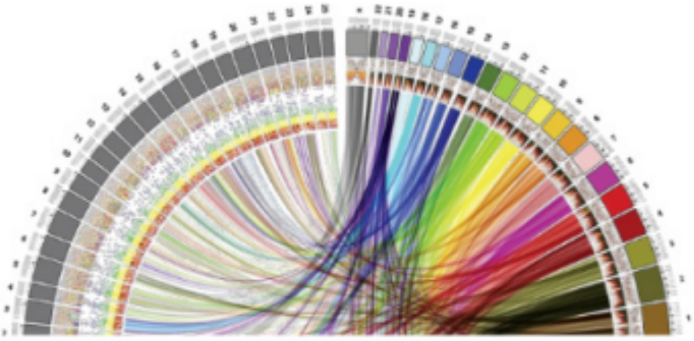


MsC in Bioinformatics

Ulcerative colitis unsupervised classification using single cell colonic data

Ángela Sanzo Machuca

supervised by

Sònia Casillas Viladerrams

Ana María Corraliza Márquez

September 2022

Ulcerative colitis unsupervised classification using single cell colonic data

Ángela Sanzo Machuca

supervised by

Sònia Casillas Viladerrams

Ana María Corraliza Márquez

September 2022

Acknoledgements

En primer lugar, me gustaría transmitir mi más sincero agradecimiento a la Dra. Azucena Salas la oportunidad de haber desarrollado este trabajo en su laboratorio. Gracias por tu confianza, por siempre estar para dar buenos consejos y estar pendiente de todo.

Me gustaría agradecer a la Dra. Ana Corraliza por la paciencia y detalle con el que ha seguido el proyecto. Gracias por escucharme, por estar siempre pendiente y el cariño con el que haces las cosas. Al Dr. Lluis Revilla, el millor profe de R i de català, gracias por ser tan paciente y explicarme las cosas una y mil veces. Inhouse bioinformatics, tengo mucha suerte de poder aprender de vosotros (¡Papel y boli!).

A Alba, Elisa, Victoria, Isa, Marisol, Iris y Miriam. Gracias por siempre estar pendiente, los consejos y echar una mano en todo lo que he necesitado. Pero, sobre todo, gracias por las risas, por el cariño con el que trabajáis y por acogerme. Sois maravillosas.

A mis padres, mi hermano Emiliano y mi familia, por su cariño, apoyo y entendimiento. A mis amigas, Marta, Ana Guillén, Ana Guerrero e Inma, por seguir creciendo juntas, aunque sea tan lejos. Estoy muy orgullosa de vosotras.

A Manuel Luna, por ser mi apoyo incondicional y esperarme siempre con los brazos abiertos. Gracias.

A la niña que fui, quien hizo de tripas corazón y me ha traído hasta aquí.

INDEX

Abstract

Ulcerative colitis is a relapsing inflammatory bowel disease affecting the gut's mucosa. The disease is characterized by rectal bleeding, abdominal pain, and diarrhea, among others, due to the damaged caused in the epithelial barrier. Its etiology remains unknown, but there are several factors that might trigger it, like the environment, genetics, and the immune system. Moreover, treatment response among patients is highly variable; even in those classified with the same severity of the disease. For this reason, a new way of classifying patients that understands patient's variability at the molecular level is needed. In this way, single-cell RNA-seq arise as a technique that can provide insights on the transcriptome of cells and help understand the way the different cell types found in the gut mucosa act in the disease.

In this dissertation we provide a proof-of-concept on the use of scRNA-seq in an unsupervised manner to stratify patients without pre-established markers. To do so, we analyzed 111000 cells from 42 samples of 28 patients, using Hierarchical k-means and Partitioning around medoids algorithms on the cellular proportions and on the pathway expression level. Validation performed through Random Forest supervised clustering algorithm showed that the hierarchical k-means algorithm performs better at both transcriptomic profiles, especially when the cell types studied were more specific. Moreover, we found out that both the cellular composition and the pathway expression are necessary for better stratification of the patients. We also found that at the cellular level, the variables that contributed the most to the clustering were those related to the gut structure, whereas at the expression level were the myeloid and stromal subsets. Finally, a Shiny web app was developed so other people in the lab could use it for their research.

Keywords: Ulcerative colitis, Single-cell RNA-seq, Unsupervised clustering, patient stratification, transcriptomic profile.

1. Introduction

1.1 Ulcerative colitis

Ulcerative Colitis (UC) is a chronic inflammatory bowel disease (IBD) characterized by the continuous and diffuse inflammation of the colonic and rectum mucosa (Feuerstein, M.D. et al, 2019). It is a relapsing disease, meaning that patients alter periods of no clinical or endoscopic manifestations (remission) with periods of active inflammation (relapse) (Raine, T. et al, 2021). This colonic inflammation can manifest as erythema, loss of vascular pattern, erosions, and the development of ulcers. Depending on the area affected, UC is further classified into the following categories (Fig. 1) (Satsangi, J. et al., 2006):

- Ulcerative proctitis: the involvement of the disease is limited to the rectum
- Left-sided UC: the involvement ranges from the rectum to the splenic flexure
- Pancolitis: the involvement extends to the splenic flexure

The classical symptoms associated with this disease are tenesmus, rectal bleeding, weight loss, and abdominal pain. Around 20% of the patients present extra-intestinal manifestations, such as anemia, arthropathy, and erythema (Magro, F. et al., 2017); negatively impacting their overall wellness. In all, this disease makes a strong burden on society (health-care costs, costly treatments, work absenteeism, etc.) (Kaplan, G.G. et al. 2015).

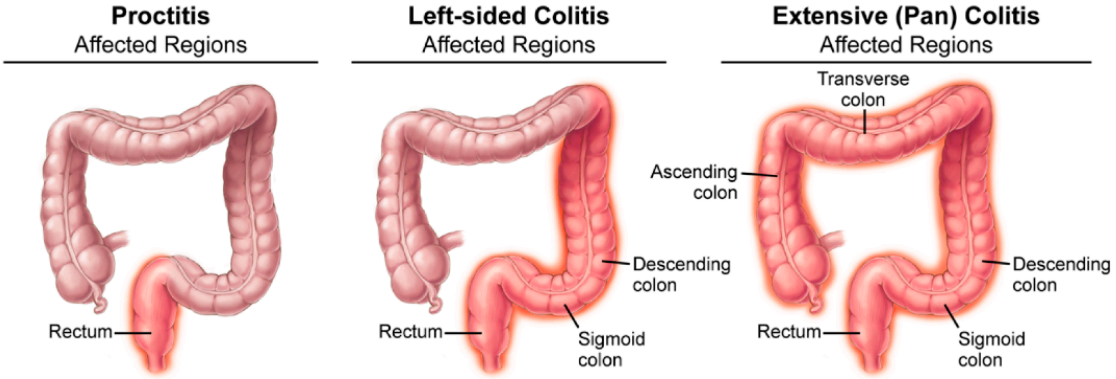


Fig 1**. Classification of Ulcerative colitis according to its extension.** The area affected can help the assessment of the disease and the patient's treatment. Image taken from (Kayal, M. et al., 2019),

1.1.1 Epidemiology

Although this disease develops at any age, it is commonly diagnosed between 15-35 years, having similar incidence among sexes. Worldwide, Ulcerative Colitis affects more than two million people in Europe and around one million in North America, but the incidence in these countries has stabilized along the years. Remarkably, the incidence increases in Asian and Latin-American countries, which could be explained by the fact that IBD has a higher incidence in industrialized areas. Environmental factors, such as changes in lifestyle, may play a relevant role in the development of IBD. However, solely these factors cannot explain the disease's pathogenesis (Du, L. et al., 2020)

1.1.2 Etiology and pathogenesis

UC is an idiopathic disease; this means its etiology has not been elucidated. However, there are several factors that have been proposed to be the drivers of the disease. There is evidence that the environment has caused changes in the incidence of the disease worldwide, but also that genetics appear to be important in the disease development (Kayal, M. et al., 2019). At the same time, the immune system is critical in UC, which explains why many treatments are currently focusing on it.

1.1.2.1 Genetic factors

Having relatives diagnosed with UC highly increases the risk of developing the disease. This risk is increased by four for first-degree relatives and by eight for siblings (Stittrich, A.B. et al., 2016). Nevertheless, studies on twins show that this risk ranges from 4% if they are dizygotic to 16% if they are monozygotic (Ungaro, R. et al., 2017).

In this sense, genetic studies have aimed to find heritable elements that would correlate to developing the disease. Genome-wide association studies (GWASs) have shed light on the genetics of UC by identifying around 163 susceptibility loci associated with IBD (Porter, R.J. et al., 2020). However, 70% of the genes are also related to other immune-mediated diseases, such as psoriasis. On the other hand, UC-specific loci are related to the human leukocyte antigen (HLA), mostly class II, in chromosome 6. Also, a new missense variant in the adenylate cyclase 7 gene (*adcy7*) is found to double the risk of UC. Moreover, a study using ChIP-seq technology has found an enrichment related to H3K27Ac in intestinal enteroids, indicating that in the disease there is an imbalance in the epithelial function at the genetic level (Mokry, M. et al., 2014)

Despite all the above, many patients do not present any genetic susceptibility to the disease, and it is estimated that around 19% of the disease heritability of the disease is explained by genetics (Chen G.B., et al., 2014).

1.1.2.2 Environmental factors

As explained before, UC’s incidence has been highly impacted by the increment of industrialization areas, as well as by westernization (Kaplan, G.G. et al., 2016). Therefore, changes in lifestyle, better hygiene, diet, sedentarism, fewer infections, and stress have been linked to UC. Nonetheless, other elements have been associated with UC for a long time, like the protective effect of smoking or getting an appendectomy before 20 years of age.

1.1.2.3 Gut microbiota

The gut microbiome consists of a diverse myriad of microorganisms found in the human digestive tract and plays a major role in the organism. Thousands of bacteria species in the gut are thought to be involved in metabolic, physiological, nutritional, and immunological functions (Guinane, C.M. et al., 2013).

It is known that alterations in the microbiota may occur due to exposure to environmental factors, like diet and drugs, and to genetic factors that module it. When these alterations happen, it can end up in an imbalance of the microorganisms found in the microbiota called dysbiosis, where there is a depletion of protective bacteria that causes an expansion of the pro-inflammatory ones. Normally, the innate and adaptative immunity prevent harmful microorganisms from proliferating. However, in a dysbiosis state, the immune response is overstimulated and eventually could cause a pathogenic condition in the host. Moreover, harmful bacteria release toxins that produce changes in the intestinal mucosa permeability, damaging it. If the epithelial function is compromised, more external agents could further worsen the injury and the inflammatory response. Still, whether gut microbiota is the triggering player of UC or not remains unclear. What is more, studies have not found any pattern that could correlate the transcriptional activity of UC to the microbiome (Moen, A.E. et al., 2018).

1.1.2.4 Epithelial dysfunction

The epithelium is the first protective barrier found in the gut. This physical barrier is composed of different cell types, including enterocytes, goblet cells, and enteroendocrine cells. These cells regulate the barrier’s permeability, which is known to be augmented in UC patients. This fact is probably due to the decrement of the mucus layer found between the epithelial cells (Turner J.R., et al., 2009). This dysfunction at the epithelial barrier can happen on account of impaired secretions produced or physical defects. Moreover, epithelial cells present receptors recognized by the immune system (like toll-like receptors or nod-like receptors). Then, when there is active inflammation, as in UC relapse, the integrity of the epithelium seems compromised because of the immune system.

Besides, *knock*-out model of *muc2* (mucine 2 gene, that produces the mucus layer in the epithelial cell) produces colitis in mice and is one of the model organisms used to study the disease (Van der Sluis, M. et al., 2006). Because of all the stated, some drugs have been developed to prevent this damage to the epithelium, like mesalamine.

1.1.2.5 Immune response

The immunological response can be classified as innate and adaptative. The innate immune system recognizes pathogens and produces an unspecific response, where the main cells involved are the natural killers (NK), eosinophils, basophils, monocytes, macrophages, and dendritic cells. Out of them, NK, macrophages, and dendritic cells are the ones that initiate the immune response against pathogens by the pattern-recognition receptors (PRRs), which are capable to recognize specific molecular patterns (Ordás, I et al., 2012). Once recognition happens, there is an activation of cytokines and chemokines, that modulate the immune cells. Then, macrophages and dendritic cells display antigens to the adaptative immune system, making them known as antigen-presenting cells (APCs).

On the other hand, the adaptative immune response main players are the lymphocytes T, which release modulator cytokines, and lymphocytes B, which produce antibodies. This response is characterized by being antigen-specific and by inducing immunological memory.

In UC, it has been demonstrated that there is an infiltration of neutrophils in the colonic epithelium, provoking the changes in the epithelial barrier's permeability mentioned (Brazil, J.C., et al., 2013, Porter, R.J. et al., 2020). This alteration promotes inflammation in the epithelium, enhancing neutrophils' survival and tissue damage due to the release of pro-inflammatory molecules. In fact, this increment in the number of neutrophils produces higher levels of the heterodimer S100A8/S100A9, denominated calprotectin, which is used as a clinical parameter to determine the level of inflammation in patients. These neutrophiles' infiltration in combination with monocytes, create a pro-inflammatory environment that induces a pathological state of the adaptative immune response, by releasing cytokines like IL-1 and TNF-α. Moreover, this environment influences the function, phenotype, and survival of new monocytes, diminishing the possibility of restoring the gut's architecture and homeostasis.

Since some UC-specific susceptibility loci are related to the HLA class II, this could also explain the pathogenic phenotype. The human leukocyte antigen of class II is a cell-surface protein complex that presents antigens to the lymphocytes T, mainly found in APCs. If there is an aberrant presentation of commensal bacteria or self-antigens, it would produce a T-cell activation that would lead to a pathogenic state (Graham D.B. et al., 2018).

1.1.3 Disease severity assessment

Activeness of the disease is defined by different indexes depending on the clinical assessment performed. One of the most used ones is the Mayo index (Schroeder, K.W. et al., 1987), which combines clinical and endoscopic parameters. Clinically, it evaluates frequency of stools, rectal bleeding, and abdominal discomfort, among others. Endoscopically, findings like erythema, vascular pattern and friability are explored. Then, the overall assessment by the clinician is considered. This index helps to evaluate the severity of the disease. Besides being a useful tool for clinical track of the disease, this score can help evaluating the responsiveness of a patient to a treatment.

However, this activity score does not consider the extent of the inflammation, which can change during the development of the disease. Since the goal in treating UC is the mucosal healing, the Modified Mayo Endoscopic Score (Lobatón, T. et al., 2015) has been developed, which does not only consider the extent of the disease but the level of inflammation. This index ranges 0-3 points, being a score of 0 defined as a segment of normal or inactive disease, a score of 1 when there is a decreased vascular pattern and erythema, a score of 2 when these sings worsen; and a score of 3 if ulcerations and bleeding appear.

1.1.4 Bowel architecture structure

Histologically, the bowel is composed of four layers: mucosa, submucosa, *muscularis propria,* and *adventitia*. In UC, only the innermost layer, mucosa, is affected. The mucosa is made from three parts: epithelium, lamina propria and *muscularis mucosae.*

Transcriptomic analysis of the mucosa in UC has classified the subsets of cells found on this layer to be wired differently, according to the health of the individual. Healthy mucosa is characterized by interactions between epithelial, fibroblasts (stromal cells) and T cells, that maintain the bowel's homeostasis. On the opposite, inflamed mucosa shifts the interactions towards macrophages, B cells and T cells, which are the principal players in adaptive immunity (Smillie C.S., et al., 2019).

It has been elucidated that cell profiles are different when comparing inflamed, non-inflamed, and healthy controls. Cell types found in the mucosa can be then classified into five subsets of cells:

- The epithelial cells have an important role in the absorption of nutrients and protection of the gut as the first immune barrier. Differentially expressed genes demonstrate that during inflammation epithelial cells activate downstream pathways related to restoring homeostasis in the gut.
- Stromal cells: give support and structure to the epithelium. During inflammation, these cells induce tissue vascularization as well as they interact with local immune cells.
- Immune cells: myeloid, T and B cells. These cells are involved in the innate and adaptive immune response explained before.

