Anàlisi del patró de disbiòsi intestinal en colitis ulcerosa a partir de mostres fecals.

Miquel Castany Roma

MU Bioinf. i Bioest.
Anàlisi de dades òmiques

Tutor/a de TF
Yolanda Guillén Montalbán

06.2023

FITXA DEL TREBALL FINAL

Títol del treball:	*Anàlisi del patró de disbiòsi intestinal en colitis ulcerosa a partir de mostres fecals.*
Nom de l'autor:	*Miquel Castany Roma*
Nom del consultor/a:	*Yolanda Guillén Montalbán*
Data de lliurament (mm/aaaa):	*06/2023*
Titulació o programa:	*Màster universitari en Bioinformàtica i bioestadística UOC-UB*
Àrea del Treball Final:	*Anàlisi de Dades Òmiques*
Idioma del treball:	*Català*
Paraules clau	*16S rRNA, metagenomics, gut microbiome*

Resum del Treball

El treball parteix de dades de seqüenciació metagenòmica de 16S rRNA de mostres fecals de pacients amb colitis ulcerosa (UC) i pacients sans (HC) de diferents estudis prèviament publicats. Els diferents estudis s'han dividit en dos cohorts: un de *discovery* (CD) i un altre de *validation* (CV). S'han realitzat les pertinents classificacions taxonòmiques de cada una de les mostres mitjançant programari especialitzat per a aquestes tasques. En el nostre cas hem utilitzat el software conegut com a *Mothur* (Schloss et al., 2009) i una base de dades de classificació taxonòmica de seqüències de 16S rRNA com és SILVA (Quast et al., 2013).

Un cop s'han obtingut les dades de classificació taxonòmica, s'ha procedit a l'anàlisi estadístic mitjançant Rstudio per tal de determinar i extreure'n un patró de disbiòsi comú a les mostres de pacients que componen la cohort *discovery*. Seguidament, aquests resultats s'han intentat extrapolar a la cohort *validation*, que ha estat compost de mostres de pacients amb UC però provinents d'un altre estudi.

Posteriorment, s'ha realitzat una tasca dedicada a la predicció de resultats clínics utilitzant les abundàncies relatives dels diferents taxons com a característica diferencial. S'han fet servir tècniques de *Machine Learning* (ML) (Serrano-Gómez et al., 2021) però també hem emprat altres aproximacions basades en models lineals generalitzats (GLMs) o regressions múltiples (Rivera-Pinto et al., 2018).

Abstract

The work is based on metagenomic sequencing data of 16S rRNA from fecal samples of patients with ulcerative colitis (UC) and healthy controls (HC) from different previously published studies. The different studies have been divided into two cohorts: a discovery cohort (CD) and a validation cohort (CV). The relevant taxonomic classifications of each sample have been performed using specialized software for these tasks. In our case, we have used the software known as *Mothur* (Schloss et al., 2009) and a taxonomic classification database for 16S rRNA sequences such as SILVA (Quast et al., 2013).

Once the taxonomic classification data has been obtained, statistical analysis has been carried out using Rstudio to determine and extract a common dysbiosis pattern to the samples of patients composing the discovery cohort. Subsequently, these results have been attempted to be extrapolated to the validation cohort, which consisted of UC patient samples from another study.

Afterwards, a task dedicated to the prediction of clinical outcomes has been carried out using the relative abundances of different taxa as differentiating features. Machine learning techniques (ML) (Serrano-Gómez et al., 2021) have been used for this purpose, but other approaches based on generalized linear models (GLMs) or multiple regressions have also been employed (Rivera-Pinto et al., 2018).

Índex

Llista de figures

1. Introducció

1.1. Context i justificació del Treball

El projecte del genoma humà, culminat l'any 2003, va suposar una fita històrica pel que fa a la biologia, la biotecnologia i la medicina en general (International Human Genome Sequencing Consortium, 2004). Aquell projecte que va durar 13 anys va permetre conèixer la seqüència, gairebé completa, del genoma del *Homo sapiens*. Les implicacions a nivell mèdic que conferia el fet de ser coneixedors de l'ordre en el qual es disposaven els 4 nucleòtids possibles al llarg dels diferents cromosomes eren enormes. Això va suposar l'aparició de la disciplina coneguda com a genòmica amb la possibilitat de poder seqüenciar i posteriorment analitzar els genomes d'altre organismes. Més tard, amb l'aparició de diferents tècniques de seqüenciació cada cop més especialitzades ha estat possible conèixer el conjunt dels genomes dels diferents microorganismes que conformen un hàbitat determinat com pugui ser, per exemple, una mostra de sòl, d'aigua de mar o bé de femta humana. Aquests conjunts de microorganismes s'anomenen microbiomes i com que gran part d'aquests microorganismes no són cultivables mitjançant tècniques de microbiologia clàssica és necessari emprar eines bioinformàtiques per al seu estudi. La disciplina que s'ocupa d'aquesta tasca es denomina metagenòmica (Breitwieser et al., 2019; Chiu & Miller, 2019; Escobar-Zepeda et al., 2015) i el principal projecte a nivell mundial que s'ocupa d'entendre com el microbioma influeix sobre la fisiologia humana és el Projecte del Microbioma Humà (HMP) (Turnbaugh et al., 2007).

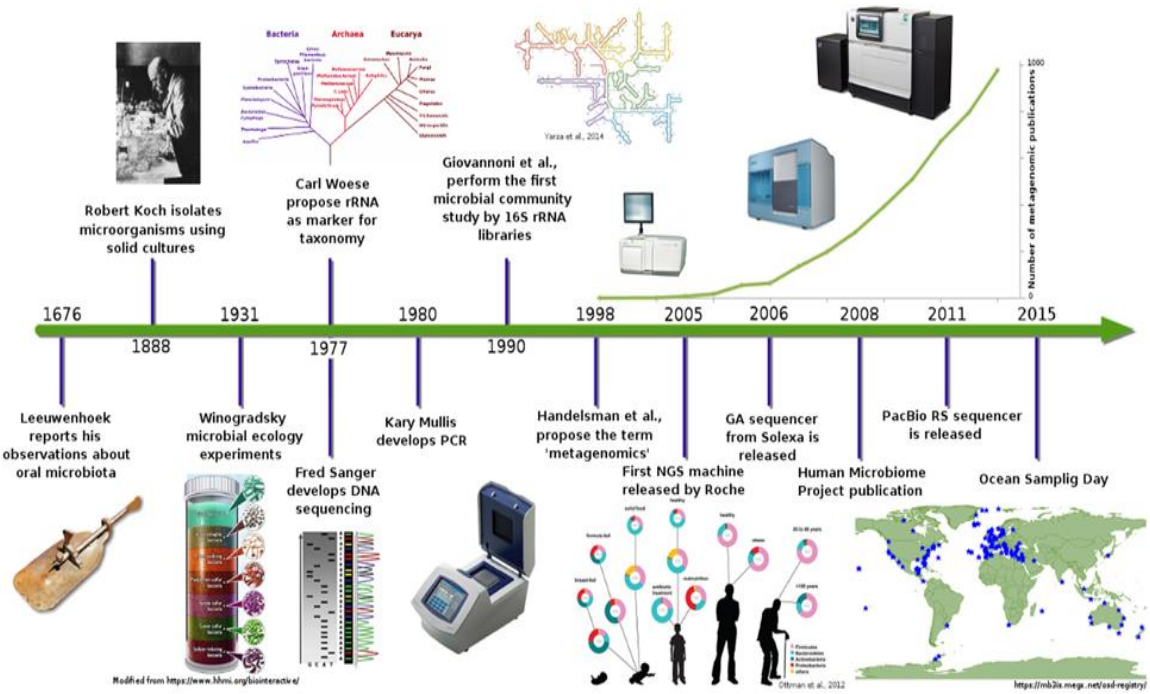


Figura 1. Evolució de la metagenòmica des del seus inicis en la microbiologia clàssica (Escobar-Zepeda et al., 2015)**.**

Un del àmbits amb més aplicabilitat d'aquestes tecnologies és el de l'estudi del microbioma humà i en concret el del microbioma intestinal humà, sobretot relacionat amb les malalties anomenades IBD (*Inflammatory Bowel Diseases* en anglès). La composició del microbioma intestinal humà no és estable ni específic a nivell d'espècie sinó que pot variar entre individus, i al llarg del temps en un mateix pacient (Amos et al., 2021; Uhr et al., 2019). No obstant, desequilibris sobtats en la composició microbiana, com per exemple, durant el tractament amb antibiòtics, pot conduir a la colonització d'altres microorganismes oportunistes i potencialment patògens. Aquests canvis en la composició del microbioma respecte a les condicions fisiològiques són anomenades disbiòsis i poden afectar la salut de l'hoste a nivell metabòlic, cardiovascular, d'inflamació o de desenvolupament de resistència a certs antibiòtics (Nagao-Kitamoto et al., 2016; Shehata et al., 2022). En aquest sentit, sorgeix la necessitat d'establir estàndards específics que serveixin com a referència per tal de identificar i classificar aquests tipus d'alteracions (Almeida et al., 2021; Forbes et al., 2018; King et al., 2019; Lloyd-Price et al., 2016; Nelson et al., 2010).

En l'actualitat existeixen nombrosos estudis publicats on s'han determinat certs patrons de disbiòsi relacionats amb diferents IBDs com puguin ser la malaltia de Crohn o la Colitis Ulcerosa (CD i UC respectivament per les seves sigles en anglès) (Amos et al., 2021; Gonzalez et al., 2022; Pisani et al., 2022). No obstant, aquests patrons no sempre coincideixen depenen de l'estudi, de la malaltia, del tipus de dades, i del tractament que s'hagi dut a terme d'aquestes. També hi juga un paper important l'edat, el sexe la regió geogràfica i altres possibles factors ambientals (Loftus, 2004). En aquest sentit, aquest treball pretén comprovar com de reproduïbles són els resultats obtinguts a partir de mostres provinents de diferents estudis previs i quines podrien ser les estratègies a seguir per tal de millorar-ne la reproductibilitat.

1.2. Objectius del Treball

Objectius generals:

1- Classificar taxonòmicament els microorganismes presents en mostres fecals de pacients amb UC i controls sans partint d'una cohort de *discovery* (CD).

2- Extreure un patró de disbiòsi a partir de la cohort *discovery* i comprovar com de reproduïble és en una cohort diferent de *validation* (CV). Comprovar també si aquest patró es pot utilitzar per fer prediccions en mostres no classificades.

Objectius específics:

1- Aconseguir suficients dades de pacients amb UC i pacients sans prou semblants com per obtenir dos cohorts *discovery* i *validation* amb les quals poder córrer el *pipeline* de classificació taxonòmica.

2- Instal·lar i utilitzar el programari necessari que ens permeti elaborar i utilitzar el *pipeline* de classificació taxonòmica.

3- Implementar un *pipeline* de classificació taxonòmica de mostres de pacients sans i amb UC.

4- Extreure un patró de disbiòsi de les mostres que conformen la cohort de *discovery* per tal de poder-lo validar creuant-lo amb la cohort de *validation*.

5- Avaluar la reproductibilitat de la disbiòsi entre diferents estudis, identificar els paràmetres que n'afecten la seva possible extrapolació i proposar possibles millores.

6- Desenvolupar un mètode de predicció a partir dels paràmetres de disbiòsi que ens permeti predir si una mostra no classificada és provinent d'un pacient amb o sense UC.

1.3. Impacte en sostenibilitat, ètic-social i de diversitat

La metagenòmica és una disciplina enormement potent per estudiar les comunitats microbianes i ha guanyat una atenció significativa en els últims anys. Permet explorar de manera exhaustiva la composició genètica dels ecosistemes microbians, permetent als investigadors investigar patrons de disbiòsi, que fan referència a desequilibris en la composició i funció de les comunitats microbianes. Tot i que l'anàlisi de metagenòmica té un enorme potencial per entendre les disbiòsis i implicacions per a la salut humana, és essencial considerar les implicacions ètiques i el compromís global associat amb aquest enfocament.

Consideracions Ètiques:

Consentiment Informat: La realització de l'anàlisi de metagenòmica requereix obtenir el consentiment informat de les persones les mostres de les quals s'estudien. Els investigadors han de garantir que els participants comprenguin plenament els riscos i beneficis potencials, la naturalesa de l'estudi i el tractament de les seves dades.

Privadesa i Protecció de Dades: La gran quantitat de dades generades per l'anàlisi de metagenòmica planteja preocupacions sobre la privadesa i la protecció de dades. Salvaguardar la informació personal i genètica de les persones és crucial per prevenir l'accés no autoritzat, l'ús indegut o l'estigmatització.

Compartir Beneficis: Els resultats de la recerca en metagenòmica poden tenir un valor comercial significatiu. Assegurar un repartiment equitatiu dels beneficis amb les comunitats i les persones de les quals s'obtenen les mostres és essencial, especialment quan es realitzen recerques en entorns de recursos limitats.

Compromís Global:

Recerca col·laborativa:. Fomentar la col·laboració internacional i compartir dades i recursos pot ajudar a accelerar el progrés científic, millorar la comprensió de la disbiòsi i fomentar els esforços col·lectius per abordar els reptes globals de salut. En aquests sentit, la cerca de possibles patrons de disbiòsi s'ha d'abordar des d'un punt de vista global de manera que el coneixement derivat pugui transcendir les fronteres geogràfiques, les diferències entre edat, sexe, ètnia, etc.

Abordar les Desigualtats en Salut: Les disbiòsis afecten sovint de manera desproporcionada a les poblacions vulnerables. Els investigadors que utilitzen els anàlisi metagenòmics han de comprometre's a abordar les desigualtats en salut i assegurar que els resultats contribueixin a millorar l'accés a l'atenció sanitària, especialment en les comunitats marginades.

Sostenibilitat Ambiental: La disbiòsi va més enllà de la salut humana i comprèn desequilibris ecològics. Incorporar aspectes de sostenibilitat ambiental en la recerca en metagenòmica pot conduir a una millor comprensió de la interacció entre la salut humana, els ecosistemes i la conservació de la biodiversitat contribuint així al concepte de ONE HEALTH.

La metagenòmica ofereix un enorme potencial per descobrir patrons de disbiòsi i aprofundir en la comprensió de les seves implicacions. No obstant això, és crucial navegar per les consideracions ètiques relacionades amb el consentiment informat, la privadesa i el repartiment de beneficis. A més, un compromís global en la recerca col·laborativa, l'abordatge de les desigualtats en salut i la sostenibilitat ambiental pot maximitzar l'impacte positiu de la metagenòmica en la millora de la salut humana i el benestar ecològic a escala global. Integrant aquests principis ètics i compromisos globals, els investigadors poden garantir que l'anàlisi de metagenòmica es realitzi de manera responsable i amb una perspectiva social més ampla.

1.4. Enfocament i mètode seguit

Com que la gran majoria de microorganismes presents en mostres fecals (o altres ambients) no són cultivables, l'única manera de conèixer-ne la composició és a través de tècniques de seqüenciació. Aquestes tècniques ens permeten determinar les seqüències dels seus genomes, ja sigui a nivell complert o parcials. A l'actualitat existeixen dues tecnologies diferents que permeten conèixer aquestes seqüències: *shotgun metagenomic sequencing* i *16S rRNA*. La primera es basa en la seqüenciació completa de tot el DNA present a la mostra, mentre que el 16S rRNA es focalitza en seqüenciar una o varies regions hipervariables del gen 16S bacterià.

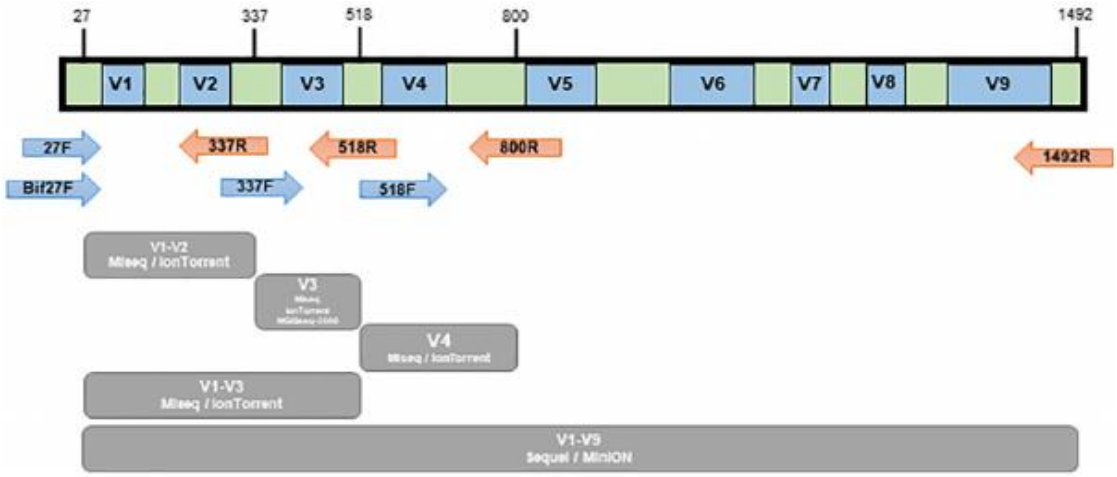


Figura 2 Esquema del gen 16S procariota (Park et al., 2021)

A la següent taula es detallen els pros i els contres de les dues tecnologies de seqüenciació 16S i *Shotgun*:

	16S rRNA	Shotgun Sequencing
Bacterial coverage	High	Limited
Cross-Domain Covergae	No	Yes
Fasle Positives	Low Risk	High Risk
Taxonomy Resolution	Genus-Species	Species-Strains
Host DNA Interference	No	Yes
Functional Profiling	No	Yes
Minimum DNA Input	10 copies of 16S	1 ng
Recommended Sample type	All	Human Microbiome
Cost per Sample	~$80	~$200

Taula 1 Avantatges i inconvenients dels mètodes 16S i *Shotgun*.

S'ha escollit treballar a partir de dades de 16S rRNA enlloc de dades de *shotgun metagenomic sequencing* per dos motius diferents: primer per una limitació de la capacitat computacional. Al no disposar d'un servidor prou potent el fet d'utilitzar dades de *shotgun metagenomic sequencing* podria fer que la tasca del seu processament fos lenta i tediosa. I en segon lloc, hi ha alguns articles que apunten que no s'observen diferències de resultats en utilitzar una naturalesa de dades o una altra (Serrano-Gómez et al., 2021; Zuo et al., 2022).

Les dades de seqüenciació metagenòmica de 16S rRNA es poden obtenir en diverses bases de dades i a partir d'estudis que les hagin publicat. En el cas que en ocupa ens centrarem en dades d'estudis previs que hagin publicat les dades crues al SRA (*Sequence Read Archive*) de la base de dades del NCBI (Amos et al., 2021; L. Dai et al., 2021; Forbes et al., 2018; Pisani et al., 2022; Zakerska-Banaszak et al., 2021; Zuo et al., 2022). El nombre de dades que se'n pot extreure és suficient per a dur a terme el treball que ens ocupa i a més tant les dades com les metadades es troben en els mateixos repositoris facilitant-ne l'exportació i la lectura de les mateixes.

Per desenvolupar el *pipeline* de classificació taxonòmica s'ha utilitzat el software conegut com a *Mothur* (Schloss et al., 2009). Per fer-ho no hem optat per la instal·lació en local de *mothur* sinó que hem utilitzat el *dockerfile* de *docker* corresponent. *Docker* és un programari similar a *VirtualBox* que conté una sèrie de màquines virtuals (contenidors) prèviament configurades anomenades *dockerfiles*, cada una amb el seu programari i les pertinents dependències necessàries (imatges) sense la necessitat d'haver d'instal·lar *Mothur* (o altres) al nostre ordinador personal.

Posteriorment hem passat a realitzar l'anàlisi estadístic i visualització dels resultats de la classificació taxonòmica. En aquests sentit, Rstudio ofereix una sèrie de paquets amb funcions especifiques per treballar amb dades metagenòmiques com poden ser *mbtools*, *microbiomeR*, *metagenomeSeq*, *phyloseq*, *HMP* o *dysbiosisR* entre d'altres (Shetty & Lahti, 2019). Segons hem anat obtenint els diferents resultats n'hem emprat uns o altres.

Per la tasca de predicció es poden utilitzar diferents aproximacions com per exemple els models lineals generalitzats (GLMs), tècniques d'aprenentatge automàtic o *Machine Learning* (ML) (Dhungel et al., 2021; Serrano-Gómez et al., 2021) o bé algoritmes basats en l'abundància relativa de taxons diferents. En el cas que ens ocupa ens hem decantat per algoritmes de ML donat que són els que s'han treballat més durant el màster. No obstant, també hem utilitzat altres metodologies com les que ofereixen els paquets de R *selbal* (Rivera-Pinto et al., 2018) i *coda4microbiome* (Calle et al., 2023) basats en mètodes estadístics.

1.5. Planificació del Treball

Recursos necessaris

A part d'un ordinador personal amb connexió a internet podem distingir entre dos tipus de recursos necessaris pel desenvolupament del treball:

- Les dades crues de les quals parteix el nostres treball.
- El programari necessari per poder dur a terme els diferents anàlisis.

Tal i com s'ha mencionat en els objectius la idea principal del treball és la de comparar dues cohorts de dades provinents d'estudis diferents, fer-ne el mateix anàlisi metagenòmic i comparar-ne els resultats. Per tant, el primer pas és recopilar aquestes dades. S'ha prioritzat la cerca d'articles científics que haguessin publicat les dades crues de seqüenciació a la base de dades NCBI concretament al SRA (*Sequence Read Archive*). El SRA és un repositori públic de dades de seqüenciació on s'hi publiquen moltes de les dades que generen grups de recerca d'arreu del món.

Com a cohort *discovery* hem escollit les dades de l'article ***A comparative study of the gut microbiota in immune-mediated inflammatory diseases—does a common dysbiosis exist?*** (Forbes et al., 2018) i com a cohort *validation* ens hem decantat per ***Dysbiosis of gut microbiota in Polish patients with ulcerative colitis: a pilot study*** (Zakerska-Banaszak et al., 2021). A continuació presentem una taula resum de les característiques de les dues cohorts.

cohort	DISCOVERY	VALIDATION
autor	Forbes et al., 2018	Zakerska-Banaszak et al., 2021
doi	10.1186/s40168-018-0603-4	10.1038/s41598-021-81628-3
any publicació	2018	2021
país	Canada	Polònia
Codi BioProject	PRJNA450340	PRJNA679275
Instrument de seqüenciació	Illumina MiSeq	Illumina MiSeq
LibraryLayout	Paired	Paired
nº de mostres HC	47 (en descartem 7 a l'atzar)	10
nº de mostres UC	40	10
regió amplificada	V4	V3 i V4

Taula 2 Resum de les característiques principals dels estudis que conformen les cohorts a comparar.

Per accedir a aquestes dades i posteriorment poder treballar-hi es requereix de software especialitzat. Tot seguit detallem els diferents programes bioinformàtics utilitzats i per a què serveixen:

SRA-toolkit: com el seu nom indica és un programa del propi SRA des d'on extraiem les dades de cadascuna de les mostres d'interès. Per fer-ho només es necessita un fitxer de text tipus .txt amb la llista dels codis d'accés de les mostres. Ens podem descarregar aquest fitxer fàcilment des de *SRA Run Selector* del NCBI. També amb SRA-toolkit hem fet la conversió dels *reads* a format fastq.

Mothur: *Mothur* és un programari informàtic dissenyat específicament per a l'anàlisi de dades de 16S rRNA per a l'estudi de la diversitat microbiana. *Mothur* facilita el processament, l'anàlisi i la interpretació de les dades de seqüenciació d'ADN en estudis de diversitat microbiana. Aquest programa ofereix un conjunt complet de funcions per realitzar tasques com el filtratge i l'eliminació de dades de baixa qualitat degudes a errors d'amplificació per PCR, l'agrupació de seqüències similars per a la formació de taules d'unitats taxonòmiques operatives (OTUs) i la generació d'anàlisis de diversitat i visualització de dades.

Figura 3 Logotip de Mothur

Com a avantatges *Mothur* permet personalitzar l'anàlisi de dades i adaptar-la a les seves necessitats específiques. Ofereix un ampli ventall d'eines i opcions de configuració per a l'anàlisi de la diversitat microbiana. També integra una gran quantitat de funcionalitats en una sola plataforma, evitant la necessitat de utilitzar diverses eines separades per a cada pas de l'anàlisi. Una altra avantatge és que *Mothur* compta amb una comunitat d'usuaris àmplia i activa, el que implica que hi ha un suport continu i actualitzacions periòdiques per part dels desenvolupadors. Això ajuda a resoldre problemes i a millorar el programari constantment.

Per contra l'ús de *Mothur* pot ser complex, especialment per a usuaris amb poca experiència en l'anàlisi de dades de seqüenciació. A més, l'anàlisi de seqüències genètiques a gran escala pot requerir una potència de càlcul significativa. És possible que l'ús de *Mothur* en grans conjunts de dades necessiti recursos computacionals addicionals per a l'execució eficient dels processos.

Per tal d'utilitzar aquests dos programes (*SRA-toolkit* i *Mothur*) ho hem fet a través de *dockerfiles* de *Docker*.

Docker és una plataforma de codi obert que permet als desenvolupadors construir, distribuir i executar aplicacions en un entorn aïllat conegut com a contenidor. Els contenidors són entitats lleugeres i portables que empaqueten tots els components necessaris per a l'execució d'una aplicació, com ara el codi, les llibreries, les dependències i les configuracions.

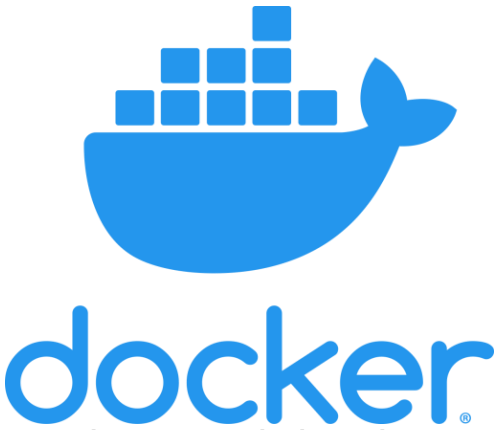


Figura 4 Logotip de Docker.

Un *Dockerfile* és un fitxer de text que conté un conjunt d'instruccions que descriuen com construir una imatge de contenidor. Una imatge de contenidor és una instantània immutable i lleugerament diferent del sistema d'arxius que s'executa en un contenidor *Docker*.

Per tal de poder treballar amb *Docker* s'ha de fer des d'un entorn *Linux*. Per tal de no haver de canviar de sistema operatiu en el nostre ordinador personal hem optat per treballar des de una màquina virtual.

Una màquina virtual és un programari que permet la creació i l'execució d'un entorn virtual complet dins d'un sistema operatiu amfitrió (*Windows* en el nostre

cas). Aquest entorn virtual actua com una màquina independent, amb el seu propi sistema operatiu, recursos de maquinari virtualitzat i aplicacions instal·lades.

VirtualBox permet als seus usuaris crear diverses màquines virtuals en un sol ordinador físic, permetent l'aïllament i la simultaneïtat de diferents sistemes operatius, facilitant així el desenvolupament, la prova de programari i l'execució d'aplicacions en entorns controlats i virtuals, sense afectar el sistema operatiu principal de l'ordinador. En el nostre cas hem creat, a través de *VirtualBox*, dos sistemes operatius *Ubuntu* (*Linux*) de 64 bits (un per cada cohort) des d'on s'ha dut a terme tota la part d'obtenció de dades i classificació taxonòmica d'aquestes.

