FACULTAT
DE CIÈNCIES
I TECNOLOGIA
UVIC | UVIC·UCC

Treball de Fi de Grau

Metodologies d'anàlisi del microbioma amb R

Guillem Vila Tubau

Grau en Biotecnologia

Tutora: Malu Calle Rosingana

Vic, juny de 2021

Agraïments

Vull agrair a la meva tutora Malu Calle Rosingana tota l'ajuda prestada al llarg del desenvolupament d'aquest treball, tant a nivell de suggeriments i resolució de problemes, com a nivell del seguiment constant del treball que ha realitzat.

Resum

El microbioma és una comunitat microbiana característica que ocupa un hàbitat ben definit, amb unes propietats fisico-químiques determinades i té un paper clau en la salut i/o el correcte funcionament del seu hoste. La recerca en el camp del microbioma, té en l'actualitat, i tindrà en un futur no molt llunyà, diverses aplicacions en el camp de la salut humana, en l'agricultura i el processament d'aliments i en la lluita contra el canvi climàtic. El microbioma s'estudia a través de la metagenòmica, principalment per seqüenciació shotgun i/o per seqüenciació d'amplicons. Les lectures resultants de la seqüenciació són processades i resumides en taules d'abundàncies d'OTU. Aquesta aproximació, però, requereix tenir en compte factors com la composicionalitat de les dades i les diferències entre les diverses metodologies d'anàlisi.

En aquest treball s'ha construït un procediment d'anàlisi del microbioma completament integrat en l'entorn d'R a través d'RStudio que inclou tant el processament bioinformàtic de les seqüències com l'anàlisi estadística posterior. S'han resumit les bases de dades de microbioma públiques més importants, i s'ha il·lustrat el procediment amb dades d'un estudi real. A més en el procés s'han avaluat els efectes de la rarefacció i el filtrat de certes mostres i OTUs/ASVs sobre els resultats obtinguts. El codi utilitzat està disponible en format tutorial amb Rmarkdown a (https://hanso3.github.io/TFG/Codi.html).

Aquest estudi em va permetre apreciar la complexitat computacional i metodològica de l'anàlisi del microbioma i la sensibilitat dels resultats a les diferents tècniques aplicades en el processament i anàlisi de les dades, tot posant de manifest la necessitat de disposar d'uns protocols estandarditzats per al processament i l'anàlisi del microbioma que facin que els resultats obtinguts siguin reproduïbles, significatius, i comparables amb els d'altres estudis.

Summary

A microbiome is a characteristic microbial community in a reasonably well-defined habitat with determined physiochemical properties, and it plays an important role in the health of the host. Microbiome research has and will have multiple applications in the field of human healthcare, agriculture, food processing and in the fight against climate change. The study of the microbiome is mostly done through Metagenomics, mainly with shotgun sequencing or/and amplicon sequencing. The reads resulting from sequencing are processed and summarised in abundance OTU tables. However, this approach requires taking into account factors such as the compositional nature of microbiome data and the differences between the various analysis methods.

In this project I constructed a procedure for the analysis of microbiome data completely integrated in R through RStudio which includes both the bioinformatics processing of the sequences and their subsequent statistical analysis. The main public microbiome databases have been summarised, and the procedure has been tested with data from a real study. Furthermore, the effects of rarefaction and filtering of certain samples and OTUs/ASVs on the results have been evaluated. The utilised code is available in Rmarkdown tutorial format on (https://hanso3.github.io/TFG/Codi.html).

This study made it possible for me to understand the computational and methodological complexity of microbiome data analysis and the sensitivity of results to the different processing and analysis techniques applied, highlighting the need for standardised microbiome processing and analysis protocols that result in reproducible, significative and comparable results.

Índex

1. Introducció 1
2. Objectius 3
3. Processament i anàlisi del microbioma 4
   3.1. Bases de dades públiques de microbioma 4
   3.2. Estudi de referència 5
   3.3. Processament de les seqüencies 5
   3.4. Anàlisi de dades de microbioma 8
      3.4.1. Diversitat alfa i beta 8
      3.4.2. Visualització de la diversitat beta 11
      3.4.3. Prova d'abundància diferencial 11
      3.4.4. Selecció de balanços 11
4. Resultats 12
   4.1. Resultats del processament 12
   4.2. Resultats de l'anàlisi 12
      4.2.1. Diversitat alfa 13
      4.2.2. Diversitat beta 16
      4.2.3. Prova d'abundàncies diferencials 22
      4.2.4. Representació dels fílums diferencialment abundants 28
      4.2.5. Selecció de balanços 29
      4.2.6. Comparació amb els resultats de l'article de referència 35
5. Conclusió 36
6. Referències 38

1. Introducció

El microbioma és una comunitat microbiana característica que ocupa un hàbitat ben definit i amb unes propietats fisico-químiques determinades. El terme no només es refereix als organismes pròpiament, sinó que també fa referència a l'hàbitat, cosa que resulta en la formació de nínxols ecològics específics (Whipps, Lewis & Cooke, 1988; Berg et al., 2020). El microbioma, que forma un micro-ecosistema dinàmic i interactiu susceptible a canviar en el temps i l'espai, és integrat en macro-ecosistemes, com podrien ser hostes eucariotes, i resulta ser crucial per a la seva salut i/o funcionament. La microbiota nucli o central és un conjunt de membres compartits entre consorcis microbians d'hàbitats similars (Berg et al., 2020).

El microbioma té un paper clau en la salut i/o el correcte funcionament del seu hoste. En el camp de la salut humana s'estan trobant associacions entre certes malalties i l'alteració del microbioma (NIH Team, 2019). Tractaments basats en el coneixement que es té sobre el microbioma, com el transplantament de microbiota fecal per tractar la infecció per *Clostridioides difficile*, ja són una realitat (van Nood et al., 2013). A part de les aplicacions potser més òbvies en el camp de la salut humana, el microbioma també jugarà un paper important en l'agricultura i el processament d'aliments (Singh & Trivedi, 2017) i en la lluita contra el canvi climàtic (Cavicchioli et al., 2019) en un futur no molt llunyà. Tot això i molt més fa que el seu estudi sigui d'interès públic i altament rellevant.

L'estudi del microbioma es fa majoritàriament mitjançant la metagenòmica, l'anàlisi de l'ADN provinent de mostres ambientals, un camp que va ser revolucionat amb el desenvolupament dels mètodes de seqüenciació NGS (Walshaw et al., 2011). Les dues aproximacions principals són la seqüenciació shotgun i la seqüenciació d'amplicons (Figura 1). La seqüenciació shotgun es basa en la seqüenciació de tot el genoma (tot l'ADN de la mostra), mentre que la seqüenciació d'amplicons es basa en la seqüenciació de gens marcadors amplificats per PCR, com el gen de l'ARNr 16S, i resulta molt més barata (Martín, Miquel, Langella & Bermúdez-Humarán, 2014). En la seqüenciació d'amplicons, les lectures obtingudes són processades bioinformàticament i resumides en taules d'abundàncies d'OTU (unitats taxonòmiques operatives) per a la seva posterior anàlisi estadística (Calle, 2019). La utilització de RStudio per fer aquesta anàlisi sembla ser molt més popular que la seva utilització per processar bioinformàticament les seqüències inicials, pas que es sol fer amb plataformes com QIIME 2™.

L'anàlisi del microbioma és complex ja que presenta una sèrie de limitacions que s'han de tenir en compte si es volen obtenir resultats sòlids. Les taules d'OTUs solen contenir un gran nombre de zeros. A més, el nombre màxim de lectures que es poden obtenir ve donat per la capacitat del seqüenciador i sol variar entre mostres, això vol dir que el nombre total de lectures no reflecteix el nombre total de microorganismes presents en la mostra, i per tant només aporta informació relativa, resultant en dependències entre les abundàncies dels diferents taxons d'una mostra. Això implica que les dades de microbioma són composicionals. La millor manera d'adreçar aquesta naturalesa composicional és treballant amb log-ratios d'Aitchison que cancel·len els biaixos composicionals. Donada una composició $(x_1, x_2, x_3, \dots, x_k)$, on cada valor x indica l'abundància de cadascun dels taxons, el log-ratio entre dues componentns, per exemple les dues primeres, es defineix com el logartime del quocient entre la primera i la segona

component, és a dir, $\log(\frac{x_1}{x_2})$. Un inconvenient d'aquesta aproximació és que requereix la substitució dels múltiples zeros presents a les taules d'abundància (Calle, 2019; Weiss et al., 2017). Altres mètodes de normalització com la rarefacció (submostreig aleatori sense substitució) són polèmics per la pèrdula d'informació que comporten (McMurdie & Holmes, 2014). L'eliminació de les mostres i/o OTUs que no sumen un nombre mínim de lectures per mitigar la dispersió de les dades i reduir la seva complexitat a través d'un pas de filtrat previ a la anàlisi també és comuna (Cao et al., 2021).

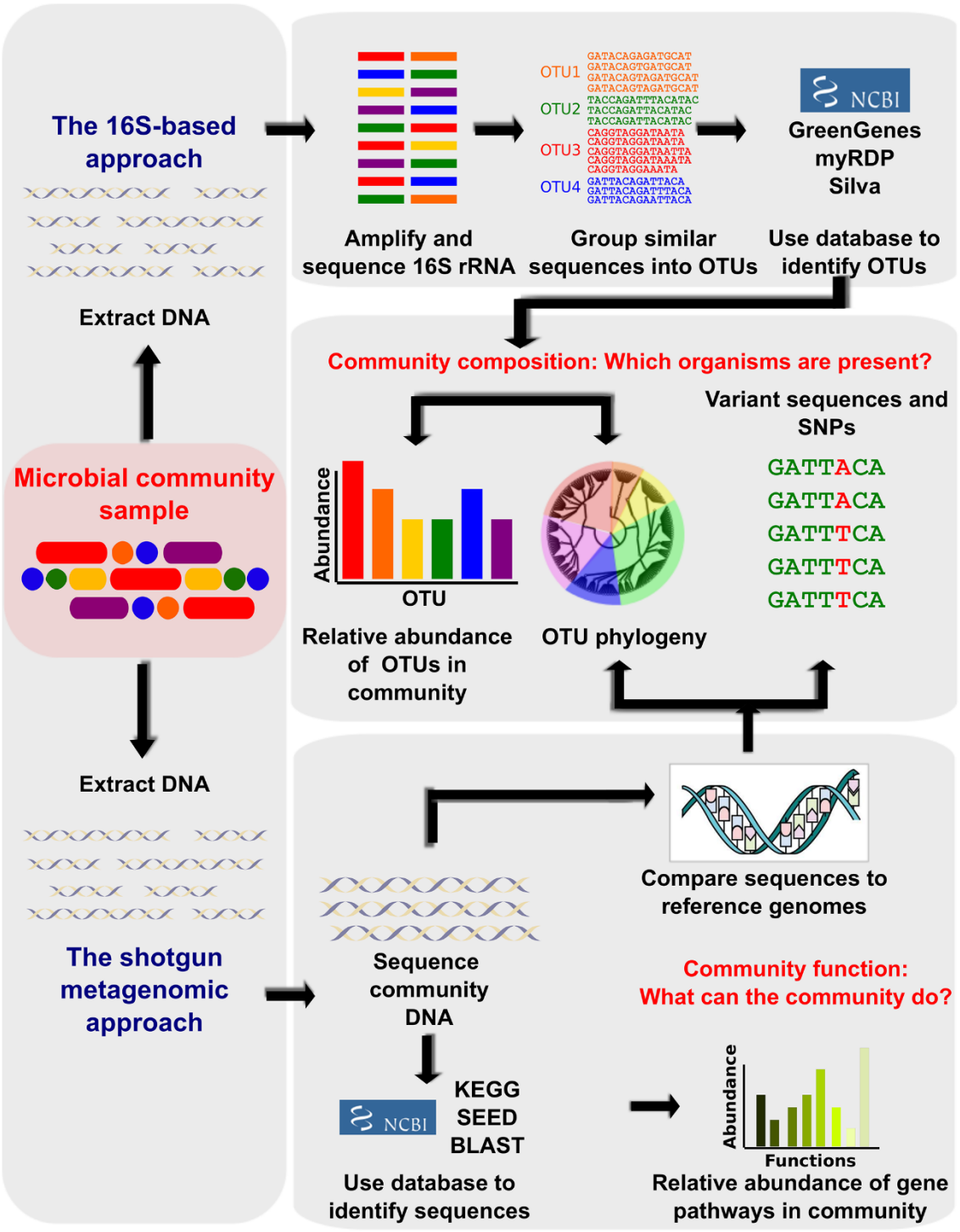

Figura 1. Resum visual del processament de les dades de microbioma obtingudes per seqüenciació shotgun i per seqüenciació d'amplicons. Crèdit a Morgan & Huttenhower (2012).

La intenció d'aquest treball és aprofundir en els mètodes de processament i anàlisi de dades de microbioma amb R i en les bases de dades públiques de microbioma. En el procés s'intentarà avaluar els efectes de la rarefacció i el filtrat sobre els resultats de l'anàlisi. Els efectes del filtrat sobre els resultats no han estat molt estudiats fins recentment (Cao et al., 2021). El codi en R utilitzat en aquest treball per al processament bioinformàtic i l'anàlisi estadística de les dades està disponible en format tutorial amb Rmarkdown a (https://hanso3.github.io/TFG/Codi.html).