1.1.5 Treatment

Treatment options in UC classically have aimed at clinical remission and the amelioration of the symptomatology. However, new strategies focus on getting endoscopic remission, mucosal healing, and a better quality of life for the patients, with reduced adverse effects. Until biological agents were developed, physicians would opt for a step-up process of amino salicylates, steroids, and immunosuppressors; due to their effect in controlling the inflammation (Ungaro, R.et al.,2017).

Aminosalicylates, also known as mesalamine, reduce inflammation by targeting prostaglandins. They can either be taken topically or orally, and it has shown double anti-inflammatory effect than placebo and rectal steroids (Kucharzik, T. et al. 2020). Normally, if it fails to induce remission, it is combined with topically administered steroids. In addition to 5-ASA, steroids are the primary option for patients with a severe acute relapse. Since they present strong adverse effects that include hormonal and kidney failures, they are not used as maintenance therapy. Thiopurines are immune suppressants that inhibit cell growth of lymphocytes T and have shown better success rate than steroids and can be used in long-term therapy (Chande, N. et al., 2015).

As stated, the discovery of biological agents has inverted the treatment pyramid due to their success in getting patients into remission, and now they are becoming the primary treatment option. Anti-TNF monoclonal antibodies (Infliximab and Adalimumab) prevent TNF molecules from binding to the cells’ receptors, reducing inflammation and granuloma formation. However, around 30% of patients have a primary failure to this treatment and do not get into remission. Also, another 30-40% fail to respond after a year.

Other biological agents are: Vedolizumab, which targets integrinα4β7, especially relevant in T cell activation; Ustekinumab which targets IL-12 and IL-23 cytokines that share the p40 unit making it possible to target different JAK-STAT pathways at once; and Tofacitinib, a JAK inhibitor used for patients that are refractory to anti-TNF therapies (Teng, M. W. et al., 2015). These biological therapies have helped reduce flare-ups, hospitalization, and surgical interventions.

Clinical parameters and signs are used to classify patients according to the severity of the disease. This classification is also used for treatment selection. However, the patient's response to them is highly variable, and this variability increases if we consider the different manifestations of the disease. Due to the heterogeneous course of the disease, new classification strategies are needed that can predict treatment's response and, eventually, improve patient's quality of life (Lai, L. et al., 2022).

1.2. Single-cell RNA-sequencing

Classical techniques like immunostaining, flow cytometry or mass cytometry (CyTOF) have identified and characterized the different cell types found on the gut. However, these techniques lack in resolution as they rely on the use of characterized well-established cell markers limiting our ability to identify new subsets or to understand each subset's contribution to the disease. Indeed, despite major advances in our knowledge and treatment of UC, we still have a very limited understanding on how each specific cell subset is individually contributing to disease and how it is regulated by different treatment strategies.

On the other hand, "omics" techniques have revolutionized biology and medicine but are commonly applied to heterogeneous tissues or bulk sorted cells, providing a low-resolution blended picture of what is going on. Moreover, whole tissue transcriptomics cannot reliably identify markers of disease progression/phenotype or predictors of response, as evidenced by the fact that none of the studies reporting such markers have translated into the clinics.

In an attempt to extract cell type-specific information from gene expression data obtained from heterogeneous tissue samples, deconvolution tools have been developed, where an estimation of the relative cellular abundance within a specific tissue can be estimated. Nonetheless, these tools are limited by the fact that they work on information obtained from cell lines or non-tissue specific cell types. Thus, current deconvolution tools cannot provide gut cell specific transcriptional techniques. Given all the stated above, techniques with higher resolution that understand tissue composition and behavior could help understand the underlying mechanisms of the disease.

In this way, Single-cell RNA-sequencing examines transcriptomes of cells at the individual level, providing new layers of information that help understand the heterogeneity of cells within a population. For instance, it is useful to identify new cell populations that otherwise would be undetected as well as elucidate the relationships among these populations in different conditions. For this reason, scRNA-sequencing can provide new insights into the pathogenesis of complex diseases (Haque, A. et al., 2017).

To perform scRNA-seq, first cells from the tissue need to be isolated and lysed to capture RNA molecules. To analyze mRNA molecules, specific primers are used, and then, mRNA is converted to complementary DNA by reverse transcription, and unique molecular identifiers are added to create cDNA libraries. There are numerous scRNA-seq methods, being Smart-seq 2 and 10X Genomics Chromium (10X; 10X Genomics, Pleasanton, CA) two of the most used ones. The main difference between them is how the cells are separated and processed. Smart-seq 2 (Switching Mechanism at 5' End of RNA template) is a cell cultured-based approach (Baran-Gale, J. et al., 2018), where cells are separated using fluorescence-activated cell sorting (FACS) and then, placed on well plates filled with Triton-X100 and ribonuclease inhibitor to stabilize RNA. On the other hand, 10X Genomics Chromium is microfluidics-based approach that captures single cells and combines them with all the needed reagents to perform reverse transcription and build barcoded libraries using a microchip. This process encapsulates single cells in oil droplets together with single gel beads containing barcoded oligonucleotides on addition to the rest of reagents needed. The result is cDNA libraries in which all molecules will contain the same barcode, thus making possible to track back the cell of origin of each of the reads after sequencing. Smart-seq-2 protocol allows to analyze more genes per cell, however, the number of cells on each experiment can vary depending on the well's size. With 10X Genomics the number of cells is not limited by the well's size and multiplexing of the samples can be performed to maximize the insights from a single experiment.

2. Hypothesis and objectives

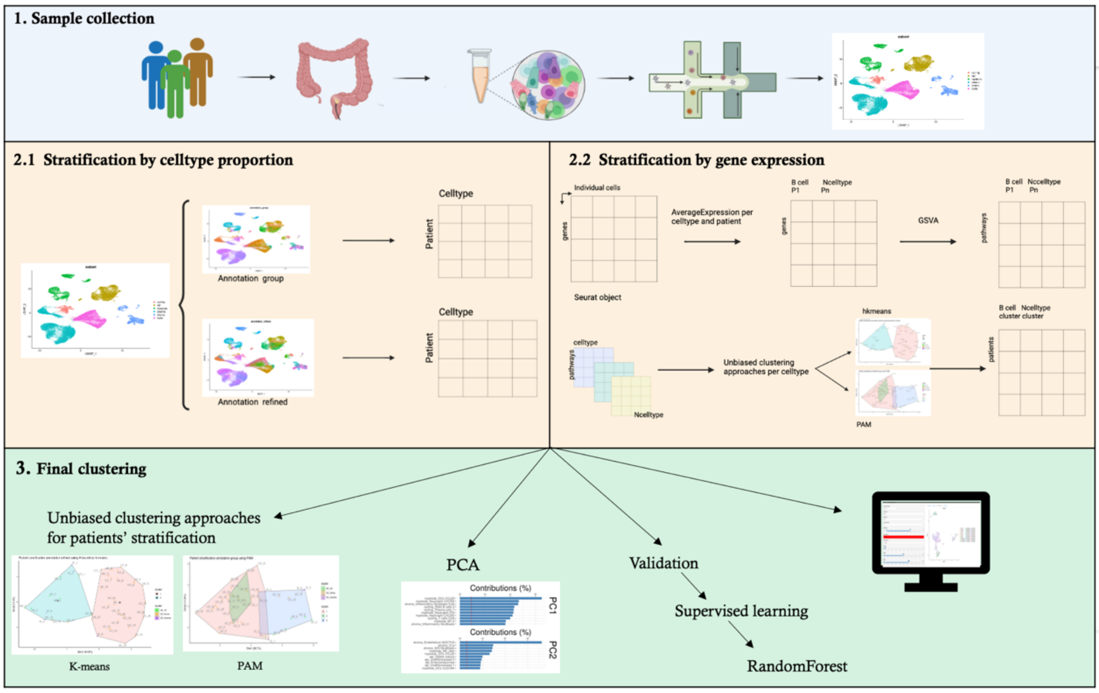
Ulcerative colitis traditional approaches classify patients according to clinical and endoscopic parameters to asset their treatments. However, their response to the available treatments is highly variable, even in patients classified within the same category. For this reason, a more accurate patient stratification method that considers a molecular basis in this multifactorial disease is needed (Selink, K. et al., 2021). Doing so would improve patients' lives and avoid unnecessary adverse effects, even surgery.

Since the molecular heterogeneity of the disease remains unknown, approaches that consider this variability can help patients' stratification, predict treatment outcomes, and find new drug targets. In this way, scRNA-seq emerges as a technique that can provide the full transcriptome of the cells found in the gut mucosa with higher resolution than classical technologies. Previous approaches have tried to classify patients and get new insights into UC. However, we still have very limited knowledge of how the cells involved are contributing to the disease. Using this new level of granularity of scRNA-seq, new players of the disease could be identified that could explain the differences among patients (Corridoni, D. et al.,2020).

On the other hand, unsupervised clustering techniques can help the unsupervised stratification of patients without focusing on the previously established markers. In this way, these techniques could uncover new players in the disease in an unbiased manner. However, since this has not been proven yet, the objectives of the project are:

- Providing a proof-of-concept of unsupervised clustering techniques in Single-cell RNA-seq sequencing from UC and healthy patients.
- Determine the best approach to assess the samples' heterogeneity using Single-cell RNA-seq transcriptomics
- Compare different unsupervised clustering methodologies and validate which one represents the known variability of the Single-cell RNA-seq dataset.
- Develop a Shiny app of the results obtained that can be used by other team members for their research purposes

3. Materials and methods



3.1 Sample collection and intestinal cell isolation

42 samples from 28 patients have been gathered for this dissertation. They come of different studies who had signed informed consent for research purposes. Out of 28 patients, 6 samples are from healthy subjects who underwent a colonoscopy for gastrointestinal symptoms, not related to IBD nor they presented any mucosal lesions, or from routine screening. On the other hand, 22 patients that suffered from UC, needed an established diagnosis of at least three months of duration. Then, follow-up biopsies from these patients were collected at different weeks for endoscopic assessment. More information on each sample can be found in Table 1 and in Table 1 of the Apex.

	Healthy controls	**Active UC**	**Inactive UC**
N samples	6	31	5
*Age**	62 (51-66) years	36 (22-55) years	34 (22-47) years
Gender	2/2/2	19/9/3	2/3/0
Segment	6/0/0/0/0/0	15/10/1/1/1/3	4/1/0/0/0/0
Mayo index	NA	0/0/1/30/0	2/1/0/0/2
Treatment	NA	7/5/7/20	0/0/0/3
Disease location	NA	7/8/12/4	0/0/2/3
Age at diagnosis	NA	4/17/5/5	0/2/1/2

Table 1. **Samples clinical information.** N samples: number of samples used in this dissertation. The following categories were added including the information provided when possible. * Age: Range values of age and the mean. Gender: Male/Female/Pending confirmation. Segment: Sigma/Rectum/Transverse/ Rectum-sigma/Descendent colon/Pending confirmation. Mayo index: 0/1/2/3/ Pending confirmation Treatment: Mesalamine/Steroids/Azathioprine/Monoclonal antibody, more than category is possible per each patient. Disease location: Left-sided colitis/Proctitis/Pancolitis/Pending confirmation. Age at diagnosis: <= 16 years/17-40 years/ >40 years/ Pending confirmation.

Once the sample was collected, it was immediately placed in cold Hank's Balanced Salt Solution (HBSS) (Gibco, MA, USA) and kept at 4ºC until processing. Then, they were washed in HBSS with 5mM DTT (Roche, Spain) for fifteen minutes, and then washed in complete medium (RPMI 1650 medium. Lonza, MD, USA) supplemented with 10% heat-inactivated bovine serum (Biosera, France), 100U/ml penicillin, 100 U/ml streptomycin and 250 ng/ml amphotericin B (Lonza), 10µg/ml gentamicin sulfate (Lonza) and 1,5mM Hepes (Lonza). Biopsies then were chopped and placed into 1.5 mL tubes containing 500 µl of digestion solution (CM + Liberase TM (0.5 Wünsch units/ml) (Roche, Spain) + DNase I (10 µg/mL) (Roche)) and incubated on an orbital shaking platform for 1h at 250 RPM at 37ºC. After digestion, the content of the tube was filtered through a 50 µm cell strainer (CellTrics, Sysmex, USA), washed with Dulbecco's Phosphate Buffered Saline (PBS; Gibco), and resuspended in FACS buffer (PBS containing 2% FBS) (Veny et al, JCC 2020).

3.2 Single-cell RNA-seq

3.2.1 10x library preparation and sequencing

Following digestion, 10x Genomics 3′ mRNA single-cell method was used. approximately 7,000 cells were loaded onto the Chromium10x Genomics platform to capture single cells, as described in the manufacturer's protocol. Generation of gel beads in emulsion (GEMs) and barcoding and GEM-reverse transcription was performed using the Chromium Single Cell 3′ and Chromium Single Cell V(D)J Reagent Kits from 10x Genomics (user guide, no. CG000086) according to manufacturer's guidance. Full-length, barcoded cDNA was amplified by PCR to generate enough mass for library construction (Nextera® PCR primers). Sequencing of the libraries was performed on HiSeq2500 (Illumina). Once sequencing is done, 10x single cell reads are processed using CellRanger software, whose outcome are feature-barcoded matrices.

3.2.2 Data processing and quality controls

Once obtained the dataset processed by CellRanger, we can follow a pipeline that allows us to get the features (gene information), cell counts, and barcodes associated with each cell of the samples. Then, together with patient's clinical information, all the samples are merged using the SeuratObject R package (version 4.0.2). After that, duplicates are removed, and low-quality cells are filtered based on their mitochondrial RNA levels. Data is normalized, and principal component analysis (PCA), as well as dimensionality reduction (UMAP), are performed. Clustering analysis is performed using the Leuven grouping algorithm, and thanks to the FindVariableFeatures() function we could identify the five major subset of cells in the gut's mucosa: epithelial, stromal, myeloid, T and B cells. Also, the cycling subet, a type of cells only found on patients recently in the literature using scRNA-seq (Smillie C.S., et al., 2019). Then, each subset was processed individually.

Inflammation and cell isolation are stressful events for the cells that can challenge their viability. For this reason, each subset underwent different quality control to preserve a sufficient cell number to perform the analysis. All cells from the subsets that had less than 200 unique features were removed. Also stromal, myeloid, and cycling cells that had more than 25% mitochondrial genes were filtered, as high mitochondrial percentage is found on low quality and dying cells. At the same time the epithelial cells that had more than 65% mitochondrial content were filtered out too . Then, subset-specific markers are used to filter cells that do not belong to it. Subsets are normalized, and the 2000 most variable features are selected. With them, k-nearest neighbor graph is calculated based on the Euclidean distance in PCA space, allowing the clustering of the cells using the Leuven algorithm (Blondel, V. et al., 2008).

3.2.3 Batch correction

Single-cell results can be influenced by the way each sample is processed, and it can concur into a batch effect in data integration. For this reason, we used Harmony package (version 0.1.0) (Korsunsky, I. et al., 2019) for sample integration. Harmony performs PCA for dimensionality reduction where it iteratively removes the batch effect selected, in our case the samples.

3.2.4 Celltype annotation

As mentioned, each subset was annotated (epithelial, stromal, cycling, myeloid, B and T cells), then, each subset is individually annotated in what is called 'annotation refined' which is a type of annotation where cell types are gathered under known markers in as many biologically relevant categories as possible. From there, a second annotation called 'annotation group' is carried out, with less granularity because there are fewer categories but with more cell types in them (Fig. 2). This kind of annotation is useful for some statistical tests.

From these two annotations we can asset which kind of annotation is more useful to perform the patient's stratification and identify if this granularity provides more insights into the disease. Finally, the different subsets were merged into a Seurat Object, PCA and UMAP were calculated, obtaining a clustering of the combination of the samples.