Figura 5 Logotips de VirtualBox i Ubuntu.

RStudio: La fase de anàlisis *downstream* s'ha dut a terme mitjançant RStudio. RStudio és un entorn de desenvolupament integrat (IDE, per les seves sigles en anglès) específicament dissenyat per a la programació en el llenguatge de programació R. És una aplicació de codi obert que proporciona un conjunt de eines i funcionalitats per a la visualització de dades, l'execució de codi, la depuració, la generació de gràfics i altres tasques relacionades amb l'anàlisi estadística i científica de dades.

Figura 6 Logotip de RStudio.

Tasques

Data gathering

Recol·lecció de dades: aquesta tasca s'ha estès fins la meitat de la fase 1 del desenvolupament del treball donat que s'han hagut de fer proves, canviar o actualitzar les dades per tal de poder desenvolupar les diferents tasques d'una manera més efectiva.

Software installation

Instal·lació del software que ens ha permès fer la part *upstream* de l'anàlisi de les dades. És a dir, tota la preparació prèvia a l'anàlisi estadístic. Aquesta part no s'ha estès més enllà de la primera setmana i mitja de la fase 1. Aquesta tasca ha estat composta de la instal·lació del *Docker* amb el *dockerfile* que ens ha permès treballar amb *Mothur*.

Upstream analysis

Aquesta part es basa en el pre-processament de les dades per tal de dur a terme la classificació taxonòmica amb la qual després s'ha fet l'anàlisi dels possibles patrons de disbiòsi. Està composta dels següents passos i no ha durat més de 3 setmanes des de la instal·lació del software. Correspon al primer dels 2 objectius generals.

Preparing data (cleaning, filtering, trimming...)

Un cop hem tingut el *Mothur* instal·lat i les dades importades les hem començat a preparar. *Mothur* té diverses funcions que permeten desfer-nos de tots aquells *reads* que no ens interessin per al nostre anàlisi.

Allining to reference sequence

Per tal de poder fer la classificació taxonòmica s'han d'alinear els *reads* a algun tipus de seqüències de referència. Existeixen diferents bases de dades per alinear *reads* de 16S rRNA. *Mothur* treballa per defecte amb la base de dades SILVA (Quast et al., 2012) però es pot canviar per alguna altra com per exemple Greengenes (DeSantis et al., 2006). Això ens retorna un arxiu en format festa dels alineaments.

Figura 7 Logotip de la base da dades SILVA.

Filter, removing duplicates, pre-cluster, chimera removing

Els alineaments també es poden filtrar per tal de eliminar tots aquells que no ens interessin.

Taxonomy Classifying

Per fer la classificació taxonòmica s'utilitza una base de dades diferent a la que s'utilitza per fer l'alineament.

Remove lineage

Aquesta part permet descartar aquelles seqüències que hagin estat classificades com a no bacterianes com puguin ser cloroplasts, mitocòndries, archaeas o eucariotes.

Prepare final fasta, final taxnomy and final count table
Es preparen els arxius finals fasta, *taxonomy* i *count table* amb els alineaments, la classificació taxonòmica i el numero de vegades que apareix cada *read* en cada mostra respectivament.

Downstream analysis

Aquesta part s'ha dut a terme mitjançant Rstudio i correspon al segon dels 2 objectius generals anteriorment esmentats (Shetty & Lahti, 2019). Inicialment no havia de durar més de les dues primeres setmanes de la fase 2 del desenvolupament però algunes tasques d'aquesta part s'han anat millorant al llarg de les setmanes posteriors.

Relative abundance visualization
Visualització mitjançant gràfics de l'abundància relativa dels diferents taxons en les diferents mostres.

Hierarchical clustering
La clusterització jeràrquica ens permet visualitzar com s'agrupen les diferents mostres segons la seva similitud/proximitat.

Alpha-diversity
La diversitat alfa és aquella que es troba dins una mateixa mostra.

Beta-diversity
La diversitat beta en canvi proporciona una mesura de similitud (o dissimilitud) entre diferents mostres.

Principal Components analysis
L'anàlisi de components principals redueix el nombre de dimensions a 2 i això ens permet visualitzar millor les similituds entre les mostres.

Differential abundance analysis
Aquest anàlisi ens permet conèixer quines unitats taxonòmiques són significativament diferents respecte els diferents grups.

Prediction

L'apartat de predicció s'ha basat en tècniques de ML com puguin ser els algoritmes supervisats basats en mètodes tipus *Random Forests Classification*. També hem optat pel paquet de R *selbal* basat en regressions múltiples i que es focalitza en buscar balanços composicionals entre grups taxonòmics que tinguin valor predictiu d'interès, ja sigui per establir un diagnòstic, una prognosi o una determinada resposta a un tractament. També s'ha utilitzat el paquet *coda4microbiome* que és com una versió actualitzada i millorada de *selbal* però que es basa en la naturalesa composicional de les dades de microbioma. Aquest apartat hauria de tenir una extensió d'uns 10-12 dies aproximadament.

Planificació temporal

El calendari proposat es mostra a continuació mitjançant un diagrama de Gantt amb les 4 PACs successives, la preparació de la defensa i les fases 1 i 2 del desenvolupament del treball desglossades en diferents tasques. La redacció de cadascuna de les PAC s'ha començat al principi de cadascuna de les dues fases de desenvolupament del treball. De la mateixa manera, la redacció de la memòria final i la presentació tenen inici al principi del propi treball i s'han anat redactant a mesura que s'ha avançat en les diferents tasques proposades.
Les diferents tasques que ocupen la fase 1 i 2 estan enllaçades mitjançant l'assoliment de diferents fites necessàries per tal de poder continuar amb el desenvolupament de la següent tasca.

Diagrama de Gantt:

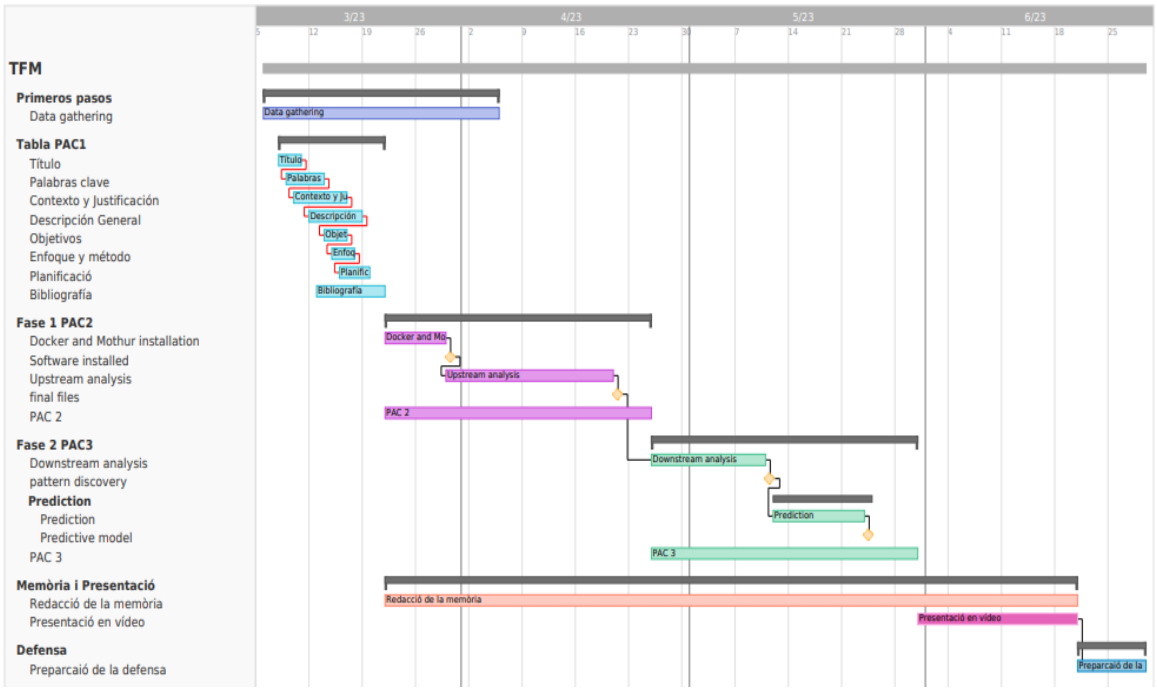


Figura 8 Diagrama de Gantt amb la planificació temporal del treball.

2. Estat de l'art

L'estudi del microbioma humà mitjançant l'anàlisi metagenòmic és un camp actiu de recerca que busca comprendre la composició i funció dels microorganismes que habiten en el cos humà. El microbioma humà és l'ecosistema complex de bacteris, virus, fongs i altres microorganismes que resideixen en diferents parts del nostre cos, com ara la pell, la boca, el tracte gastrointestinal i altres nitxos ecològics potencials.

L'anàlisi metagenòmic implica l'extracció de mostres biològiques, com ara mostres fecals o mostres de saliva, i la seqüenciació de l'ADN o l'ARN present en aquestes mostres. Aquesta seqüenciació proporciona informació sobre el genoma dels microorganismes presents en el microbioma, permetent identificar-los i caracteritzar-los.

L'ús de tècniques metagenòmiques ha revolucionat l'estudi del microbioma humà, ja que permet analitzar l'ADN o l'ARN de tots els microorganismes presents en una mostra sense la necessitat de cultivar-los en el laboratori, fet que tot sovint és inviable donat que molts microorganismes presents en aquests ambients a vegades no cultivables. Això ha permès descobrir una gran diversitat de microorganismes que abans eren difícils de detectar i estudiar. A continuació es mostra una representació esquemàtica del diferents processos d'anàlisi metagenòmic existent:

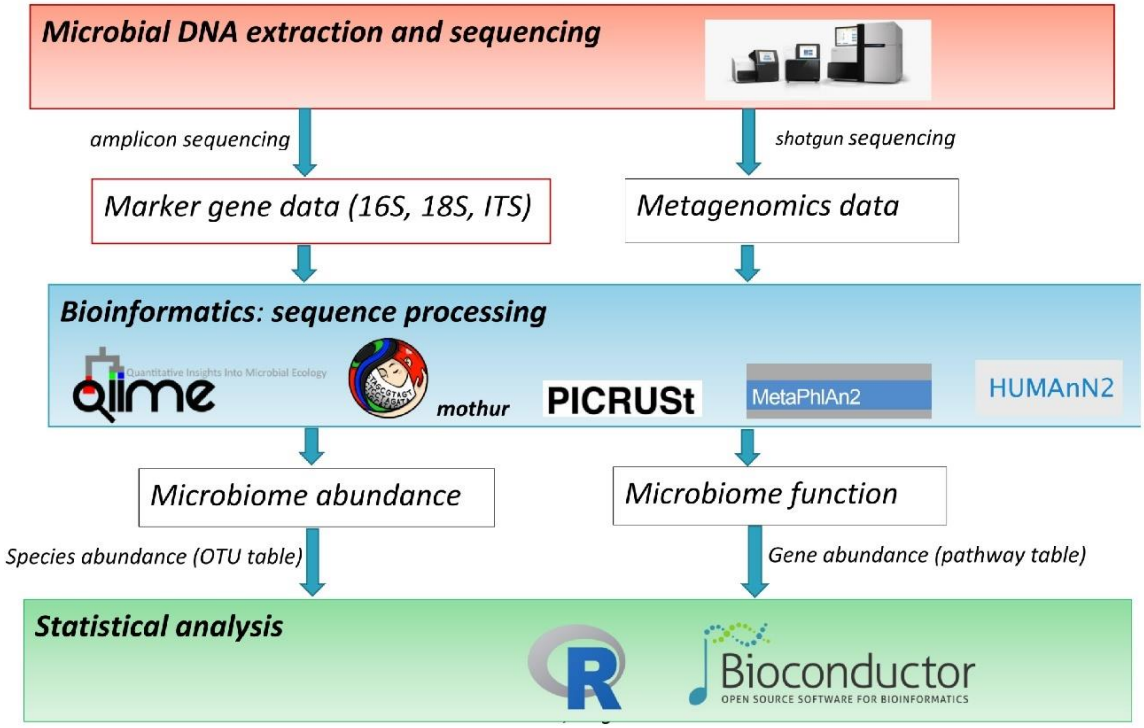


Figura 9 Esquema de l'estudi del microbioma (Luz Calle, 2019).

Actualment, l'estudi del microbioma humà mitjançant l'anàlisi metagenòmic s'ha aplicat a diverses àrees de recerca, incloent la salut gastrointestinal, la salut de

la pell, les malalties autoimmunes, les malalties cardiovasculars i fins i tot la salut mental. A continuació, es presenten algunes àrees destacades d'estudi:

Salut gastrointestinal:

S'ha demostrat que l'anàlisi metagenòmic del microbioma intestinal pot ajudar a comprendre la relació entre els microorganismes intestinals i la salut humana. S'han realitzat estudis per identificar les diferències en la composició del microbioma entre persones sanes i pacients amb malalties com ara les anomenades malalties inflamatòries intestinals (IBD), la síndrome del còlon irritable, la malaltia de Crohn, o la colitis ulcerosa com en el cas que ens ocupa (Bakir-Gungor et al., 2022; Y. Ma et al., 2021; Xu et al., 2022).

Dermatologia:

El microbioma cutani ha estat objecte d'estudi per comprendre el seu paper en condicions com l'acne, l'èczema i altres trastorns de la pell. L'anàlisi metagenòmic ha permès identificar les espècies bacterianes i fúngiques presents en la pell i estudiar les seves interaccions amb l'amfitrió humà (Ferretti et al., 2017; Godlewska et al., 2020; Nagar & Hasija, 2018).

Malalties autoimmunes:

S'ha investigat la possible relació entre el microbioma i malalties autoimmunes com l'artritis reumatoide o la psoriasi. L'anàlisi metagenòmic ha revelat diferències en la composició del microbioma entre pacients amb aquestes malalties i individus sans, suggereix que els microorganismes poden influir en la resposta immune i l'aparició de les malalties autoimmunes (Gupta et al., 2021; Korotky & Peslyak, 2020).

Salut mental:

S'ha proposat una possible connexió entre el microbioma intestinal i la salut mental, incloent els trastorns com la depressió, l'ansietat i l'autisme. L'anàlisi metagenòmic ha revelat que les diferències en la composició del microbioma poden estar associades a alteracions en el funcionament cerebral i en els neurotransmissors. Això obre la porta a noves investigacions sobre com els microorganismes intestinals poden influir en la salut mental i com els canvis en la composició del microbioma poden tenir efectes en els trastorns mentals.

Alzheimer:

Hi ha evidències cada vegada més forts que suggereixen una relació entre el microbioma i l'Alzheimer. El microbioma és la col·lecció de microorganismes que viuen en simbiosi en el nostre cos, especialment en el tracte gastrointestinal. Estudis recents han demostrat que les alteracions en el microbioma estan associades a canvis en el cervell i l'aparició de la malaltia d'Alzheimer. S'ha observat una disminució en la diversitat bacteriana en individus amb Alzheimer, així com la presència de certes espècies bacterianes específiques en el cervell dels pacients amb la malaltia. Aquestes evidències suggereixen que el

microbioma pot tenir un paper clau en el desenvolupament i progressió de l'Alzheimer, obrint noves possibilitats per a la prevenció i el tractament de la malaltia. Tanmateix, són necessàries més investigacions per comprendre millor aquesta relació complexa (Cammann et al., 2023; Jung et al., 2022; Paley, 2019; Sun et al., 2022).

En general, l'estudi del microbioma humà mitjançant l'anàlisi metagenòmic ha proporcionat una comprensió més profunda de la diversitat microbiana que coexisteix amb nosaltres i dels seus possibles efectes en la salut humana. Aquest enfocament ha contribuït a l'avanç de la medicina personalitzada, ja que permet identificar marcadors microbiològics associats a diferents condicions de salut i malalties. Amb el temps, es preveu que l'ús de l'anàlisi metagenòmica en l'estudi del microbioma humà continuï evolucionant i aportant noves aportacions en àmbits com la prevenció, el diagnòstic i el tractament de diverses malalties (Butler et al., 2023; Lai et al., 2022; Taniya et al., 2022; F. Zhu et al., 2020).

En qualsevol estudi científic i també en els metagenomics, la reproductibilitat és un element clau per garantir la confiança i la validesa dels resultats obtinguts. La capacitat de reproduir els resultats és fonamental per a la comunitat científica, ja que permet verificar i validar les troballes, així com avaluar la consistència dels resultats en diferents contextos i poblacions. La reproductibilitat també facilita la comparació i integració de dades de diferents estudis, fomentant la col·laboració i l'avanç del coneixement en aquest camp. A més, en l'àmbit clínic, la reproductibilitat és essencial per a la traducció de l'anàlisi metagenòmica a la pràctica mèdica i al diagnòstic, ja que assegura la fiabilitat dels resultats utilitzats per prendre decisions en matèria de salut. Per tant, l'establiment de protocols estandarditzats, el registre adequat de les metodologies utilitzades i la compartició de dades són elements clau per assegurar la reproductibilitat i promoure l'avenç en els estudis metagenòmics del microbioma humà.

3. Materials i mètodes

L'anàlisi metagenòmic dut a terme es pot dividir en dues parts principals:

- L'anàlisi *Upstream* (o en contra de la corrent).
- L'anàlisi *Downstream* (o a favor de la corrent).

En aquest treball però, també s'ha dut a terme una última tasca de predicció mitjançant models estadístics i tècniques de *Machine Leearning* (ML). A mode esquemàtic es pot resumir el fluxe de treball dut a terme en la següent figura:

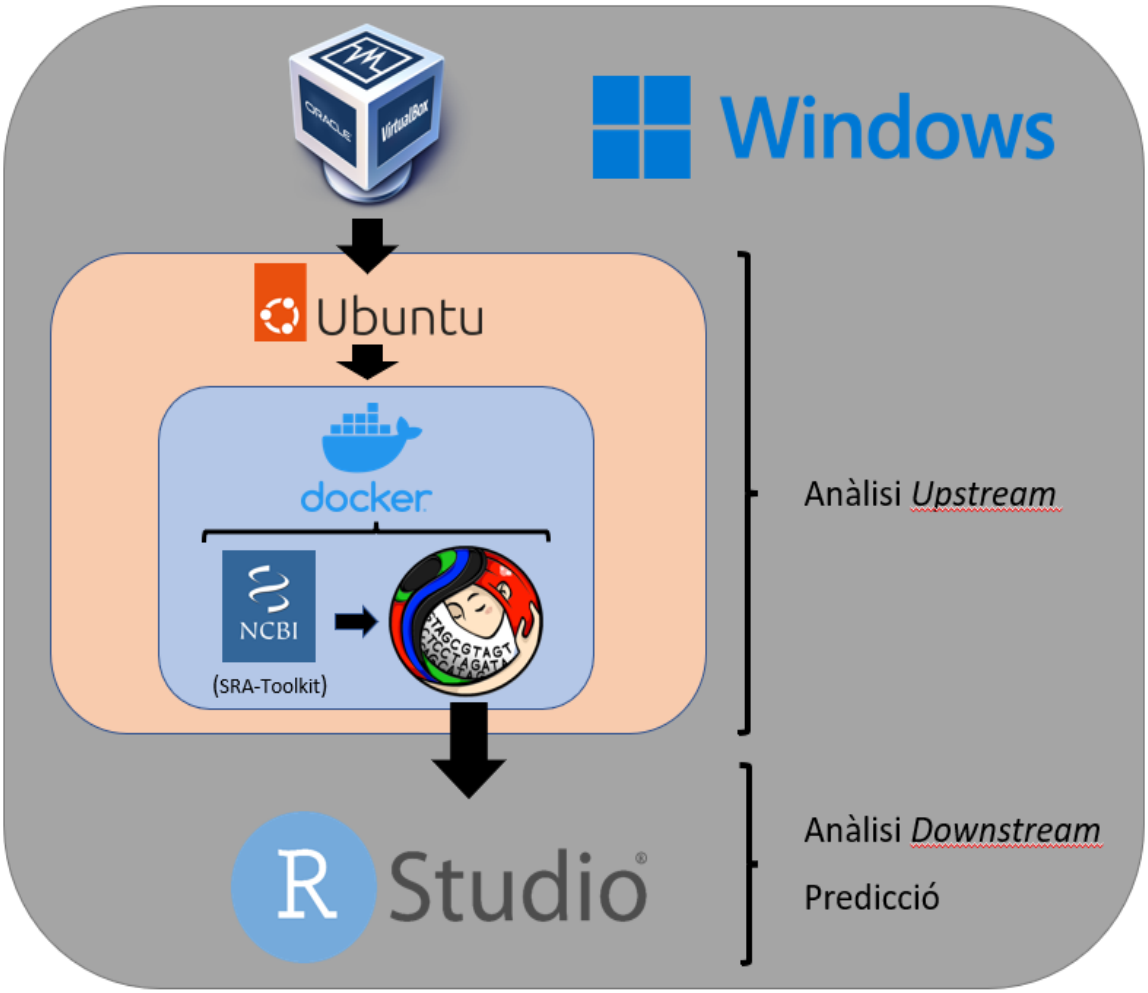


Figura 10 Esquema del *workflow* utilitzat.

No obstant, abans de començar amb l'anàlisi hem d'obtenir les dades crues sobre les quals treballarem i instal·lar els programes necessaris per poder treballar. Tot seguit descrivim aquesta part.

Instal·lació dels softwares i obtenció de les dades

Creació d'una màquina virtual mitjançant *VirtualBox*:

Hem creat una màquina virtual amb suficient espai de memòria a partir d'una imatge d'*Ubuntu* 22.04.2 de 64 bits https://ubuntu.com/download. Per tal de dur a terme una correcta instl·lació podem seguir qualsevol dels nombrosos tutorials existents a internet.

Instal·lació de Docker:

Dins la màquina virtual d'Ubuntu hem instal·lat el *Docker* des de la terminal del mateix sistema operatiu. Per fer-ho correctament hem seguit el tutorial https://www.digitalocean.com/community/tutorials/how-to-install-and-use-docker-on-ubuntu-20-04-es

Pulling dels dockerfiles de *SRA-toolkit* i *mothur*:

Un cop hem tingut el *Docker* instal·lat al sistema operatiu d'*Ubuntu* ens hem baixat les imatges dels dos programes necessaris per tal de poder-nos descarregar les dades i fer-ne la classificació taxonòmica.

Des del repositori de *DockerHub* podem accedir a totes les imatges disponibles d'aquests dos softwares. En el nostre cas hem optat per:

https://hub.docker.com/r/pegi3s/sratoolkit → En el cas de *SRA-toolkit*
https://hub.docker.com/r/biocontainers/mothur → En el cas de *mothur*

Cerca de les dades crues a SRA:

Per tal d'indicar a *sra-toolkit* quines són les mostres que volem obtenir ho hem de fer a través d'un fitxer de text tipus en format .txt que ens podem descarregar fàcilment des de *SRA* de *NCBI*.

Tant en l'article de la CD com en el de la cohort de validació hi ha indicat el codi de *BioProject que* un cop dins del *NCBI* et condueix a *SRA Run Selector*.

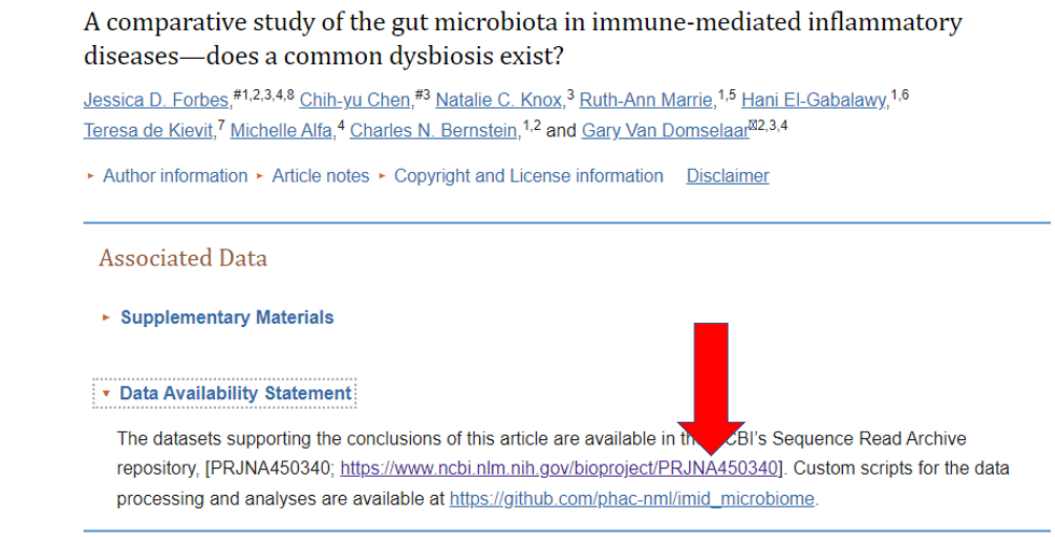
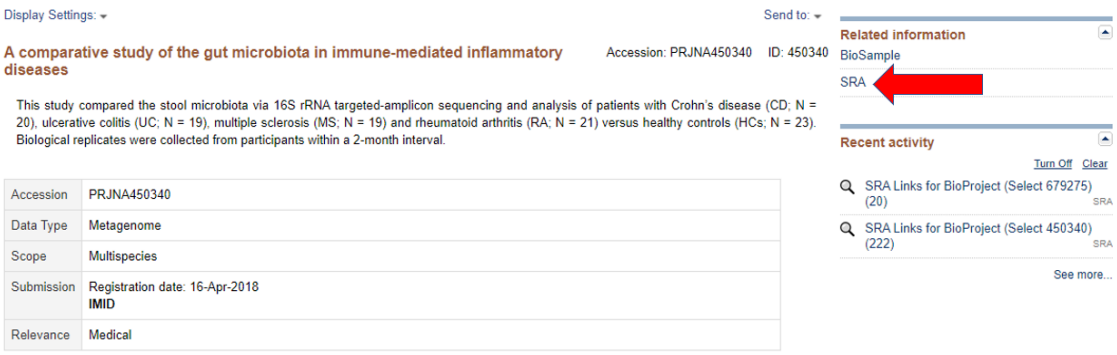
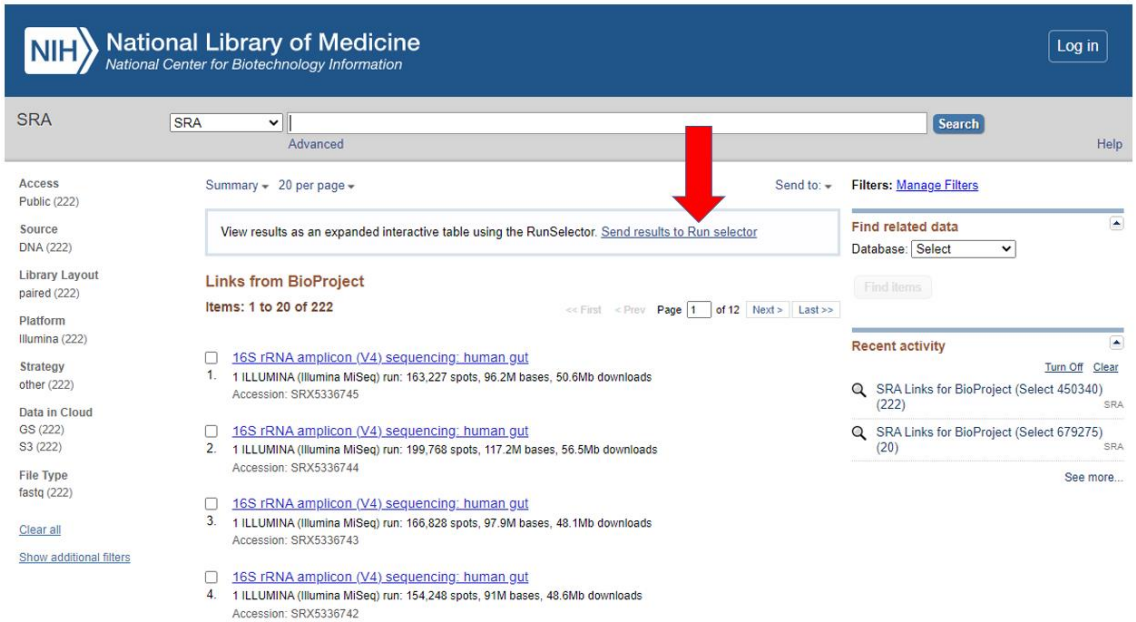


Figura 11 Captura de pantalla de l'article corresponent a la CD. La fletxa indica l'enllaç a la pàgina de BioProject.