2. Objectius

En aquest treball es vol construir un procediment d'anàlisi del microbioma completament integrat en l'entorn d'R a través d'RStudio que permeti tant processar bioinformàticament les lectures resultants de la seqüenciació d'amplicons com analitzar-les estadísticament un cop processades, permetent així la realització de tots els passos post-seqüenciació utilitzant una sola eina. Aquest procediment s'il·lustrarà amb unes dades d'un estudi real obtingudes d'una de les bases de dades públiques de microbioma. En el procés s'intentarà avaluar els efectes de la rarefacció i el filtrat sobre els resultats de l'anàlisi, i observar si divergeixen molt dels resultats obtinguts en l'estudi de referència de les dades. Així, en resum, es vol:

- Fer una recerca bibliogràfica sobre els mètodes de processament i anàlisi que permeti la construcció del procediment amb R i la seva posterior presentació en format tutorial a través d'Rmarkdown, a poder ser utilitzant mètodes d'anàlisi que tinguin en compte la naturalesa composicional de les dades de microbioma.
- Localitzar i investigar les principals bases de dades públiques de microbioma, tant a nivell de funcionament com a nivell de contingut, i en el procés obtenir unes dades crues per posar a prova el funcionament del procediment.
- Avaluar els efectes de la rarefacció i el filtrat sobre els resultats de les anàlisis.
- Comparar els resultats obtinguts amb els publicats en l'estudi de referència.

Amb aquests objectius s'intentarà donar resposta a les següents preguntes:

- Quines són les principals bases de dades públiques de microbioma, com funcionen i quin tipus d'informació contenen?
- Quins són els principals mètodes de processament i anàlisi de dades de microbioma? És possible processar i analitzar les dades només amb R? Quines són les seves principals limitacions?
- Com afecten la rarefacció i el filtrat als resultats de l'anàlisi?
- Canvien molt els resultats obtinguts respecte els resultats publicats en l'estudi de referència?

3. Processament i anàlisi del microbioma

3.1. Bases de dades públiques de microbioma

La primera part del treball va ser la cerca d'unes dades apropiades per posar a prova el procediment de processament i anàlisi amb R que prepararia posteriorment. Per fer-ho, en primer lloc, vaig buscar les principals bases de dades de microbioma públiques i vaig analitzar les seves característiques en profunditat, les principals bases de dades són resumides a continuació. En segon lloc vaig buscar unes dades apropiades per posar a prova el procediment, i les vaig descarregar a través de la base de dades amb un funcionament més senzill.

Base	**Tipus de dades***	**Emmagatzemat dades**	**Descàrrega múltiple**	**Disponibilitat metadades**	**Observacions**	**URL**
MDB: Microbiome Database	Crues	No, només enllaç a altres bases	No	No	Moltes pàgines donaven error, només conté enllaços a les dades	https://db.cngb.org/microbiome/
Human Microbiome Project Data Portal	Crues	Sí	Requereix programa	No	Dificultat per a la descàrrega de múltiples mostres	https://portal.hmpdacc.org/
MicrobiomeDB	Processades	Sí	Taula d'OTUs	Sí	Format taula d'OTUs estrany	https://microbiomedb.org/mbio/app/
GMrepo	Processades	Sí	Taula d'OTUs	Sí	Problemes amb carregar algunes pàgines	https://gmrepo.humangut.info/home
MGnify	Processades	Sí	Taula d'OTUs	No	Millor opció per dades processades	https://www.ebi.ac.uk/metagenomics/
NCBI	Crues i Processades	Sí	Requereix registre	Sí	Funcionament poc clar	https://www.ncbi.nlm.nih.gov/biosample/
ENA	Crues	Sí	Sí, fàcil i clara	No	Dificultat per obtenir les metadades	https://www.ebi.ac.uk/ena/browser/home

*Taula 1. Taula resum de les principals bases de dades públiques de microbioma i les seves característiques més importants. *Tipus de Dades: "crues" quan estan disponibles els fitxers fastq o similar resultats de la seqüenciació; "processades" quan només estan disponibles les taules d'abundància d'OTUs.*

3.2. Estudi de referència

Les dades que finalment vaig seleccionar van ser les del projecte PRJEB13092 "Reduced diversity and altered composition of the gut microbiome in individuals with myalgic encephalomyelitis/chronic fatigue syndrome" (Giloteaux et al., 2016), originalment trobades a la base de dades MGnify en la seva forma processada i finalment descarregades a través de l'ENA en la seva forma crua, per haver resultat ser la base de dades amb més facilitats per a la descàrrega de totes les mostres sense processar d'un estudi. Les metadades van ser descarregades directament d'un enllaç referenciat a l'article associat al projecte ja que no estaven disponibles a l'ENA. Les dades em van semblar interessants pel seu nombre de mostres (87), que era un compromís sà entre grans estudis que contenien milers de mostres, i altres que en contenien una dotzena o menys. A més d'això les dades contenien una mitjana de lectures per mostra molt més gran que la de la resta de dades que vaig considerar.

L'objectiu d'aquest estudi era avaluar si hi havia una associació entre la síndrome de la fatiga crònica i l'alteració del microbioma intestinal. Les dades contenien les mostres fecals de 48 pacients amb síndrome de fatiga crònica i 39 controls sans, seqüenciades per seqüenciació d'amplicons del gen que codifica per l'ARNr 16S amb la plataforma Illumina MiSeq 2x250 bp. A part d'això, a les metadades hi havia els nivells en sang de proteïna C-reactiva (CRP), proteïna d'unió als àcids grassos intestinals (I-FABP), lipopolisacàrids (LPS), proteïna d'unió als lipopolisacàrids (LBP), i CD14 soluble (sCD14) de cada individu.

A l'article de referència, pel que fa a l'anàlisi del microbioma, van estudiar si hi havia diferències significatives en la diversitat alfa de controls i pacients amb les mesures Shannon, Chao1 i PD utilitzant una prova de Wilcoxon-Mann-Whitney. També van estudiar la seva diversitat beta mesurant les distàncies UniFrac no ponderades i UniFrac ponderades i mirant si hi havia agregació entre controls i pacients representant-les amb una anàlisi de les components principals (PCA). Finalment van avaluar si hi havia OTUs que fossin significativament diferencialment abundants entre controls i pacients mitjançant una prova U de Wilcoxon-Mann-Whitney.

Van observar una diversitat alfa significativament reduïda en pacients respecte controls, no van observar agregació de les dades representant les distàncies UniFrac no ponderades i UniFrac ponderades, i van observar diferències significatives en l'abundància de diverses OTUs entre pacients i controls. Van concloure que hi havia una associació entre la síndrome de la fatiga crònica i una disbiosis del microbioma intestinal.

3.3. Processament de les seqüencies

Per poder construir un procediment de processament de les lectures obtingudes amb la seqüenciació utilitzant R vaig fer una recerca bibliogràfica extensiva sobre els diferents mètodes de processament, tant a nivell teòric, com a nivell de com implementar-los en R. Com he explicat breument a la introducció les lectures resultats de la seqüenciació són processades i resumides en taules d'unitats taxonòmiques operatives (OTU) per a la seva

posterior anàlisi estadística. Aquestes taules solen contenir les diferents mostres en les files i les diferents OTUs (grups de seqüencies llegides amb un percentatge donat d'identitat de seqüencia, normalment per sobre del 97%) en les columnes, amb el nombre de lectures de cada OTU per cada mostra (nombre de cops que s'han llegit les seqüencies que formen part de la OTU per cada mostra). Aquetes OTUs es solen classificar taxonòmicament mitjançant machine learning o a través d'un alineament amb seqüencies conegudes del gen amplificat procedents de diverses bases de dades com Greengenes o SILVA (Galloway-Peña & Hanson, 2020).

El primer pas en la construcció d'una taula d'OTUs a partir dels fitxers resultants de la seqüenciació és retallar les lectures perquè tinguin una llargada consistent i filtrar les de baixa qualitat. És recomanable fer una inspecció de la qualitat de les dades per ajustar els paràmetres de retall i filtrat a les característiques particulars de les lectures. Per exemple les lectures obtingudes mitjançant Illumina tendeixen a patir un descens de la qualitat cap al final de les lectures (Callahan et al., 2017).

Normalment després del filtrat s'agrupen les lectures de la seqüenciació en OTUs, que són conjunts de lectures que divergeixen entre elles per sota d'un llindar de dissimilaritat marcat de manera arbitrària, procés anomenat clustering. Una altra opció és fer servir el DADA2, un mètode d'alta resolució per deduir les variants de seqüència ribosòmica de manera exacta, tot adaptant el clustering a les particularitats de cada conjunt de dades i evitant la utilització de llindars arbitraris. El mètode DADA2 fa servir un model parametritzat de substitució d'errors per distingir errors de seqüenciació de variació biològica. Per tenir en compte la variació en les taxes d'error entre les diferents mostres, els paràmetres del model es dedueixen fent servir aprenentatge no supervisat alternant la inferència de la mostra amb l'estimació del paràmetre fins que els dos són consistents. La inferència es pot fer independentment per cada mostra o amb lectures aleatòries de totes les mostres. Els passos per agrupar les seqüències provinents de seqüenciació paired-end amb el paquet DADA2 són la desreplicació (ajuntar les seqüencies idèntiques), l'estimació de les taxes d'error, la deducció de les variants de seqüència d'amplicó amb la funció dada, la unió de les lectures forward i reverse amb els errors corregits tot eliminant les lectures que tenen bases diferents en les zones de sobreposició, i finalment la construcció d'una taula amb el nombre de lectures per cada amplicon sequence variant (ASV) en comptes de OTUs (Callahan, n.d.; Callahan et al., 2017).

Una vegada s'ha aconseguit la taula de seqüencies s'eliminen les seqüències quimèriques (que són el resultat de la unió de dues o més seqüències durant el procés d'amplificació per PCR que es fa abans de la seqüenciació). A continuació s'assigna una taxonomia a les seqüències restants. Amb el paquet DADA2 l'assignació de la taxonomia es fa mitjançant un classificador Bayesià que compara les diferents seqüències amb les ja classificades d'un training set. Hi ha diversos training sets disponibles actualitzats regularment. El resultat d'aquets procés és la obtenció d'una taula de taxonomia que conté la taxonomia que s'ha assignat a cada ASV. Així hi ha una taula amb el nombre de cops que apareix cada ASV en cada mostra, i una taula que conté l'assignació taxonòmica de cada ASV (Callahan et al., 2017).

Finalment queda l'opció de construir un arbre filogenètic per veure la relació entre les diferents seqüències i ajuntar tots els components en un objecte phyloseq. La construcció

de l'arbre filogenètic es fa des de zero, primer amb un alineament múltiple fent servir el paquet DECIPHER, i després amb la construcció de l'arbre pròpiament mitjançant el paquet phangorn. La construcció de l'objecte phyloseq es fa ajuntat les metadades, la taula de seqüències, la taula de taxonomia i l'arbre filogenètic en un objecte phyloseq utilitzant el paquet phyloseq (McMurdie & Holmes, 2013). La construcció de l'objecte phyloseq resulta molt útil a l'hora d'exportar i importar les dades des de RStudio, facilitant, per exemple, el treball amb més d'un ordinador o la compartició de les dades.

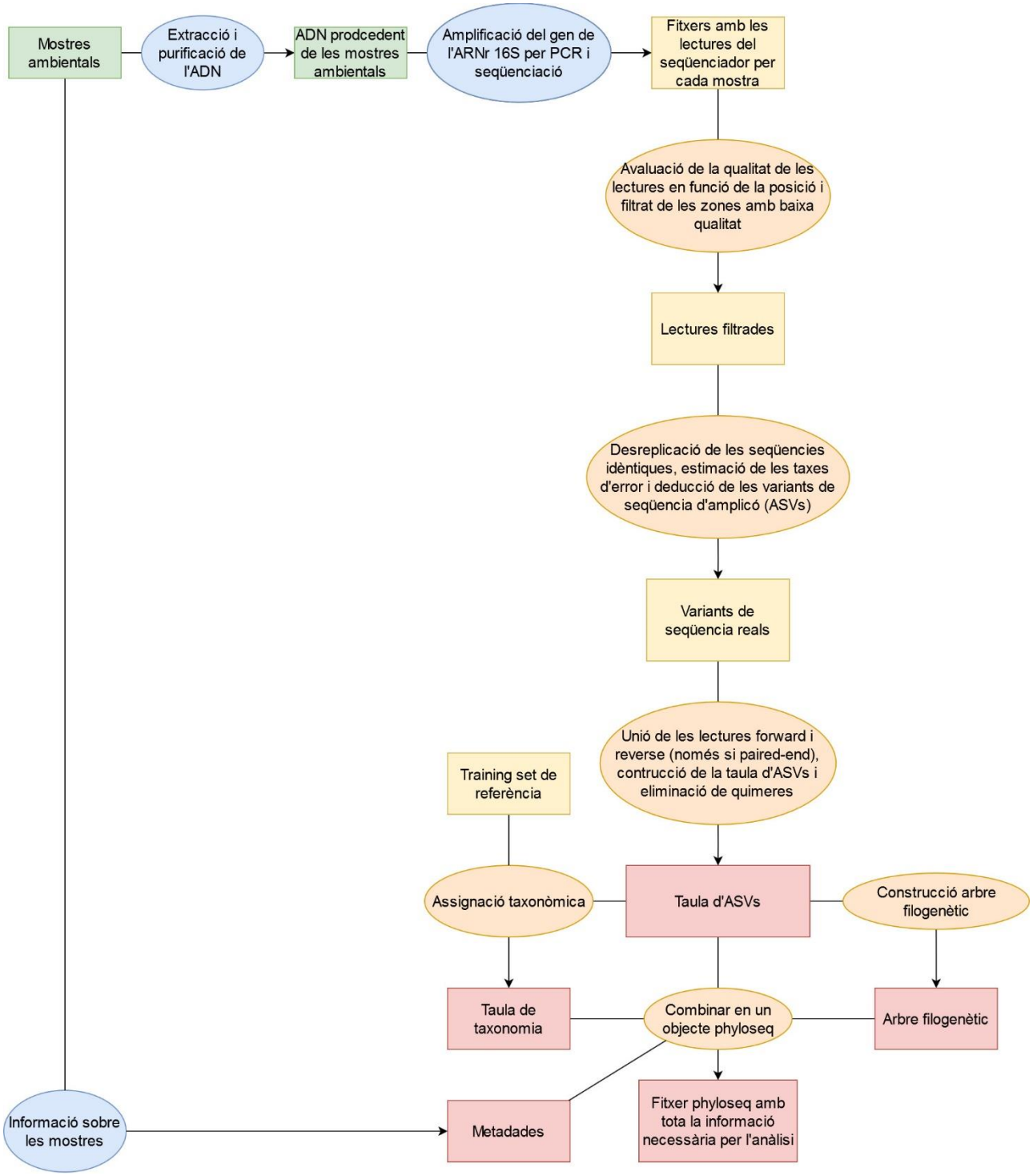

Figura 2. Esquema del processament de les seqüencies amb el paquet DADA2.

