptome, (2) between the superblock and the CGH, (3) between the host transcriptome and the localization, and (4) between CGH and the host transcriptome were 1, 1, 1 and 1/3, respectively (See table 4.6. To see if the superblock could classify the sample by location, we plotted the first two components of the superblock.

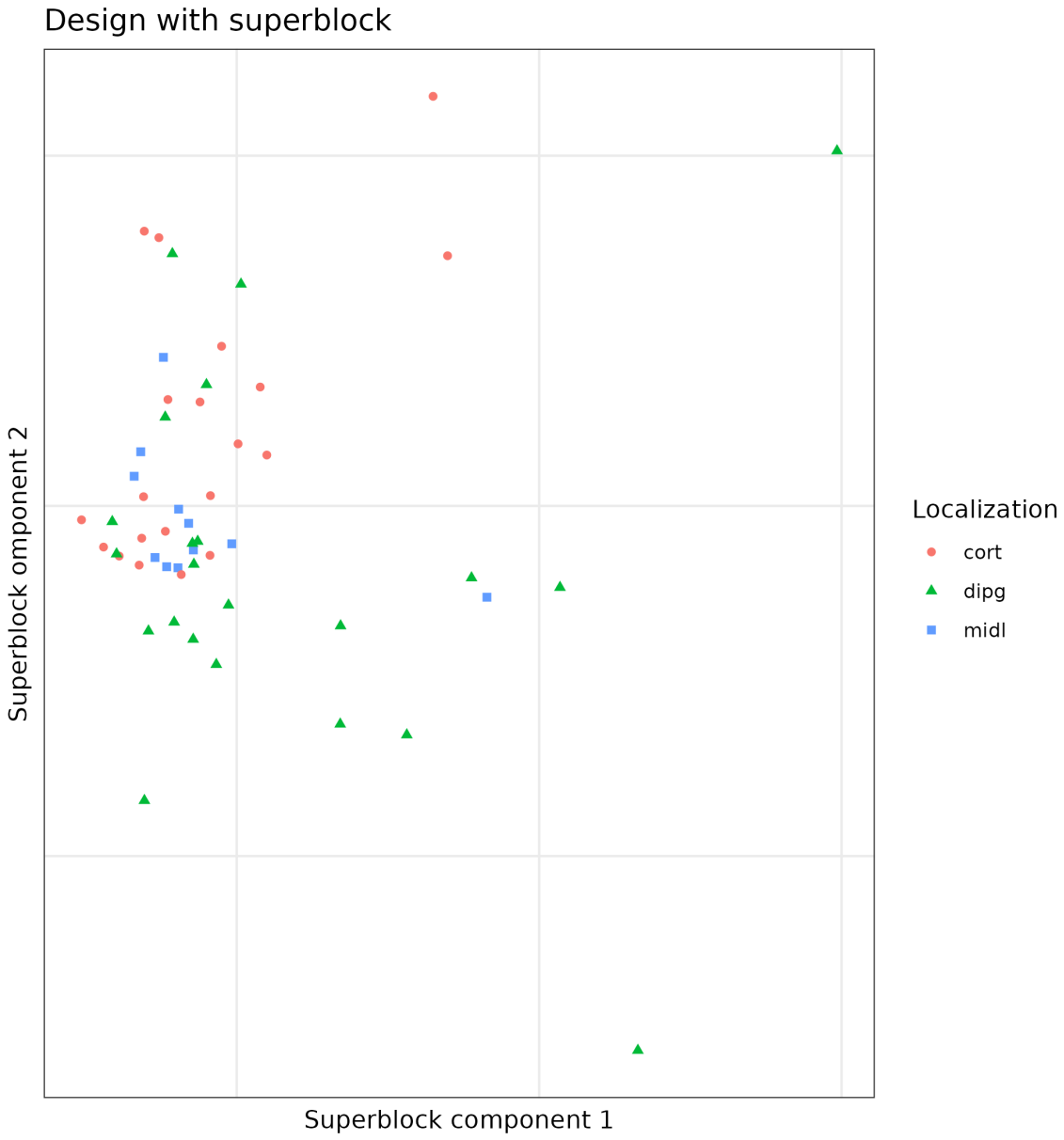

Figure 4.4: First components of the superblock which has all the data of the samples on the Puget's dataset.

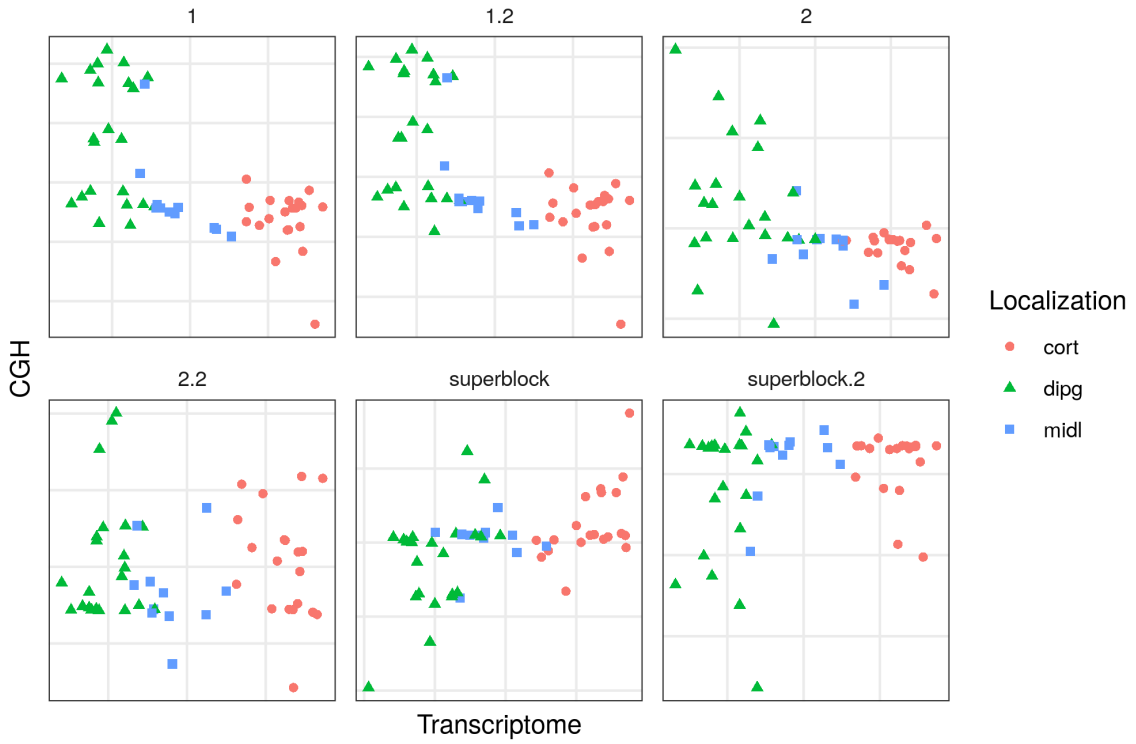

Figure 4.5: Different RGCCA models in the Puget's dataset. The different models with the same data showing the sample position on the first components of the CGH and the transcriptome block. Model 1 and 1.2 with transcriptomics, CGH data and all the data about the samples together. Model 2 and 2.2 with transcriptomics, CGH data and all the data bout the samples on different blocks. Model superblock and superblock.2 have all the data in different blokcs and one block with all the data.

We can clearly see that they do not classify the samples according to the location of the tumor, which is known to affect the tumor phenotype [149].

Adding one block containing the age of the patient and the severity of the tumor to the model, decreased the inner AVE. The best model with these blocks, according to the inner AVE, was that in which the interactions (1) within the host transcriptome, (2) between the host transcriptome and the localization, (3) between the host transcriptome and(4) the CGH and between the CGH and the other variables were 1, 1, 1/3 and 1/3, respectively. The first components of each model can be seen in the figure:

We can observe on the figure 4.5, the strong dependency between gene expression and location since the first model while the weaker relationship with the CGH assay [149]. On the other hand, the major difference is the dispersion on the CGH component on each model.

The effect of the superblock and weights on different models to the inner AVE. There are significant differences between having the superblock and not having it.

The different models resulted on the following AVE values as reported on 4.7:

Figure 4.6: Effect of superblock and weights on the inner AVE on Puget's dataset. Designs with the superblock showed higher inner AVe scores than without it. Interaction yes/no indicates RNA and RNA interaction.

Table 4.7: AVE values of RGCCA models in the Puget's dataset. Values for both inner and outer AVE of the first canonical component of models 1, 1.2, 2 2.2 and superblock and superblock.2 are shown.

Model	***inner AVE***	***outer AVE***
1	0.6333592	0.0692097
1.2	0.8512360	0.0692319
2	0.2791546	0.0738695
2.2	0.6902329	0.0692707
superblock	0.7055847	0.0734578
superblock.2	0.8047477	0.0695821

4.2.2 HSCT dataset

The permanova analysis was performed on this dataset to estimate which were the variables that are more relevant. From the many variables the location, sex, patient id and others were found to be related to the variability of the microbiome or the

transcriptome on this dataset.

With the permanova analysis we found that more of the 50% of the variance of normalized RNA-seq data and microbiome data respectively is explained by the variables of location, disease, sex, and the interaction between disease and sex. On the transcriptome the most important factor is location which is more than 15% of the variance, while on the microbiome data the most important factor is the patient id followed by location of the sample.

Table 4.8: Permanova analysis of transcriptome. The variables and their interactions (shown with :) and the R^2 values and the associated p-value. The higher the R^2 the more variance is explained by that factor.

Factor	***R***	***p-value***
Location	0.18057	0.00020
IBD	0.03120	0.00020
Sex	0.01306	0.00120
IBD:Sex	0.01279	0.11798
Location:IBD	0.02427	0.11458
Location:Sex	0.00816	0.03519
Location:IBD:Sex	0.00486	0.52190
Residuals	0.72508	

Table 4.9: Permanova analysis of microbiome. The variables and their interactions (shown with :) and the R^2 values and the associated p-value. The higher the R^2 the more variance is explained by that factor.

Factor	***R***	***p-value***
Location	0.06061	0.0002000
IBD	0.04967	0.0002000
Sex	0.01712	0.0003999
IBD:Sex	0.01091	0.6604679
Location:IBD	0.02089	0.8476305
Location:Sex	0.01139	0.0075918
Location:IBD:Sex	0.00289	0.9994001
Residuals	0.82652	

With globaltest the results were similar: sex, ibd, location, age and time since diagnosis were able to explain the SESCD score (p-value $5.7 \cdot 10^{-21}$). The resulting p-value was well below the 0,05 threshold defined for RNA-seq data on the models including the segment of the sample, sex and treatment.

On the microbiome data the results were similar but the p-value was considerably higher but still below the threshold.

Diversity indices of the samples were explored and compared for several subsets. Splitting by location of the sample and disease provided the highest differences and the diversity index along time did not change much.

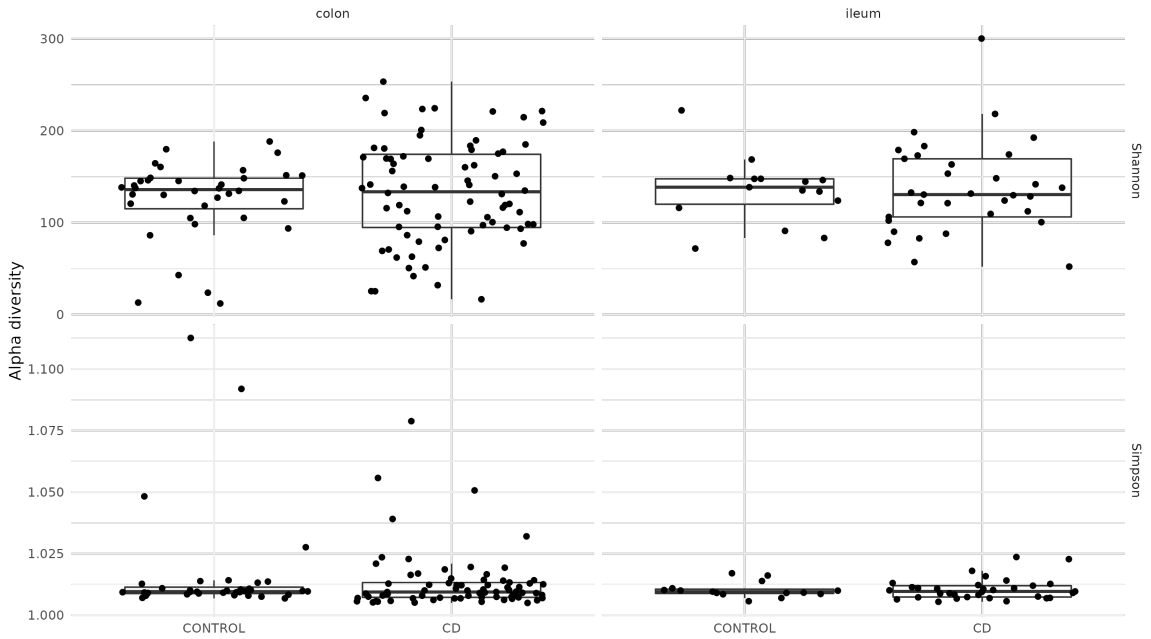

Figure 4.7: Microbiome diversity in the HSCT dataset. On the upper section the Shannon effective and on the lower row the Simpson effective diversity splitted by colon and ileum and controls and CD.

The PCA didn't show any pattern on the microbiome according to the location, as can be seen on the first dimensions of the PCA on figure 4.8:

The PCA on the transcriptome shows a clear pattern splitting by location of the sample on the first dimension (see figure 4.9).

Weighted gene co-expression network analysis did not provide relevant links between bacteria and transcriptome as it failed to find an acceptable scale free degree. As can be seen on the Figure 4.10, the scale free topology does not reach the recommended threshold of 0.9 and the mean connectivity is also very low even for the first power.

STATegRa was used between stool 16S data and intestinal 16S data under the assumption that there is a shared common factor without influence of other categorical variables. However, it did not find a good agreement between these two data sources and 16S data source was not longer used on the analysis. In addition, the model is fixed, so it did not allow to find new or other relationships that are not one to one.

With RGCCA we could select different models and use all the data available without much assumptions. The models with the highest inner AVE of the family 1 and the family 2 models were similar to those on the Häsler dataset.

The weights of these models can be observed here:

Table 4.10: Relationships between the different blocks in the HSCT dataset for model 0. 0 indicates no relationship and 1 indicates a strong relationship.

Model 0	***Transcriptome***	***Microbiome***
Transcriptome	0	1
Microbiome	1	0

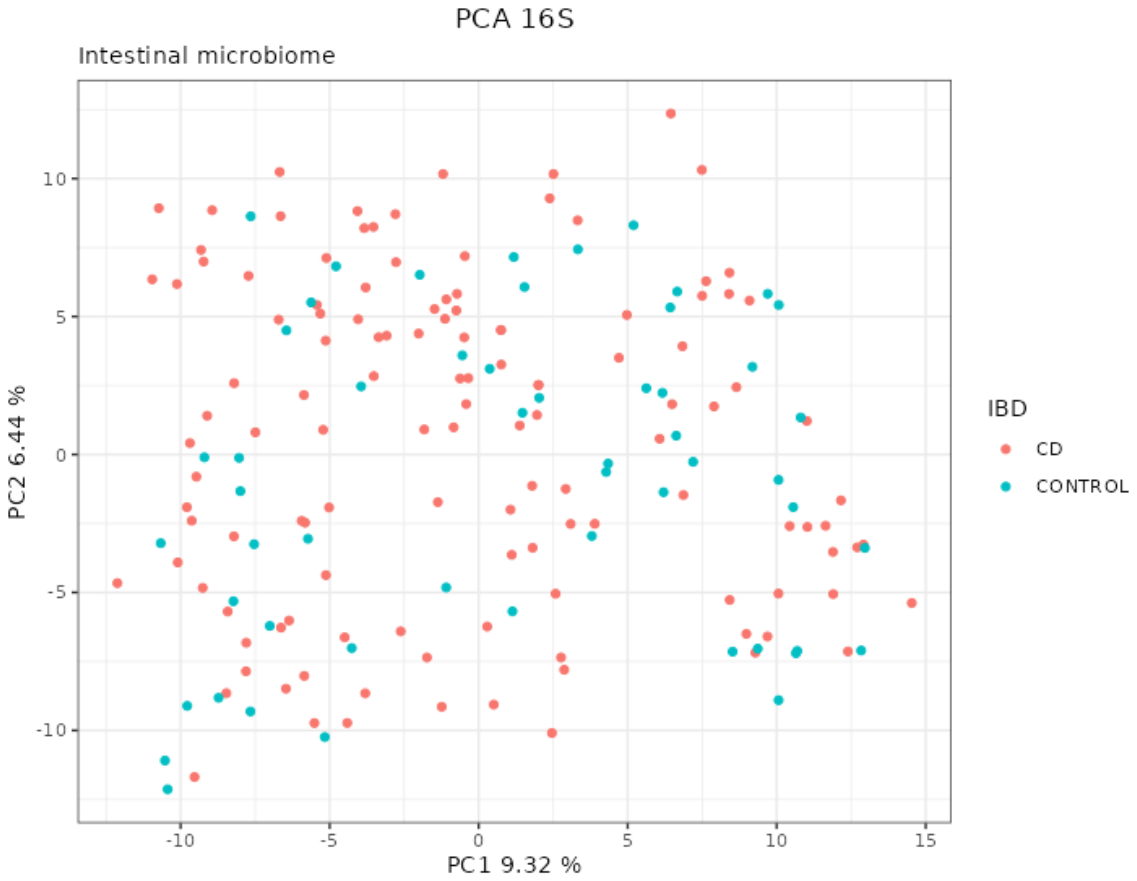

Figure 4.8: PCA of the 16S data of the HSCT dataset. Samples colored by location of the segment. There are no clear patterns according to the location.

If we include the information about the samples all together in a block called metadata we can start from this model on 4.11:

Table 4.11: elationships between the different blocks in the HSCT dataset for model 1.1. 0 indicates no relationship and 1 indicates a strong relationship.

Model 1.1	***Transcriptome***	***Microbiome***	***metadata***
Transcriptome	0	0	1
Microbiome	0	0	1
metadata	1	1	0

When looking for the model that adjust better following this blocks we arrived to model 1.2 thanks to `inteRmodel`, described below on table 4.12:

Table 4.12: Relationships between the different blocks in the HSCT dataset for model 1.2. 0 indicates no relationship and 1 indicates a strong relationship.

Model 1.2	***Transcriptome***	***Microbiome***	***metadata***
Transcriptome	0	0.0	1.0
Microbiome	0	0.0	0.1
metadata	1	0.1	0.0

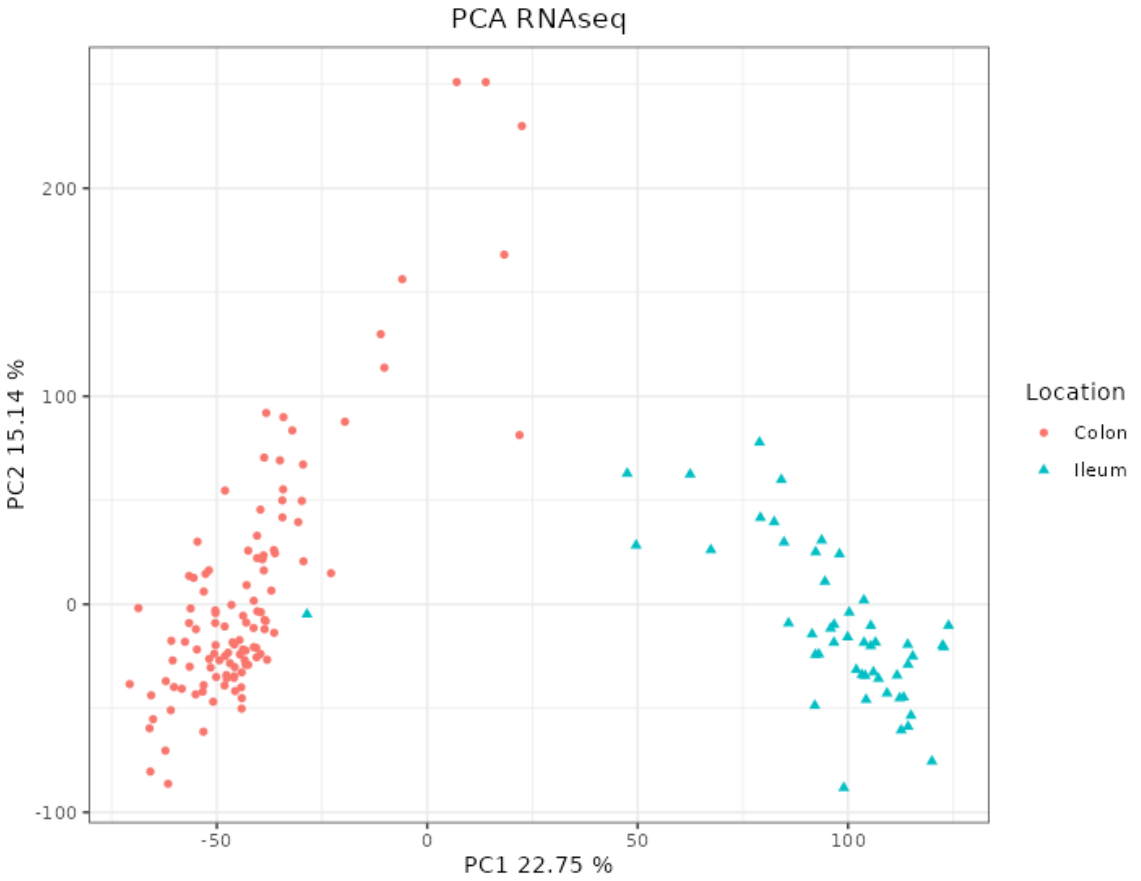

Figure 4.9: PCA of the RNAseq data of the HSCT dataset. Samples colored by location of the segment. The samples separate according to the location.

On model two we split the invariable variables from those related to the location:

Table 4.13: Relationships between the different blocks in the HSCT dataset for model 2. 0 indicates no relationship and 1 indicates a strong relationship.

Model 2	***Transcriptome***	***Microbiome***	***Demographic***	***Location***	***Time***
Transcriptome	0	1	1	1	0
Microbiome	1	0	1	1	0
Demographic	1	1	0	0	1
Location	1	1	0	0	0
Time	0	0	1	0	0

Following this split, we used `inteRmodel` (See section above 4.1.3) to find the model that has higher inner AVE for these blocks is the one on table 4.14:

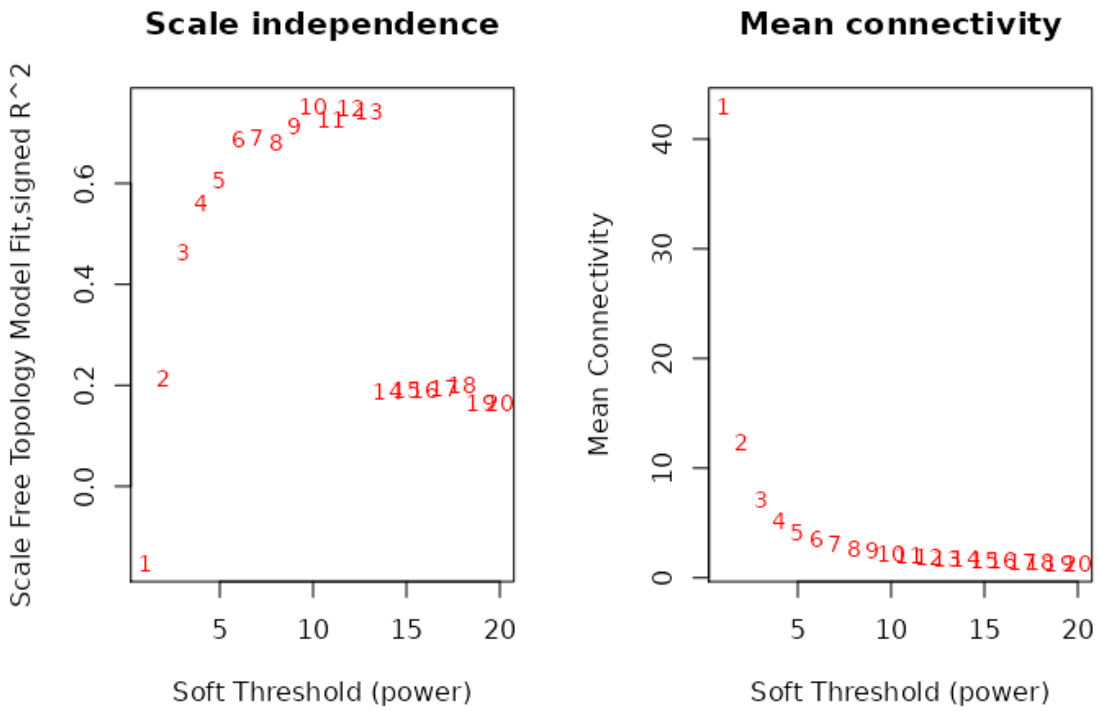

Figure 4.10: Power evaluation of WGCNA of the HSCT dataset. On the ordinate the power on the abscissa on the left the scale free topology model fit; on the right the mean connectivity. There is a low fit even on large power and the mean connectivity is below 100 from the very first value.

Table 4.14: Relationships between the different blocks in the HSCT dataset for model 2.2. 0 indicates no relationship and 1 indicates a strong relationship.

Model 2.2	*Transcriptome*	*Microbiome*	*Demographic*	*Location*	*Time*
Transcriptome	0	0.0	0.0	1.0	0.0
Microbiome	0	0.0	0.2	0.1	0.0
Demographic	0	0.2	0.0	0.0	0.6
Location	1	0.1	0.0	0.0	0.0
Time	0	0.0	0.6	0.0	0.0

We also tested specifically a model from the family 2.3, which can be seen on table 4.15:

Table 4.15: Relationships between the different blocks in the HSCT dataset for model 2.3. 0 indicates no relationship and 1 indicates a strong relationship.

Model 2.3	*Transcriptome*	*Microbiome*	*Demographic*	*Location*	*Time*
Transcriptome	0.0	0.1	0.0	1.0	0
Microbiome	0.1	0.0	0.1	0.1	0
Demographic	0.0	0.1	0.0	0.0	1
Location	1.0	0.1	0.0	0.0	0
Time	0.0	0.0	1.0	0.0	0

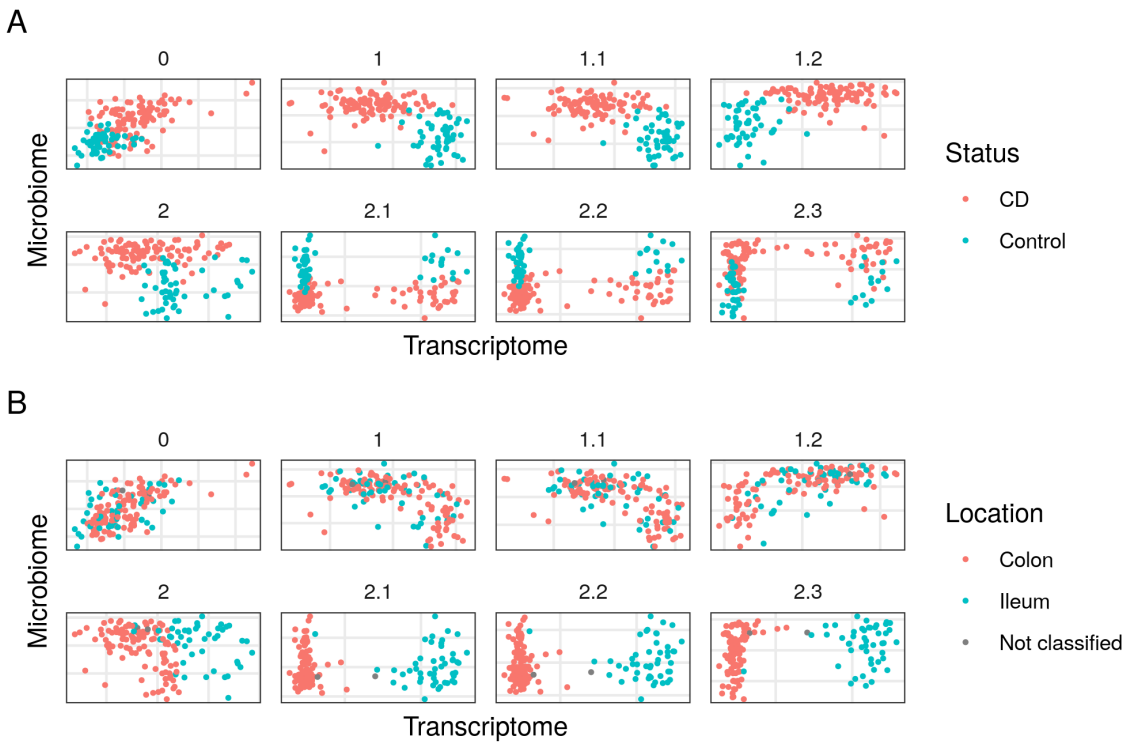

Figure 4.11: Models in the HSCT dataset. On the abscissa the transcriptome, on the ordinate the Microbiome. Each square represents a different model of the HSCT dataset. On panel A colored by disease status, on panel B colored by sample location. Model 0 has only transcriptome and microbiome data, models 1 to 1.2 with data about the samples and models 2.1 to 2.3 with data about the samples split in 3 blocks.

The best model of the family 2 confirmed a relationship between the host transcriptome and the location-related variables, while the microbiome was associated with the demographic and location-related variables (see figure 4.11 and S2 data of [197]). Overall, we see that the relationships in the model affected the distribution of samples on the components of both the host transcriptome and the microbiome.

Table 4.16: The models in the HSCT and their AVE values. For each model the inner AVE and the outer AVE is presented.

Model	***inner AVE***	***outer AVE***
0.0	0.3999234	0.1001689
1.0	0.6230190	0.0842333
1.1	0.5678189	0.0848714
1.2	0.7043881	0.0775766
2.0	0.2517363	0.0982050
2.1	0.6940253	0.0940266
2.2	0.8187640	0.0941628
2.3	0.7761846	0.0943938

The different models selected different variables, some of which are shared between models. The most similar models are those that have split the metadata into 3 blocks,

followed by those that have the metadata in a single block.

In order to analyze the accuracy of the models, one thousand bootstraps were used to integrate the data from the HSCT CD dataset. Each bootstrap had its own dispersion on the variables according to the samples selected, the distribution of the bootstraps used are represented here:

Figure 4.12: Dispersion of the bootstraped samples on age and percentage of colon and controls samples.

Evaluating the same model on each bootstrap lead to a dispersion on the inner AVE of the model. The lower the dispersion, the more robust the model was to different conditions than in the initial testing.

With the bootstrapped models we used BaseSet to estimate the probability that each variable to be relevant for the association with a disease. However, due to big amount of small probabilities when using the BaseSet package to calculate which variables are more relevant it could not provide a good estimation on time.

MCIA was applied as a baseline of the integration, the first two components were represented similarly to those of the blocks when using RGCCA 4.14.

The AUC of classifying the transcriptome in colon or ileum segments was compared between the models (see 4.15 and with MCIA.

These models have the following AUC to classify the location of the sample according to the first component of the transcriptome block.

Figure 4.13: Bootstrap of the models 0, 1.2 and 2.2. The point with the black circle is the AVE of the original data. The dispersion is shown by the ellipses. Model 0 and 1.2 have lower inner and outer AVE score, model 2.2 has lower outer score but higher inner value than the bootstrapped.

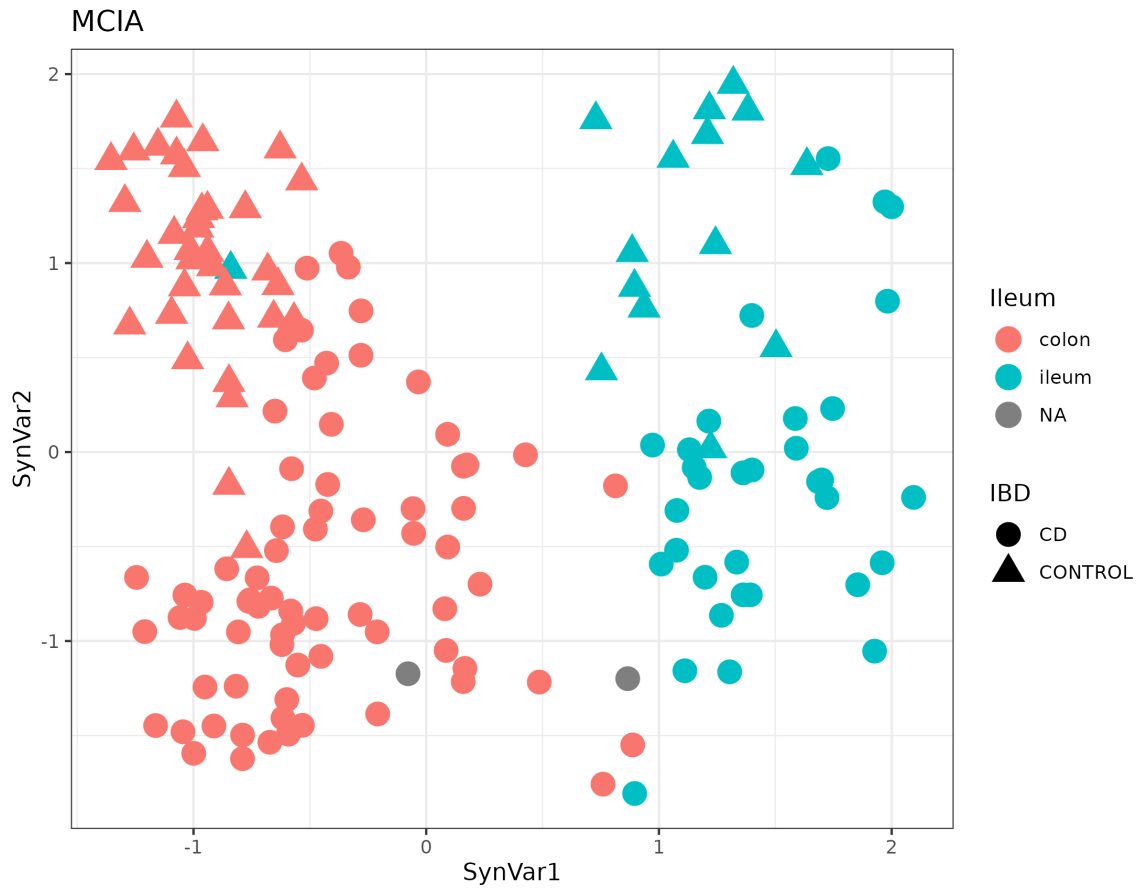

Figure 4.14: MCIA dimensions in the HSCT dataset. MCIA first two synthetic variables. In red circles the colon and in blue triangles the ileum.

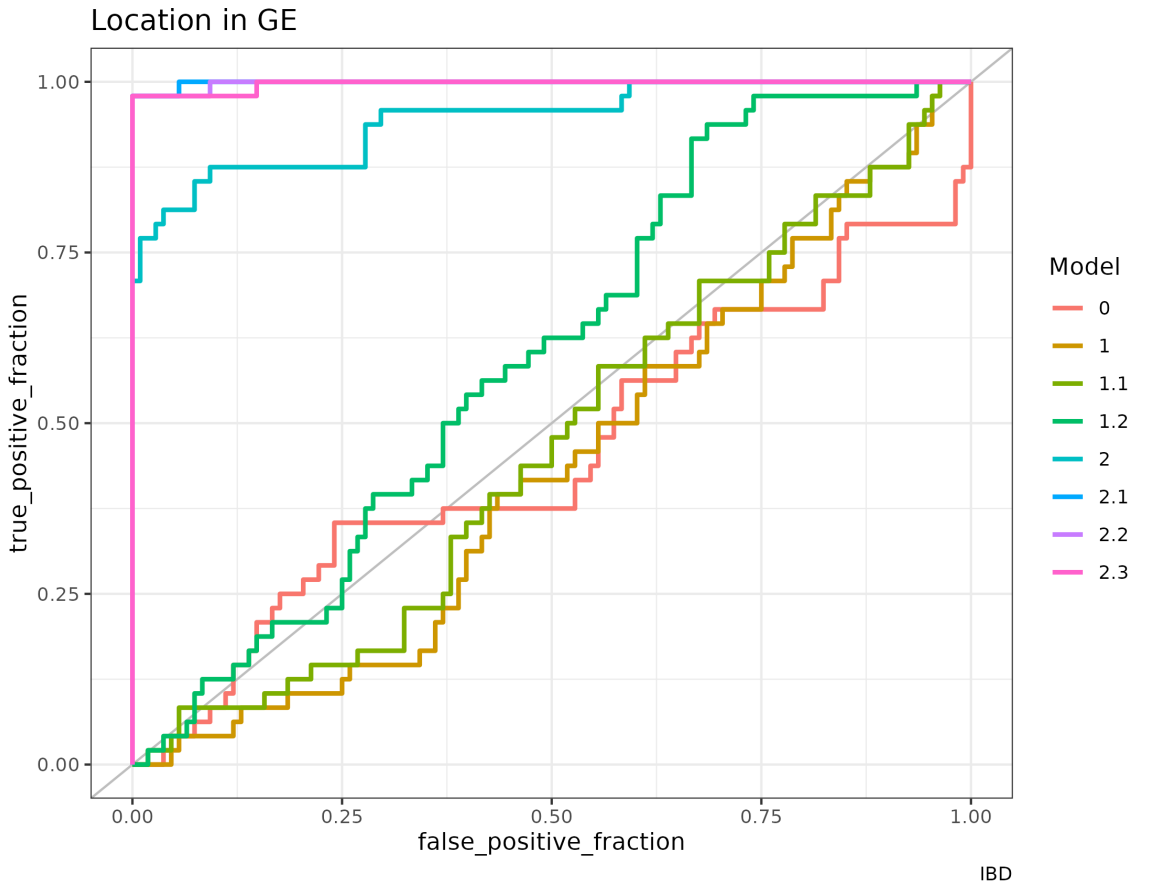

Figure 4.15: AUC of the RGCCA models in the HSCT dataset. The classification of the localization of the sample according to the first component of the gene expression of the models generated with RGCCA on the HSCT dataset.