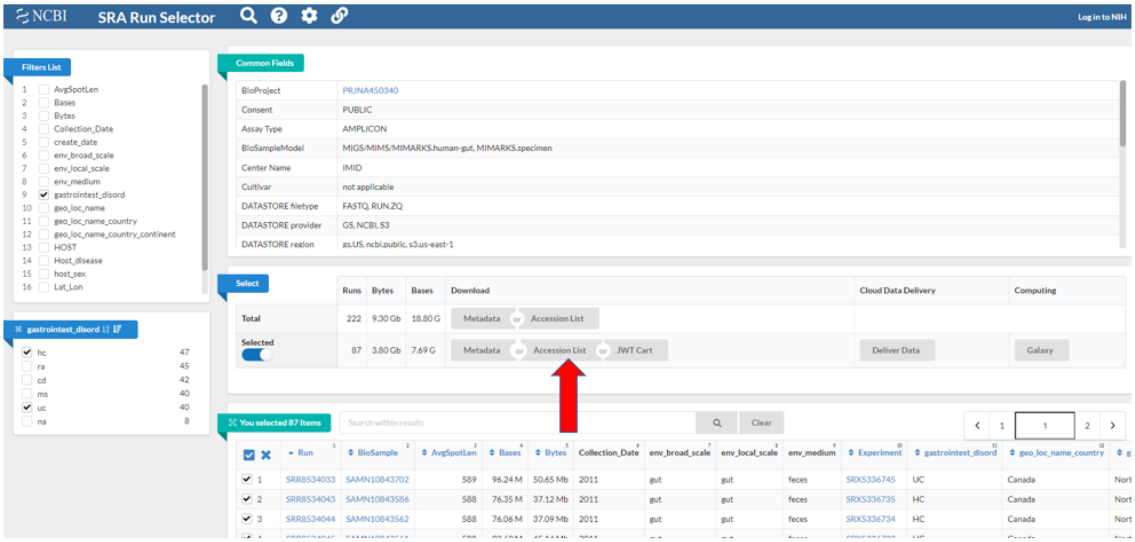
Al seguir l'enllaç arribem a la següent pàgina:



Cliquem a "SRA".



I enviem els resultats al *Run Selector*.



Tal i com es pot veure en els panells de múltiple selecció de la part esquerra de la pantalla, es poden seleccionar les mostres segons les variables de les metadades que més ens interessin. En aquest cas es pot veure com hem seleccionat "*gastrointestinal_disord*" i despres "*hc*" i "*uc*" per tal de seleccionar només les mostres de pacients sans i amb colitis ulcerosa. Un cop feta la selecció cliquem a *Accession List* i automàticament se'ns descarrega el fitxer amb la llista de *Accession numbers* de cada una de les mostres.

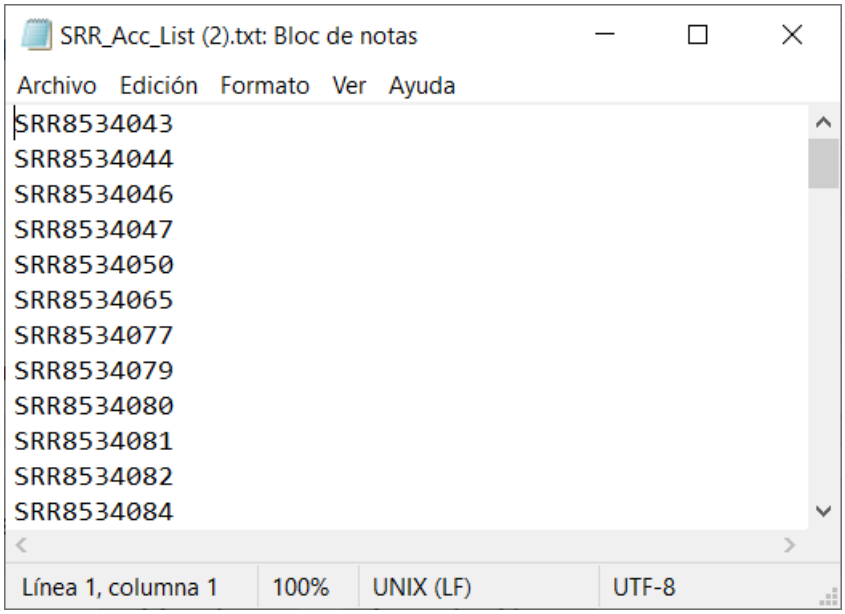


Figura 12 Captura de pantalla del fitxer de text amb la llista de *Accession numbers* de les mostres d'una de les cohorts.

Cada un d'aquests codis correspon a una mostra, és a dir, un pacient de l'estudi. També des de la mateixa pàgina podem descarregar-nos les metadades associades a les mostres seleccionades on s'indica si cada mostra correspon al grup control HC o al grup de pacients amb colitis ulcerosa UC o bé altres variables que puguin ser d'interès, com el sexe, l'edat, etc.

Retrieving de les dades crues des de SRA (Sequence Read Archive):

En aquest punt és on hem començat a treballar amb *Docker*. Es poden consultar les instruccions utilitzades al repositori de *github* d'aquest treball: https://github.com/miquelcastany/metagenomics_TFM. Cridant la imatge de *sra-toolkit* anterior i indicant els directoris d'entrada i de sortida dels corresponents fitxers des de la terminal d'*Ubuntu* ens hem descarregat les dades crues dels dos estudis corresponents a les cohorts de *discovery* i *validation*. Mitjançant *SRA-toolkit* hem emprat dues funcions:

- ***prefetch*** → permet fer el *retrieving* de les dades crues des de NCBI. Aquestes dades es troben en format .sra i els arxius que les contenen no es poden obrir.

- ***fasterqdump*** → converteix els arxius que contenen els *reads* a format fastq. Per a cada fitxer .sra es generen 2 fitxers .fastq corresponents a cada una de les cadenes de ADN de cada *read*.

Un cop tenim les dades descarregades ja podem començar amb l'anàlisi.

Anàlisi *upstream*

Classificació taxonòmica mitjançant Mothur:

Hem implementat un pipeline de classificació taxonòmica amb Mothur basat en el *MiSeq SOP* de la seva pàgina web https://mothur.org/wiki/miseq_sop/. Aquest apartat de la web de *Mothur* explica pas a pas com dur a terme una classificació taxonòmica a partir de les seqüencies del gen 16S bacterià generades amb la plataforma de seqüenciació *MiSeq* de *Illumina* com és el cas de les nostres dades.

Els diferents passos s'han mencionat ja a l'apartat de tasques de la planificació del treball però tot seguit en farem una presentació amb més profunditat. També es poden consultar les diferents comandes del *dockerfile* de *Mothur* emprades al *github* del treball al igual que amb sra-toolkit https://github.com/miquelcastany/metagenomics_TFM.

- ***make.file***() → Aquesta instrucció aparella cadascun dels dos fitxers .fastq pertanyents a una mostra. Per fer-ho tant sols genera un fitxer de text amb 3 columnes: la primera amb el nom de la mostra i la segona i la tercera amb els noms dels dos fitxers .fastq.

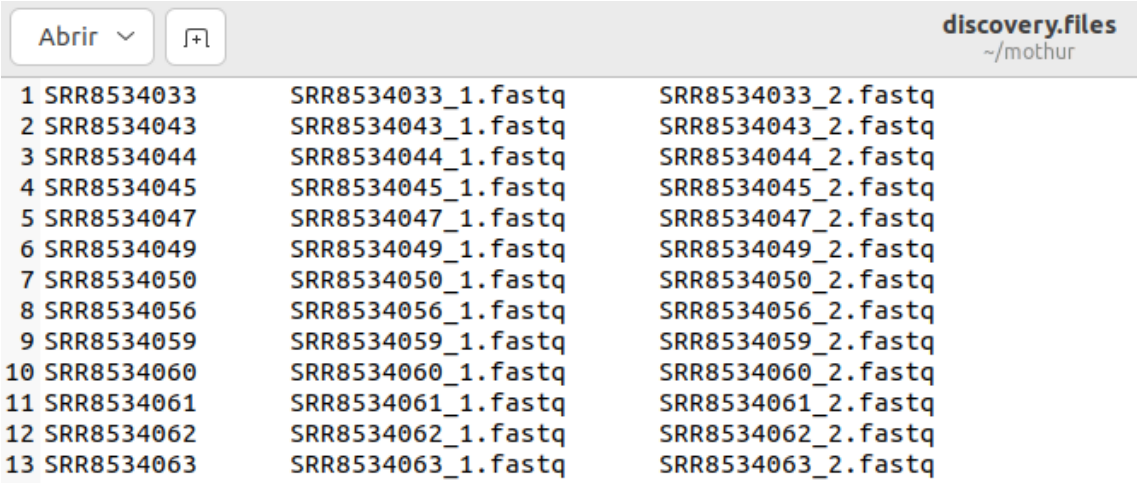


Figura 13 Captura de pantalla del fitxer .file.

- ***make.contigs***() → El que fa make.contigs() és primer combinar els dos grups de reads de cada mostra i després combinar les dades de totes les mostres, és a dir, crea els contigs a partir d'una seqüència i la seva complementaria. Per fer-ho crea la seqüència complementaria de la seqüència *reverse* i els ajunta per formar els contigs. Make.contigs() genera diversos fitxers però el més important és el trim.contigs.fasta, amb el corresponent prefix que haguem indicat al make.file():

Figura 14 Captura de pantalla del fitxer fasta resultant de make.contigs().

- ***summary.seqs()*** → Amb summary.seqs() el que fem és presentar un resum de les característiques dels *reads* compresos en el fitxer que l'hi indiquem. Característiques com el numero de base on comença i on acaba cada *read* el nombre de bases ambigües o la longitud màxima dels homopolímers presents. Summary.seqs() és una instrucció que es va repetint en diversos punts de l'anàlisi.

- ***screen.seqs()*** → Mitjançant screen.seqs() podem filtrar tots aquells *reads* que no ens interessen. Per exemple, podem filtrar per una llargada màxima o mínima, eliminar aquells *reads* que presenten bases ambigües o les que presenten homopolímers més llargs de 8 nucleòtids. Aquests paràmetres varien segons els tipus de dades amb les que estiguem treballant, la seva naturalesa i el propòsit que ens haguem marcat. A l'apartat de resultats descriurem com hem fixat aquests paràmetres d'acceptació o filtratge tant en la cohort *discovery* com *validation*. La instrucció screen.seqs() també l'hem utilitzar

després de fer l'alineament per eliminar aquells alineaments que no ens interessin.

- ***unique.seqs()*** → unique.seqs identifica cada un dels *reads* únics presents. És a dir, elimina les repeticions i es queda amb una sola representació de cada un dels *reads*. També s'executa un cop realitzat l'alineament.

- ***count.seqs()*** → a partir del fitxer resultant de l'screen.seqs() i un altre fitxer .groups generat al principi amb el make.contigs() generem la taula de comptatges o *count_table*. La taula de comptatges és una taula que conté el nombre de vegades que apareix cada *read* a cada una de les mostres que configuren la cohort. La *count_table* és essencial de cara a l'anàlisi *downstream* ja que ens basarem sobretot en les abundàncies relatives.

1	Representative_Sequence	total	SRR8534033		SRR8534043		SRR8534044		SRR8534045		SRR8534047		SRR8534049	
	SRR8534050	SRR8534056	SRR8534059		SRR8534060		SRR8534061		SRR8534062		SRR8534063		SRR8534064	
	SRR8534065	SRR8534069	SRR8534075		SRR8534077		SRR8534079		SRR8534080		SRR8534081		SRR8534084	
	SRR8534087	SRR8534090	SRR8534099		SRR8534100		SRR8534109		SRR8534114		SRR8534123		SRR8534126	
	SRR8534128	SRR8534133	SRR8534143		SRR8534146		SRR8534147		SRR8534148		SRR8534159		SRR8534160	
	SRR8534173	SRR8534182	SRR8534186		SRR8534187		SRR8534199		SRR8534201		SRR8534202		SRR8534204	
	SRR8534205	SRR8534206	SRR8534207		SRR8534208		SRR8534209		SRR8534210		SRR8534213		SRR8534215	
	SRR8534216	SRR8534217	SRR8534218		SRR8534219		SRR8534220		SRR8534221		SRR8534222		SRR8534223	
	SRR8534224	SRR8534225	SRR8534232		SRR8534234		SRR8534235		SRR8534240		SRR8534242		SRR8534243	
	SRR8534244	SRR8534245	SRR8534246		SRR8534247		SRR8534248		SRR8534249		SRR8534251		SRR8534252	
	SRR8534253	SRR8534254												
2	SRR8534064.76778	498	0	0	33	0	7	1	8	0	5	11	0	0
	22	32	0	0	0	0	0	0	10	1	21	0	0	0	0
	9	3	0	14	10	41	4	0	1	21	0	0	5	12	0	25
	0	36	15	0	0	13	4	0	1	0	0	0	0	0	0	10
	11	6	1	4	0	0	1	0	3	4	2	1	0	36	2	20
	0	0	4	28												
3	SRR8534064.107073	1	0	0	0	0	0	0	0	0	0	0	0	0
	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0												
4	SRR8534064.137445	1	0	0	0	0	0	0	0	0	0	0	0	0
	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0												
5	SRR8534064.14795	12958	0	1	862	0	156	46	349	0	222	212	2	0
	717	347	10	2	4	0	0	1	2	253	283	579	0	2	0	1
	332	42	1	190	119	1613	51	1	23	593	2	0	47	236	2	688
	1	441	146	1	0	305	40	0	51	1	1	0	0	1	0	295
	151	82	15	94	0	0	1	0	198	169	43	44	21	355	115	911
	0	0	316	1169												

Figura 15 Captura de pantalla de la taula de comptatges count_table. Es pot veure com algunes seqüències apareixen repetides vegades i en diverses mostres mentre que d'altres tan sols apareixen un únic cop en tot el data set.

- ***align.seqs()*** → En aquesta part procedim a fer l'alineament contra una base de dades d'alineaments de referència, és a dir, prèviament classificats. Això ens permet classificar taxonòmicament cada un dels *reads* almenys fins a rang de gènere. La base de dades escollida és SILVA https://www.arb-silva.de/. Podem descarregar-nos l'arxiu que conté aquesta base de dades a la mateixa pàgina del *MiSeq SOP* de *Mothur*. Hi ha vàries versions disponibles. Nosaltres hem empleat la que apareix enllaçada al principi del *MiSeq SOP*.

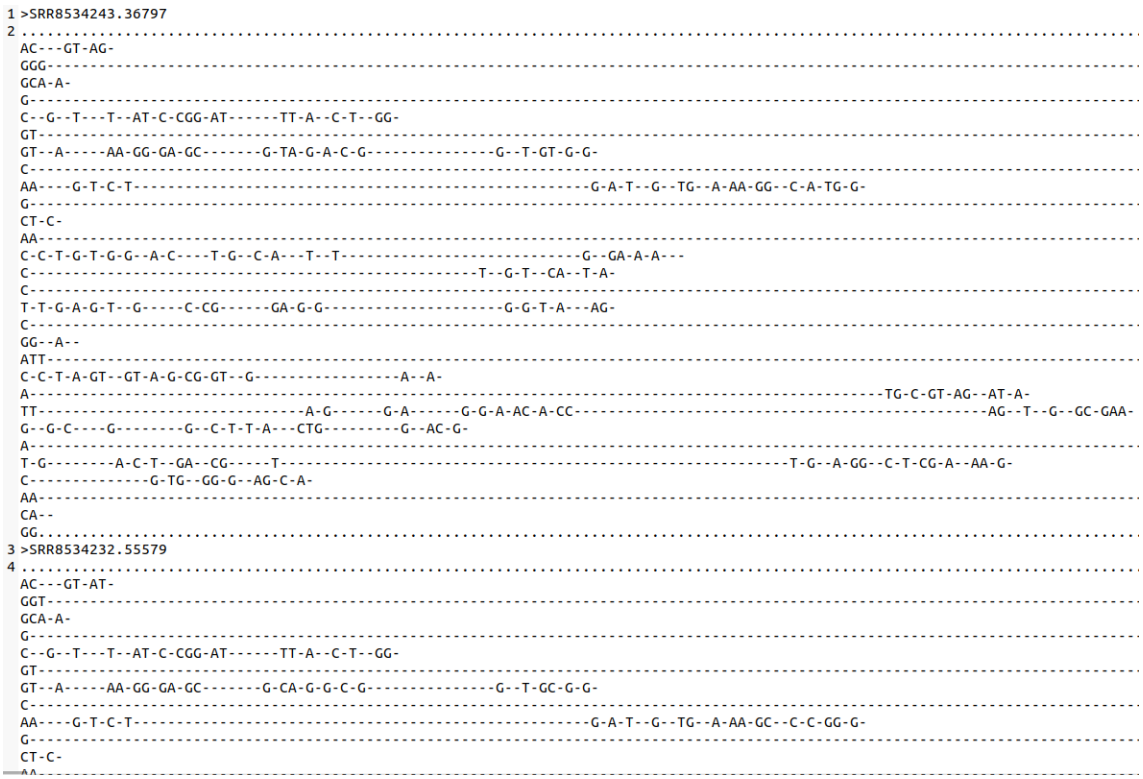


Figura 16 Captura de pantalla del fitxer corresponent a l'alineamnet.

- ***filter.seqs()*** → Aquesta instrucció té dues funcions: la primera és eliminar les regions *overhangs*, és a dir, aquells trams de l'inici i el final que sobresurten de l'alineament pròpiament. També elimina les posicions de l'alineament que contenen el símbol "-" que representa els gaps. Al final de la comanda *Mothur* ens indica la llargada total de l'alineament havent eliminat les columnes amb gaps.

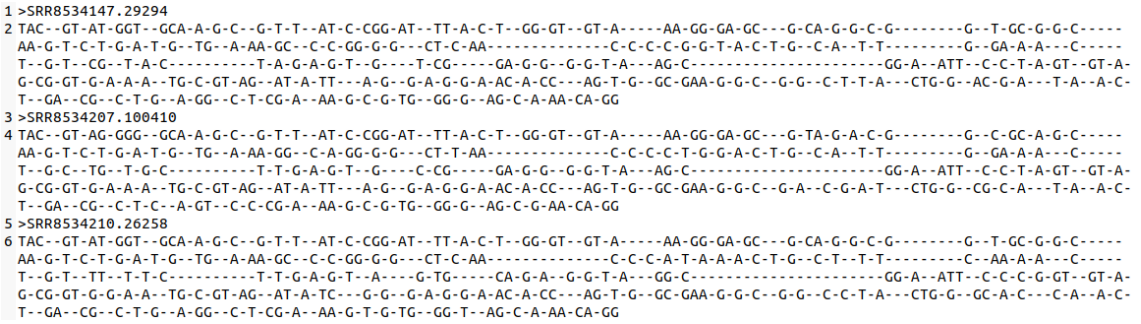


Figura 17 Captura de pantalla de l'alineament després del procés de filtratge. Es pot observar com s'han eliminat la majoria de gaps "-".

- ***pre.cluster()*** → pre.cluster() es basa en un algoritme que ordena les seqüències segons l'abundància en què apareixen. Després les compara amb la resta de seqüències i les agrupa amb aquelles que es troben dins d'un límit en quant a similitud. En el nostre cas hem indicat a la mateixa comanda mitjançant "diffs=2" que agrupi les seqüències amb un màxim de 2 nucleòtids de diferència entre elles.

- ***classify.seqs()*** → Arribats en aquest punt ja hem fet tot el filtratge i pre-processament necessari per poder dur a terme la classificació

taxonòmica. Per fer-ho, *classify.seqs*() utilitza un algoritme *k-nearest neighbor* basat en distàncies entre les seqüències a classificar i les seqüències consens de referència. A part de l'arxiu d'entrada en format fasta es requereix una base de dades de seqüències i un arxiu taxonòmic amb les seqüències de referència.

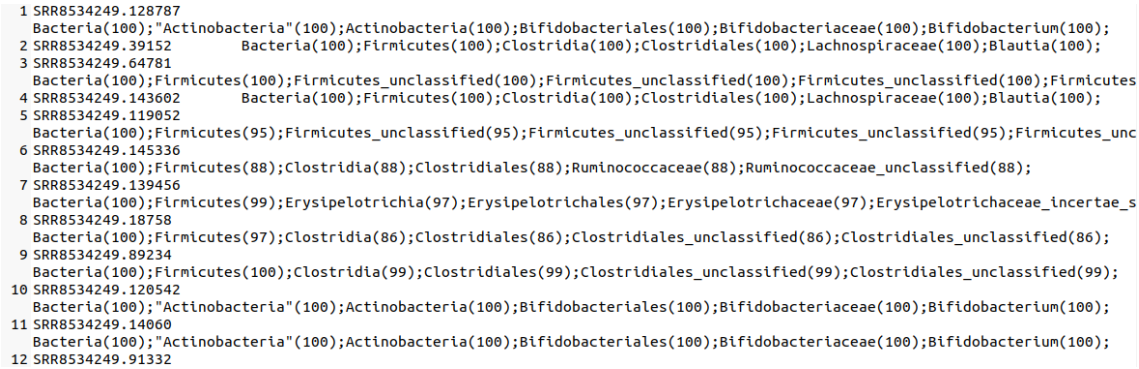


Figura 18 Captura de pantalla del fitxer .taxonomy indicant la classificació taxonòmica de cada una de les seqüències.

- ***Remove.lineage()*** → Per últim només queda eliminar totes aquelles seqüències classificades com a no bacterianes. Podrien ser cloroplast, mitocòndries, archeas o eucariotes. Podria ser interessant incloure aquestes seqüències a l'estudi però tenint en compte que el gen 16S rRNA és exclusivament bacterià estaríem cometen un error.

Al final d'aquest *pipeline* hem obtingut tres fitxers finals, dos dels quals són els que s'han utilitzat per poder desenvolupar l'anàlisi *downstream*. El fitxer fasta final, la taula de comptatges i l'arxiu amb la classificació taxonòmica de cada un dels *reads* presents en el data set. Els dos últims junt amb el fitxer de metadades són els que hem necessitat per crear l'objecte *phyloseq* en R per tal de fer tot l'anàlisi. A continuació en destaquem els passos seguits.

Tot aquest *pipeline* de classificació taxonòmica s'ha dut a terme per a cada una de les cohorts.

Anàlisi *downstream*

L'anàlisi *downstream* s'ha basat en el tutorial explicat aquí https://www.nicholas-ollberding.com/post/introduction-to-the-statistical-analysis-of-microbiome-data-in-r/ (*Introduction to the Statistical Analysis of Microbiome Data in R | Academic*, n.d.) però també en el tutorial del paquet microbiome de R https://microbiome.github.io/tutorials/ (*Introduction to the Microbiome R Package*, n.d.) entre d'altres.

De la mateixa manera que amb *Mothur* el codi complet de l'anàlisi es pot trobar al repositori de GitHub d'aquest treball https://github.com/miquelcastany/metagenomics_TFM .

A part de la importació a R dels fitxers finals resultants de *Mothur*, la creació de l'objecte *phyloseq* y el pre-processament de les dades i metadades, l'anàlisi ha constat dels següents passos descrits a continuació.

Visualització de les abundàncies relatives i anàlisi estadístic

Hem començat visualitzant les abundàncies relatives a nivell de phyllum mitjançant la funció tax_glom() del paquet *phyloseq* de R. Per fer-ho hem optat per dues estratègies diferents. Fer un diagrama de barres apilat amb les abundàncies relatives dels phylums presents en cada mostra i segons si pertanyen al grup control HC o amb colitis ulcerosa UC.

També hem fet un diagrama de caixes o boxplot independent per a cada un dels phylums. A més hem dut a terme un test de Wilcoxon mitjançant la funció wilcox.test() per avaluar-ne les possibles diferències entre HC i UC. Els diferents p-valors obtinguts s'han mostrat al mateix gràfic.

Aquesta part és la més informativa a nivell de poder extreure un possible patró de disbiòsi que sigui comú al grup UC respecte del HC.

Clustering jeràrquic

A partir de les abundàncies relatives s'ha emprat el mètode de *Bray-Curtis* i el paquet *vegan* de R per calcular les distàncies entre les diferents mostres i poder-les agrupar entre elles segons aquests valors. També hem aplicat el mètode Ward de clusterització i hem indicat el grup al qual pertany cada mostra amb un codi de colors.

Diversitat Alpha

La diversitat alpha és aquella diversitat inherent a cada una de les mostres però que a més contempla la distribució de les diferents unitats taxonòmiques. Es quantifica a partir de la riquesa, és a dir el nombre d'espècies presents a la mostra, però també té en compte la distribució d'aquestes. És per aquest fet que es diferencia del concepte de riquesa tot i que sovint aquests dos conceptes es poden confondre.

A través de la funció alpha() del paquet *microbiome* de R hem calculat els diferents index de diversitat alpha. Posteriorment hem fet un boxplot dels índex de diversitat *Shannon* del dos grups HC i UC. També s'han calculat els p-valors dels test de *Kolmogorov-Smirnov* i *Wilcoxon* entre els dos grups per a avaluar-ne les diferències.

Diversitat Beta

La diversitat beta en comptes de ser inherent a una sola mostra o comunitat microbiana és una mesura de similitud (o dissimilitud) entre mostres o poblacions d'un mateix grup. L'hem calculat per a cada un dels dos grups com la dissimilitud de cada una de les mostres respecte a la mitjana del grup al

qual pertanyen. Per al seu càlcul s'ha emprat la funció divergence() del paquet microbiome i s'ha graficat en boxplots.

Posteriorment també s'ha fet un test de Wilcoxon per veure si les diferències observades al gràfic són prou significatives o no a nivell estadístic.