3.4. Anàlisi de dades de microbioma

Per poder construir el procediment d'anàlisi de les dades amb R vaig fer una recerca bibliogràfica extensiva sobre els diferents mètodes d'anàlisi de dades de microbioma, tant a nivell teòric, com a nivell de com implementar-ho en R. A continuació presento els principals mètodes d'anàlisi de dades de microbioma que vaig identificar al llarg de la recerca bibliogràfica i la seva implementació en R. Per als mètodes d'anàlisi de la diversitat alfa i beta he utilitzat unes taules resum. La resta de mètodes són descrits breument.

3.4.1. Diversitat alfa i beta

La diversitat alfa és una mesura unidimensional que resumeix la diversitat d'una comunitat ecològica, en altres paraules i en el context de l'anàlisi de microbioma, és una mesura de la diversitat dins d'una mostra. Hi ha diverses maneres de mesurar la diversitat alfa, la més simple, la riquesa, només té en compte el nombre de OTUs (o ASV) en la mostra. Altres més complexes també tenen en compte l'abundància d'aquestes OTUs per ajustar la seva contribució (Willis, 2019; Kim et al., 2017). Per exemple una mostra amb 4 OTUs amb un 25% d'abundància cadascuna seria més diversa que si una OTU representa el 95% de la composició de la mostra i les tres restants només sumen el 5%.

El problema és que els mètodes per mesurar la diversitat alfa van ser desenvolupats per estudis d'ecologia tradicionals, i a vegades el seu significat no és clar quan són aplicats a estudis de seqüenciació d'amplicons mitjançant NGS, ja que aquests són altament variables i no són del tot fiables per estudiar els microorganismes més minoritaris i/o desconeguts (Edgar, n.d.). Les característiques de les principals mesures de la diversitat alfa són resumides a la taula 2 a continuació.

Si la diversitat alfa és una mesura de la diversitat dins d'una mostra, la diversitat beta és una mesura de la diversitat entre mostres. Per ser més precisos és una mesura de la distància entre mostres, de fins a quin punt dues mostres són semblants o diferents a nivell d'organismes (en el cas de l'anàlisi de dades de microbioma OTUs o ASVs) que les componen (Stephenson, 2015; Scholz, n.d.). Les característiques de les principals mesures de la diversitat beta són resumides a la taula 3 a continuació.

Mesures	**Càlcul**	**Informació**	**Implementació R**	**Prova d'hipòtesi utilitzada**
Riquesa observada	Nombre d'espècies (en el context de l'anàlisi de dades de microbioma nombre de taxons o OTUs)	No té en compte el pes de cada OTU, si una domina molt no es veurà.	Funció "estimate_richness" de la llibreria "phyloseq"	Prova de suma de rangs de Wilcoxon a través de "wilcox.test"
Chao1	$C = \frac{N + S^2}{2D}$ On N és el nombre d'OTUs, S és el nombre d'OTUs amb només una lectura, i D és el nombre d'OTUs amb només dues lectures.	Intenta extreure informació sobre les taxonomies menys abundants a partir dels singletons i doubletons		
Shannon	$H' = -\sum_{i=1}^{R} p_i \ln p_i$ On p_i és la proporció d'individus que pertanyen a una espècie i.	Té en compte nombre d'OTUs i la seva proporció. Com més igualment distribuïdes estiguin les OTUs més gran serà l'índex.		
Simpson	$\lambda = \sum_{i=1}^{R} {p_i}^2$ On p_i és la proporció d'individus que pertanyen a una espècie i.	Té en compte nombre d'OTUs i la seva proporció. Si s'acosta a 1 (max) indica que hi ha una OTU molt dominant i que la resta són poc comunes.		
PD	Suma de la llargada de les branques de l'arbre filogenètic.	Valors més grans indiquen diversitats més grans.	Funció "pd" de la llibreria "picante"	

Taula 2. Taula resum amb les mesures més comunes de la diversitat alfa i la seva implementació en el codi R d'anàlisi amb la prova d'hipòtesi estadística utilitzada per avaluar si les diferències entre els resultats de controls i pacients són significatives.

Distància	**Càlcul**	**Informació**	**Implementació R**	**Prova d'hipòtesi utilitzada**
Bray-Curtis	$BC_{ij} = 1 - \frac{2C_{ij}}{S_i + S_j}$ Donades una mostra i i una mostra j, C_{ij} és la suma del nombre de lectures menor per cada OTU que tenen en comú. S_i i S_j són el nombre de lectures total de cada mostra. Si treballem amb proporcions la suma de S_i i S_j sempre donarà 2 i per tant ens quedarà que l'índex és 1- C_{ij}.	En cas de restar a 1 el resultat de la divisió, com més s'acosti a 1 l'índex més desiguals són les mostres i com més s'acosti a 0 més iguals són. Si no es fa la resta la interpretació del resultat és la inversa: proximitat a 1 indica similaritat i proximitat a 0 indica dissimilaritat.	Funció "distance" de la llibreria "phyloseq"	PERMANOVA a través de la funció "adonis" de la llibreria "vegan"
UniFrac	Es calcula dividint la suma de la llargada de les branques de l'arbre filogenètic compartides entre les dues mostres per la suma de la llargada de totes les branques. La ponderada té en compte les abundàncies de cada OTU a l'hora de fer el càlcul.	Com més gran més similaritat. No té en compte les abundàncies de les OTUs.		
UniFrac Ponderada		Com més gran més similaritat. Té en compte les abundàncies de les OTUs.		
Jaccard	$J(A,B) = \frac{\lvert A \cap B \rvert}{\lvert A \cup B \rvert}$ Donades una mostra A i una mostra B, l'índex és el resultat de dividir el nombre d'OTUs presents en les dues mostres per la suma del nombre d'OTUs de A i de B menys el nombre d'OTUs que es troben en les dues mostres.	Com més gran sigui la intersecció entre A i B, més s'acostarà l'índex a 1, indicant similaritat entre les mostres. Com més petita sigui més s'acostarà l'índex a 0, indicant dissimilaritat entre les mostres. No té en compte les abundàncies de les OTUs.		

Taula 3. Taula resum amb les distàncies més comunes utilitzades per mesurar la diversitat beta i la seva implementació en el codi R d'anàlisi amb la prova d'hipòtesi estadística utilitzada per avaluar si les diferències entre els resultats de controls i pacients són significatives.

3.4.2. Visualització de la diversitat beta

Per la visualització de la diversitat beta vaig utilitzar l'escalament multidimensional (MDS), també anomenat anàlisi de coordenades principals, que és un mètode d'ordinació per poder visualitzar gràficament les distàncies entre els components de dues o més mostres amb l'objectiu d'identificar possibles agrupacions o tendències. Per fer aquesta representació s'han de mesurar les distàncies entre les diferents mostres. Es poden fer servir una gran varietat de distàncies: des de la distància euclidiana fins a la distància de Bray-Curtis o la UniFrac, resumides a la taula de l'apartat anterior. Llavors s'ha de reduir la seva dimensionalitat mitjançant tècniques d'ordinació, en aquest cas l'MDS, cosa que permet la seva representació en dues dimensions. La seva implementació en R és a través de la funció "ordinate" del paquet "phyloseq". A diferència de l'MDS mètric, el no mètric fa servir una regressió isotònica per estimar una transformació de les dissimilaritats (Wickelmaier, n.d.; Calle, 2019), l'MDS no mètric només apareix al tutorial Rmarkdown.

3.4.3. Prova d'abundància diferencial

L'objectiu de la prova d'abundància diferencial és identificar ASVs (o OTUs) que siguin diferencialment abundants de manera significativa entre dues mostres o grups de mostres. Canvis significatius en l'abundància de certs taxons s'han associat a diverses condicions com diarrea, obesitat, sida, etc. La seva utilització en el context de l'anàlisi de dades de microbioma mitjançant seqüenciació d'amplicons és controvertida per diversos problemes com la utilització de mètodes que no s'ajusten a la naturalesa composicional de les dades, la pobre representació de l'ambient original per part de les mostres, submostres i taules OTU, la dificultat en la normalització i la general falta de reproductibilitat dels experiments (Morton et al., 2019; Calle, 2019).

Vaig decidir utilitzar aldex, que és una funció d'R del paquet ALDEX2 que permet analitzar la presencia de OTUs diferencialment abundants entre dues o més condicions / grups de mostres. La funció treballa amb log-ratios, ajustant-se a la naturalesa composicional de les dades de microbioma. Una vegada executada, la funció d'una sola línia genera instàncies de Monte Carlo de la distribució de Dirichlet per a cada mostra, converteix cada instància mitjançant una transformació a log-ratios, i després retorna els resultats de les proves per dues mostres (t de Welch, Wilcoxon) o per múltiples mostres (glm, Kruskal-Wallice). Aquesta funció també mesura la mida de l'efecte per l'anàlisi de dues mostres (Gloor et al., 2021).

3.4.4. Selecció de balanços

Un altre mètode d'anàlisi que té en compte la naturalesa composicional de les dades és la identificació de grups de taxons anomenats signatures microbianes, l'abundància relativa entre les quals es pugui associar a un fenotip d'interès, utilitzant la funció selbal del paquet selbal. Els passos que el selbal segueix per seleccionar balanços són: la substitució de zeros per poder portar a terme la transformació logarítmica de les dades, la selecció de dos taxons de manera que el seu log-ratio estigui el màxim d'associat a la variable

resposta, i l'adició de nous taxons al balanç (al numerador o al denominador del log-ratio) de tal manera que el criteri especificat d'optimització (àrea sota ROC o MSE) sigui millorat mitjançant un procés de selecció avançada (forward stepwise regression) fins que no quedi cap taxó que millori el criteri especificat d'optimització. També s'atura el procés d'adició de taxons al balanç quan s'arriba al nombre màxim de components. Aquest nombre màxim és definit a través d'un procés de validació creuada que també s'utilitza per avaluar com de robust és el balanç (Rivera-Pinto et al., 2018; Calle, 2019).

A part del selbal, una altra eina per seleccionar balanços és el CoDaCoRe, implementat en R a través de la funció "codacore" del paquet codacore. El mètode utilitza la discretització, que facilita la interpretació dels resultats transformant les mitjanes geomètriques ponderades sobre totes les covariants en balanços reals sobre un nombre petit de covariants. El model es pot regularitzar a través del paràmetre λ. El model també es pot estendre per fer aprenentatge supervisat, funcions de regressió no linear, etc (Gordon-Rodriguez, Quinn & Cunningham, n.d.).

4. Resultats

4.1. Resultats del processament

Amb unes dades apropiades descarregades i la informació teòrica necessària sobre el processament de seqüències de microbioma, vaig procedir a aplicar el codi de processament bioinformàtic de les lectures descrit en el tutorial "Workflow for Microbiome Data Analysis: from raw reads to community analyses" (Callahan et al., 2017). La principal limitació durant el processament va ser la falta de capacitat computacional del meu ordinador, que només tenia 8 GB de RAM, quantitat que no era ni molt menys suficient i resultava en la congelació de l'RStudio quan provava d'executar el codi de processament amb les 87 mostres. Així per la construcció i prova del codi vaig utilitzar un petit subconjunt de mostres per veure que tot anés bé. Una vegada el procediment va estar a punt, vaig demanar a la meva tutora que l'executés amb totes les mostres i m'enviés l'objecte phyloseq obtingut. El codi va funcionar sense problemes i va resultar en un objecte phyloseq de 300 kB amb el qual vaig poder continuar amb el procediment d'anàlisi sense cap altre problema a nivell de limitació computacional.

El codi resultant de la construcció del procediment de processament i que va ser utilitzat per processar les seqüencies està disponible en format tutorial amb Rmarkdown a (https://hanso3.github.io/TFG/Codi.html).

4.2. Resultats de l'anàlisi

A continuació estan resumits els resultats d'aplicar el codi d'anàlisi estadística a les dades procedents de l'estudi "Reduced diversity and altered composition of the gut microbiome in individuals with myalgic encephalomyelitis/chronic fatigue syndrome" (Giloteaux et al., 2016) després del seu processament bioinformàtic. Per avaluar l'efecte de la rarefacció sobre els resultats vaig treballar amb dades on s'havien filtrat totes les mostres i ASVs

que no sumaven, com a mínim, 10, 500 i 1000 lectures i amb les dades sense filtrar. Per avaluar l'efecte del filtrat sobre els resultats vaig treballar amb dades sense filtrar, i amb dades on s'havien filtrat totes les mostres i ASVs que no sumaven, com a mínim, 10, 500, i 1000 lectures. Finalment vaig comparar els resultats obtinguts amb els de l'estudi de referència.

4.2.1. Diversitat alfa

Els mètodes d'anàlisi de la diversitat alfa que vaig implementar van ser, com ja he explicat, la riquesa observada, Chao1, Shannon, Simpson i phylogenetic diversity (PD). Per la primera execució vaig utilitzar directament les dades sense filtrar, com recomana la funció "estimate_richness".