Table 4.17: AUC values of RGCCA models of the HSCT dataset classyfing the location of the sample according to the first component of the transcriptome block. From model 0 to model 2.3, the best classification is achieved with model 2.1. Note that this is removing two samples for which the location is unkown.

Model	***AUC***
0	0.4537037
1	0.4309414
1.1	0.4639275
1.2	0.5958719
2	0.9450231
2.1	0.9988426
2.2	0.9980710
2.3	0.9969136

On MCIA the AUC for the classification of ileum or colon samples is of 0.9851 once those two samples with unknown location are excluded. This is on par with the models of family 2 as can be seen on the table 4.17.

The different models selected different variables as can be seen below:

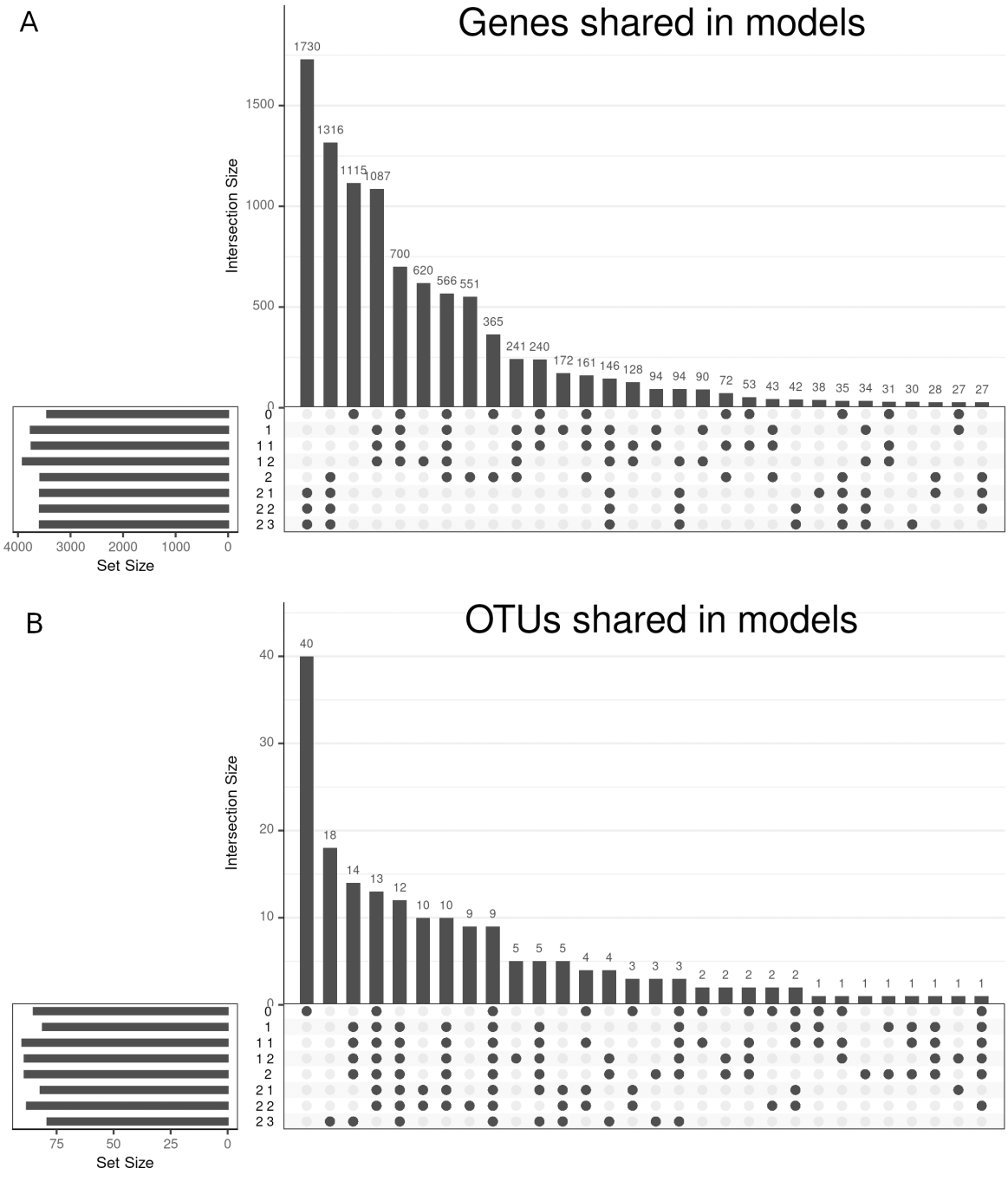

Figure 4.16: Upset plot of variables selected in the HSCT dataset. The variables selected on each model from 0 to 2.3 showing the intersection between them regarding genes, panel A, and OTUs, panel B.

Differences and similarities between the selected features of each model can be observed on Figure 4.16. Genes are very similar between model 0 to 1.2 and between 2 to 2.3, while OTUs are very unique on model 0 and others shared between most models.

4.2.3 Häsler's dataset

If we look at the dataset 16S, the PCA doe not show a pattern as can be seen on figure 4.17:

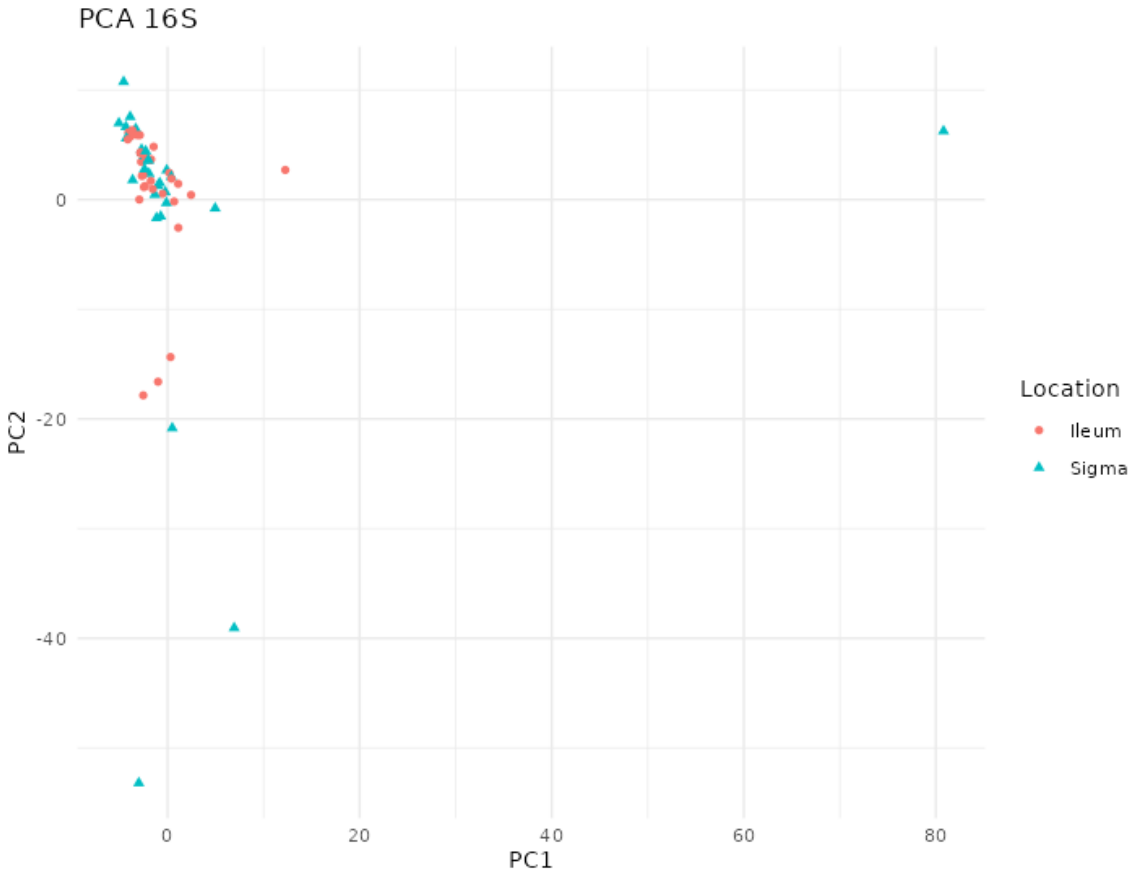

Figure 4.17: PCA of 16S of Häsler's dataset colored by the location of the samples.

The PCA on the transcriptome shows a distinction between colon and ileum according to the first dimension of the pca that can be seen on 4.18:

In this dataset, the parameter tau behaved slightly differently than with the previous dataset but the value from the Schäfer's method for tau was close to the best value.

Models for Häsler dataset are:

Table 4.18: Model 0 of Häsler dataset. 0 indicates no relationship and 1 indicates a strong relationship.

Model 0	***RNAseq***	***micro***
RNAseq	0	1
micro	1	0

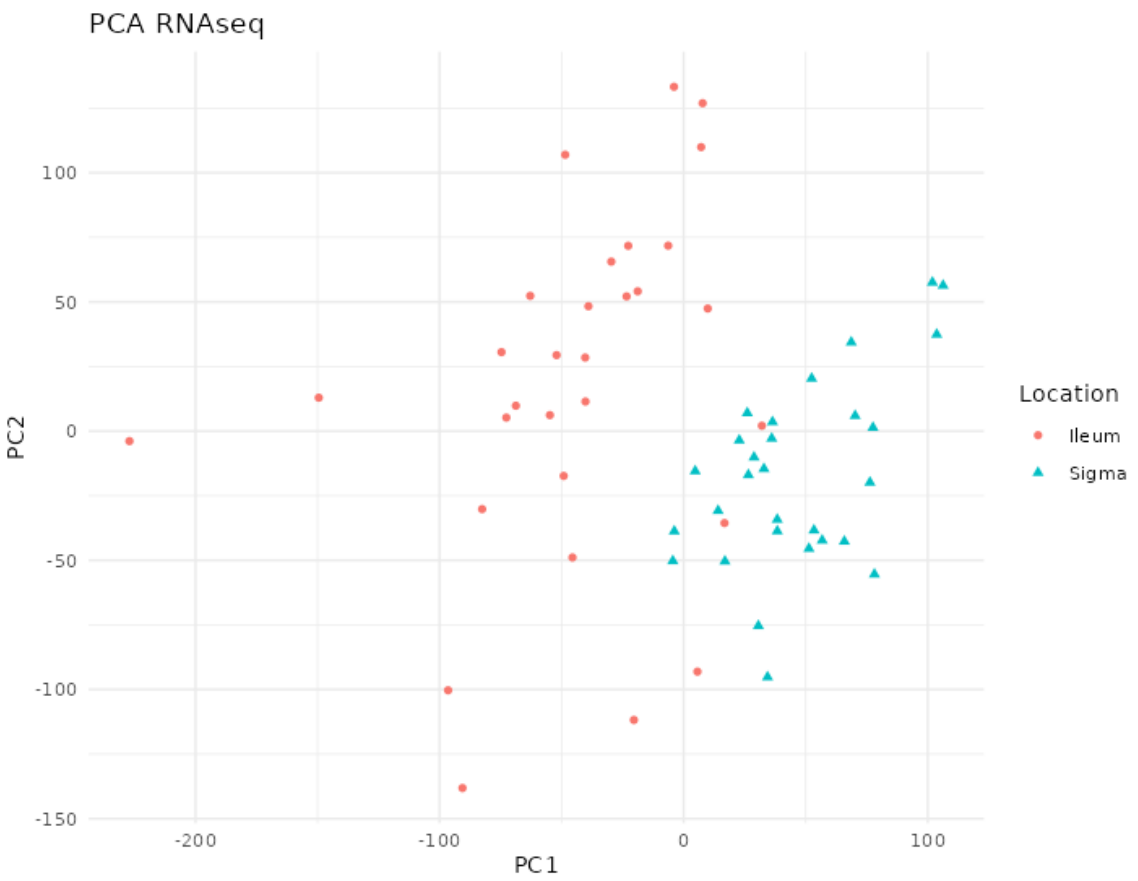

Figure 4.18: PCA of RNAseq of the Häsler's dataset. The samples are colored by location sigma and ileum. There are two clear groups according to the location.

The first model for family 1 is on table 4.19:

Table 4.19: Model 1.1 of the Häsler dataset. 0 indicates no relationship and 1 indicates a strong relationship.

Model 1.1	***RNAseq***	***micro***	***meta***
RNAseq	0	0	1
micro	0	0	1
meta	1	1	0

The after optimization of the model of family 1, the best one is on table 4.20:

Table 4.20: Model 1.2 of the Häsler dataset. 0 indicates no relationship and 1 indicates a strong relationship.

Model 1.2	***RNAseq***	***micro***	***meta***
RNAseq	0	0.0	1.0
micro	0	0.0	0.1
meta	1	0.1	0.0

The first model for family 2 is on table 4.21:

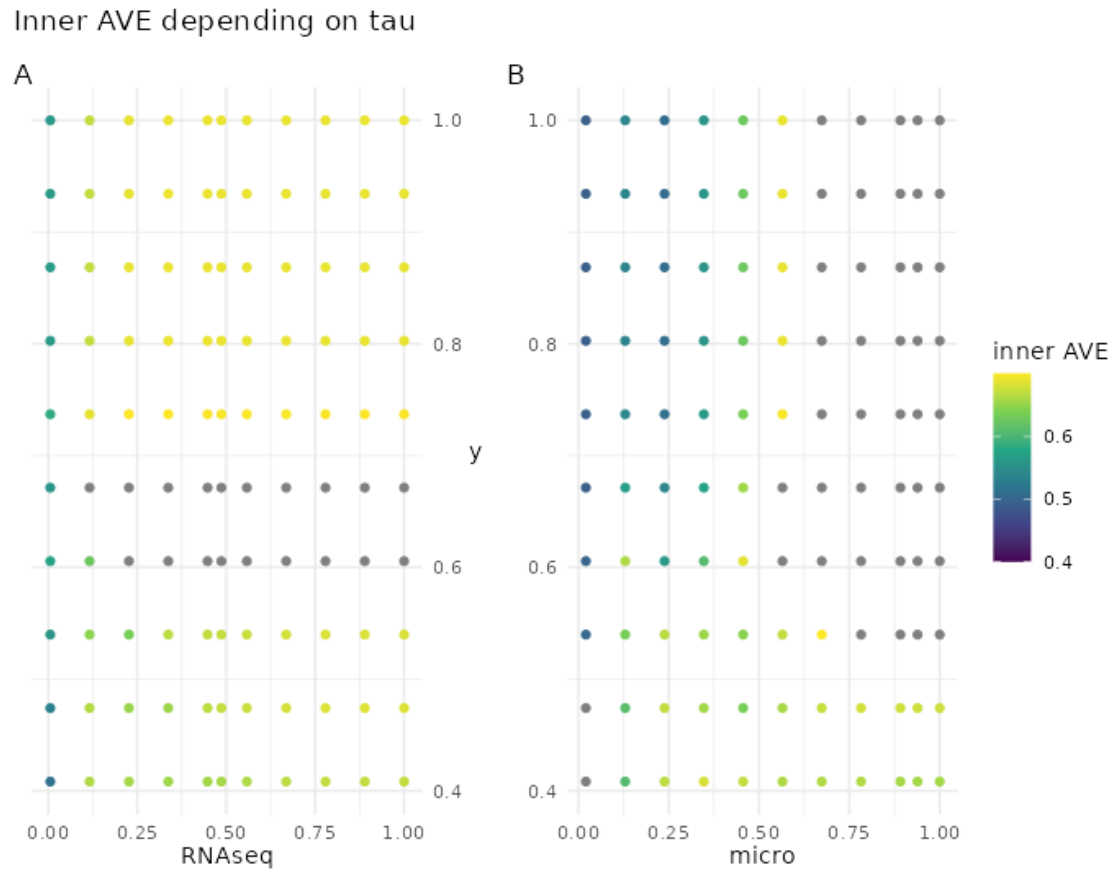

Figure 4.19: Changes on tau on the centroid scheme in the Häsler dataset affect the inner AVE score on the model 1. The panel A shows on the ordinate the RNAseq tau value, the panel B on the right, shows the tau of the microorganism; both of them show the y's tau on the abscissa.

Table 4.21: Model 2.1 of the Häsler dataset. 0 indicates no relationship and 1 indicates a strong relationship.

Model 2.1	***RNAseq***	***micro***	***Location***	***Demographic***	***Time***
RNAseq	0	0.0	1.0	0	0
micro	0	0.0	0.5	1	0
Location	1	0.5	0.0	0	0
Demographic	0	1.0	0.0	0	1
Time	0	0.0	0.0	1	0

After optimization of models of family 1 the best model according to the inner AVE score is on table 4.22:

Table 4.22: Model 2.2 of the Häsler dataset. 0 indicates no relationship and 1 indicates a strong relationship.

Model 2.2	*RNAseq*	*micro*	*Location*	*Demographic*	*Time*
RNAseq	0.0	0.1	1	0.0	0.0
micro	0.1	0.0	0	0.1	1.0
Location	1.0	0.0	0	0.0	0.0
Demographic	0.0	0.1	0	0.0	0.1
Time	0.0	1.0	0	0.1	0.0

On table 4.23, we can see here the AVE scores of each of the previous models:

Table 4.23: AVE values of RGCCA models in Häsler's dataset. The inner and the outer AVE scores of multiple models tested on the Häsler dataset are shown. The model with the hightest inner AVE is model 1.2.

Model	*inner AVE*	*outer AVE*
0	0.8217371	0.0961236
1.1	0.7461423	0.1024148
1.2	0.8349410	0.1025486
2.1	0.4980681	0.1008395
2.2	0.7513065	0.1009915

In contrast to the HSCT's dataset (table 4.16), the model with the highest inner AVE was model 1.2 but model 2.2 was close to it (see table 4.23). Model 2.2 has a relationship of 0.1 between microbiome and the host transcriptome and of 1 between the location and the host transcriptome. The microbiome block is also related by a factor of 0.1 with the demographic block and of 1 with the time block. Lastly, the time and the demographic block are related by a factor of a 0.1. In either case the family 1 and family 2 models can correctly separate by sample location (colon or ileum) but not by disease type or inflammation status as can be seen on figure 4.20.

There is no observable cluster of IBD samples and the other samples, showing that on this dataset the differences of the microbiome between the different type of samples are less stark. The classification of samples was very accurate in all the models, specially on model 2.2, see figure 4.21:

This accuracy resulted on high AUC values for all the models, as can be seen on table 4.24:

Table 4.24: AUC of the RGCCA models in Häsler's dataset. The AUC was calculated with the first dimension of the gene expression block ability to predict location of the sample.

Model	*AUC*
0	0.8011494
1.1	0.9781609
1.2	0.9977011
2.1	1.0000000
2.2	1.0000000

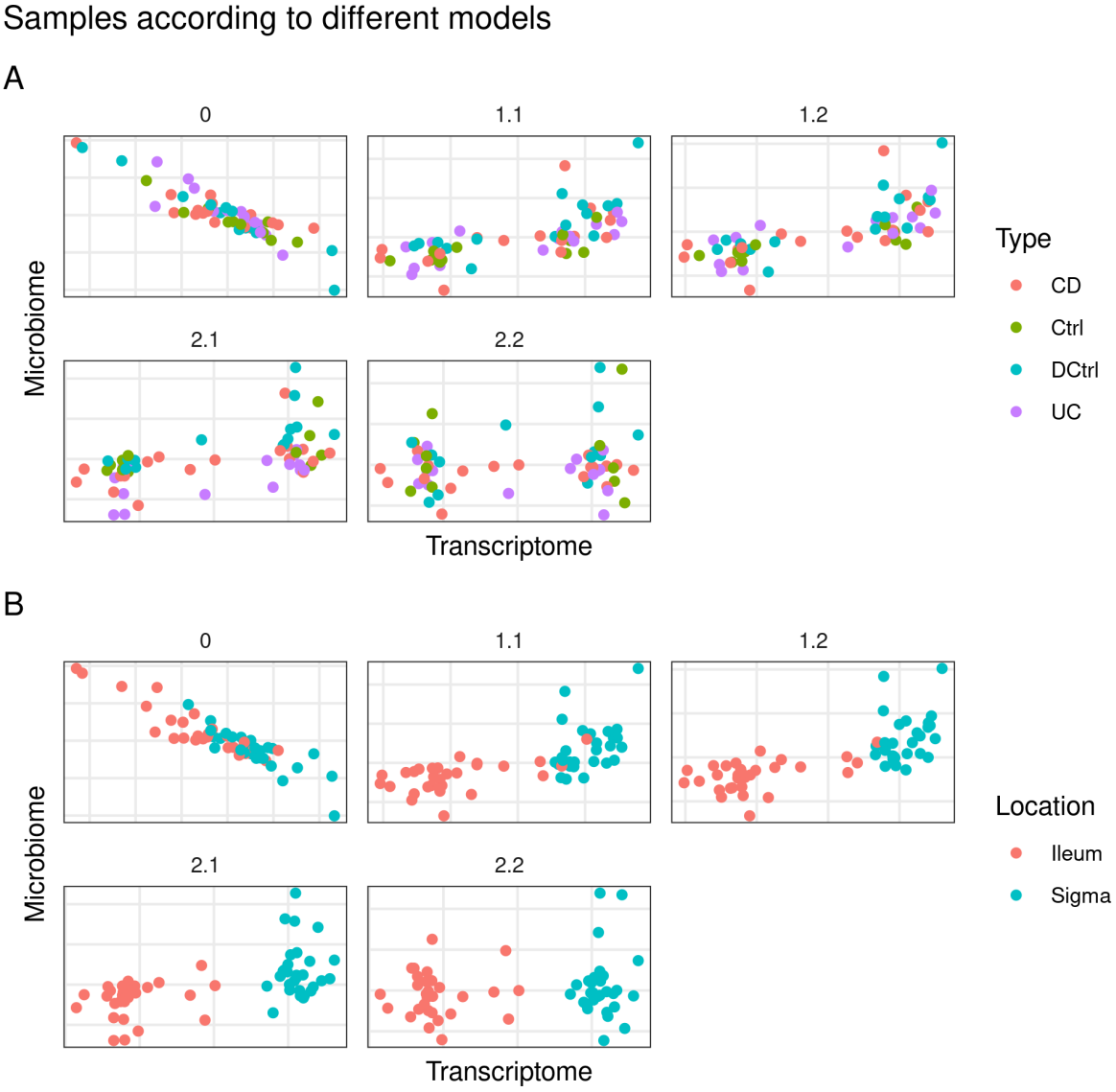

Figure 4.20: Models from inteRmodel in the Häsler's dataset. Model 0 with just the transcriptome and microbiome data. Models 1.1 to 1.2 with transcriptome, microbiome and sample data in a single block. Models 2.1 and 2.2 with transcriptome, microbiome and sample data in multiple blocks. On the A panel colored by disease on the B panel colored by location of the sample.

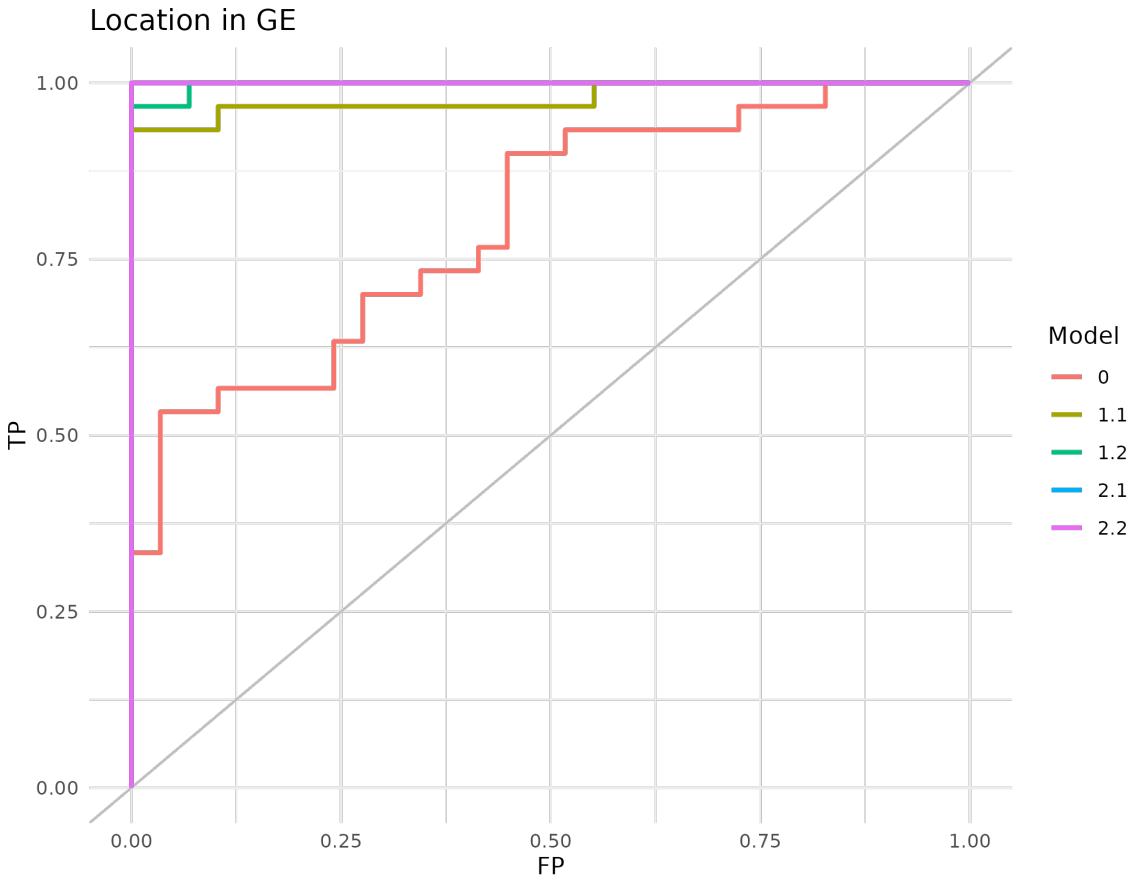

Figure 4.21: AUC of models with RGCCA in the Häsler's dataset. The classification of the localization of the sample according to the first component of the gene expression of the models generated with RGCCA on the Häsler's dataset.

MCIA was applied as a baseline of the integration and compared to the different models to know which one separates best colon and ileum samples. The result on the first two dimension is shown on figure 4.22:

MCIA's AUC results was as high as the model 2.2 to classify samples according to their location. It was even better to classify the samples according to the type of sample they are: 0.6248 vs 1 the best AUC from RGCCA that corresponds to model 1.2.

4.2.4 Morgan's dataset

To explore this dataset that is different from the others with IBD.

The PCA didn't show any pattern on the microbiome according to the location, see figure 4.23.

The PCA on the transcriptome did not show a clear distinction between the prepouch ileum and the pouch on figure 4.24 but there is a pattern:

We tested if results of inteRmodel were consistent on this dataset with the other datasets. The first model we tried is model 0 as in table 4.25:

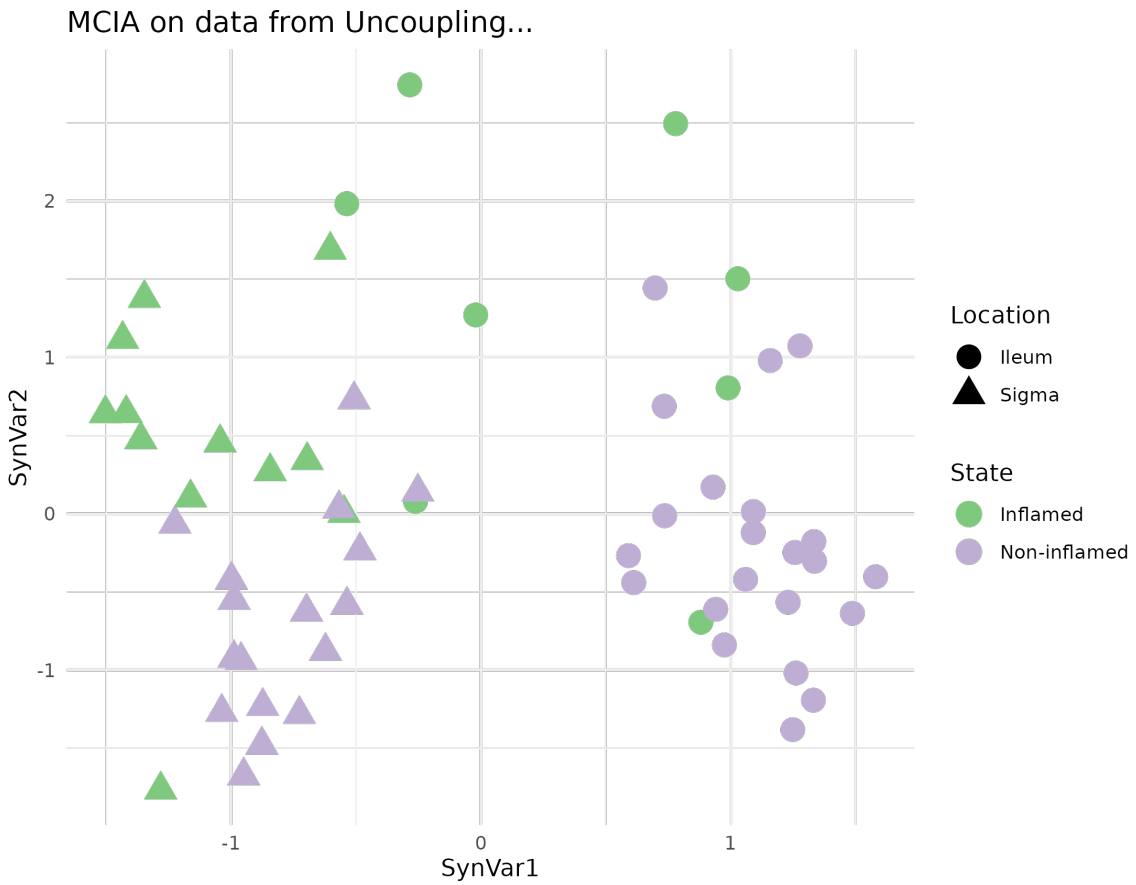

Figure 4.22: MCIA dimensions in the Häsler's dataset. MCIA first two dimensions of the dataset colored by state, the shape is according to the location of the samples. Shows two vertical groups on the first syntethic dimension according to the location of the samples.

Table 4.25: Relationships between the different blocks in the Morgan's dataset for model 0. 0 indicates no relationship and 1 indicates a strong relationship.

Model 0	***Transcriptome***	***Microbiome***
Transcriptome	0	1
Microbiome	1	0

We then added the data about the samples as provided 3.1.4, on a simple model as in table 4.26:

Table 4.26: Relationships between the different blocks in the Morgan's dataset for model 1. 0 indicates no relationship and 1 indicates a strong relationship.

Model 1	***Transcriptome***	***Microbiome***	***metadata***
Transcriptome	0	0	1
Microbiome	0	0	1
metadata	1	1	0

When looking for the model that adjust better following this structure we arrived to

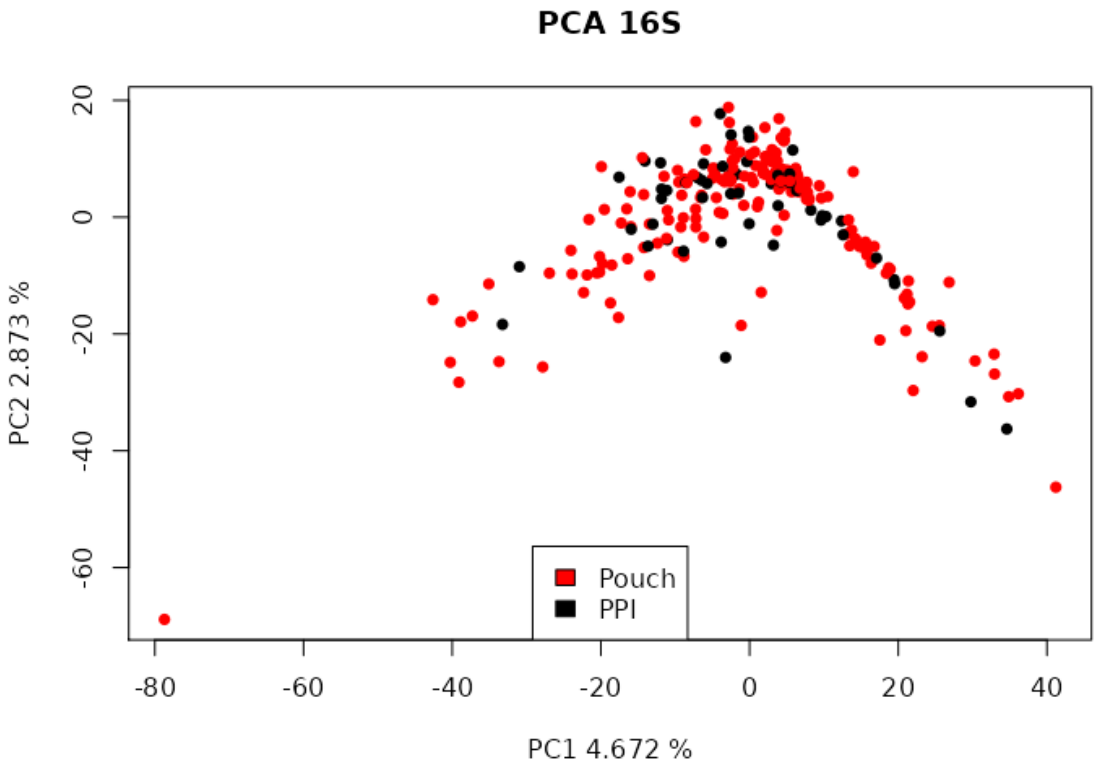

Figure 4.23: PCA of 16S of the Morgan's dataset. The sample are colored by location.

model 1.2, described below:

Table 4.27: Relationships between the different blocks in the Morgan's dataset for model 1.2. 0 indicates no relationship and 1 indicates a strong relationship.

Model 1.2	***Transcriptome***	***Microbiome***	***metadata***
Transcriptome	0.0	0.1	0
Microbiome	0.1	0.0	1
metadata	0.0	1.0	0

On model two we split the invariable variables from those related to the location (see 4.28):

Table 4.28: Relationships between the different blocks in the Morgan's dataset for model 2. 0 indicates no relationship and 1 indicates a strong relationship.

Model 2	***Transcriptome***	***Microbiome***	***Demographic***	***Location***
Transcriptome	0	1	1	1
Microbiome	1	0	1	1
Demographic	1	1	0	0
Location	1	1	0	0

The model that has higher inner AVE for these blocks is the following:

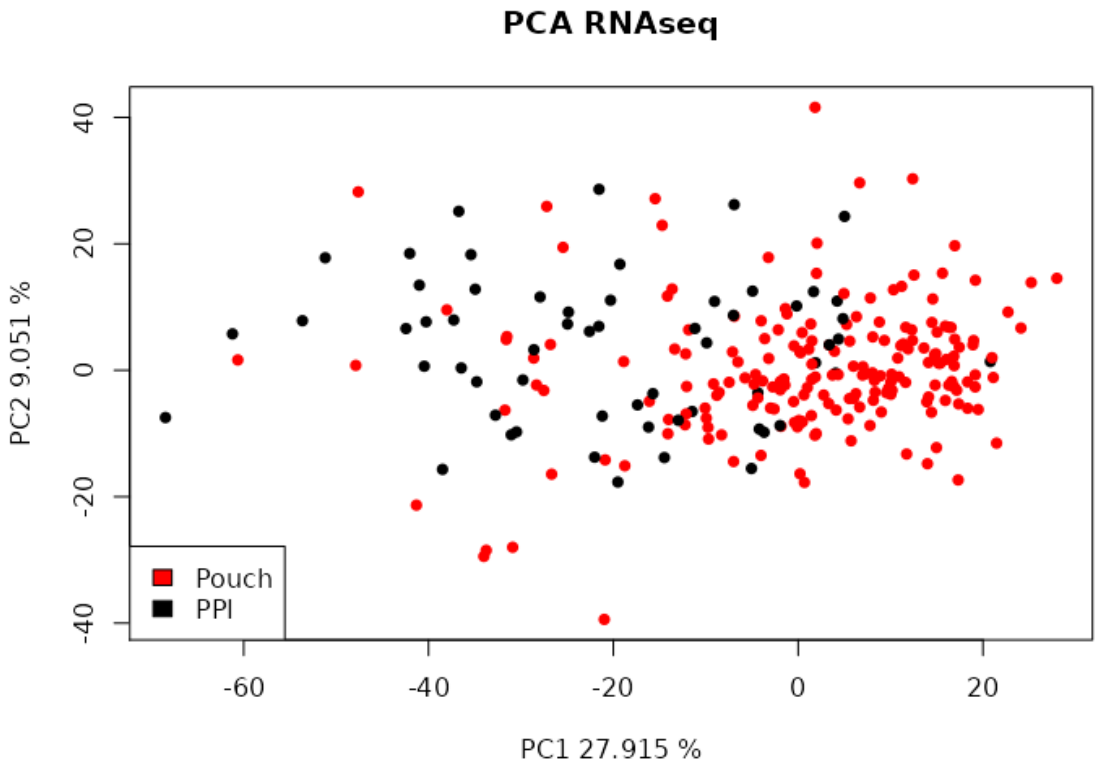

Figure 4.24: PCA of RNAseq of the Morgan's dataset. The sample are colored by location.

Table 4.29: Relationships between the different blocks in the Morgan's dataset for model 2.2. 0 indicates no relationship and 1 indicates a strong relationship.

Model 2.2	*Transcriptome*	*Microbiome*	*Demographic*	*Location*
Transcriptome	0.0	0.1	1	0.1
Microbiome	0.1	0.0	0	0.0
Demographic	1.0	0.0	0	0.0
Location	0.1	0.0	0	0.0

Each model is different from previous models. After model 2.2 we looked on the model similar to model 2.3 in the HSCT dataset showed but it is the same as in model 2.2. However, it is kept on the further analysis.