La diferència entre diversitat alpha i beta s'exemplifica a la següent figura:

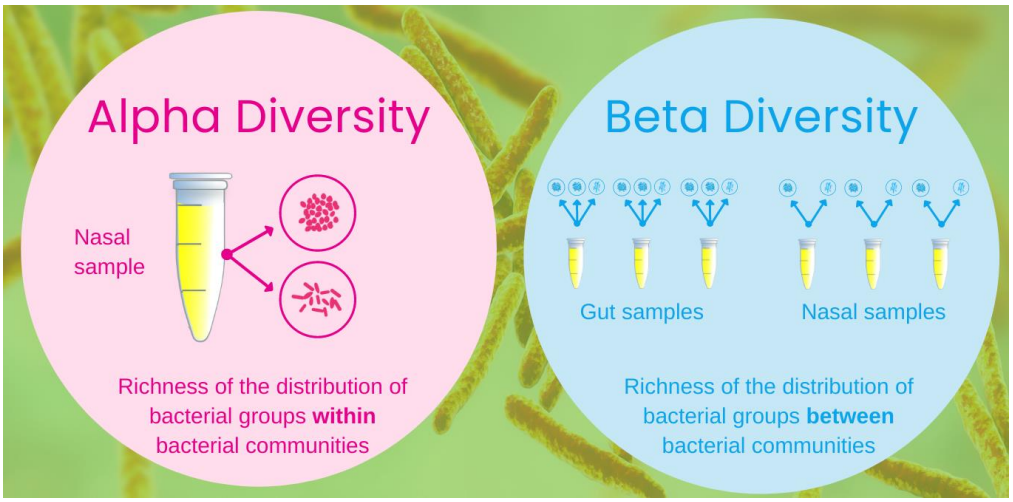


Figura 19 Diferència entre alpha i beta diversity (*The Microbial World of Our Pets* | *GoldBio*, n.d.)

Anàlisi de components principals

L'anàlisi de components principals (PCA) és una tècnica estadística utilitzada per reduir la dimensió de les dades i identificar patrons o estructures subjacents en un conjunt de variables. L'objectiu de la PCA és transformar un conjunt de variables originals en un nou conjunt de variables no correlacionades anomenades components principals. Aquests components principals són combinacions lineals de les variables originals i estan ordenats per la seva variança, de manera que el primer component principal representa la major part de la variança de les dades, el segon component principal captura la segona major part, i així successivament.

S'ha utilitzat la funció transform() del paquet *microbiome* per tal de transformar els comptatges a CLR (*centered log-ratio*) y la funció ordinate() del paquet *phyloseq* per poder fer un *scree plot* i visualitzar la proporció de variació que és capaç d'explicar cada una de les components principals. Posteriorment hem graficat les dues primeres components principals de en un gràfic per a cada una de les cohorts.

Anàlisi d'abundàncies relatives diferencial

Mitjançant varis test estadístics de Wilcoxon y fixant un nivell de significació de α=0.05 hem fet l'anàlisi d'abundàncies relatives diferencial a diferents nivells taxonòmics entre els grups HC i UC. Tot i que al primer apartat de visualització d'abundàncies relatives ja hem fet un test Wilcoxon per al rang

de phylum i n'hem mostrat els p-valors al gràfic, aquí els hem calculat per als nivells de classe i ordre.

Per dur-ho a terme és necessari primer aglomerar les dades al nivell taxonòmic pertinent a través de la funció tax_glom() del paquet *phyloseq*. Després ens hem basat en el mètode emprat a https://www.nicholas-ollberding.com/post/introduction-to-the-statistical-analysis-of-microbiome-data-in-r/ però basant-nos en les abundàncies relatives en comptes dels valors transformats a CLR. En aquest tutorial s'utilitzen dataframes niats.

Predicció

L'objectiu de la part de predicció és implementar diferents models predictius i poder-ne avaluar les diferències i el rendiment de cadascun. En aquest treball s'han emprat els mètodes *selbal* (Rivera-Pinto et al., 2018b) i *coda4microbiome* (Calle et al., 2023) basats en regressions múltiples i anàlisis composicionals respectivament. També s'ha utilitzat un model de *Machine Learning* basat en *Random Forest Classification*. Tots tres models s'han implementat a nivell taxonòmic de família.

A continuació exposem les característiques principals de cada un d'ells.

Selbal

El paquet de R *selbal* implementa un mètode de selecció avançada per a la identificació de dos grups de diferents taxons l'abundància relativa (o equilibri) dels quals està associat amb la variable resposta d'interès.

Mitjançant la funció selbal.cv() el *selbal* pretén respondre les següents preguntes:

- Quin és el nombre òptim de variables per incloure en l'equilibri?
- Quins són els grups de taxons l'abundància relativa dels quals està més associada a la variable de resposta?
- L'equilibri proposat és robust?

Per tal de poder córrer la funció selbal.cv() hem hagut de filtrar les nostres dades per tal d'eliminar-ne les columnes que no contenien suficient informació. Per dur-ho a terme hem aplicat un punt de tall específic per a cada cohort i la funció prune_taxa() del paquet *phyloseq*. Per a cada cohort s'ha hagut d'aplicar un punt de tall determinat.

Coda4microbiome

coda4microbiome és un altre paquet de R i és la versió millorada de *selbal*. De fet, els dos paquets ha estat desenvolupat pels mateixos investigadors.

coda4microbiome té la particularitat que analitza les dades de microbiomes com a dades composicionals CoDA (*Compositional Data Analysis*). Les dades composicionals es refereixen a dades que representen les proporcions relatives o percentatges de les diferents parts o components d'un conjunt. En les dades composicionals, els valors dels components individuals no es consideren de manera independent, sinó com a parts d'un tot que sumen una constant o el 100%. En l'anàlisi de dades composicionals, és important tenir en compte la naturalesa composicional de les dades i abordar els reptes específics associats, com la restricció de la suma constant i la presència d'estructures de correlació. S'han desenvolupat diverses tècniques estadístiques per analitzar eficaçment les dades composicionals, incloent-hi transformacions de log-ràtio, coordenades log-ràtio isomètriques (ILR) i models de regressió dissenyats per a dades composicionals.

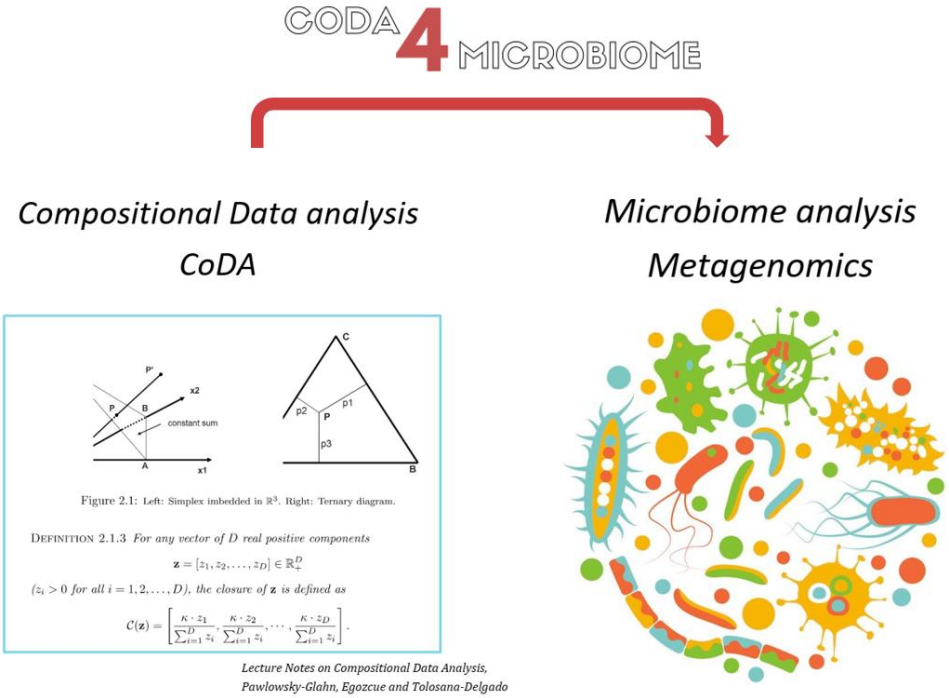


Figura 20 (Calle et al., 2023)

Al igual que amb *selbal*, *coda4microbiome* té com a objectiu la predicció de signatures microbianes que puguin ser emprades com a potencial biomarcadors.

Mitjançant la funció coda_glmnet() l'algoritme realitza la selecció de variables mitjançant regressió penalitzada en el conjunt de totes les log-ratios aparellades, tant per a resultats binaris (regressió logística) com per a resultats continus (regressió lineal). El resultat s'expressa com un equilibri entre dos grups de taxons.

Machine Learning

A part dels dos mètodes explicats més amunt i que es basen en models estadístics, també hem utilitzat metodologies basades en aprenentatge automàtic o *Machine Learning* (ML).

L'aprenentatge automàtic és una branca de la intel·ligència artificial (IA) que es centra en el desenvolupament d'algoritmes i models que permeten als ordinadors aprendre i fer prediccions o preses de decisions sense ser programats explícitament per a tasques específiques. Implica el disseny i entrenament de sistemes computacionals per aprendre i millorar automàticament a partir de les dades, en lloc de dependre d'instruccions explícites.

En ML, es desenvolupen algoritmes per analitzar i extreure patrons o idees a partir de conjunts de dades, que són utilitzats per entrenar models. Llavors aquests models són capaços de fer prediccions o prendre decisions sobre noves dades noves basant-se en els patrons i coneixements apresos durant la fase d'entrenament.

En el cas que ens ocupa hem utilitzat un mètode de classificació basat en la tècnica de *Random Forest*. La classificació mitjançant *Random Forest* és una tècnica de ML que combina diversos arbres de decisió per fer prediccions. Cada arbre es construeix utilitzant subconjunts aleatoris de dades i característiques. En el procés de predicció cada arbre retorna la seva predicció de manera independent, i el resultat final es determina per votació majoritària. Els *Random Forest* són coneguts per la seva precisió, robustesa i capacitat de gestionar dades sorolloses. També proporcionen idees sobre la importància de les característiques.(*Understanding Random Forest. How the Algorithm Works and Why It Is… | by Tony Yiu | Towards Data Science*, n.d.)

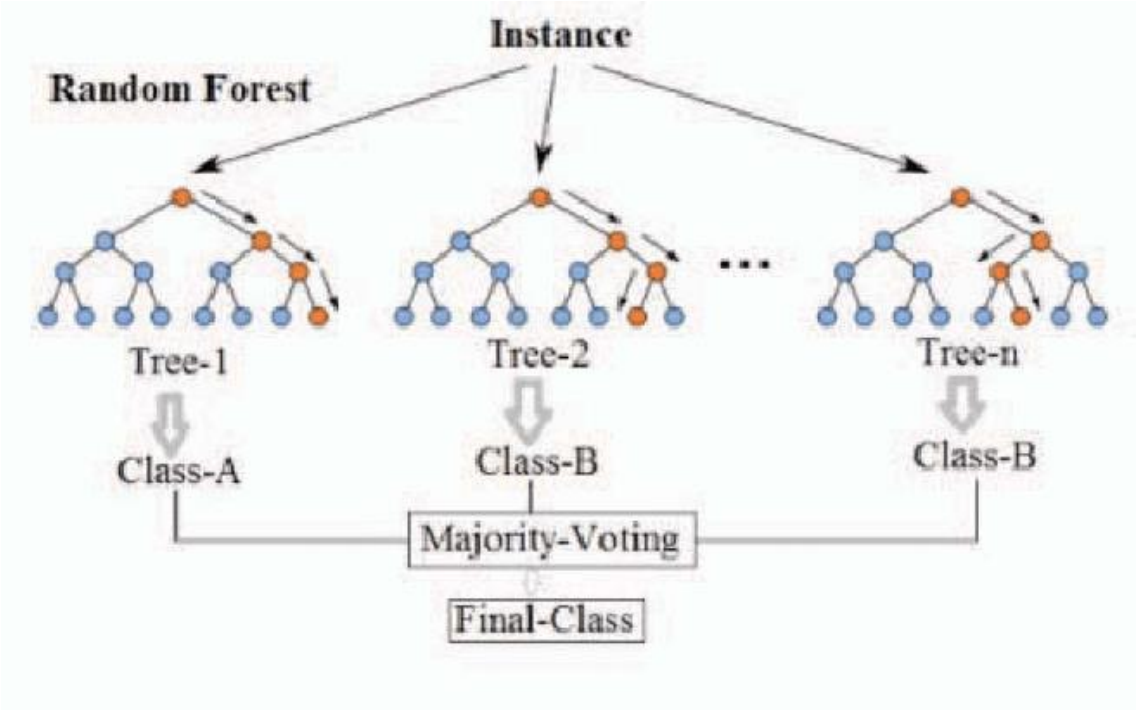


Figura 21 Esquema d'un model de classificació basat en *Random Forest* (Chugh et al., 2020).

Per tal d'implementar la metodologia de *Random Forest* hem utilitzat la funció randomForest() del paquet *caret* de R. Hem dividit cada una de les cohorts

en subconjunts d'entrenament i de test en una proporció del 70% per al grup d'entrenament i posteriorment se n'ha avaluat el rendiment testejant el 30% restant de les dades. La funció randomForest() permet triar el nombre d'arbres de decisió i l'hem fixat en 50.

4. Resultats

Tot seguit presentem els resultats més significatius tant de l'anàlisi *upstream* com del *downstream*.

Resultats *Upstream Analysis*

Seguint les instruccions del *MiSeq SOP* de *Mothur* hem anat eliminant soroll de fons de les dades baixades amb *SRA-toolkit*.

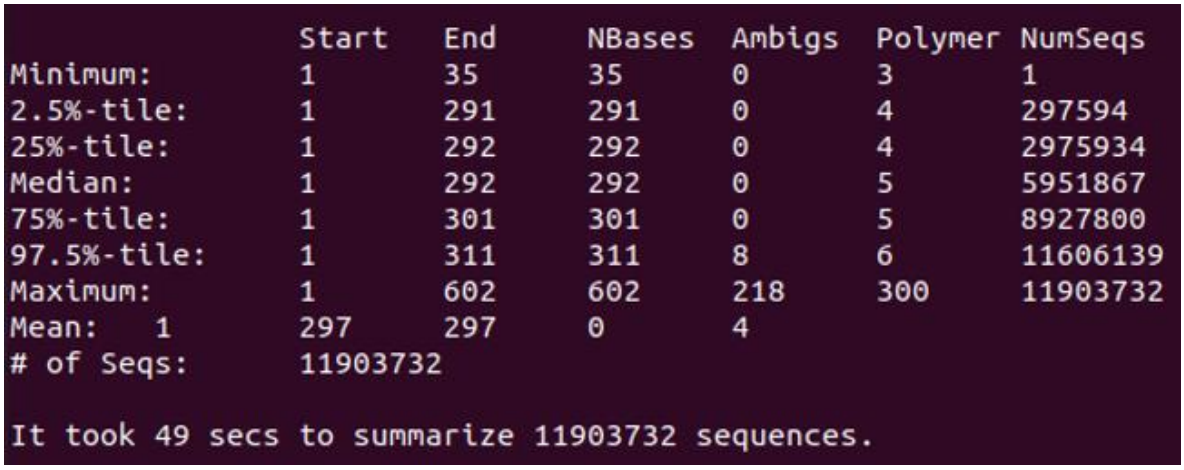
La millor manera de presentar els resultats obtinguts a cada un dels passos de l'anàlisi *upstream* és resumir-los en una taula comparativa entre les dues cohorts:

	Discovery	**Validation**
sample number	80	20
Amplicon region	v4	v3-v4
initial contigs	11903732	3147325
Maxlength screen	321	520
contigs removed	1871705	2277946
Unique contigs	2846826	663093
count.seqs()	10032027	869379
pcr.seqs()	oligos	start=6388, end= 25318
screen.seqs(align)	start= 13862, end=23444	start=1, end=18928
Alignments w/ too many bases	13303	88
Sequences reversed	11974	46
sequences removed after screening alignment	2681	23900
alignment length	570	1017
number of filtered sequences	2844145	639193
Unique sequences classified before pre-cluster	368362	634605
Unique sequences classified after pre-cluster	232391	565051
unclassified	9	0
after remove lineage	230657	539402
percentage of final seqs from initials	1,94%	17,14%

Taula 3 Resum dels paràmetres emprats en el procés de filtratge i resultats obtinguts en l'anàlisi upstream.

No obstant, també podem comentar pas a pas alguns dels *findings* més interessants.

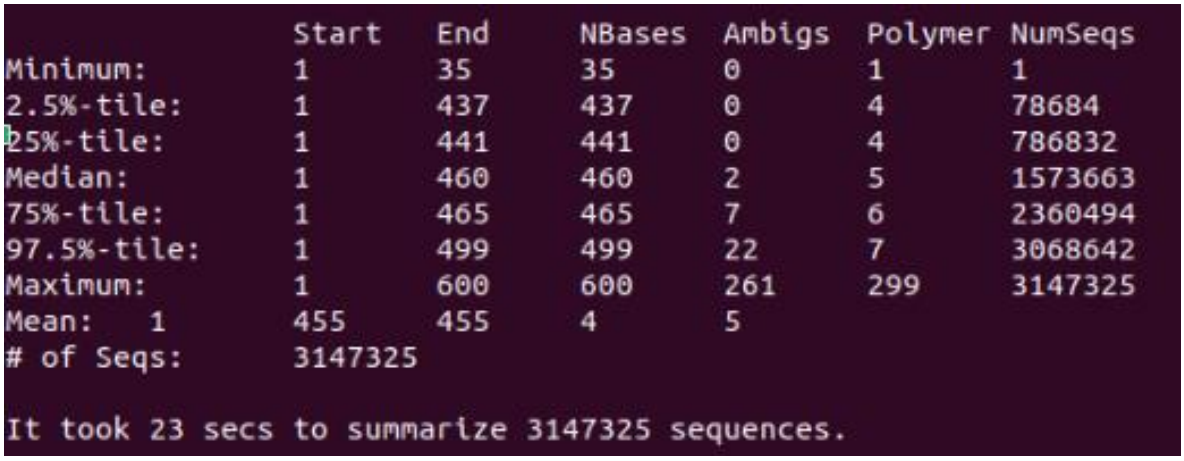
Comparant els resultats del summary.seqs() després d'haver creat els contigs:



	Start	End	NBases	Ambigs	Polymer	NumSeqs
Minimum:	1	35	35	0	3	1
2.5%-tile:	1	291	291	0	4	297594
25%-tile:	1	292	292	0	4	2975934
Median:	1	292	292	0	5	5951867
75%-tile:	1	301	301	0	5	8927800
97.5%-tile:	1	311	311	8	6	11606139
Maximum:	1	602	602	218	300	11903732
Mean: 1	297	297	0	4		
# of Seqs:	11903732					

It took 49 secs to summarize 11903732 sequences.

Figura 22 Summary.seqs() dels contigs inicials de la cohort *discovery*.



	Start	End	NBases	Ambigs	Polymer	NumSeqs
Minimum:	1	35	35	0	1	1
2.5%-tile:	1	437	437	0	4	78684
25%-tile:	1	441	441	0	4	786832
Median:	1	460	460	2	5	1573663
75%-tile:	1	465	465	7	6	2360494
97.5%-tile:	1	499	499	22	7	3068642
Maximum:	1	600	600	261	299	3147325
Mean: 1	455	455	4	5		
# of Seqs:	3147325					

It took 23 secs to summarize 3147325 sequences.

Figura 23 Summary.seqs() dels contigs inicials de la cohort *validation*.

Podem observar la diferència que hi ha en nombre de contigs generats entre les dues cohorts. Tenint en compte que a la CD tenim 4 vegades més de mostres és plausible trobem aproximadament 4 vegades més de contigs.

Veiem que la mitjana de llargada dels contigs és de 297 i 455 per la CD i la CV respectivament. Això concorda amb el fet que en la CD s'havia seqüenciat només la regió V4 del gen 16S mentre que a la CV s'havia fet també junt amb la regió V3 del mateix gen. Ens ha interessat treure totes aquelles seqüències amb bases ambigües, homopolímers més llargs de 8 nucleòtids i fixant un punt de tall de llargada màxima lleugerament superior al 97,5%-tile de cada una de les cohorts: 321 per la CD i 520 per la CV.

Un altre fet diferencial entre les dues cohorts és el mètode que hem emprat per reduir a la regió d'interès l'alineament de la base de dades SILVA. Com que en l'article de la CD es mencionen els primers utilitzats per a l'amplificació els hem indicat a la funció pcr.seqs() de *Mothur*. En canvi, a la CV hem indicat les posicions d'inici i final de les regions V3 i V4 respectivament dins de l'alineament. Per indicar els oligos, o primers els hem hagut de copiar en un fitxer de text.

Curiosament al fer el unique.seqs() després del filtratge de l'alineament en la CD passem de 2844145 a 232391 seqüències. Això representa menys del 10% de les seqüències inicials abans del unique.seqs(). Aquest fet no passa en el cas de la CV on passem de 639193 a 565051 seqüències.

Resultats *Downstream Analysis*

Passem a comentar els resultats obtinguts a partir de cada part de l'anàlisi estadístic de les dades prèviament processades amb *Mothur*.

Relative abundance visualization and testing

Presentem els resultats dels diagrames de barres apilades de les abundàncies relatives a nivell de phylum

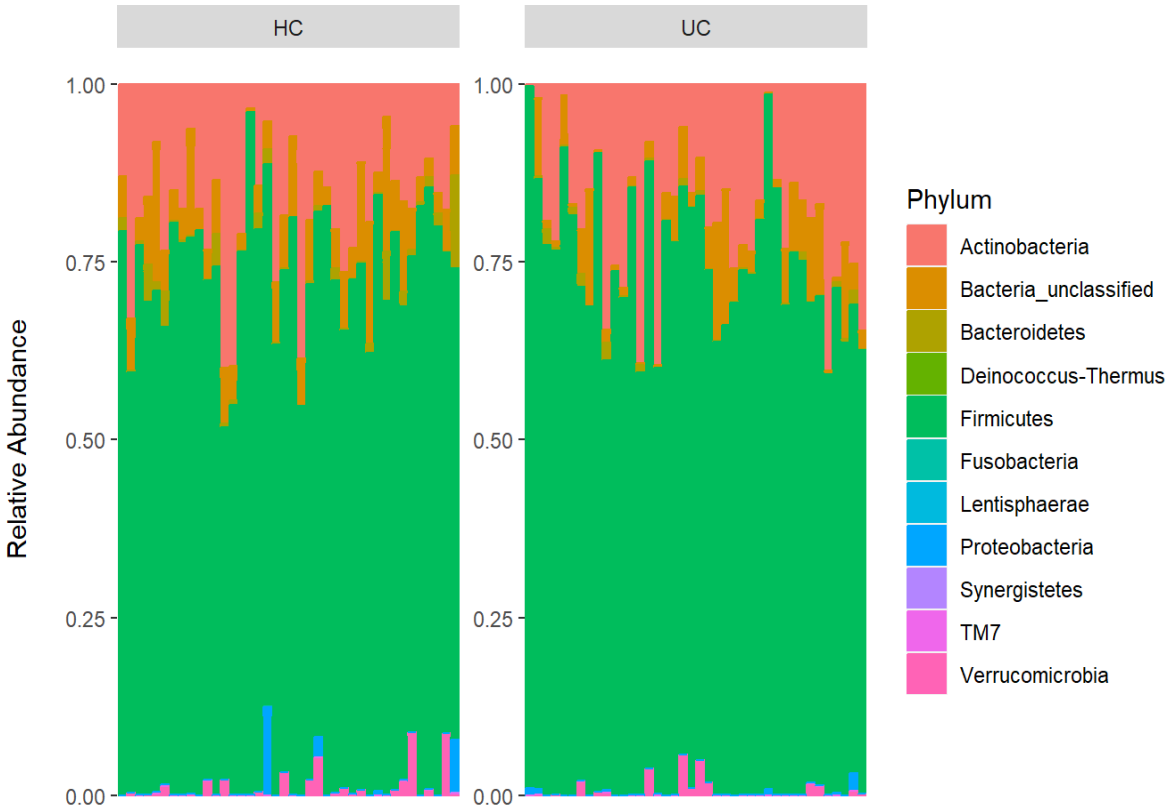


Figura 24 Comparació entre HC i UC d'abundàncies relatives a nivell de *phylum* de la CD.

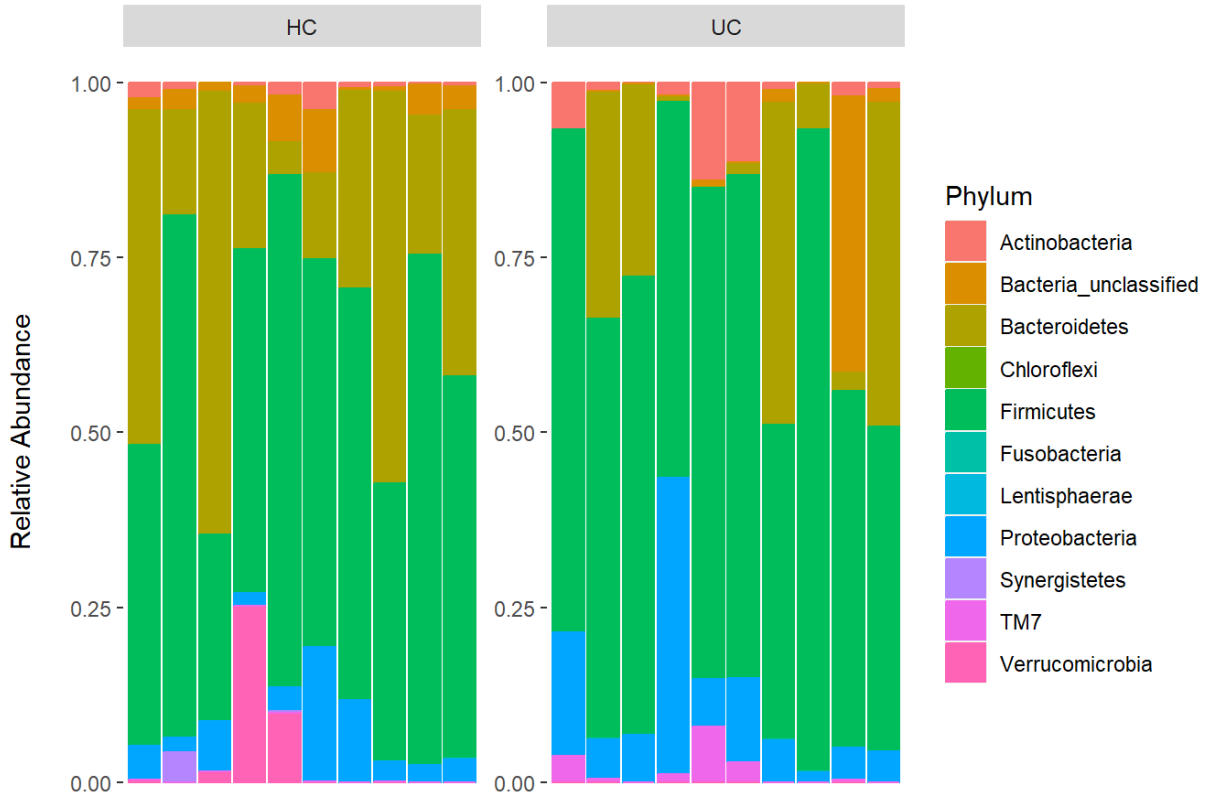


Figura 25 Comparació entre HC i UC d'abundàncies relatives a nivell de *phylum* de la CV.

Diagrames de caixes de cada un dels phyllums comparant les dues condicions. S'indiquen també els p-valors dels test de *Wilcoxon* entre HC i UC per a cada *phyllum*.

Figura 26 Boxplots de cada un dels *phylums* de la CD.