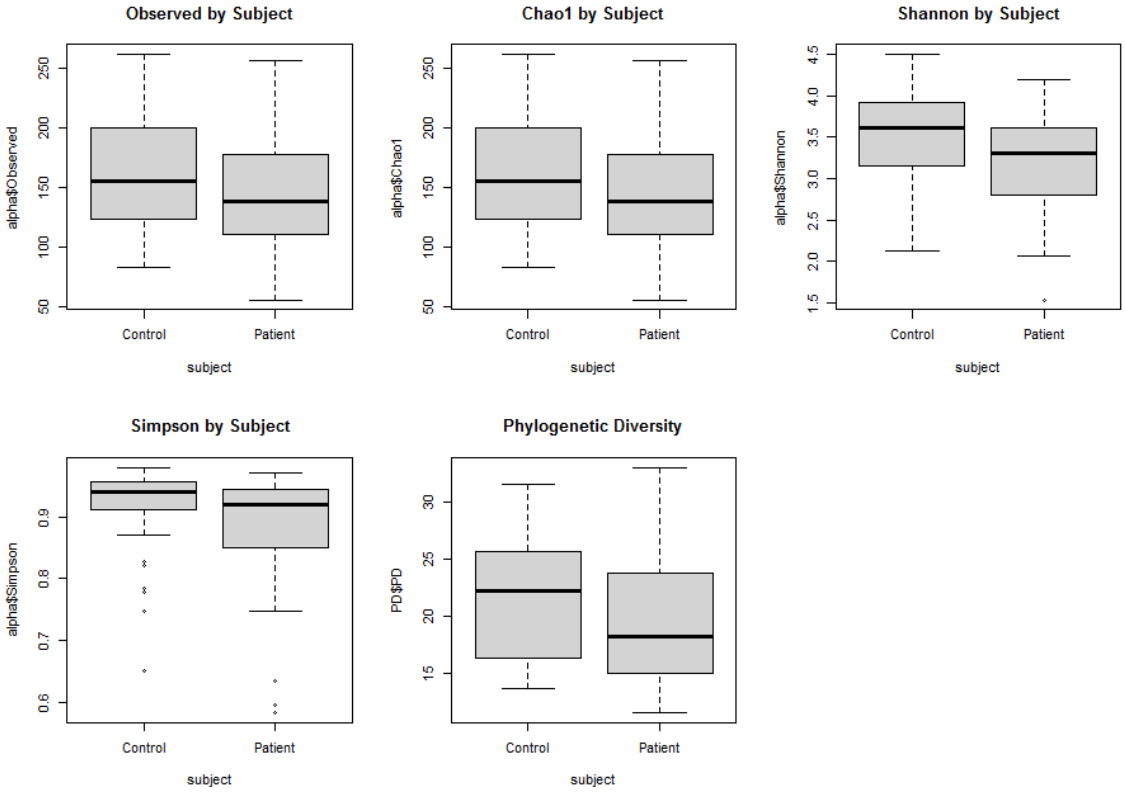

Figura 3. Diagrames de caixes dels valors de les diferents mesures de la diversitat amb pacients i controls separats (variable "Subject" de les metadades) utilitzant les dades sense filtrar. La diversitat tendeix a ser menor en els pacients (dreta).

S'observava una tendència a una diversitat reduïda en els pacients respecte els controls. Per avaluar si les diferències entre controls i pacients eren significatives vaig utilitzar una prova dels signes de Wilcoxon considerant significatius els p-valors per sota de 0,05. Sense filtrat, només resultaven significatives les diferències entre controls i pacients utilitzant els índex de Shannon i Simpson. Els valors de la riquesa observada i de l'índex de Chao1 semblaven coincidir pel baix nombre de singletons i doubletons (ASVs amb

només una i dues lectures respectivament) presents en aquestes dades (en absència d'aquests, el resultat de Chao1 coincideix amb el valor de la riquesa observada).

A continuació vaig voler observar l'efecte de la utilització de diferents paràmetres de filtrat sobre la significació de les diferències entre els dos grups. Vaig calcular les diferents mesures de diversitat i avaluar la significació de les diferencies entre els dos grups després d'haver filtrar les ASVs i mostres que no tenien un mínim de 10, 500 i 1000 lectures totals. Els resultats obtinguts estan resumits a la taula 4 juntament amb els p-valors obtinguts utilitzant les dades sense filtrar.

Mesura	**P-valor sense filtrat**	**P-valor Filtrat 10**	**P-valor Filtrat 500**	**P-valor Filtrat 1000**
Riquesa Observada	0,1216	0,128	0,0104	0,00575
Chao1	0,1216	0,128	0,0104	0,00575
Shannon	0,0128	0,0125	0,011	0,00968
Simpson	0,0209	0,0204	0,0185	0,0209
PD	0,0631	0,0583	0,0138	0,00704

Taula 4. Variació dels p-valors resultats d'aplicar una prova de Wilcoxon per avaluar si les diferències entre les diversitats alfa de pacients i controls eren estadísticament significatives en funció dels paràmetres de filtrat utilitzats.

Utilitzant 10 com a paràmetre de filtrat inicial la significació dels resultats no semblava variar de manera excessiva (± 0,0064 era la variació més gran observada), les mesures que resultaven en diferències significatives entre controls i pacients hi seguien resultant i les que no, seguien sense resultar-hi. Utilitzant 500 s'observava un canvi radical en els resultats: per totes les mesures hi havia diferències significatives entre els dos grups, amb el p-valor de les 3 primeres mesures sent 10 vegades més petit que amb paràmetre de filtrat 10. Augmentat el paràmetre de filtrat fins a 1000 hi havia un augment de la significació similar. Així, per aquestes dades, semblava haver-hi una clara relació entre la utilització de paràmetres de filtrat més grans i la obtenció de diferències més significatives entre els dos grups pel que fa a la diversitat alfa. Cal destacar que l'índex de Simpson va ser la mesura de la diversitat alfa menys afectada per la utilització de diferents paràmetres de filtrat (variació màxima de ±0,0024 amb les diferències entre els dos grups sempre sent significatives).

Per comprovar l'efecte de la utilització de dades rarificades sobre la diversitat alfa vaig repetir el procés utilitzant les dades rarificades amb rngseed = 123 sense filtrat, i amb els diferents paràmetres de filtrat utilitzats anteriorment.

Mesura	**P-valor sense filtrat**	**P-valor Filtrat 10**	**P-valor Filtrat 500**	**P-valor Filtrat 1000**
Riquesa Observada	0,1476	0,1525	0,0113	0,0061
Chao1	0,1501	0,1393	0,0094	0,0067
Shannon	0,0141	0,0145	0,0102	0,0096
Simpson	0,0194	0,0199	0,0153	0,0185
PD	0,071	0,0781	0,0094	0,0029

Taula 5. Variació dels p-valors resultats d'aplicar una prova de Wilcoxon per avaluar si les diferències entre les diversitats alfa de pacients i controls eren estadísticament significatives en funció dels paràmetres de filtrat utilitzats i fent servir les dades rarificades amb rngseed = 123.

Amb les dades rarificades, malgrat haver obtingut p-valors lleugerament diferents, els resultats no van canviar a nivell de significació de les diferències. La rarefacció no va mitigar gens l'efecte del filtrat. Les mesures més afectades per la rarefacció van ser l'índex de Simpson i PD.

4.2.2. Diversitat beta

A nivell de diversitat beta vaig utilitzar les distancies de Bray-Curtis, UniFrac, UniFrac ponderada i la de Jaccard. Com en el cas anterior vaig voler avaluar l'efecte de diferents paràmetres de filtrat inicial sobre els resultats. En primer lloc vaig mesurar les distàncies utilitzant les dades sense filtrar, i a continuació les dades amb les ASVs i mostres filtrades per sota de 10, 500 i 1000 lectures. En tots els casos llavors vaig fer la ordinació MDS i representar les distàncies per veure si les mostres de controls i pacients tendien a agrupar-se.

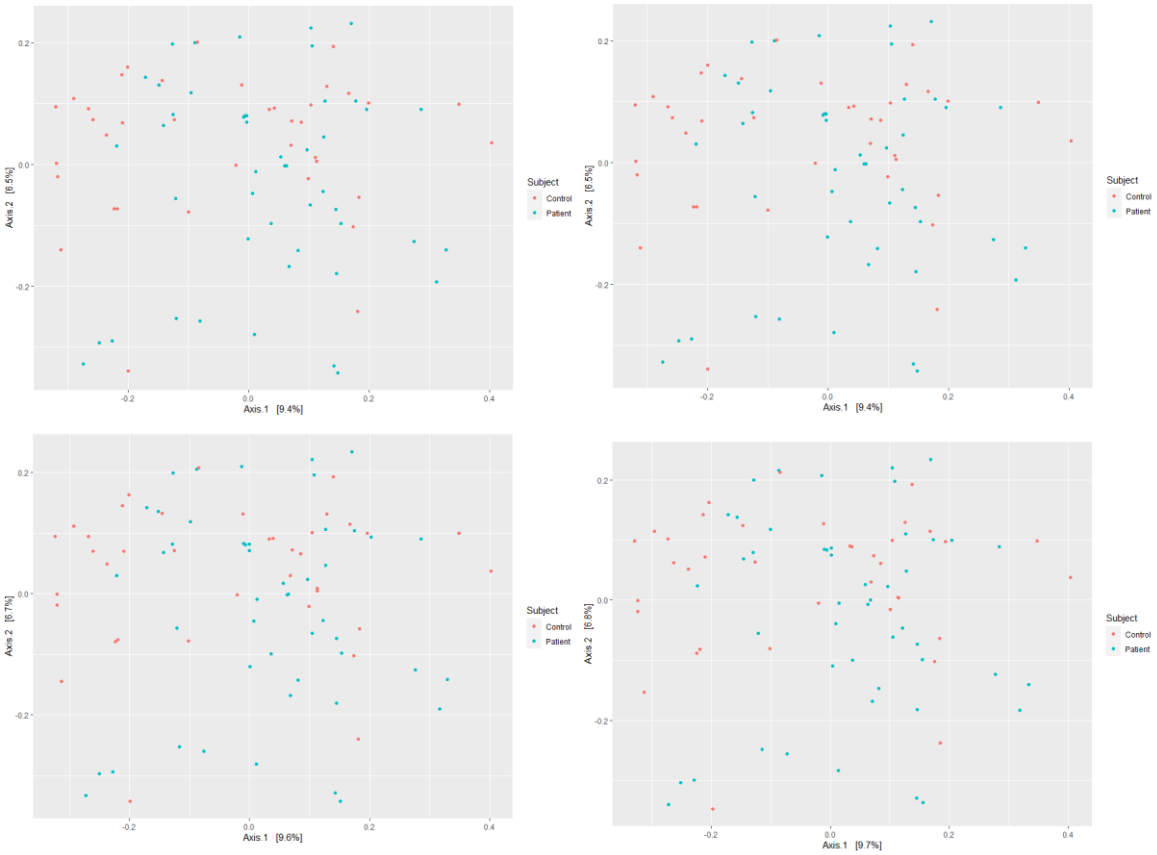

Figura 4. Representació mitjançant un MDS de les distàncies Bray-Curtis. A dalt a l'esquerra hi ha representades les distàncies de les mostres sense filtrar, a dalt a la dreta amb filtrat 10, a baix a l'esquerra amb filtrat 500 i a baix a la dreta amb filtrat 1000.

Amb les distàncies de Bray-Curtis no hi havia una agregació clara entre els dos grups. La utilització de diferents paràmetres de filtrat no semblava variar gaire el resultat. La variança explicada per la primera coordenada no passava en cap cas del 10% i la segona no passava del 7%. Així, com a mínim a nivell d'ordinació, no semblava que la utilització de diferents paràmetres de filtrat afectés massa el resultat.

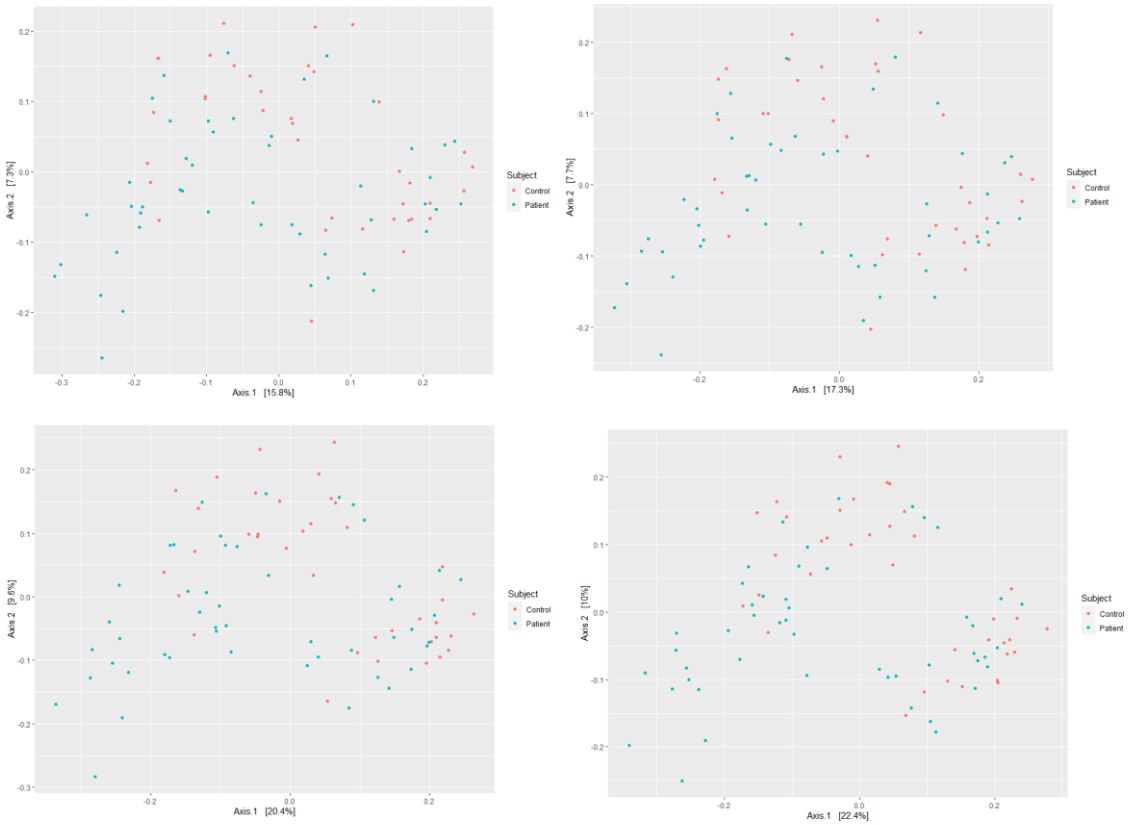

Figura 5. Representació mitjançant un MDS de les distàncies UniFrac no ponderades. A dalt a l'esquerra hi ha representades les distàncies de les mostres sense filtrar, a dalt a la dreta amb filtrat 10, a baix a l'esquerra amb filtrat 500 i a baix a la dreta amb filtrat 1000.