The different models were not able to separate the samples neither by location or sex.

Nevertheless, we compared the classification with the MCIA algorithm and still resulted that model 2.2 provide a better classification than MCIA.

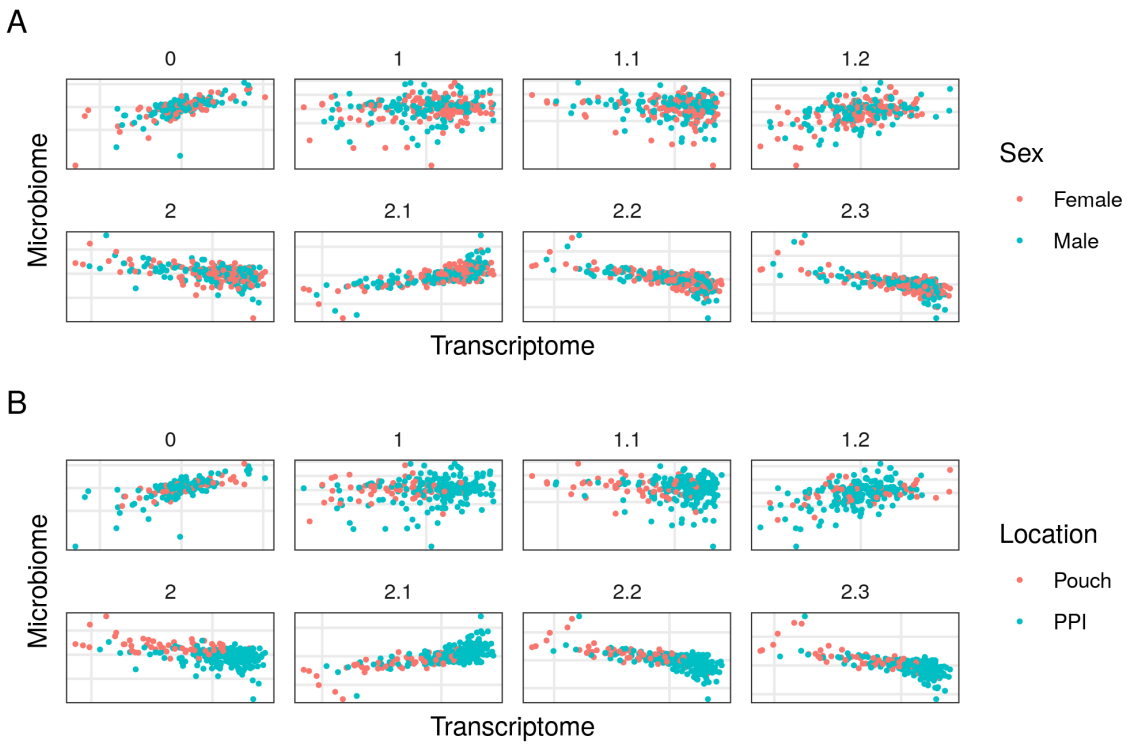

Figure 4.25: Models from inteRmodel in the Morgan's dataset. First component of the transcriptome and microbiome of models on the Morgan's dataset. Model 0 without sample data. Model 1 to 1.2 with all the sample data in a single block and models 2.1 to 2.3 with sample data in several blocks. Panel A shows samples colored by sex and panel B by segment of the sample. There is no clear classification neither by location nor sex on any of the models.

Table 4.30: AVE values of RGCCA for the Morgan's dataset. The inner and the outer AVE scores of multiple models tested on the Morgan dataset are shown. The model with the hightest inner AVE is model 1.2.

Model	***inner AVE***	***outer AVE***
0.0	0.4735601	0.1098639
1.0	0.6333592	0.1152280
1.1	0.2448234	0.1104746
1.2	0.7868443	0.0422660
2.0	0.4404123	0.1088730
2.1	0.6052598	0.1074900
2.2	0.6895661	0.1081315
2.3	0.6895661	0.1081315

When exploring the bootstraps of the data we found that model 1.2 is highly variable:

In addition the model 2.2 usually has a lower inner AVE compared to model 1.2.

The area under the curve for these models is:

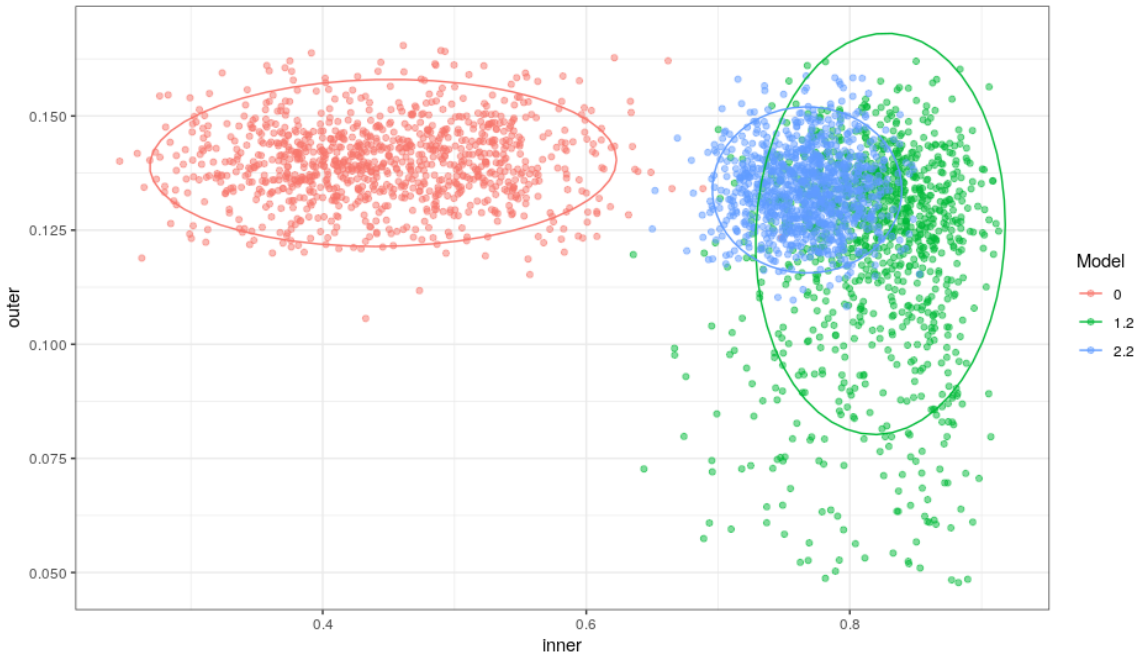

Figure 4.26: Inner and outer AVE scores of the bootstrapped models 0 1.2 and 2.2. Model 0 does not have sample data. Model 1.2 has microbiome, transcriptonme and sample data in a single block and model 2.2 has microbiome, transcriptome and the sample data split in several blocks.

Table 4.31: AUC for the Morgan's dataset classificating the localization of the sample according to the first component of the gene expression of the models generated with RGCCA.

Model	***AUC***
0	0.4969734
1	0.7934971
1.1	0.8161536
1.2	0.5606192
2	0.8546351
2.1	0.8473712
2.2	0.8352646
2.3	0.8352646

With MCIA:

If we quantify this separation by the first dimension of MCIA, the AUC is 0.818, which is slightly worst than the models of family 2.

4.2.5 Howell's dataset

The 16S of this dataset doesn't show any clear pattern regarding the location of the samples according to the firsts dimensions of the PCA on 4.29:

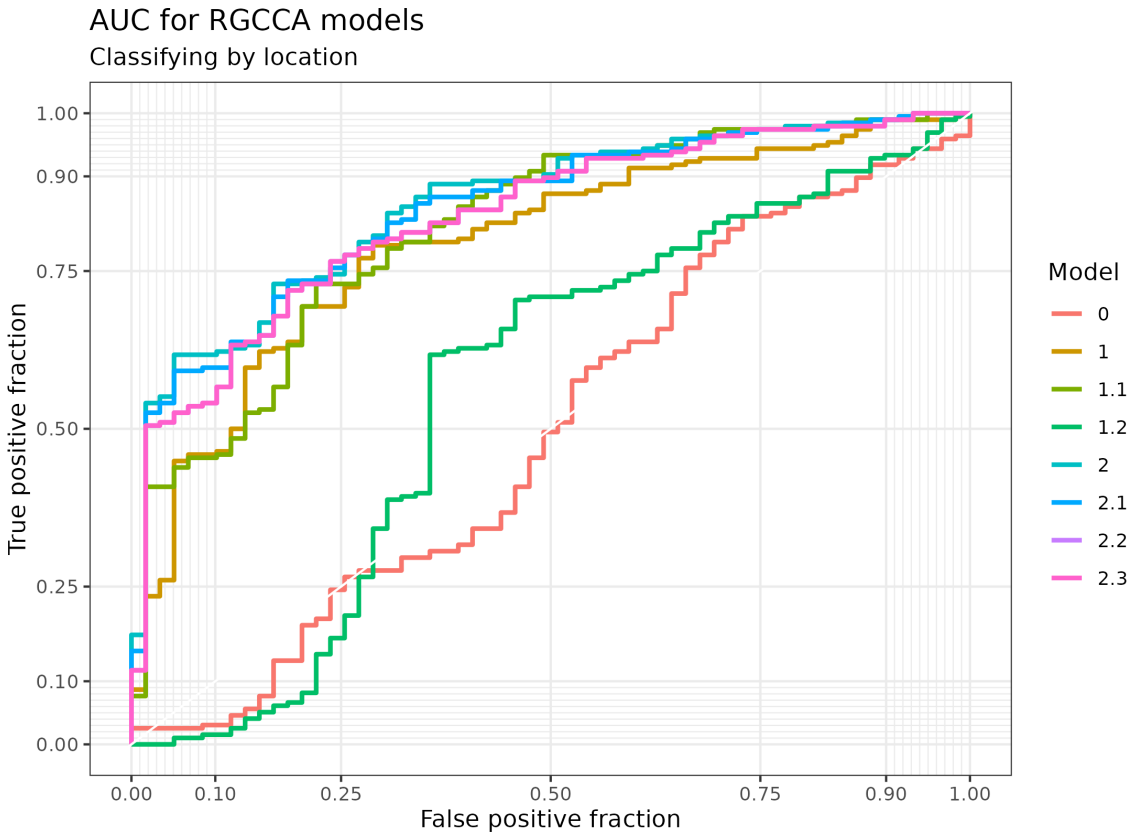

Figure 4.27: AUC of RGCCA models in the Morgan's dataset. The classification of the localization of the sample according to the first component of the gene expression of the models generated with RGCCA on the Morgan's dataset.

The PCA on the transcriptome shows a distinction between colon and ileum 4.30 there are almost two distinct groups according to location:

This dataset was processed to confirm the results on the previous datasets. As always first we started with model 0, connecting both the RNAseq and the 16S blocks:

Later we look for the best model of family 1 (without looking at any previous model of family 1). This resulted on the following model 4.32:

Table 4.32: Relationships between the different blocks in the Howell's dataset for model 1.2. 0 indicates no relationship and 1 indicates a strong relationship.

Model 1.2	***RNAseq***	***16S***	***metadata***
RNAseq	0.0	0.1	1
16S	0.1	0.0	0
metadata	1.0	0.0	0

Model 1.2 4.32 was the best according to the AVE score but perform worse when attempting to recreate known biological differences via classifying samples as we can see below 4.31:

Model 2.2 was selected for further analysis as it describes more accurately the biology of the dataset it. Model 2.2 can be seen on 4.33.

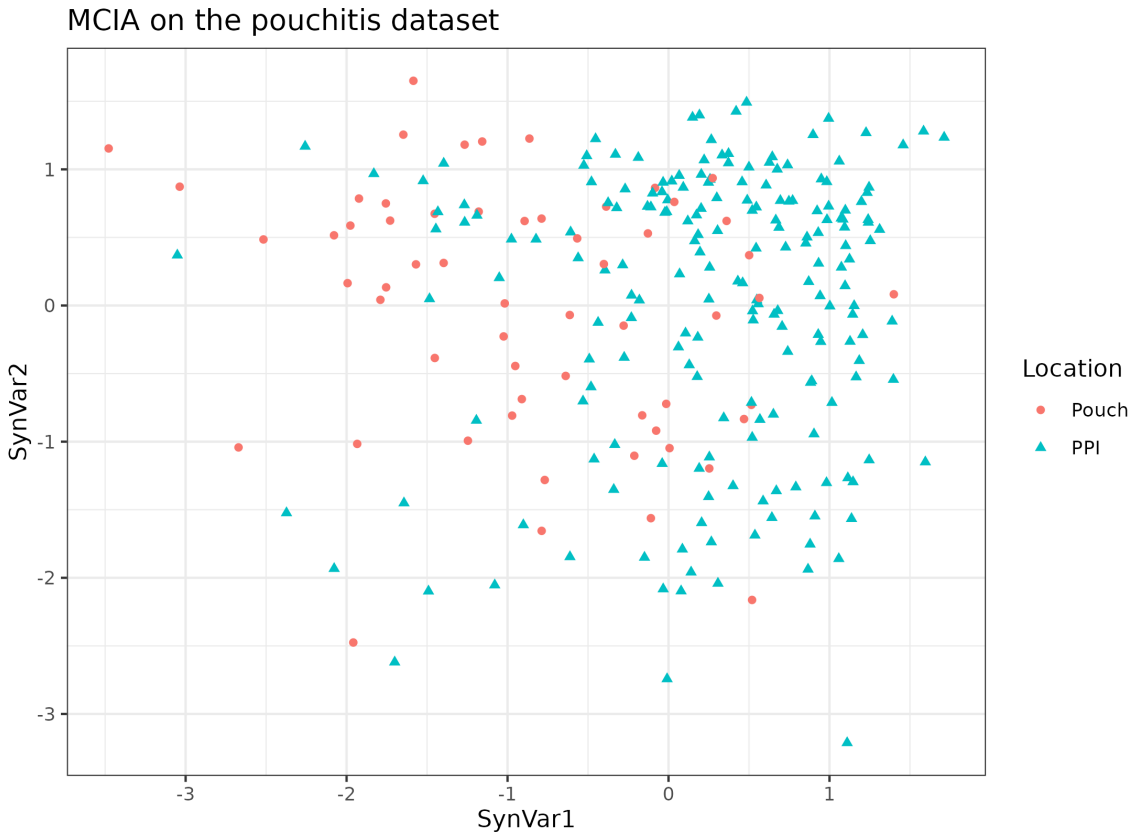

Figure 4.28: MCIA dimensions in the Morgan's dataset. MCIA first two dimensions of the dataset. The color and shape match the location of the sample. There is some separation between samples by location.

Table 4.33: Relationships between the different blocks in the Howell's dataset for model 2.2. 0 indicates no relationship and 1 indicates a strong relationship.

Model 2.2	***RNAseq***	***16S***	***demographics***	***location***
RNAseq	0	0	0.0	1.0
16S	0	0	1.0	0.0
demographics	0	1	0.0	0.1
location	1	0	0.1	0.0

Model 1.2 has a 0.1 relationship between the ASV and the transcriptome and 1 between transcriptome and metadata. While model 2.2 has a relationship of 1 between location and transcriptome and demographics and ASV but only of 0.1 between demographics and location.

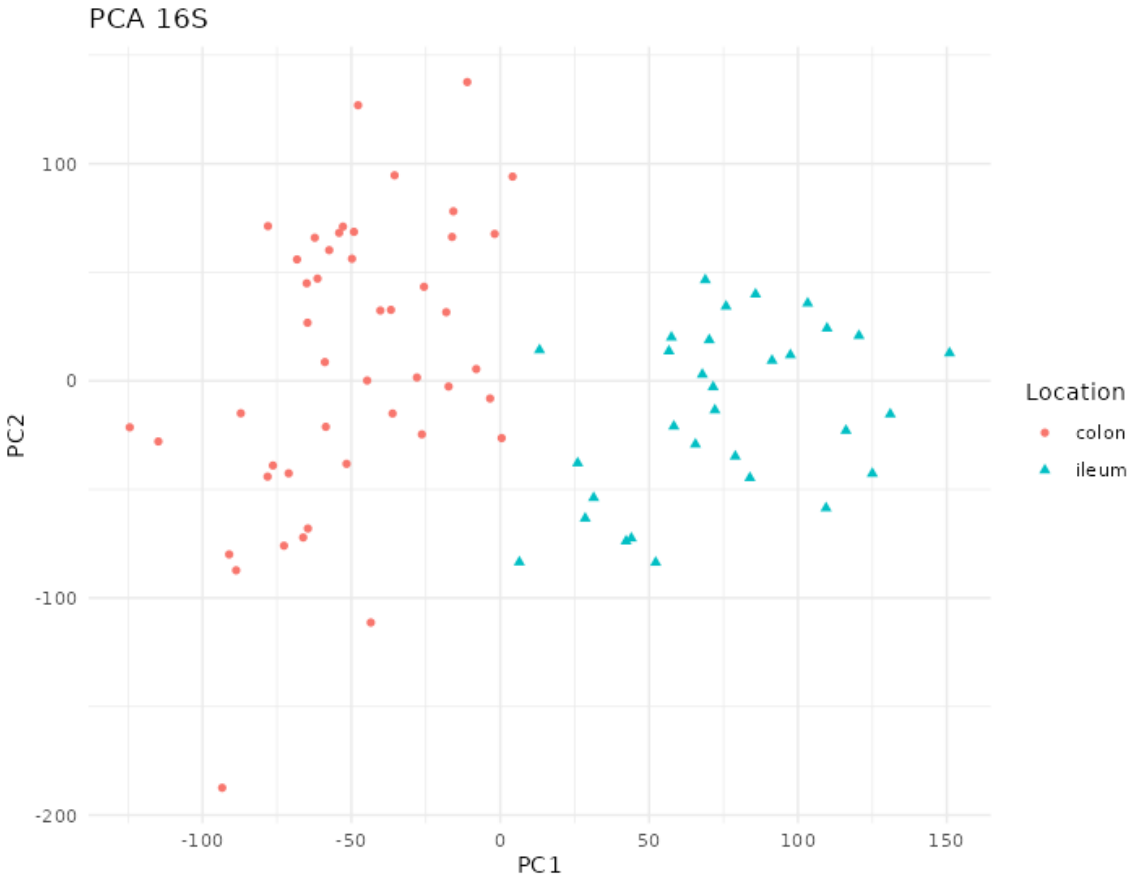

Figure 4.29: PCA of 16S data of the Howell's dataset. Samples are colored by location. There are no pattern on any of the first do dimensions.

Table 4.34: AVE values of RGCCA for the Howell's dataset. The inner and the outer AVE scores of multiple models tested in the Howell dataset are shown. The model with the hightest inner AVE is model 1.2.

Model	***inner AVE***	***outer AVE***
0.0	0.7180980	0.1112390
1.2	0.8972258	0.1660267
2.2	0.8433274	0.1659844

The bootstrapping showed that model 1.2 has indeed higher inner AVE values than model 2.2 and is more stable than model 1.2. While model 0 shows a high variation according to which samples are selected.

If we look at the classification of the models on figure 4.34, we can see that models 1.2, 2 and 2.2 classify perfectly the samples by the transcriptome into the location of the sample.

The AUC of each model can be seen on 4.35:

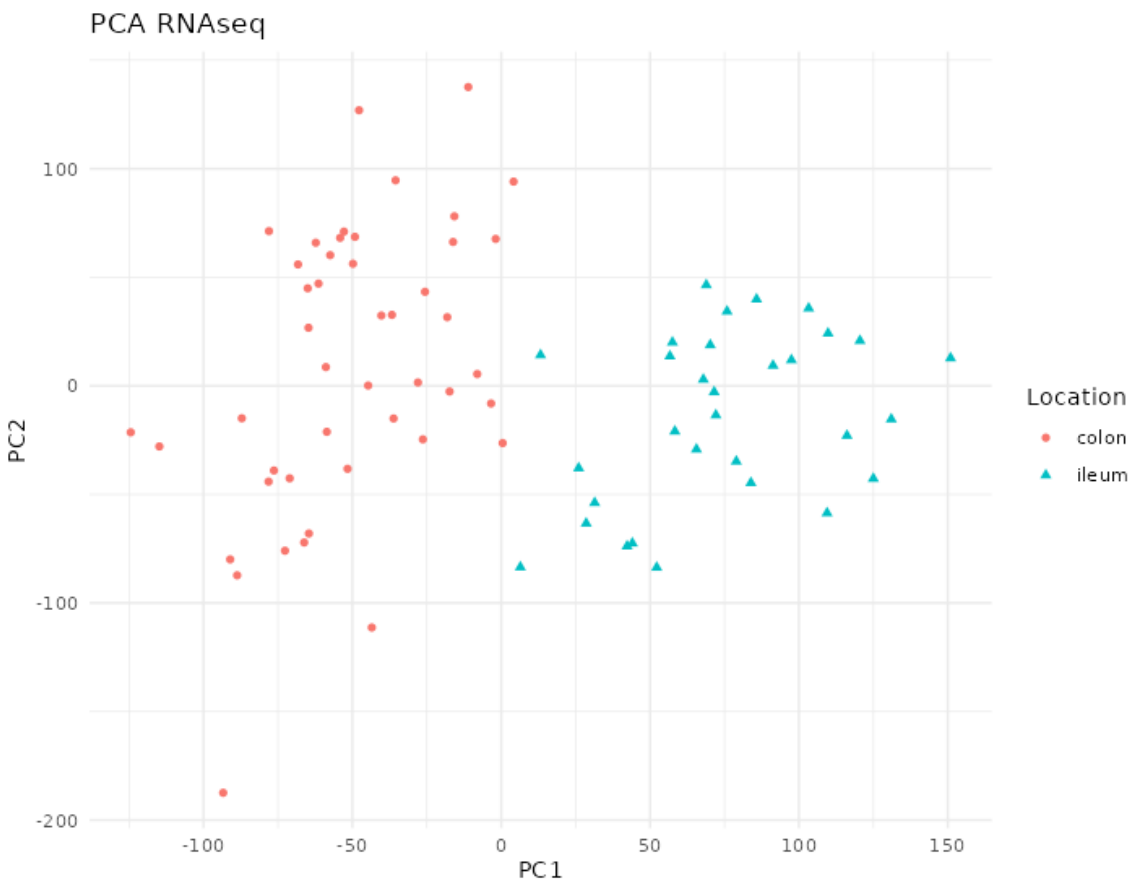

Figure 4.30: PCA of RNAseq data of the Howell's dataset. Samples are colored by location. There are two groups of samples according to their location.

Table 4.35: AUC of the RGCCA models in the Howell's dataset. The classification of the location of the sample according to the first componenent of the models shown. Model 0, 2 and 2.2 have a perfect classification of the samples to their respective location.

model	***AUC***
0	0.6255259
1	0.5974755
1.2	1.0000000
2	1.0000000
2.2	1.0000000

In this dataset we also focused on the most important ASV according to the model 2.2 that were present in more than 2 samples that in total were present in the whole dataset. These ASV were summarized to a single value and then used to calculate the AUC, which was 0.85. The dot product of the ASV and genes were also calculated and used to find out which ASV are related to which genes.

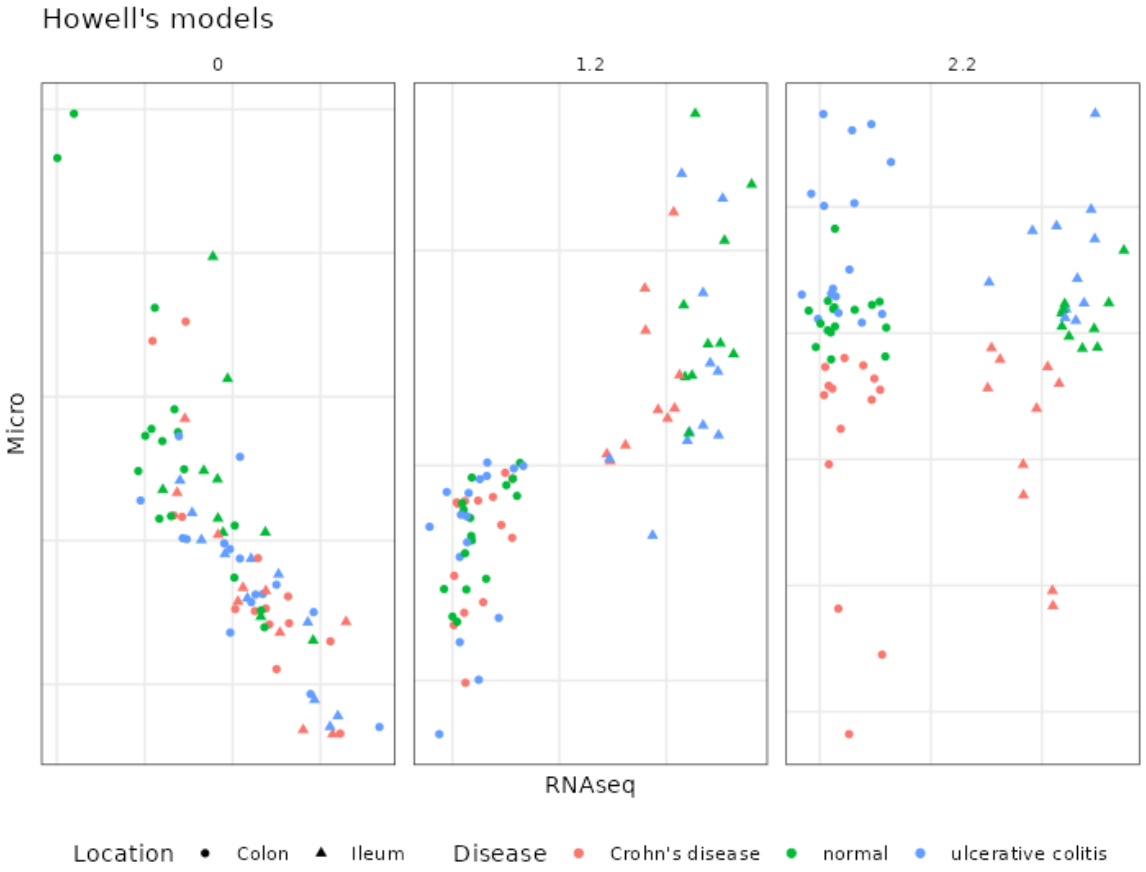

Figure 4.31: Models from inteRmodel in the Howell's datset. The three main models, model 0, 1.2 and 2.2 on the Howell's dataset colored by section colon, ileum and shape according to the disease: square, ulcerative colitis; triangle, normal; circle, Crohn's disease. Model 0 has just trancriptomic and microbiome data, model 1.2 has transcriptomic, microbiome and sample data and model 2.2 has transcriptomic, microbiome and sample data split in different blocks.

4.2.6 Between datasets

The HSCT genes were compared to the Howell's genes from model 2.2. There are 3580 selected on model 2.2 in the HSCT dataset and 2189 genes on the Howell's dataset. From them the 1228 genes in common were analyzed for which GO terms and pathways they are enriched. The results is represented on 4.35.

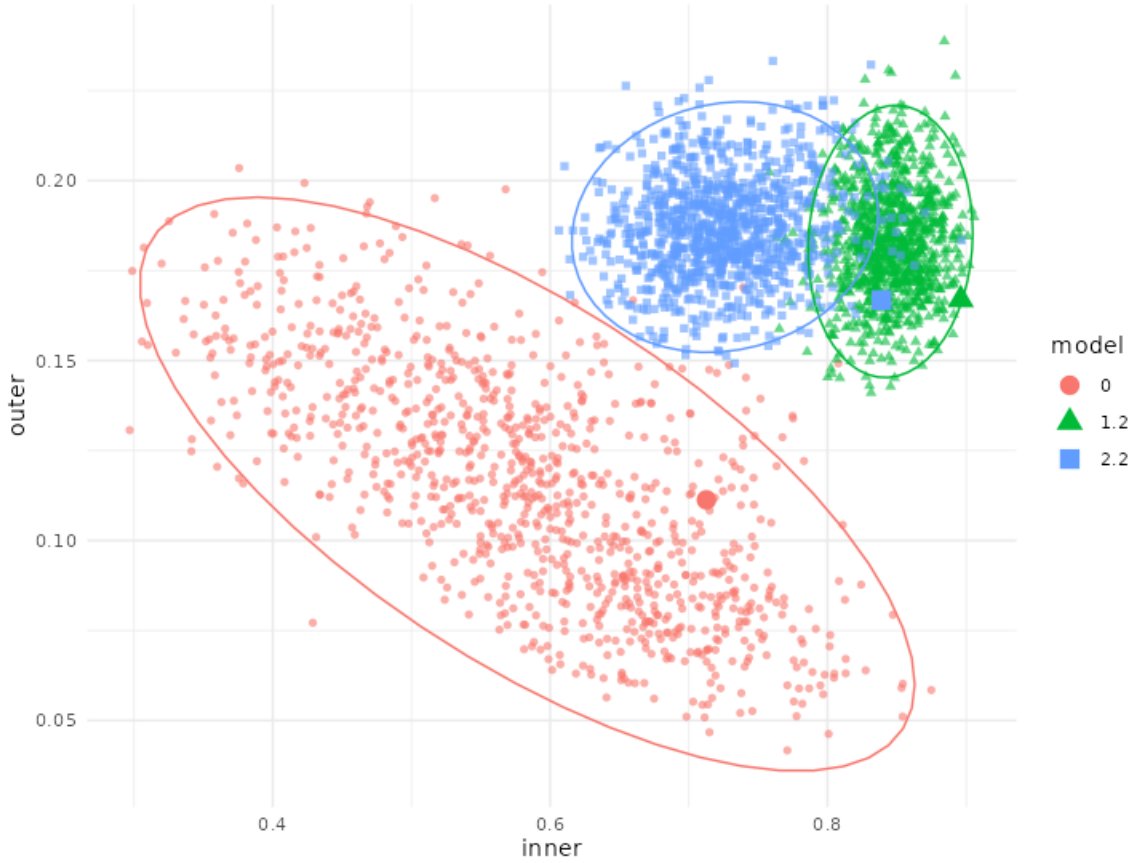

Figure 4.32: Bootstrap of models in Howell's datset. Bootstrap of the different models on the inner and outer AVE: Model 0 has just trancriptomic and microbiome data, model 1.2 has transcriptomic, microbiome and sample data and model 2.2 has transcriptomic, microbiome and sample data split in different blocks. The bigger points are the models on the original dataset.

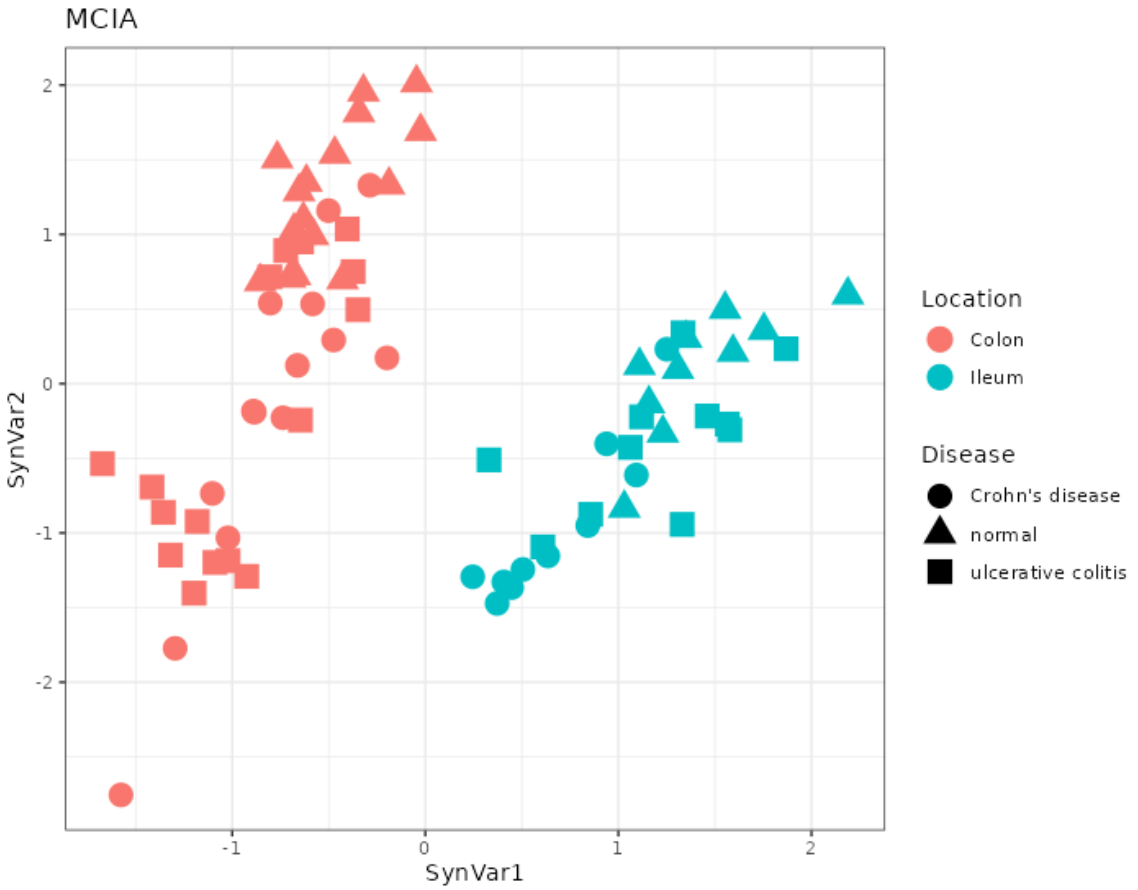

Figure 4.33: MCIA dimensions on the Howell's dataset. The first dimensions separates by location.

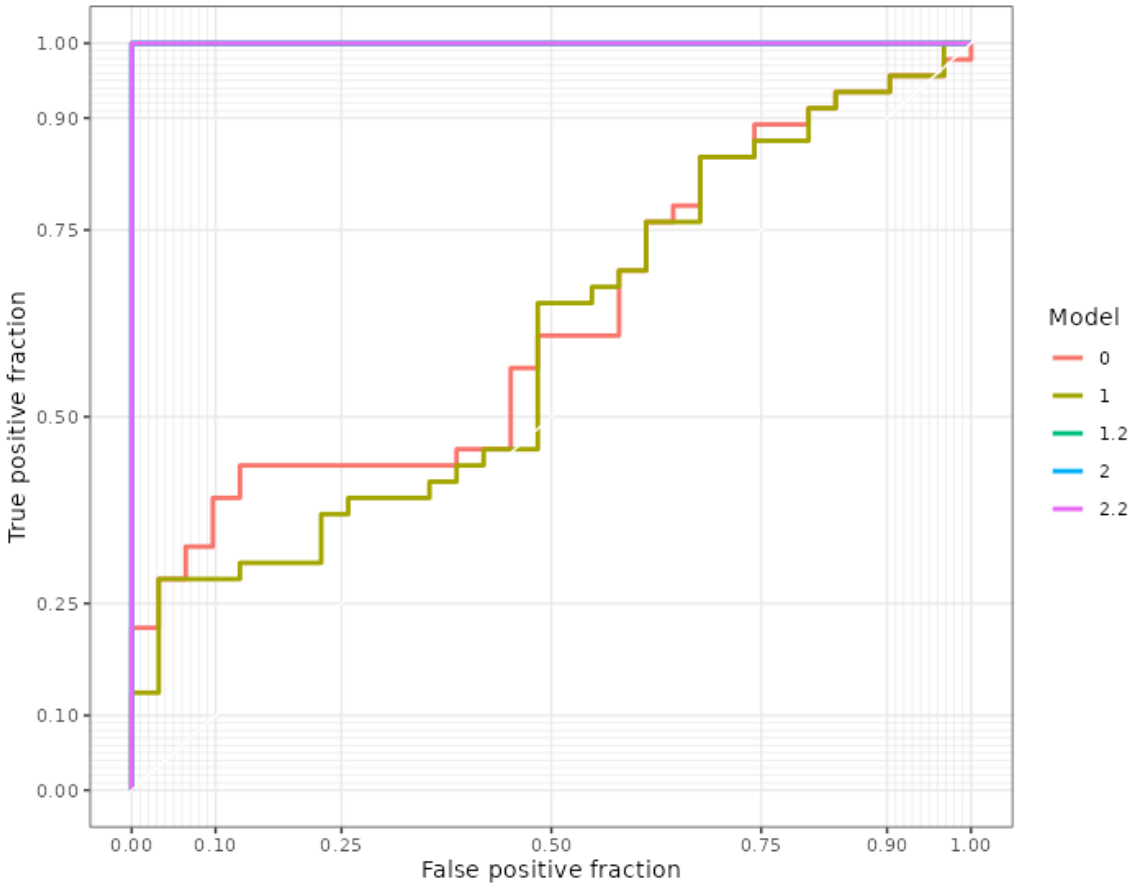

Figure 4.34: AUC of the RGCCA models in the Howell's dataset. The classification of the localization of the sample according to the first component of the gene expression of the models generated with RGCCA on the Howell's dataset.