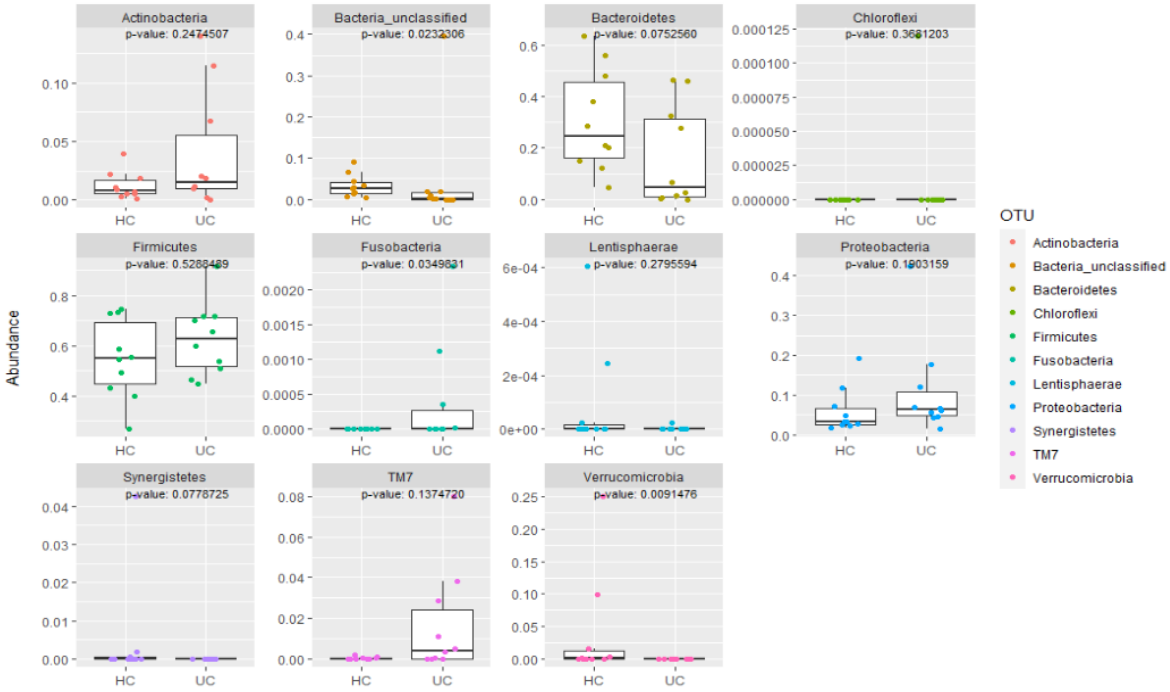


Figura 27 Boxplots de cada un dels *phylums* de la CV.

Hierarchical Clustering

Clústers jeràrquics de CD i CV.

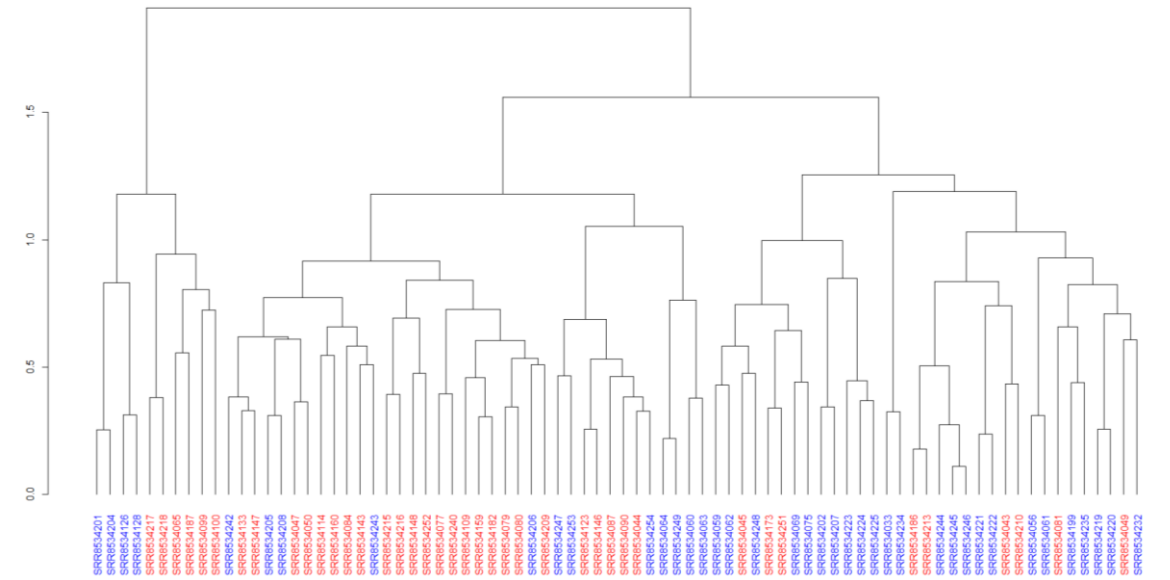


Figura 28 Clúster jeràrquic de les mostres de la CD.

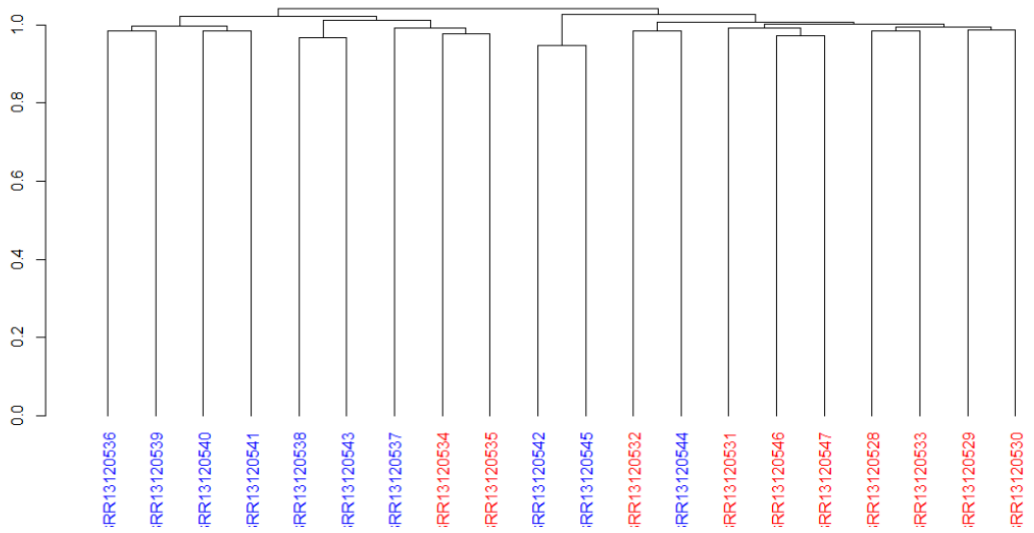


Figura 29 Clúster jeràrquic de les mostres de la CV.

Alpha Diversity

Mostrem els boxplots de les diversitats alpha de cada un dels grup i per a cada cohort i els p-valors obtinguts dels test de Wilcoxon entre els índex de cada condició.

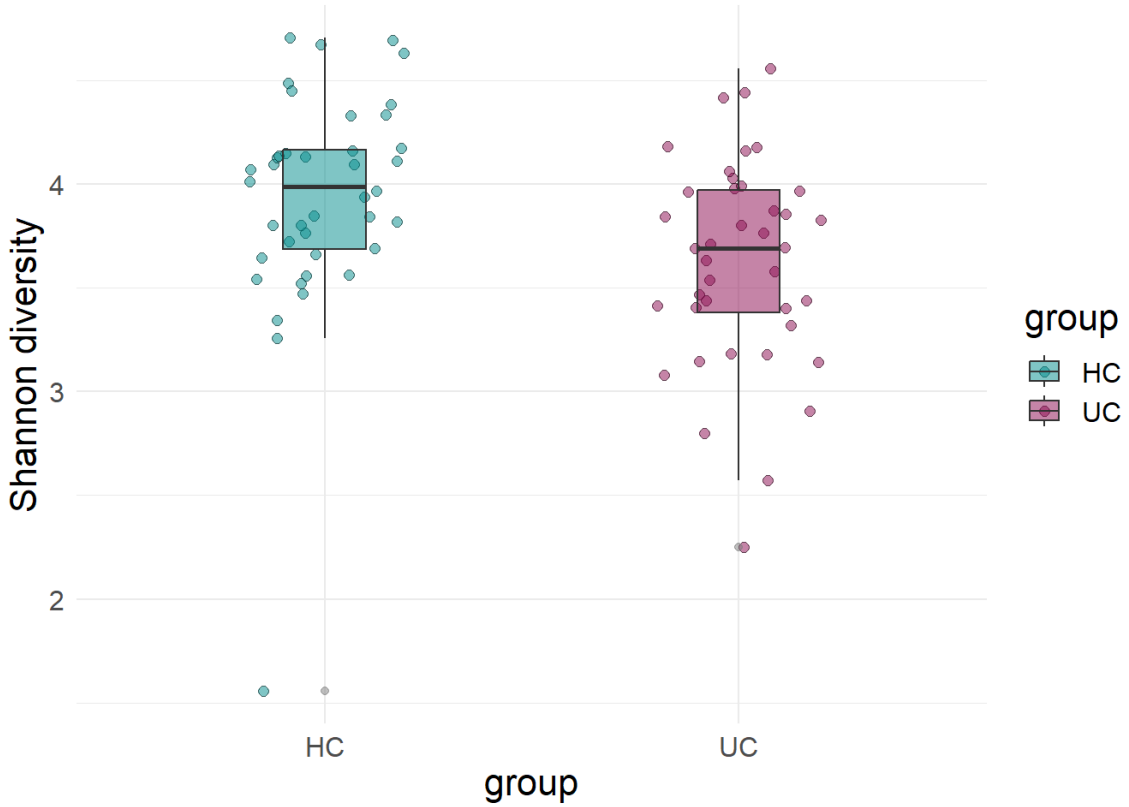


Figura 30 Alpha diversities de HC i UC de la CD.

En el cas de la CD el p-valor per a les diversitats alpha de la CD és de 0.002948961. Per tant, per un valor de significació de α = 0.05 la diferència es significativa.

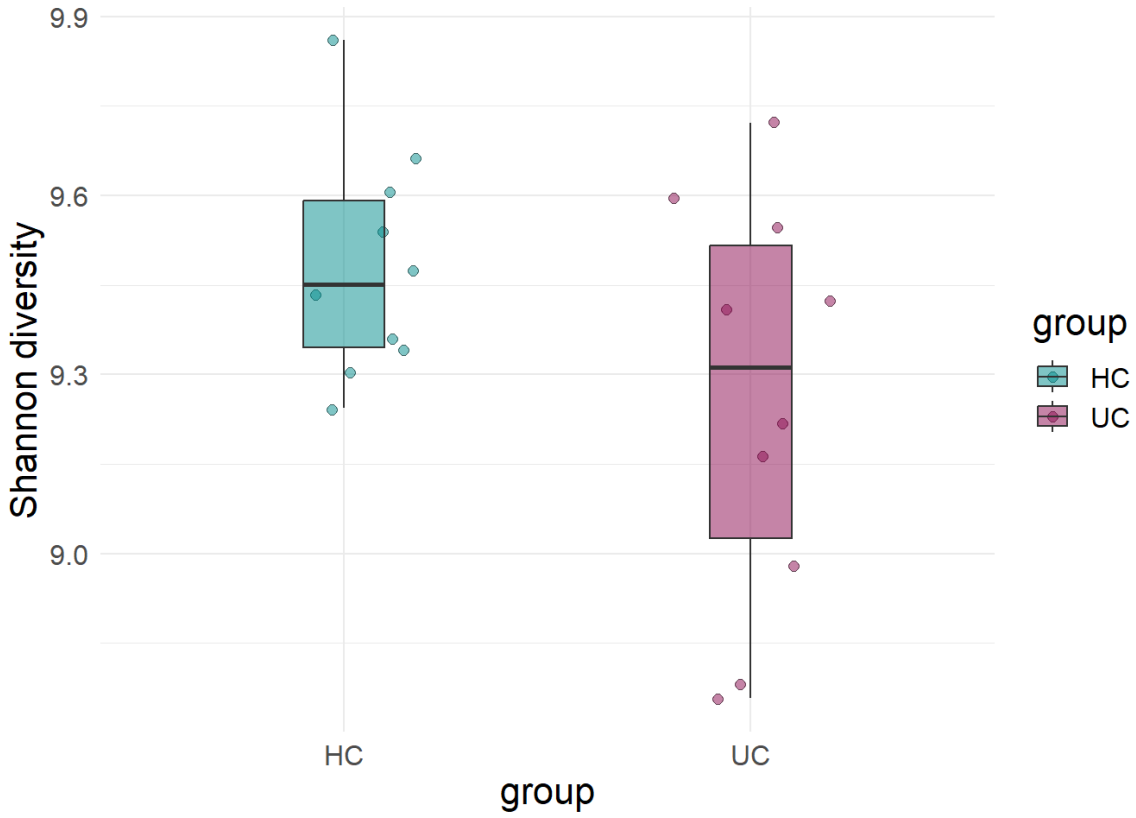


Figura 31 Alpha diversities de HC i UC de la CV.

En canvi per a la CV el p-valor obtingut amb el test de Wilcoxon és de 0.1654939. Per tant, per a α = 0.05 no podem rebutjar la hipòtesi nul·la de no diferències entre els dos grups HC i UC.

Beta Diversity

Amb la diversitat beta procedirem de la mateixa manera que amb la alpha. Presentem tot seguit els gràfics i indiquem el p-valor dels test Wilcoxon.

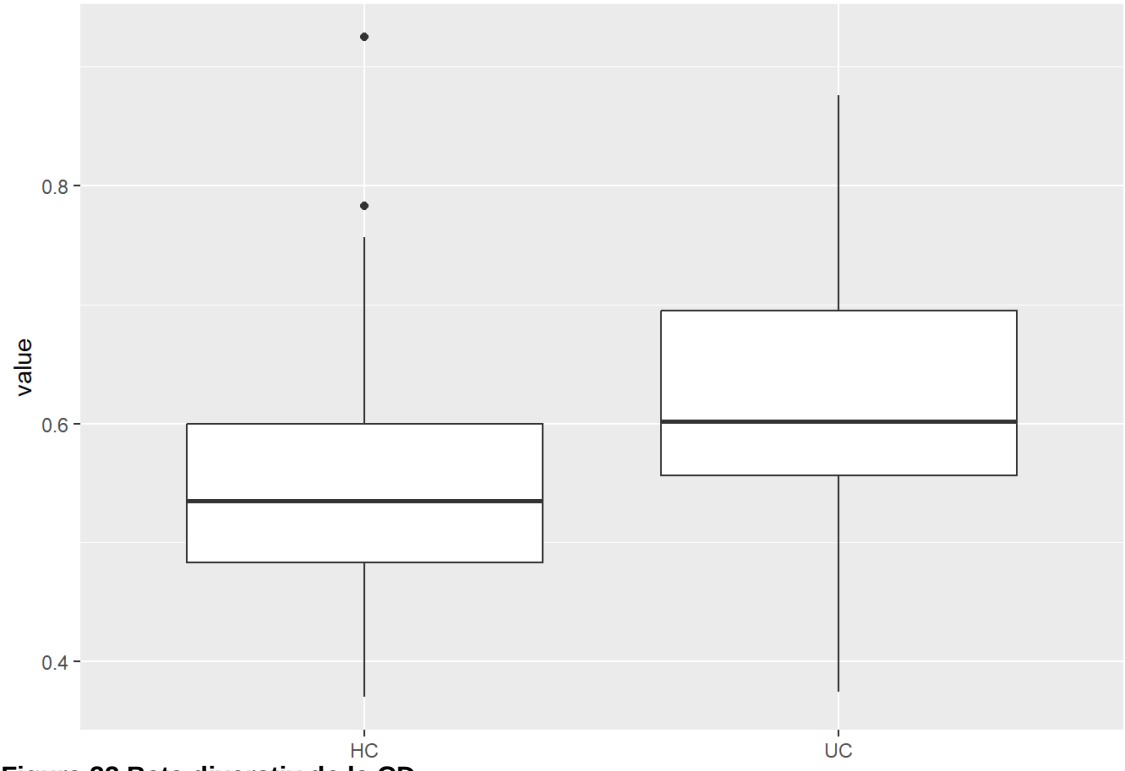


Figura 32 Beta diverstiy de la CD.

Mitjançant el test de Wilcoxon obtenim un p-valor de 0.001011187, suficientment baix com per rebutjar la hipòtesi nul·la de no diferències entre HC i UC.

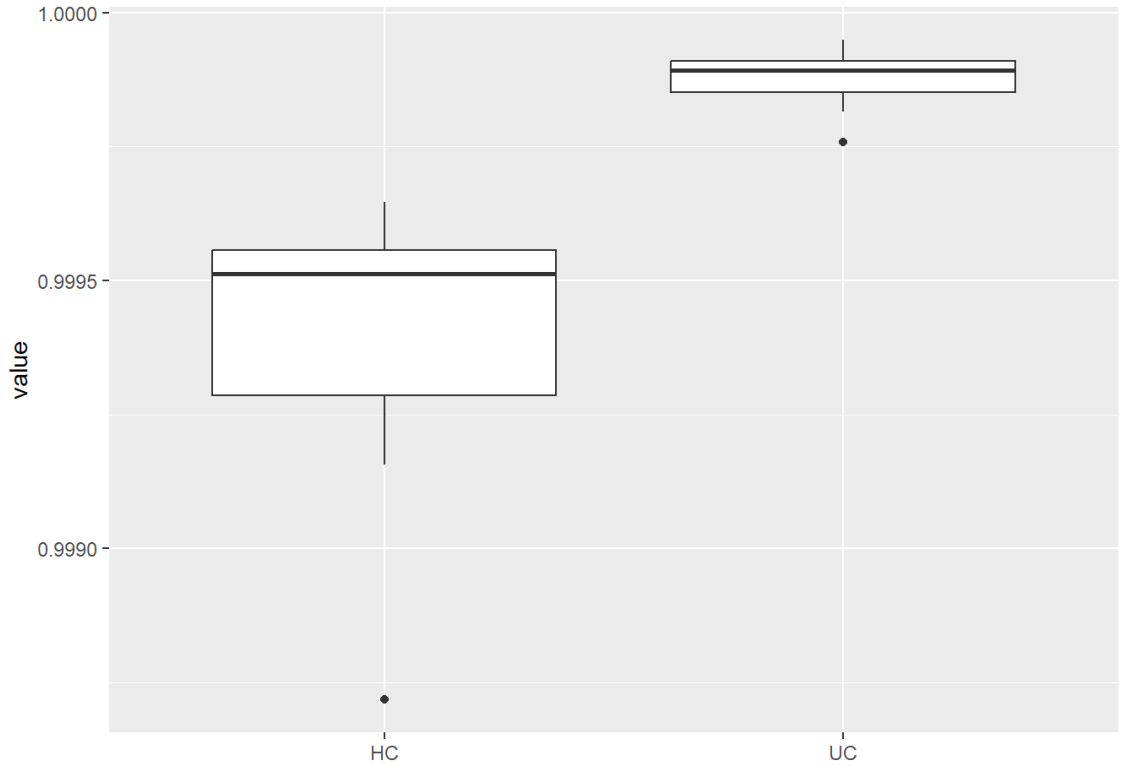


Figura 33 Beta diversity per la CV.

En el cas de la CV obtenim un p-valor de 1.082509e-05 encara més significatiu.

Principal Component Analysis

Gràfic de les dues primeres components principals de la CD:

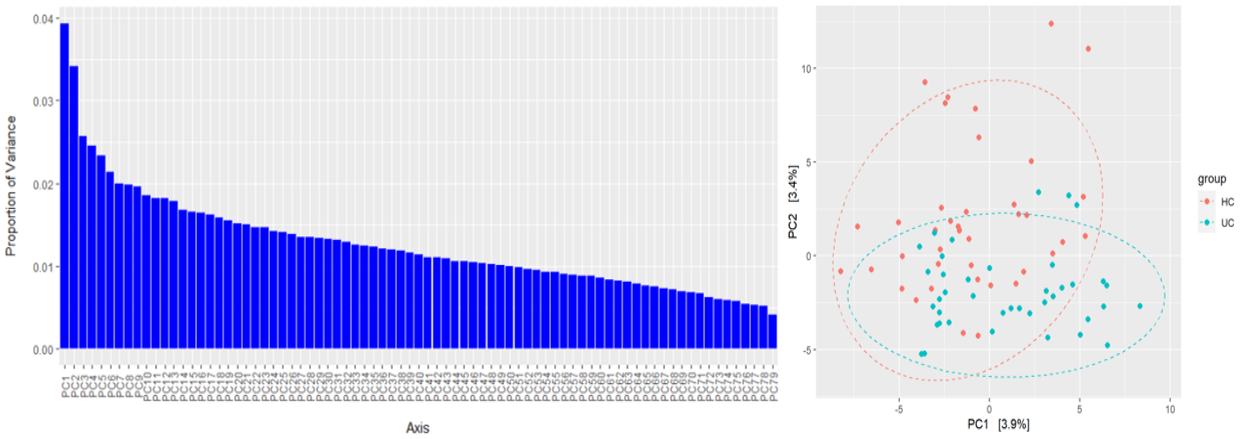


Figura 34 PCA de la CD.

En el PCA observem una certa separació entre les mostres de cada cohort, però també es nota un notable solapament entre elles. Ara centrem-nos en el PCA de la CV per analitzar-ne els resultats amb més detall.

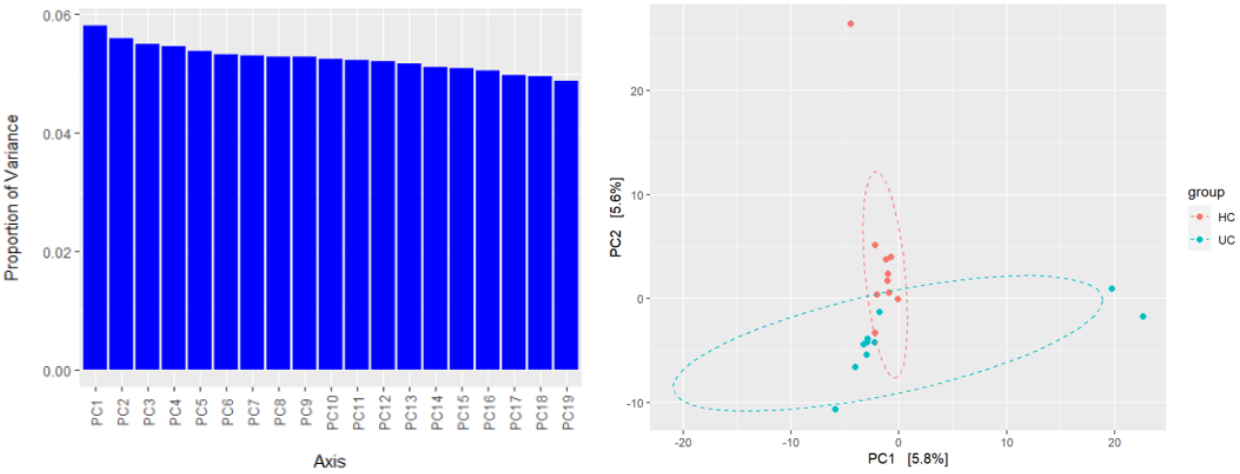


Figura 35 PCA de la CV.

En el cas de la CV, s'observa que hi ha alguns *outliers* tant en el grup HC com en el grup UC, però la separació entre els dos és molt més evident en comparació amb la CD. Hi ha una clara distinció entre les mostres del grup HC i del grup UC, amb pràcticament cap solapament entre elles. Aquesta diferenciació més marcada suggereix una major variabilitat entre les mostres de la CV en comparació amb les de la CD.

Differential abundance analysis

Mostrem a continuació els diferents rangs taxonòmics identificats com a significativament diferents entre les dues condicions i per a cada cohort.

A tibble: 2 × 10 Groups: OTU [2]

OTU <chr>	p_value <dbl>	BH_FDR <dbl>	Kingdom <chr>	Phylum <chr>
Bacteria_unclassified	0.009256287	0.009256287	Bacteria	Bacteria_unclassified
Bacteroidetes	0.014946960	0.014946960	Bacteria	Bacteroidetes

2 rows

Figura 36 Phylums amb p-valors inferiors a 0.05 en la CD.

A tibble: 4 × 10 Groups: OTU [4]

OTU <chr>	p_value <dbl>	BH_FDR <dbl>	Kingdom <chr>	Phylum <chr>	Class <chr>
SRR8534187.152050	0.007535674	0.007535674	Bacteria	Bacteroidetes	Bacteroidia
SRR8534242.99833	0.009256287	0.009256287	Bacteria	Bacteria_unclassified	Bacteria_unclassified
SRR8534252.29939	0.019032073	0.019032073	Bacteria	Proteobacteria	Deltaproteobacteria
SRR8534033.21838	0.043375341	0.043375341	Bacteria	Firmicutes	Bacilli

4 rows

Figura 37 Classes amb p-valors inferiors a 0.05 en la CD.

A tibble: 9 × 10 Groups: OTU [9]

OTU <chr>	p_value <dbl>	BH_FDR <dbl>	Kingdom <chr>	Phylum <chr>	Class <chr>	Order <chr>
SRR8534062.136017	0.000216775	0.000216775	Bacteria	Firmicutes	Clostridia	Clostridia_unclassified
SRR8534187.152050	0.007535674	0.007535674	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales
SRR8534242.99833	0.009256287	0.009256287	Bacteria	Bacteria_unclassified	Bacteria_unclassified	Bacteria_unclassified
SRR8534060.43072	0.019282242	0.019282242	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales
SRR8534252.29939	0.021054581	0.021054581	Bacteria	Proteobacteria	Deltaproteobacteria	Desulfovibrionales
SRR8534223.71062	0.021814008	0.021814008	Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales
SRR8534218.37862	0.027340814	0.027340814	Bacteria	Actinobacteria	Actinobacteria	Coriobacteriales
SRR8534033.21838	0.041382905	0.041382905	Bacteria	Firmicutes	Bacilli	Lactobacillales
SRR8534224.147477	0.048709164	0.048709164	Bacteria	Actinobacteria	Actinobacteria	Bifidobacteriales

9 rows | 1-9 of 10 columns

Figura 38 Ordres amb p-valors inferiors a 0.05 en la CD.

A tibble: 3 × 10 Groups: OTU [3]

OTU <chr>	p_value <dbl>	BH_FDR <dbl>	Kingdom <chr>	Phylum <chr>
Verrucomicrobia	0.009147582	0.009147582	Bacteria	Verrucomicrobia
Bacteria_unclassified	0.023230639	0.023230639	Bacteria	Bacteria_unclassified
Fusobacteria	0.034983096	0.034983096	Bacteria	Fusobacteria

1-3 of 3 rows | 1-8 of 10 columns

Figura 39 Phylums amb p-valors inferiors a 0.05 en la CV.