Amb les distàncies UniFrac no ponderades semblava haver-hi una certa separació entre els dos grups, especialment utilitzant els paràmetres de filtrat 500 i 1000. Les variàncies explicades per la primera i segona coordenades semblaven augmentar amb la utilització de paràmetres de filtrat més estrictes, passant del 15,8% de variança explicada per la primera coordenada utilitzant les dades sense filtrar a el 22,4% de la variança explicada per la primera coordenada utilitzant les dades amb les ASVs i mostres que no sumaven 1000 lectures filtrades. Així, semblava ser que per la distància UniFrac no ponderada la utilització de paràmetres de filtrat més estrictes resultava en una més clara agrupació de les dades i en explicacions de la variança més altes per part de les dues primeres coordenades.

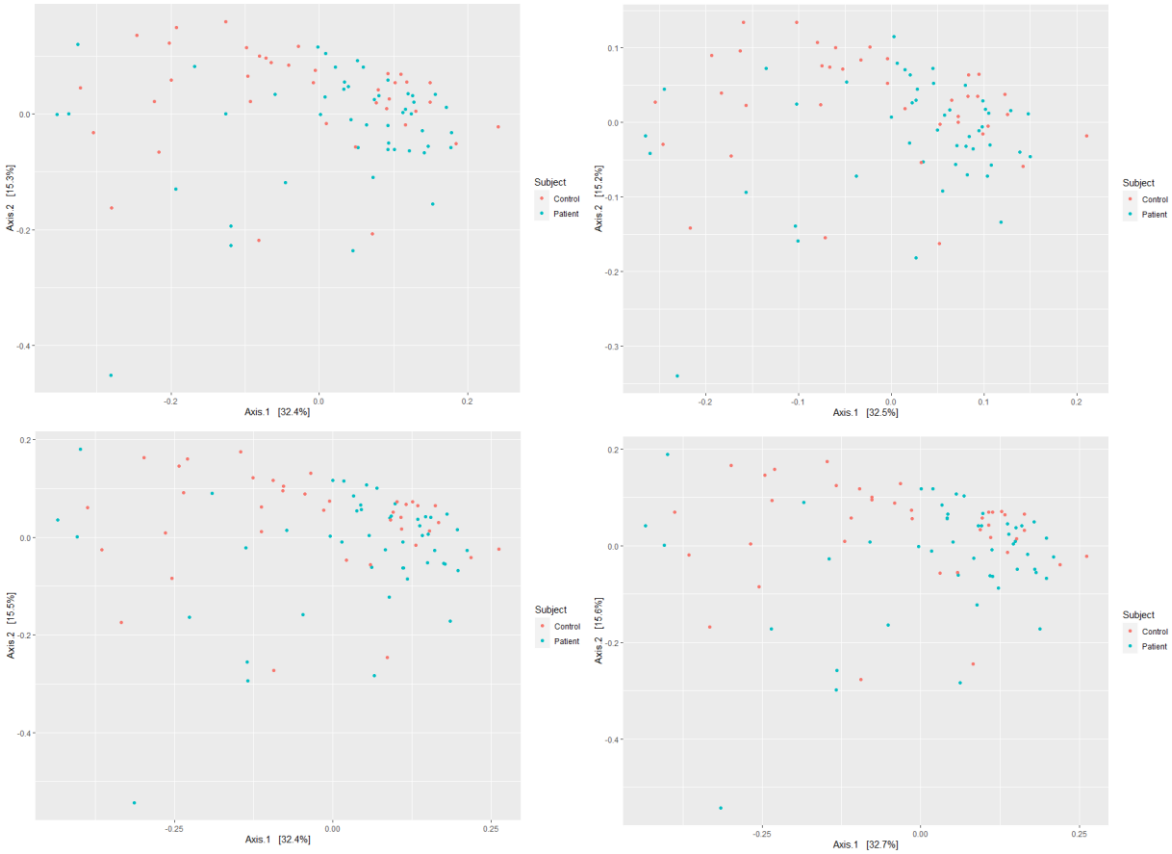

Figura 6. Representació mitjançant un MDS de les distàncies UniFrac ponderades. A dalt a l'esquerra hi ha representades les distàncies de les mostres sense filtrar, a dalt a la dreta amb filtrat 10, a baix a l'esquerra amb filtrat 500 i a baix a la dreta amb filtrat 1000.

Amb les distàncies UniFrac ponderades, malgrat que la primera coordenada expliqués com a mínim el 30% de la variança independentment del paràmetre de filtrat utilitzat, no hi havia, a simple vista, una agregació clara dels dos grups com la observada amb les distàncies UniFrac no ponderades. També a diferència de les dues distàncies anteriors, la variança explicada per les dues primeres coordenades pràcticament no variava en funció del paràmetre de filtrat utilitzat (variació màxima de 0,3% per la primera coordenada i de 0,4% per la segona). Així, per les distàncies UniFrac ponderades, no semblava que la utilització de diferents paràmetres de filtrat afectés massa el resultat.

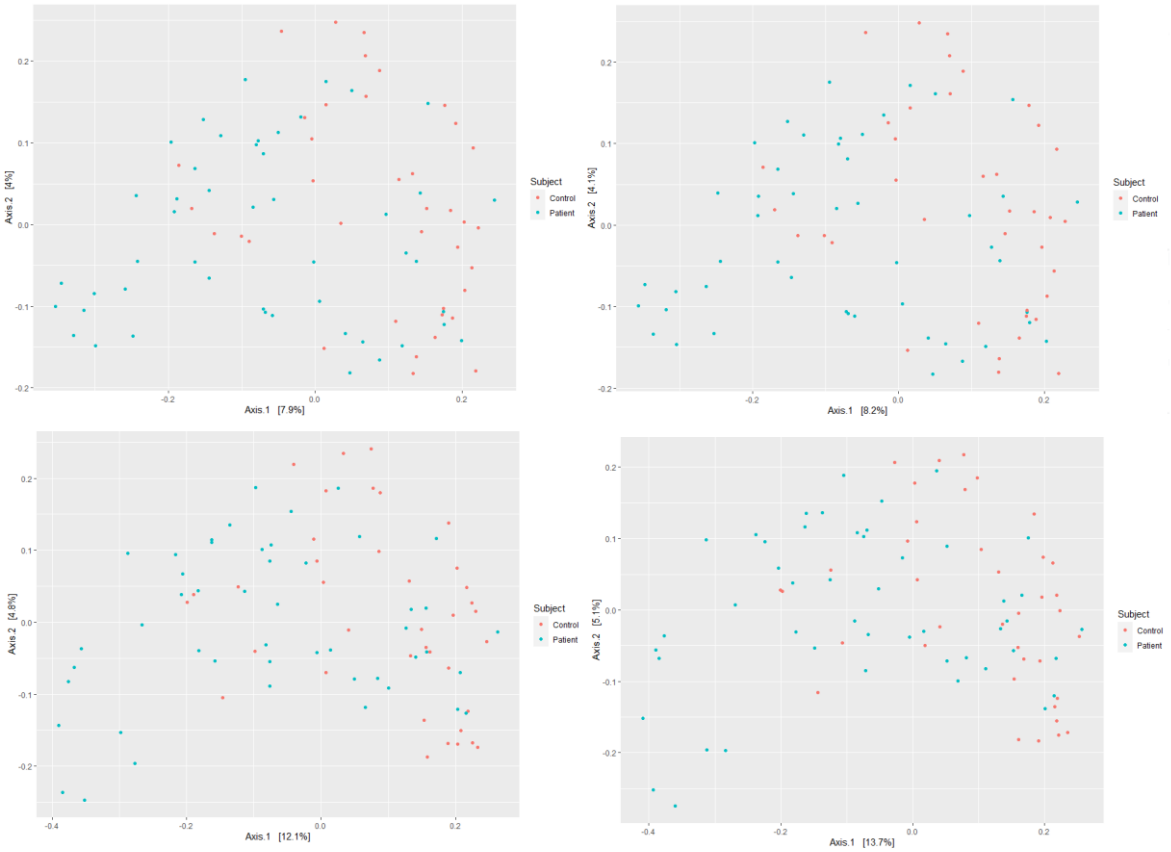

Figura 7. Representació mitjançant un MDS de les distàncies de Jaccard. A dalt a l'esquerra hi ha representades les distàncies de les mostres sense filtrar, a dalt a la dreta amb filtrat 10, a baix a l'esquerra amb filtrat 500 i a baix a la dreta amb filtrat 1000.

Finalment, amb les distàncies de Jaccard, s'observava una certa agrupació de les dades independentment del paràmetre de filtrat utilitzat. La variança explicada per les dues primeres coordenades semblava augmentar amb la utilització de paràmetres de filtrat més estrictes, passant d'un 7,9% de la variança explicada per la primera coordenada amb les dades sense filtrar, a un 13,7% utilitzant les dades amb les ASVs i mostres que no sumaven 1000 lectures filtrades. En aquest sentit, la distància de Jaccard sembla seguir la tendència també observada en la distàncies UniFrac no ponderades a un augment de la variança explicada per les dues primeres coordenades amb l'augment dels valors dels paràmetres de filtrat utilitzats.

També vaig comprovar, utilitzant les distàncies UniFrac ponderades mesurades amb les dades amb totes les ASVs i mostres que no sumaven 500 lectures filtrades, si s'observava una agrupació clara de les dades en funció d'altres variables disponibles a les metadades com edat o índex de massa corporal. No vaig observar cap agrupació clara, a continuació hi ha les representacions amb les mostres marcades en funció de l'edat i de l'índex de massa corporal com a exemples.

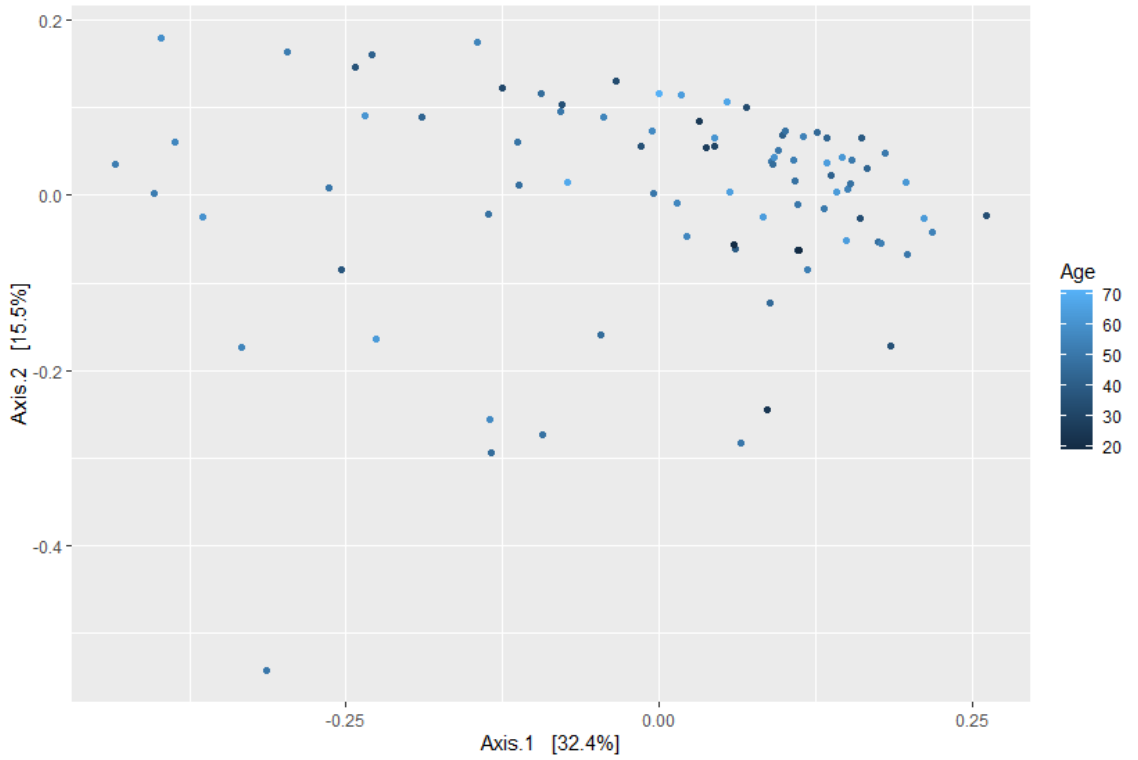

Figura 8. Representació mitjançant un MDS de les distàncies UniFrac ponderades filtrant les mostres i ASVs amb menys de 500 lectures totals. No s'observa agrupació entre les mostres en funció de l'edat.