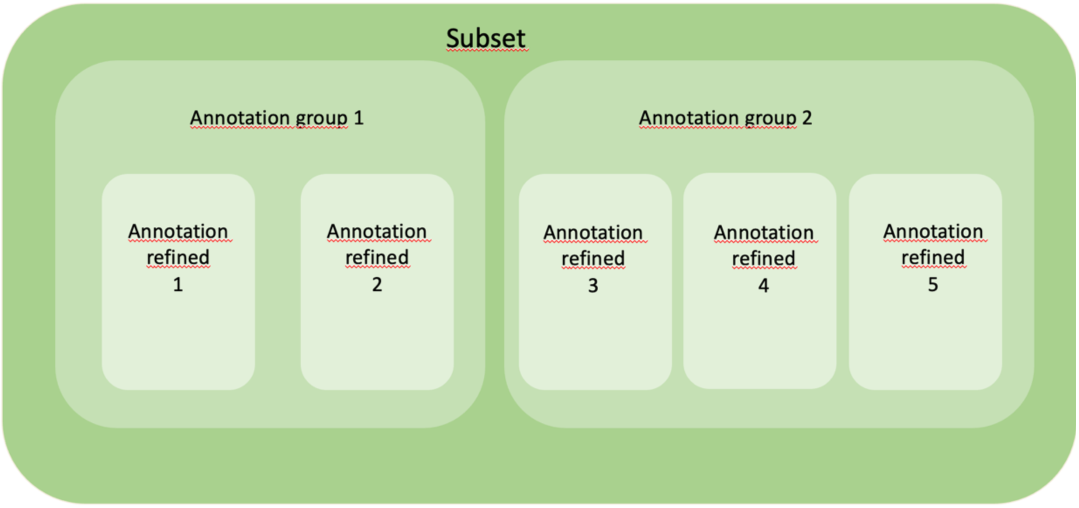


Fig. 2 **Annotation levels according to their granularity**. The outer classification is each one of the six subsets. Within each of them we find the annotation group, that covers multiples annotation refined categories, which are the ones with most granularity.

3.3 Ulcerative Colitis stratification

3.3.1 Clustering methods

Since we want to classify patients in an unsupervised manner, partitioning clustering is used to classify the observations of our dataset based on their similarities. In our model, we compare the outcome of two different clustering algorithms to check differences between them at classifying patients, or if there are batch effects in the data that we should consider. The clustering methodologies used are Hierarchical k-means and Partition around medoids (PAM).

Hierarchical k-means is a hybrid method that combines k-means and hierarchical clustering. As k-means is highly sensitive to the initial selection of clusters, this approach avoids that by using hierarchical clustering. It works by computing hierarchical clustering on the dataset and cutting the tree into k-clusters. Then, it computes the mean of each cluster and solves it by computing k-means using the set of cluster centers as the initial cluster center.

To define the k to use, we used the NbClust package (version 3.0.1) which determines the best number of clusters to perform the analysis. It uses around 30 statistical indices and proposes the best k based on the quality of the clustering.

On the other hand, PAM is not sensitive to outliers like k-means and can be a robust alternative clustering approach. PAM works by looking for representative objects or medoids in the dataset. Then, it assigns each observation to the nearest medoid. After that, it searches if any object of the cluster decreases the average dissimilarity coefficient, and the highest is selected as the medoid of the cluster. It keeps repeating until the medoids do not change. The number of clusters was selected in the same way.

For both, the final clustering was visualized using the Factoextra package (version 1.0.7)

3.3.2 Stratification by cell proportion

To understand the variability between patients and healthy individuals, a cell type composition analysis was performed. In this way, we could identify what cells are most important to classify patients and compare the result obtained with the clustering done by the gene expression and check if the depth of the annotations is important for the stratification.

To do so, we calculated the proportion of cells per patient. In this way, data is standardized, and we obtain a matrix where rows are the individuals and columns are the variables where no missing values are found.

3.3.3 Stratification by gene expression

Single-cell technologies provide further insights into transcriptomics that could help understand the heterogeneity of the pathology. The normalized data from our Seurat object is a matrix of transcriptomic expression whose columns are the cells, and the rows are the genes. From this, AverageExpression() function from Seurat calculates the average gene expression on each cell type per patient. In this way, the matrix would have cell types per patient as columns and genes as rows.

Gene Set Variation Analysis (GSVA) is a non-parametric, unsupervised gene set enrichment method that calculates gene set expression from the gene expression matrix. By doing so, information on pathway enrichment can be obtained. To perform this method, we used the GSVA package (version 3.15), using the previous matrix as input. At the same time, gene sets to perform the analysis were obtained from the Molecular Signatures Database (MSigDB, v7.5.1). Specifically, the gene sets used were the biological processes from the ontology gene sets, and the ones from the canonical pathways. Previous papers in the literature state that pathway database choice is extremely relevant for the resulting output (Mubeen, S. et al., 2019, Zhang, C. et al., 2021). For this reason, this integrative database is chosen (Emert-Streib, F. et al., 2011). Also, it is known that equivalent pathways that come from different databases can provide divergent results. To prevent this and an over-dimensionality of cell functions, gene sets that were similar in 90% of the genes were filtered out.

After performing GSVA, we obtained a matrix where columns are the cell type per patient and the rows are the pathways. Next, we created individual matrixes per each cell type (1:N cell type). From here, we performed the unsupervised clustering (both hierarchical k-means and PAM) per each cell type. Then, we created a matrix where columns belonged to cell types and rows to patients. Finally, we performed the unsupervised classification on this matrix that gave us the final stratification per patient.

3.3.4 Contribution variables

Principal Component Analysis is a dimensionality-reduction method that is used to transform a large set of variables into a smaller one that still holds most of the information in the large set. The function fviz_**c**ontrib() from the Factoextra package allows to visualize the contribution of the columns (cell types) to the result of a PCA analysis. In this way, we can check the variables that are taking more variability in the dataset and, hence, are affecting more to the final clustering result.

3.3.5 Model validation process

Supervised machine learning classificatory algorithms use training sets to assign test data into specific categories. From this training set it recognizes patterns and try to understand how the data tested can fit in them and label the data according to it. One of the most used is Random Forest (RF), a non-parametric model that builds decision trees on different samples and counts the majority vote to classify the data tested (Saric, A. et al., 2017). To perform this analysis the RandomForest package was used (version 4.7-1.1).

The way RF works is by randomly dividing the data set in two, 30% of the samples go to the test set and 70% to the training set. RF training algorithm applies bootstrapping technique to construct the tree and decide the classification. Since the dataset is relatively small and the samples do not overlap, the error rate of the classification is obtained from the *out-of-bag* error (OOB error). To keep the OOB error low, the strength of each tree needs to be increased. For this reason, the number of random variables used on each tree (mtry) and the number of trees used in the forest (ntry) need to be adjusted to be optimal. After optimizing the values, the RF is run and the OOB is observed to find the rate at which it stabilizes and reaches its minimum.

Then, RF allows to calculate which variables help a better prediction of the classification, this can be asset in two ways. First, by the Mean Decrease Accuracy, which is how much the model accuracy decreases if that variable is removed. Second, by the Mean Decrease Gini, which measures the variable importance according to the Gini impurity index.

3.4 Data accessibility

The analysis code is available on GitHub (https://github.com/ibd-bcn/ibd-bcn_all_colitis). Also, an applicative Shiny has been developed for this project, as it helps display the results obtained. However, the code associated with the Shiny app cannot be shared due to internal protocols.

4. Results

ScRNA-seq analysis of colonic biopsies from healthy controls (HC) and active Ulcerative Colitis (UC) patients identified 112 cell types in 111000 cells. As described in table 2, the analysis of the samples resulted in the differentiation of six subsets: Epithelial, Cycling, Myeloid, Plasma, Stroma and T cells (Fig. 3). Within these subsets, 112 cell types could be identified due to known markers, and then, a second annotation with les granularity, called 'group' is identified.

Subset	**N cells**	**Annotation refined**	**Annotation group**
Cycling	3110	12	5
Epithelial	20199	21	13
Myeloid	14333	22	11
Plasma	38072	14	3
Stroma	9284	21	14
T cells	26002	22	7

Table 2. **Final subset annotated**. Different subsets that were identified during the quality control of the samples, as well as the number of cell types found on each annotation.

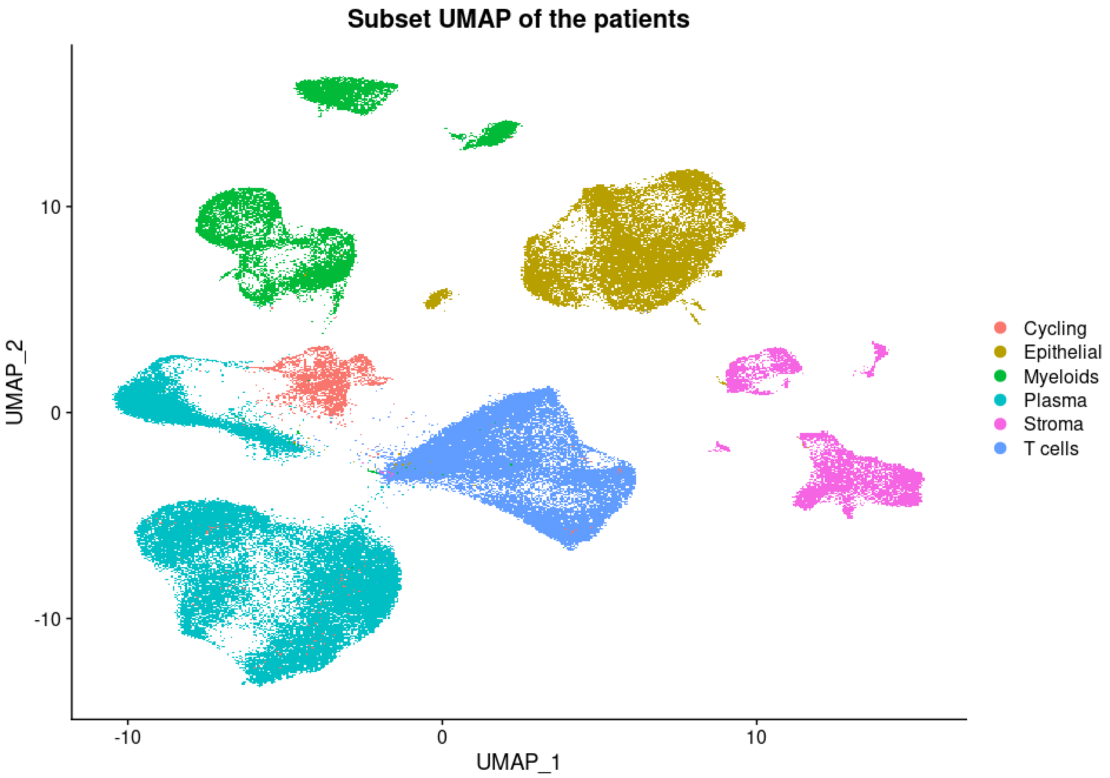


Fig. 3. **Single-cell object dimensional reduction plot**. UMAP result of the samples after the quality control and integration pipeline, where each dot represents a cell.

4.1 Stratification by cell proportions

Cell type proportions showed differences among the patients. We grouped cell types in a larger annotation (annotation group) to help us understand different patterns between health status. For instance, inflammation damages epithelial cells, which is reflected in the annotation group category (Fig 4 a), where we see that samples from healthy individuals happen to have more colonocytes and other cell types implicated in the gut's epithelial barrier than the patients that present an active state of the disease. Moreover, the patients that are in an inactive state recover the proportion of epithelial cells, meaning that there is a recovering of the mucosa of these patients.

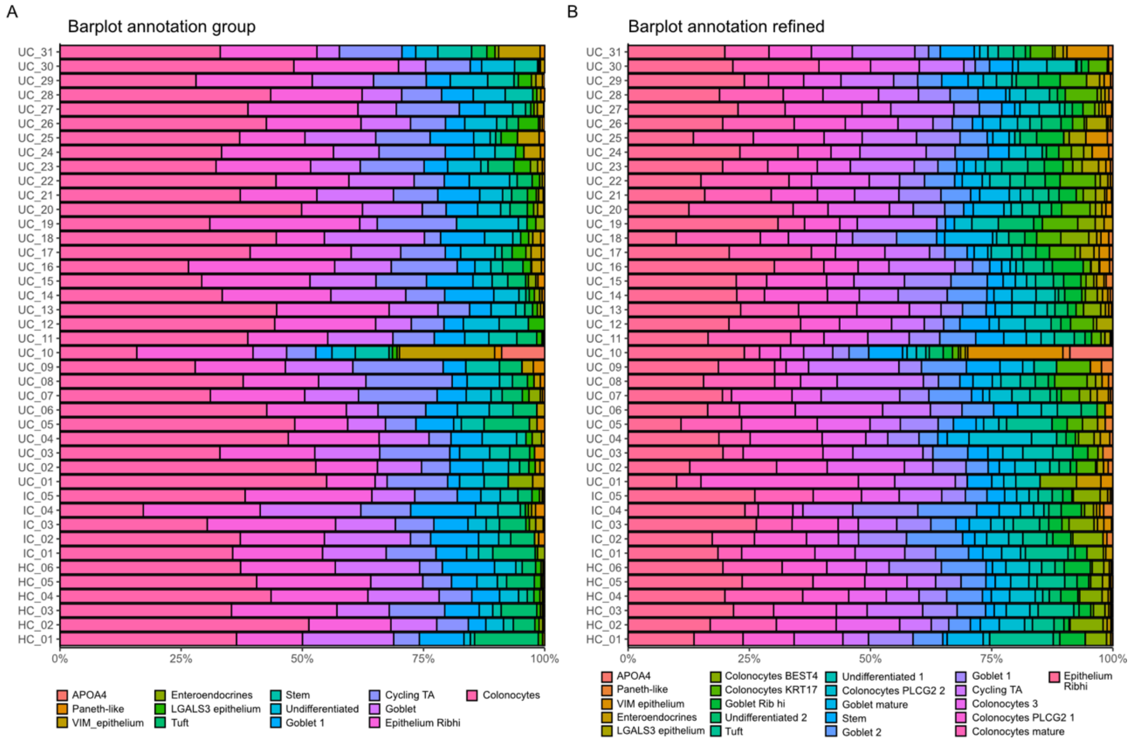


Fig 4. A. **Proportions per annotation group**. Barplot of the proportion of each cell type by this annotation per patient. B. **Proportions per annotation refined**. Barplot of the proportion of each cell type by this annotation per patient.

4.1.1 Clustering on annotation group

NbClust() function determined that 3 was the appropriate number k to perform the Hierarchical k-means by the rule of the majority. When plotted, neither of the three clusters was clearly defined by health status. Most healthy controls would be in cluster 3. However, 5 samples from active colitis are present there too.