A tibble: 10 × 10 Groups: OTU [10]

OTU <chr>	p_value <dbl>	BH_FDR <dbl>	Kingdom <chr>	Phylum <chr>	Class <chr>
SRR13120536.41090	7.577562e-05	7.577562e-05	Bacteria	Firmicutes	Bacilli
SRR13120546.65478	7.511794e-04	7.511794e-04	Bacteria	Proteobacteria	Proteobacteria_unclassified
SRR13120541.118592	3.886207e-03	3.886207e-03	Bacteria	Proteobacteria	Gammaproteobacteria
SRR13120535.102826	8.218413e-03	8.218413e-03	Bacteria	Bacteroidetes	Bacteroidetes_unclassified
SRR13120531.52840	9.147582e-03	9.147582e-03	Bacteria	Verrucomicrobia	Verrucomicrobiae
SRR13120533.99799	1.468964e-02	1.468964e-02	Bacteria	Firmicutes	Firmicutes_unclassified
SRR13120544.120733	2.323064e-02	2.323064e-02	Bacteria	Bacteria_unclassified	Bacteria_unclassified
SRR13120537.7451	3.498310e-02	3.498310e-02	Bacteria	Fusobacteria	Fusobacteria
SRR13120533.64145	3.546299e-02	3.546299e-02	Bacteria	Proteobacteria	Alphaproteobacteria
SRR13120545.65993	4.234621e-02	4.234621e-02	Bacteria	Proteobacteria	Deltaproteobacteria

10 rows

Figura 40 Classes amb p-valors inferiors a 0.05 en la CV.

A tibble: 13 × 10 Groups: OTU [13]

OTU <chr>	p_value <dbl>	BH_FDR <dbl>	Kingdom <chr>	Phylum <chr>	Class <chr>	Order <chr>
SRR13120539.171497	0.0001299011	0.0001299011	Bacteria	Firmicutes	Bacilli	Bacillales
SRR13120541.118592	0.0007252809	0.0007252809	Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonadales
SRR13120546.65478	0.0007511794	0.0007511794	Bacteria	Proteobacteria	Proteobacteria_unclassified	Proteobacteria_unclassified
SRR13120532.84609	0.0059719636	0.0059719636	Bacteria	Proteobacteria	Alphaproteobacteria	Alphaproteobacteria_unclassified
SRR13120535.102826	0.0082184128	0.0082184128	Bacteria	Bacteroidetes	Bacteroidetes_unclassified	Bacteroidetes_unclassified
SRR13120534.35305	0.0089306978	0.0089306978	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales
SRR13120531.52840	0.0091475820	0.0091475820	Bacteria	Verrucomicrobia	Verrucomicrobiae	Verrucomicrobiales
SRR13120533.99410	0.0102773606	0.0102773606	Bacteria	Firmicutes	Clostridia	Clostridia_unclassified
SRR13120533.99799	0.0146896447	0.0146896447	Bacteria	Firmicutes	Firmicutes_unclassified	Firmicutes_unclassified
SRR13120530.27029	0.0149313896	0.0149313896	Bacteria	Proteobacteria	Deltaproteobacteria	Deltaproteobacteria_unclassified
SRR13120544.120733	0.0232306393	0.0232306393	Bacteria	Bacteria_unclassified	Bacteria_unclassified	Bacteria_unclassified
SRR13120546.15993	0.0349830964	0.0349830964	Bacteria	Proteobacteria	Deltaproteobacteria	Bdellovibrionales
SRR13120537.7451	0.0349830964	0.0349830964	Bacteria	Fusobacteria	Fusobacteria	Fusobacteriales

1-13 of 13 rows

Figura 41 Ordres amb p-valors inferiors a 0.05 en la CV.

Les diferències trobades entre grups HC i UC de les dues cohorts es resumeixen a la següent taula. També es marquen en verd aquells taxons en que la diferència s'ha trobat tant en la CD com en la CV.

Taxonomic Level	Discovery Cohort			Validation Cohort		
	p-value	OTU	group with more abundance	p-value	OTU	group with more abundance
Phylum				0.009147582	Verrucomicrobia	HC
	0.009256287	Bacteria_unclassified	HC	0.023230639	Bacteria_unclassified	HC
	0.014946960	Bacteroidetes	HC			
				0.034983096	Fusobacteria	UC
Class	0.007535674	Bacteroidia	HC			
	0.043375341	Bacilli	UC	7,57756E-05	Bacilli	UC
				7,51E-04	Proteobacteria_unclassified	HC
				3,89E-03	Gammaproteobacteria	UC
				8,22E-03	Bacteroidetes_unclassified	HC
				9,15E-03	Verrucomicrobiae	HC
				1,47E-02	Firmicutes_unclassified	HC
	0.009256287	Bacteria_unclassified	HC	2,32E+04	Bacteria_unclassified	HC
				3,50E+04	Fusobacteria	UC
				3,55E+04	Alphaproteobacteria	HC
	0.019032073	Deltaproteobacteria	HC	4,23E+04	Deltaproteobacteria	HC
Order	0.000216775	Clostridia_unclassified	HC			

	0.007535674	Bacteroidales	HC			
	0.009256287	Bacteria_unclassified	HC	0.0232306393	Bacteria_unclassified	HC
	0.019282242	Rhizobiales	HC			
	0.021054581	Desulfovibrionales	HC			
	0.021814008	Rhodobacterales	UC			
	0.027340814	Coriobacteriales	HC			
	0.041382905	Lactobacillales	UC			
	0.048709164	Bifidobacteriales	UC			
				0.0001299011	Bacillales	UC
				0.0007252809	Pseudomonadales	UC
				0.0007511794	Proteobacteria_unclassified	UC
				0.0059719636	Alphaproteobacteria_unclassified	
				0.0082184128	Bacteroidetes_unclassified	
				0.0089306978	Enterobacteriales	
				0.0091475820	Verrucomicrobiales	HC
				0.0102773606	Clostridia_unclassified	HC
				0.0146896447	Firmicutes_unclassified	UC
				0.0149313896	Deltaproteobacteria_unclassified	
				0.0349830964	Bdellovibrionales	
				0.0349830964	Fusobacteriales	HC

Taula 4 Resum dels diferents taxons amb diferències significatives d'abundància relativa.

En general podem observar com hem trobat més diferències a la CV que a la CD en tots tres rangs taxonòmics explorats.

A part de *Bacteria_unclassified* que trobem diferencialment abundant a les dues cohorts i en tots els rangs taxonòmics només hi ha la classe *Deltaproteobacteria* més abundant al grup sa tant en una cohort com en l'altre. També hi ha la classe *Bacilli* que la trobem més abundant al grup de pacients amb UC.

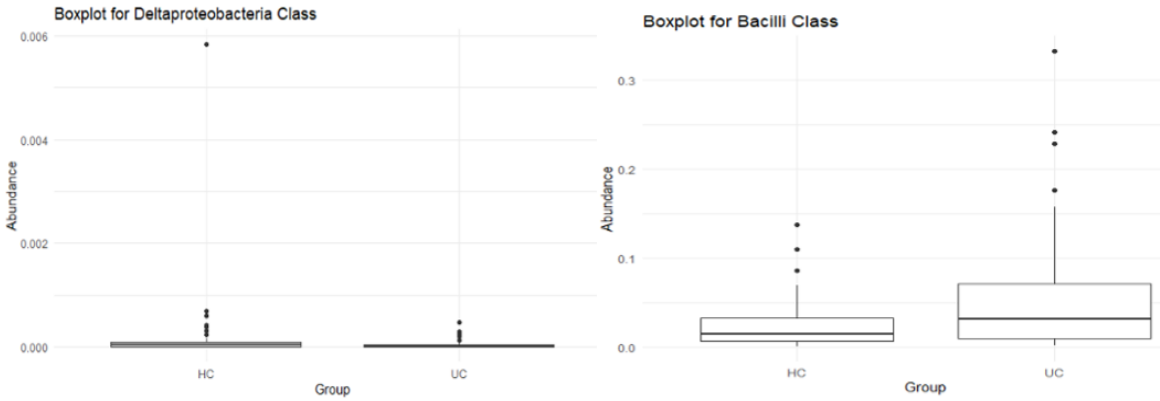


Figura 42 Diferències entres les classes *Deltaproteobacteria* i *Bacilli* a la CD.

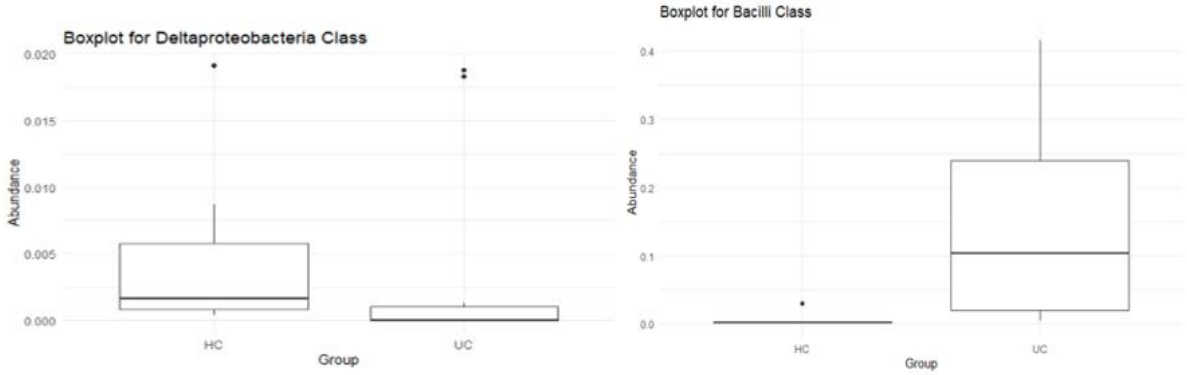


Figura 43 Diferències entres les classes *Deltaproteobacteria* i *Bacilli* a la CV.

Resultats de la Predicció

Selbal

Mitjançant el gràfic global.plot del procés de cross-validation de *selbal* podem visualitzar quants taxons, en aquest cas famílies, ha trobat l'algoritme com a nombre òptim per tal de maximitzar-ne la capacitat predictiva. *selbal* fa la predicció de la condició comprovant la relació que s'estableix entre aquests taxons. En aquest sentit, els classifica en el denominador o en el numerador del balanç.

Per la CD *selbal* troba 8 taxons amb la capacitat predictiva òptima. 4 d'ells en el denominador (*Rickenellaceae*, *Corynebacteriaceae*, *Acidaminococcaceae* i *Bacteria_unclassified*) i els 4 restants al numerador (*Carnobacteriaceae*, *Eubacteriaceae*, *Sutterellaceae* i *Betaproteobacteria_unclassified*).

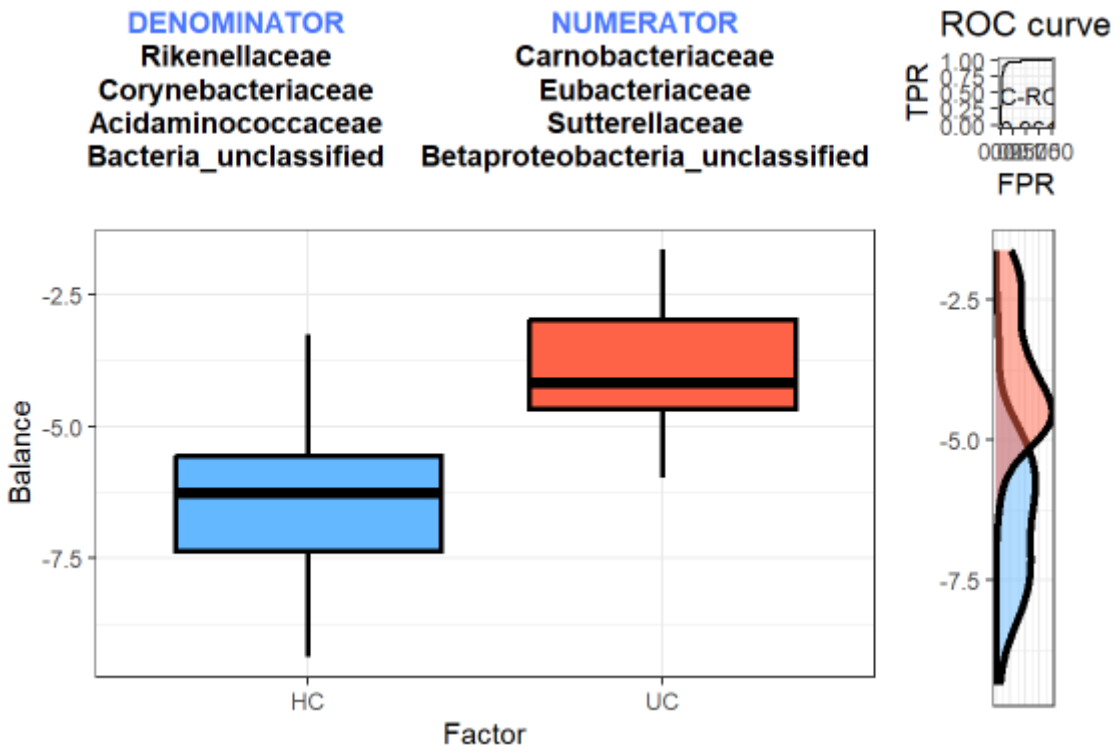


Figura 44 Plot global de la CD. Indicant els taxons en el numerador i el denominador del balanç.

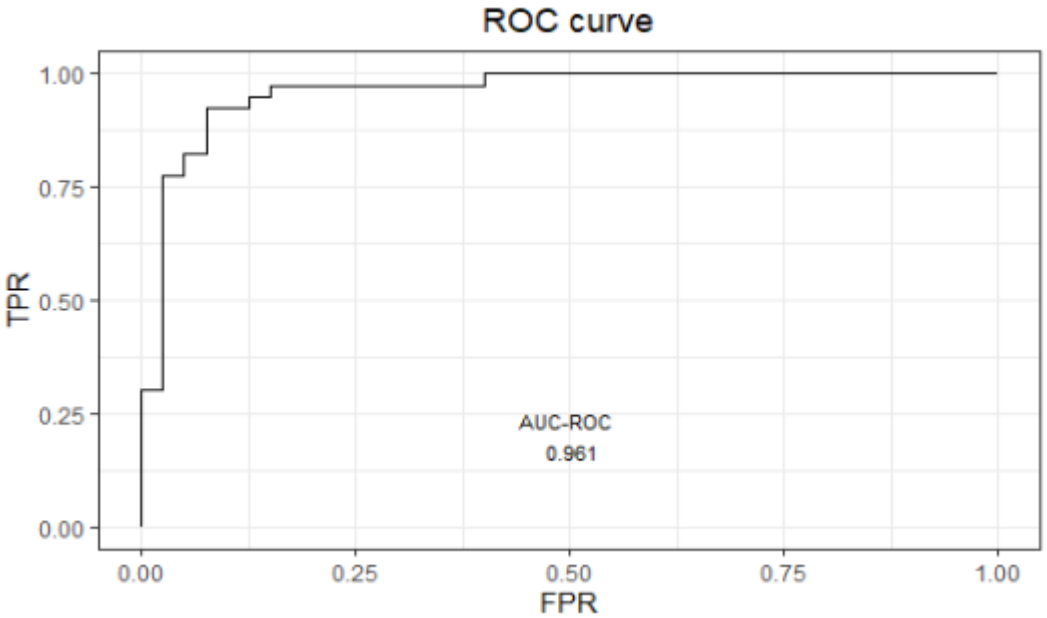


Figura 45 AUC del selbal per la CD.

El valor de la AUC per la CD és de 0.961.

Per altra banda a la CV *selbal* només ha trobat dos grups taxonòmics amb capacitat predictiva: *Verrucomicrobiaceae* en el denominador i *Bacillaceae_1* en el numerador. Sorprenentment però, la AUC és igual a 1 indicant una confiança del 100% en la capacitat predictiva del model.

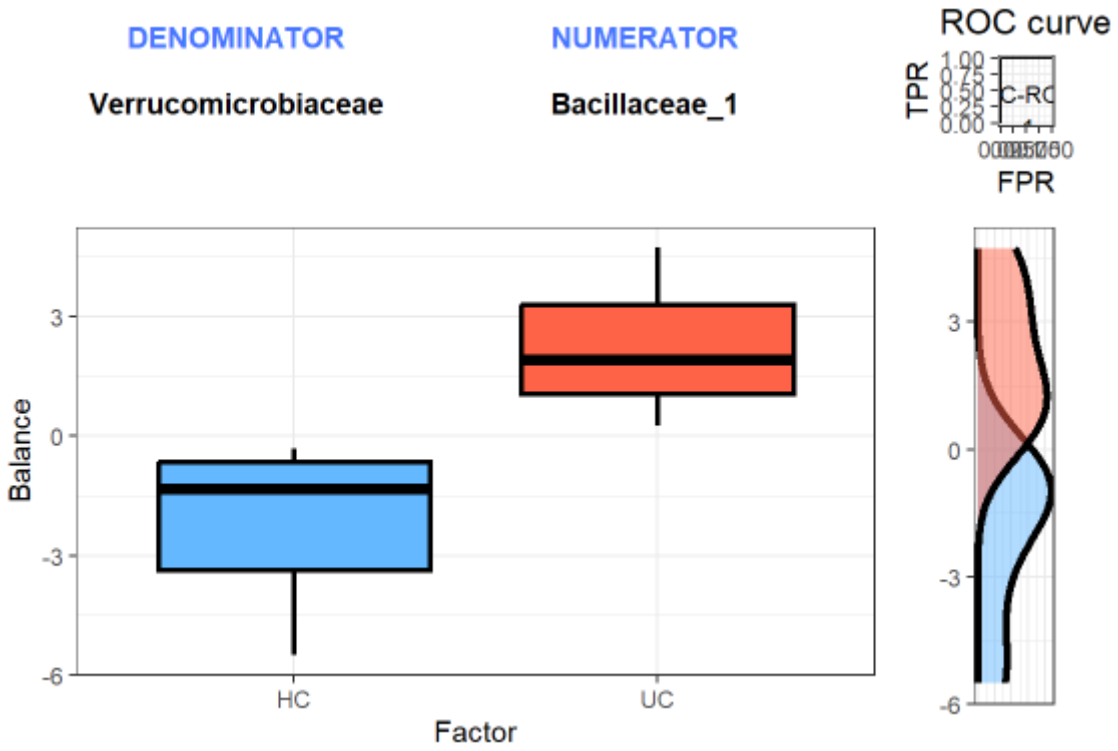


Figura 46 Plot global de la CV. Indicant els taxons en el numerador i el denominador del balanç.

Tant per la CD com per la CV hem hagut de fixar un valor determinat de *threshold* per tal de poder córrer la funció selbal.cv. Per la CD aquest *threshold* ha estat 6 i per la CV 131. Aquest valor s'han trobat mitjançant assaig error intentant trobar el valor més baix possible que admetia la funció. Això es deu a que les nostres dades, i especialment les de la CV contenen nombroses columnes amb valors igual a 0. El missatge d'advertència de R és:

```
Warning: Column(s) containing more than 80% zeros/unobserved values
were found
```

Per aquesta raó els resultats que retorna *selbal* per la CV no són fiables. El valor de AUC = 1 tampoc ho és.

Coda4microbiome.

Mostrem els resultats més significatius del mètode *coda4microbiome* en les dues cohorts a la següent taula:

coda4microbiome	Cohort Discovery	Cohort Validation
n-fold utilitzat	8	3
bionamial deviance	No	Yes
number of taxa	20	23
apparent AUC	0,973125	1
mean cv-AUC	0,8458361	0,230254
sd cv-AUC	0,06784436	0,0479

Taula 5 Resultats comparatius de coda4microbiome entre CD i CV.

Tal i com ha succeït també amb *selbal* la AUC per la CD és molt pròxima a 1 però en la CV és exactament igual a 1.

En el cas de coda4microbiome R també ens retorna el següent missatge de *wanring* per la CV:

```
Warning: one multinomial or binomial class has fewer than 8
observations;
```

És a dir que no tenim prou dades en un dels grups i per tant els resultats de l'algoritme no seran creïbles. De fet, amb aquesta cohort l'algoritme no ens ha retornat el valor de AUC sinó que ha utilitzat la *binomial deviance* que mesura l'allunyament de les nostres dades del model. Tot i baixar el nombre de n-folds a 3 (és el valor mínim que admet coda4microbiome) no hem obtingut la AUC, indicant així que no tenim prou dades i per tant que els resultats obtinguts no són fiables.

En el següent gràfic es mostren els taxons, que en aquest cas són les famílies, més importants a nivell de valor predictiu. S'indica el signe negatiu o positiu de cada un d'ells segons la banda on s'hagin establert del balanç/relació entre ells.

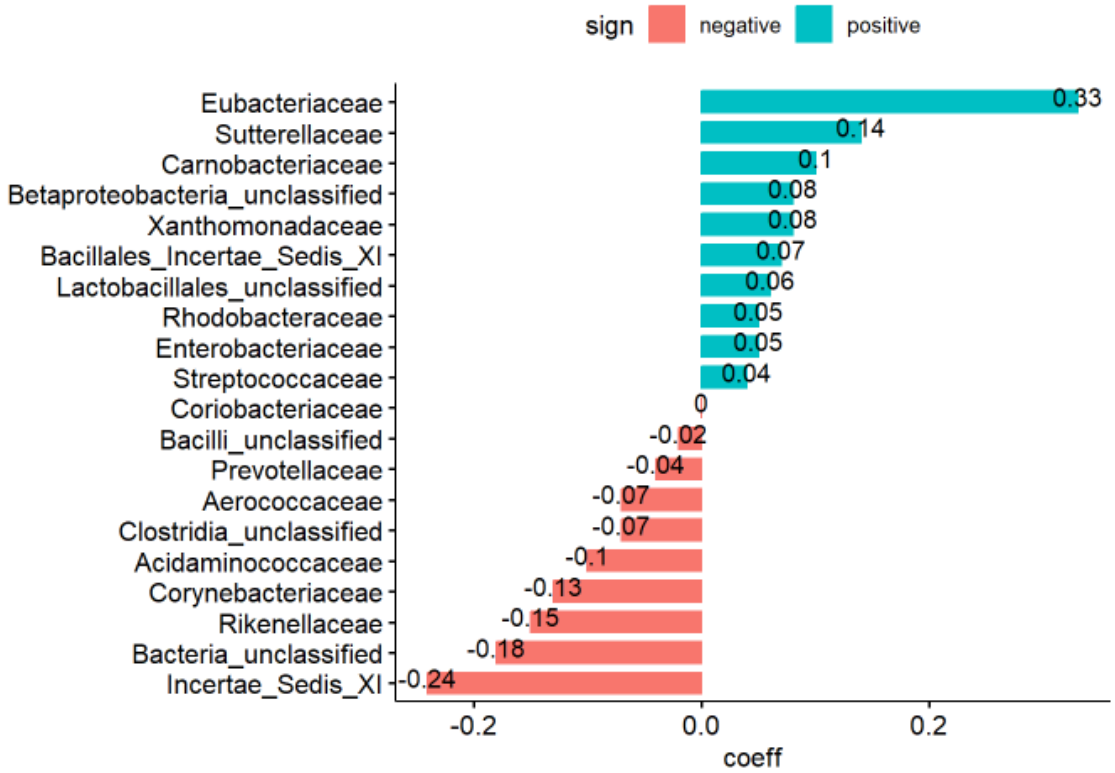


Figura 47 Famílies amb un valor predictiu més important en la CD.

Mostrem el mateix per la CV tot i que els resultats per aquest conjunt de dades no és creïble.

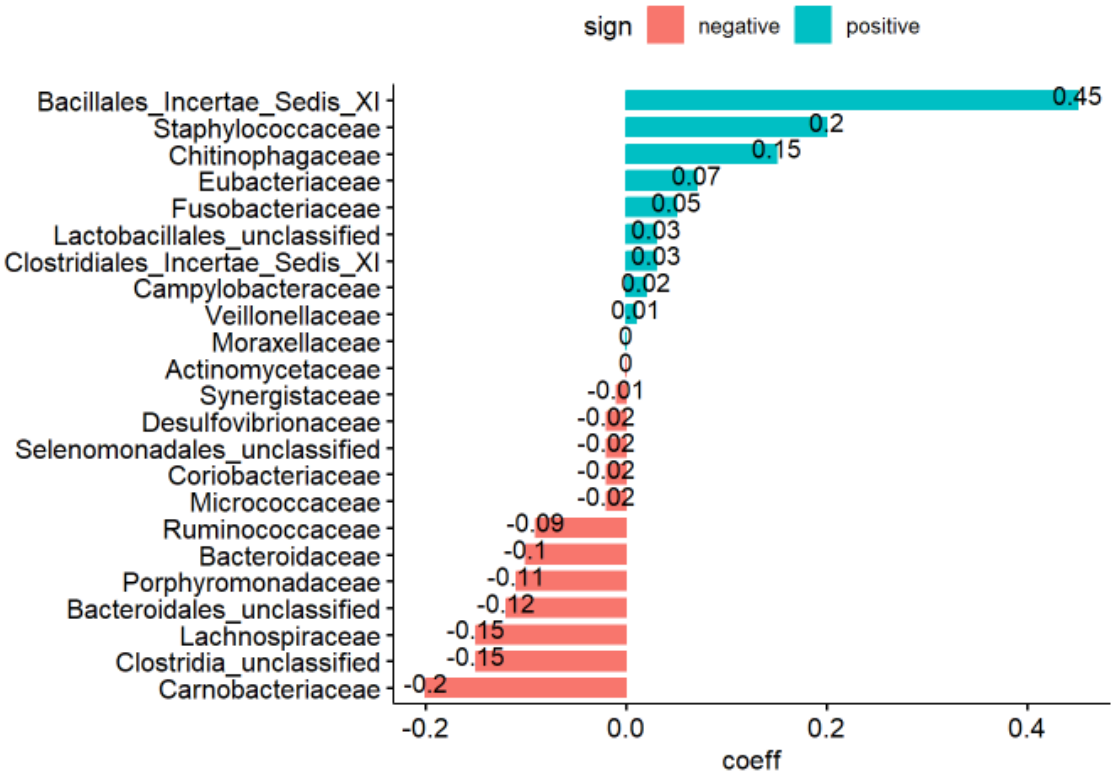


Figura 48 Famílies amb un valor predictiu més important en la CV

Random Forests (*Machine Learning*)

Mostrem els resultats de rendiment del model entrenat contra el subconjunt test en forma de matriu de confusió.

Primer per la CD:

```
Confusion Matrix and Statistics

          Reference
Prediction HC UC
        HC 11  4
        UC  1  8

               Accuracy : 0.7917
                 95% CI : (0.5785, 0.9287)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : 0.003305

                  Kappa : 0.5833

 Mcnemar's Test P-Value : 0.371093

            Sensitivity : 0.9167
            Specificity : 0.6667
         Pos Pred Value : 0.7333
         Neg Pred Value : 0.8889
             Prevalence : 0.5000
         Detection Rate : 0.4583
   Detection Prevalence : 0.6250
      Balanced Accuracy : 0.7917

       'Positive' Class : HC
```

El subconjunt test de la CD és de 24 mostres. Una de les mostres que pertanyen a la condició HC és classificada pel model com a UC (1 fals positiu) i 4 que són

realment de pacients amb UC són classificades com a sanes (4 falsos negatius). Això ens dona una sensibilitat del 91.67% i una especificitat del 66.67%. El model ha estat capaç de classificar correctament el 79.17& de les mostres.

El p-valor [Acc > NIR] retornat de 0.003305 ens informa de que l'exactitud del model és significativament millor que el fet hipotètic de fer la classificació de manera aleatòria.

Kappa és una mesura de concordança entre les prediccions del model i les classes reals, corregida pel fet que podria ocórrer per casualitat. El valor Kappa de 0.5833 indica una correlació moderada entre les prediccions del model i les classes reals.