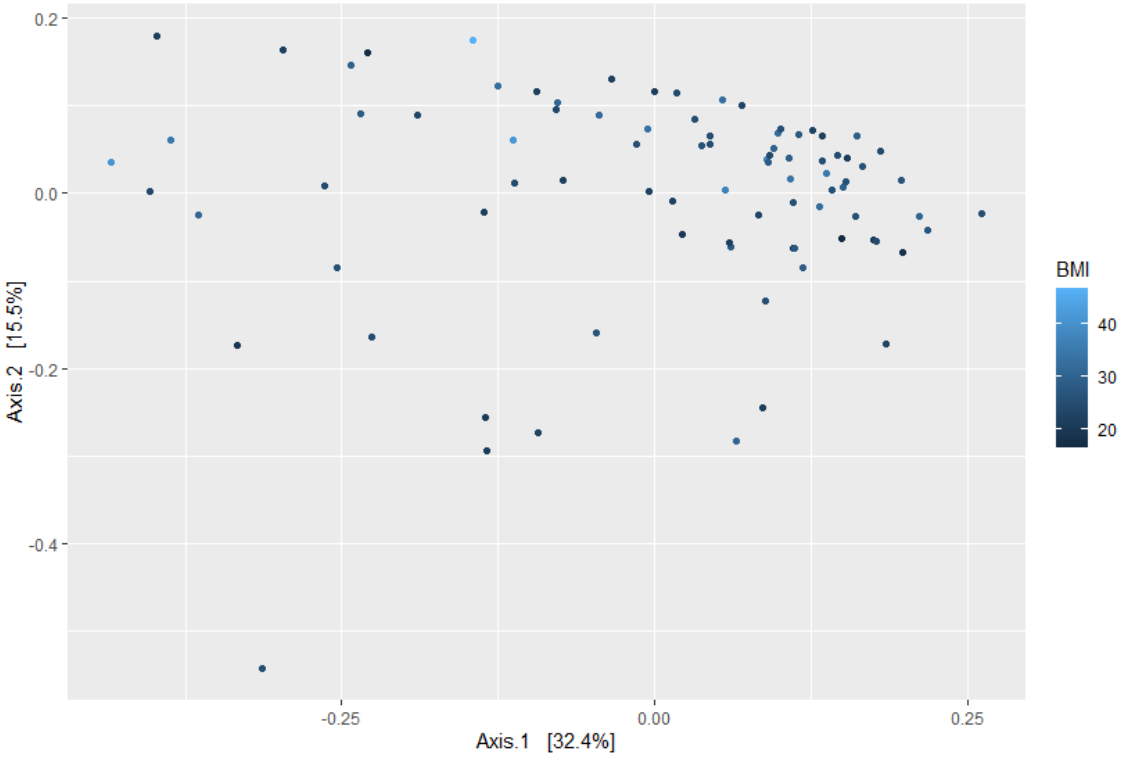

Figura 9. Representació mitjançant un MDS de les distàncies UniFrac ponderades filtrant les mostres i ASVs amb menys de 500 lectures totals. No s'observa agrupació entre les mostres en funció de l'índex de massa corporal.

Per comprovar si hi havia diferències estadísticament significatives entre les distàncies de les mostres dels controls i les dels pacients, i si aquesta significació variava utilitzant diferents paràmetres de filtrat, vaig realitzar un test PERMANOVA utilitzant la funció adonis del paquet vegan. Els p-valors resultants d'aplicar la prova estan resumits a la taula 3.

Distància	**No Filtrat**	**Filtrat 10**	**Filtrat 500**	**Filtrat 1000**
Bray-Curtis	0,003 (±0,002)	0,004 (±0,003)	0,003 (±0,002)	0,003 (±0,002)
UniFrac no ponderada	0,001 (+0,002)	0,002 (±0,001)	0,002 (±0,001)	0,001 (+0,001)
UniFrac Ponderada	0,028 (±0,01)	0,02 (±0,01)	0,02 (±0,005)	0,02 (±0,004)
Jaccard	0,001	0,001	0,001	0,001

Taula 6. Variació dels p-valors resultants d'aplicar 5 PERMANOVES per avaluar si hi havia diferències estadísticament significatives entre les distàncies de les mostres dels controls i les dels pacients en funció dels paràmetres de filtrat utilitzats.

A diferència de en la diversitat alfa, en la diversitat beta les diferències entre controls i pacients eren significatives tant amb les mostres sense filtrar com amb les mostres filtrades, i la significació d'aquestes diferències no variava amb la variació del paràmetre de filtrat utilitzat. Les úniques diferències que vaig poder observar al llarg de la comparació venien més donades per les permutacions a l'atzar de la PERMANOVA que per la variació causada per la utilització de diferents paràmetres de filtrat.

Finalment, per avaluar l'efecte de la rarefacció, per simplicitat només vaig mirar el seu efecte sobre les distàncies UniFrac no ponderades utilitzant diferents paràmetres de filtrat. La representació resultant va ser la mateixa cada vegada independentment del paràmetre de filtrat utilitzat, i coincidia tant en distribució com en explicació de la variança per part de les dues primeres coordenades amb la representació de les distàncies UniFrac no ponderades utilitzant 1000 com a paràmetre de filtrat amb les dades sense rarificar. La significació de les diferències entre les distàncies de pacients i controls tampoc va variar utilitzant diferents paràmetres de filtrat i els resultats obtinguts van ser pràcticament idèntics als de la taula 6.

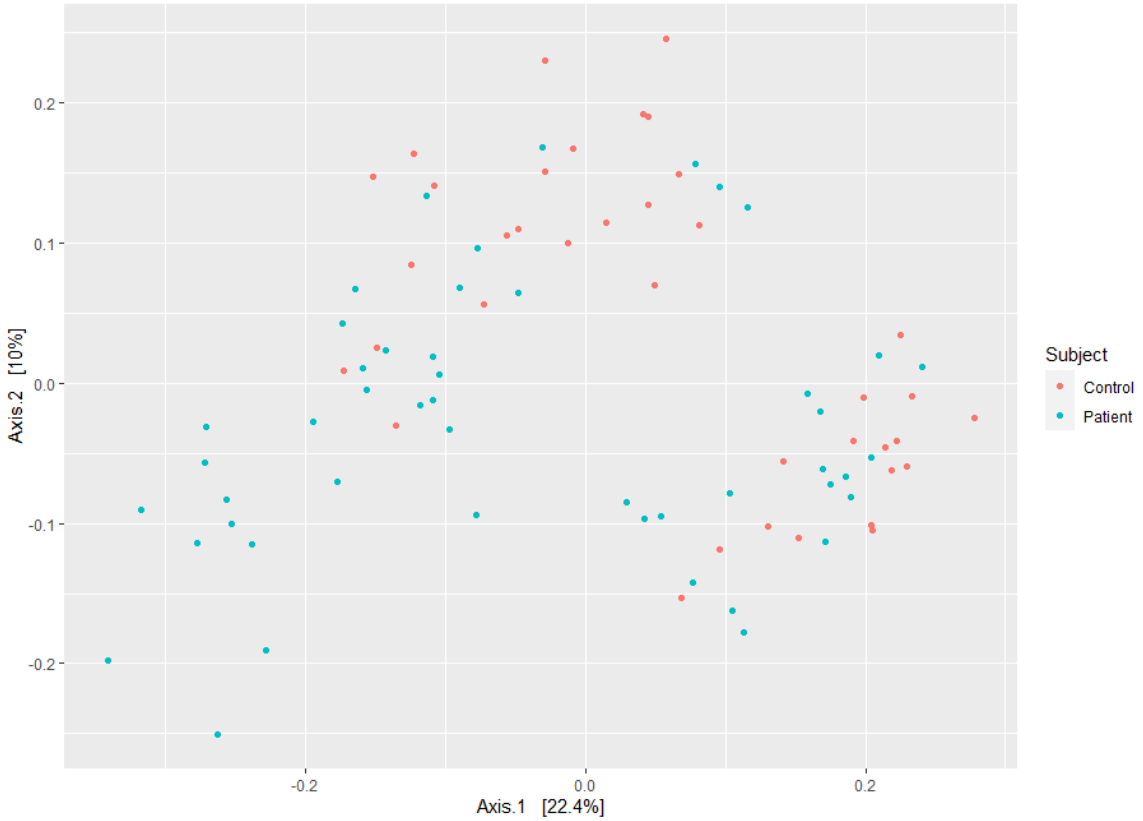

Figura 10. Representació mitjançant un MDS de les distàncies UniFrac no ponderades utilitzant les dades rarificades amb rngseed = 123 i sense filtrar, el resultat no va variar gens utilitzant diferents paràmetres de filtrat.

4.2.3. Prova d'abundàncies diferencials

També vaig fer un estudi de les composicions taxonòmiques de les mostres i de si hi havia diferències significatives entre les abundàncies de certes ASVs en funció de si les mostres provenien de pacients o de controls. Per fer-ho vaig utilitzar una prova U de Mann-Whitney a través de la funció aldex del paquet ALDEx2, dissenyat expressament per l'anàlisi d'aquest tipus de dades. En primer lloc vaig observar les composicions taxonòmiques a simple vista amb un diagrama de caixes, per fer-lo vaig utilitzar les abundàncies relatives filtrant totes les mostres i ASVs amb menys de 500 lectures i vaig aglomerar les ASVs al nivell taxonòmic fílum per facilitar la seva visualització. Vaig agrupar els fílums que no sumaven una proporció total de 0,4 entre totes les mostres a la categoria "rare", aquests van ser Fusobacteria, Lentisphaerae, Euryarchaeota, Cyanobacteria/Chloroplast, Campilobacterota, Proteobacteria, Plantae, Synergistetes. La presència d'organismes relacionats evolutivament als cianobacteris en el microbioma intestinal humà (Di Rienzi et al., 2013) podria explicar que hi hagi ASVs als quals s'han assignat taxons corresponents al fílum Cyanobacteria. Una altra explicació podria ser la presencia de restes vegetals en les mostres que també explicaria la presencia de ASVs a les quals es va assignar Plantae, per ser precisos se'ls hi va assignar el gènere *Zea*, gènere al qual pertany el blat de moro. Així també podria ser que a algunes mostres hi hagués restes de blat de moro i cloroplasts.

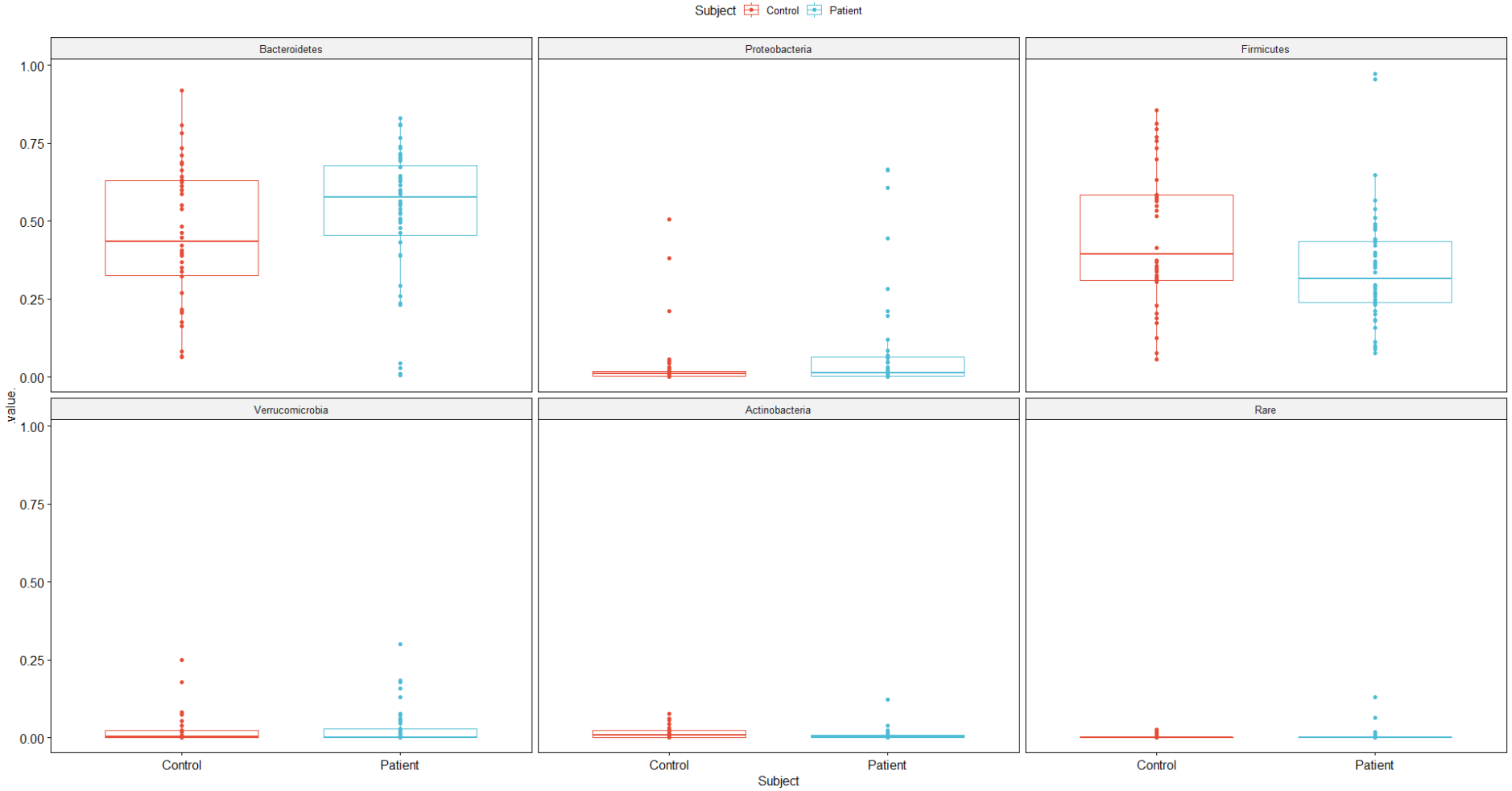

Figura 11. Diagrames de caixes de les abundàncies relatives de cada fílum en controls i pacients filtrant totes les mostres i ASVs amb menys de 500 lectures.