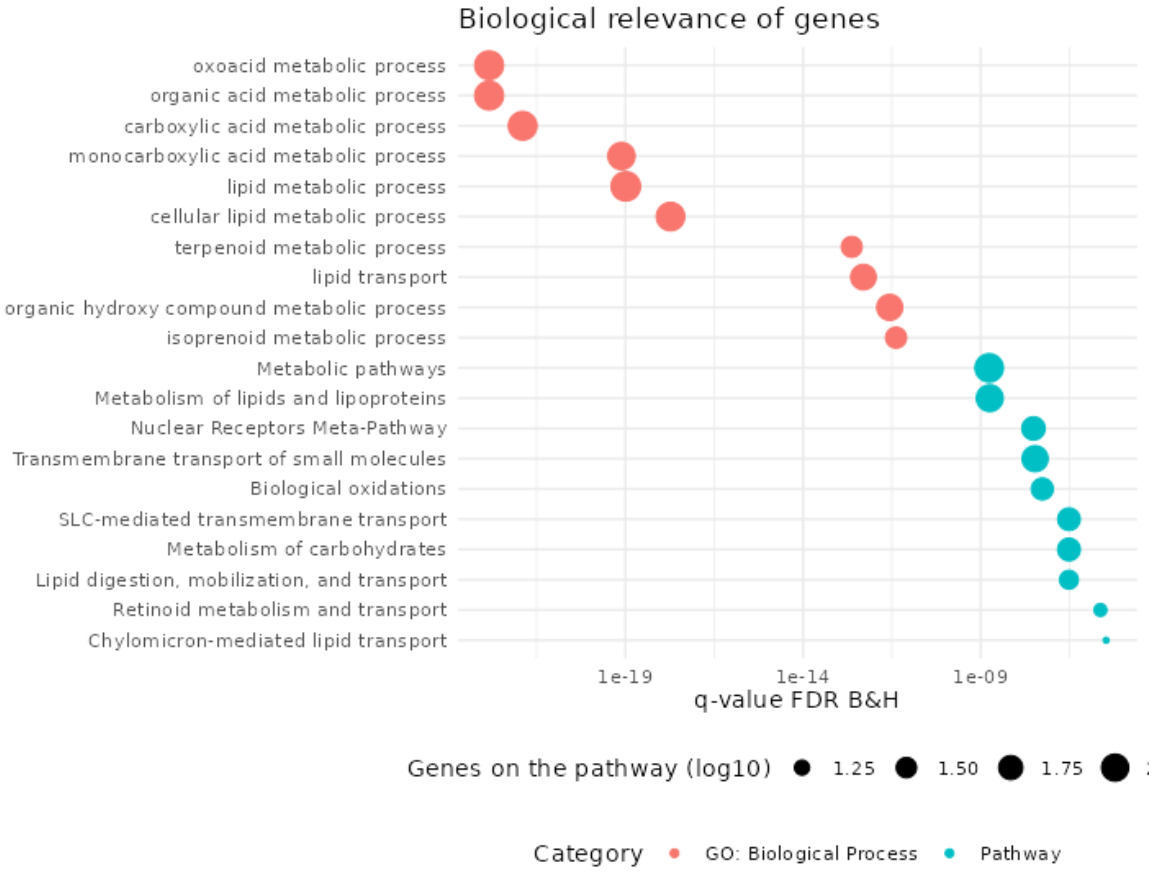

Figure 4.35: Significance of pathways on common genes in HSCT and Howell's dataset ordered by p-value, size according to the number of genes on the pathway found on the dataset and color blue for pathways and red for gene ontologies of biological process.

Discussion

In this chapter we will summarize the main findings in relation to the broad research community and other work as well as the impact of the results on further research in or clinical practice.

5.1 Preliminary steps

The quality of the initial RNAseq and 16S data is crucial for a valid integration analysis. In sequencing data, it is very important to avoid contamination and have enough amount of data for the analysis. To avoid contamination it is very important to control the protocol used to process the samples before the sequencing. In addition control samples can be added, a blank to correct for extraneous material and a sample with known content to confirm that the sequencing worked correctly.

Our lab uses a well established protocol to isolate and sequence RNA (Described on methods' section 3.2) and we rarely encounter problems during this phase. In contrast, DNA extraction and 16S sequencing had to be established and optimized, initially with the guidelines and support from collaborators especially Dr. Ilias Lagkouvardos. Indeed, as we encountered some problems we added blank samples to control for bacterial contamination during sample pre-processing or sequencing.

We initially processed the 16S data to obtain OTUs. OTUs were the standard some years back [198]. However, OTUs are not comparable between studies even those that use the same primers. For this reason, and after a suggestion from a reviewer (on the review process of [197]), we moved to process later 16S sequencing datasets to obtain ASV. ASV allow the comparison of the taxonomic imputation between studies using the same primers. However, comparing 16S taxa ASV or OTUs is hard and selecting the right tool to compare them is important [199].

It is also worth keeping in mind that several species have variable number of 16S rRNA genes. Higher sequence counts of certain 16S might not mean higher abundance of that species compared to another with lower 16S counts. However, the exact copy number status may change even within the same bacterial species, making the correction difficult [200]. It is however possible to accurately correct for copy number variation on mock populations where the species and the genomes are known (or at least the 16S rRNA), but harder or impossible on samples whose composition is unknown. To our knowledge there is no available method yet to do this.

Furthermore, the precision of 16S sequencing for classification of taxa is not enough to understand its role in the gut. Adherent invasive cells, are bacteria share genetic content with but show different behavior [201]. These bacteria are known to be more abundant on patients with IBD [201]. Thus, bacteria with the same complete sequence can play a different role *in vivo*.

Most of these steps were, however, out of my control as they were performed by a lab teammate or by collaborators. Other concerns, like primers used, 16S regions amplified are currently unavoidable.

5.1.1 The datasets

There is not an established methodology to calculate the size of cohorts for interaction or multi-omic studies. Indeed, there are no clear rules on which alpha power or which kind of relations can be tested. This might be due to the lack of mathematical background and modelization of the relevant relations in biology. Further research in this area might help finding which kind of relationships can be expected given certain dataset sizes. Usually size of the datasets is determined based on practical reasons: either costs or access to patients.

There have been some efforts to create artificial multi-omic datasets [140–145, 202, 203]. This could allow benchmarking different tools, impute missing data and do pilot studies. These efforts usually focus on RNA-seq, ATAC-seq (DNase-seq), ChIP-seq, small RNA-seq, Methyl-seq or proteomics but not 16S, microbiome or metagenomics data. In addition, they have also focused on finding relationships between samples missing relationships within cells (regulated DNA, protein recruitment, transcription factors, siRNA regulation, etc.) and between cells (immune response, adherence, cytokines signaling, etc.) that might surface at the sample level. To my knowledge there is no accurate artificial method to create 16S datasets or related microbiome datasets. This made impossible to simulate and compare different tools on a synthetic dataset with known relationships on this thesis to evaluate performance of different methods.

There is also no reference dataset for integration in IBD that accurately represents the disease. There is a large consortium effort with many samples [4], but this does not contain 16S and RNAseq data from the same location at the same time for as many samples as the datasets used on this thesis.

In order to validate results of methods we are limited to compare their results on different datasets. However, each dataset is collected with different goals and processed differently. In addition, authors upload their data to different, not centralized, data repositories, such as, gene expression omnibus (GEO), European Genome-Phenome Archive (EGA), European Nucleotide Archive (ENA), among others. Authors might also provide the processed data as supplementary material on their articles. Some projects whose primary purpose is providing data for the community establish their own dedicated sites to store the data [4].

There is not a resource were datasets of publications are collected with their characteristics, age, sex, sample location, extraction method, sequencing protocol, etc.. Finding different datasets with comparable data is thus challenging. Furthermore when pooling together datasets batch effect correction will most likely be necessary. There are tools to overcome this, but currently only work if the datasets share some features or samples in common [204]. These methods usually require datasets with similar RNA sequencing procedures.

In this thesis there were two in house datasets collected and sequenced: the HSCT and the BARCELONA dataset. In addition, we looked up for the most similar datasets

published to confirm our results outside our own cohorts. We found several with intestinal 16S and RNAseq data from the same location, to compare with our own datasets.

As explained in the methods' section 3.1.2, the HSCT dataset is a very unique cohort of CD patients undergoing hematopoietic stem cell transplant. This treatment is reserved to patients for which all the other treatment failed and this may be the only way to reach remission. These patients are closely monitored and followed up for several years. This treatment is not indicated for UC, so this dataset includes only CD patients.

The BARCELONA dataset (See appendix section D.1) includes samples from CD and UC collected prospectively in patients starting biological treatment and followed up for up to a year. Patients had shorter disease duration but were not refractory to biologic treatments and may even be naive to any biologic. Thus, analysis of this cohort could have provided us with a better understanding on the initial relationships between the microbiome and the mucosa of IBD patients.

Unfortunately, the 16S microbial sequence data from the BARCELONA cohort was not of enough qualtiy to make reliable analysis and confidently extract hypothesis or relationships (See appendix section D.1). It is not clear what happened and therefore we cannot hypothesize how to avoid such problems in the future. It could have been a problem with the DNA sample and/or of the sequencing process which did not include both positive and negative controls. Nevertheless, sample sequencing was repeated in an independent sequencing facility to overcome the limitations of the first dataset but problems persisted rendering this data suboptimal for further analysis.

What happened with the BARCELONA dataset highlights the importance of data quality checks. In the BARCELONA cohort despite no batch effect as checked with `experDesign` on the second sequencing, the low diversity indices suggested that there was some problem with the microbiome data. Caution and communication between all team members (i.e, clinicians, technicians and bioinformaticians) is important to discover this kind of problems and ensure the quality of the data.

Other datasets used from published sourced were assumed they had already passed enough controls and were of good quality. Nevertheless, they were screened to avoid quality issues by visually inspecting several principal components colored by several factors, checking gene expression and the microbiome profile and looking for known gene markers and microbiome taxonomy that might indicate quality problems. RNAseq data was usually compared using previously described methods on 3.6.1.

The Puget's dataset provided a good benchmark to test the methods and performance of RGCCA works. CNV/CGH is not comparable to microbial data, but the microarray data is similar to the RNAseq data from other datasets. We were not as much interested on the biology as on learning about the RGCCA method and its applicability to integrate different omics.

The Häsler IBD dataset was obtained using the same sequencing techniques from endoscopic biopsies as our dataset HSCT and BARCELONA. The 16S data was very similar to the HSCT dataset. We could confirm that the inteRmodel approach works on more than one dataset.

The taxonomy analysis of the different datasets was done differently. On HSCT following our collaborators advise we used IMNGS to annotate its microbiota data [159]. Afterwards we used SILVA database to annotate the microbiome data on their corresponding taxa.

The Morgan pouchitis dataset was also related to IBD but it included patients that underwent colectomy and samples are specific of the pouch or pre-pouch ileum. No healthy samples are included and there is no follow up. Thus, it was unlikely that a classification of the location could be achieved and that classification of the samples according to the microbiota could not be based on the disease (as all of them had undergone the same procedure and had the same disease type). Nonetheless, there seems to be a partial separation by location on the models on 4.25 that could partially be explained by different degrees of inflammation.

The Howell dataset includes both pediatric CD and UC (and non-IBD) samples. It is very similar to our BARCELONA dataset. As expected, time of the disease duration of these patients is lower, however, they were not followed up at different time points.

Overall, while challenging we were able to identify a few published datasets that were used to validate and compare the different models. Running such tests is essential to establish the best approach for data integration in the future an validate its results.

5.1.2 The methods

As seen on the introduction there are many different methods available and new tools and methods are frequently being released. The most up to date list of tools can be added to a collaborative list that was created with the purpose of providing access to the growing list of methodology of the scientific community. Methods differ in their quality, usage and the quality of the software. Some of them have been tested on several datasets but the most important validation process a method has to undergo is the mathematical validation.

With the increase of available analytical tools side by side comparison has become increasingly important. There are reviews that apply different tools to the same datasets [130], some are more theoretical [205], others are focused on a different area like metabolomics [100].

Few of these methods have been applied on IBD datasets. However, recently there has been a review focused on integration on IBD [148]. In this publication the authors suggest that one must be mindful of the gap between the experimental conditions and the real world. It also encourages to collect more data about the exposome (the environment patients are exposed to). It ends up advising to set up guidelines for multi-omic studies tailored to the field, coordinate a global framework to prevent redundant studies and to ensure efficient funding and resources and disseminate training and education on computational approaches to analyze multi-omic datasets.

Current methodological approaches focus on comparing tools on (several) previously published datasets [130], an approach that we have taken too. This approach was not used to compare different tools, but to validate findings of one dataset in other datasets. This is especially important due to the lack of golden datasets or a way to reliable simulate datasets as discussed on the previous section 5.1.1.

In IBD, many studies focus on finding some genes or bacteria to answer a narrow question they have in mind, like which bacteria are related to inflammation [81] or disease activity. In this thesis, the focus was on finding a good representation of the relationships that identify groups of genes and bacteria that were relevant to the disease in an orchestrated manner. We made the assumptions that the microbiome composition and the host's transcriptome were related. This assumption is backed up by several other previous studies supporting this relationship [41, 42, 206–208].

Tools that relate the variance of a block with other variables, both numeric and categoric, are needed to search which variables are important. PERMANOVA and `globaltest` served that purpose, but they do not give any insight into which specific microbial species are driving the association between the microbiome and the variables [209]. In addition, we could be missing some other important variables. It is known that other factors beyond the omics data collected, mainly environmental factors, genetic susceptibility and the immune response may play a role [210]. Sudhakar *et. al.* [148] recommend being conscious of the gap between the data available and the biological process. One variable we did not keep track was the microbial load which is linked to the gut's microbiome community variation [211]. This needs to be quantified at the time of DNA isolation and most studies, including our own lack this data.

We tried to find which genes and bacteria are correlated between them using `WGCNA`, a tool designed to find common co-occurring patterns based on correlations. It requires homogeneous samples, with a minimum of 12 samples per condition. However, when applied to the whole dataset there might be too much variance in order of WGCNA to find the proper signal. As we have patients of different ages and samples from different intestinal segments, both variables are highly related to the intestinal microbiota, this might be the reason why this method failed to achieve a good fit on the scale free topology with around 100 of mean connectivity. In addition, having microbiome and RNASeq in the same matrix, would probably be hard for the process to find good relationships if we applied the same normalization process to both of them without escalation. We could have tried to make smaller groups and then compare the modules between them but groups may have been too small since samples were from multiple segments and conditions.

We briefly considered using `STATegRa`. However, it is not possible to model specific interactions between blocks. The method implemented on `STATegRA` might be useful for cases where there is a great agreement between blocks or were environmental factor do not play a huge role on any of the blocks of data.

To identify related variables other methods use correlations between the variables [212]. On our dataset we visually explored the correlations for all the datasets, but the significant correlations were usually driven by an outlier, or there was not a good fit of the data due to missing data (data not shown). We tried filtering those correlations identified by the models, being less restrictive by removing those samples that did not have microbiome presence on at least 5 samples, and removing those samples without that microbiome presence, separating based on intestinal segment and a combination of all. None of these variations provided a clear insight over which genes were correlated to which bacterial signatures. This is in contrast to other publications that relied purely on the Spearman ρ metric [80].

We developed and applied `BaseSet` to find these relationships using a different ap-

proach. However, it failed because it is computationally expensive to calculate the likelihood of 1500 variables; there are too many combinations. In addition, the numeric precision of said calculations suffers from the floating point problem and must be carefully considered [213]. To support multiplying more than 1000 float numbers a different strategy such as using log values might be better. We could not come up with a better strategy to find all the combinations needed, perhaps a better method exists that could be used to find which are the terms more influential to the end result. During the peer-review process of the package for its acceptance on rOpenSci, some concerns were raised about conflating probabilities with fuzzy-sets. For all these reasons this approach was no longer pursued. However, the package was mentioned as top 40 packages added on CRAN that month and it will be useful under other circumstances or when less combinations are possible.

The development of the `experDesign` package helped us to avoid batch effects on the sequencing step. However, as seen in the previous section, batch effects were not completely prevented. As discussed in the related publication [193], there are several tools already focused on this problem: `OSAT` ([214]), `anticlust` ([215]) and `Omixer` ([216]). But these tools have some shortcomings, that are covered by `experDesign`: `OSAT` cannot handle missing values and does not work well for arbitrary batches, `anticlust` only accepts numerical variables but it is based on a powerful mathematical theorem and `Omixer` has bugs that prevent comparing with the other tools with no possible workaround.
In addition, `experDesign` received requests to have a new feature for expanding experiments. This might help improve the quality of bigger datasets to ensure they can be extended in several sequencing runs. This feature would be useful for multi-omics datasets or in big cohorts to minimize batch effects associated with long running collection of samples.

We selected `MCIA` as a baseline to compare our method because it works well on a wide range of datasets, has a good documentation as well and it is widespread used. The method was developed after `RGCCA` and recently there have been publications that show that it outperforms other methods on its versatility on different contexts [130]. On the dataset analyzed we found that it performed similarly to `inteRmodel` but this will be discussed on section 5.3.

5.2 Designing models

Previous publications using `RGCCA` in IBD have focused on validating genes *DUOX2* and *APOA1* as inflammation predictors ([81]) from previously published articles [217]. Some publication tried to summarize the existing relationships in IBD [80], but none were focused on finding the relationships in IBD using `RGCCA` as we did.

There are many variables outside transcriptomics and the microbiome that may be relevant in disease homeostasis. These variables should be included on the models to find which genes and bacteria are truly related and not confounded by other factors. In addition, the relationships between the blocks are unknown on both the strength and interaction. To model the relationships the connections between blocks had to be selected on `RGCCA`. Last, the variables that belong to a block should be carefully

considered as the assumption is that the whole block is correlated with other blocks connected.

If one has preexisting theories about the data, a specific model can be used stating these known or hypothetical relationships. However, if new relationships are being explored or no prior assumptions on the data are held the models should be created with random links between blocks, and evaluate which model is better.

The connections tested required that all models should have all blocks indirectly connected to other blocks with no blocks left unconnected. This avoids optimizing two different networks of blocks that are not connected between them, thus, forcing the model to represent all the information.

Typically blocks are defined by each omic data origin and no other information is included. However, we knew that the transcriptome is mainly related to the location of the samples, as we have seen it on the PCAs, and we expected that the microbiome would be more related to the patients demographic characteristics and influenced by dietary and other environmental factors. This is especially important because in our datasets we have samples from the same patient and time from multiple locations. On other studies there are less samples per individual and timepoint (if there are several timepoints) [80, 150, 151].

On models of family 2, variables were grouped according to type into the Demographic block, Time block, and Location block (see tables on design of Puget's 4.4, 4.3; HSCT 4.13 4.14, 4.15; Häsler: 4.21, 4.22, Morgan's 4.28, 4.29, Howell's 4.33 models).

The exploration of design on Puget's dataset, and the datasets analysis of HSCT, Häsler and Morgan's datasets was published after peer review [197].

5.3 Evaluating models

To evaluate a model `RGCCA` provides the inner and outer average variance explained (AVE); (see section 3.3.1.2). As the inner AVE measures how good does the data fit in the model, we used it to evaluate and compare the models.

Furthermore, bootstrapping was used to evaluate the fitness of a design on a diverse collection of datasets. Although on `inteRmodel` there is the option to use a leave-one-out procedure, we did not use it to evaluate the fitness of the models.

Using an external cohort to validate the same model, or using a different method to see if it finds the same relationships or explains the data as accurately is also a common approach to evaluate and validate models. Using the same approach on different data helps to ensure the replicability of the results [218].

We used the same approach on four different cohorts, with different origins and types of samples, but all related to the IBD popultation including 16S data and intestinal transcriptome. Some of them have multiple samples from the same individuals while others do not [80].

We also compared our method with a different one to see how generalizable are the results. Of the multiple methods available we used `MCIA` [219]. We compared it with

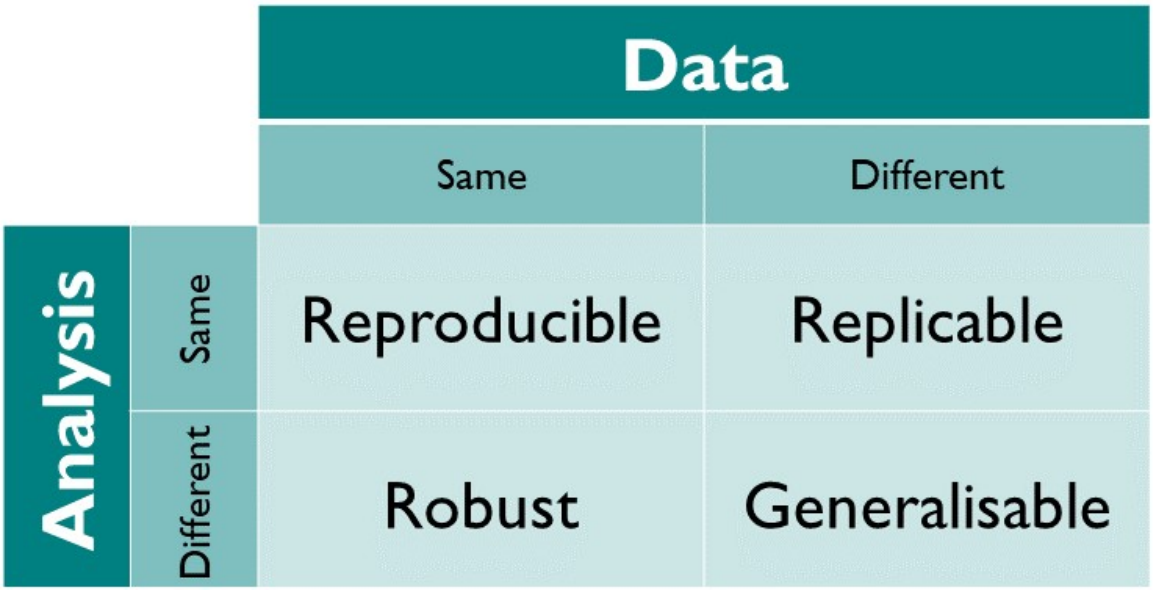

Figure 5.1: Reproducibility matrix indicating the terminology used between using the same method and the same data. Figure from The Turing Way: A Handbook for Reproducible Data Science (Version v1.0.1).

`inteRmodel` by looking at the area under the curve for classifying the samples with the canonical component of the transcriptome data according to their location.

The procedure of separating independent variables in their own block of data and later search the best model that fits the data provides a good strategy that should be considered for integration efforts. The procedural method of searching a model and testing it is implemented on `inteRmodel`. But the most important thing is to consider which variables are independent of which and if they can be separated into a block for later usage on the modeling (See [220]). If not done properly it can lead to undesired conclusions like the effect known as Simpson's paradox [221]. Unfortunately this cannot be automated and it is left for the practitioner.

Splitting the variables of each sample into several blocks forces `RGCCA` to adjust for a new canonical dimension. The omics block such as gene expression and 16S data could be split, as the expression of some genes influence other genes, such as transcriptor factors (First mentioned on [222]) , miRNA [223], and siRNA [224]. All these interactions and regulations could distort the canonical correlations. However, the nature of this relationships is not linear and the interaction between them is multiple and very complex. Its complexity has prevented to accurately account for all subcellular reactions at speeds that could be useful (it has only been recently accomplished for a prokaryote cell [225]). In addition, these interactions are time dependent, not linear and are highly interconnected between many variables. For these reasons gene expression and 16S blocks were not split into several blocks.

Besides comparing the results of different methods, these models need to be evaluated by the insights they provide on the biological system they are being applied to, in our case IBD. So far the models were only discussed on their technical merits.

It is known that the mucosal transciptome is related to sample localization (of)i.e. small or large intestine) and disease activity [226]. This can be seen on the PCAs

which separate colon an ileum on the first dimension: See the PCAs of the different datasets 4.9, 4.18 and 4.30. The difference is so great that many times the colon and ileum samples are analyzed separately. As such, it was a reasonable assumption to expect models to reflect these differences on the gene expression canonical component.

The variability of microbiome difficults finding clear patterns (See the PCAs of the different datasets: 4.8, 4.17, 4.23 and 4.29). In healthy humans, it has been suggested that there is a shared microbial patterns among groups of individuals [227]. Such theory proposes three groups of similar microbiota composition, named enterotypes. However, the enterotypes classification is not unanimously accepted [228, 229]. Some suggest it is an artifact of the methodology used.

There are many factors that influence the gut microbiome [230]. But the role of microbiome in IBD has gained a lot of attention in the last 10 years [17, 231, 232]. The microbiome has been associated for instance with treatment response in CD and suggested it could be useful for building an improved classifier for CD [233].

In addition, it is known that microbiome of patients in remission is different from non-IBD patients [234]. So even if two people are apparently healthy at a given time their microbiome might not be similar.

The relapsing nature of IBD, suggests that at different time points the gene expression or the microbiome might be different. For this reason, time is an important variable when modelling the disease. If samples at multiple timepoints are available the time difference should be taken into account to identify the state of the disease for each patient. Even in healthy non-IBD patients, multiple samples with microbiota from the same patients could provide common microbiome signature, which can help identify altered microbiome states later on.

Of the cohorts analyzed in this thesis, the HSCT and the Häsler dataset are the only ones including time related variables. However, the Häsler dataset only had age at date of sampling, and there were not several samples for the same patients at different timepoints. This leaves the HSCT dataset as the only one with multiple timepoints from the same patients. Having that many samples from the same patient might explain why the classification of the disease on this cohort works so well, despite being a cohort of complicated patients with several background treatments, refractory to all previous treatments, some of them having undergone surgery too and undergoing severe immunosupression during follow up. Despite all of this, there was large agreement in the genes and pathways identified as relevant between datasets (See figure 4.35).

Comparing the microbiome between datasets is less straight forward. If the same primers are used the ASV could be directly compared. However, the ASV' length might be different. In addition, in this thesis we used dataset obtained with different primers and also OTUs. For these reasons a direct comparison is not possible. A comparison of the taxonomy of the annotated microbiome on the datasets is the next possibility.

However, some datasets used did not provide the annotation (Häsler's and Morgan's datasets). So we are are left with comparing OTUs (HSCT dataset) with ASVs (Howell's). Despite the errors on ASV annotation ([235]), OTUs from the HSCT dataset can be compared to ASV from the Howell's dataset. There were very few

common taxonomic levels selected on both datasets (data not shown). This was not a surprise, as there are many factors that influence one's microbiome profile and the diversity indexes already showed high differences between samples of the same dataset.

The microbiome of each dataset seemed capable to classify the samples according to the disease (data not shown). However, on further evaluation via bootstrapping this classification was not significant, as different relatively big amount of ASVs were able to classify the samples. This might be due to the uniqueness of microbial composition to each sample.

In Howell's and Häsler dataset, the best model based on AVE is 1.2; which separates the microbiome component by location too (See Figure 4.20 and table 4.23 for Häsler dataset and figures 4.31 and table 4.34 for Howell dataset). The 2.2 models according to the inner AVE were not that far away from models 1.2. This indicates that the relationship of the microbiome is stronger than with the location, but both factors should be considered when looking for relationships of genes with microbiome.

5.4 Implications

On this thesis several methods has been developed to help multi-omic data integration. `experDesign` was implemented during the initial steps when moving from bench to in-silico analysis. `BaseSets` and `inteRmodel` are useful for computational analysis. `BaseSets` might help beyond integration analysis such as single cell annotation[1].

Some studies using host transcriptomics of fecal wash infer inflammation without colonoscopy [236]. This would help patients to avoid an unpleasant experience, and reduce the usage of clinical facilities. It could be possible that just sequencing the intestinal microbiome could be enough to identify the patients' disease. However, this requires further validation to ensure that the diagnosis is accurate enough on a diverse and big population. There are already studies on this direction, not only for IBD or intestine but for several different human regions [4].

With our studies we hoped that we would provide which bacteria played a role on the disease, or which genes and bacteria are related on IBD. We obtained a list of putative genes and bacteria but not a clear pairing of which genes interact with which bacteria. This could mean that the microbiome community is related to all the genes. It could also mean that the methods are not powerful enough to find more tailored relationships as normalizations and generalizations present on `RGCCA` do not allow to subgroup or classify the variables already detected. Maybe a different method that would not depend on the same principles might be able to detect finer relationships on the variables selected.

There is a disconnect between the computational side and the experimental side, driven by the difficulty to design and perform an experiment to test the new information that multi-omic experiments provide. This is referred as the gap in other publications [148]. Closing this gap between the computational methods and the data

[1] See an exploration of this using AUCell https://bioconductor.org/packages/AUCell on this website: https://llrs.github.io/BaseSet_scRNAseq/AUCell.html.

origin and practical usages would potentially require closer collaboration between clinicians, statisticians, bioinformaticians and research software engineers; in addition to creative ideas accounting for the standard procedure on hospitals and points of care.

However, it is also possible that further developments or creatively applying statistical methods might help closing the gap. For instance, there are many combinations of possible interactions. These interactions between microbiome and genes are currently hard to explore statistically. Further research on how to reduce the space of possible combinations of bacteria or evaluate which combinations are more important might be useful on the future. This might involve using more network integration methods.

Krassowski's *et. al.* ([95]) advise on software engineering and reproducibility practices to share awareness with new researchers in multi-omics for end-to-end workflow. In addition the recent recommendation about integration on IBD ([148]) suggests that there is still much to be done to increase the results on these projects and the effectivity. We agree with them and encourage other researchers to carefully consider the proposed methodology before starting their own projects.

There seems to be a tendency of multi-omics projects to focus on metagenomics, metatranscriptomics and metaproteomics and abandon plain 16S sequencing [237]. This could be explained by the numerous problems that 16S sequencing has. Some of these are avoided or solved by using these other omics techniques. For instance the metatranscriptomics and metaproteomics of bacteria could help detect what is actively being produced by the microbiota and the metagenomics, what genes are actually there not just restricted to 16S rRNA sequences. But these methods also have their shortcomings. Metagenomics still does not detect adherent invasive cells and metatranscriptomics and metaproteomics can not explain which bacteria is expressing which genes (even if paired with metagenomics or 16S data). The single cell revolution might provide more insights in the future, but at the moment the first studies of single-cell multi-omics are being published.

Perhaps as suggested in other publications network methods might be able to provide more detailed information about the relationships [238]. But, it is unclear how complete and valid these networks are. Current literature focus on already known and studied genes instead of on more novel and with higher relevant genes [239]. By extension, this translates in bias in networks and pathways resources available for pathway analysis and integration methods. This could explain why there are many genes selected by the models with few pathways or gene sets.

To study complex systems, new technologies like "organ-on-chip" or more specifically "gut-on-chip" are being developed [240]. These systems expand on the already useful technique of using organoids to better mimic the epithelial cross-talk in the laboratory. Latest developments include the addition of separate environments and distinct flow on these environments. By introducing gut microbiota, these systems will help studying the interaction of the gut microbiota and the intestinal epithelium. However, it is not clear how accurately account for other factors such as the presence of inflammation.

In summary, as shown with the PERMANOVA approach, several factors affect the disease and the relationship between gut's microbiome and gut's mucosa. Analysis or comparisons without taking them into account might provide misleading or false results. To our knowledge this is the first study using these variables as part of the

integration study with canonical correlations. We hope this provides a first approximation to accurately understand the relationships between genes and bacteria on IBD.

Conclusions

In this doctoral thesis, we assessed the relationships between the microbiota and the gut transcriptome in inflammatory bowel disease.

It allows us to conclude the following:

6.1 Study 1: Multi-omic modelling of inflammatory bowel disease with regularized canonical correlation analysis

1. Applying methods that use information about the samples show better results than omitting this information.
2. Correlations are not enough to identify relationship between genes and microorganisms.
3. The `inteRmodel` method provides consistent connections between blocks on different datasets of inflammatory bowel disease.

6.2 Study 2: Genes and microbiome relationship on inflammatory bowel disease

1. The host's transcriptomics is heavily influenced by location and to a lesser degree by the type of activity.
2. The microbiome composition relates better to the sample location than to the disease status but both are almost equally important
3. Microorganisms highly tied to intestinal location might indicate the disease type, (Crohn's disease, ulcerative colitis or non-IBD) of the patients, but microorganisms related to the disease are common in all the locations.

Acknowledgments

A la directora, tutora, jefa i tants papers que fas, moltes gràcies Azu per haver estat a sobre de la tesis, he aprés molt de tot sobre com treballar amb intensitat i eficiencia, tocant de peus a terra. També per preguntar i tenir presents sobre com es pot fer servir això i com pot ajudar cada idea que tenia.

Gràcies Juanjo per donar-me la oportunitat d'entrar a IDIBAPS i per tota la ajuda al llarg d'aquests anys per ser un bioinformàtic independent. Sempre has estat disponible quan no sabia com seguir i aportaves alguna idea o paquet per probar.

A tot el laboratori per el bon ambient que hi ha hagut fent les alegries més grans al llarg de aquests anys i les adversitats més lleugeres. Cadascú ho ha fet a la seva manera: a l'Ana per ser el complement perfecte de l'equip bioinformàtic i sempre estar disponible per aconsellar-me. L'Alba pel seu optimisme i perseverancia, también por tu buen criterio diseñando, espero que algo se me pegue. A l'Isa pel seu detallisme i exemple de fer bé les coses, les orquidies siempre están preciosas! La Marisol per la seva senzillesa i estar sempre pendent dels petits detalls de la paperasa. La Elisa per la seva alegria i il · lusió que mai defalleixen i s'encomanen. La Victòria per la seva paciència i consells sobre com descriure els projectes. L'Angela per acceptar amb paciència les meves sortides friquis. La Míriam perqué sempre ens toca moure trastos, a veure si algun dia montem el wallapop científic! i l'Iris que sempre té alguna cosa a aportar.
Moltes gràcies a totes, només lamento no haver aprés més de vosaltres.

Hi ha persones que estan sempre com la Pepa, gràcies per no tancar-me mai el email, i altres que marxen a altres projectes. Gràcies per compartir una part d'aquest projecte amb mí: l'Aida per tota l'empenta que ha posat en aquests anys, moltes gràcies per introduir-me en el món de la microbiota! també per la teva acollida al lab. La Montse per organitzar els primers viatges fòra i ensenyar-me a organitzar-me. La Núria per introduir-me en el grup i animar-me a fer el doctorat amb l'Azu. També en Daniel pel curset de processament de RNAseq.

A tots de l'equip de l'hospital, per intentar entendre aquestes caixes negres de les tesis i ajudar a que no perdi de vista l'objectiu de la recerca; en especial al Julià, per fer sempre les millors preguntes i a la Maica per resoldre molts dubtes a l'inici.

A tot l'equip del Pau que em van acompanyar en els meus primers passos en la investigació, la Bea, la Júlia, la Délia, la Mar, en Luis, la Elisa... Alguns que ja no estan per la planta 3 com en Josep o la Carolina.

A tota la familia, els que estan més aprop i els que estan més lluny. Als pares per seguir amb interès tot el procés i procurar entendre alguna cosa. A la Maria per convidar-me sovint a casa seva a descansar. A la Maria del Mar per la teva alegria i plantejar reptes amb el python. A la Marta pels consells, el cotxe i les excursions. A l'Ana per seguir amb interés la investigació i estar pendent dels avanços sobre la

microbiota. A l'Albert per convidar-me a dinar sovint i conèixer racons i històries de Barcelona o la familia.

Als amics de Xaloc, en Mario per posar la casa tantes vegades al llarg de la tesis, un día tienes que venir a la mía! L'Albert Mateo per les refrescants sortides en bici i excursions, a ver si hacemos más y con todos! A, l'Alejandro que sempre ha estat un suport moral com a doctorand. En Bruno per acollir-me a casa seva (he de repetir la visita a Munich!). A l'Edu per estar sempre disposat a acostar-me en cotxe o trobar un lloc fàcil d'arribar en transport públic. Als amics de la carrera, en Frederic per compartir molts dinars, penes i alegries. En Joan per marcar el camí, sempre has estat per davant! i l'Aleix, espero que la bioinformàtica vagi a més!

Els amics d'Obenc, el Cristian per tantes aventures que em viscut aquests anys. En Daniel per ensenyar-me que el temps no es barrera per fer coses (Encara no m'has presentat els teus gossos!). A en Lluís que tampoc no pares mai i m'has transmés la teva gran ambició. A en Ricard per totes les discussions sobre el futur profesional, de la programació, els jocs, películes i tantes coses, segur que trobes una manera de passar al seguent nivell! A en Joaquín per ensenyar-me el món empresarial de software i tants detalls tècnics. A tots els que han vingut a fútbol, en especial al Josep M^{a} per ser l'incansable impulsor i treure gent de sota les pedres i a en Josep per ser el més constant, ja chutaré més sovint!

I also received many help from numerous and anonymous people from several corners of the world. They might have shared some interesting tip, advice or tale on twitter, or answered my questions on Bioconductor's support forum, Biostars, Bioinformatics SE or Stackexchange network. Also those that made and provided the software for free that made all this possible on your own time, R core, CRAN, Bioconductor and rOpenSci team, package developers and maintainers of the ~500 packages I use.

Dr. Häsler for kindly providing data of the samples from their cohort and Dr. Cristian Hernández and Dr. Mark S. Silverber for their collaboration and selfless sharing the data of their cohort.

Last but not least, those without all this could not be possible. Thanks to all the patients that allowed clinicians to collect samples and data for research. I hope that the knowledge we gain with on this thesis and other studies will bring us closer to heal you.