PAM clustering algorithm also selected three as the optimal k. However, this approach also failed to classify patients according to health status. Although it improved how healthy and inactive patients were clustered, samples were mixed in clusters two and three, which did not explain the inter-patient variability

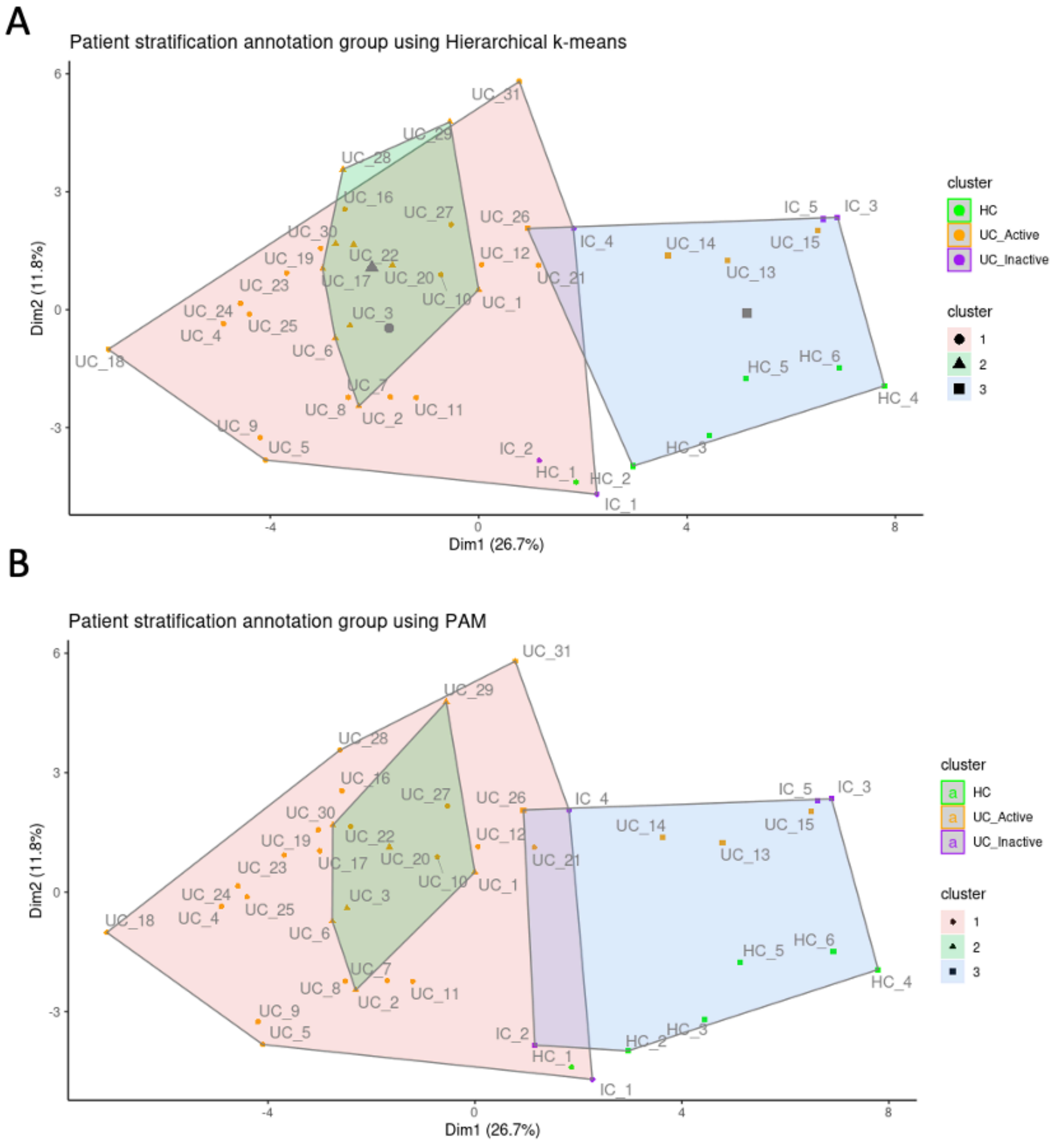


Fig. 5. **Patient stratification by cell proportions in annotation group. A. Hierarchical k-means algorith**m. Each cluster is presented with a different color. Inside them, each dot represents a sample that is colored according to their health, green for healthy controls (HC), orange for the active UC samples, and purple for inactive ones. **B. PAM algorithm**. Each cluster is presented with a different color. Inside them, each dot represents a sample that is colored according to their health, green for healthy controls (HC), orange for active state, and purple for the inactive ones.

After PCA analysis, we checked the contribution of each cell type to the dataset's variability (Fig 6). As a result, we obtained that Principal Component 1 is mainly affected by the Colonocytes proportion, a part of the epithelial cells that are highly damaged when the disease is in an active state. Moreover, other important variables are the cell types of the plasma subset that infiltrate the gut mucosa of UC patients.

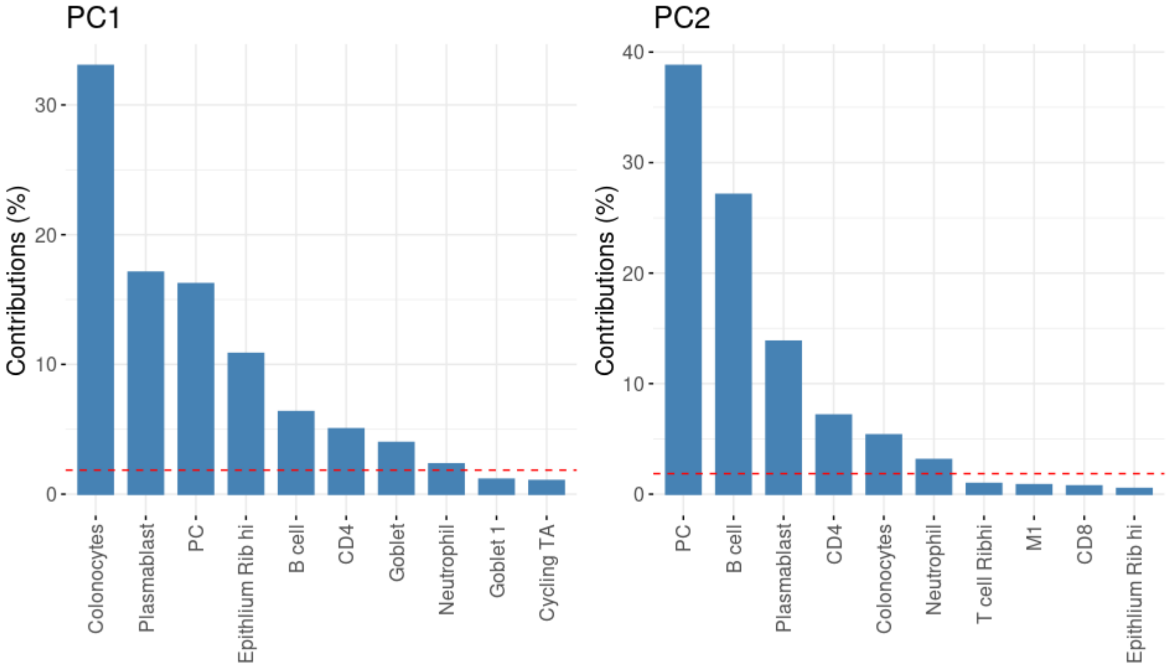


Fig. 6. **Variable contributions on annotation group**. Result of the PCA analysis performed on the samples according to the annotation that groups the cell types in major categories. PC1 is mainly explained by the Colonocytes and plasma cells, so does the PC2. The red line indicates the expected average contribution.

We performed a RF classification analysis to validate the results obtained. As a result, the Hierarchical k-means clustering at different mtry values (3, 7, and 14) had an OOB error of 12.5%, indicating that the 87.5% of the samples were classified correctly in the assay. Moreover, we assessed algorithm performance by conducting a confusion matrix of the train and test sets.

As expected, the accuracy of the train set was 1, whereas in the test set was 0.9. Also, the test set had an accuracy of 0.9, p-value of 0.011 and kappa value of 0.8462, indicating that the samples' classification is considered accurate. Then, the PAM algorithm, obtained similar values; 12.9 OOB error, an accuracy of 0.9, a p-value of 0.046 and kappa of 0.8077. A summary on the statistics variables obtained of the test set can be found on table 3

Algorithm	**OOB error**	**Accuracy**	**p-value**	**Kappa**
Hierarchical k-means	12.5%	0.9	0.011	0.846
PAM	12.9%	0.9	0.046	0.8077

Table 3. **Annotation group Random Forest**. Random forest statistics outcomes of the test set on the classification of the cell proportions group annotation.

Then, the Mean Decrease Gini allows to plot the variables of importance that explain the model produced by RF, in the case of the clustering performed using Hierarchical k-means, the variables that explain the clustering performed, considering the top 10, are the epithelial cell types (5/10), the plasma ones (2/10), the T cells (2/10) and the cycling (1/10) (Fig 7 A). On the other hand, in the clustering performed by the PAM algorithm, the variables that contribute most to the model are the epithelial (5/10) cell types, the plasma ones (3/10) and the T cells (2/10) (Fig 7 B)

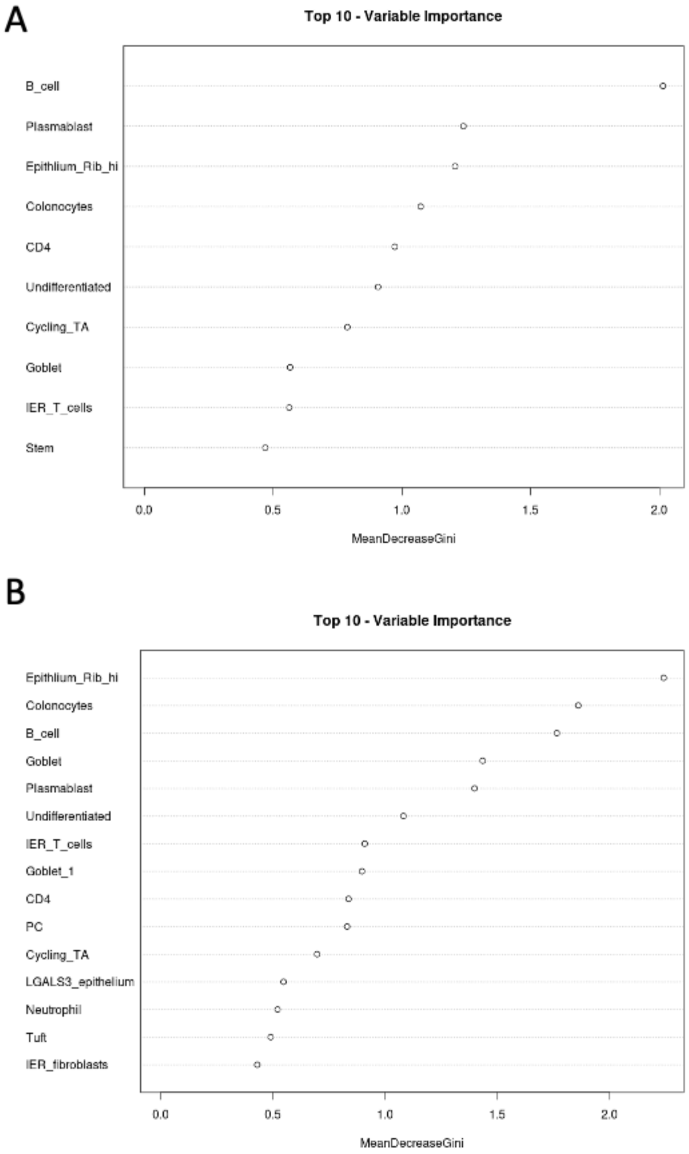


Fig. 7. **Top ten variable of importance of the RandomForest model on the annotation group A. Hierarchical k-means algorithm.** This index explains the earlier classification by the epithelial and plasma subsets. **B. PAM algorithm**. This index explains the earlier classification by the epithelial and plasma subsets.

4.1.2 Clustering on annotation refined

This annotation relies on a more specific approach to the cell types found in each subset. The number of k chosen for the hierarchical k-means clustering analysis is 2. As a result, we obtained that most patients with an active UC were gathered in cluster 2, whereas healthy and inactive patients were in cluster 1. Interestingly, one UC inactive sample was inside cluster 2. This sample has higher levels of cycling cells than the other inactive ones. (Fig. 8).

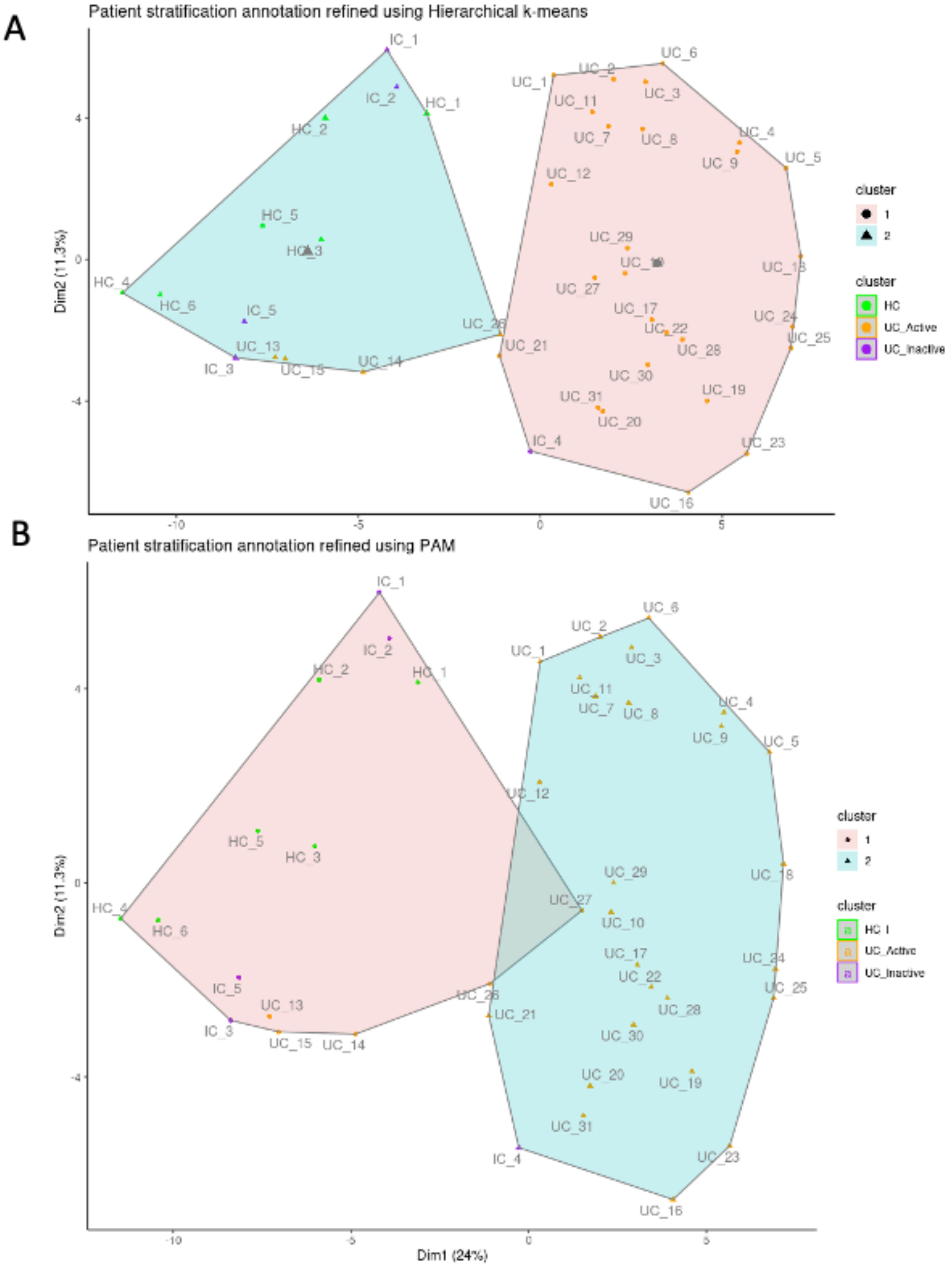


Fig. 8 **Patient stratification by cell proportions in annotation refined. A Hierarchical k-means algorithm**. Each cluster is presented by a different color. Inside them, each dot represents a sample that is colored according to their health, green for healthy controls (HC), orange for active state, and purple for the inactive ones. B. **PAM algorithm**. Each cluster is presented by a different color. Inside them, each dot

represents a sample that is colored according to their health, green for healthy controls (HC), orange for active state, and purple for the inactive ones.

In the same way, clustering analysis using PAM on this annotation also concurred in k =2. The outcome was highly similar to the Hierarchical k-means algorithm outcome, although in the PAM clustering method more samples that are in an active state are found in the same cluster as the healthy ones. In this way, Hierarchical k-means is a better model to represent the differences in disease status using this annotation. Both clustering algorithms aimed for two clusters, one of them almost entirely composed of all the active samples. For this reason, this annotation could be considered a better approach than the annotation group to cluster patients in an unsupervised manner.

Following the pipeline, we used PCA analysis to understand the variability in the dataset and the variables that contribute the most to this classification. Even though the cell types that contribute to each principal component are different from the annotation group, the subtypes that contribute the most are the epithelial and plasma cells.