En la figura anterior es pot observar un aparent lleuger increment en la proporció de Proteobacteria i Bacteroidetes, i una disminució en la proporció de Firmicutes en els pacients respecte els controls. La resta no sembla canviar massa a simple vista. Per estudiar si hi havia diferències significatives entre les abundàncies dels fílums de pacients i controls i observar l'efecte de la utilització de diferents paràmetres de filtrat sobre la significació d'aquestes diferències vaig utilitzar la funció aldex amb les ASVs aglomerades al nivell taxonòmic fílum (resultats a la taula 7).

Fílum	**P-valor sense filtrar**	**P-valor filtrat 10**	**P-valor filtrat 500**	**P-valor filtrat 1000**
Actinobacteria	0,0259	0,0682	0,0124	0,0030
Firmicutes	0,0889	0,2743	0,0920	0,0761
Verrucomicrobiota	0,2724	0,3675	0,2071	0,1598
Fusobacteria	0,4484	0,5798	0,4525	-
Lentisphaerae	0,4935	0,6337	-	-
Euryarchaeota	0,6194	0,6748	0,5334	0,5001
Bacteroidetes	0,5732	0,7860	0,5932	0,5422
Cyanobacteria/Chloroplast	0,5002	0,5176	0,7161	0,6909
Campilobacterota	0,6021	0,6861	-	-
Proteobacteria	0,8736	0,8064	0,8739	0,8983
Plantae	0,6569	0,6701	-	-
Synergistetes	0,6757	0,7101	-	-

Taula 7. P-valors corregits per Benjamini-Hochberg de la prova ALDEx2 portada a terme per veure si hi havia diferències significatives entre les abundàncies dels fílums de les mostres de controls i pacients provant diferents paràmetres de filtrat.

Només hi havia diferències significatives entre pacients i controls en l'abundància del fílum Actinobacteria. En aquest cas les diferències eren més significatives utilitzant les dades sense filtrar que utilitzant les dades amb totes les mostres i ASVs que no sumaven 10 lectures totals filtrades. Malgrat aquesta petita anomalia sembla ser que la utilització de paràmetres de filtrat més grans (500 i 1000 respectivament) resultava en diferències més significatives com en el cas de la diversitat alfa. Cap dels fílums amb diferències aparents en els diagrames de caixes apareix com a diferencialment abundant. A continuació vaig aplicar la mateixa prova a les dades sense aglomerar, és a dir, al nivell taxonòmic més baix possible corresponent a l'assignat originalment a les ASVs. També vaig voler comprovar l'efecte d'utilitzar diferents paràmetres de filtrat sobre la significació de les diferències.

Dades no rarificades	Sense filtrat	Filtrat 10	Filtrat 500	Filtrat 1000
Nombre ASVs diferencialment abundants	0	0	7	14
Taxons assignats a les ASVs	-	-	Gèneres *Collinsella*, *Monoglobus*, *Erysipelatoclostridium*, *Dorea*, *Mediterraneibacter*, *Faecalibacterium* i família Ruminococcaceae	Gèneres *Collinsella*, *Monoglobus*, *Mediterraneibacter*, *Dorea*, *Faecalibacterium* (x2), *Erysipelatoclostridium*, *Ruminococcus* (x2), *Gemmiger*, *Fusicatenibacter*, *Coprococcus*, *Flavonifractor* i família Ruminococcaceae

Taula 8. Nombre d'ASVs identificades com a diferencialment abundants en funció del paràmetre de filtrat utilitzat i els seus taxons assignats utilitzant les dades sense rarificar.

Utilitzant les dades sense filtrar i les dades amb totes les mostres i ASVs que no sumaven 10 lectures filtrades no hi havia ni una sola ASV diferencialment abundant de manera significativa entre controls i pacients, amb els p-valors més baixos sent 0,067 i 0,058 respectivament, en ambdós casos corresponents a l'ASV80 a la qual s'havia assignat el gènere *Collinsella* de la família Coriobacteriaceae.

Utilitzant les dades amb totes les mostres i ASVs que no sumaven 500 lectures filtrades hi havia 7 ASVs amb abundàncies significativament diferents entre controls i pacients. A totes elles s'havien assignat taxons del fílum Firmicutes menys a la més diferent (p-valor de 0,008) que corresponia al gènere *Collinsella* pertanyent al fílum Actinobacteria i que era responsable de la major part de les diferències entre controls i pacients pel que feia a l'abundància del fílum Actinobacteria. La resta eren: el gènere *Monoglobus* (p-valor de 0,0306), família Ruminococcaceae (p-valor de 0,0375), gènere *Erysipelatoclostridium* (p-valor de 0, 0405), gènere *Dorea* (p-valor de 0,0428), gènere *Mediterraneibacter* (p-valor de 0,0434) i gènere *Faecalibacterium* (p-valor de 0,0467). Els gèneres *Monoglobus* i *Faecalibacterium* pertanyen a la família Ruminococcaceae. Els gèneres *Dorea* i *Mediterraneibacter* pertanyen a la família Lachnospiraceae. Finalment el gènere *Erysipelatoclostridium* pertany a la família Erysipelatoclostridiaceae.

Utilitzant 1000 com a paràmetre de filtrat les ASVs diferencialment abundants passaven a ser 14, el doble que amb filtrat 500. Seguien apareixent totes les ASVs que apareixien amb filtrat 500, només que amb p-valors més baixos (per exemple el p-valor de l'ASV corresponent al gènere *Collinsella* baixava fins a 0,003), i n'apareixien 7 de noves. Els taxons assignats a les 7 ASVs noves eren gènere *Ruminococcus* (p-valor de 0,034), gènere *Gemmiger* (p-valor de 0,0396), gènere *Faecalibacterium* (p-valor de 0,0397), gènere *Fusicatenibacter* (p-valor de 0,0416), gènere *Coprococcus* (p-valor de 0,044), gènere *Ruminococcus* (p-valor de 0,047), i gènere *Flavonifractor* (p-valor de 0,049). Els gèneres *Ruminococcus*, *Gemmiger*, *Faecalibacterium* i *Flavonifractor* pertanyen a la família Ruminococcaceae. Els gèneres *Fusicatenibacter* i *Coprococcus* pertanyen a la família Lachnospiraceae.

Tots els taxons que s'havien assignat a les ASVs identificades com a diferencialment abundants entre pacients i controls, excepte els gèneres *Collinsella* i *Erysipelatoclostridiaceae* pertanyien a l'ordre Clostridiales dins la classe Clostridia del fílum Firmicutes, al qual també pertany el gènere *Erysipelatoclostridiaceae* malgrat formar part de l'ordre Erysipelotrichales dins la classe Erysipelotrichia, sent el gènere *Collinsella* (de l'ordre Coriobacteriales, dins la classe Coriobacteriia del fílum Actinobacteria) l'únic que no formava part del fílum Firmicutes. Això semblaria indicar que si hi ha alguna associació entre la síndrome de la fatiga crònica i una alteració del microbioma, aquesta alteració vindria majoritàriament donada per diversos membres del fílum Firmicutes i pel gènere *Collinsella*. Independentment d'això cal destacar que la significació de les diferències entre controls i pacients pel que fa a l'abundància d'aquests taxons va variat molt en funció del paràmetre de filtrat utilitzat, amb una aparent correlació entre diferencies més significatives amb paràmetres de filtrat més grans, fent que aquests resultats siguin, com a mínim, qüestionables.

Per acabar vaig voler avaluar l'efecte sobre els resultats d'utilitzar les dades rarificades amb els diferents paràmetres de filtrat. Aparentment la seva utilització va reduir la significació de les diferències entre controls i pacients.

Dades rarificades	Sense filtrat	Filtrat 10	Filtrat 500	Filtrat 1000
Nombre ASVs diferencialment abundants	0	0	4	9
Taxons assignats a les ASVs	-	-	Gèneres *Collinsella*, *Monoglobus*, *Erysipelatoclostridium* i *Mediterraneibacter*	Gèneres *Collinsella*, *Monoglobus*, *Mediterraneibacter*, *Dorea*, *Faecalibacterium*, *Erysipelatoclostridium*, *Ruminococcus*, *Flavonifractor* i família Ruminococcaceae

Taula 9. Nombre d'ASVs identificats com a diferencialment abundants en funció del paràmetre de filtrat utilitzat i els seus taxons assignats utilitzant les dades rarificades.

Utilitzant les dades rarificades sense filtrar i les dades rarificades amb totes les mostres i ASVs que no sumaven 10 lectures filtrades, no hi havia ni una sola ASV diferencialment abundant de manera significativa entre controls i pacients, amb els p-valors més baixos sent 0,107 i 0,069 respectivament, en ambdós casos corresponents a l'ASV80 a la qual s'havia assignat el gènere *Collinsella* de la família Coriobacteriaceae. Aquests p-valors eren més alts que els obtinguts amb les dades sense rarificar.

Amb filtrat 500 i les dades rarificades va passar d'haver-hi 7 ASVs identificades com a diferencialment abundants de manera significativa a haver-n'hi 4, totes elles presents entre les 7 que s'identificaven sense rarificar les dades amb filtrat 500. L'ASV a la qual

s'havia assignat el gènere *Collinsella* seguia sent la més diferencialment abundant amb un p-valor de 0,017 (comparat amb 0,008 sense rarificar), seguida per les ASVs a les quals s'havia assignat el gènere *Erysipelatoclostridium* amb un p-valor de 0,035 (comparat amb 0, 0405 sense rarificar), el gènere *Monoglobus* amb un p-valor de 0,0474 (comparat amb 0,0306 sense rarificar), i el gènere *Mediterraneibacter* amb un p-valor de 0,0496 (comparat amb 0,0434 sense rarificar).

Finalment amb filtrat 1000 i les dades rarificades va passar d'haver-hi 14 ASVs identificades com a diferencialment abundants de manera significativa a haver-n'hi 9, totes elles presents entre les 14 que s'identificaven sense rarificar les dades amb filtrat 1000. L'ASV a la qual s'havia assignat el gènere *Collinsella* seguia sent la més diferencialment abundant amb un p-valor de 0,009 (comparat amb 0,003 sense rarificar), seguida per les ASVs a les quals s'havia assignat el gènere *Monoglobus* amb un p-valor de 0,022 (comparat amb 0,0197 sense rarificar), la família Ruminococcaceae amb un p-valor de 0,031 (comparat amb 0,023 sense rarificar), el gènere *Erysipelatoclostridium* amb un p-valor de 0,0325 (comparat amb 0,0307 sense rarificar), el gènere *Faecalibacterium* amb un p-valor de 0,0339 (comparat amb 0,0314 sense rarificar), el gènere *Dorea* amb un p-valor de 0,0399 (comparat amb 0,0306 sense rarificar), el gènere *Ruminococcus* amb un p-valor de 0,0402 (comparat amb 0,0372 sense rarificar), el gènere *Mediterraneibacter* amb un p-valor de 0,0411 (comparat amb 0,0241 sense rarificar), i el gènere *Flavonifractor* amb un p-valor de 0,044 (comparat amb 0,049 sense rarificar).

Aparentment la rarefacció va disminuir la significació de les diferències entre controls i pacients, resultant en menys ASVs sent diferencialment abundants de manera significativa entre controls i pacients que utilitzant les dades no rarificades. Malgrat això, si es mira l'efecte sobre els p-valors obtinguts, aquest va ser major o menor segons l'ASV: en algunes va resultar en pujades i en altres en baixades respecte els p-valors obtinguts amb les dades sense rarificar. Així l'efecte de la rarefacció sobre els resultats a nivell de proves d'abundàncies diferencials també semblava ser considerable, però no tant constant com l'efecte de la utilització de diferents paràmetres de filtrat.

4.2.4. Representació dels fílums diferencialment abundants

Mitjançant la seva representació en un arbre filogenètic, vaig voler observar la distribució dels diferents taxons dins els fílums on vaig observar diferències significatives, és a dir, dins de Firmicutes i Actinobacteria a través de la funció "plot_tree" del paquet "phyloseq". La representació d'Actinobacteria va resultar molt més senzilla pel baix nombre de ASVs a les quals s'havíen assignat taxons pertanyents a aquest fílum que apareixen a les dades, fent possible la seva representació directament al nivell taxonòmic més baix disponible (gènere).

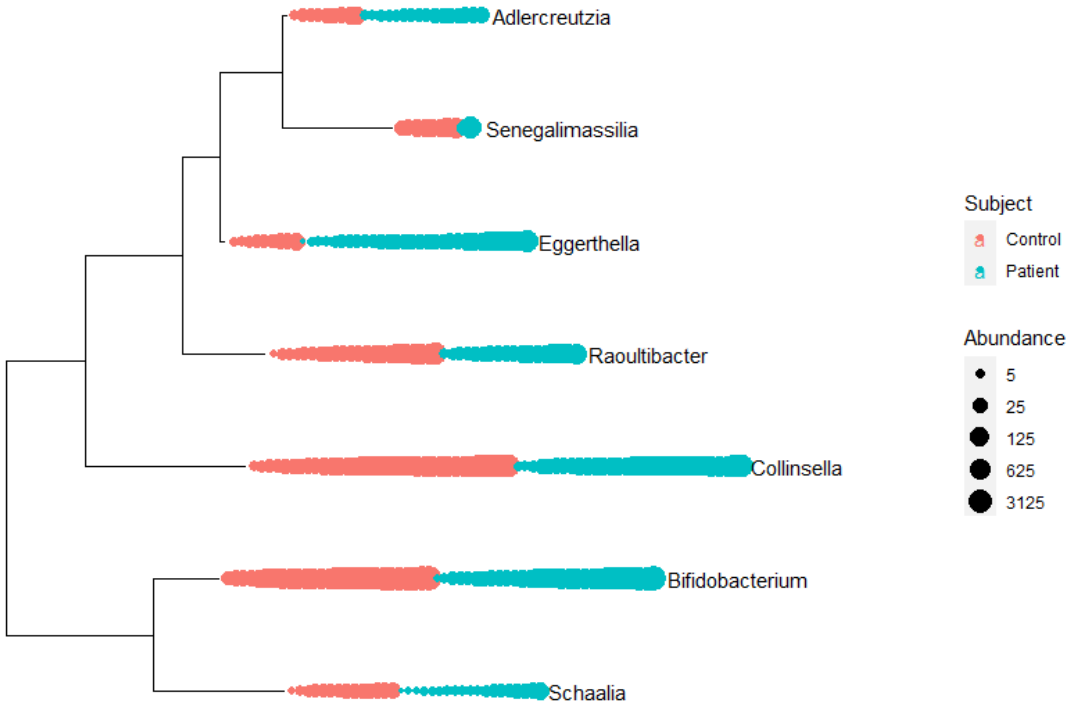

Figura 12. Representació mitjançant un arbre filogenètic dels gèneres del fílum Actinobacteria amb les seves abundàncies en controls i pacients. Filtrades les ASVs i mostres amb menys de 500 lectures totals. S'utilitza el nombre de lectures en comptes de l'abundància relativa (lectures/lectures totals).