References

1. Raine T, Verstockt B, Kopylov U, Karmiris K, Goldberg R, Atreya R, et al. ECCO topical review: Refractory inflammatory bowel disease. Journal of Crohn's and Colitis. 2021;15:1605–20.
2. Jairath V, Feagan BG. Global burden of inflammatory bowel disease. The Lancet Gastroenterology & Hepatology. 2020;5:2–3.
3. Burisch J, Munkholm P. The epidemiology of inflammatory bowel disease. Scandinavian Journal of Gastroenterology. 2015;50:942–51.
4. Human Microbiome Project Consortium BA, Nelson KE, Pop M, Creasy HH, Giglio MG, Huttenhower C, et al. A framework for human microbiome research. Nature. 2012;486:21521.
5. Shaw KA, Bertha M, Hofmekler T, Chopra P, Vatanen T, Srivatsa A, et al. Dysbiosis, inflammation, and response to treatment: A longitudinal study of pediatric subjects with newly diagnosed inflammatory bowel disease. Genome Medicine. 2016;8:75.
6. Mukhopadhya I, Hansen R, El-Omar EM, Hold GL. IBD—what role do Proteobacteria play? Nature Reviews Gastroenterology & Hepatology. 2012;9:219–30.
7. Zhang Y-Z, Li Y-Y. Inflammatory bowel disease: pathogenesis. World Journal of Gastroenterology. 2014;20:91–9.
8. Hugot J-P. Genetic origin of IBD. Inflammatory Bowel Diseases. 2004;10:S11–5.
9. Silva FAR, Rodrigues BL, Ayrizono M de LS, Leal RF. The Immunological Basis of Inflammatory Bowel Disease. Gastroenterology Research and Practice. 2016;2016:2097274.
10. de Mattos BRR, Garcia MPG, Nogueira JB, Paiatto LN, Albuquerque CG, Souza CL, et al. Inflammatory Bowel Disease: An Overview of Immune Mechanisms and Biological Treatments. Mediators of Inflammation. 2015;2015:493012.
11. McGovern DPB, Kugathasan S, Cho JH. Genetics of Inflammatory Bowel Diseases. Gastroenterology. 2015;149:1163–1176.e2.
12. Satsangi J, Silverberg MS, Vermeire S, Colombel J-F. The Montreal classification of inflammatory bowel disease: controversies, consensus, and implications. Gut. 2006;55:749–53.
13. Horowitz JE, Warner N, Staples J, Crowley E, Gosalia N, Murchie R, et al. Mutation spectrum of NOD2 reveals recessive inheritance as a main driver of early onset crohn's disease. Scientific Reports. 2021;11:5595.
14. Kumar M, Garand M, Al Khodor S. Integrating omics for a better understanding of inflammatory bowel disease: A step towards personalized medicine. Journal of Translational Medicine. 2019;17:419.
15. Jostins L, Ripke S, Weersma RK, Duerr RH, McGovern DP, Hui KY, et al. Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature. 2012;491:119–24.
16. Momozawa Y, Dmitrieva J, Théâtre E, Deffontaine V, Rahmouni S, Charloteaux B, et al. IBD risk loci are enriched in multigenic regulatory modules encompassing putative causative genes. Nature Communications. 2018;9.
17. Khanna S, Tosh PK. A Clinician's Primer on the Role of the Microbiome in Human Health and Disease. Mayo Clinic Proceedings. 2014;89:107–14.
18. Swidsinski A, Ladhoff A, Pernthaler A, Swidsinski S, LoeningBaucke V, Ortner M,

et al. Mucosal flora in inflammatory bowel disease. Gastroenterology. 2002;122:44–54.
19. Tamboli CP, Neut C, Desreumaux P, Colombel JF. Dysbiosis in inflammatory bowel disease. Gut. 2004;53:1–4.
20. Ott SJ. Reduction in diversity of the colonic mucosa associated bacterial microflora in patients with active inflammatory bowel disease. Gut. 2004;53:685–93.
21. Kostic AD, Xavier RJ, Gevers D. The microbiome in inflammatory bowel disease: Current status and the future ahead. Gastroenterology. 2014;146:1489–99.
22. Sender R, Fuchs S, Milo R. Revised Estimates for the Number of Human and Bacteria Cells in the Body. PLOS Biology. 2016;14:e1002533.
23. Sankarasubramanian J, Ahmad R, Avuthu N, Singh AB, Guda C. Gut microbiota and metabolic specificity in ulcerative colitis and crohn's disease. Frontiers in Medicine. 2020;7.
24. Lopez-Siles M, Martinez-Medina M, Busquets D, Sabat-Mir M, Duncan SH, Flint HJ, et al. Mucosa-associated Faecalibacterium prausnitzii and Escherichia coli co-abundance can distinguish Irritable Bowel Syndrome and Inflammatory Bowel Disease phenotypes. International journal of medical microbiology: IJMM. 2014;304:464–75.
25. Ferrer-Picón E, Dotti I, Corraliza AM, Mayorgas A, Esteller M, Perales JC, et al. Intestinal Inflammation Modulates the Epithelial Response to Butyrate in Patients With Inflammatory Bowel Disease. Inflammatory Bowel Diseases. 2020;26:43–55.
26. Darfeuille-Michaud A, Neut C, Barnich N, Lederman E, Di Martino P, Desreumaux P, et al. Presence of adherent Escherichia coli strains in ileal mucosa of patients with Crohn's disease. Gastroenterology. 1998;115:1405–13.
27. Tsilingiri K, Barbosa T, Penna G, Caprioli F, Sonzogni A, Viale G, et al. Probiotic and postbiotic activity in health and disease: comparison on a novel polarised ex-vivo organ culture model. Gut. 2012;61:1007–15.
28. Vanderpool C, Yan F, Polk BD. Mechanisms of probiotic action: Implications for therapeutic applications in inflammatory bowel diseases. Inflammatory Bowel Diseases. 2008;14:1585–96.
29. Isaacs K, Herfarth H. Role of probiotic therapy in IBD. Inflammatory Bowel Diseases. 2008;14:1597–605.
30. Plaza-Diaz J, Ruiz-Ojeda FJ, Gil-Campos M, Gil A. Mechanisms of action of probiotics. Advances in Nutrition. 2019;10:S49–66.
31. Morelli L, Capurso L. FAO/WHO guidelines on probiotics: 10 years later. Journal of Clinical Gastroenterology. 2012;46:S1.
32. Okumura R, Takeda K. Roles of intestinal epithelial cells in the maintenance of gut homeostasis. Experimental & Molecular Medicine. 2017;49:e338–8.
33. Faria AMC, Mucida D, McCafferty D-M, Tsuji NM, Verhasselt V. Tolerance and inflammation at the gut mucosa. Clinical & Developmental Immunology. 2012;2012:738475.
34. Hisamatsu T, Kanai T, Mikami Y, Yoneno K, Matsuoka K, Hibi T. Immune aspects of the pathogenesis of inflammatory bowel disease. Pharmacology & Therapeutics. 2013;137:283–97.
35. Michielan A, D'Incà R. Intestinal Permeability in Inflammatory Bowel Disease: Pathogenesis, Clinical Evaluation, and Therapy of Leaky Gut. Mediators of Inflammation. 2015;2015:628157.
36. Mayorgas A. Human Primary Organoid-Derived Epithelial Monolayers as a Novel Strategy for the Study of Adherent Invasive Escherichia coli pathogenicity and the

effects of Postbiotics on Intestinal Epithelial Function. PhD thesis. 2021.
37. Neurath MF, Fuss I, Kelsall BL, Presky DH, Waegell W, Strober W. Experimental granulomatous colitis in mice is abrogated by induction of TGF-beta-mediated oral tolerance. Journal of Experimental Medicine. 1996;183:2605–16.
38. Corraliza AM, Ricart E, López-García A, Carme Masamunt M, Veny M, Esteller M, et al. Differences in Peripheral and Tissue Immune Cell Populations Following Haematopoietic Stem Cell Transplantation in Crohn's Disease Patients. Journal of Crohn's and Colitis. https://doi.org/10.1093/ecco-jcc/jjy203.
39. Strachan DP. Hay fever, hygiene, and household size. BMJ : British Medical Journal. 1989;299:1259–60.
40. Scudellari M. News feature: Cleaning up the hygiene hypothesis. Proceedings of the National Academy of Sciences of the United States of America. 2017;114:1433–6.
41. Thomas GAO, Rhodes J, Green JT. Inflammatory bowel disease and smoking—a review. Official journal of the American College of Gastroenterology | ACG. 1998;93:144149.
42. Cornish JA, Tan E, Simillis C, Clark SK, Teare J, Tekkis PP. The risk of oral contraceptives in the etiology of inflammatory bowel disease: a meta-analysis. The American Journal of Gastroenterology. 2008;103:2394–400.
43. Kaufmann HJ, Taubin HL. Nonsteroidal anti-inflammatory drugs activate quiescent inflammatory bowel disease. Annals of Internal Medicine. 1987;107:513–6.
44. Bitton A, Dobkin PL, Edwardes MD, Sewitch MJ, Meddings JB, Rawal S, et al. Predicting relapse in Crohn's disease: a biopsychosocial model. Gut. 2008;57:1386–92.
45. Baumgart DC, Sandborn WJ. Crohn's disease. The Lancet. 2012;380:1590–605.
46. Khanna S, Shin A, Kelly CP. Management of clostridium difficile infection in inflammatory bowel disease: Expert review from the clinical practice updates committee of the AGA institute. Clinical Gastroenterology and Hepatology. 2017;15:166–74.
47. Guardiola J, Lobatón T, Cerrillo E, Ferreiro-Iglesias R, Gisbert JP, Domènech E, et al. Recommendations of the Spanish Working Group on Crohn's Disease and Ulcerative Colitis (GETECCU) on the utility of the determination of faecal calprotectin in inflammatory bowel disease. Gastroenterología y Hepatología (English Edition). 2018;41:514–29.
48. Sands BE. Biomarkers of Inflammation in Inflammatory Bowel Disease. Gastroenterology. 2015;149:1275–1285.e2.
49. Corraliza Márquez A. Immune mechanisms involved in inducing remission in Crohn's disease patients undergoing hematopoietic stem cell transplant. PhD thesis. 2019.
50. Bassolas Molina H. Resposta T específica contra antígens de la microbiota comensal en la malaltia de Crohn i inhibició de ROR?t com a estratègia terapèutica. PhD thesis. 2018.
51. Daperno M, D'Haens G, Van Assche G, Baert F, Bulois P, Maunoury V, et al. Development and validation of a new, simplified endoscopic activity score for crohn's disease: The SES-CD. Gastrointestinal Endoscopy. 2004;60:505–12.
52. Best WR, Becktel JM, Singleton JW, Kern F. Development of a Crohn's Disease Activity Index: National Cooperative Crohn's Disease Study. Gastroenterology. 1976;70:439–44.
53. Bhattacharya A, Rao BB, Koutroubakis IE, Click B, Vargas EJ, Regueiro M, et al. Silent crohn's disease predicts increased bowel damage during multiyear follow-

up: The consequences of under-reporting active inflammation. Inflammatory Bowel Diseases. 2016;22:2665–71.
54. Peyrin-Biroulet L, Loftus EVJ, Colombel J-F, Sandborn WJ. The natural history of adult crohn's disease in population-based cohorts. Official journal of the American College of Gastroenterology | ACG. 2010;105:289297.
55. Etchevers MJ, Aceituno M, García-Bosch O, Ordás I, Sans M, Ricart E, et al. Risk factors and characteristics of extent progression in ulcerative colitis. Inflammatory Bowel Diseases. 2009;15:1320–5.
56. Boonstra K, van Erpecum KJ, van Nieuwkerk KMJ, Drenth JPH, Poen AC, Witteman BJM, et al. Primary sclerosing cholangitis is associated with a distinct phenotype of inflammatory bowel disease. Inflammatory Bowel Diseases. 2012;18:2270–6.
57. Mark-Christensen A, Laurberg S, Haboubi N. Dysplasia in Inflammatory Bowel Disease: Historical Review, Critical Histopathological Analysis, and Clinical Implications. Inflammatory Bowel Diseases. 2018;24:1895–903.
58. Schieffer KM, Williams ED, Yochum GS, Koltun WA. Review article: the pathogenesis of pouchitis. Alimentary Pharmacology & Therapeutics. 2016;44:817–35.
59. Schroeder KW, Tremaine WJ, Ilstrup DM. Coated Oral 5-Aminosalicylic Acid Therapy for Mildly to Moderately Active Ulcerative Colitis. New England Journal of Medicine. 1987;317:1625–9.
60. Irvine EJ. Development and Subsequent Refinement of the Inflammatory Bowel Disease Questionnaire: A Quality-of-Life Instrument for Adult Patients with Inflammatory Bowel Disease. Journal of Pediatric Gastroenterology & Nutrition. 1999;28 Supplement:S23–7.
61. Travis SPL, Schnell D, Krzeski P, Abreu MT, Altman DG, Colombel J-F, et al. Developing an instrument to assess the endoscopic severity of ulcerative colitis: The ulcerative colitis endoscopic index of severity (UCEIS). Gut. 2012;61:535–42.
62. Travis SPL, Stange EF, Lémann M, Öresland T, Chowers Y, Forbes A, et al. European evidence based consensus on the diagnosis and management of Crohn's disease: current management. Gut. 2006;55 suppl 1:i16–35.
63. Akobeng AK, Zhang D, Gordon M, MacDonald JK. Oral 5-aminosalicylic acid for maintenance of medically-induced remission in Crohn's disease. Cochrane Database of Systematic Reviews. 2016. https://doi.org/10.1002/14651858.CD003715.pub3.
64. Feller M, Huwiler K, Schoepfer A, Shang A, Furrer H, Egger M. Long-term antibiotic treatment for crohn's disease: Systematic review and meta-analysis of placebo-controlled trials. Clinical Infectious Diseases. 2010;50:473–80.
65. Prantera C, Scribano ML. Antibiotics and probiotics in inflammatory bowel disease: Why, when, and how. Current Opinion in Gastroenterology. 2009;25:329333.
66. Rezaie A, Kuenzig ME, Benchimol EI, Griffiths AM, Otley AR, Steinhart AH, et al. Budesonide for induction of remission in Crohn's disease. Cochrane Database of Systematic Reviews. 2015. https://doi.org/10.1002/14651858.CD000296.pub4.
67. Lichtenstein GR, Hanauer SB, Sandborn WJ, Gastroenterology TPPC of the AC of. Management of crohn's disease in adults. Official journal of the American College of Gastroenterology | ACG. 2009;104:465483.
68. Ouellette AJ, Bevins CL. Paneth cell defensins and innate immunity of the small bowel. Inflammatory Bowel Diseases. 2001;7:43–50.
69. Warner B, Johnston E, Arenas-Hernandez M, Marinaki A, Irving P, Sanderson J. A practical guide to thiopurine prescribing and monitoring in IBD. Frontline Gastroenterology. 2018;9:10–5.

70. Chande N, Patton PH, Tsoulis DJ, Thomas BS, MacDonald JK. Azathioprine or 6-mercaptopurine for maintenance of remission in Crohn's disease. Cochrane Database of Systematic Reviews. 2015. https://doi.org/10.1002/14651858.CD000067.pub3.
71. Gisbert JP, Linares PM, Mcnicholl AG, Maté J, Gomollón F. Meta-analysis: the efficacy of azathioprine and mercaptopurine in ulcerative colitis. Alimentary Pharmacology & Therapeutics. 2009;30:126–37.
72. Peyrin-Biroulet L, Lémann M. Review article: remission rates achievable by current therapies for inflammatory bowel disease. Alimentary Pharmacology & Therapeutics. 2011;33:870–9.
73. Billioud V, Sandborn WJ, Peyrin-Biroulet L. Loss of response and need for adalimumab dose intensification in crohn's disease: A systematic review. Official journal of the American College of Gastroenterology | ACG. 2011;106:674684.
74. Hwang JM, Varma MG. Surgery for inflammatory bowel disease. World Journal of Gastroenterology : WJG. 2008;14:2678–90.
75. Gardiner KR, Dasari BVM. Operative Management of Small Bowel Crohn's Disease. Surgical Clinics of North America. 2007;87:587–610.
76. Lewis RT, Maron DJ. Efficacy and complications of surgery for crohn's disease. Gastroenterology & Hepatology. 2010;6:587–96.
77. Corraliza AM, Ricart E, López-García A, Carme Masamunt M, Veny M, Esteller M, et al. Differences in Peripheral and Tissue Immune Cell Populations Following Haematopoietic Stem Cell Transplantation in Crohn's Disease Patients. Journal of Crohn's and Colitis. https://doi.org/10.1093/ecco-jcc/jjy203.
78. Weingarden AR, Vaughn BP. Intestinal microbiota, fecal microbiota transplantation, and inflammatory bowel disease. Gut Microbes. 2017;8:238–52.
79. Beck LC, Granger CL, Masi AC, Stewart CJ. Use of omic technologies in early life gastrointestinal health and disease: From bench to bedside. Expert Review of Proteomics. 2021;18:247–59.
80. Häsler R, Sheibani-Tezerji R, Sinha A, Barann M, Rehman A, Esser D, et al. Uncoupling of mucosal gene regulation, mRNA splicing and adherent microbiota signatures in inflammatory bowel disease. Gut. 2017;66:2087–97.
81. Tang MS, Bowcutt R, Leung JM, Wolff MJ, Gundra UM, Hudesman D, et al. Integrated Analysis of Biopsies from Inflammatory Bowel Disease Patients Identifies SAA1 as a Link Between Mucosal Microbes with TH17 and TH22 Cells. Inflammatory Bowel Diseases. 2017;23:1544–54.
82. Hernández-Rocha C, Borowski K, Turpin W, Filice M, Nayeri S, Raygoza Garay JA, et al. Integrative analysis of colonic biopsies from inflammatory bowel disease patients identifies an interaction between microbial bile-acid inducible gene abundance and human angiopoietin-like 4 gene expression. Journal of Crohn's and Colitis. 2021. https://doi.org/10.1093/ecco-jcc/jjab096.
83. Hu S, Vila AV, Gacesa R, Collij V, Stevens C, Fu JM, et al. Whole exome sequencing analyses reveal gene–microbiota interactions in the context of IBD. Gut. 2021;70:285–96.
84. Mayorgas A, Dotti I, Salas A. Microbial Metabolites, Postbiotics, and Intestinal Epithelial Function. Molecular Nutrition & Food Research. 2021;65:2000188.
85. Planell N, Lozano JJ, Mora-Buch R, Masamunt MC, Jimeno M, Ordás I, et al. Transcriptional analysis of the intestinal mucosa of patients with ulcerative colitis in remission reveals lasting epithelial cell alterations. Gut. 2013;62:967–76.
86. Leal RF, Planell N, Kajekar R, Lozano JJ, Ordás I, Dotti I, et al. Identification of

inflammatory mediators in patients with Crohn's disease unresponsive to anti-TNF? therapy. Gut. 2015;64:233–42.
87. Massimino L, Lamparelli LA, Houshyar Y, D'Alessio S, Peyrin-Biroulet L, Vetrano S, et al. The Inflammatory Bowel Disease Transcriptome and Metatranscriptome Meta-Analysis (IBD TaMMA) framework. Nature Computational Science. 2021;1:511–5.
88. Knights D, Lassen KG, Xavier RJ. Advances in inflammatory bowel disease pathogenesis: Linking host genetics and the microbiome. Gut. 2013;62.
89. Repnik K, Potočnik U. eQTL analysis links inflammatory bowel disease associated 1q21 locus to ECM1 gene. Journal of Applied Genetics. 2016;57:363–72.
90. Hu S, Uniken Venema WT, Westra H-J, Vich Vila A, Barbieri R, Voskuil MD, et al. Inflammation status modulates the effect of host genetic variation on intestinal gene expression in inflammatory bowel disease. Nature Communications. 2021;12:1122.
91. Jung S, Liu W, Baek J, Moon JW, Ye BD, Lee H-S, et al. Expression quantitative trait loci (eQTL) mapping in korean patients with crohn's disease and identification of potential causal genes through integration with disease associations. Frontiers in Genetics. 2020;11.
92. Dai Y, Pei G, Zhao Z, Jia P. A convergent study of genetic variants associated with crohn's disease: Evidence from GWAS, gene expression, methylation, eQTL and TWAS. Frontiers in Genetics. 2019;10.
93. Ahmed I, Roy BC, Khan SA, Septer S, Umar S. Microbiome, metabolome and inflammatory bowel disease. Microorganisms. 2016;4:20.
94. Gallagher K, Catesson A, Griffin JL, Holmes E, Williams HRT. Metabolomic Analysis in Inflammatory Bowel Disease: A Systematic Review. Journal of Crohn's & Colitis. 2021;15:813–26.
95. Krassowski M, Das V, Sahu SK, Misra BB. State of the field in multi-omics research: From computational needs to data mining and sharing. Frontiers in Genetics. 2020;11:1598.
96. Yannakoudakis H, Cummins R. Evaluating the performance of automated text scoring systems. Denver, Colorado: Association for Computational Linguistics; 2015. p. 213223.
97. HOTELLING H. RELATIONS BETWEEN TWO SETS OF VARIATES. Biometrika. 1936;28:321–77.
98. Biancolillo A. Method development in the area of multi-block analysis focused on food analysis. PhD thesis.
99. Wu C, Zhou F, Ren J, Li X, Jiang Y, Ma S. A Selective Review of Multi-Level Omics Data Integration Using Variable Selection. High-Throughput. 2019;8:4.
100. Cavill R, Jennen D, Kleinjans J, Briedé JJ. Transcriptomic and metabolomic data integration. Briefings in Bioinformatics. 2016;17:891–901.
101. Chong J, Xia J. Computational approaches for integrative analysis of the metabolome and microbiome. Metabolites. 2017;7:62.
102. Huang S, Chaudhary K, Garmire LX. More is better: Recent progress in multi-omics data integration methods. Frontiers in Genetics. 2017;8.
103. Rohart F, Mason EA, Matigian N, Mosbergen R, Korn O, Chen T, et al. A molecular classification of human mesenchymal stromal cells. PeerJ. 2016;4:e1845.
104. Ibrahim EC, Guillemot V, Comte M, Tenenhaus A, Zendjidjian XY, Cancel A, et al. Modeling a linkage between blood transcriptional expression and activity in

brain regions to infer the phenotype of schizophrenia patients. npj Schizophrenia. 2017;3:25.
105. Wheeler HE, Aquino-Michaels K, Gamazon ER, Trubetskoy VV, Dolan ME, Huang RS, et al. Poly-Omic Prediction of Complex Traits: OmicKriging. Genetic Epidemiology. 2014;38:402–15.
106. Yin L, Chau CKL, Sham P-C, So H-C. Integrating Clinical Data and Imputed Transcriptome from GWAS to Uncover Complex Disease Subtypes: Applications in Psychiatry and Cardiology. The American Journal of Human Genetics. 2019;105:1193–212.
107. Tarazona S, Arzalluz-Luque A, Conesa A. Undisclosed, unmet and neglected challenges in multi-omics studies. Nature Computational Science. 2021;1–8.
108. Tarazona S, Balzano-Nogueira L, Gómez-Cabrero D, Schmidt A, Imhof A, Hankemeier T, et al. Harmonization of quality metrics and power calculation in multi-omic studies. Nature Communications. 2020;11:3092.
109. Massoni-Badosa R, Iacono G, Moutinho C, Kulis M, Palau N, Marchese D, et al. Sampling time-dependent artifacts in single-cell genomics studies. Genome Biology. 2020;21:112.
110. Zhu Y, Wang L, Yin Y, Yang E. Systematic analysis of gene expression patterns associated with postmortem interval in human tissues. Scientific Reports. 2017;7:5435.
111. Ferreira PG, Muñoz-Aguirre M, Reverter F, Sá Godinho CP, Sousa A, Amadoz A, et al. The effects of death and post-mortem cold ischemia on human tissue transcriptomes. Nature Communications. 2018;9:490.
112. Jacob F, Monod J. Genetic regulatory mechanisms in the synthesis of proteins. Journal of Molecular Biology. 1961;3:318–56.
113. Koh HWL, Fermin D, Vogel C, Choi KP, Ewing RM, Choi H. iOmicsPASS: Network-based integration of multiomics data for predictive subnetwork discovery. npj Systems Biology and Applications. 2019;5:1–0.
114. Yule GU. On the theory of correlation for any number of variables, treated by a new system of notation. Proceedings of the Royal Society of London Series A, Containing Papers of a Mathematical and Physical Character. 1907;79:182–93.
115. Tenenhaus A, Tenenhaus M. Regularized Generalized Canonical Correlation Analysis. Psychometrika. 2011;76:257–84.
116. Tenenhaus A, Philippe C, Guillemot V, Le Cao K-A, Grill J, Frouin V. Variable selection for generalized canonical correlation analysis. Biostatistics. 2014;15:569–83.
117. Culhane AC, Perri‘ere G, Higgins DG. Cross-platform comparison and visualisation of gene expression data using co-inertia analysis. BMC Bioinformatics. 2003;4:59.
118. Vito RD, Bellio R, Trippa L, Parmigiani G. Multi-study factor analysis. Biometrics. 2019;75:337–46.
119. Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T, Marioni JC, et al. Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. Molecular Systems Biology. 2018;14:e8124.
120. Gomez-Cabrero D, Tarazona S, Ferreirós-Vidal I, Ramirez RN, Company C, Schmidt A, et al. STATegra, a comprehensive multi-omics dataset of B-cell differentiation in mouse. Scientific Data. 2019;6:1–5.
121. Franzosa EA, McIver LJ, Rahnavard G, Thompson LR, Schirmer M, Weingart G, et al. Species-level functional profiling of metagenomes and metatranscriptomes. Nature Methods. 2018;15:962.

122. Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nature Methods. 2015;12:902–3.
123. Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, et al. Metagenomic biomarker discovery and explanation. Genome Biology. 2011;12:R60.
124. Didier G, Valdeolivas A, Baudot A. Identifying communities from multiplex biological networks by randomized optimization of modularity. F1000Research. 2018;7.
125. Valdeolivas A, Tichit L, Navarro C, Perrin S, Odelin G, Levy N, et al. Random walk with restart on multiplex and heterogeneous biological networks. Bioinformatics (Oxford, England). 2019;35:497–505.
126. Pio-Lopez L, Valdeolivas A, Tichit L, Remy É, Baudot A. MultiVERSE: A multiplex and multiplex-heterogeneous network embedding approach. arXiv:200810085 [cs, q-bio]. 2021.
127. Bayes T, Price null. An essay towards solving a problem in the doctrine of chances. By the late rev. Mr. Bayes, f. R. s. Philosophical Transactions of the Royal Society of London. 1763;53:370–418.
128. Zhu J, Sova P, Xu Q, Dombek KM, Xu EY, Vu H, et al. Stitching together Multiple Data Dimensions Reveals Interacting Metabolomic and Transcriptomic Networks That Modulate Cell Regulation. PLOS Biology. 2012;10:e1001301.
129. Lock EF, Dunson DB. Bayesian consensus clustering. Bioinformatics. 2013;29:2610–6.
130. Cantini L, Zakeri P, Hernandez C, Naldi A, Thieffry D, Remy E, et al. Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer. Nature Communications. 2021;12:124.
131. Tenenhaus M, Tenenhaus A, Groenen PJF. Regularized Generalized Canonical Correlation Analysis: A Framework for Sequential Multiblock Component Methods. Psychometrika. 2017;82:737–77.
132. Rohart F, Gautier B, Singh A, Cao K-AL. mixOmics: An R package for 'omics feature selection and multiple data integration. PLOS Computational Biology. 2017;13:e1005752.
133. Virtanen S, Klami A, Khan S, Kaski S. Bayesian Group Factor Analysis. PMLR; 2012. p. 1269–77.
134. Novoa-del-Toro E-M, Mezura-Montes E, Vignes M, Magdinier F, Tichit L, Baudot A. A Multi-Objective Genetic Algorithm to Find Active Modules in Multiplex Biological Networks. bioRxiv. 2020;2020.05.25.114215.
135. Lock EF, Hoadley KA, Marron JS, Nobel AB. JOINT AND INDIVIDUAL VARIATION EXPLAINED (JIVE) FOR INTEGRATED ANALYSIS OF MULTIPLE DATA TYPES. The annals of applied statistics. 2013;7:523–42.
136. Yang Z, Michailidis G. A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data. Bioinformatics. 2016;32:1–8.
137. Sherry A, Henson RK. Conducting and Interpreting Canonical Correlation Analysis in Personality Research: A User-Friendly Primer. Journal of Personality Assessment. 2005;84:37–48.
138. Sherry A, Henson RK. Conducting and Interpreting Canonical Correlation Analysis in Personality Research: A User-Friendly Primer. 1981.
139. Chung R-H, Kang C-Y. A multi-omics data simulator for complex disease studies and its application to evaluate multi-omics data analysis methods for disease classification. GigaScience. 2019;8.

140. Martínez-Mira C, Conesa A, Tarazona S. MOSim: Multi-Omics Simulation in R. 2018.
141. Patuzzi I, Baruzzo G, Losasso C, Ricci A, Di Camillo B. metaSPARSim: A 16S rRNA gene sequencing count data simulator. BMC Bioinformatics. 2019;20:416.
142. Fritz A, Hofmann P, Majda S, Dahms E, Dröge J, Fiedler J, et al. CAMISIM: Simulating metagenomes and microbial communities. Microbiome. 2019;7:17.
143. Fu J, Frazee AC, Collado-Torres L, Jaffe AE, Leek JT. Ballgown: Flexible, isoform-level differential expression analysis. Bioconductor version: Release (3.13); 2021.
144. Frazee AC, Jaffe AE, Kirchner R, Leek JT. Polyester: Simulate RNA-seq reads. Bioconductor version: Release (3.13); 2021.
145. McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Research. 2012;40:4288–97.
146. De Souza HSP, Fiocchi C, Iliopoulos D. The IBD interactome: An integrated view of aetiology, pathogenesis and therapy. 2017;14.
147. Valles-Colomer M, Darzi Y, Vieira-Silva S, Falony G, Raes J, Joossens M. Meta-omics in inflammatory bowel disease research: Applications, challenges, and guidelines. Journal of Crohn's and Colitis. 2016;10:735–46.
148. Sudhakar P, Alsoud D, Wellens J, Verstockt S, Arnauts K, Verstockt B, et al. Tailoring multi-omics to inflammatory bowel diseases: All for one and one for all. Journal of Crohn's and Colitis. 2022;jjac027.
149. Puget S, Philippe C, Bax DA, Job B, Varlet P, Junier M-P, et al. Mesenchymal Transition and PDGFRA Amplification/Mutation Are Key Distinct Oncogenic Events in Pediatric Diffuse Intrinsic Pontine Gliomas. PLOS ONE. 2012;7:e30313.
150. Morgan XC, Kabakchiev B, Waldron L, Tyler AD, Tickle TL, Milgrom R, et al. Associations between host gene expression, the mucosal microbiome, and clinical outcome in the pelvic pouch of patients with inflammatory bowel disease. Genome Biology. 2015;16:67.
151. Howell KJ, Kraiczy J, Nayak KM, Gasparetto M, Ross A, Lee C, et al. DNA Methylation and Transcription Patterns in Intestinal Epithelial Cells From Pediatric Patients With Inflammatory Bowel Diseases Differentiate Disease Subtypes and Associate With Outcome. Gastroenterology. 2018;154:585–98.
152. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnetjournal. 2011;17:10–2.
153. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics (Oxford, England). 2013;29:15–21.
154. Li B, Dewey CN. RSEM: Accurate transcript quantification from RNA-seq data with or without a reference genome. BMC Bioinformatics. 2011;12:323.
155. Corraliza AM, Ricart E, L'opez-García A, Carme Masamunt M, Veny M, Esteller M, et al. Differences in Peripheral and Tissue Immune Cell Populations Following Haematopoietic Stem Cell Transplantation in Crohn's Disease Patients. Journal of Crohn's and Colitis. https://doi.org/10.1093/ecco-jcc/jjy203.
156. Berry D, Ben Mahfoudh K, Wagner M, Loy A. Barcoded primers used in multiplex amplicon pyrosequencing bias amplification. Applied and Environmental Microbiology. 2011;77:7846–9.
157. Klindworth A, Pruesse E, Schweer T, Peplies J, Quast C, Horn M, et al. Evalua-

tion of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. Nucleic Acids Research. 2013;41:e1–1.
158. Edgar RC. UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nature Methods. 2013;10:996–8.
159. Lagkouvardos I, Joseph D, Kapfhammer M, Giritli S, Horn M, Haller D, et al. IMNGS: A comprehensive open resource of processed 16S rRNA microbial profiles for ecology and diversity studies. Scientific Reports. 2016;6.
160. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: High-resolution sample inference from Illumina amplicon data. Nature Methods. 2016;13:581–3.
161. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. Nucleic Acids Research. 2013;41 Database issue:D590–6.
162. Jordan C. Essai sur la géométrie à n dimensions. Bulletin de la Société Mathématique de France. 1875;3:103–74.
163. Tenenhaus M. Component-based Structural Equation Modelling. Total Quality Management & Business Excellence. 2008;19:871–86.
164. Tenenhaus M, Hanafi M. A Bridge Between PLS Path Modeling and Multi-Block Data Analysis. In: Esposito Vinzi V, Chin WW, Henseler J, Wang H, editors. Handbook of Partial Least Squares. Berlin, Heidelberg: Springer Berlin Heidelberg; 2010. p. 99–123.
165. Tenenhaus A, Tenenhaus M. Regularized generalized canonical correlation analysis for multiblock or multigroup data analysis. European Journal of Operational Research. 2014;238:391–403.
166. Tenenhaus A, Philippe C, Frouin V. Kernel Generalized Canonical Correlation Analysis. Computational Statistics & Data Analysis. 2015;90:114–31.
167. Gloaguen A, Philippe C, Frouin V, Gennari G, Dehaene-Lambertz G, Le Brusquet L, et al. Multiway generalized canonical correlation analysis. Biostatistics. 2020. https://doi.org/10.1093/biostatistics/kxaa010.
168. Schäfer J, Strimmer K. A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics. Statistical Applications in Genetics and Molecular Biology. 2005;4.
169. Horst P. Relations among m sets of measures. Psychometrika. 1961;26:129–49.
170. KETTENRING JR. Canonical analysis of several sets of variables. Biometrika. 1971;58:433–51.
171. Van de Geer JP. Linear relations among k sets of variables. Psychometrika. 1984;49:79–94.
172. Meng C, Kuster B, Culhane AC, Gholami AM. A multivariate approach to the integration of multi-omics datasets. BMC Bioinformatics. 2014;15:162.
173. Planell N, Lagani V, Sebastian-Leon P, van der Kloet F, Ewing E, Karathanasis N, et al. STATegra: Multi-omics data integration – a conceptual scheme with a bioinformatics pipeline. Frontiers in Genetics. 2021;12.
174. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. The Innovation. 2021;2:100141.
175. Richter FC, Friedrich M, Pohin M, Alsaleh G, Guschina I, Wideman SK, et al. Cell-extrinsic autophagy in mature adipocytes regulates anti-inflammatory response to intestinal tissue injury through lipid mobilization. 2021;2021.10.25.465200.
176. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA,

et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences. 2005;102:15545–50.
177. Protiva P, Pendyala S, Nelson C, Augenlicht LH, Lipkin M, Holt PR. Calcium and 1,25-dihydroxyvitamin D3 modulate genes of immune and inflammatory pathways in the human colon: A human crossover trial. The American Journal of Clinical Nutrition. 2016;103:1224–31.
178. Korotkevich G, Sukhov V, Budin N, Shpak B, Artyomov MN, Sergushichev A. Fast gene set enrichment analysis. 2021. https://doi.org/https://doi.org/10.1101/060012.
179. Fabregat A, Sidiropoulos K, Garapati P, Gillespie M, Hausmann K, Haw R, et al. The reactome pathway knowledgebase. Nucleic Acids Research. 2016;44.
180. Hänzelmann S, Castelo R, Guinney J. GSVA: Gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics. 2013;14:7.
181. Escudero-Hernández C, Beelen Granlund A van, Bruland T, Sandvik AK, Koch S, Østvik AE, et al. Transcriptomic Profiling of Collagenous Colitis Identifies Hallmarks of Nondestructive Inflammatory Bowel Disease. Cellular and Molecular Gastroenterology and Hepatology. 2021;12:665–87.
182. Anderson MJ. A new method for non-parametric multivariate analysis of variance. Austral Ecology. 2001;26:32–46.
183. Warton DI, Wright ST, Wang Y. Distance-based multivariate analyses confound location and dispersion effects. Methods in Ecology and Evolution. 2012;3:89–101.
184. Goeman JJ, van de Geer SA, van Houwelingen HC. Testing against a high dimensional alternative. Journal of the Royal Statistical Society Series B (Statistical Methodology). 2006;68:477–93.
185. Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, et al. Vegan: Community ecology package. 2020.
186. Ritchie MEM, Phipson B, Wu D, Hu Y, Law CWC, Shi W, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic acids research. 2015;43:e47.
187. Law CW, Chen Y, Shi W, Smyth GK. Voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biology. 2014;15:R29.
188. Law CW, Alhamdoosh M, Su S, Dong X, Tian L, Smyth GK, et al. RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR. F1000Research. 2018;5.
189. Yoav Benjamini, Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B (Methodological). 57.
190. Langfelder P, Horvath S. WGCNA: An r package for weighted correlation network analysis. BMC Bioinformatics. 2008;9:559.
191. Filzmoser P, Viertl R. Testing hypotheses with fuzzy data: The fuzzy p -value. Metrika. 2004;59:21–9.
192. Dubois D, Prade H. [Proceedings 1993] second IEEE international conference on fuzzy systems. 1993. p. 1059–1068 vol.2.
193. Revilla Sancho L, Lozano J-J, Salas A. experDesign: stratifying samples into batches with minimal bias. Journal of Open Source Software. 2021;6:3358.
194. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, et al. pROC: An open-source package for r and s+ to analyze and compare ROC curves. BMC Bioinformatics. 2011;12:77.