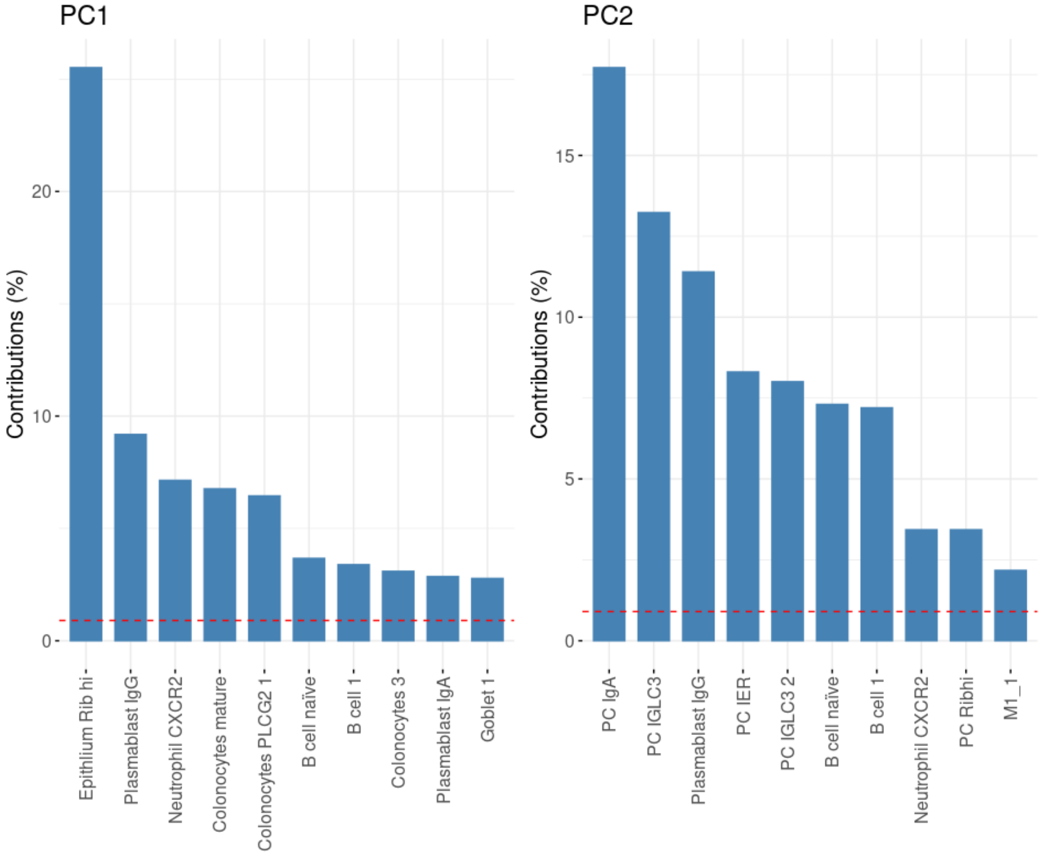


Fig. 9. **Variable contributions on annotation refined**. Result of the PCA analysis performed on the samples according to the annotation refined. PC1 is mainly explained by the epithelial es and plasma cells. PC2 is explained by the plasma cells.

By performing the RF algorithm with the clustering resulting from the Hierarchical k-means, the classification of the test set had an accuracy of 0.9, a p-value of 0.05, and a kappa value of 0.7925. Moreover, the OOB estimate error was 3.23%, meaning that samples were well classified in a 96.77%. The variables that were more important for the model according to the Mean Decrease Gini were related to the epithelial (5/10), myeloid (2/10), stromal (1/10), and plasma (1/10) subsets.(Fig 10). Then, the validation process on the PAM algorithm resulted in an accuracy of 0.9, p-value of 0.05, and kappa had a value of 0.8. However, this clustering increased the OOB error to 12.9%. Interestingly, the variables of importance that were highly ranked belonged to the epithelial subset (6/10), T (2/10), myeloid (1/10) and stromal (1/10) cells.

As it can be seen on both clustering methodologies, the colonocytes were the cell type with higher relevance, probably due to the damage these cells suffer during the disease activity.

Algorithm	**OOB error**	**Accuracy**	**p-value**	**Kappa**
Hierarchical k-means	3.23%	0.9	0.05	0.7925
PAM	12.9%	0.9	0.05	0.8

Table 4. **Annotation refined Random Forest**. Random forest statistics outcomes of the test set on the classification of the cell proportions annotation refined.

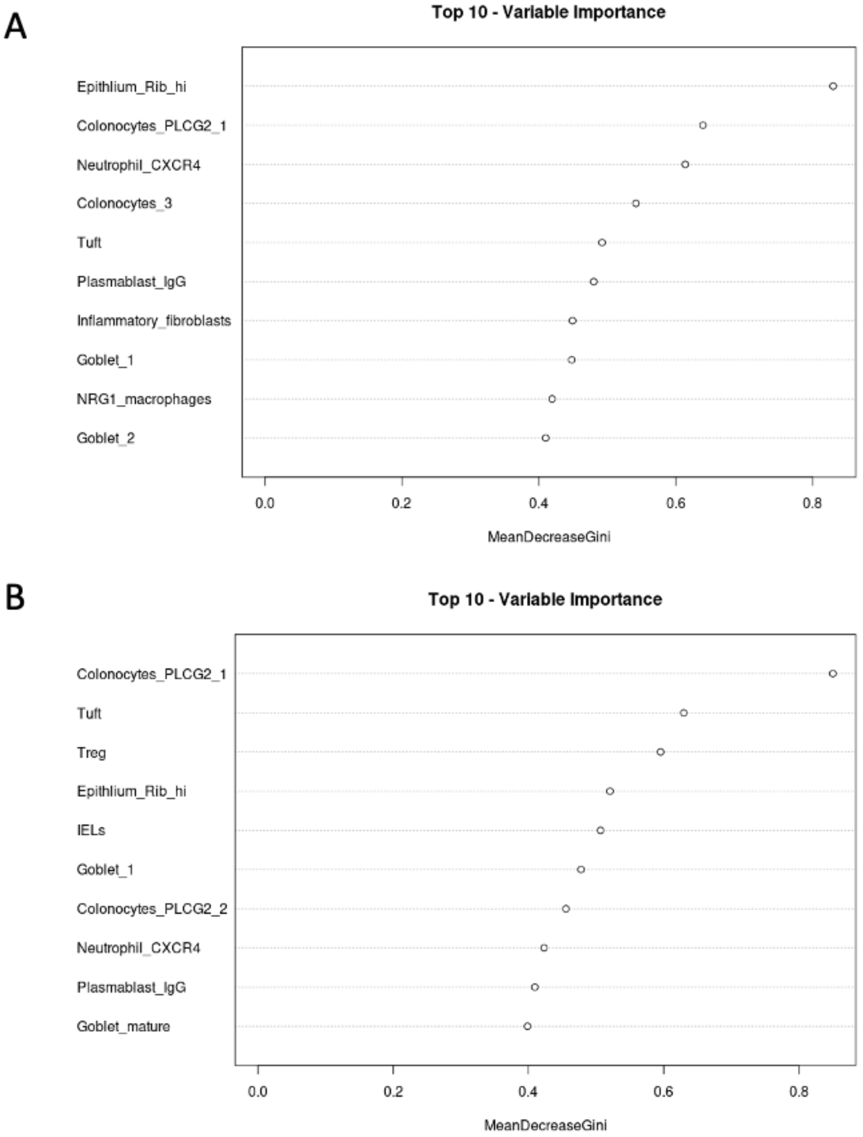


Fig. 10. **Top ten variable of importance of the Random Forest on the annotation refined. A. Hierarchical k-means algorithm**. This index explains that the earlier classification mainly relies on the epithelial and plasma subsets on this algorithm. B. **PAM algorithm**. This index explains that the earlier classification mainly relies on the epithelial, T cell and myeloid subsets on this algorithm

To measure the performance of the RF model, we created an Area under the curve – receiver operating characteristic (AUC-ROC) curve, which can help evaluate the RF performance. As it can be seen, both RF models show a good performance of the classification (Fig 10). Nevertheless, our number of samples should be increased to obtain a smoother curve.

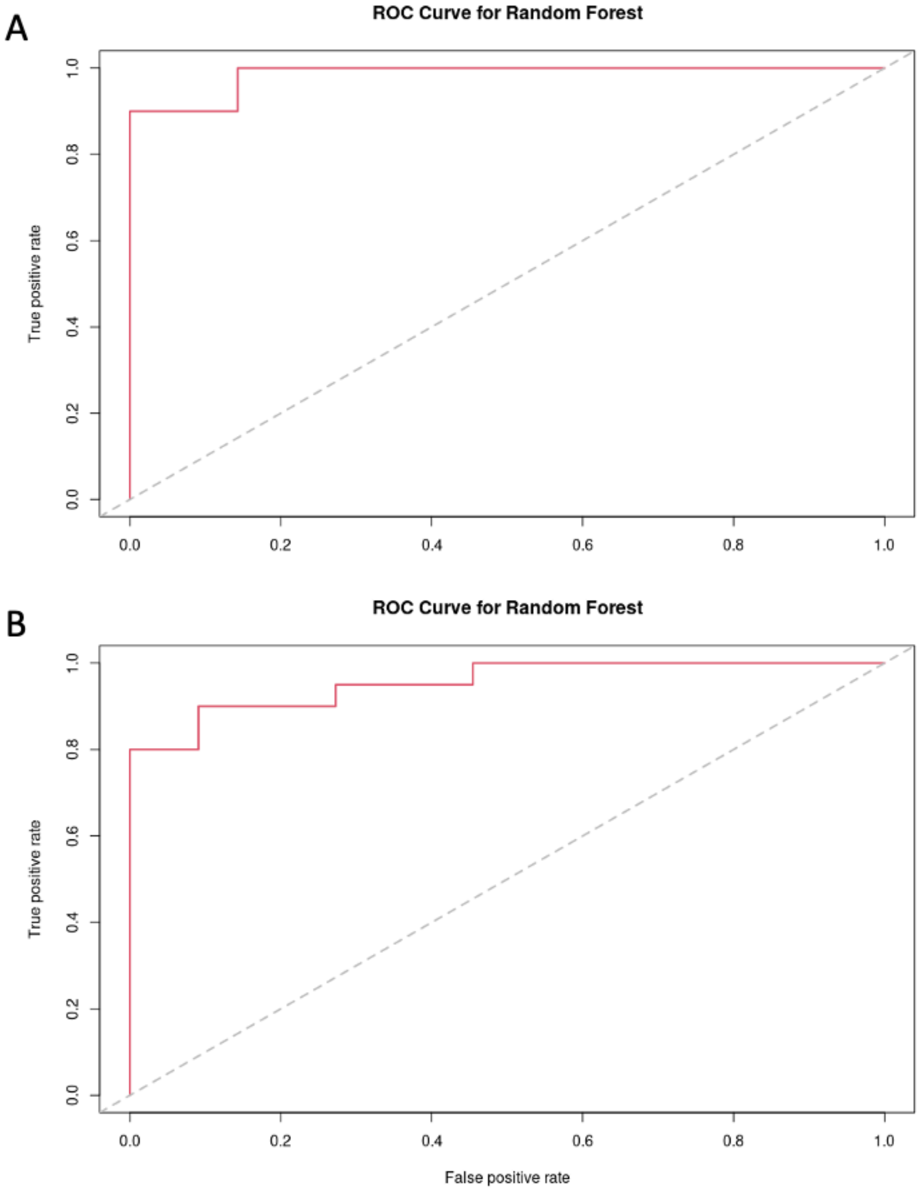


Fig. 11. **ROC curve for Random Forest A. Hierarchical k-means on annotation refined. B. PAM on annotation refined**

4.2 Stratification by gene expression

Given the previous results summarized on table 3 and 4, when samples are labeled under a refined annotation, we obtain better results at clustering, and the parameters associated to the validation of RF show lower OOB error on the classification. For this reason, for the following analysis, this annotation is the one used to perform the clustering.

Briefly, as further detailed in the Material and Methods section, we obtained a summarized expression per cell type and sample using the GSVA R package. Using this method, we obtained a matrix where rows are pathways and columns are cell types per sample. We performed unsupervised sample stratification per each cell type using Hierarchical k-means and PAM clustering techniques. After that, for each method, we generated a matrix containing the stratification results per cell type, where the rows were the patients, and the columns were the cell types. Unsupervised clustering algorithms in these resulting matrixes were carried out using k = 2 for both methodologies, according to the results of the statistical indexes provided by NbClust.

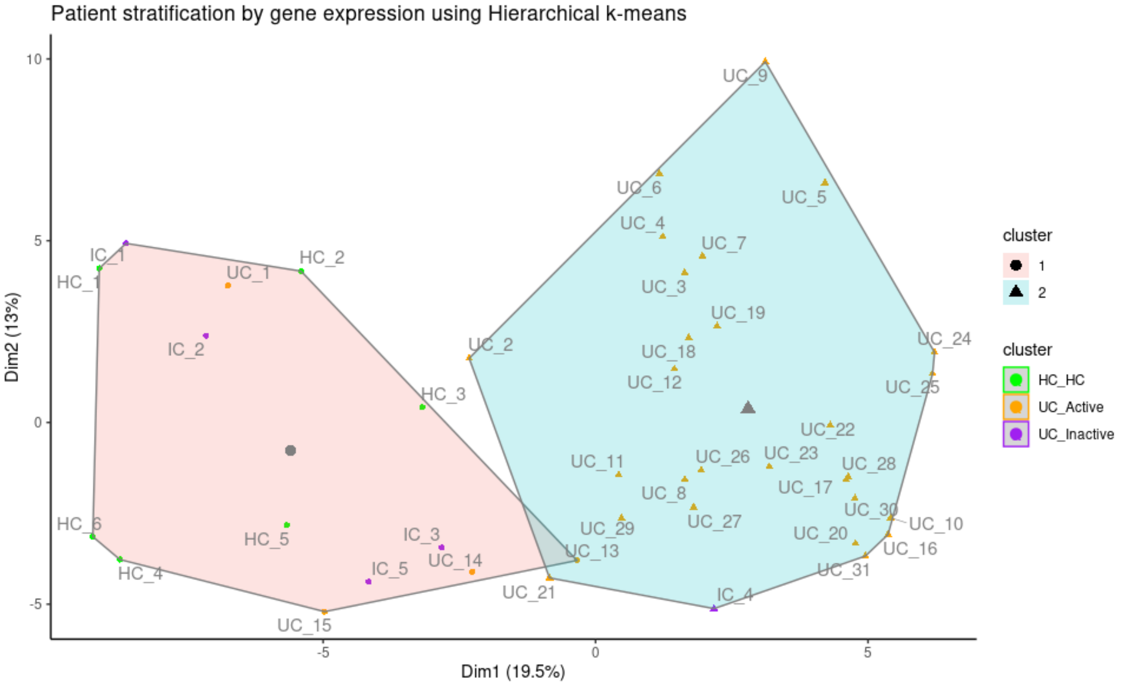


Fig. 12. **Patient stratification by gene expression using Hierarchical k-means algorithm.** Each cluster is presented with a different color. Inside them, each dot represents a sample that is colored according to

their health, green for healthy controls (HC), orange for the active UC samples, and purple for inactive ones.

Hierarchical k-means clustering performed considering pathway expression provides a good classification method (Fig 12), similar to the one done by cell proportions. However, this method provides better insights into the molecular granularity of the samples, as we are considering the molecular basis of the disease to classify patients. As it can be seen, cluster one is formed by the healthy and inactive samples, except for the samples UC_1, UC_13, UC_14 and UC_15. However, samples UC_13 and UC_14 come from patients that present a mild disease activity. In the same way as in the clustering carried out on the annotation refined of the cell proportions, IC_4 is found on cluster two.

Moreover, the clustering performed using PAM showed a similar classification to the one performed using the Hierarchical k-means algorithm. However, as it can be seen in the PCA of the PAM outcome, the two clusters are more distanced in the first dimension.