La prova de *McNemar* és una prova estadística per determinar si els errors comesos per un model depenen de la classe. El valor-p de *Mcnemar* és de 0,371093 i per tant, suggereix que no hi ha una dependència significativa en els errors comesos pel model.

Ara passem a la CV:

```
Confusion Matrix and Statistics

          Reference
Prediction HC UC
        HC  2  0
        UC  1  3

               Accuracy : 0.8333
                 95% CI : (0.3588, 0.9958)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : 0.1094

                  Kappa : 0.6667

 Mcnemar's Test P-Value : 1.0000

            Sensitivity : 0.6667
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 0.7500
             Prevalence : 0.5000
         Detection Rate : 0.3333
   Detection Prevalence : 0.3333
      Balanced Accuracy : 0.8333

       'Positive' Class : HC
```

En aquest model, tot i donar un valor de AUC més elevat que l'anterior veiem que té un p-valor de [Acc > NIR] de 0.1094, per tant, ens indica que la precisió del model no és significativament millor que classificar aleatòriament la classe majoritària.

5. Conclusions i treballs futurs

Durant el desenvolupament d'aquest treball hem après a implementar un *pipeline* de classificació taxonòmica a nivell bacterià a partir de dades de seqüenciació de 16S rRNA de mostres fecals.

A més, hem aconseguit exportar els resultats a un software de tractament estadístic com és R i poder-ne extreure informació de valor i avaluar la reproductibilitat entre dades de diferents estudis.

No obstant, durant el transcurs d'aquest treball han aparegut diferents reptes a nivell tècnic que s'han pogut superar. D'entrada, hem partit amb el *handycap* de la limitació de la capacitat computacional que ha limitat l'abast del treball.

El fet de no disposar de servidors externs ens ha obligat a treballar amb els recursos computacionals més propis de l'àmbit domèstic. Aquest fet, en certa mesura, ha condicionat el procés de selecció de les mostres que acabarien configurant les dues cohorts que s'han comparat al llarg de tot l'estudi.

Durant el procés de classificació taxonòmica a partir de les dades crues s'han anat generant nombrosos fitxers d'extensió considerable. Sobretot al generar els fitxers resultants dels alineaments. En aquest procés s'han generat més de 200GB de dades en la CD en format .txt, entre d'altres fitxers. Per tant, a l'hora de triar les dades per efectuar el mateix anàlisi de la cohort de validació s'ha tingut en compte que el nombre de mostres a processar no fos tant extens com en la de descobriment. Tot i haver pogut desenvolupar totes les parts del treball amb ambdues cohorts hem vist com per exemple durant la part de predicció els resultats obtinguts de la CV no són fiables degut a la falta de dades.

Passem ara a comentar els resultats obtinguts. Dividirem aquesta secció en el que han estat les dues fases de desenvolupament del treball. Anàlisi *upstream*, *downstream* i predicció.

Conclusions de l'anàlisi *Upstream*

Si ens fixem en la taula 2 que resumeix els resultats de l'anàlisi *upstream* amb *Mothur*, destaca la notable diferència en el nombre de *reads* finals. És sorprenent com, després de tots els passos de filtratge successius i eliminació de soroll de fons, només conservem un 1,94% dels *reads* inicials per les mostres de la CD, mentre que per les mostres de la CV aconseguim mantenir fins al 17,14%.

Aquesta disparitat pot explicar-se pel fet que les mostres de la CD contenien inicialment moltes menys seqüències úniques en comparació amb les mostres de la CV. En aquest cas, tot i que s'hagués realitzat una seqüenciació més profunda, no hauríem descobert noves seqüències. Bàsicament, estaríem seqüenciant *reads* que ja havíem seqüenciat anteriorment.

Cal destacar que aquesta explicació és només una hipòtesi i no tenim evidències concretes per confirmar-la. És possible que hi hagi altres factors que

contribueixin a aquesta diferència de *reads* finals entre les mostres de CD i CV com les diferents regions amplificades durant el procés de seqüenciació. Recordem que en l'estudi de la CD s'han utilitzat primers específics només per la regió hipervariable V4. Mentre que en l'estudi de la CV s'ha amplificat V3 i V4 conjuntament. En aquest sentit hi ha nombrosos estudis que fan referència a la importància d'escollir la regió a seqüenciar i els efectes que poden tenir prendre una decisió o una altra (Bukin et al., 2022; Fadeev et al., 2021)

És important considerar també la possibilitat que pugui haver-hi algun tipus d'error no detectat durant l'anàlisi, ja sigui en els passos de filtratge o en altres aspectes del procés.

De cara a un futur s'hauria de realitzar una revisió exhaustiva de tots els passos de l'anàlisi i comprovar si hi ha alguna anomalia o error en les dades obtingudes. Això podria implicar la verificació de les etapes de filtratge, el control de qualitat de les seqüències i la identificació de possibles factors que puguin influir en la discrepància observada. L'exploració de diferents possibilitats i la investigació més a fons podrien ajudar-nos a comprendre millor aquesta disparitat entre les mostres de la CD i la CV i a descartar o confirmar la presència d'errors o altres explicacions.

Conclusions de l'anàlisi *Downstream*

Al llarg de l'anàlisi *dowsnstream* hem trobat similituds i deferències entre les dades que componen les dues cohorts d'estudi.

Durant l'inici de l'anàlisi, vam observar, mitjançant la visualització de les abundàncies relatives a nivell de phylum, una clara diferència en la composició taxonòmica de les mostres entre els dos estudis. Es va observar una presència notablement més alta del phylum *Actinobacteria* a la CD en comparació amb la CV, mentre que podem observar el fenomen invers amb el phylum *Bacteroidetes*. D'altra banda, és evident que el phylum dominant en la majoria de les mostres tant de la CD com de la CV és *Firmicutes*. Aquesta informació ens proporciona indicis importants sobre les diferències en la composició taxonòmica entre els dos estudis i destaca el paper significatiu del phylum *Firmicutes* en ambdues cohorts.

En aquest sentit és interessant considerar el que es coneix com a *Firmicutes*/*Bacteroidetes ratio*. S'ha descrit a la literatura com aquests dos phylums junts són els més representatius pel que fa a la composició del microbioma humà i junts que poden arribar fins al 90% del *coverage* total en algunes mostres i com s'estableix una relació directe amb factors com l'edat o patologies com ara l'obesitat (Magne et al., 2020; Mariat et al., 2009; Qin et al., 2010).

Per aquest motiu pensem que les composicions que observem a la figura 23 corresponents a les abundàncies relatives de la CD no són del tot fiables o si més no, no reflecteixen el paradigma més estès dintre de la comunitat científica. Això, juntament amb les consideracions ja exposades sobre les diferències

observades entre les dues cohorts durant l'anàlisi *upstream*, ens fan sospitar de l'existència d'algun error durant aquesta part de l'estudi.

Però si mirem el que es diu a l'article original d'on provenen les dades podem llegir el següent: "Gram-negative microorganisms (i.e., Bacteroidetes, Proteobacteria, and Verrucomicrobia) were detected in a lower than expected abundance (…)". Això significa que els mateixos investigadors que vàrem generar les dades ja es van trobar amb uns resultats similars als nostres pel què fa a la baixa abundància relativa del phylum *Bacteroidetes*, entre d'altres. Per tant, la hipotèsi de que hi hagués hagut algun error durant la classificació taxonòmica queda descartada. No obstant, encara no queda clar si el baix *coverage* observat al final de l'anàlisi upstream es deu a un error del mètode emprat o a una eventual naturalesa corrompuda de les dades crues.

Seguint amb les consideracions de l'anàlisi downstream podem veure que pel que fa a les diversitats alpha i beta es reprodueix el mateix patró a les dues cohorts. Si bé és cert que no podem afirmar que hi hagi diferències en la diversitat alpha de la CV, sí que mitjançant el gràfic de caixes es pot apreciar una certa tendència a una menor diversitat al grup amb UC. La CV al estar composta de moltes menys mostres requeriria d'una diferència molt més pronunciada per poder arribar a rebutjar la hipotèsi nul·la de no diferències entre grup control sa i grup malalt. A la bibliografia podem trobar estudis amb resultats sobre diversitat alpha similars als trobats aquí (Abdel-Rahman & Morgan, 2023; Barberio et al., 2022; Dahal et al., 2023; S. Zhu et al., 2022). Però a més, a l'etudi de la CD les diferències en la alpha diversiy van en el mateix sentit: "*Significant differences in diversity (Shannon, Simpson) were similarly significant across both disease cohorts [CD and UC] and HC (pFDR < 0.001)*".

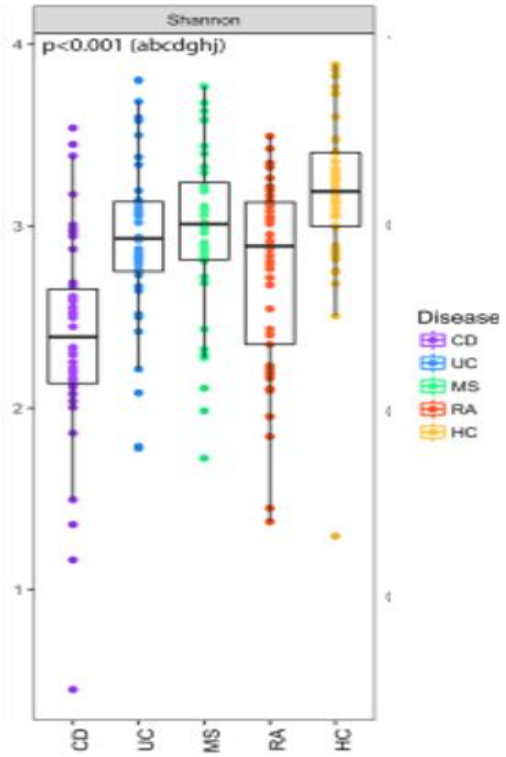


Figura 49 Índex shannon de diversitat alpha de l'estudi que compon la CD.

Al contrari, observem un patró diferencial de diversitat beta que sí que pot extrapolar estadísticament d'una cohort a l'altra. Tant en els diagrames de caixa com en les estadístiques, es pot apreciar una major diversitat beta en el grup amb UC en les dues cohorts, però amb una diferència més notable a la CV. A més, s'han trobat referències a la bibliografia on s'observa que, a partir de l'estudi de la diversitat beta, es pot apreciar una afectació en la composició del microbioma intestinal dels pacients amb malalties inflamatòries de l'intestí (IBD) (Abdel-Rahman & Morgan, 2023). Al mateix estudi de les dades de la CD també es menciona que: "*(…) whereas the HC cohort showed the least variability*". Fent referència a què en tots els altres grups, incloent el de pacients amb UC, es pot observar una variabilitat més gran respecte del grup control.

En consonància amb aquesta troballa i les diferències observades en la diversitat beta, també podem observar com la diferenciació mitjançant l'anàlisi de components principals és visible en les dues cohorts, però amb una presència més evident a la CV. De fet, en aquest últim conjunt de dades pràcticament no hi ha solapament entre els dos grups, quedant clarament separats en el gràfic. Aquesta clara separació reflecteix una major diferència en la composició del microbioma entre els grups de la CV.

Un aspecte destacable és la proporció de variança explicada per les principals components en cada cohort. En el cas de la CV, les dues principals components expliquen una proporció més gran de la variança (5,8% + 5,6%) en comparació amb la CD (3,9% + 3,4%). Aquesta diferència indica que les principals components de la CV tenen un major poder explicatiu sobre la variabilitat present en les mostres.

Una observació curiosa és la uniformitat o poca diferència en la proporció de variança explicada per cada una de les components en la CV. A diferència de la CD, on la primera component destaca, a la CV, qualsevol de les components té una millor capacitat per explicar la proporció de variança en el conjunt de dades. Aquesta consistència en l'explicació de la variança pot suggerir una major homogeneïtat en la distribució de les mostres i una major influència de múltiples factors en la composició del microbioma en la CV.

Més amunt, ja hem abordat la diferència en l'abundància relativa en el rang taxonòmic de phylum. A més d'això, hem realitzat l'anàlisi en els següents dos nivells taxonòmics: classe i ordre. En aquesta anàlisi, s'han identificat diversos grups taxonòmics amb abundàncies relatives significativament diferents entre els grups HC i UC, especialment en el cas de la CV. No obstant això, només s'han trobat dues classes que presenten aquesta diferència en les dues cohorts: *Deltaproteobacteria* i *Bacilli*. S'ha observat una major abundància de *Deltaproteobacteria* en els pacients sans i una major abundància de *Bacilli* en els pacients amb colitis ulcerosa en les dues cohorts. Això suggereix un patró de disbiòsi en el qual la presència de *Deltaproteobacteria* està associada a pacients sans, mentre que *Bacilli* és més abundant en pacients amb colitis ulcerosa. Aquestes troballes concorden amb algunes referències bibliogràfiques (Dahal et al., 2023; Graspeuntner et al., 2019; Maslennikov et al., 2021; Nascimento et al., 2020), tot i que d'altres estudis presenten resultats contraris (Z. F. Dai et al., 2021; Maldonado-Arriaga et al., 2021).

També al mateix estudi de les dades de la CV es mencionen findings que van en el mateix sentit: "*We demonstrated that diversity in bacterial composition among Firmicutes concerned a considerable increase in Bacillaceae, Clostridiaceae 1, Enterococcaceae, Staphylococcaceae, Streptococcaceae and Veillonellaceae, (...)*". La mojria d'aquestes famílies pertanyen a la classe *Bacilli*. En concret *Bacillaceae, Enterococcaceae, Staphylococcaceae* i *Streptococcaceae* estan dintre del grup *Bacilli*.

Malgrat existeixi a la literatura diferents evidències en sentit contrari sobre l'abundància relativa de les classes *Deltaproteobacteria* i *Bacilli* en pacients amb UC, és important tenir en compte diverses consideracions per solucionar aquesta discrepància. Primerament, seria essencial avaluar la consistència i la reproductibilitat dels resultats mitjançant estudis addicionals amb un nombre més gran de mostres i una diversitat de poblacions més àmplia.

D'altra banda, s'ha observat que un altre grup, *Bacteria_unclassified*, colonitza diferencialment la microbiota intestinal dels dos grups HC i UC. No obstant això, com indica el seu nom, aquest grup és una categoria general que inclou les seqüències bacterianes que encara no han estat classificades taxonòmicament. Per tant, trobar una diferència d'abundància en aquest grup no aporta informació significativa ja que pot contenir bacteris diferents en cada cohort.

Conclusions sobre la predicció

En l'apartat de predicció hem comparat el rendiment de dues metodologies basades en models estadístics i una altra basada en aprenentatge automàtic (ML).

El primer que se'n desprèn dels resultats obtinguts és que amb les tres aproximacions diferents la CV no conté suficients dades com per poder utilitzar aquests mètodes i obtenir uns resultats fiables. Hem vist que tant amb l'algoritme *selbal* com *coda4microbiome* teníem una sèrie d'inconvenients a l'hora de córrer les pertinents funcions amb les dades d'aquesta cohort. De la mateixa manera també hem vist com mitjançant un algoritme basat en *Random Forest Classification* tampoc obtenim resultats fiables. En tots els casos, tot i obtenir valors de AUC lleugerament millors en la CV que en la CD hem vist que aquests nivells de precisió dels models no tenien validesa. Ho hem vist clarament en el model de ML on el p-valor de la [Acc > NIR] és de 0.1094 en la CV mentre que és del 0.003305 en la CD. Aquest p-valor ens indica si la precisió del model és significativament millor que la taxa de no informació. Aquesta taxa en el nostre cas és igual a 0.5 donat que només tenim dues classes possibles. Per tant, la precisió del model en el cas de la CV no és millor que el fet de suposar aleatòriament la classe majoritària dintre del conjunt de dades.

A l'estudi d'on provenen les dades de la CD també entrenen un model mitjançant un algoritme basat en *Random Forest Classification* i en avaluar-ne els resultats arriben a la següent conclusió: "*The low abundance of some taxa (e.g., Bacteroidetes and Proteobacteria) did not noticeably reduce the ability of our machine learning classifiers to assign diseased and healthy samples. These*

study findings reveal that microbiota characterization can be performed using "imperfect" datasets, which may be further overcome through the use of robust machine learning approaches." Durant el nostre anàlisi de la CD, hem observat resultats inesperats en les dades, indicant la possible presència d'un problema durant el procés de seqüenciació. Malgrat això, l'estudi d'on provenen les dades confirma que, utilitzant les metodologies emprades, és possible obtenir nivells de precisió significativament alts i fiables per a tasques de predicció, sempre que l'algoritme sigui prou robust.

Si tenim aquest fet en compte i per tant prescindim d'avaluar els models generats a partir de la CV podem comparar les diferents metodologies i valorar quina d'elles té una millor *performance*.

	AUC
selbal	0,961
coda4microbiome	0.973125
Random Forests	0.7917

Taula 6 Comparació de valor de AUC entre els 3 models predictius utilitzats en la CD.

Veiem que tant *selbal* com *coda4microbiome* donen valors de precisió molt alts i propers al 100%. Hem de tenir en compte però que el resultat del *Random Forests* es veu fortament influït per dos factors diferents: la llavor d'aleatorització amb què s'inicií el procés d'entrenament i el *ratio* de mostres destinades a traint/test.

Si mirem la bibliografia publicada sobre aquestes metodologies i en concret l'article on es presenta *coda4microbiome* veiem que aquesta metodologia és plantejada com una versió actualitzada i millorada de *selbal* en termes de rapidesa i flexibilitat (Calle et al., 2023). Per tant, podem afirmar que mitjançant el paquet de R *coda4microbiome* hem assolit els millors resultats en quant a capacitat predictiva.

Assoliment dels Objectius

A l'inici del treball plantejàvem 2 objectius generals que es veiem desglossats en 6 objectius més específics.

A continuació mencionem cada un dels objectius i el grau d'assoliment de cadascun d'ells.

1- **Recopilació de dades**: A partir d'estudis publicats hem recopilat dades de pacients amb colitis ulcerosa (UC) i pacients sans per crear dues cohorts, una de descoberta (*discovery*) i una de validació (*validation*). Hem assegurat que les dues cohorts tinguin suficients dades de pacients prou semblants per garantir la validesa dels resultats.

2- **Instal·lació i utilització de programari**: Hem instal·lat i utilitzat el programari necessari per a la creació i utilització del *pipeline* de classificació taxonòmica. Aquests *softwares* ens han permès processar

les dades i aplicar els algoritmes de classificació necessaris per a l'objectiu 3.

3- **Implementació del pipeline de classificació taxonòmica**: Hem implementat un *pipeline* de classificació taxonòmica utilitzant les dades de les mostres de pacients sans i amb UC. Aquest *pipeline* s'ha realitzat mitjançant el programari *Mothur* de classificació de dades de 16S rRNA.

4- **Extreure un patró de disbiòsi de les mostres que conformen la cohort de *discovery*:** Si bé és cert que s'ha pogut dur a terme tot l'anàlisi estadístic, i també la part de predicció amb aquesta cohort, hem d'admetre que les dades de la CD presenten una certa anomalia en quant al patró d'abundàncies relatives de *phylums* que presenten i no concorden amb el que s'esperava obtenir. Ja s'ha exposat a la part de conclusions de l'anàlisi *downstream* que el fet de no observar un ratio *Firmicutes*/*Bacteroidetes* típic en la CD és indici de que hi pugui haver algun tipus de problema amb aquestes dades. També hem mencionat però que en el propi estudi d'on hem extret les dades s'han trobat amb un problema similar i tot i així han seguit amb l'anàlisi estadístic.

5- **Avaluar la reproductibilitat de la disbiòsi entre diferents estudis**: Tot i no haver obtingut un patró clarament diferenciat de disbiòsi a partir de la cohort de *discovery* sí que hem trobat punts en comú com algunes diferències a nivell de composició d'algunes classes que hem vist que es reproduïen en els dos conjunts de dades. També hem vist el mateix en quan a la diferència entre diversitat beta entre els grups sans i malalts en totes dues cohorts.

6- **Implementació d'un o més mètodes de predicció**: Per tal d'avaluar el funcionament i el rendiment de les diferents metodologies de predicció utilitzades, hem realitzat diverses proves. Aquests mètodes ens permeten determinar si una nova mostra no classificada pertany a un grup específic o a un altre. En el cas de *coda4microbiome*, hem obtingut resultats de precisió excepcionalment bons, el que demostra que és una eina de gran valor diagnòstic. També hem observat que, malgrat utilitzar conjunts de dades imperfectes com en el cas de la CD, podem aplicar aquestes metodologies amb èxit. No obstant això, en el cas de la CV, ens hem trobat amb un grup amb un nombre molt limitat de dades, cosa que ha afectat negativament el funcionament.

Considerant tot l'exposat fins ara, creiem que la metodologia i l'enfocament utilitzats durant el treball han estat realment adequats per aconseguir els objectius que ens havíem proposat inicialment, malgrat haver-nos enfrontat a certes deficiències en la naturalesa de les dades, tant en la CD com en la CD. De fet, gairebé no hem hagut de fer canvis significatius en la planificació del desenvolupament del treball, excepte alguna petita desviació temporal sense gaire importància.

Propostes de Futur

Proposem diverses aproximacions per tal de dur a terme en un futur que ens permetin millorar i/o ampliar les tasques que hem pogut desenvolupar en aquest treball:

Reproduir tot l'anàlisi amb altres conjunts de dades:

Hem vist que tant en una cohort com en l'altre ens hem trobat amb diferents problemàtiques que o bé no n'hem identificat l'origen (com en la classificació taxonòmica de la CD) o bé hem arribat a la conclusió que les dades no eren prou potents com per realitzar determinades tasques (com en el cas de la predicció amb la CV). Per tant, una primera aproximació, i la més senzilla, és buscar altres dades que hagin estat publicades i intentar fer tots els anàlisis pertinents de nou.

Utilitzar altres versions de SILVA o altres bases de dades:

Durant l'etapa de classificació taxonòmica es du a terme un alineament de seqüències contra una base de dades d'alineaments previs. En el nostre cas hem utilitzat la base de dades SILVA. Aquesta base de dades, com tantes d'altres, es troba en contínua millora i de fet n'existeixen diferents versions. Per tant, es podria emprar altres versions ja existents més complertes o bé mantenir-se al corrent de les actualitzacions periòdiques que hi van havent. També existeixen altres bases de dades similars com *Greengenes*.

Modificacions en la imatge de *Mothur*:

Per tal de fer la classificació taxonòmica s'ha utilitzat el programa *Mothur* a través d'una imatge de *Docker*. Aquestes imatges són modificables i de fet hi ha algunes funcions descrites en el MiSeq SOP de *Mothur* que no les hem pogut dur a terme com per exemple la funció cimera.vsearch(). Aquesta funció no pertany pròpiament a *Mothur* sinó a un altre software anomenat *vsearch*. El que fa és buscar possibles seqüències quimèriques resultants d'errors durant el procés d'amplificació per PCR. Per tant, una altra opció a intentar realitzar de cara a un futur és instal·lar aquestes modificacions a la imatge de *Mothur* amb la qual treballem i córrer tot el *pipeline* incorporant també aquest pas.

Altres tecnologies de seqüenciació:

En aquest apartat podríem incloure la tecnologia *shotgun sequencing* que ja s'ha esmentat a l'apartat 1.4 de la introducció però en aquest cas s'hauria de canviar tot el *pipeline* de classificació taxonòmica degut a la naturalesa del tipus de dades que genera aquesta tècnica.

Una altra aproximació més recent és la que ofereix la nova tecnologia *nanopore sequencing*. *Nanopore sequencing* és un mètode de seqüenciació d'ADN o ARN que utilitza nanoporus, que són petits canals o porus pels quals es fa passar les molècules d'ADN o ARN. Durant la seqüenciació, les propietats elèctriques del nanoporus canvien a mesura que diferents nucleòtids passen a través d'ell, permetent la identificació i seqüenciació dels nucleòtids.

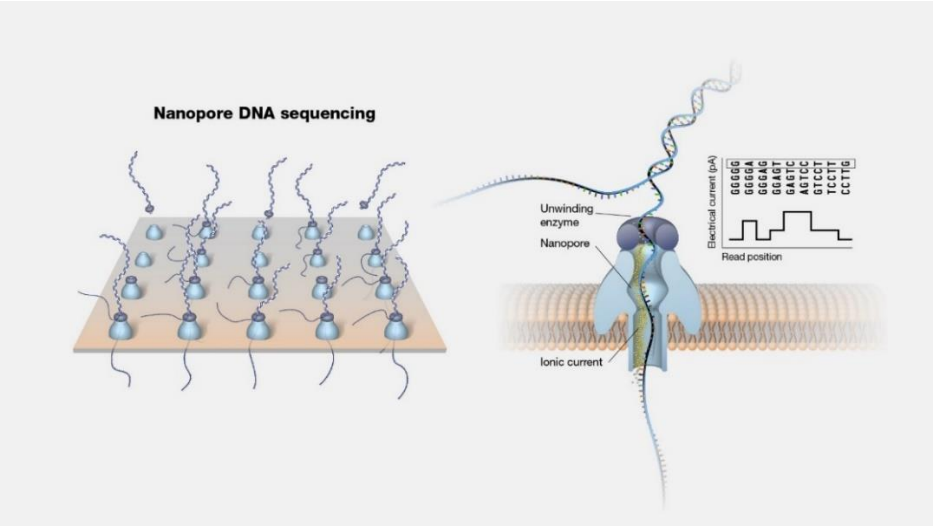


Figura 50 Esquema del funcionament del mètode nanopore sequencing (*Nanopore DNA Sequencing*, n.d.).

Aquesta tecnologia permetria poder utilitzar el mateix *pipeline* de classificació taxonòmica que hem emprat. Dintre de la comunitat científica la seqüenciació per nanoporus cada cop està més estesa i la metagenòmica és na de les disciplines on s'està usant més (Heikema et al., 2020; X. Ma et al., 2017; Rodríguez-Pérez et al., 2021; Rozas et al., 2022).