195. Wickham H. Advanced r. Second Edition. Boca Raton: Chapman & Hall; 2019.
196. Greenland S, Brumback B. An overview of relations among causal modelling methods. International Journal of Epidemiology. 2002;31:1030–7.
197. Revilla L, Mayorgas A, Corraliza AM, Masamunt MC, Metwaly A, Haller D, et al. Multi-omic modelling of inflammatory bowel disease with regularized canonical correlation analysis. PLOS ONE. 2021;16:e0246367.
198. Callahan BJ, McMurdie PJ, Holmes SP. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. The ISME Journal. 2017;11:2639–43.
199. Nearing JT, Douglas GM, Hayes MG, MacDonald J, Desai DK, Allward N, et al. Microbiome differential abundance methods produce different results across 38 datasets. Nature Communications. 2022;13:342.
200. Louca S, Doebeli M, Parfrey LW. Correcting for 16S rRNA gene copy numbers in microbiome surveys remains an unsolved problem. Microbiome. 2018;6:41.
201. Nadalian B, Yadegar A, Houri H, Olfatifar M, Shahrokh S, Asadzadeh Aghdaei H, et al. Prevalence of the pathobiont adherent-invasive Escherichia coli and inflammatory bowel disease: a systematic review and meta-analysis. Journal of Gastroenterology and Hepatology. 2021;36:852–63.
202. Chalise P, Raghavan R, Fridley BL. InterSIM: Simulation tool for multiple integrative 'omic datasets'. Computer methods and programs in biomedicine. 2016;128:69–74.
203. Chung R-H, Kang C-Y. A multi-omics data simulator for complex disease studies and its application to evaluate multi-omics data analysis methods for disease classification. GigaScience. 2019;8.
204. Ugidos M, Tarazona S, Prats-Montalbán JM, Ferrer A, Conesa A. MultiBaC: A strategy to remove batch effects between different omic data types: Statistical Methods in Medical Research. 2020. https://doi.org/10.1177/0962280220907365.
205. Bersanelli M, Mosca E, Remondini D, Giampieri E, Sala C, Castellani G, et al. Methods for the integration of multi-omics data: Mathematical aspects. BMC Bioinformatics. 2016;17:S15.
206. Holmes E, Li JV, Marchesi JR, Nicholson JK. Gut Microbiota Composition and Activity in Relation to Host Metabolic Phenotype and Disease Risk. Cell Metabolism. 2012;16:559–64.
207. Stappenbeck TS, Hooper LV, Gordon JI. Developmental regulation of intestinal angiogenesis by indigenous microbes via Paneth cells. Proceedings of the National Academy of Sciences of the United States of America. 2002;99:15451–5.
208. Brand EC, Klaassen MAY, Gacesa R, Vich Vila A, Ghosh H, Zoete MR de, et al. Healthy Cotwins Share Gut Microbiome Signatures With Their Inflammatory Bowel Disease Twins and Unrelated Patients. Gastroenterology. 2021;160:1970–85.
209. Susin A, Wang Y, Lê Cao K-A, Calle ML. Variable selection in microbiome compositional data analysis. NAR Genomics and Bioinformatics. 2020;2:lqaa029.
210. Sartor RB. Mechanisms of disease: pathogenesis of Crohn's disease and ulcerative colitis. Nature Clinical Practice Gastroenterology & Hepatology. 2006;3:390–407.
211. Vandeputte D, Kathagen G, D'hoe K, Vieira-Silva S, Valles-Colomer M, Sabino J, et al. Quantitative microbiome profiling links gut community variation to microbial load. Nature. 2017;551:507–11.
212. Vila-Casadesús M, Gironella M, Lozano JJ. MiRComb: An R Package to Analyse miRNA-mRNA Interactions. Examples across Five Digestive Cancers. PloS One.

2016;11:e0151127.
213. Goldberg D. What every computer scientist should know about floating-point arithmetic. ACM Computing Surveys. 1991;23:548.
214. Yan L, Ma C, Wang D, Hu Q, Qin M, Conroy JM, et al. OSAT: A tool for sample-to-batch allocations in genomics experiments. BMC Genomics. 2012;13:689.
215. Papenberg M, Klau GW. Using anticlustering to partition data sets into equivalent parts. Psychological Methods. 2020. https://doi.org/10.1037/met0000301.
216. Sinke L, Cats D, Heijmans BT. Omixer: Multivariate and reproducible sample randomization to proactively counter batch effects in omics studies. Bioinformatics. 2021. https://doi.org/10.1093/bioinformatics/btab159.
217. Haberman Y, Tickle TL, Dexheimer PJ, Kim M-O, Tang D, Karns R, et al. Pediatric Crohn disease patients exhibit specific ileal transcriptome and microbiome signature. The Journal of Clinical Investigation. 2014;124:3617–33.
218. Community TTW. The turing way: A handbook for reproducible, ethical and collaborative research. The Turing Way Community; 2021.
219. Meng C, Kuster B, Culhane AC, Gholami AM. A multivariate approach to the integration of multi-omics datasets. BMC Bioinformatics. 2014;15:162.
220. Pearl J. Transportability across studies: A formal approach. 2011.
221. Simpson EH. The Interpretation of Interaction in Contingency Tables. Journal of the Royal Statistical Society: Series B (Methodological). 1951;13:238–41.
222. Stillman DJ. Interactions of a Eukaryotic RNA Polymerase III Transcription Factor with Promoters. Biology; 1984.
223. Lee RC, Feinbaum RL, Ambros V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell. 1993;75:843–54.
224. Hamilton AJ, Baulcombe DC. A species of small antisense RNA in posttranscriptional gene silencing in plants. Science. 1999;286:950–2.
225. Thornburg ZR, Bianchi DM, Brier TA, Gilbert BR, Earnest TM, Melo MCR, et al. Fundamental behaviors emerge from simulations of a living minimal cell. Cell. 2022;185:345–360.e28.
226. Criss ZK, Bhasin N, Di Rienzi SC, Rajan A, Deans-Fielder K, Swaminathan G, et al. Drivers of transcriptional variance in human intestinal epithelial organoids. Physiological Genomics. 2021;53:486–508.
227. Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, et al. Enterotypes of the human gut microbiome. Nature. 2011;473:174–80.
228. Koren O, Knights D, Gonzalez A, Waldron L, Segata N, Knight R, et al. A Guide to Enterotypes across the Human Body: Meta-Analysis of Microbial Community Structures in Human Microbiome Datasets. PLOS Computational Biology. 2013;9:e1002863.
229. Cheng M, Ning K. Stereotypes About Enterotype: the Old and New Ideas. Genomics, Proteomics & Bioinformatics. 2019;17:4–12.
230. Hasan N, Yang H. Factors affecting the composition of the gut microbiota, and its modulation. PeerJ. 2019;7:e7502.
231. Bringiotti R, Ierardi E, Lovero R, Losurdo G, Leo AD, Principi M. Intestinal microbiota: The explosive mixture at the origin of inflammatory bowel disease? World Journal of Gastrointestinal Pathophysiology. 2014;5:550–9.
232. Lloyd-Price J, Arze C, Ananthakrishnan AN, Schirmer M, Avila-Pacheco J, Poon TW, et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature. 2019;569:655.

233. Douglas GM, Hansen R, Jones CMA, Dunn KA, Comeau AM, Bielawski JP, et al. Multi-omics differentially classify disease state and treatment outcome in pediatric crohn's disease. Microbiome. 2018;6:13.
234. Halfvarson J, Brislawn CJ, Lamendella R, Vázquez-Baeza Y, Walters WA, Bramer LM, et al. Dynamics of the human gut microbiome in inflammatory bowel disease. Nature Microbiology. 2017;2:1–7.
235. Edgar R. Taxonomy annotation and guide tree errors in 16S rRNA databases. PeerJ. 2018;6:e5030.
236. Ungar B, Yavzori M, Fudim E, Picard O, Kopylov U, Eliakim R, et al. Host transcriptome signatures in human faecal-washes predict histological remission in patients with IBD. Gut. 2022. https://doi.org/10.1136/gutjnl-2021-325516.
237. Zhang X, Li L, Butcher J, Stintzi A, Figeys D. Advancing functional and translational microbiome research using meta-omics approaches. Microbiome. 2019;7:154.
238. Jiang D, Armour CR, Hu C, Mei M, Tian C, Sharpton TJ, et al. Microbiome multi-omics network analysis: Statistical considerations, limitations, and opportunities. Frontiers in Genetics. 2019;10.
239. Haynes WA, Tomczak A, Khatri P. Gene annotation bias impedes biomedical research. Scientific Reports. 2018;8:1362.
240. Collij V, Klaassen MAY, Weersma RK, Vila AV. Gut microbiota in inflammatory bowel diseases: moving from basic science to clinical applications. Human Genetics. 2021;140:703–8.
241. Parkhomenko E, Tritchler D, Beyene J. Sparse Canonical Correlation Analysis with Application to Genomic Data Integration. Statistical Applications in Genetics and Molecular Biology. 2009;8.
242. Waaijenborg S, Hamer PCV de W, Zwinderman AH. Quantifying the Association between Gene Expressions and DNA-Markers by Penalized Canonical Correlation Analysis. Statistical Applications in Genetics and Molecular Biology. 2008;7.
243. Witten DM, Tibshirani RJ. Extensions of sparse canonical correlation analysis with applications to genomic data. Statistical Applications in Genetics and Molecular Biology. 2009;8:127.
244. Lê Cao K-A, Martin PG, Robert-Granié C, Besse P. Sparse canonical methods for biological data integration: Application to a cross-platform study. BMC Bioinformatics. 2009;10:34.
245. Hwang H. Regularized Generalized Structured Component Analysis. Psychometrika. 2009;74:517–30.
246. Soneson C, Lilljebjörn H, Fioretos T, Fontes M. Integrative analysis of gene expression and copy number alterations using canonical correlation analysis. BMC Bioinformatics. 2010;11:191.
247. Zhang S, Li Q, Liu J, Zhou XJ. A novel computational framework for simultaneous integration of multiple types of genomic data to identify microRNA-gene regulatory modules. Bioinformatics. 2011;27:i401–9.
248. Lee W, Lee D, Lee Y, Pawitan Y. Sparse Canonical Covariance Analysis for High-throughput Data. Statistical Applications in Genetics and Molecular Biology. 2011;10.
249. Abdi H, Williams LJ, Valentin D, Bennani-Dosse M. STATIS and DISTATIS: optimum multitable principal component analysis and three way metric multidimensional scaling. WIREs Computational Statistics. 2012;4:124–67.
250. Zhang S, Liu C-C, Li W, Shen H, Laird PW, Zhou XJ. Discovery of multi-

dimensional modules by integrative analysis of cancer genomic data. Nucleic Acids Research. 2012;40:9379–91.
251. Li W, Zhang S, Liu C-C, Zhou XJ. Identifying multi-layer gene regulatory modules from multi-dimensional genomic data. Bioinformatics. 2012;28:2458–66.
252. Abdi H, Williams LJ, Valentin D. Multiple factor analysis: principal component analysis for multitable and multiblock data sets. WIREs Computational Statistics. 2013;5:149–79.
253. Schlauch D, Paulson JN, Young A, Glass K, Quackenbush J. Estimating gene regulatory networks with pandaR. Bioinformatics. 2017;33:2232–4.
254. Ray P, Zheng L, Lucas J, Carin L. Bayesian joint analysis of heterogeneous genomics data. Bioinformatics. 2014;30:1370–6.
255. Bunte K, Leppäaho E, Saarinen I, Kaski S. Sparse group factor analysis for biclustering of multiple data sources. Bioinformatics. 2016;32:2457–63.
256. Chen M, Gao C, Ren Z, Zhou HH. Sparse CCA via precision adjusted iterative thresholding. arXiv:13116186 [math, stat]. 2013.
257. Leppäaho E, Ammad-ud-din M, Kaski S. GFA: Exploratory analysis of multiple data sources with group factor analysis. Journal of Machine Learning Research. 2017;18:1–5.
258. Klami A, Bouchard G, Tripathi A. Group-sparse embeddings in collective matrix factorization. arXiv:13125921 [cs, stat]. 2014.
259. Meng C, Basunia A, Peters B, Gholami AM, Kuster B, Culhane AC. MOGSA: integrative single sample gene-set analysis of multiple omics data. 2018.
260. Zhao S, Gao C, Mukherjee S, Engelhardt BE. Bayesian group latent factor analysis with structured sparsity. arXiv:14112698 [q-bio, stat]. 2015.
261. Voillet V, Besse P, Liaubet L, San Cristobal M, González I. Handling missing rows in multi-omics data integration: Multiple imputation in multiple factor analysis framework. BMC Bioinformatics. 2016;17:402.
262. Beaton D, Dunlop J, Abdi H, Alzheimer's Disease Neuroimaging Initiative. Partial least squares correspondence analysis: A framework to simultaneously analyze behavioral and genetic data. Psychological Methods. 2016;21:621–51.
263. Singh A, Shannon CP, Gautier B, Rohart F, Vacher M, Tebbutt SJ, et al. DIABLO: An integrative approach for identifying key molecular drivers from multi-omics assays. Bioinformatics. 2019;35:3055–62.
264. Yoon G, Carroll RJ, Gaynanova I. Sparse semiparametric canonical correlation analysis for data of mixed types. arXiv:180705274 [stat]. 2019.
265. Gaynanova I, Li G. Structural learning and integrative decomposition of multi-view data. arXiv:170706573 [stat]. 2017.
266. Madrigal P. fCCAC: Functional canonical correlation analysis to evaluate covariance between nucleic acid sequencing datasets. Bioinformatics. 2017;33:746–8.
267. Yoshida K, Yoshimoto J, Doya K. Sparse kernel canonical correlation analysis for discovery of nonlinear interactions in high-dimensional data. BMC Bioinformatics. 2017;18:108.
268. Kawaguchi A, Yamashita F. Supervised multiblock sparse multivariable analysis with application to multimodal brain imaging genetics. Biostatistics. 2017;18:651–65.
269. Feng Q, Jiang M, Hannig J, Marron JS. Angle-based joint and individual variation explained. arXiv:170402060 [stat]. 2018.
270. Argelaguet R, Arnol D, Bredikhin D, Deloro Y, Velten B, Marioni JC, et al. MOFA+: A statistical framework for comprehensive integration of multi-modal single-

cell data. Genome Biology. 2020;21:111.
271. Brown BC, Bray NL, Pachter L. Expression reflects population structure. 2018. https://doi.org/10.1101/364448.
272. Zhang Y, Gaynanova I. Joint association and classification analysis of multi-view data. arXiv:181108511 [cs, stat]. 2020.
273. Tang TM, Allen GI. Integrated principal components analysis. arXiv:181000832 [stat]. 2021.
274. Min EJ, Safo SE, Long Q. Penalized co-inertia analysis with applications to -omics data. Bioinformatics (Oxford, England). 2019;35:1018–25.
275. Safo SE, Li S, Long Q. Integrative analysis of transcriptomic and metabolomic data via sparse canonical correlation analysis with incorporation of biological information. Biometrics. 2018;74:300–12.
276. Min W, Liu J, Zhang S. Sparse weighted canonical correlation analysis. arXiv:171004792 [cs, stat]. 2017.
277. Bouhaddani S el, Uh H-W, Jongbloed G, Hayward C, Klarić L, Kiełbasa SM, et al. Integrating omics datasets with the OmicsPLS package. BMC Bioinformatics. 2018;19:371.
278. Pimentel H, Zhiyue H, Huang H. Biclustering by sparse canonical correlation analysis. 2018;6:11.
279. Kim Y, Bismeijer T, Zwart W, Wessels LFA, Vis DJ. Genomic data integration by WON-PARAFAC identifies interpretable factors for predicting drug-sensitivity in vivo. Nature Communications. 2019;10:5034.
280. Lock EF, Park JY, Hoadley KA. Bidimensional linked matrix factorization for pan-omics pan-cancer analysis. arXiv:200202601 [cs, q-bio, stat]. 2020.
281. Ronen J, Hayat S, Akalin A. Evaluation of colorectal cancer subtypes and cell lines using deep learning. Life Science Alliance. 2019;2.
282. Shi WJ, Zhuang Y, Russell PH, Hobbs BD, Parker MM, Castaldi PJ, et al. Unsupervised discovery of phenotype-specific multi-omics networks. Bioinformatics. 2019;35:4336–43.
283. Csala A, Zwinderman AH, Hof MH. Multiset sparse partial least squares path modeling for high dimensional omics data analysis. BMC Bioinformatics. 2020;21:9.
284. Fan Z, Zhou Y, Ressom HW. MOTA: Network-Based Multi-Omic Data Integration for Biomarker Discovery. Metabolites. 2020;10:144.
285. Shu H, Wang X, Zhu H. D-CCA: A decomposition-based canonical correlation analysis for high-dimensional datasets. Journal of the American Statistical Association. 2020;115:292–306.
286. Hawinkel S, Bijnens L, Cao K-AL, Thas O. Model-based joint visualization of multiple compositional omics datasets. NAR Genomics and Bioinformatics. 2020;2.
287. Gundersen G, Dumitrascu B, Ash JT, Engelhardt BE. Uncertainty in Artificial Intelligence. PMLR; 2020. p. 945–55.
288. Velten B, Braunger JM, Arnol D, Argelaguet R, Stegle O. Identifying temporal and spatial patterns of variation from multi-modal data using MEFISTO. bioRxiv. 2020;2020.11.03.366674.

Online resources

Some links that we found useful on the thesis and could be useful if you are interested on the multi-omics field.

- Awesome multi-omics https://github.com/mikelove/awesome-multi-omics : An online repository of references to multi-omics methods. Reproduced here with their references [1]:

Table A.1: Integration methods available and their references.

Method	Publication
SCCA	[241]
PCCA	[242]
PMA	[243]
sPLS	[244]
gesca	[245]
Regularized dual CCA	[246]
RGCCA	[115]
SNMNMF	[247]
scca	[248]
STATIS	[249]
joint NMF	[250]
sMBPLS	[251]
Bayesian group factor analysis	[133]
RIMBANET	[128]
FactoMineR	[252]
JIVE	[135]
pandaR	[253]
omicade4	[172]
STATegRa	[173]
Joint factor model	[254]
GFAsparse	[255]
Sparse CCA	[256]
CCAGFA	[257]
CMF	[258]
MOGSA	[259]
iNMF	[136]
BASS	[260]
imputeMFA	[261]

[1] Consulted on 2021/11/10 from https://github.com/mikelove/awesome-multi-omics

Method	Publication
PLSCA	[262]
mixOmics	[263]
mixedCCA	[264]
SLIDE	[265]
fCCAC	[266]
TSKCCA	[267]
SMSMA	[268]
AJIVE	[269]
MOFA	[270]
PCA+CCA	[271]
JACA	[272]
iPCA	[273]
pCIA	[274]
sSCCA	[275]
SWCCA	[276]
OmicsPLS	[277]
SCCA-BC	[278]
WON-PARAFAC	[279]
BIDIFAC	[280]
maui	[281]
SmCCNet	[282]
msPLS	[283]
MOTA	[284]
D-CCA	[285]
COMBI	[286]
DPCCA	[287]
MEFISTO	[288]
MultiPower	[108]

- Bookdown: The book about how to write this type of books.
- Bioconductor: The project about bioinformatics on R mostly related to sequencing technologies.
- CRAN: The main archive of R extensions/packages for R.
- GitHub: Company which allows to freely host remote git repositories of many projects, including some used or developed on this thesis.

Software

Along the years of this thesis several pieces of software have been generated as well as packages. Here they are listed for easier retrieval. They are listed on two ways, one with a brief explanation and another one ordered by what software piece is used on each analysis.

B.1 STAR

The parameters and options used with STAR are:

```
STAR \
    --outSAMtype BAM SortedByCoordinate \
    --outFilterIntronMotifs RemoveNoncanonical \
    --outSAMattributes All \
    --outReadsUnmapped Fastx \
    --outSAMstrandField intronMotif \
    --outFilterScoreMinOverLread 0.5 \
    --outFilterMatchNminOverLread 0.5 \
    --outFilterType BySJout \
    --alignSJoverhangMin 8 \
    --alignSJDBoverhangMin 1 \
    --outFilterMismatchNmax 999 \
    --outFilterMismatchNoverLmax 0.04 \
    --genomeDir "$genome/STAR" \
    --limitBAMsortRAM 10000000000 \
    --runMode alignReads \
    --genomeLoad NoSharedMemory \
    --quantMode TranscriptomeSAM \
    --outFileNamePrefix $output \
    --runThreadN "$threads" \
    --readFilesCommand zcat \
    --readFilesIn "$file1" "$file2"
```

The `$genome` is the path to the location on the computer where the genome is, `$output` is the prefix of the output file, `$threads` is the number of threads used and `$file1` and `$file2` are the paired fastq files.

B.2 RSEM

Code used for RSEM where `$threads` is the number of threads used, `$rseminp` is the input file in BAM format, `$genome` is the path to the location on the computer where the genome is, and `$rsem` is the output file.

```
rsem-calculate-expression \
    --quiet \
    --paired-end \
    -p "$threads"  \
    --estimate-rspd \
    --append-names \
    --no-bam-output \
    --bam "$rseminp" "$genome/RSEM/RSEM" "$rsem"
```

B.3 Listed

An improved/tested version of RGCCA, some modifications on the internal functions to ease the maintenance as well as adding tests and sometimes improving the documentation. Also modified so that it is possible to provide a vector of models so that the model of the first dimension is not the same as the model on the second dimension (not sure if mathematically speaking makes sense but from a biological one we think might be interesting to have it).

Designed to be used with RGCCA we wrote the package inteRmodel to ease the bootstrapping and model selection.

A package to design batches to avoid batch effect experDesign and its website on GitHub.

Explore the effects of the hyperparameters on RGCCA on the provided dataset of gliomaData (Originally provided here) there is this repository sgcca_hyperparameters.

We used a pouchitis cohort published in this article[150] that was used to compare how performs our method in other's dataset. The code used can be found in this repository.

Some functions used to explore the TRIM dataset ended up in the integration package.This include functions for correlation, network analysis, enrichment, normalization of metadata...

We developed a package to analyze sets and fuzzy sets BaseSet (based on what we learned from a previous iteration of the package). This package was meant to be used with the probabilities that arise from bootstrapping the models. However, due to the long times of calculation that it would require it was not used.

To analyze the BARCELONA cohort (also named antiTNF) a different repository was created to analyze the data using the previously developed packages.

B.4 By project/publication

All code of the analysis of the publications is available (in his messed state and complicated history) and a brief description as to why they were used:

Multi-omic modelling of inflammatory bowel disease with regularized canonical correlation analysis:

- TRIM: Mangle with the sample, dataset, explore several methods...
- Puget's: Explore the effects of the hyperparameters on RGCCA on the provided dataset.
- inteRmodel: Package for easy repeating the methodology developed on TRIM.
- Morgan's: Work with the pouchitis cohort used in this article.
- Häsler's: Work with the UC/CD dataset used in this article.
- integration: Package with functions that we wrote or used on different parts of exploring the TRIM dataset ended up here.

BaseSet:

- BaseSet: Fuzzy logic implementation, available on rOpenSci too and its documentation website .

experDesign:

- experDesign: Help design experiments in batches. It has a documentation website too.

BARCELONA:

- BARCELONA: Code to analyze the BARCELONA's dataset

Validation:

- Howell's: Work with Howell's 2018 dataset.
- Cristian's: Work with Cristian's 2020 dataset.

Articles

Articles published on peer-review journals about this thesis:

C.1 Multi-omic modelling of inflammatory bowel disease with regularized canonical correlation analysis

Article peer-reviewed published on 2021, freely available online.

The same article can be find below.

PLOS ONE

RESEARCH ARTICLE

Multi-omic modelling of inflammatory bowel disease with regularized canonical correlation analysis

Lluís Revilla[1,2], Aida Mayorgas[2], Ana M. Corraliza[2], Maria C. Masamunt[2], Amira Metwaly[3], Dirk Haller[3,4], Eva Tristán[1,5], Anna Carrasco[1,5], Maria Esteve[1,5], Julian Panés[1,2], Elena Ricart[1,2], Juan J. Lozano[1], Azucena Salas[2]*

1 Centro de Investigación Biomédica en Red de Enfermedades Hepática y Digestivas (CIBERehd), Barcelona, Spain, **2** Department of Gastroenterology, IDIBAPS, Hospital Clínic, Barcelona, Spain, **3** Chair of Nutrition and Immunology, Technical University of Munich, Freising-Weihenstephan, Germany, **4** ZIEL Institute for Food and Health, Technical University of Munich, Freising-Weihenstephan, Germany, **5** Department of Gastroenterology, Hospital Universitari Mútua Terrassa, Barcelona, Spain

* asalas1@clinic.cat

OPEN ACCESS

Citation: Revilla L, Mayorgas A, Corraliza AM, Masamunt MC, Metwaly A, Haller D, et al. (2021) Multi-omic modelling of inflammatory bowel disease with regularized canonical correlation analysis. PLoS ONE 16(2): e0246367. https://doi.org/10.1371/journal.pone.0246367

Editor: Franck Carbonero, Washington State University - Spokane, UNITED STATES

Received: July 23, 2020

Accepted: January 18, 2021

Published: February 8, 2021

Data Availability Statement: Glioma data available from https://biodev.cea.fr/sgcca/ RNA-seq data from CD dataset available at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE139179 Microbiome data from CD dataset available at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE139680 data from pouchitis dataset available at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE65270 Analysis code of our HSCT CD dataset available at: https://github.com/llrs/TRIM repository. Parameter studies performed at https://github.com/llrs/sgcca_hyperparameters.

Abstract

Background

Personalized medicine requires finding relationships between variables that influence a patient's phenotype and predicting an outcome. Sparse generalized canonical correlation analysis identifies relationships between different groups of variables. This method requires establishing a model of the expected interaction between those variables. Describing these interactions is challenging when the relationship is unknown or when there is no pre-established hypothesis. Thus, our aim was to develop a method to find the relationships between microbiome and host transcriptome data and the relevant clinical variables in a complex disease, such as Crohn's disease.

Results

We present here a method to identify interactions based on canonical correlation analysis. We show that the model is the most important factor to identify relationships between blocks using a dataset of Crohn's disease patients with longitudinal sampling. First the analysis was tested in two previously published datasets: a glioma and a Crohn's disease and ulcerative colitis dataset where we describe how to select the optimum parameters. Using such parameters, we analyzed our Crohn's disease data set. We selected the model with the highest inner average variance explained to identify relationships between transcriptome, gut microbiome and clinically relevant variables. Adding the clinically relevant variables improved the average variance explained by the model compared to multiple co-inertia analysis.

Conclusions

The methodology described herein provides a general framework for identifying interactions between sets of omic data and clinically relevant variables. Following this method, we found

CD/UC dataset analysis available at: https://github.com/llrs/Uncoupling Pouchitis dataset analysis available at: https://github.com/llrs/pouchitis Helper package required for the analysis available at https://github.com/llrs/integration-helper. Package with the methodology: Project name: inteRmodel Project home page: https://llrs.github.io/inteRmodel/ Operating system: Platform independent Programming language: R License: MIT.

Funding: This work was supported by the Leona and Harry Helmsley Charitable grant 2015PG-IBD005; Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBERehd,); AiM is supported by grant BES-2016-076642 from the Ministerio de Ciencia, Innovación y Universidades, Spain. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: We confirm that the following competing interests' statement does not alter our adherence to PLOS ONE policies on sharing data and materials: I have read the journal's policy and the authors of this manuscript have the following competing interests: JP has received consulting fees from AbbVie, Arena Pharmaceuticals, Boehringer Ingelheim, Celgene, Celltrion, Ferring, Genentech, Janssen, Merck, Sharp & Dohme (MSD), Oppilan, Pfizer, Progenity, Robarts, Roche, Takeda, Theravance, and TiGenix; speaker's fees from AbbVie, Ferring, Janssen, MSD, Shire Pharmaceuticals, and Takeda; and research funding from AbbVie and MSD. AS reports research grants from Roche, Genentech, Boehringer Ingelheim and AbbVie, lecture fees from Roche, Boehringer Ingelheim and Pfizer, and consultancy fees from Genentech and GSK.

genes and microorganisms that were related to each other independently of the model, while others were specific to the model used. Thus, model selection proved crucial to finding the existing relationships in multi-omics datasets.

Introduction

The creation of datasets from different high-throughput sequencing technologies on the same samples provides an opportunity to identify relationships between datasets and improve our understanding of diseases. This approach has been used in several diseases, such as cancer, inflammatory bowel disease (IBD) and pouchitis, among others [1–3].

IBD is comprised of Crohn's disease (CD) and ulcerative colitis (UC). Around 4.2 million individuals suffer from IBD in Europe and North America combined [4]. The chronic inflammatory response observed suggests an interaction between host genetic factors and the intestinal microbiota. Several studies support the concept that CD arises from an exacerbated immune response against commensal gut microorganisms in genetically predisposed individuals. Nonetheless, the disease might result from imbalanced microbial composition, leading to dysbiosis [5, 6].

Understanding the contribution of the gut microbiota to CD pathogenesis and maintenance of the disease is an ongoing field of research [7–9]. These alterations could be shaped by a genetic predisposition and environmental factors (i.e., bacterial or viral infection, diet, usage of antibiotic, or the socioeconomic status) [10]. On the other hand, pouchitis refers to the inflammation of the ileal pouch, an artificial rectum surgically created out of ileal gut tissue in patients who have undergone a colectomy. One possible underlying cause of pouchitis might also be the an imbalance in the gut microbiome [11]. However, the cause-effect relation between dysbiosis and intestinal inflammatory disease remains unclear [12–14].

The most common method for analyzing the relationship between microorganisms and the gut mucosa is to sequence both the 16S rRNA gene of the microbiome and the patient's transcriptome, respectively. Gut microbial DNA can be sequenced from feces or intestinal tissue, while human RNA is isolated from endoscopic biopsies or surgical samples. In some cases, patients are followed up for long periods and longitudinal samples can be obtained [15]. Multivariate methods are used to integrate DNA and RNA data, and therefore can identify relationships between the intestinal microbiome and the gut epithelium [8, 14, 16, 17]. Correlations, which are multivariate, are the predominant method used to find relationships between two omics datasets [7, 17–19]. A recent study revealed more significant correlations between host RNA and microbial DNA in samples from healthy controls than in patients with IBD, and suggests an "uncoupling" or breakup of these "homeostatic" correlations in diseased subjects [7]. Although their analysis used correlations, as well as univariate methods, these method do not consider confounders such as age, diet or sample localization in the gut, which could lead to false conclusions [20, 21].

Other multivariate methods provide frameworks with an unlimited number of variables involved [22, 23]. These methods summarize the variability of the datasets and select features in order to obtain loading factors for a new coordinate system where samples are represented. They summarize the largest amount of variability found among the samples' variables [24]. Those multivariate methods are capable of summarizing several variables from the same sample. Some multivariate methods work when variables are grouped in in a block. Multi-block methods allow to analyze variables obtained from different technical origins [25–29]. These

multi-block methods assume the existence of relationships between variables of the different blocks.

An example of a multi-block method is the regularized generalized canonical correlation analysis (RGCCA) which enables reducing the dimensions of an arbitrary number of blocks for data derived from the same sample [30–32]. RGCCA has already been used in the context of IBD with RNA-seq and 16S rRNA data [16]. However, it was used to select human genes and microorganism related to the inflammation predictors DUOX2 and APOA1. To our knowledge, a concrete description of the relationship between the gut's mucosal host transcriptome and microbiome in CD using RGCCA has not been performed.

In this study, we evaluate the effect of the parameters of RGCCA on the canonical components and we identify a strategy of analysis that better explains two previously published datasets. We then used this method, as well as multiple co-inertia analysis (MCIA), to compare two datasets, our hematopoietic stem cell transplant CD dataset and an online available pouchitis dataset in order to identify interactions between microorganisms and the host transcriptome of the gut epithelium [33]. Overall, we believe that our approach constitutes an innovative method for identifying multiple relationships present in multi-omics datasets and their most relevant variables. Identifying those relevant variables will lead to discover the cross-talk between microorganisms and the host and enhance our knowledge of the inflammatory bowel disease.

Methods

Patients and biopsies processing

Samples from the CD dataset included in this study were from a cohort of patients with severe refractory CD undergoing hematopoietic stem cell transplant (HSCT). Patients were treated in the Department of Gastroenterology (Hospital Clínic de Barcelona–Spain–). The protocol was approved by the Catalan Transplantation Organization and by the Institutional Ethics Committee of the Hospital Clinic de Barcelona (Study Number 2012/7244). All patients provided written consent following extensive counselling. Colonic and ileal biopsies were obtained at several time points during ileocolonoscopy. Patients were followed-up for 4 years and biopsies were collected every six or twelve months after HSCT. Samples were obtained when possible from both uninvolved and involved areas. In addition, biopsies were taken from the ileum and colon regions of 19 non-IBD controls consisting of individuals with no history of IBD and who presented no significant pathological findings following endoscopic examination for colon cancer surveillance (Hospital Univesitari Mútua de Terrassa–Spain–). The protocol was approved by the Institutional Ethics Committee of the Hospital Univesitari Mútua de Terrassa (Study Number NA1651). At least one biopsy was collected and fresh-frozen at -80˚C for microbial DNA extraction. The remaining biopsies were placed in RNAlater RNA Stabilization Reagent (Qiagen, Hilde, Germany) and stored at -80˚C until total RNA extraction.

Mucosal transcriptome

Total RNA from mucosal samples (HSCT CD cohort) was isolated using the RNeasy kit (Qiagen, Hilde, Germany). RNA sequencing libraries were prepared for paired-end sequencing using HighSeq-4000 platform. Later, cutadapt (version 1.7.1) was used for quality filtering and the libraries were mapped against the human reference genome using the STAR aligner (2.5.2a) with Ensembl annotation (release GRCh38.10). Read counts per gene were obtained with RSEM (version 1.2.31) as previously described [15]. Analysis was performed using R (version 3.6.1) and Bioconductor (Version 3.10) on Ubuntu 18.04. The host transcriptome was visually inspected for batch effects in PCA. Outliers and the top 10% genes using the coefficient

of variation were removed (20593, with remaining 37685 genes). Data was normalized using the trimmed mean of M-values and log transformed into counts per millions using edgeR (version 3.28).

Microbial DNA extraction from mucosal samples

Biopsies from the HSCT CD cohort were resuspended in 180 μl TET (TrisHCl 0.02M, EDTA 0.002M, Triton 1X) buffer and 20mg/ml lysozyme (Carl Roth, Quimivita, S.A.). Samples were incubated for 1h at 37˚C and vortexed with 25 μl Proteinase K before incubating at 56˚C for 3h. Buffer B3 (NucleoSpin Tissue Kit–Macherey-Nagel) was added followed by a heat treatment for 10 min at 70˚C. After adding 100% ethanol, samples were centrifuged at 11000 x *g* for 1 min. Two washing steps were performed before eluting DNA. Concentrations and purity were checked using NanoDrop One (Thermo Fisher Scientific). Samples were immediately used or placed at -20˚C for long-term storage.

High throughput 16S ribosomal RNA (rRNA) gene sequencing

Library preparation and sequencing were performed at the Technische Universität München. Briefly, volumes of 600μL DNA stabilization solution (STRATEC biomedical) and 400μL Phenol:choloform:isoamyl alcohol (25:24:1, Sigma-Aldrich) were added to the aliquots. Microbial cells were disrupted by mechanical lysis using FastPrep-24.: Heat tratment and centrifugation were conducted after adding a cooling adaptor. Supernatatnts were treated with RNase to eliminate RNA. Total DNA was purified using gDNA columns as described in detail previously [34]. Briefly, the V3-V4 regions of 16S rRNA gene were amplified (15x15 cycles) following a previously described two-step protocol [35] using forward and reverse primers 341F-785R [36]. Purification of amplicons was performed by using the AMPure XP system (Beckmann). Next, sequencing was performed with pooled samples in paired-end modus (PE275) using an MiSeq system (Illumina, Inc.) according to the manufacturer's instructions and 25% (v/v) PhiX standard library.

Microbial profiling

Data analysis was carried out as previously described [37]. Processing of raw-reads was performed by using the IMNGS (version 1.0 Build 2007) pipeline based on the UPARSE approach [38]. Sequences were demultiplexed, trimmed to the first base with a quality score <3 and then paired. Sequences with less than 300 and more than 600 nucleotides and paired reads with an expected error >3 were excluded from the analysis. Trimming of the remaining reads was done by trimming 5 nucleotides from each end to avoid GC bias and non-random base composition. Operational taxonomic units (OTUs) were clustered at 97% sequence similarity. Taxonomy assignment was performed at 80% confidence level using the RDP classifier [39] and the SILVA ribosomal RNA gene database project [34]. Later the data was normalized using the same method as for RNA-seq described above. The microbiome was visually inspected for batch effects in PCA; none were found. The resulting OTUs table was normalized using edgeR (Version 3.28).

Datasets

Table 1 shows all datasets included in the study. The glioma dataset came from diffuse intrinsic pontine glioma patients that included the host transcriptome analyzed with Agilent 44K Whole Human Genome Array G4410B and G4112F, patients copy number variation processed with the ADM-2 algorithm, and data from comparative genomic hybridization (CGH)

Table 1. Summary of samples and characteristics of the datasets used.

	Glioma	CD/UC	HSCT CD	Pouchitis
Samples (non-disease/diseased)	0/53	33/26	51/107	0/255
Sex (female/male)	28/25	42/17	22/15	101/102
Location	Cort: 20	Ileum:30	Ileum: 48	Pouch: 59
	Dipg: 22	Colon:29	Colon: 108	PPI: 196
	Midl: 11		Unknown: 2	
SES-CD local (mean (min-max))	NA		2.15 (0–12)	NA
CDAI mean (min-max)	NA		120 (0–450)	NA
Age at diagnostic (<16/16<x<40/x>40 years)			7/11/0	
Years of disease: mean (min-max)			14 (8–28)	

PPI: pre-pouch ileum. Cort: supratentorial, midl: central nuclei, dipg: brain stem. NA not applicable; an empty cell signifies unknown. Only the HSCT CD dataset was generated by the authors, all the other datasets were previously made publicly available.

https://doi.org/10.1371/journal.pone.0246367.t001

analyzed using Mutation Surveyor software. In addition, this dataset contained information on age, localization of the tumor, sex and a numerical grading of the severity of the tumor(see Table 1) [40, 41].

An IBD-related dataset was obtained from Prof. Dr. Rosentiel and Prof. Dr. Robert Häsler. It included samples from the terminal ileum and sigma from CD, UC, infectious disease-controls and healthy controls (see Table 1) [7]. The provided data included location, gender, location, age, and the status (inflamed or non-inflamed) of the region from which the biopsy was taken. The HSCT CD cohort involved 158 samples (both host RNA and microbial DNA) from 18 CD patients undergoing HSCT in our center and 19 non-IBD controls (Table 1) [15]. In addition to the samples, clinical information such as age, sex, treatment, years since disease diagnosis, prior surgery, location of the biopsies, segmental simple endoscopic score for Crohn's disease (SES-CD), time of the HSCT and response to treatment were collected. A previously published dataset from a pouchitis study was also analyzed (Table 1) [33]. A total of 255 samples from 203 patients were used containing data for both host transcriptome and microbiome. This dataset included identifiers for the patients, whether the sample was from the pre-pouch ileum or from the pouch, the sex, the outcome of the procedure and an inflammatory severity score ISCORE. The pouch ileum might be inflamed or not.