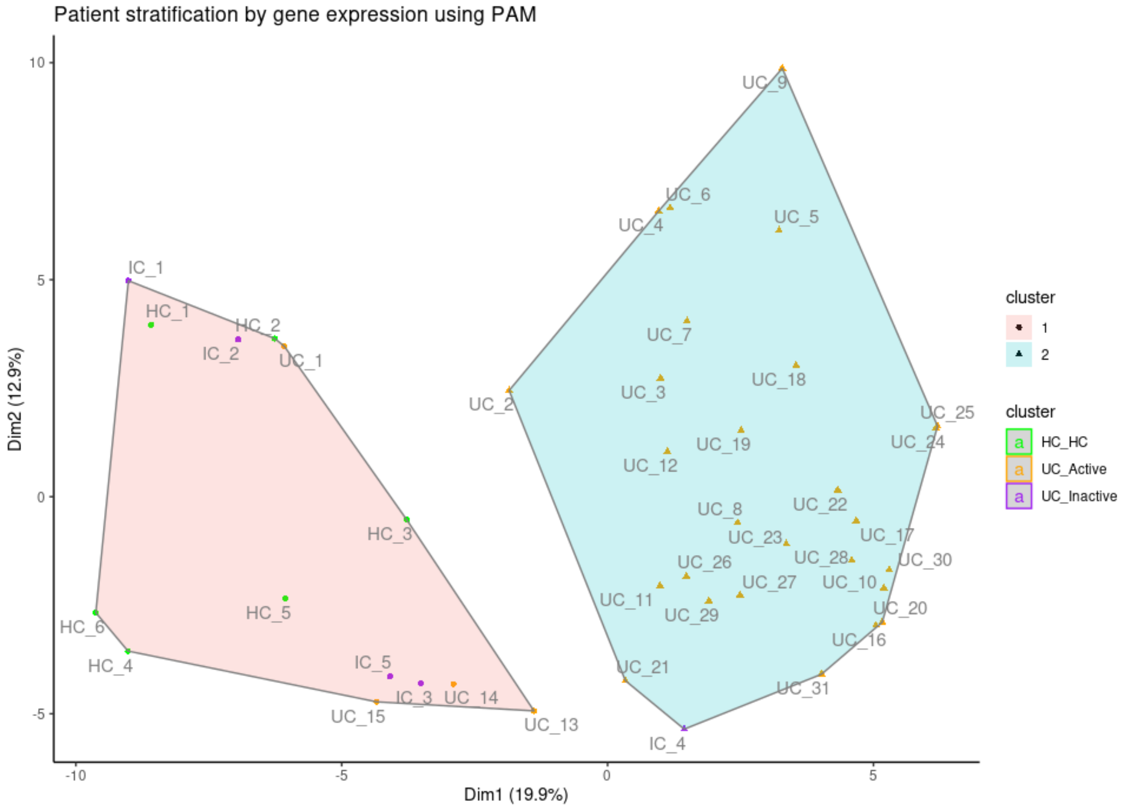


Fig. 13 **Patient stratification by gene expression using PAM algorithm**. Each cluster is presented by a different color. Inside them, each dot represents a sample that is colored according to their health, green for healthy controls (HC), orange for active state, and purple for the inactive ones.

Moreover, the variables that showed higher contribution to the principal components in both stratification methods belonged to clusters found in the myeloid and stromal subsets. However, the cell types involved are different within these subsets.

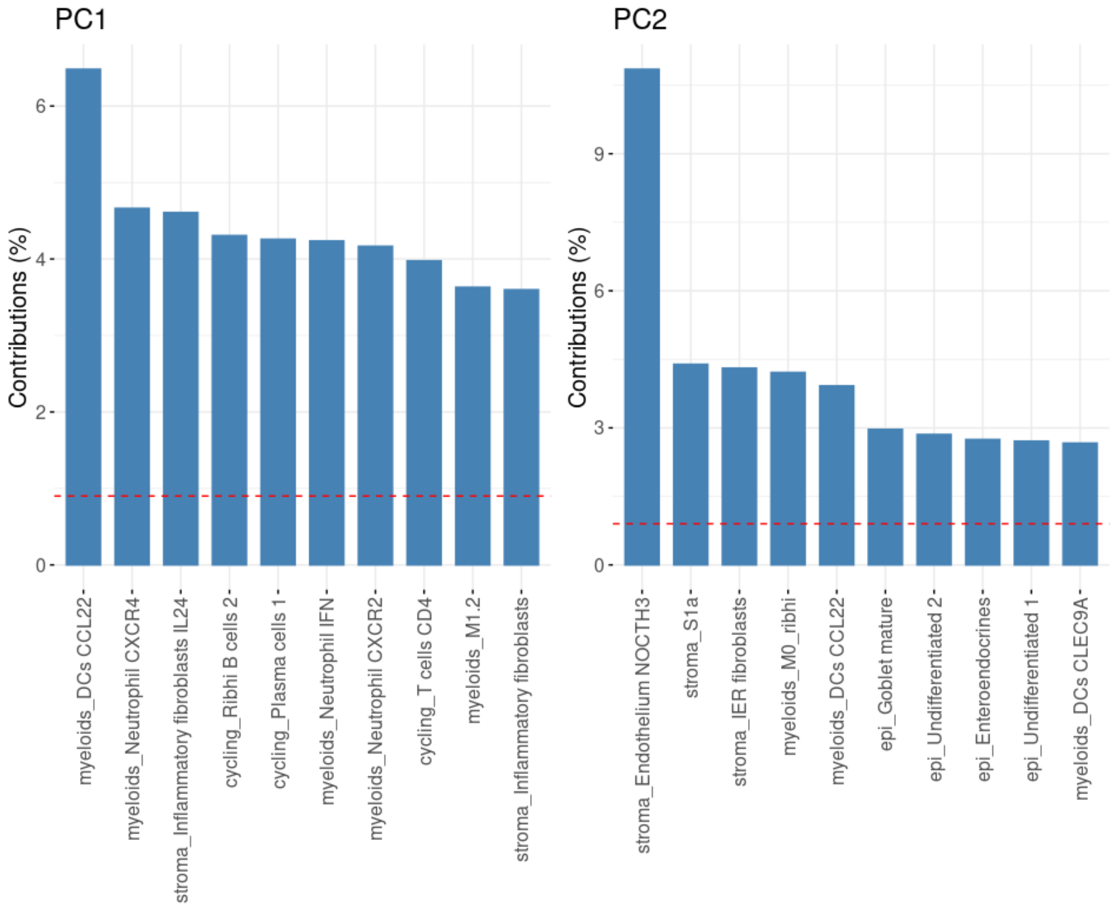


Fig. 14. **Variable contributions on gene expression using Hierarchical k-means approach**. Result of the PCA analysis performed on the samples according to their gene expression using the Hierarchical k-means algorithm. PC1 is mainly explained by myeloid cells. PC2 is explained by stromal cells.

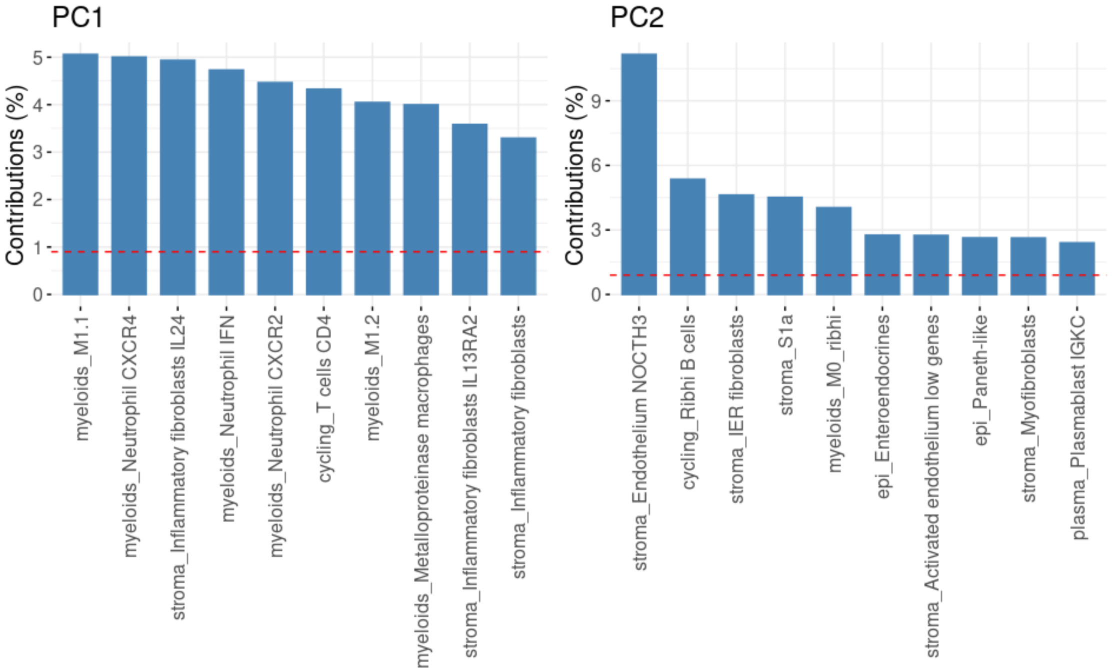


Fig. 15. **Variable contributions on gene expression using PAM approach**. Result of the PCA analysis performed on the samples according to their gene expression using the PAM algorithm. PC1 is mainly explained by myeloid cells. PC2 is explained by stromal cells.

Remarkably, the different cell types clustering algorithms individually could not correctly classify the patients according to their health. Only by obtaining the bigger picture we can focus on how the patients are classified.

4.2.1. Validation process

As mentioned, the RF algorithm was used to validate the classification provided by the unsupervised clustering algorithms. Considering sample clusterization as the variable of study, the RF algorithm was performed, obtaining on the Hierarchical k-means clustering technique an OOB estimate error of 6.45%. This OOB is low enough to accept that the classification has been done correctly. After adjusting the tree, the variables of importance considered by the Gini index to explain the model mostly rely on the myeloid (5/10) but also on the stromal (1/10), cycling (2/10), and plasma (2/10) ones.

The PAM clustering method resulted in an RF classification with a 3.23% OOB error. The accuracy decreased to 0.81, the p-value to 0.175 and the kappa value of 0.6071, making it the worst model in classifying the patients. The variables of importance mainly rely on the myeloid subset (7/10), but we can also find cycling (1/10), stromal (1/10) and plasma (1/10) variables on top.

Algorithm	**OOB error**	**Accuracy**	**p-value**	**Kappa**
Hierarchical k-means	6.45%	1	0.006	1
PAM	3.23%	0.9	0.175	0.6071

Table 5. **Gene expression Random Forest**. Random forest statistics outcomes of the test set on the classification by gene expression on both algorithms.

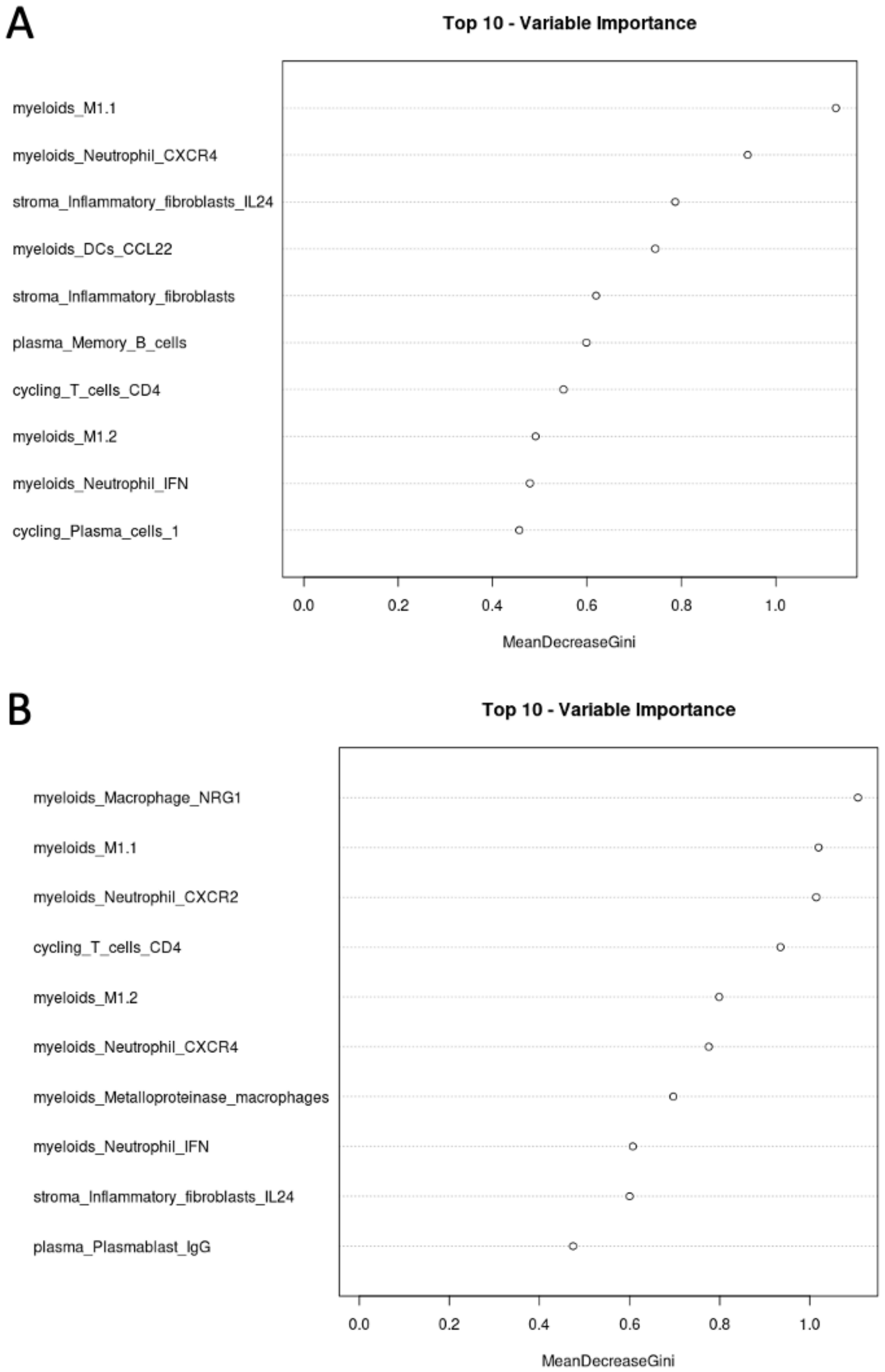


Fig. 16. **Mean decrease Gini top 10 variable of the stratification by gene expression. A Hierarchical k-means algorithm**. Result of the variable of importance of the RF analysis. Remarkably, the cell types found in the myeloid subset are the ones found to be more important in the classification. **B PAM algorithm**. Result of the variable of importance of the RF analysis. The myeloid subset mainly explains the classification by this partitioning method.