Una de les avantatges que presenta aquesta tecnologia és que no depèn del procés d'amplificació per PCR i teòricament permet seqüenciar *reads* sense límit de llargada (Leggett & Clark, 2017; Matsuo et al., 2021). Això també comporta que sigui una tecnologia d'especial interès en altres camps com la epigenètica donat que permet seqüenciar el DNA nadiu, és a dir, respectant tots aquells nucleòtids que presentin modificacions de tipus epigenètica (Liu et al., 2021). No obstant, per seguir un protocol de classificació taxonòmica mitjançant la seqüenciació per nanoporus hem de seguir fent un pas d'amplificació del gen 16S:

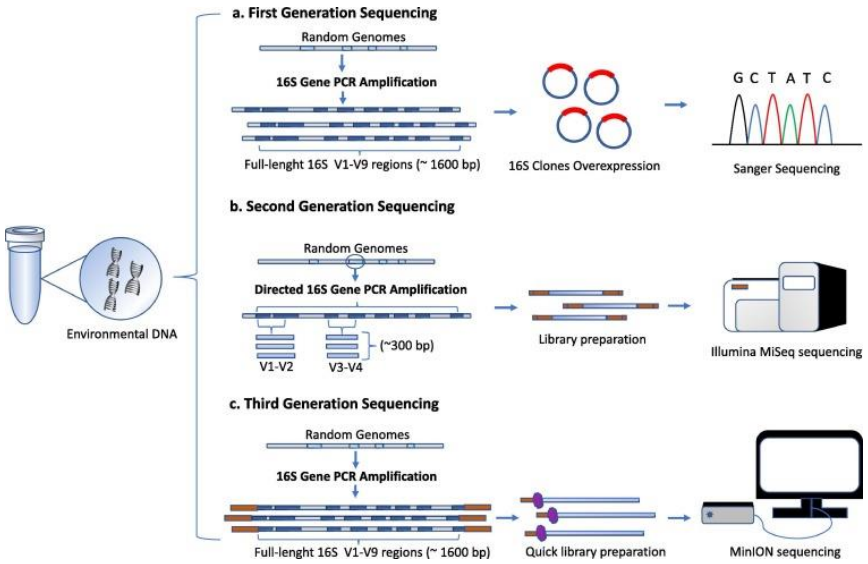


Figura 51 Comparació del diferents processos de seqüenciació existents (Santos et al., 2020)**.**

6. Glossari

AUC = Area Under the Curve

CD = Cohort Discovery

CV = Cohort Validation

DNA = Deoxyribonucleic Acid

GLM = Generalized Linear Models

HC = Healthy Controls

HMP = Human Microbiome Project

IBD = Inflammatory Bowel Diseases

ML = Machine Learning

OTU = Operational Taxonomic Unit

PCA = Principal Component Analysis

PCR = Polymerase Chain Reaction

RNA = Ribonucleic Acid

rRNA = ribosomic Ribonucleic Acid

SOP = Standard Operations Protocol

UC = Ulcerative Colitis

7. Bibliografia

Abdel-Rahman, L. I. H., & Morgan, X. C. (2023). Searching for a Consensus Among Inflammatory Bowel Disease Studies: A Systematic Meta-Analysis. *Inflammatory Bowel Diseases*, *29*(1), 125–139. https://doi.org/10.1093/ibd/izac194

Almeida, A., Nayfach, S., Boland, M., Strozzi, F., Beracochea, M., Shi, Z. J., Pollard, K. S., Sakharova, E., Parks, D. H., Hugenholtz, P., Segata, N., Kyrpides, N. C., & Finn, R. D. (2021). A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nature Biotechnology*, *39*(1), 105–114. https://doi.org/10.1038/S41587-020-0603-3

Amos, G. C. A., Sergaki, C., Logan, A., Iriarte, R., Bannaga, A., Chandrapalan, S., Wellington, E. M. H., Rijpkema, S., & Arasaradnam, R. P. (2021). Exploring how microbiome signatures change across inflammatory bowel disease conditions and disease locations. *Scientific Reports*, *11*(1), 18699. https://doi.org/10.1038/s41598-021-96942-z

Bakir-Gungor, B., Hacilar, H., Jabeer, A., Nalbantoglu, O. U., Aran, O., & Yousef, M. (2022). Inflammatory bowel disease biomarkers of human gut microbiota selected via different feature selection methods. *PeerJ*, *10*. https://doi.org/10.7717/PEERJ.13205/SUPP-10

Barberio, B., Facchin, S., Patuzzi, I., Ford, A. C., Massimi, D., Valle, G., Sattin, E., Simionati, B., Bertazzo, E., Zingone, F., & Savarino, E. V. (2022). A specific microbiota signature is associated to various degrees of ulcerative colitis as assessed by a machine learning approach. *Gut Microbes*, *14*(1). https://doi.org/10.1080/19490976.2022.2028366

Breitwieser, F. P., Lu, J., & Salzberg, S. L. (2019). A review of methods and databases for metagenomic classification and assembly. *Briefings in Bioinformatics*, *20*(4), 1125. https://doi.org/10.1093/BIB/BBX120

Bukin, Yu. S., Galachyants, Yu. P., Morozov, I. V., Bukin, S. V., Zakharenko, A. S., & Zemskaya, T. I. (2022). Author Correction: The effect of 16S rRNA region choice on bacterial community metabarcoding results. *Scientific Data*, *9*(1), 94. https://doi.org/10.1038/s41597-022-01246-0

Butler, M. I., Bastiaanssen, T. F. S., Long-Smith, C., Morkl, S., Berding, K., Ritz, N. L., Strain, C., Patangia, D., Patel, S., Stanton, C., O'Mahony, S. M., Cryan, J. F., Clarke, G., & Dinan, T. G. (2023). The gut microbiome in social anxiety disorder: evidence of altered composition and function. *Translational Psychiatry*, *13*(1), 95. https://doi.org/10.1038/s41398-023-02325-5

Calle, M. L., Pujolassos, M., & Susin, A. (2023). coda4microbiome: compositional data analysis for microbiome cross-sectional and longitudinal studies. *BMC Bioinformatics*, *24*(1), 1–19. https://doi.org/10.1186/S12859-023-05205-3/TABLES/3

Cammann, D., Lu, Y., Cummings, M. J., Zhang, M. L., Cue, J. M., Do, J., Ebersole, J., Chen, X., Oh, E. C., Cummings, J. L., & Chen, J. (2023). Genetic correlations between Alzheimer's disease and gut microbiome genera. *Scientific Reports*, *13*(1), 5258. https://doi.org/10.1038/s41598-023-31730-5

Chiu, C. Y., & Miller, S. A. (2019). Clinical metagenomics. *Nature Reviews Genetics 2019 20:6*, *20*(6), 341–355. https://doi.org/10.1038/s41576-019-0113-7

Chugh, R. S., Bhatia, V., Khanna, K., & Bhatia, V. (2020). A Comparative Analysis of Classifiers for Image Classification. *2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, 248–253. https://doi.org/10.1109/Confluence47617.2020.9058042

Consortium, I. H. G. S. (2004). Finishing the euchromatic sequence of the human genome. *Nature*, *431*(7011), 931–945. https://doi.org/10.1038/nature03001

Dahal, R. H., Kim, S., Kim, Y. K., Kim, E. S., & Kim, J. (2023). Insight into gut dysbiosis of patients with inflammatory bowel disease and ischemic colitis. *Frontiers in Microbiology*, *14*. https://doi.org/10.3389/fmicb.2023.1174832

Dai, L., Tang, Y., Zhou, W., Dang, Y., Sun, Q., Tang, Z., Zhu, M., & Ji, G. (2021). Gut Microbiota and Related Metabolites Were Disturbed in Ulcerative Colitis and Partly Restored After Mesalamine Treatment. *Frontiers in Pharmacology*, *11*, 2337. https://doi.org/10.3389/FPHAR.2020.620724/BIBTEX

Dai, Z. F., Ma, X. Y., Yang, R., Wang, H., Xu, D., Yang, J., Guo, X., Meng, S., Xu, R., Li, Y., Xu, Y., Li, K., & Lin, X. (2021). Intestinal flora alterations in patients with ulcerative colitis and their association with inflammation. *Experimental and Therapeutic Medicine*, *22*(5), 1322. https://doi.org/10.3892/etm.2021.10757

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., Huber, T., Dalevi, D., Hu, P., & Andersen, G. L. (2006). Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. *Applied and Environmental Microbiology*, *72*(7), 5069–5072. https://doi.org/10.1128/AEM.03006-05

Dhungel, E., Mreyoud, Y., Gwak, H. J., Rajeh, A., Rho, M., & Ahn, T. H. (2021). MegaR: an interactive R package for rapid sample classification and phenotype prediction using metagenome profiles and machine learning. *BMC Bioinformatics*, *22*(1), 1–12. https://doi.org/10.1186/S12859-020-03933-4/TABLES/3

Escobar-Zepeda, A., De León, A. V. P., & Sanchez-Flores, A. (2015). The road to metagenomics: From microbiology to DNA sequencing technologies and bioinformatics. *Frontiers in Genetics*, *6*(DEC), 348. https://doi.org/10.3389/FGENE.2015.00348/BIBTEX

Fadeev, E., Cardozo-Mino, M. G., Rapp, J. Z., Bienhold, C., Salter, I., Salman-Carvalho, V., Molari, M., Tegetmeyer, H. E., Buttigieg, P. L., & Boetius, A. (2021). Comparison of Two 16S rRNA Primers (V3–V4 and V4–V5) for Studies of Arctic Microbial Communities. *Frontiers in Microbiology*, *12*. https://doi.org/10.3389/fmicb.2021.637526

Ferretti, P., Farina, S., Cristofolini, M., Girolomoni, G., Tett, A., & Segata, N. (2017). Experimental metagenomics and ribosomal profiling of the human skin microbiome. *Experimental Dermatology*, *26*(3), 211–219. https://doi.org/10.1111/EXD.13210

Forbes, J. D., Chen, C. Y., Knox, N. C., Marrie, R. A., El-Gabalawy, H., De Kievit, T., Alfa, M., Bernstein, C. N., & Van Domselaar, G. (2018). A comparative study of the gut microbiota in immune-mediated inflammatory diseases—does a common dysbiosis exist? *Microbiome*, *6*(1). https://doi.org/10.1186/S40168-018-0603-4

Godlewska, U., Brzoza, P., Kwiecień, K., Kwitniewski, M., & Cichy, J. (2020). Metagenomic Studies in Inflammatory Skin Diseases. *Current Microbiology*, *77*(11), 3201. https://doi.org/10.1007/S00284-020-02163-4

Gonzalez, C. G., Mills, R. H., Zhu, Q., Sauceda, C., Knight, R., Dulai, P. S., & Gonzalez, D. J. (2022). Location-specific signatures of Crohn's disease at a multi-omics scale. *Microbiome*, *10*(1), 1–15. https://doi.org/10.1186/S40168-022-01331-X/FIGURES/4

Graspeuntner, S., Waschina, S., Künzel, S., Twisselmann, N., Rausch, T. K., Cloppenborg-Schmidt, K., Zimmermann, J., Viemann, D., Herting, E., Göpel, W., Baines, J. F., Kaleta, C., Rupp, J., Härtel, C., & Pagel, J. (2019). Gut Dysbiosis With Bacilli Dominance and Accumulation of Fermentation Products Precedes Late-onset Sepsis in Preterm Infants. *Clinical Infectious Diseases*, *69*(2), 268–277. https://doi.org/10.1093/cid/ciy882

Gupta, V. K., Cunningham, K. Y., Hur, B., Bakshi, U., Huang, H., Warrington, K. J., Taneja, V., Myasoedova, E., Davis, J. M., & Sung, J. (2021). Gut microbial determinants of clinically important improvement in patients with rheumatoid arthritis. *Genome Medicine*, *13*(1), 149. https://doi.org/10.1186/s13073-021-00957-0

Heikema, A. P., Horst-Kreft, D., Boers, S. A., Jansen, R., Hiltemann, S. D., de Koning, W., Kraaij, R., de Ridder, M. A. J., van Houten, C. B., Bont, L. J., Stubbs, A. P., & Hays, J. P. (2020). Comparison of Illumina versus Nanopore 16S rRNA Gene Sequencing of the Human Nasal Microbiota. *Genes*, *11*(9), 1105. https://doi.org/10.3390/genes11091105

Introduction to the microbiome R package. (n.d.). Retrieved March 18, 2023, from https://microbiome.github.io/tutorials/

Introduction to the Statistical Analysis of Microbiome Data in R | Academic. (n.d.). Retrieved March 18, 2023, from https://www.nicholas-ollberding.com/post/introduction-to-the-statistical-analysis-of-microbiome-data-in-r/

Jung, J. H., Kim, G., Byun, M. S., Lee, J. H., Yi, D., Park, H., & Lee, D. Y. (2022). Gut microbiome alterations in preclinical Alzheimer's disease. *PLOS ONE*, *17*(11), e0278276. https://doi.org/10.1371/journal.pone.0278276

King, C. H., Desai, H., Sylvetsky, A. C., LoTempio, J., Ayanyan, S., Carrie, J., Crandall, K. A., Fochtman, B. C., Gasparyan, L., Gulzar, N., Howell, P., Issa, N., Krampis, K., Mishra, L., Morizono, H., Pisegna, J. R., Rao, S., Ren, Y., Simonyan, V., … Mazumder, R. (2019). Baseline human gut microbiota profile in healthy people and standard reporting template. *PLOS ONE*, *14*(9), e0206484. https://doi.org/10.1371/JOURNAL.PONE.0206484

Korotky, N., & Peslyak, M. (2020). Blood Metagenome in Health and Psoriasis. *Frontiers in Medicine*, *7*. https://doi.org/10.3389/fmed.2020.00333

Lai, J., Li, A., Jiang, J., Yuan, X., Zhang, P., Xi, C., Wu, L., Wang, Z., Chen, J., Lu, J., Lu, S., Mou, T., Zhou, H., Wang, D., Huang, M., Dong, F., Li, M. D., Xu, Y., Song, X., & Hu, S. (2022). Metagenomic analysis reveals gut bacterial signatures for diagnosis and treatment outcome prediction in bipolar depression. *Psychiatry Research*, *307*, 114326. https://doi.org/10.1016/j.psychres.2021.114326

Leggett, R. M., & Clark, M. D. (2017). A world of opportunities with nanopore sequencing. *Journal of Experimental Botany*, *68*(20), 5419–5429. https://doi.org/10.1093/JXB/ERX289

Liu, Y., Rosikiewicz, W., Pan, Z., Jillette, N., Wang, P., Taghbalout, A., Foox, J., Mason, C., Carroll, M., Cheng, A., & Li, S. (2021). DNA methylation-calling tools for Oxford Nanopore sequencing: a survey and human epigenome-

wide evaluation. *Genome Biology*, *22*(1), 295. https://doi.org/10.1186/s13059-021-02510-z

Lloyd-Price, J., Abu-Ali, G., & Huttenhower, C. (2016). The healthy human microbiome. *Genome Medicine 2016 8:1*, *8*(1), 1–11. https://doi.org/10.1186/S13073-016-0307-Y

Loftus, E. V. (2004). Clinical epidemiology of inflammatory bowel disease: incidence, prevalence, and environmental influences. *Gastroenterology*, *126*(6), 1504–1517. https://doi.org/10.1053/J.GASTRO.2004.01.063

Luz Calle, M. (2019). Statistical Analysis of Metagenomics Data. *Genomics & Informatics*, *17*(1). https://doi.org/10.5808/GI.2019.17.1.E6

Ma, X., Stachler, E., & Bibby, K. (2017). Evaluation of Oxford Nanopore MinIONTM Sequencing for 16S rRNA Microbiome Characterization. *BioRxiv*, 099960. https://doi.org/10.1101/099960

Ma, Y., Zhang, Y., Xiang, J., Xiang, S., Zhao, Y., Xiao, M., Du, F., Ji, H., Kaboli, P. J., Wu, X., Li, M., Wen, Q., Shen, J., Yang, Z., Li, J., & Xiao, Z. (2021). Metagenome Analysis of Intestinal Bacteria in Healthy People, Patients With Inflammatory Bowel Disease and Colorectal Cancer. *Frontiers in Cellular and Infection Microbiology*, *11*, 48. https://doi.org/10.3389/FCIMB.2021.599734/BIBTEX

Magne, F., Gotteland, M., Gauthier, L., Zazueta, A., Pesoa, S., Navarrete, P., & Balamurugan, R. (2020). The Firmicutes/Bacteroidetes Ratio: A Relevant Marker of Gut Dysbiosis in Obese Patients? *Nutrients*, *12*(5), 1474. https://doi.org/10.3390/nu12051474

Maldonado-Arriaga, B., Sandoval-Jiménez, S., Rodríguez-Silverio, J., Lizeth Alcaráz-Estrada, S., Cortés-Espinosa, T., Pérez-Cabeza de Vaca, R., Licona-Cassani, C., Gámez-Valdez, J. S., Shaw, J., Mondragón-Terán, P., Hernández-Cortez, C., Suárez-Cuenca, J. A., & Castro-Escarpulli, G. (2021). Gut dysbiosis and clinical phases of pancolitis in patients with ulcerative colitis. *MicrobiologyOpen*, *10*(2). https://doi.org/10.1002/mbo3.1181

Mariat, D., Firmesse, O., Levenez, F., Guimarães, V., Sokol, H., Doré, J., Corthier, G., & Furet, J.-P. (2009). The Firmicutes/Bacteroidetes ratio of the human microbiota changes with age. *BMC Microbiology*, *9*(1), 123. https://doi.org/10.1186/1471-2180-9-123

Maslennikov, R., Ivashkin, V., Efremova, I., Alieva, A., Kashuh, E., Tsvetaeva, E., Poluektova, E., Shirokova, E., & Ivashkin, K. (2021). Gut dysbiosis is associated with poorer long-term prognosis in cirrhosis. *World Journal of Hepatology*, *13*(5), 557–570. https://doi.org/10.4254/wjh.v13.i5.557

Matsuo, Y., Komiya, S., Yasumizu, Y., Yasuoka, Y., Mizushima, K., Takagi, T., Kryukov, K., Fukuda, A., Morimoto, Y., Naito, Y., Okada, H., Bono, H., Nakagawa, S., & Hirota, K. (2021). Full-length 16S rRNA gene amplicon analysis of human gut microbiota using MinION™ nanopore sequencing confers species-level resolution. *BMC Microbiology*, *21*(1), 35. https://doi.org/10.1186/s12866-021-02094-5

Nagao-Kitamoto, H., Shreiner, A. B., Gillilland, M. G., Kitamoto, S., Ishii, C., Hirayama, A., Kuffa, P., El-Zaatari, M., Grasberger, H., Seekatz, A. M., Higgins, P. D. R., Young, V. B., Fukuda, S., Kao, J. Y., & Kamada, N. (2016). Functional Characterization of Inflammatory Bowel Disease–Associated Gut Dysbiosis in Gnotobiotic Mice. *Cellular and Molecular Gastroenterology and Hepatology*, *2*(4), 468–481. https://doi.org/10.1016/J.JCMGH.2016.02.003

Nagar, P., & Hasija, Y. (2018). Metagenomic approach in study and treatment of various skin diseases: a brief review. *Biomedical Dermatology 2018 2:1*, *2*(1), 1–8. https://doi.org/10.1186/S41702-018-0029-4

Nanopore DNA Sequencing. (n.d.). Retrieved June 17, 2023, from https://www.genome.gov/genetics-glossary/Nanopore-DNA-Sequencing

Nascimento, R. de P. do, Machado, A. P. da F., Galvez, J., Cazarin, C. B. B., & Maróstica Junior, M. R. (2020). Ulcerative colitis: Gut microbiota, immunopathogenesis and application of natural products in animal models. *Life Sciences*, *258*, 118129. https://doi.org/10.1016/j.lfs.2020.118129

Nelson, K. E., Weinstock, G. M., Highlander, S. K., Worley, K. C., Creasy, H. H., Wortman, J. R., Rusch, D. B., Mitreva, M., Sodergren, E., Chinwalla, A. T., Feldgarden, M., Gevers, D., Haas, B. J., Madupu, R., Ward, D. V., Birren, B. W., Gibbs, R. A., Methe, B., Petrosino, J. F., … Zhu, D. (2010). A catalog of reference genomes from the human microbiome. *Science (New York, N.Y.)*, *328*(5981), 994–999. https://doi.org/10.1126/SCIENCE.1183605

Paley, E. L. (2019). Discovery of Gut Bacteria Specific to Alzheimer's Associated Diseases is a Clue to Understanding Disease Etiology: Meta-Analysis of Population-Based Data on Human Gut Metagenomics and Metabolomics. *Journal of Alzheimer's Disease*, *72*(1), 319–355. https://doi.org/10.3233/JAD-190873

Park, C., Kim, S. B., Choi, S. H., & Kim, S. (2021). Comparison of 16S rRNA Gene Based Microbial Profiling Using Five Next-Generation Sequencers and Various Primers. *Frontiers in Microbiology*, *12*. https://doi.org/10.3389/fmicb.2021.715500

Pisani, A., Rausch, P., Bang, C., Ellul, S., Tabone, T., Cordina, C. M., Zahra, G., Franke, A., & Ellul, P. (2022). Dysbiosis in the Gut Microbiota in Patients with Inflammatory Bowel Disease during Remission. *Microbiology Spectrum*, *10*(3). https://doi.org/10.1128/SPECTRUM.00616-22

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., Mende, D. R., Li, J., Xu, J., Li, S., Li, D., Cao, J., Wang, B., Liang, H., Zheng, H., … Zoetendal, E. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, *464*(7285), 59–65. https://doi.org/10.1038/NATURE08821

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., & Glöckner, F. O. (2012). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*, *41*(D1), D590–D596. https://doi.org/10.1093/nar/gks1219

Rivera-Pinto, J., Egozcue, J. J., Pawlowsky-Glahn, V., Paredes, R., Noguera-Julian, M., & Calle, M. L. (2018a). Balances: a New Perspective for Microbiome Analysis. *MSystems*, *3*(4), e00053-18. https://doi.org/10.1128/mSystems.00053-18

Rivera-Pinto, J., Egozcue, J. J., Pawlowsky-Glahn, V., Paredes, R., Noguera-Julian, M., & Calle, M. L. (2018b). Balances: a New Perspective for Microbiome Analysis. *MSystems*, *3*(4), 53–71. https://doi.org/10.1128/MSYSTEMS.00053-18/SUPPL_FILE/SYS004182245SF8.PDF

Rodríguez-Pérez, H., Ciuffreda, L., & Flores, C. (2021). NanoCLUST: a species-level analysis of 16S rRNA nanopore sequencing data. *Bioinformatics*, *37*(11), 1600–1601. https://doi.org/10.1093/bioinformatics/btaa900

Rozas, M., Brillet, F., Callewaert, C., & Paetzold, B. (2022). MinION™ Nanopore Sequencing of Skin Microbiome 16S and 16S-23S rRNA Gene Amplicons. *Frontiers in Cellular and Infection Microbiology*, *11*. https://doi.org/10.3389/fcimb.2021.806476

Santos, A., van Aerle, R., Barrientos, L., & Martinez-Urtaza, J. (2020). Computational methods for 16S metabarcoding studies using Nanopore sequencing data. *Computational and Structural Biotechnology Journal*, *18*, 296–305. https://doi.org/10.1016/J.CSBJ.2020.01.005

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A., Oakley, B. B., Parks, D. H., Robinson, C. J., Sahl, J. W., Stres, B., Thallinger, G. G., Van Horn, D. J., & Weber, C. F. (2009). Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, *75*(23), 7537–7541. https://doi.org/10.1128/AEM.01541-09/ASSET/91BD47E1-E1DA-4980-B8B3-5DFA9C4F1FE7/ASSETS/GRAPHIC/ZAM0230904840002.JPEG

Serrano-Gómez, G., Mayorga, L., Oyarzun, I., Roca, J., Borruel, N., Casellas, F., Varela, E., Pozuelo, M., Machiels, K., Guarner, F., Vermeire, S., & Manichanh, C. (2021). Dysbiosis and relapse-related microbiome in inflammatory bowel disease: A shotgun metagenomic approach. *Computational and Structural Biotechnology Journal*, *19*, 6481–6489. https://doi.org/10.1016/j.csbj.2021.11.037

Shehata, E., Parker, A., Suzuki, T., Swann, J. R., Suez, J., Kroon, P. A., & Day-Walsh, P. (2022). Microbiomes in physiology: insights into 21st-century global medical challenges. *Experimental Physiology*, *107*(4), 257. https://doi.org/10.1113/EP090226

Shetty, S. A., & Lahti, L. (2019). Microbiome data science. *Journal of Biosciences*, *44*(5), 115. https://doi.org/10.1007/s12038-019-9930-2

Sun, P., Zhu, H., Li, X., Shi, W., Guo, Y., Du, X., Zhang, L., Su, L., & Qin, C. (2022). Comparative Metagenomics and Metabolomes Reveals Abnormal Metabolism Activity Is Associated with Gut Microbiota in Alzheimer's Disease Mice. *International Journal of Molecular Sciences*, *23*(19), 11560. https://doi.org/10.3390/ijms231911560

Taniya, M. A., Chung, H.-J., Al Mamun, A., Alam, S., Aziz, Md. A., Emon, N. U., Islam, Md. M., Hong, S.-T. shool, Podder, B. R., Ara Mimi, A., Aktar Suchi, S., & Xiao, J. (2022). Role of Gut Microbiome in Autism Spectrum Disorder and Its Therapeutic Regulation. *Frontiers in Cellular and Infection Microbiology*, *12*. https://doi.org/10.3389/fcimb.2022.915701

The Microbial World of Our Pets | GoldBio. (n.d.). Retrieved June 11, 2023, from https://goldbio.com/articles/article/Microbial-World-Our-Pets

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., & Gordon, J. I. (2007). The human microbiome project: exploring the microbial part of ourselves in a changing world. *Nature*, *449*(7164), 804. https://doi.org/10.1038/NATURE06244

Uhr, G. T., Dohnalová, L., & Thaiss, C. A. (2019). The Dimension of Time in Host-Microbiome Interactions. *MSystems*, *4*(1). https://doi.org/10.1128/MSYSTEMS.00216-18

Understanding Random Forest. How the Algorithm Works and Why it Is… | by Tony Yiu | Towards Data Science. (n.d.). Retrieved June 11, 2023, from

https://towardsdatascience.com/understanding-random-forest-58381e0602d2

Xu, X., Ocansey, D. K. W., Hang, S., Wang, B., Amoah, S., Yi, C., Zhang, X., Liu, L., & Mao, F. (2022). The gut metagenomics and metabolomics signature in patients with inflammatory bowel disease. *Gut Pathogens*, *14*(1), 1–18. https://doi.org/10.1186/S13099-022-00499-9/FIGURES/7

Zakerska-Banaszak, O., Tomczak, H., Gabryel, M., Baturo, A., Wolko, L., Michalak, M., Malinska, N., Mankowska-Wierzbicka, D., Eder, P., Dobrowolska, A., Slomski, R., & Skrzypczak-Zielinska, M. (2021). Dysbiosis of gut microbiota in Polish patients with ulcerative colitis: a pilot study. *Scientific Reports 2021 11:1*, *11*(1), 1–13. https://doi.org/10.1038/s41598-021-81628-3

Zhu, F., Ju, Y., Wang, W., Wang, Q., Guo, R., Ma, Q., Sun, Q., Fan, Y., Xie, Y., Yang, Z., Jie, Z., Zhao, B., Xiao, L., Yang, L., Zhang, T., Feng, J., Guo, L., He, X., Chen, Y., … Ma, X. (2020). Metagenome-wide association of gut microbiome features for schizophrenia. *Nature Communications*, *11*(1), 1612. https://doi.org/10.1038/s41467-020-15457-9

Zhu, S., Han, M., Liu, S., Fan, L., Shi, H., & Li, P. (2022). Composition and diverse differences of intestinal microbiota in ulcerative colitis patients. *Frontiers in Cellular and Infection Microbiology*, *12*. https://doi.org/10.3389/fcimb.2022.953962

Zuo, W., Wang, B., Bai, X., Luan, Y., Fan, Y., Michail, S., & Sun, F. (2022). 16S rRNA and metagenomic shotgun sequencing data revealed consistent patterns of gut microbiome signature in pediatric ulcerative colitis. *Scientific Reports 2022 12:1*, *12*(1), 1–13. https://doi.org/10.1038/s41598-022-07995-7

8. Annexos

Tots els fitxers, scripts i diversos resultats d'aquest treball es troben publicats al repositori https://github.com/miquelcastany/metagenomics_TFM.

Voldria agrair sincerament tota l'ajuda, idees, consells i expertesa de la directora d'aquest TFM, la Yolanda Guillén.

Gràcies també a la Malu Calle per reseoldre els dubtes sobre *selbal* i *coda4microbiome*.

I moltíssimes gràcies a la Mireia.

Miquel Castany Roma
20.06.2023