Integration

Sparse regularized generalized canonical correlation analysis (SRGCCA), implemented in RGCCA package (version 2.12), was used for this integration analysis [42]. This variation of the RGCCA method is better suited for biological data with sparsity such as the results obtained by RNA sequencing. The scheme used to add the different canonical components was the centroid scheme, which allows one to determine the positive and negative related variables. The regularization parameters used were those suggested by the tau.estimate, which is a compromise between correlation and covariance also known as Schäfer's method [43]. When looking for the covariance from phenotypic categorical variables in order to maximize the covariance instead of the correlation 1 was used for regularization.

Numeric values from the same assay were set on the same block. Relevant clinical variables were grouped in one block unless otherwise indicated. Categorical data was encoded as binary (dummy) variables for each factor, where 0 indicates not present and 1 indicates present omitting one level. Each block was standardized to zero mean and unit variances, and then divided by the square root of the number of variables of the block with the function scale2.

MCIA was also performed on the CD/UC, CD and pouchitis dataset using only the experimental data [28]. RGCCA was compared to MCIA by examining the area under the curve (AUC) of both methods when classifying localization on the first component of the shared latent space of MCIA and the first component of the host transcriptome on the RGCCA method.

Parameter testing

The sparse canonical correlation analysis involved three parameters besides the input data: the regularization parameter (tau) the model and the scheme. To evaluate the effect of each parameter, the one being tested was changed while keeping constant all the others. This model included weights indicating the relationship between the blocks. These parameters were tested on the glioma dataset and on the CD/UC dataset.

All models were analyzed using weights from 0 to 1 by 0.1 intervals in the relationship between blocks. These weights indicate the strength of the relationship between the variables of two blocks, the higher it is, the stronger is the relationship between the variables. To test the effect of the model, all combinations of weights were analyzed. The indicators of methods quality consist of the inner average variance explained (AVE) the outer AVE and the AVE of each block. The inner AVE is defined by how well the components of each block correlate with one other [31]. The outer AVE is defined by how well the variables of a block correlate with the component for all of the blocks. As we were interested in discovering the relationships between blocks, the inner AVE was used to select the best model, the higher the inner AVE is, the better the model.

The scheme controls how the different correlations of the canonical components are summarized. The three schemes available (horst, centroid and factorial) are compared using a simple model regarding their inner AVE and the selected genes.

Tau was tested on the glioma and the CD/UC dataset between the minimum accepted value and 1 for each block.

Models were validated using 1000 bootstraps with resampling to assess the stability of the inner and outer AVE.

Models used

Different models were tested for the integration of the data from the CD or the pouchitis dataset. The first model, model 0, used only two blocks, the microbiome and the host transcriptome data with interaction between them, but with no within interactions (Model not shown).

The second family of models (models 1, 1.1 and 1.2), family 1, in addition to the microbiome and host transcriptome data, included those variables we considered clinically relevant variables including some that were related to disease activity. This model was explored because it takes into account already known information that could help reveal relevant relationships. For instance, the HSCT CD dataset included the following variables: patient ID, sex, age, age at diagnosis, previous surgery, current treatment, time after HSCT and location of the sample. Including these variables could potentially help to reveal a relationship that changes with patient's characteristics, time and location.

The last family of models (models 2, 2.1, 2.2 and 2.3), family 2, used the same information as that for family 1 models, but grouped the clinical variables into three blocks, one for demographics, one for time-related variables and one for variables related to localization of the sample. Although, this family of models is more complex than family 1 the relationships found can potentially occur independently of time, clinical variables and location, thus revealing other relationships that could not be identified using the family 1 models. All models can be found

on S1 Data. Models 1 to 2.3 were modeled to utilize known, clinically relevant variables with the host transcriptome and microbiome data available.

With the glioma dataset, the microbiome block was replaced by the CGH block. In addition to the previously mentioned models, the glioma dataset was also analyzed considering all the variables from the different blocks as a single entity, which is known as a superblock [44]. A superblock is a block created with all the variables on the system usually connected with each individual block of the system being analyzed.

Only the models in which all the blocks were part of a single connected network were analyzed, thus, 31 of all possible models were filtered out. For models 1 to 2.3, all the combinations of different weights on the model matrix were analyzed. First weights 0, 0.5 and 1 were used to select the model with the highest inner AVE. To further describe the interactions of models 1.1 and 2.1, different weights from 0 to 1 by 0.1 intervals were tested; the best model of each family resulted from model 1.2 and 2.2, respectively. By taking into account a direct interaction between the microbiome and the host transcriptome we could confirm whether the results of model 2.2 had improved in model.

Results

Parameters on the glioma dataset

We first determine the best strategy to obtain the right values of the parameters on SRGCCA using the glioma dataset. This was the dataset originally used to develop and test the SRGCCA method [39]. By parameters we mean the scheme used, the regularization effect, and the models as constructed by weights, all of which can affect the final solution of the SRGCCA (See Parameters testing in Methods).

Tau controls the number of variables selected from each block, regulating the stringency of the model. Tau can be estimated using Schäfer's method [43], which tries to balance both the correlation and the covariance for selecting the variables of the block. When estimated by this method, the tau provides a good intermediate solution for numeric variables. For those blocks that encode categorical variables as numeric values, the covariance of the block with the other block is the only relevant meaning; thus, a tau value of 1 is more appropriate although several values were explored. The effect of tau on the inner AVE is shown in Fig 1A, where usually an increase on tau increases the inner AVE as well, although Schäfer's method provided result is close to the optimum value.

All the weights between 0 and 1 (by 0.1 intervals) in the glioma dataset were analyzed using all three schemes: horst, centroid and factorial. The horst and the centroid scheme were similar while the factorial resulted in the most different AVE values (see S1 Data). The centroid scheme takes into account all the relationship regardless of the canonical correlation sign. This, together with its similarity to horst scheme, prompted its selection as the best scheme.

The three blocks with the best tau and the centroid scheme were analyzed by changing the weights between 0 and 1 by 0.1 intervals. According to the inner AVE, the best model was the one in which the weights (1) between the host transcriptome and location, (2) the host transcriptome and the CGH, and (3) the CGH block were linked to variables related to the location with weights of 1, 0.1 and 0.1, respectively.

When we added a superblock to the data, there was an increase of 0.01 on the inner AVE of the model (See Methods section Models used and [44]). The model with the superblock that explained most of the variance was that in which the weights of the interaction within (1) the host transcriptome, (2) between the superblock and the CGH, (3) between the host transcriptome and the localization, and (4) between CGH and the host transcriptome were 1, 1, 1 and 1/3, respectively. To see if the superblock could classify the sample by location, we plotted the

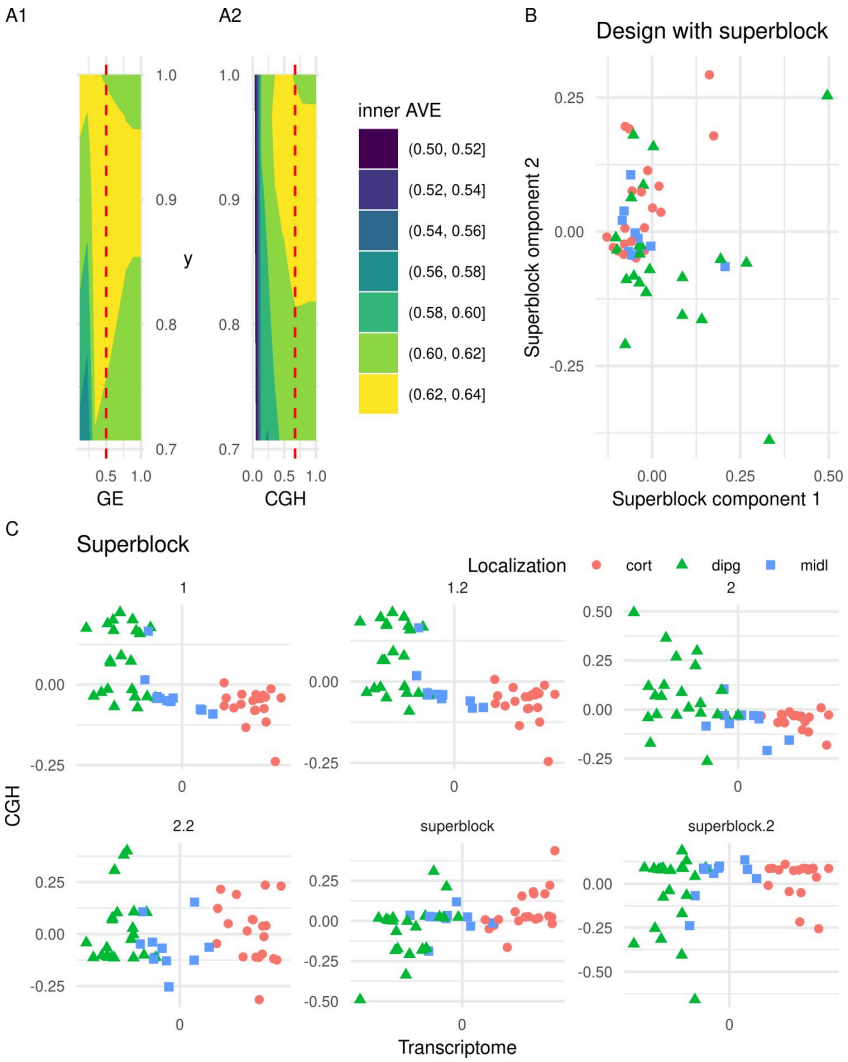

Fig 1. Analysis of the parameters on the glioma dataset. A1 and A2: A contour plot of the median of the inner AVEresult of an SRGCCA with different tau values for each block (GE, gene expression of the host transcriptome, CGH (comparative genomic hybridization) for the copy number variation and y for the location). Higher tau normally increases the inner AVE, Schäfer's approximation is marked with the red vertical line. B: First two dimensions of the superblock on the glioma dataset. The first two components of the superblock within the best model, according to the inner AVE from the glioma dataset. C: First dimensions of the host transcriptome and the CGH block of models on the glioma dataset are represented. Comparison of the different models by visualizing the first components of the host transcriptome gene expression (GE) and the copy number variation (CGH) blocks from the glioma dataset. Each point represents a sample (colored by location). Cort: supratentorial, dipg: brain stem, midl: central nuclei.

https://doi.org/10.1371/journal.pone.0246367.g001

first two components of the superblock (see Fig 1B). We can clearly see that they do not classify the samples according to the location of the tumor, which is known to affect the tumor phenotype [40].

Adding one block containing the age of the patient and the severity of the tumor to the model, decreased the inner AVE. The best model with these blocks, according to the inner

AVE, was that in which the interactions (1) within the host transcriptome, (2) between the host transcriptome and the localization, (3) between the host transcriptome and(4) the CGH and between the CGH and the other variables were 1, 1, 1/3 and 1/3, respectively (see S2 Data, Glioma sheet). The first components of each model can be seen in Fig 1C. We can observe on the figure, the strong dependency between gene expression and location since the first model while the weaker relationship with the CGH assay [40]. On the other hand, the major difference is the dispersion on the CGH component on each model.

As the model with a superblock did not help explain the relationships between blocks, we decided not to apply it to the other datasets. The scheme selected was the centroid, which takes the absolute value of the relation between components. These parameters were used for further analysis on the CD/UC, the CD and pouchitis datasets.

Parameters on the CD/UC dataset

After an exploratory analysis of the parameters on the glioma dataset, we analyzed the CD/UC dataset, which was similar to our CD dataset and include information on both the host transcriptomics and bacterial genomics. These data were obtained using the same sequencing techniques from endoscopic biopsies.

In this dataset, the parameter tau behaved slightly differently than with the previous dataset but the value from the Schäfer's method for tau was close to the best value (see S1 Fig).

In contrast to the glioma dataset, the model with the highest inner AVE was model 1.2 (S2 Data). Model 2.2 has a relationship of 0.1 between microbiome and the host transcriptome and of 1 between the location and the host transcriptome. The microbiome block is also related by a factor of 0.1 with the demographic block and of 1 with the time block. Lastly the time and the demographic block are related by a factor of a 0.1. In either case the family 1 and family 2 models can correctly separate by sample location (colon or ileum) but not by disease type (see Fig 2) or inflammation status (data not shown).

Analyzing the models on the HSCT CD and pouchitis datasets

Having established the best parameters for analyzing a related IBD dataset, we studied our HSCT CD dataset using SRGCCA. Model 1.2 had the highest inner AVE of the family 1 model. A search for the highest inner AVE within the family 2 models resulted in model 2.2 (S2 Data). This model revealed a direct relationship between the host transcriptome and the location-related variables, while the microbiome was associated with the demographic and location-related variables (see Fig 3 and S2 Data). Overall, we see that the relationships in the model affected the distribution of samples on the components of both the host transcriptome and the microbiome.

Finally, we used another related cohort to confirm the applicability of SRGCCA to an independent dataset (see Fig 4). Model 1.2 had the highest inner AVE. A search for the highest inner AVE among the family 2 models resulted in model 2.2, although it did not have a higher inner AVE than model 1.2. Moreover, no direct relationship between the host transcriptome and the clinically relevant variables was apparent (S2 Data). Family 2 models better stratified the samples by location (pouch vs pre-pouch) than did those of family 1. Nonetheless, they were separated by location-related variables in some models, albeit not as clearly as with the HSCT CD dataset. This might indicate that while sex does not affect the interaction, the location-related variables do affect the pouchitis.

Of all these models, as described above, the best according to the inner AVE on the HSCT CD dataset was model 2.2. This model explained known differences between the host

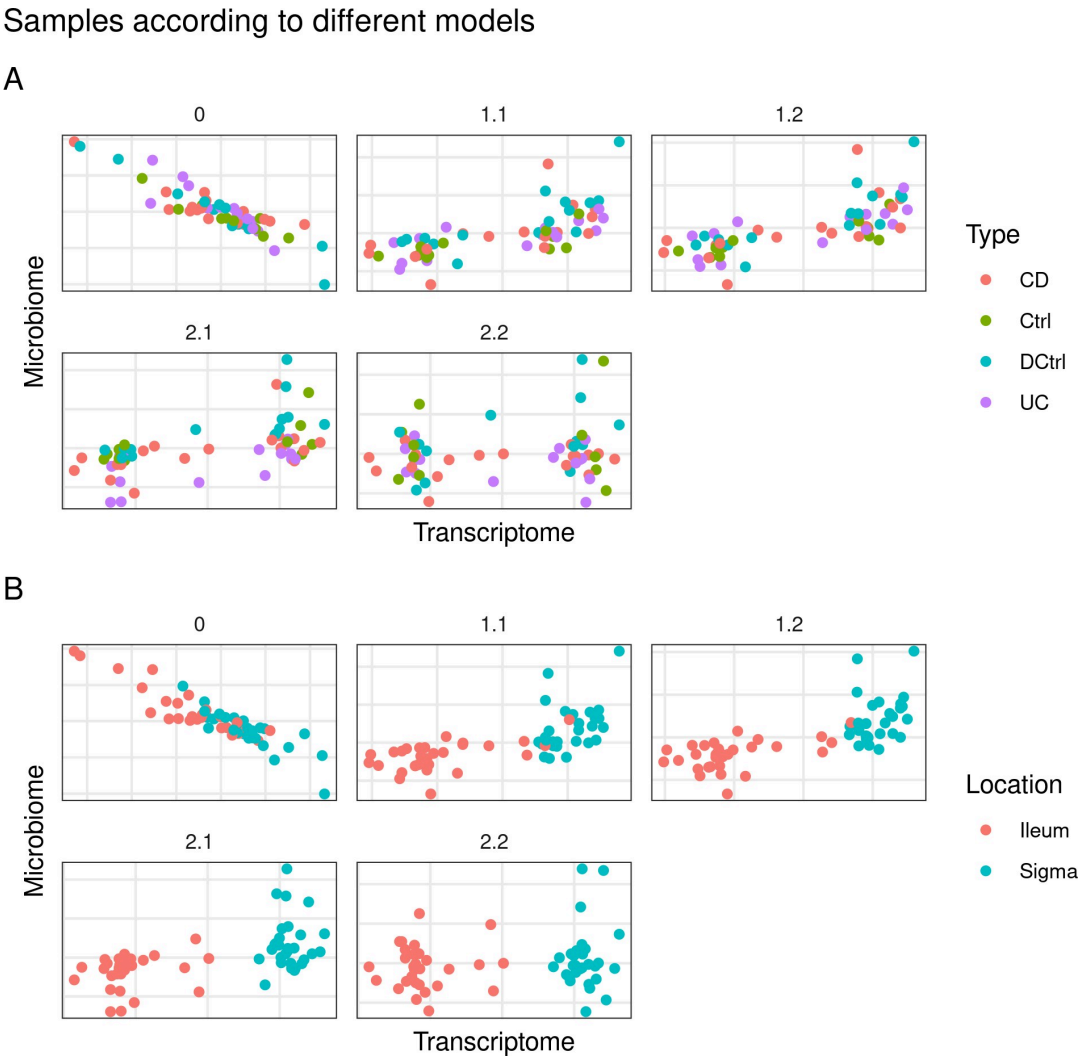

Fig 2. First dimensions of the host transcriptome and the microbiome block of models on the Crohn's disease ulcerative colitis/ dataset. Comparison of the models that better explained the interaction between the microbiome and the host transcriptome data on the CD/UC dataset. Each point represents a sample colored according to a characteristic: A) samples are colored by disease type, CD Crohn's disease, Ctrl, control; DCtrl diseased control, inflamed but not from IBD patients, UC ulcerative Colitis; and B, by location, colon or ileum, on the first components of the host transcriptome and the microbiome. Better models separate samples by tissue location using the host transcriptome component.

https://doi.org/10.1371/journal.pone.0246367.g002

transcriptome gut regions [15]. The microbiome separated the samples by disease status, indicating that it was highly relevant for the relationship with the host transcriptome.

Using the HSCT CD dataset we also looked for the best model using a single block for the clinically relevant variables, following the family model 1 structure. The model from family 1 models with the highest AVE was that in which the transcriptomics was related to the phenotype by 0.1, while the microbiome was related to the clinically relevant variables by 1. This model revealed that the relationship between the microbiome and the clinically relevant

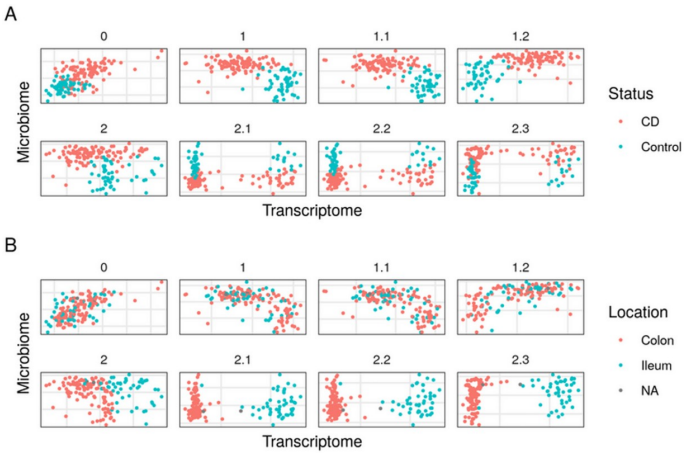

Fig 3. First dimensions of the host transcriptome and the microbiome block of models on the hematopoietic stem cell transplant Crohn's disease dataset. Comparison of the models that better explained the interaction between the microbiome and the host transcriptome data on the HSCT CD dataset. Each point represents a sample (colored by disease status): A, non-CD (Control) or CD; and B, by location, colon or ileum, on the first components of the host transcriptome and the microbiome. Better models separate samples by tissue location by the host transcriptome component and the diseased and controls samples by the microbiome component.

https://doi.org/10.1371/journal.pone.0246367.g003

variables carried more weight than that between the clinically relevant variables and the transcriptomics on the HSCT CD dataset.

In addition, the host transcriptome was related to location-dependent variables by a weight of 1, while the microbiome was related to demographic variables, and to location related variables, by a weight of 1 and 0.5, respectively. Demographic variables were also linked by 1 to the time variables block (see S2 Data, HSCT_CD sheet).

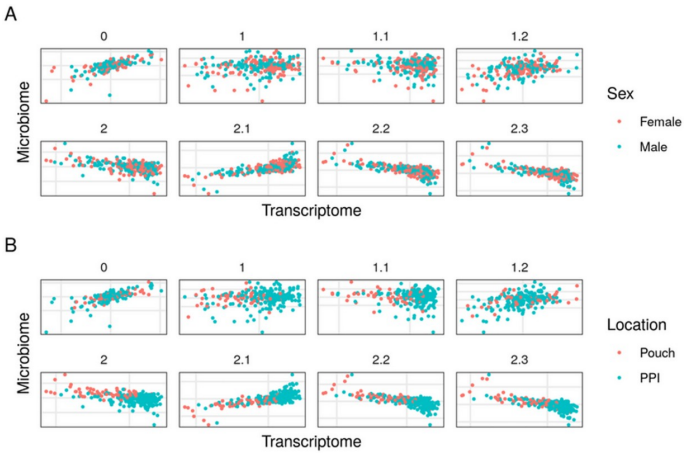

Fig 4. First dimensions of the host transcriptome and the microbiome block of models on the pouchitis dataset. Comparison of the models vis-à-vis on the pouchitis dataset by the first component of the host transcriptome and the microbiome from the HSCT CD dataset. Each point represents a sample colored by sex (A), where females are in red and males in blue, and by location (B), where the pouch is the red, and PPI is the pre-pouch ileum. The samples do not show a sex-specific pattern but on the best models the host transcriptome partially separates pouch and pre-pouch ileum samples.

https://doi.org/10.1371/journal.pone.0246367.g004

The interaction of genes within the host transcriptome was also analyzed on the HSCT CD dataset. Adding this interaction increased the inner AVE score between 0.10 and 0.03 depending on the model. However, it was not deemed important to find the relationships between the host transcriptome and the microbiome and thus was not compared between datasets.

Genes selected by SRGCCA as related to the microbiome in our HSCT CD dataset were different between the family 1 and 2 models (see Fig 5A), suggesting that the relationship between microorganisms and genes is independently influenced by location, time and demographic-related variables. The influence of the microbiome remained constant as indicated by the high number of OTUs shared between family 1 and 2 models suggesting that previously observed differences might have been due to covariates since the microorganisms identified by multiple models remained unchanged (Fig 5B).

Comparison of models

As expected, when analyzing the same dataset with different models the output results in different relevant variables. In order to analyze the accuracy of the models, one thousand bootstraps were used to integrate the data from the HSCT CD dataset (Fig 4 and Table 2). Each model had its own dispersion on the same bootstrapped samples (Fig 6). The lower the dispersion, the more robust the model was to different conditions than in the initial testing.

Model 2.2 had both higher inner and outer AVE mean values and less standard deviation (Fig 4 and Table 2). This indicates that it was more robust than the other models, regardless of the input data.

The bootstrap analysis of the one thousand bootstraps on the pouchitis dataset showed that model 1.2 had the highest mean inner AVE, while model 0 had the highest mean outer AVE (Table 3). Overall, model 1.2 was considered the most robust.

The models with the highest inner AVE were more robust to different data, which indicates that they can be applied more generally and not solely to these samples.

Comparison of methods

We have seen that this method provides robust models of the interactions on the datasets. However, given the many methods available for integration multiple omics, we sought to determine how these methods would perform compared to other existing approaches. In particular, we ran a comparison with MCIA, which is a newer method that requires less parameters while still being conceptually similar to SRGCCA.

Applying MCIA to the CD/UC, HSCT CD and pouchitis datasets produced similar distribution on the synthetic space compared to our method (Fig 7). This method was able to classify the samples by their location on the first component in a manner similar to our own method with the first transcriptomic component. On the pouchitis dataset neither method could separate the samples by location while MCIA did worse than our best model according to the AUC. In all three datasets the best model outperformed MCIA when classifying the samples according to their location (Fig 7), with the greater difference involving the pouchitis dataset (data not shown).

Discussion

This study provides a framework for identifying interactions between blocks of data, a step towards understanding biological relationships between datasets or between datasets and other particularly relevant variables. First, we studied the parameters' influence on a glioma and CD/UC dataset, adjusting their values and testing how generalizable they are. Then, we

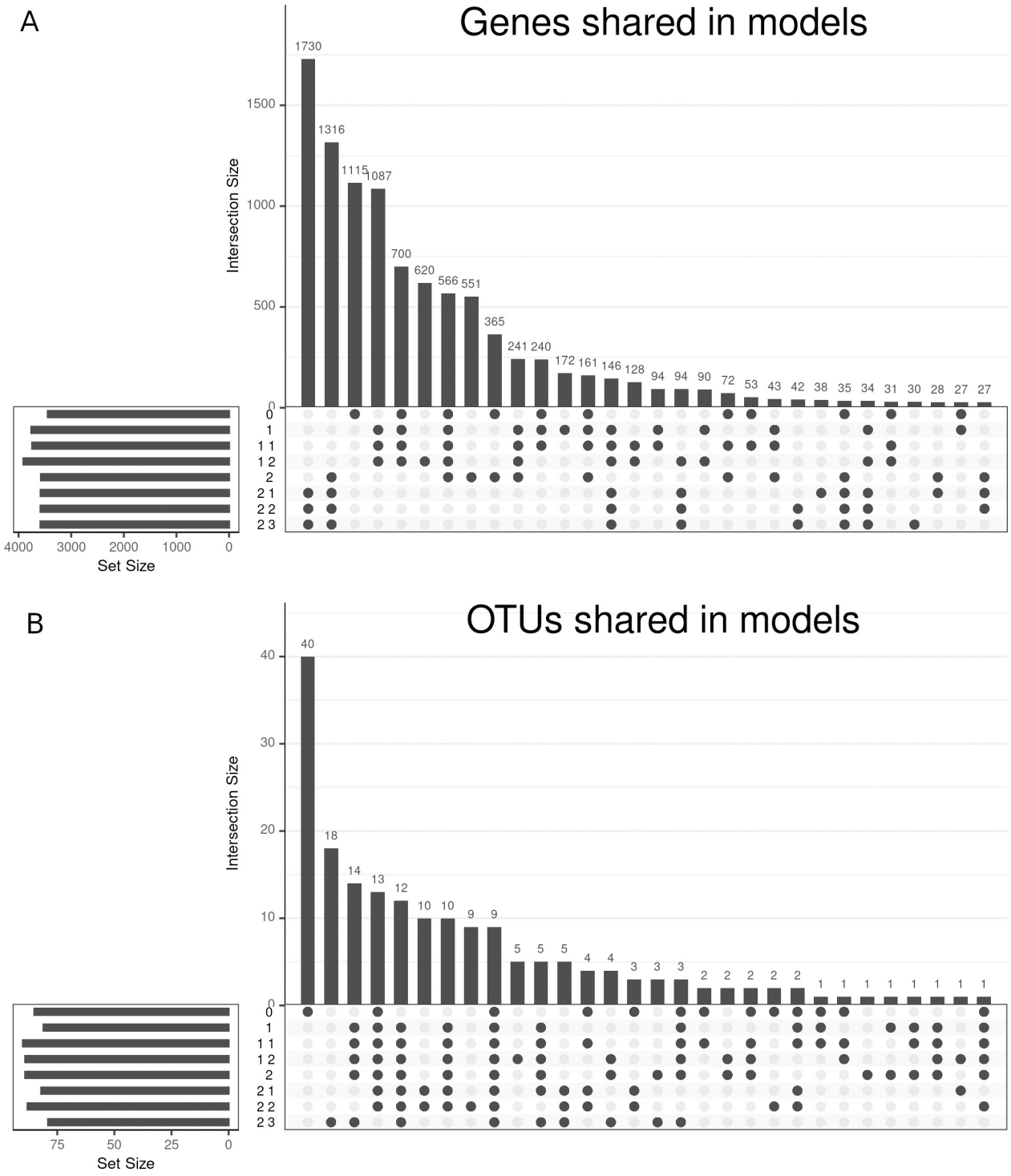

Fig 5. UpSet plot of the of the models on the hematopoietic stem cell transplant Crohn's disease dataset. The heights of the bars represent the genes (A) or OTUs (B) shared between the models selected by the points; 30 intersections are shown.

https://doi.org/10.1371/journal.pone.0246367.g005

Table 2. Bootstrapped mean and standard deviation of inner and outer AVE values on the HSCT CD dataset.

Model	AVE	Mean	Sd
0	inner	0,550	0,0469
1.2	inner	0,768	0,0223
2.2	**inner**	**0,785**	**0,0163**
0	outer	0,104	0,0132
1.2	outer	0,088	0,0106
2.2	**outer**	**0,105**	**0,0069**

The best models according to the mean are shown in bold.

https://doi.org/10.1371/journal.pone.0246367.t002

developed a method to find the best model for the relationships between blocks. Lastly, we validated the method in two independent datasets.

We explored the regularization of the blocks on two previously published datasets from glioma and IBD patients. The regularization of a block modulates how many variables are selected and whether correlation or covariance have to be used when looking for the canonical correlation with other blocks [28, 30]. A tau value of 1 allowed us to select all variables, which maximized their covariance. On blocks that included only clinically relevant categorical variables, regularization must be equal to 1, since correlations with categorical variables have a different meaning. As the host transcriptome and microbiome blocks contain many variables, a shrinkage parameter closer to 0 was expected, as observed with the glioma and the CD/UC datasets. In addition, estimating tau for the quantitative blocks resulted in higher inner AVE scores since the quantitative variables that contributed most to the data variation were selected.

Based on the regularization obtained, we explored different schemes of integration on the glioma dataset. The resulting canonical components of the centroid and horst schemes did differ in some models. In fact, the canonical correlations between blocks were likely positive, making the differences between these two schemes unobservable. The centroid scheme was selected to analyze the CD and the pouchitis datasets, since canonical correlations are not always positive.

Independently of the scheme involved, a superblock not only aids in interpretation, but also helps account for the possibility of interactions between variables of the same block. The increase observed in the inner AVE may have stemmed from the interaction between variables of the same block. However, such an interpretation is not as clear as with blocks generated by a single assay or from closely related variables [30]. The superblock, which is used for redundancy analysis, did not help in terms of grouping different samples [44]. Moreover, if the goal of the model is to accurately represent the system under study, the superblock is not necessary, regardless of the assistance it provides in improving the inner AVE.

The superblock is usually related to all the other blocks. Typically, a weight of 1 is used to indicate a direct relationship between two blocks. Modifying the weights of the model influenced the result by changing AVE scores and the variables selected from each block. The highest inner AVE score was not defined by the highest weights on all the relationships.

The weights of the models represent how much one block interacts with another if the interactions are linear, an assumption of any canonical correlation [31]. In such cases, the weights are representative of the interactions between blocks.

The weights define the relationships between blocks in SRGCCA, which together determine the model of the components. Other methods like MCIA and joint and individual variation-explained (JIVE) assume a common relationship between all components, which results in a

Fig 6. Bootstrap results of three models on the hematopoietic stem cell transplant Crohn's disease dataset. Variance of AVE using the same samples on three models with the HSCT CD dataset. Each point shows the AVE for each analysis performed. The brighter colors reflect the result of this model on the original data (including all samples). Dispersion on the bootstrapped samples is reduced as a model more accurately represents the relationships present on the dataset.

https://doi.org/10.1371/journal.pone.0246367.g006

common space for the samples [27, 28]. This difference is crucial for exploring the role of the components; for example, in our manuscript each model represents the same system with different interactions and assumptions. Comparing different models after the SRGCCA led to explanations for different aspects of the same system. Here, we also show that compared to our method, MCIA can results in similar samples' classification on a latent space. However, it was not always as good as was evident by the AUC when classifying the samples by location. In addition, the interpretation of the MCIA was not as straight forward as with SRGCCA. Furthermore, with our method the observed classification of the samples according to their

Table 3. Bootstrapped mean and standard deviation of inner and outer AVE values on the pouchitis dataset.

Model	AVE	Mean	Sd
0	inner	0,448	0,0811
1.2	**inner**	**0,820**	**0,0457**
2.2	inner	0,767	0,0332
0	**outer**	**0,140**	**0,0087**
1.2	outer	0,120	0,0227
2.2	outer	0,134	0,0085

The models with the higher mean AVE values are shown in bold.

https://doi.org/10.1371/journal.pone.0246367.t003

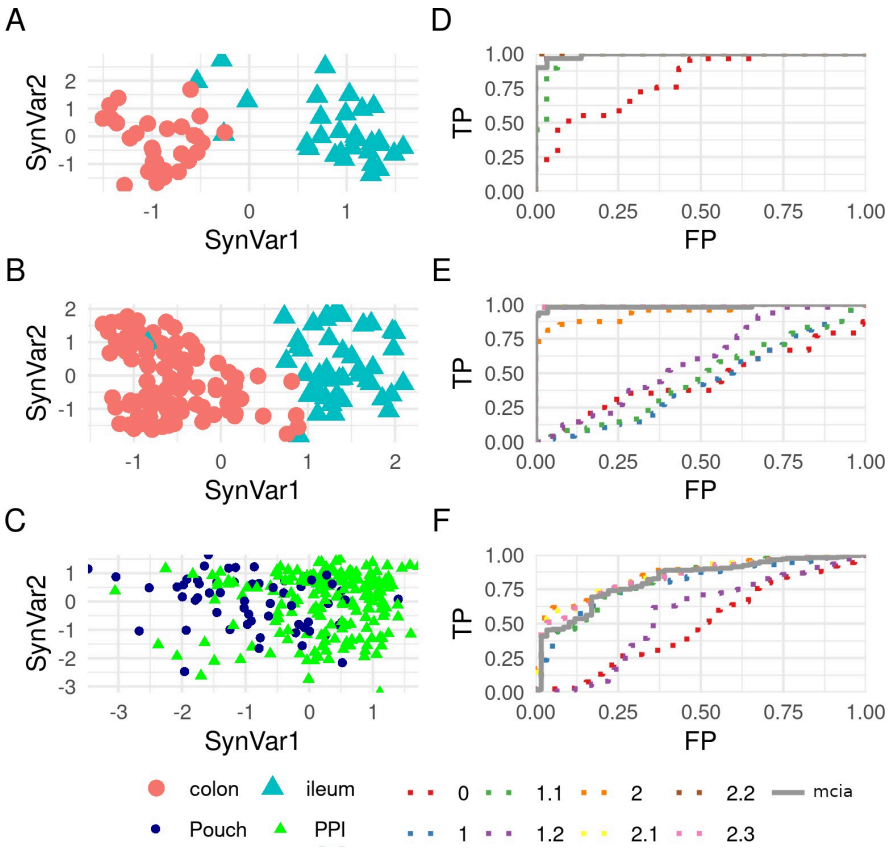

Fig 7. Multiple co-inertia analysis and area under the curve for the location of the Crohn's disease/ulcerative colitis, the hematopoietic stem cell transplant Crohn's disease and pouchitis dataset. A, B, C plots are the results of applying multiple co-inertia analysis (MCIA) where the horizontal and vertical axis represent the synthetic variable 1 and 2 respectively. D, E, F plots are the area under the curve (AUC) for all the methods applied on this dataset. The first row (A, D) is the analysis of CD/UC dataset, the second one (B, E) the HSCT CD dataset, and the third one (C, F) the pouchitis dataset.

https://doi.org/10.1371/journal.pone.0246367.g007

location can be directly attributed to the host transcriptome while with MCIA that effect could result from either the host transcriptome or the microbiome.

Looking at the glioma data, the best model according to the inner AVE was that with the superblock. As previously explained, this model might represent the hierarchical relationships present in the data. However, the superblock did not provide more interpretable results in the glioma dataset.

In the glioma dataset, the model lacking the superblock but with the highest inner AVE indicated that the localization of a tumor influences the host transcriptome to a greater degree than the copy number variations, if the relationships are linear. Adding supplementary information on the samples' localization did not increase the inner AVE, suggesting that there was a high dependence between localization and the tumor host transcriptome.

Interactions within the host transcriptome usually increase the inner AVE of the models. With the CD and the pouchitis datasets, self-interaction increased the inner AVE, as well as the selected features, except in models 0 to 1.2 in the CD data set. This suggests that the interactions within the same omic block become relevant if the model does not take into account

the interaction between other clinically relevant variables. If other relevant variables are included, then the effect of this interaction is significantly less.

Model 0 looked for direct relationships between the microbiome and the host transcriptome. Confounders that influence both host transcriptome and microbiome, such as age or the localization and inflammation status, were not taken into account in this model. This is due to the fact that they can bias the relations found with this model [45]. Nonetheless, this model was capable of grouping the samples of the CD dataset according to their disease status, though this was not true of the pouchitis dataset.

Family 1 models use three blocks, including one for clinically important information about the samples. This new block was added to avoid biasing the integration by known factors of the samples such as sex, or location. In the best model of this family, the microbiome block had a weak relationship with the host transcriptome. This weak relationship was possibly an indicative of not lineal relations. If the relationships were not lineal, then they could not be fully identified by RGCCA [31]. Another possibility is that the microbiome was related to other variables not included on the dataset.

Finally, family 2 models, compared to those of family 1, were designed to explain the relationship between the microbiome and host transcriptome, allowing for the presence of independent interactions with location, age and other demographic-related variables. In family 1 models all the relevant variables were mixed together. In order to allow for such interactions, unrelated variables were separated in different blocks.

In the HSCT CD dataset, a cursory analysis confirmed that the genes selected by SRGCCA with model 2.2 were related to the sample location [15]. The selected microorganisms previously linked to CD dysbiosis were *Faecalibacterium sp*. and *Bacteroides sp*. (see S3 Data) [46]. This suggests that the variables selected were relevant for their role in both the tissue and the disease. Thus, the genes and microorganisms that have significant relationships were likely to be present in this context.

There are several previously known interactions between the variables collected on the multiple datasets. For instance the butyrate produced by the microbiome affects the state of the epithelial cells, implying a relationship between the microbiome and the host transcriptome [47]. It is also known that the microbiome changes along the gastrointestinal tract; thus, the microbiome and host transcriptome blocks must be connected [48]. Moreover, the microbiome is influenced by diet, which would imply a relationship between demographics and the microbiome [49]. In addition, there are some studies that observe changes over time, with perhaps additional links to changes in diet. With our method we could a connection between all of these blocks.