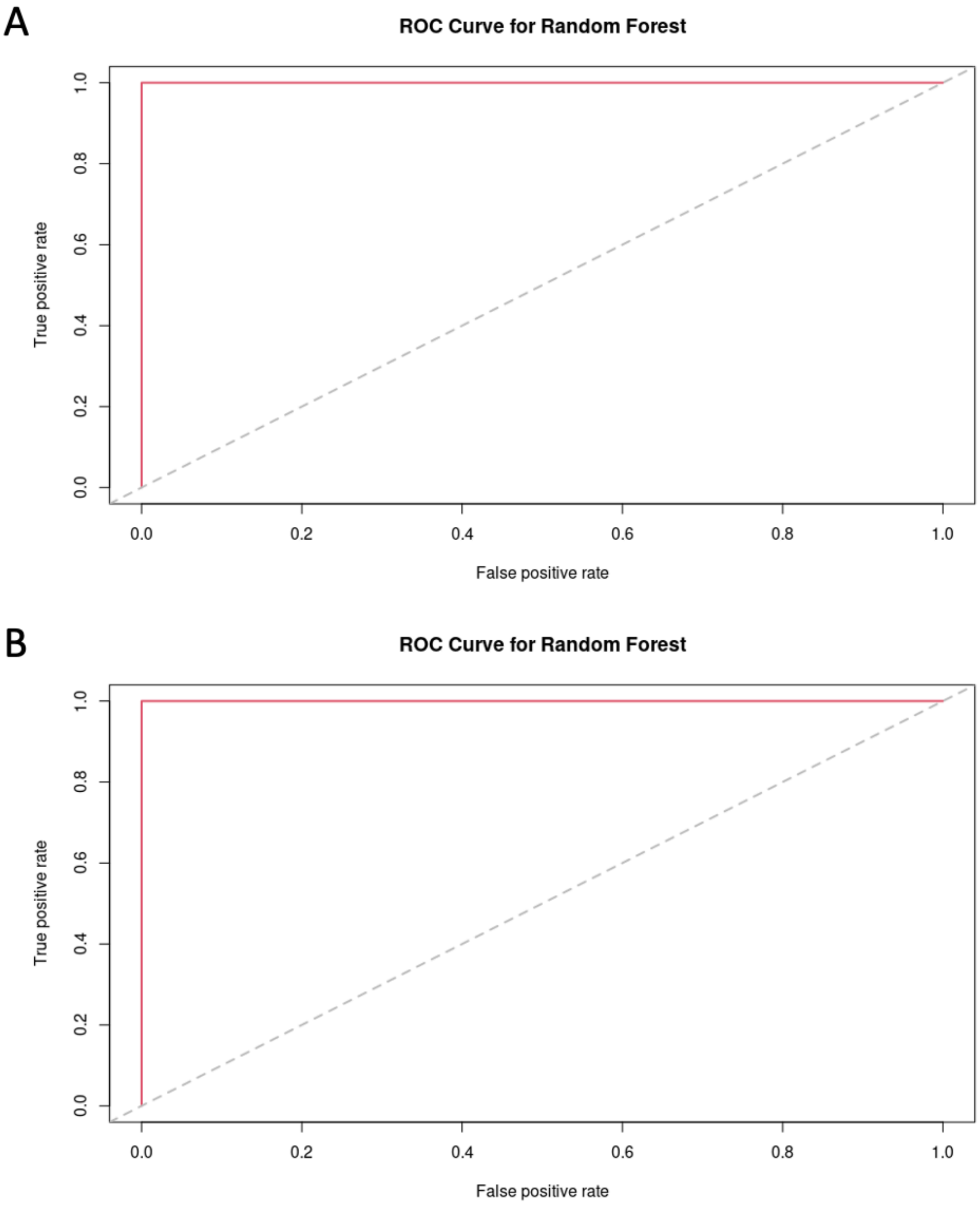
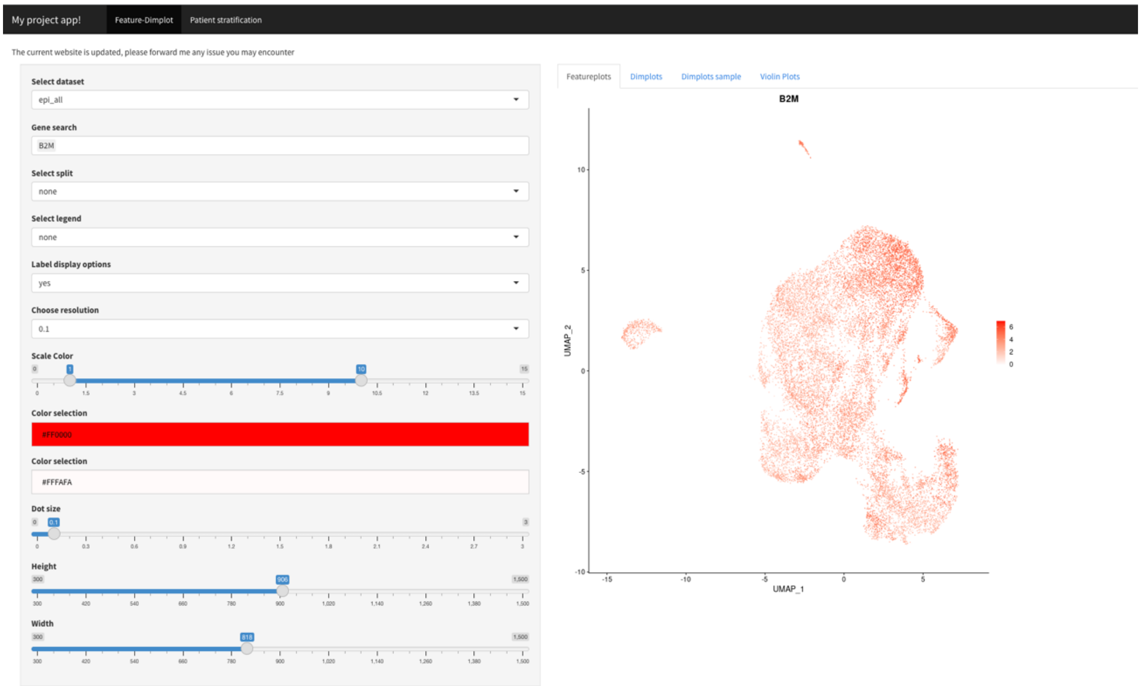


Fig. 17. ROC **curve for Random Forest on gene expression approach. A Hierarchical k-means algorithm**. B. **PAM algorithm**. The curve represents the how well the model can classify the patients.

Given the outcome the AUROC model, we can conclude that both models model has a good performance on the patients' classification. However, if we check the other parameters mentioned on table 5.

4.3 Results display

Since the code will not be available for this part of the dissertation, the following images are a way to portray the outcome of the shiny app developed. There are different panels, the plots related to the samples and gene expression and those to the results of the clustering. Different plot dimension were added as a tool bar so any one can easily select what they want to check and customize size, legend, labels, and plot size among others. To access the web app, a security panel has also been added.



The current website is updated, please forward me any issue you may encounter

Select dataset: epi_all

Gene search: B2M

Select split: none

Select legend: none

Label display options: yes

Choose resolution: 0.1

Scale Color: 1 – 10 (0 to 15)

Color selection: #FF0000

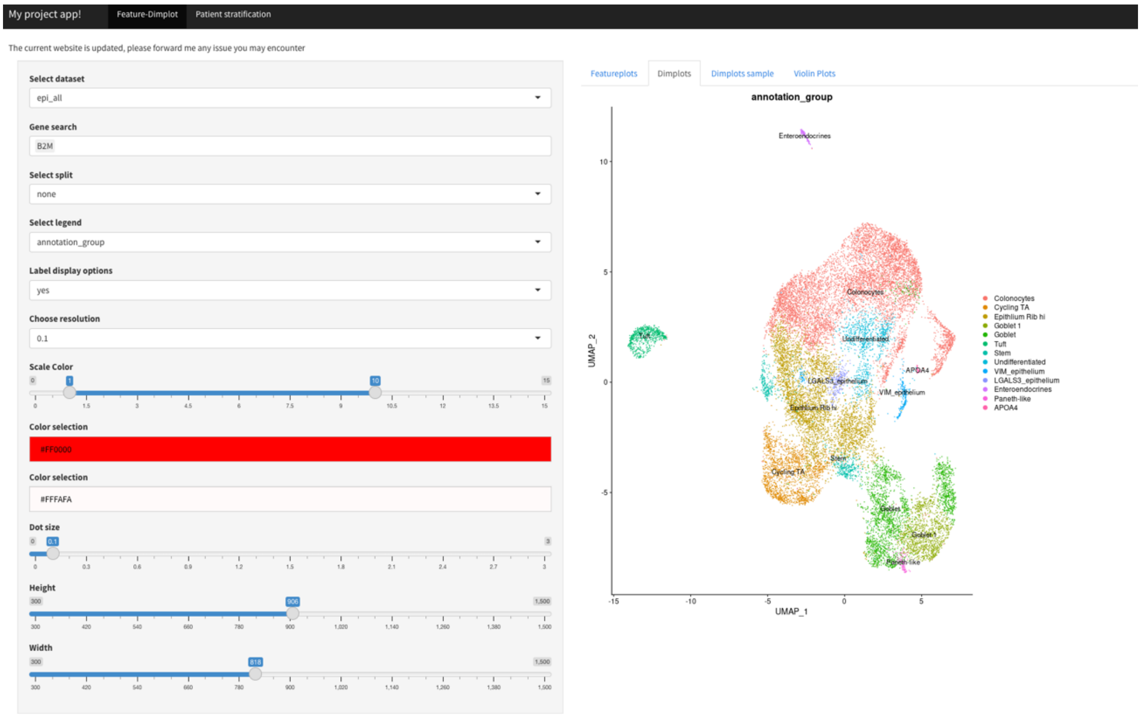
Color selection: #FFFAFA

Dot size: 0.1 (0 to 3)

Height: 906 (300 to 1,500)

Width: 818 (300 to 1,500)

Fig. 18. Shiny app gene expression on epithelial cells



My project app! | Feature-Dimplot | Patient stratification

The current website is updated, please forward me any issue you may encounter

Select dataset
epi_all

Gene search
B2M

Select split
none

Select legend
annotation_group

Label display options
yes

Choose resolution
0.1

Scale Color
0 – 15: 1 to 10

Color selection
#FF0000

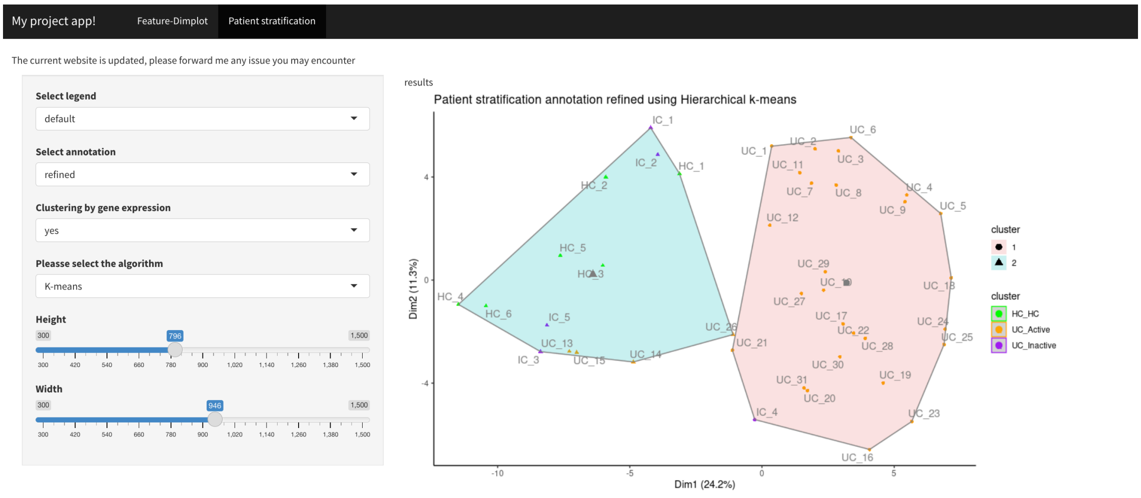
Color selection
#FFFAFA

Dot size
0 – 3: 0.1

Height
300 – 1,500: 906

Width
300 – 1,500: 818

Fig. 19. Shiny app annotation group on epithelial cells



My project app! | Feature-Dimplot | Patient stratification

The current website is updated, please forward me any issue you may encounter

Select legend
default

Select annotation
refined

Clustering by gene expression
yes

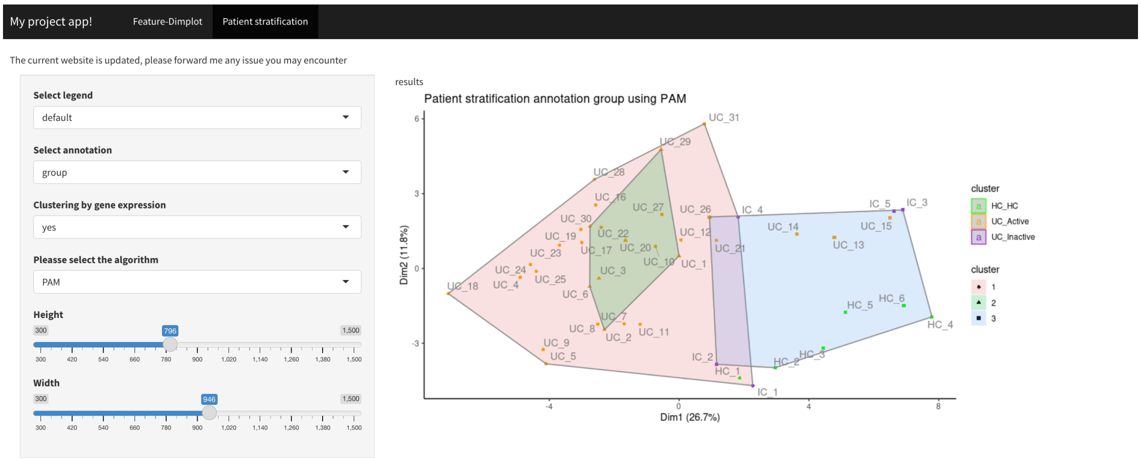
Pleasse select the algorithm
K-means

Height
300 – 1,500: 796

Width
300 – 1,500: 946

results

Fig. 20. Shiny app annotation refined clustering by cell proportions using Hierarchical k-means algorithm



The current website is updated, please forward me any issue you may encounter

Select legend
default

Select annotation
group

Clustering by gene expression
yes

Pleasse select the algorithm
PAM

Height
796

Width
946

Fig. 21. Shiny app annotation group clustering by cell proportions using PAM algorithm

5. Discussion

Ulcerative colitis is a complex relapsing disease characterized by the inflammation of the gut mucosa. scRNA-seq provides a valuable tool to understand the molecular scheme that comprises this disease and could help shed light on its underlying mechanisms. In this dissertation, we have analyzed 111000 cells from a dataset of 42 samples previously collected in the lab from 28 patients using scRNA-seq technology to provide a proof-of-concept of the patients' stratification using unsupervised clustering methods. These samples already had already been clinically characterized, and then categorized by their disease status (UC, IC, and HC).

To acknowledge the patient's heterogeneity, first, cell proportions were analyzed. From there, we encountered that cell type annotation is of the highest relevance to the result obtained. If we only consider global categories under each subset, probably we are missing information on the key players of the disease, and that is why a more refined annotation achieved a better clustering result of samples according to their disease status.

Partitioning clustering methods classify data based on their clusters' similarities and have been recently used to understand variability among patient samples (Zhun, X. et al., 2021). For this reason, we used and compared Hierarchical k-means and Partitioning Around Medoids techniques, and then performed a supervised approach to classify patients by their health category (Random Forest) to validate the clustering. Since the Hierarchical k-means approach can be sensitive to outliers, PAM algorithm can solve this by its dissimilarity score, which can help understand more our dataset.

When only considering the cell proportions, the predominant variables that explained the heterogeneity of the disease were cell types associated with the gut mucosa structure. During inflammation, epithelial cells found on the gut surface go through massive destruction due to ulceration that leads to the ultimate collapse of the gut cell barrier during the disease. This makes them highly variable between healthy controls and UC patients. For this reason, as expected, colonocytes, among other epithelial cells, are found to be more variable between the patient stratification clusters.

On the other hand, a remarkable number of immune cells infiltrate colonic mucosa during the disease, including plasma cells. Compared to HC, UC patients have a higher amount of IgG producing plasma ells, which are known to be related with inflammatory response (Boland, B.S. et al., 2020). As expected, immune cells, and especially plasma cells, have also a presence in the top variable features explaining the samples' variability when considering cell proportions.

ScRNA-seq provides valuable insight not only into the cell type proportions but also transcriptomic cell profiles at magnitudes that other traditional techniques cannot. This new level of transcriptomic information can help elucidate a better stratification of the patients.

Moreover, to classify patients by their gene expression, we used the MSigDB database, but only the pathways that were canonical or belonged to the biological processes categories. In future approaches, we could use a more extensive dataset to provide new insights into the patient's pathway expression. However, some genes are still not well characterized in the literature. Hence, we will not be able to understand their possible role in the pathogenesis of UC. Still, using pathways to understand heterogeneity proves to be a powerful tool that helps explain the differences between phenotypes. This way, changes in a set of genes coordinated in a cell function could explain the mechanisms underlying UC complexity.

The classification obtained from the pathway expression helped uncover new possible key players in the disease. As it can be seen on figure X, myeloid cells, specifically M1 macrophages, have high relevance in the clustering using hierarchical k-means algorithm. This methodology was statistically significant and was the one that presented the best OOB. Some studies have associated these macrophages to the inflammatory chronic state. (Zhu, W. et al., 2014) At the same time, we have seen that stromal cells are also relevant to the health status of the disease. Studies on enriched genes on these two subsets (especially in inflammatory monocytes and DCs) have associated them to resistance to Anti-TNF therapy (Steinbach, E.C. et al., 2015). In this sense, our model reflects the rewiring that takes place during inflammation of the gut mucosa that promotes chronicity, like monocytes, fibroblasts, and the immunoglobin-mediated response (Smillie C.S., et al., 2019).

Moreover, this methodology could be validated using Random Forest classification with more than 90% accuracy in classifying the samples (table 5). However, future protocols would require a higher number of samples to be able to generalize the results obtained. Besides, future approaches should aim to classify only patients with active disease to find patterns that can help elucidate the underlying molecular mechanisms and predict their response to treatment. Interestingly, the classification achieved using gene expression highly resembles the achieved by the one done using cell proportions. Even though the main variables that contributed to the classification are different in each case, in the cell proportions the ones related to the gut structure had more relevance, whereas in the pathway expression the myeloid and stromal subsets were the most important ones. Overall, both approaches provide valuable insights that are needed to stratify the samples and need to be considered in future studies.

To conclude, this dissertation provides a proof-of-concept of the use of unsupervised clustering methodologies to classify patients. scRNA-seq techniques provided new levels of granularity that support stromal and myeloid cells as key players in the disease. Finally, by portraying the results obtained from this work in an interactive web applicative made with the Shiny R package, we aim for a collaborative environment between experimental and bioinformatic teams to pave the way for better scientific research, which ultimately improves patients' quality of life.

6. Conclusions

- Unsupervised clustering techniques in Single-cell RNA-seq sequencing are useful to stratify UC and healthy patients.
- Both cell proportions and gene expression protocols are necessary to assess samples' heterogeneity.
- Hierarchical k-means performed better in stratification protocols.
- Delivering results by an interactive web application has led to better communication between researchers in the lab.

7. References

1. Feuerstein, J. D., Moss, A. C., & Farraye, F. A. (2019). Ulcerative Colitis. Mayo Clinic proceedings, 94(7), 1357–1373. https://doi.org/10.1016/j.mayocp.2019.01.018
2. Raine, T., Verstockt, B., Kopylov, U., Karmiris, K., Goldberg, R., Atreya, R., Burisch, J., Burke, J., Ellul, P., Hedin, C., Holubar, S. D., Katsanos, K., Lobaton, T., Schmidt, C., & Cullen, G. (2021). ECCO Topical Review: Refractory Inflammatory Bowel Disease. Journal of Crohn's & colitis, 15(10), 1605–1620. https://doi.org/10.1093/ecco-jcc/jjab112
3. Satsangi, J., Silverberg, M. S., Vermeire, S., & Colombel, J. F. (2006). The Montreal classification of inflammatory bowel disease: controversies, consensus, and implications. Gut, 55(6), 749–753. https://doi.org/10.1136/gut.2005.082909
4. Magro, F., Gionchetti, P., Eliakim, R., Ardizzone, S., Armuzzi, A., Barreiro-de Acosta, M., Burisch, J., Gecse, K. B., Hart, A. L., Hindryckx, P., Langner, C., Limdi, J. K., Pellino, G., Zagórowicz, E., Raine, T., Harbord, M., Rieder, F., & European Crohn's and Colitis Organisation [ECCO] (2017). Third European Evidence-based Consensus on Diagnosis and Management of Ulcerative Colitis. Part 1: Definitions, Diagnosis, Extra-intestinal Manifestations, Pregnancy, Cancer Surveillance, Surgery, and Ileo-anal Pouch Disorders. Journal of Crohn's & colitis, 11(6), 649–670. https://doi.org/10.1093/ecco-jcc/jjx008
5. Kaplan G. G. (2015). The global burden of IBD: from 2015 to 2025. Nature reviews. Gastroenterology & hepatology, 12(12), 720–727. https://doi.org/10.1038/nrgastro.2015.150
6. Du, L., & Ha, C. (2020). Epidemiology and Pathogenesis of Ulcerative Colitis. Gastroenterology clinics of North America, 49(4), 643–654. https://doi.org/10.1016/j.gtc.2020.07.005
7. Kayal, M., & Shah, S. (2019). Ulcerative Colitis: Current and Emerging Treatment Strategies. Journal of clinical medicine, 9(1), 94. https://doi.org/10.3390/jcm9010094
8. Stittrich, A. B., Ashworth, J., Shi, M., Robinson, M., Mauldin, D., Brunkow, M. E., Biswas, S., Kim, J. M., Kwon, K. S., Jung, J. U., Galas, D., Serikawa, K., Duerr, R. H., Guthery, S. L., Peschon, J., Hood, L., Roach, J. C., & Glusman, G. (2016). Genomic architecture of inflammatory bowel disease in five families with multiple affected individuals. Human genome variation, 3, 15060. https://doi.org/10.1038/hgv.2015.60
9. Ungaro R, Mehandru S, Allen PB, et al. Ulcerative colitis. Lancet 2017;389:1756–70.
10. Porter, R. J., Kalla, R., & Ho, G. T. (2020). Ulcerative colitis: Recent advances in the understanding of disease pathogenesis. F1000Research, 9, F1000 Faculty Rev-294. https://doi.org/10.12688/f1000research.20M805.1

11. Mokry M, Middendorp S, Wiegerinck CL, et al. Many inflammatory bowel disease risk loci include regions that regulate gene expression in immune cells and the intestinal epithelium. Gastroenterology 2014;146:1040-7.
12. Chen GB, Lee SH, Brion MJ, et al.: Estimation and partitioning of (co)heritability of inflammatory bowel disease from GWAS and immunochip data. Hum Mol Genet. 2014; 23(17): 4710–20.
13. Kaplan, G. G., & Ng, S. C. (2016). Globalisation of inflammatory bowel disease: perspectives from the evolution of inflammatory bowel disease in the UK and China. The lancet. Gastroenterology & hepatology, 1(4), 307–316. https://doi.org/10.1016/S2468-1253(16)30077-2
14. Villumsen, M., Aznar, S., Pakkenberg, B., Jess, T., & Brudek, T. (2019). Inflammatory bowel disease increases the risk of Parkinson's disease: a Danish nationwide cohort study 1977-2014. Gut, 68(1), 18–24. https://doi.org/10.1136/gutjnl-2017-315666
15. C. M. Guinane and P. D. Cotter, “Role of the gut microbiota in health and chronic gastrointestinal disease: understanding a hidden metabolic organ,” Therapeutic advances in gastroenterology, vol. 6, no. 4, pp. 295–308, 2013.
16. Moen AE, Lindstrøm JC, Tannæs TM, et al.: The prevalence and transcriptional activity of the mucosal microbiota of ulcerative colitis patients. Sci Rep. 2018; 8(1): 17278.
17. Turner JR: Intestinal mucosal barrier function in health and disease. Nat Rev Immunol. 2009; 9(11): 799–809. PubMed Abstract | Publisher Full Text
18. M. Van der Sluis, B. A. De Koning, A. C. De Bruijn, A. Velcich, J. P. Meijerink, J. B. Van Goudoever, H. A. Bu ̈ller, J. Dekker, I. Van Seunin- gen, I. B. Renes, et al., “Muc2-deficient mice spontaneously develop colitis, indicating that muc2 is critical for colonic protection,” Gastroenterology, vol. 131, no. 1, pp. 117–129, 2006

    K. W. Schroeder, W. J. Tremaine, and D. M. Ilstrup, “Coated oral 5- aminosalicylic acid therapy for mildly to moderately active ulcerative co- litis,” New England Journal of Medicine, vol. 317, no. 26, pp. 1625–1629, 1987.

19. Ordás, I., Eckmann, L., Talamini, M., Baumgart, D. C., & Sandborn, W. J. (2012). Ulcerative colitis. Lancet (London, England), 380(9853), 1606–1619. https://doi.org/10.1016/S0140-6736(12)60150-0
20. J. C. Brazil, N. A. Louis, and C. A. Parkos, “The role of polymorphonuclear leukocyte trafficking in the perpetuation of inflammation during inflamma- tory bowel disease,” Inflammatory bowel diseases, vol. 19, no. 7, p. 1556, 201
21. Graham DB, Luo C, O’Connell DJ, et al.: Antigen discovery and specification of immunodominance hierarchies for MHCII-restricted epitopes. Nat Med. 2018;24(11):1762–72. 10.1038/s41591-018-0203-7

22. 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.
23. Lobatón, T., Bessissow, T., De Hertogh, G., Lemmens, B., Maedler, C., Van Assche, G., Vermeire, S., Bisschops, R., Rutgeerts, P., Bitton, A., Afif, W., Marcus, V., & Ferrante, M. (2015). The Modified Mayo Endoscopic Score (MMES): A New Index for the Assessment of Extension and Severity of Endoscopic Activity in Ulcerative Colitis Patients. Journal of Crohn's & colitis, 9(10), 846–852. https://doi.org/10.1093/ecco-jcc/jjv111
24. Smillie, C. S., Biton, M., Ordovas-Montanes, J., Sullivan, K. M., Burgin, G., Graham, D. B., Herbst, R. H., Rogel, N., Slyper, M., Waldman, J., Sud, M., Andrews, E., Velonias, G., Haber, A. L., Jagadeesh, K., Vickovic, S., Yao, J., Stevens, C., Dionne, D., Nguyen, L. T., … Regev, A. (2019). Intra- and Inter-cellular Rewiring of the Human Colon during Ulcerative Colitis. Cell, 178(3), 714–730.e22. https://doi.org/10.1016/j.cell.2019.06.029
25. Kucharzik, T., Koletzko, S., Kannengiesser, K., & Dignass, A. (2020). Ulcerative Colitis-Diagnostic and Therapeutic Algorithms. Deutsches Arzteblatt international, 117(33-34), 564–574. https://doi.org/10.3238/arztebl.2020.0564
26. Chande, N., Patton, P. H., Tsoulis, D. J., Thomas, B. S., & MacDonald, J. K. (2015). Azathioprine or 6-mercaptopurine for maintenance of remission in Crohn's disease. The Cochrane database of systematic reviews, (10), CD000067. https://doi.org/10.1002/14651858.CD000067.pub3
27. Teng, Michele W L; Bowman, Edward P; McElwee, Joshua J; Smyth, Mark J; Casanova, Jean-Laurent; Cooper, Andrea M; Cua, Daniel J (2015). IL-12 and IL-23 cytokines: from discovery to targeted therapies for immune-mediated inflammatory diseases. Nature Medicine, 21(7), 719–729. doi:10.1038/nm.3895
28. Lai, L., Li, H., Feng, Q., Shen, J., & Ran, Z. (2021). Multi-factor mediated functional modules identify novel classification of ulcerative colitis and functional gene panel. Scientific reports, 11(1), 5669. https://doi.org/10.1038/s41598-021-85000-3
29. Selin, K. A., Hedin, C., & Villablanca, E. J. (2021). Immunological Networks Defining the Heterogeneity of Inflammatory Bowel Diseases. Journal of Crohn's & colitis, 15(11), 1959–1973. https://doi.org/10.1093/ecco-jcc/jjab085
30. Corridoni, D., Chapman, T., Antanaviciute, A., Satsangi, J., & Simmons, A. (2020). Inflammatory Bowel Disease Through the Lens of Single-cell RNA-seq Technologies. Inflammatory bowel diseases, 26(11), 1658–1668. https://doi.org/10.1093/ibd/izaa089

31. Louvain Blondel, V., Guillaume, J., Lambiotte, R., and Lefebvre, E. (2008). Fast Unfolding of Communities in Large Networks. J. Stat. Mech. Theor. Exp. 83, 10008. doi:10.1088/1742-5468/2008/10/p10008
32. Korsunsky Ilya, Millard Nghia, Fan Jean, Slowikowski Kamil, Zhang Fan, Wei Kevin, Baglaenko Yuriy, Brenner Michael, Loh Po-ru, Raychaudhuri Soumya. Fast, sensitive and accurate integration of single-cell data with Harmony. Nature Methods. 2019;16(12):1289–1296. doi: 10.1038/s41592-019-0619-0.
33. Sarica, A., Cerasa, A., & Quattrone, A. (2017). Random Forest Algorithm for the Classification of Neuroimaging Data in Alzheimer's Disease: A Systematic Review. *Frontiers in aging neuroscience*, *9*, 329. https://doi.org/10.3389/fnagi.2017.00329
34. Baran-Gale, J., Chandra, T., & Kirschner, K. (2018). Experimental design for single-cell RNA sequencing. *Briefings in functional genomics*, *17*(4), 233–239. https://doi.org/10.1093/bfgp/elx035
35. Haque, A., Engel, J., Teichmann, S. A., & Lönnberg, T. (2017). A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome medicine*, *9*(1), 75. https://doi.org/10.1186/s13073-017-0467-4
36. Mubeen, S., Hoyt, C. T., Gemünd, A., Hofmann-Apitius, M., Fröhlich, H., & Domingo-Fernández, D. (2019). The Impact of Pathway Database Choice on Statistical Enrichment Analysis and Predictive Modeling. *Frontiers in genetics*, *10*, 1203. https://doi.org/10.3389/fgene.2019.0120
37. Zhang, C., Gao, L., Wang, B., & Gao, Y. (2021). Improving Single-Cell RNA-seq Clustering by Integrating Pathways. *Briefings in bioinformatics*, *22*(6), bbab147. https://doi.org/10.1093/bib/bbab147
38. Emmert-Streib, F., & Glazko, G. V. (2011). Pathway analysis of expression data: deciphering functional building blocks of complex diseases. *PLoS computational biology*, *7*(5), e1002053. https://doi.org/10.1371/journal.pcbi.1002053
39. Zhu, W., Yu, J., Nie, Y., Shi, X., Liu, Y., Li, F., & Zhang, X. L. (2014). Disequilibrium of M1 and M2 macrophages correlates with the development of experimental inflammatory bowel diseases. Immunological investigations, 43(7), 638–652. https://doi.org/10.3109/08820139.2014.909456
40. Steinbach, E. C., Gipson, G. R., & Sheikh, S. Z. (2015). Induction of Murine Intestinal Inflammation by Adoptive Transfer of Effector CD4+ CD45RB high T Cells into Immunodeficient Mice. Journal of visualized experiments : JoVE, (98), 52533. https://doi.org/10.3791/52533
41. Boland, B. S., He, Z., Tsai, M. S., Olvera, J. G., Omilusik, K. D., Duong, H. G., Kim, E. S., Limary, A. E., Jin, W., Milner, J. J., Yu, B., Patel, S. A., Louis, T. L., Tysl, T., Kurd, N. S., Bortnick, A.,

Quezada, L. K., Kanbar, J. N., Miralles, A., Huylebroeck, D., … Chang, J. T. (2020). Heterogeneity and clonal relationships of adaptive immune cells in ulcerative colitis revealed by single-cell analyses. Science immunology, 5(50), eabb4432. https://doi.org/10.1126/sciimmunol.abb4432

APEX

Patient	**Health activity**	**Mayo index (0-3)**	**N cells**
UC_1	Active	3	1386
UC_2	Active	3	2537
UC_3	Active	3	2045
UC_4	Active	3	1562
UC_5	Active	3	2877
UC_6	Active	3	2751
UC_7	Active	3	2269
UC_8	Active	3	2752
UC_9	Active	3	4028
UC_10	Active	3	3552
UC_11	Active	3	2915
UC_12	Active	3	1359
UC_13	Active	3	1788
UC_14	Active	3	1456
UC_15	Active	3	2279
UC_16	Active	2	5312
UC_17	Active	3	2789
UC_18	Active	3	2731
UC_19	Active	3	760
UC_20	Active	3	3007
UC_21	Active	3	2368
UC_22	Active	3	1215
UC_23	Active	3	2866
UC_24	Active	3	4373
UC_25	Active	3	5045
UC_26	Active	3	2214
UC_27	Active	3	2316
UC_28	Active	3	2157
UC_29	Active	3	3731
UC_30	Active	3	3335
UC_31	Active	3	5337

HC_1	Healthy Control	NA	1498
HC_2	Healthy Control	NA	2483
HC_3	Healthy Control	NA	3463
HC_4	Healthy Control	NA	2256
HC_5	Healthy Control	NA	3305
HC_6	Healthy Control	NA	1888
IC_1	Inactive	NA	1041
IC_2	Inactive	NA	921
IC_3	Inactive	0	1677
IC_4	Inactive	1	2845
IC_5	Inactive	0	4511

Table 1**. Patient information**. Samples used for this dissertation and their associated health activity, Mayo score and number of cells. Patient code used comes from the projects where the samples were obtained.