In the pouchitis dataset, model 1.2 captured a greater degree of variance than model 2.2, contrary to the results obtained with the HSCT CD dataset. This might be because potentially important variables, such as age, were lacking and possibly because the model was confounded. In addition, we could not make direct comparisons with the HSCT CD dataset as it did not include non-diseased samples although it did include non-inflamed samples. This is due to the fact that the model differentiates by subgroups of patients instead of by a distinct relationship between healthy and diseased samples.

The findings of this study have to be assessed in light of certain limitations. RGCCA cannot describe a causal relationship or the mechanisms underlying the relationships between RNA transcriptomics and the microbiome. However, models for RGCCA can be used to select variables for further studies and experiments in order to validate these relationships. This method has been implemented in an R package, called inteRmodel, which can be found at https://github.com/llrs/inteRmodel/. This package implements the methodology described in this manuscript and also incorporate some help functions for the analysis.

When examining an interaction within a block, we only assumed the existence of an interaction within the host transcriptome. However, it must be noted that microorganisms create communities for which the interactions of several microorganisms is essential and we did not consider interaction within the microbiome in the present study [50]. Knowing how microbial communities rise and interact remains an open question that could affect any interpretation of the results [50, 51]. In addition, the taxonomy imputation can be biased by the copy number variation of the 16S rRNA present on the microbiome. This problem has not yet been solved, and the workflow used could over-estimate the abundance of some taxonomies [52].

In the present study, as we did not use a simulated data set with known relationships between blocks, we could not assess the specificity or sensitivity of our approach. In addition, we could not confirm by further analysis and experiments whether the selected variables were necessary to start or maintain CD or pouchitis.

Conclusions

RGCCA is a powerful integration tool. We have shown that the model is the most important parameter when selecting variables. The weights of the model represent the strengths of the relationships between blocks. Here we propose a robust methodology implemented with inteRmodel, to identify the best models guided by the inner AVE when there is no prior knowledge of the existing relationship.

This method can identify relationships in complex systems such as Crohn's disease by taking into account the interactions between the microbiome, host transcriptome and the relevant clinical variables. The resulting analysis can improve our understanding of the biological relationships between different omics datasets and other relevant (clinical) variables.

Supporting information

S1 Data. Schemes types and AVE on the glioma dataset.
(XLSX)

S2 Data. Models design representations.
(XLSX)

S3 Data. Variables selected by the model 2.2 on the HSCT CD dataset.
(XLSX)

S1 Fig. Tau effect on the CD/UC dataset.
(TIF)

Acknowledgments

We thank Daniel Aguilar for RNA-seq analysis assistance and Ilias Lagkouvardos for his assistance on 16S-seq analysis. Many thanks to Robert Häsler and Philip Rosenstiel for providing the processed data and the metadata of the UC/CD datasets. We are grateful to Joe Moore for English-language assistance.

Author Contributions

Conceptualization: Lluís Revilla, Juan J. Lozano, Azucena Salas.

Data curation: Lluís Revilla, Ana M. Corraliza, Maria C. Masamunt.

Formal analysis: Lluís Revilla.

Funding acquisition: Dirk Haller, Julian Panés, Azucena Salas.

Investigation: Lluís Revilla, Aida Mayorgas, Maria C. Masamunt, Amira Metwaly.

Methodology: Lluís Revilla, Azucena Salas.

Project administration: Maria C. Masamunt.

Resources: Maria C. Masamunt, Eva Tristán, Anna Carrasco, Maria Esteve, Julian Panés, Elena Ricart, Juan J. Lozano.

Software: Lluís Revilla, Ana M. Corraliza, Juan J. Lozano.

Supervision: Dirk Haller, Juan J. Lozano, Azucena Salas.

Visualization: Lluís Revilla.

Writing – original draft: Lluís Revilla.

Writing – review & editing: Lluís Revilla, Aida Mayorgas, Ana M. Corraliza, Maria C. Masamunt, Amira Metwaly, Dirk Haller, Eva Tristán, Anna Carrasco, Maria Esteve, Julian Panés, Elena Ricart, Juan J. Lozano, Azucena Salas.

References

1. Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp Oncol Poznan Pol. 2015; 19: A68–77. https://doi.org/10.5114/wo.2014.47136 PMID: 25691825
2. Human Microbiome Project Consortium BA, Nelson KE, Pop M, Creasy HH, Giglio MG, Huttenhower C, et al. A framework for human microbiome research. Nature. 2012; 486: 215–21. https://doi.org/10.1038/nature11209 PMID: 22699610
3. Beale DJ, Karpe AV, Ahmed W. Beyond Metabolomics: A Review of Multi-Omics-Based Approaches. In: Beale DJ, Kouremenos KA, Palombo EA, editors. Microbial Metabolomics: Applications in Clinical, Environmental, and Industrial Microbiology. Cham: Springer International Publishing; 2016. pp. 289–312. https://doi.org/10.1007/978-3-319-46326-1_10
4. Holmberg FE, Seidelin JB, Yin X, Mead BE, Tong Z, Li Y, et al. Culturing human intestinal stem cells for regenerative applications in the treatment of inflammatory bowel disease. EMBO Mol Med. 2017; 9: 558–570. https://doi.org/10.15252/emmm.201607260 PMID: 28283650
5. McIlroy J, Ianiro G, Mukhopadhya I, Hansen R, Hold GL. Review article: the gut microbiome in inflammatory bowel disease-avenues for microbial management. Aliment Pharmacol Ther. 2018; 47: 26–42. https://doi.org/10.1111/apt.14384 PMID: 29034981
6. Øyri SF, Műzes G, Sipos F. Dysbiotic gut microbiome: A key element of Crohn's disease. Comp Immunol Microbiol Infect Dis. 2015; 43: 36–49. https://doi.org/10.1016/j.cimid.2015.10.005 PMID: 26616659
7. Häsler R, Sheibani-Tezerji R, Sinha A, Barann M, Rehman A, Esser D, et al. Uncoupling of mucosal gene regulation, mRNA splicing and adherent microbiota signatures in inflammatory bowel disease. Gut. 2016; gutjnl–2016–311651. https://doi.org/10.1136/gutjnl-2016-311651 PMID: 27694142
8. Haberman Y, Tickle TL, Dexheimer PJ, Kim M-O, Tang D, Karns R, et al. Pediatric Crohn disease patients exhibit specific ileal transcriptome and microbiome signature. J Clin Invest. 2014; 124: 3617–3633. https://doi.org/10.1172/JCI75436 PMID: 25003194
9. Loganathan P, Catinella AP, Hashash JG, Gajendran M, Loganathan P, Catinella AP, et al. A comprehensive review and update on Crohn's disease. Dis Mon. 2018; 64: 20–57. Available: http://www.ncbi.nlm.nih.gov/pubmed/28826742 http://linkinghub.elsevier.com/retrieve/pii/S0011502917301530 https://www-sciencedirect-com.sire.ub.edu/science/article/pii/S0011502917301530?via%3Dihub https://doi.org/10.1016/j.disamonth.2017.07.001 PMID: 28826742
10. Azimi T, Nasiri MJ, Chirani AS, Pouriran R, Dabiri H. The role of bacteria in the inflammatory bowel disease development: a narrative review. APMIS. 2018; 126: 275–283. https://doi.org/10.1111/apm.12814 PMID: 29508438
11. Hata K, Ishihara S, Nozawa H, Kawai K, Kiyomatsu T, Tanaka T, et al. Pouchitis after ileal pouch-anal anastomosis in ulcerative colitis: Diagnosis, management, risk factors, and incidence. Dig Endosc Off J Jpn Gastroenterol Endosc Soc. 2017; 29: 26–34. https://doi.org/10.1111/den.12744 PMID: 27681447

12. De Souza HSP, Fiocchi C, Iliopoulos D. The IBD interactome: An integrated view of aetiology, pathogenesis and therapy. 2017; 14. https://doi.org/10.1038/nrgastro.2017.110 PMID: 28831186
13. Gaujoux R, Starosvetsky E, Maimon N, Vallania F, Bar-Yoseph H, Pressman S, et al. Inflammatory bowel disease Cell-centred meta-analysis reveals baseline predictors of anti-TNFα non-response in biopsy and blood of patients with IBD. Gut. 2018; 0: 1–11. https://doi.org/10.1136/gutjnl-2017-315494 PMID: 29618496
14. Huang H, Vangay P, McKinlay CE, Knights D. Multi-omics analysis of inflammatory bowel disease. Immunol Lett. 2014; 162: 62–68. https://doi.org/10.1016/j.imlet.2014.07.014 PMID: 25131220
15. Corraliza AM, Ricart E, López-García A, Carme Masamunt M, Veny M, Esteller M, et al. Differences in Peripheral and Tissue Immune Cell Populations Following Haematopoietic Stem Cell Transplantation in Crohn's Disease Patients. J Crohns Colitis. [cited 25 Jan 2019]. https://doi.org/10.1093/ecco-jcc/jjy203 PMID: 30521002
16. Tang MS, Bowcutt R, Leung JM, Wolff MJ, Gundra UM, Hudesman D, et al. Integrated Analysis of Biopsies from Inflammatory Bowel Disease Patients Identifies SAA1 as a Link Between Mucosal Microbes with TH17 and TH22 Cells. Inflamm Bowel Dis. 2017; 23: 1544–1554. https://doi.org/10.1097/MIB.0000000000001208 PMID: 28806280
17. Gevers D, Kugathasan S, Denson LA, Vázquez-Baeza Y, Van Treuren W, Ren B, et al. The Treatment-Naive Microbiome in New-Onset Crohn's Disease. Cell Host Microbe. 2014; 15: 382–392. https://doi.org/10.1016/j.chom.2014.02.005 PMID: 24629344
18. Presley LL, Ye J, Li X, LeBlanc J, Zhang Z, Ruegger PM, et al. Host-Microbe Relationships in Inflammatory Bowel Disease Detected by Bacterial and Metaproteomic Analysis of the Mucosal-Luminal Interface. Inflamm Bowel Dis. 2012; 18: 409–417. https://doi.org/10.1002/ibd.21793 PMID: 21698720
19. Lopez-Siles M, Enrich-Capó N, Aldeguer X, Sabat-Mir M, Duncan SH, Garcia-Gil LJ, et al. Alterations in the Abundance and Co-occurrence of Akkermansia muciniphila and Faecalibacterium prausnitzii in the Colonic Mucosa of Inflammatory Bowel Disease Subjects. Front Cell Infect Microbiol. 2018; 8. https://doi.org/10.3389/fcimb.2018.00281 PMID: 30245977
20. Saccenti E, Hoefsloot HCJ, Smilde AK, Westerhuis JA, Hendriks MMWB. Reflections on univariate and multivariate analysis of metabolomics data. Metabolomics. 2014; 10: 361–374. https://doi.org/10.1007/s11306-013-0598-6
21. Lovell D, Pawlowsky-Glahn V, Egozcue JJ, Marguerat S, Bähler J. Proportionality: A Valid Alternative to Correlation for Relative Data. PLOS Comput Biol. 2015; 11: e1004075. https://doi.org/10.1371/journal.pcbi.1004075 PMID: 25775355
22. Cavill R, Jennen D, Kleinjans J, Briedé JJ. Transcriptomic and metabolomic data integration. Brief Bioinform. 2016; 17: 891–901. https://doi.org/10.1093/bib/bbv090 PMID: 26467821
23. Chong J, Xia J. Computational Approaches for Integrative Analysis of the Metabolome and Microbiome. Metabolites. 2017; 7: 62. https://doi.org/10.3390/metabo7040062 PMID: 29156542
24. Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, et al. Metagenomic biomarker discovery and explanation. Genome Biol. 2011; 12: R60. https://doi.org/10.1186/gb-2011-12-6-r60 PMID: 21702898
25. Rohart F, Gautier B, Singh A, Cao K-AL. mixOmics: An R package for 'omics feature selection and multiple data integration. PLOS Comput Biol. 2017; 13: e1005752. https://doi.org/10.1371/journal.pcbi.1005752 PMID: 29099853
26. Deun KV, Mechelen IV, Thorrez L, Schouteden M, Moor BD, Werf MJ van der, et al. DISCO-SCA and Properly Applied GSVD as Swinging Methods to Find Common and Distinctive Processes. PLOS ONE. 2012; 7: e37840. https://doi.org/10.1371/journal.pone.0037840 PMID: 22693578
27. Lock EF, Hoadley KA, Marron JS, Nobel AB. Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. Ann Appl Stat. 2013; 7: 523–542. https://doi.org/10.1214/12-AOAS597 PMID: 23745156
28. Meng C, Kuster B, Culhane AC, Gholami AM. A multivariate approach to the integration of multi-omics datasets. BMC Bioinformatics. 2014; 15: 162. https://doi.org/10.1186/1471-2105-15-162 PMID: 24884486
29. Witten DM, Tibshirani R, Hastie T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics. 2009; 10: 515–534. https://doi.org/10.1093/biostatistics/kxp008 PMID: 19377034
30. Tenenhaus A, Tenenhaus M. Regularized generalized canonical correlation analysis for multiblock or multigroup data analysis. Eur J Oper Res. 2014; 238: 391–403. https://doi.org/10.1093/biostatistics/kxu001 PMID: 24550197
31. Tenenhaus A, Tenenhaus M. Regularized Generalized Canonical Correlation Analysis. Psychometrika. 2011; 76: 257–284. https://doi.org/10.1007/s11336-011-9206-8

32. Löfstedt T, Hadj-Selem F, Guillemot V, Philippe C, Raymond N, Duchesney E, et al. A general multi-block method for structured variable selection. ArXiv161009490 Stat. 2016 [cited 29 Jun 2018]. Available: http://arxiv.org/abs/1610.09490

33. Lagkouvardos I, Kläring K, Heinzmann SS, Platz S, Scholz B, Engel K-H, et al. Gut metabolites and bacterial community networks during a pilot intervention study with flaxseeds in healthy adult men. Mol Nutr Food Res. 2015; 59: 1614–1628. https://doi.org/10.1002/mnfr.201500125 PMID: 25988339

34. Berry D, Ben Mahfoudh K, Wagner M, Loy A. Barcoded primers used in multiplex amplicon pyrosequencing bias amplification. Appl Environ Microbiol. 2011; 77: 7846–7849. https://doi.org/10.1128/AEM.05220-11 PMID: 21890669

35. Klindworth A, Pruesse E, Schweer T, Peplies J, Quast C, Horn M, et al. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. Nucleic Acids Res. 2013; 41: e1–e1. https://doi.org/10.1093/nar/gks808 PMID: 22933715

36. Lagkouvardos I, Joseph D, Kapfhammer M, Giritli S, Horn M, Haller D, et al. IMNGS: A comprehensive open resource of processed 16S rRNA microbial profiles for ecology and diversity studies. Sci Rep. 2016;6. https://doi.org/10.1038/s41598-016-0015-2 PMID: 28442741

37. Edgar RC. UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nat Methods. 2013; 10: 996–998. https://doi.org/10.1038/nmeth.2604 PMID: 23955772

38. Wang Q, Garrity GM, Tiedje JM, Cole JR. Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. Appl Environ Microbiol. 2007; 73: 5261–5267. https://doi.org/10.1128/AEM.00062-07 PMID: 17586664

39. Puget S, Philippe C, Bax DA, Job B, Varlet P, Junier M-P, et al. Mesenchymal Transition and PDGFRA Amplification/Mutation Are Key Distinct Oncogenic Events in Pediatric Diffuse Intrinsic Pontine Gliomas. PLOS ONE. 2012; 7: e30313. https://doi.org/10.1371/journal.pone.0030313 PMID: 22389665

40. Morgan XC, Kabakchiev B, Waldron L, Tyler AD, Tickle TL, Milgrom R, et al. Associations between host gene expression, the mucosal microbiome, and clinical outcome in the pelvic pouch of patients with inflammatory bowel disease. Genome Biol. 2015; 16: 67. https://doi.org/10.1186/s13059-015-0637-x PMID: 25887922

41. Sparse Generalized Canonical Correlation Analysis. [cited 26 Sep 2018]. Available: http://biodev.cea.fr/sgcca/

42. Tenenhaus A, Philippe C, Guillemot V, Le Cao K-A, Grill J, Frouin V. Variable selection for generalized canonical correlation analysis. Biostatistics. 2014; 15: 569–583. https://doi.org/10.1093/biostatistics/kxu001 PMID: 24550197

43. Schäfer J, Strimmer K. A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics. Stat Appl Genet Mol Biol. 2005; 4. https://doi.org/10.2202/1544-6115.1175 PMID: 16646851

44. Tenenhaus M, Tenenhaus A, Groenen PJF. Regularized Generalized Canonical Correlation Analysis: A Framework for Sequential Multiblock Component Methods. Psychometrika. 2017; 82: 737–777. https://doi.org/10.1007/s11336-017-9573-x PMID: 28536930

45. Aleman FDD, Valenzano DR. Microbiome evolution during host aging. PLOS Pathog. 2019; 15: e1007727. https://doi.org/10.1371/journal.ppat.1007727 PMID: 31344129

46. Wen L, Duffy A. Factors Influencing the Gut Microbiota, Inflammation, and Type 2 Diabetes. J Nutr. 2017; 147: 1468S–1475S. https://doi.org/10.3945/jn.116.240754 PMID: 28615382

47. Schirmer M, Franzosa EA, Lloyd-Price J, McIver LJ, Schwager R, Poon TW, et al. Dynamics of metatranscription in the inflammatory bowel disease gut microbiome. Nat Microbiol. 2018; 3: 337–346. https://doi.org/10.1038/s41564-017-0089-z PMID: 29311644

48. Ferrer-Picón E, Dotti I, Corraliza AM, Mayorgas A, Esteller M, Perales JC, et al. Intestinal Inflammation Modulates the Epithelial Response to Butyrate in Patients With Inflammatory Bowel Disease. Inflamm Bowel Dis. 2020; 26: 43–55. https://doi.org/10.1093/ibd/izz119 PMID: 31211831

49. Hillman ET, Lu H, Yao T, Nakatsu CH. Microbial Ecology along the Gastrointestinal Tract. Microbes Environ. 2017;advpub. https://doi.org/10.1264/jsme2.ME17017 PMID: 29129876

50. Stubbendieck RM, Vargas-Bautista C, Straight PD. Bacterial Communities: Interactions to Scale. Front Microbiol. 2016; 7. https://doi.org/10.3389/fmicb.2016.01234 PMID: 27551280

51. Koren O, Knights D, Gonzalez A, Waldron L, Segata N, Knight R, et al. A Guide to Enterotypes across the Human Body: Meta-Analysis of Microbial Community Structures in Human Microbiome Datasets. PLOS Comput Biol. 2013; 9: e1002863. https://doi.org/10.1371/journal.pcbi.1002863 PMID: 23326225

52. Louca S, Doebeli M, Parfrey LW. Correcting for 16S rRNA gene copy numbers in microbiome surveys remains an unsolved problem. Microbiome. 2018; 6: 41. https://doi.org/10.1186/s40168-018-0420-9 PMID: 29482646

C.2 experDesign: stratifying samples into batches with minimal bias

Article peer-reviewed published on 2021, freely available online.

The same article can be find below.

The Journal of Open Source Software

experDesign: stratifying samples into batches with minimal bias

Lluís Revilla Sancho[1, 2], **Juan-José Lozano**[1], **and Azucena Salas**[2]

1 Centro de Investigación Biomédica en Red, Enfermedades Hepáticas y Digestivas **2** Institut d'Investigacions Biomèdiques August Pi i Sunyer, IDIBAPS

DOI: 10.21105/joss.03358

Software

- Review
- Repository
- Archive

Editor: Lorena Pantano

Reviewers:

- @abartlett004
- @stemangiola

Submitted: 23 April 2021
Published: 27 November 2021

Summary

The design of an experiment is critical to its success. Nonetheless, even when correctly designed, the process leading up to the moment of measuring a given variable is critical. At any one of the several steps, from sample collection to measurement of a variable, various errors and problems can affect the experimental results. Failure to take such variability into account can render an experiment inconclusive. *experDesign* provides tools to minimize the risk of inconclusive results by assigning samples to batches to reduce potential batch effects.

Introduction

To design an experiment that can support conclusive results upon analysis, the source of the variation between samples must be identified. Typically, one can control the environment in which the study or experiment is being conducted. Sometimes, however, this is not possible. In such cases, techniques to control variations must to be applied. There are three methods used to decrease the uncertainty of the unwanted variation: blocking, randomization and replication (Klaus 2015).

Blocking is a method that groups samples that are equal according to one or more variables, allowing the estimation of the differences between each batch by comparing measurements within the blocks. **Randomization** minimizes the variation in the measurements by randomly mixing the potential confounding variables. **Replication** increases the number of samples used in an experiment to better estimate the variation of the experiment. In some settings these techniques can be applied together to enhance the robustness of the study.

Between the designing of an experiment and the measurement of the samples, some samples might be lost, contaminated, or degraded below the quality threshold. In addition, experiments will occasionally need to be carried out in batches. The later might be needed for technical reasons; for example, the device cannot measure more than a given number of samples at the same time. Practical reasons can also be a factor; for instance, it may not be possible to obtain additional measurements in the field during the allotted time.

This divergence from the original design might cause batch effects, thereby perturbing the analysis. There are several techniques to identify and assess batch effects when analyzing an already measured experiment (Leek et al. 2010). It would be better to avoid such batch effects before executing an experiment. By taking into account the differences between the original design and the state before the measurement is conducted, confounding effects can be minimized.

To prevent the batch effect from confounding the analysis after the initial design of the experiment, there are two options: randomization and replication. Randomization, consists of

The Journal of Open Source Software

shuffling the samples in order to mix different attributes, which can help reduce variations across groups. In contrast, replication helps estimate the variation of the measurements or samples, thus increasing the precision of the estimates of the true value obtained by the analysis. Replications consist of increasing the number of measurements with similar attributes. When a sample is measured multiple times, this is referred to as a technical replicate. Technical replicates help estimate the variation of the measurement method, and thus the possible batch effect (Blainey, Krzywinski, and Altman 2014).

Randomization and replication can be used to prevent batch effects that might confound the analysis. By examining how the variables are distributed across each batch, proper randomization can be ensured, thus minimizing batch effects. This is known as randomized block experimental design or stratified random sampling experimental design.

State of the art

There are certain tools that can minimize batch effects on the R language in multiple fields, particularly for biological research (R Core Team 2014). Here we briefly describe the currently available packages:

- *OSAT*, at Bioconductor, first allocates the samples from each batch according to a variable; it then shuffles the samples from each batch in order to randomize the other variables (Yan et al. 2012). This algorithm relies on categorical variables and cannot use numerical variables (e.g., those that are age- or time-related) unless they are treated as categorical variables.
  OSAT provides templates for plates that hold 2, 4, 8 Illumina BeadChip chips, having 24, 48 or 96 wells. Moreover, it works for both numeric and categorical variables but OSAT might return less rows than the input provided because they might have NA value.
- *anticlust*, at CRAN, divides the samples into similar groups, ensuring similarity by enforcing heterogeneity within groups (Papenberg and Klau 2020). Conceptually it is similar to the clustering method k-means.
  anticlust does not handle all types of variables, it only accepts numeric variables.
- Recently, *Omixer*, a new package, has been made available at Bioconductor (Sinke, Cats, and Heijmans 2021). It tests whether the random assignments made by it are homogeneous by transforming all variables to numeric values and using the Kendall's correlation when there are more than 5 samples; otherwise, it utilizes the Pearson's chi-squared test.
  There is a bug in the Omixer that prevents it from working unless specific conditions are meet. This precluded any comparisons of Omixer with other tools using the same settings.

For completeness a description and comparison of the usage of the different software packages currently available on CRAN and Bioconductor is presented below. First, we start with some real data obtained from a survey. This data set has three variables of interest; Sex, Smoke and Age are a mix of categorical and numeric variables.

Statement of need

Current solutions for stratifying samples to reduce and control batch effect do not work for all cases. They are either specialized to a particular type of data, they omit some conditions that

The Journal of Open Source Software

are usually met, or they only work under a specific subset of conditions. The new package ***experDesign*** works with all data types and does not require a spatial distribution making it suitable for all kind of experiments. This package is intended for people needing a quick and easy solution that will provide reasonable suggestions on how to best distribute the samples for analysis.

Description

The package ***experDesign*** provides the function `design` to arrange the samples into multiple batches such that a variable's distribution remains homogeneous within each batch. Each batch is set to have some centrality and dispersion statistics to match as closely as possible with the original input design data. The statistics used are the mean, the standard deviation, the median absolute deviation, variables with no value number, the entropy and the independence of the categorical variables. With each iteration if the random distribution of the sample statistics for each batch has fewer differences vis-à-vis the original distribution than the last stored sample distribution then it replaces it as the best sample distribution. Upon completion of the iterations the best sample distribution is returned to the user.

Users can examine the following flowchart to decide what function(s) they need to use:

The Journal of Open Source Software

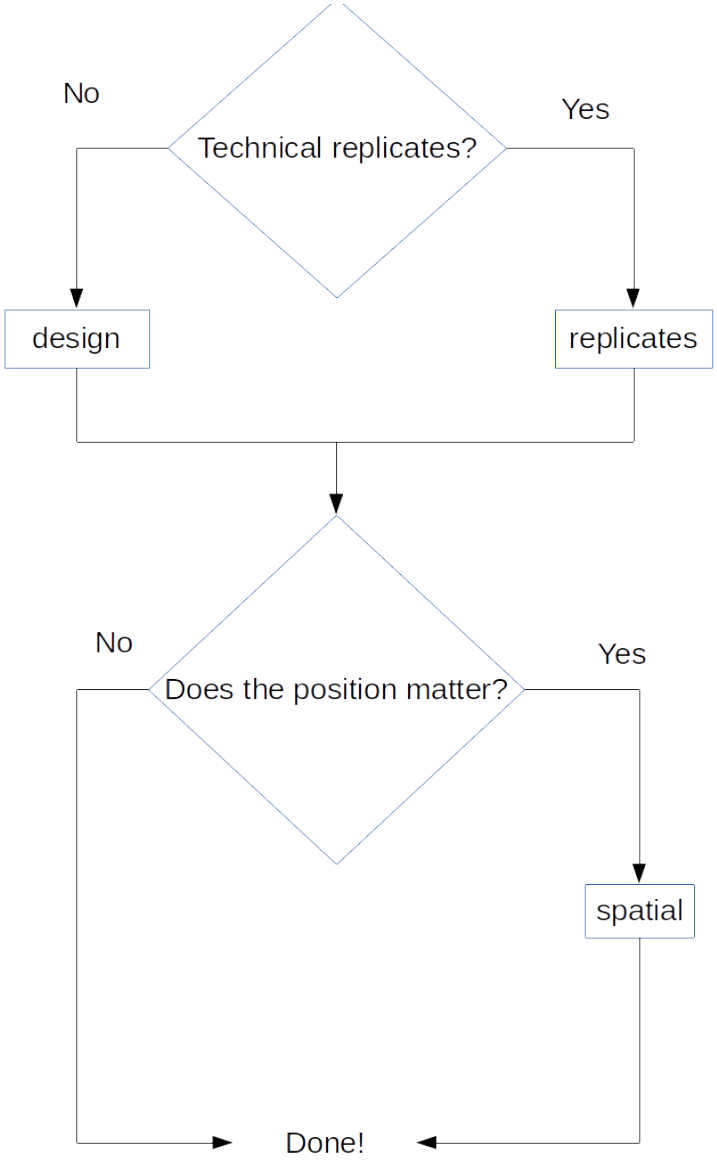

Figure 1: Flow chart to decide which functions are needed

If users want a design without replicates but the batches have some spatial distribution, we must use `design` to allocate the samples on each batch, followed by the `spatial` function to randomly distribute the samples homogeneously by position within each batch. See the example in `inspect` and the vignette:

The Journal of Open Source Software

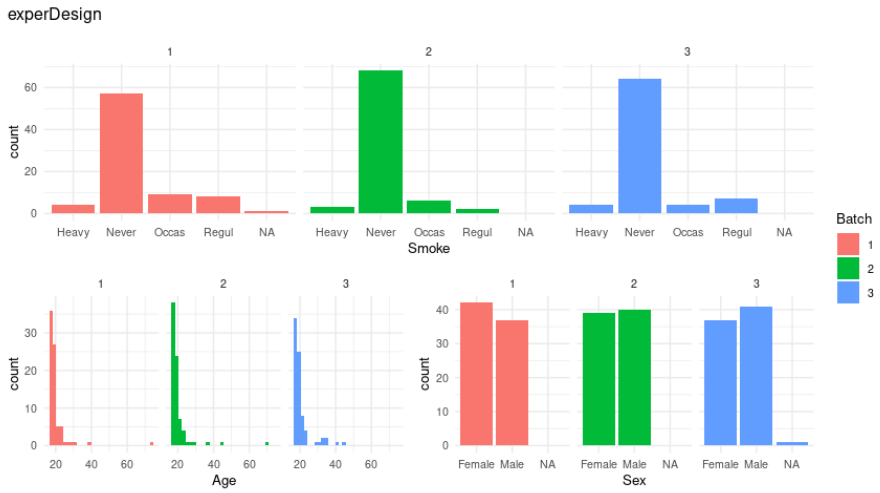

Figure 2: Example of distributions generated with ***experDesign***.

On Figure 2 one can see that the distribution index generated by experDesign uses both numeric and categorical variables keeping also the samples with missing (NA) information.

The statistics of the index can be checked for multiple statistics, as shown on the help pages of `evaluate_na`, `evaluate_entropy`, `evaluate_mad`, `evaluate_sd` and `evaluate_mean`. We can also compare our results with the original distribution via `evaluate_orig`.

In addition to distributing the samples into batches, ***experDesign*** provides tools to add technical replicates. In order to choose them from the available samples, the function `extreme_cases` is provided. For easier usage, the `replicates` function designs an experiment with the desired number of replicates per batch.

experDesign also provides several small utilities to make it easier to design the experiment in batches. For instance, a function called `sizes_batches` helps calculate the number of samples in order to distribute them across the required batches. Furthermore, `optimum_batches` calculates the minimal number of batches required. Examples of all this methods can be found on the manual page of each function and on the vignette.

In conclusion ***experDesign*** offers a fast method for preparing a batched experiment. It can use as many numeric and categorical variables as needed to stratify the experimental design based on batches including spatial distributions.

Acknowledgments

We are grateful to Joe Moore for English-language assistance and we would like to thank reviewers and the editor for taking the time and effort necessary to review the manuscript.

References

Blainey, Paul, Martin Krzywinski, and Naomi Altman. 2014. "Replication." *Nature Methods* 11 (9): 879–80. https://doi.org/10.1038/nmeth.3091.

Klaus, Bernd. 2015. "Statistical Relevancerelevant Statistics, Part i." *The EMBO Journal* 34 (22): 2727–30. https://doi.org/10.15252/embj.201592958.

Leek, Jeffrey T., Robert B. Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W. Evan Johnson, Donald Geman, Keith Baggerly, and Rafael A. Irizarry. 2010. "Tackling the Widespread and Critical Impact of Batch Effects in High-Throughput Data." *Nature Reviews. Genetics* 11 (10). https://doi.org/10.1038/nrg2825.

Papenberg, Martin, and Gunnar W. Klau. 2020. "Using Anticlustering to Partition Data Sets into Equivalent Parts." *Psychological Methods*. https://doi.org/10.1037/met0000301.

R Core Team. 2014. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://R-project.org/.

Sinke, Lucy, Davy Cats, and Bastiaan T Heijmans. 2021. "Omixer: Multivariate and Reproducible Sample Randomization to Proactively Counter Batch Effects in Omics Studies." *Bioinformatics*, no. btab159 (March). https://doi.org/10.1093/bioinformatics/btab159.

Yan, Li, Changxing Ma, Dan Wang, Qiang Hu, Maochun Qin, Jeffrey M. Conroy, Lara E. Sucheston, et al. 2012. "OSAT: A Tool for Sample-to-Batch Allocations in Genomics Experiments." *BMC Genomics* 13 (1): 689. https://doi.org/10.1186/1471-2164-13-689.

Other datasets

D.1 BARCELONA dataset

All patients with an established diagnosis of IBD, including Crohn's disease, ulcerative colitis, unclassified IBD, indeterminate colitis, or pouchitis, starting treatment with a biologic agent were monitored following the schedule of clinical visits, laboratory tests, imaging procedures and biologic sampling at the beginning of their treatment with anti-TNF therapy and after 14 weeks and 46 weeks a biopsy from an ileocolonoscopy. The protocol was approved by the Institutional Ethics Committee of the Hospital Clinic de Barcelona (Study Number HCB/2012/7845 and HCB/2012/7956).

Patients that were referred to the Hospital Clínic de Barcelona IBD unit, who had already started treatment with a biologic agent in another center, were also included adapting to the corresponding time-schedule of their treatment. In all patients, starting anti-TNF treatment will be decided before the protocol entry decision according to medical clinical practice.

Anonymized identification of the patients, disease, sex age at diagnostic, age at the moment of the sample taking, time since the start of the treatment and sample segment was collected.

Table D.1: Samples included from the BARCELONA dataset characteristics.

Characteristic	BARCELONA
Individuals	62
Status (non-IBD/CD/UC)	8/33/21
Sex (female/male)	29/33
Age at diagnostic (<17/<40/>40 years)	2/44/8
Years of disease: mean (min-max)	7.6 (0-32)
Age: mean (min-max)	41 (18-68)
Time (0/14/46 weeks)	41/40/32
Sample segment (ileum/colon)	39/87

The process of DNA extraction and sequencing was different for this dataset. We used different 16S-V3V4 primers pair 341f/806r on a MiSeq Nano sequencing by the RTSF Genomics Core at Michigan State University, United State of America. The sequence of the primers used was:

341f: 5'-CCTACGGGAGGCAGCAG-3'
806r: 5'-GGACTACHVHHHTWTCTAAT-3'

The result of MiSeq Nano were processed with bcl2fastq (v1.8.4).

This dataset was processed as usual but as part of the quality controls of the dataset the diversity measures of the samples was analyzed on D.1:

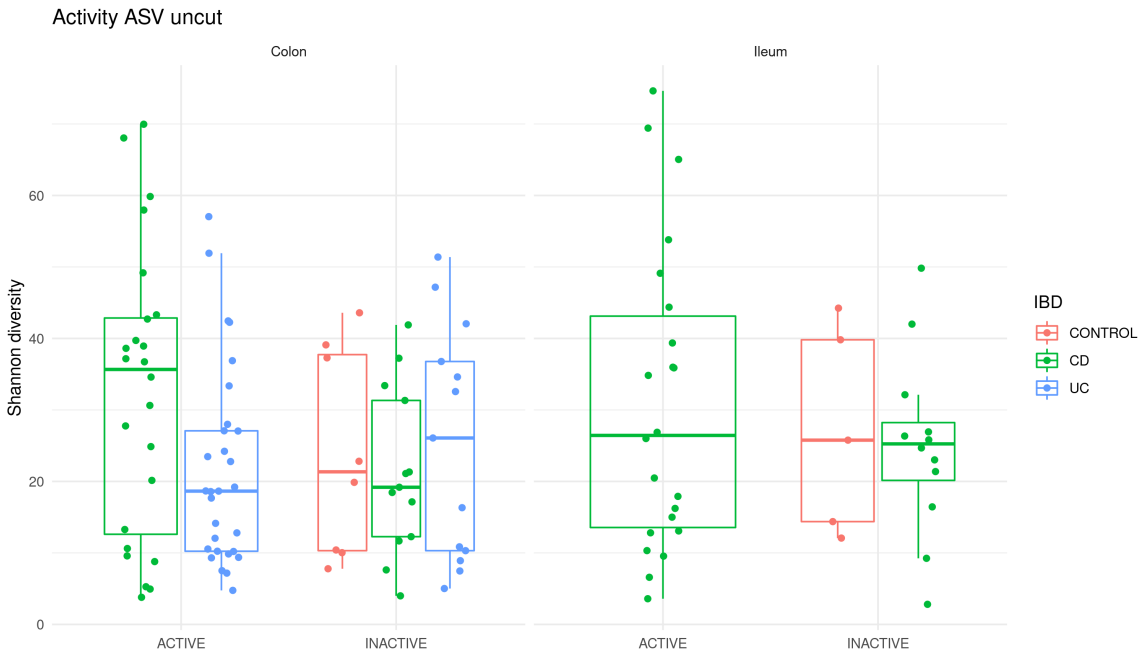

Figure D.1: Diversity indices of Barcelona according to the location and disease status. There is a lot of diversity between different groups but importantly the control samples overlap with the patients with inflammatory bowel disease.

Control samples diversity should be lower and not on the same range as with samples of patients with IBD. The dataset's 16S was sequenced several times by different platforms. Despite the pilots and the negative controls on the sequencing process, each time there were different problems: contamination, low quality and then this suspicious diversity. It does not seem to be a problem of the sequencing facility, so this data was abandoned as unreliable.

D.2 Hernández' dataset

This dataset was obtained from collaborators at Mount Sinai, Toronto, Canada [82].

Patients with UC, CD were recruited when attending regularly scheduled visits or surveillance. In addition, asymptomatic healthy controls (HC) were recruited during routine, age-related colorectal cancer screening by colonoscopy. 290 samples were collected together with information about the disease, age at diagnosis, age at the moment of the sampling, sex, sample location and smoking status.

Table D.2: Characteristics of samples included from Hernández's dataset.

Characteristic	Hernández'
Disease (non-IBD/CD/UC)	46/54/66
Age at diagnostic (<17/<40/>40 years)	29/73/18

Characteristic	Hernández'
Age: mean (min-max)	40 (17-71)
Sex (female/male)	81/85
Smoking (never/ex/current)	115/34/16
Sample Location (ileum/colon)	97/193

D.2.1 Results

It substitutes the BARCELONA dataset to confirm the results on the previous datasets. However, at the time of writing the process is not complete.

Models output

E.1 HSCT

E.1.1 Genes

E.1.2 Microbiome

E.2 Häsler

E.2.1 Genes

E.2.2 Microbiome

E.3 Morgan

E.3.1 Genes

E.3.2 Microbiome

E.4 Howell

E.4.1 Genes

E.4.2 Microbiome