La representació no va resultar molt informativa ja que el gènere *Collinsella*, que era el més diferencialment abundant d'entre tots els taxons, es trobava, aparentment, en proporcions similars en els dos grups, mentre que altres generes com *Senegalimassilia* i *Eggerthella*, variaven molt de manera aparent entre el grup de controls i el de pacients malgrat no aparèixer com a diferencialment abundants de manera significativa aplicant la funció aldex. El resultat de la representació no semblava variar utilitzant les abundàncies relatives en comptes de les absolutes. Utilitzant paràmetres de filtrat més baixos anaven apareixent més gèneres poc abundants la informació sobre els quals es perd completament amb paràmetres de filtrat més grans (per exemple, utilitzant 10 en comptes de 500, apareixien 19 gèneres, més que el doble dels 7 que apareixien utilitzant 500), això va servir per posar de manifest tota la informació sobre taxons més minoritaris que es perd amb el filtrat.

La representació del fílum Firmicutes, fins i tot a nivell família, va resultar molt poc informativa pel gran nombre d'ASVs als quals s'havien assignat taxons d'aquest fílum.

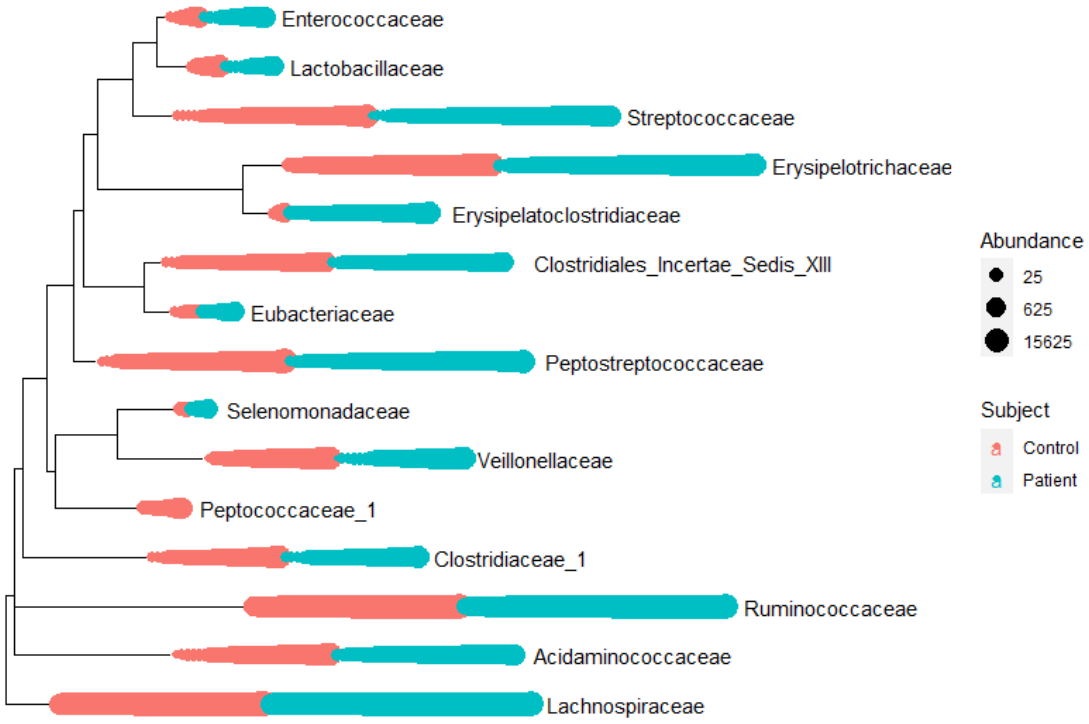

Figura 13. Representació mitjançant un arbre filogenètic dels gèneres del fílum Firmicutes amb les seves abundàncies en controls i pacients. Filtrades les ASVs i mostres amb menys de 500 lectures totals. S'utilitza el nombre de lectures en comptes de l'abundància relativa (lectures/lectures totals).

Igual que en el cas anterior les aparences semblaven tornar a enganyar. La família Ruminococcaceae, que era diferencialment abundant de manera estadísticament significativa, i a la qual pertanyien dos dels gèneres que també ho eren (també utilitzant filtrat 500), aparentava ser igualment abundant en controls que en pacients, igual que la família Lachnospiraceae, a la qual pertanyien els gèneres diferencialment abundants *Dorea* i *Mediterraneibacter*. L'única diferència que es podia observar de manera clara a simple vista i que resultava ser significativa a nivell estadístic era la de la família Erysipelatoclostridiaceae a la qual pertany el gènere diferencialment abundant *Erysipelatoclostridium*.

4.2.5. Selecció de balanços

L'últim mètode d'anàlisi que vaig utilitzar va ser la selecció de balanços mitjançant les funcions selbal i CoDaCoRe. Amb selbal, després de tenir diversos problemes amb la part de la substitució de zeros de la funció i posteriorment executant la funció amb les dades al nivell taxonòmic més baix, vaig acabar utilitzant les dades amb les ASVs aglomerades al nivell taxonòmic família sumant 1 a totes les lectures per substituir els 0 per 1 d'una

manera “manual”. Vaig executar la funció amb tots els paràmetres predeterminats i utilitzant les dades amb les ASVs i mostres amb menys de 500 lectures totals filtrades.

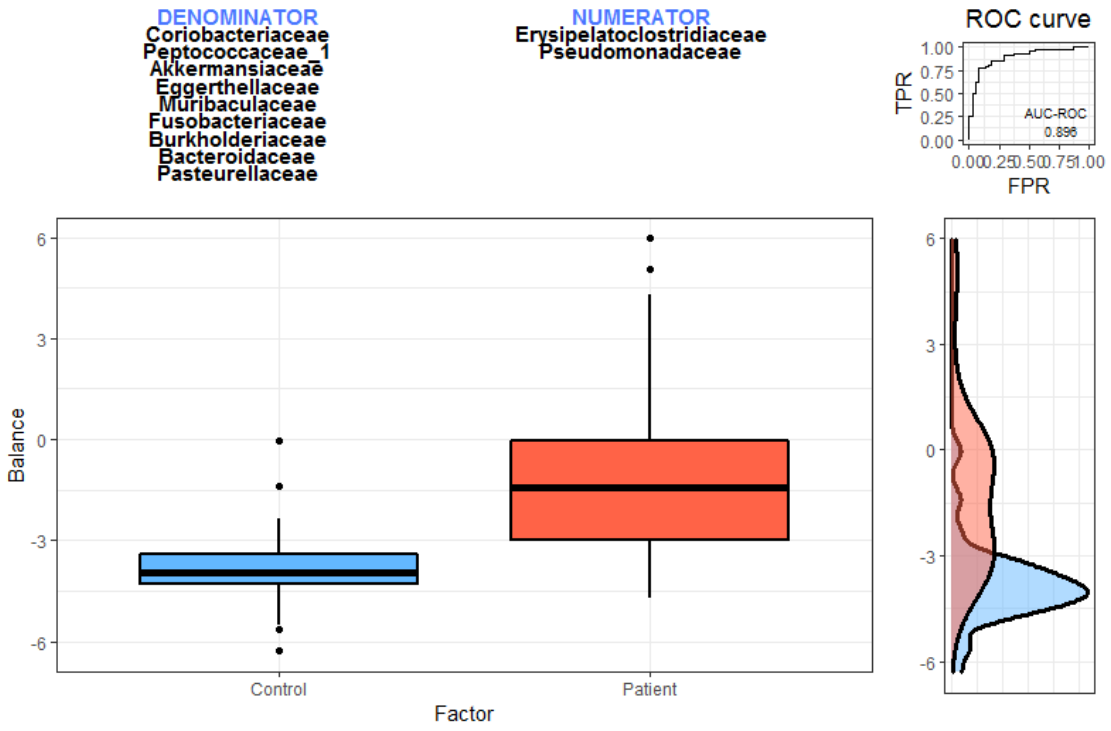

Figura 14. Resultat d’executar la funció selbal amb les dades aglomerades a nivell família i la substitució dels 0 per 1. Descripció del balanç global: les taxonomies sota denominador i numerador són els grups de famílies que formen el balanç, els diagrames de caixes mostren la distribució de les puntuacions del balanç per controls i pacients. A dalt a la dreta hi ha la corba de la característica operativa del receptor amb el seu valor de l’àrea sota la corba (0,896), a sota hi ha la corba de densitat per controls (blau) i pacients (vermell).

En aquest cas el balanç global o signatura microbiana venia definit pels grups de famílies {Coriobacteriaceae, Peptococcaceae, Akkermansiaceae, Eggerthellaceae, Muribaculaceae, Fusobacteriaceae, Burkholderiaceae, Bacteroidaceae i Pasteurellaceae} i {Erysipelatoclostridiaceae i Pseudomonadaceae}. Els controls tenien puntuacions del balanç més baixes que els pacients, indicant que en ells l’abundància mitjana relativa de les famílies Erysipelatoclostridiaceae i Pseudomonadaceae era més baixa que l'abundància mitjana de les famílies Coriobacteriaceae, Peptococcaceae, Akkermansiaceae, Eggerthellaceae, Muribaculaceae, Fusobacteriaceae, Burkholderiaceae, Bacteroidaceae i Pasteurellaceae. No vaig estimar la solidesa del balanç global seleccionat mitjançant validació creuada per problemes a l’hora d’executar la funció selbal.cv amb aquestes dades.

De les famílies que apareixen en el balanç només dues havien estat identificades com a diferencialment abundants de manera significativa entre controls i pacients a través de la funció aldex del paquet ALDEx2, la dels Erysipelatoclostridiaceae i la dels Coriobacteriaceae, la resta no s'havien detectat.

Una altra eina per la selecció de balanços que vaig utilitzar és el CoDaCoRe. En aquest cas no vaig tenir problemes en l’execució, malgrat que la substitució de 0 per utilitzar la

funció s'havia de fer manualment (amb el mateix mètode que vaig utilitzar en el selbal de sumar 1 a totes les lectures), almenys amb R. Vaig fer servir balanços com a tipus de log ratio i lambda 1. També vaig utilitzar les dades amb les ASVs i mostres amb menys de 500 lectures totals filtrades.

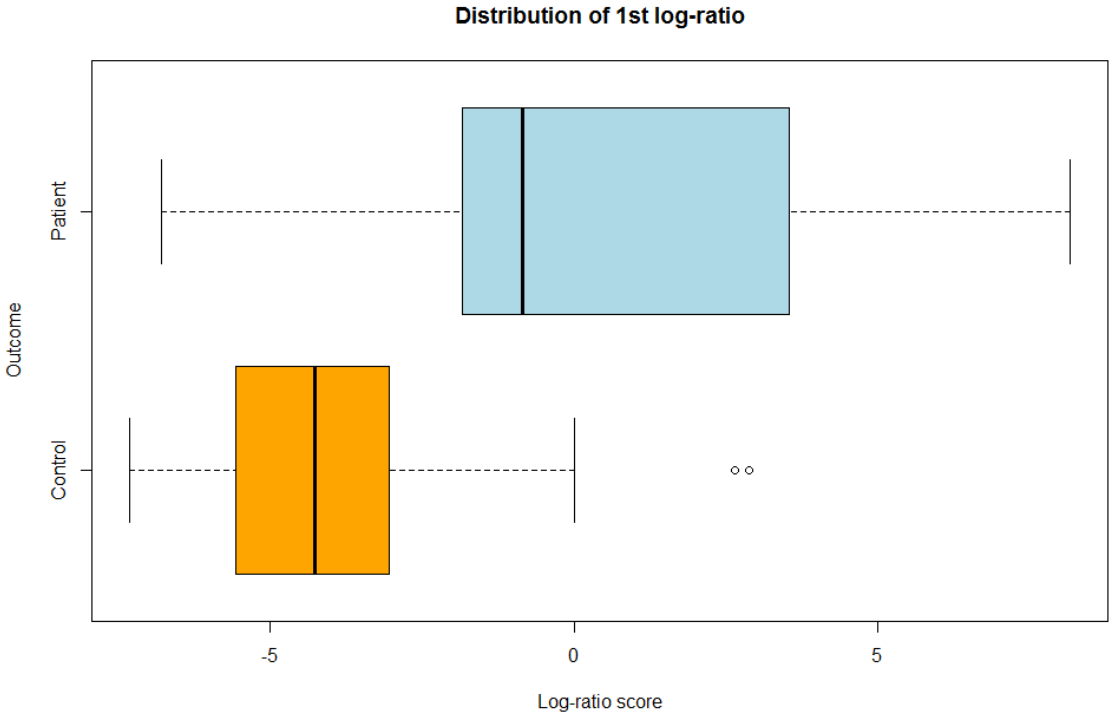

Figura 15. Resultat de construir un model amb CoDaCoRe.

Amb aquesta funció els resultats s'imprimeixen a la consola. El balanç, utilitzant les dades sense cap tipus d'aglomeració, en aquest cas va venir definit únicament per la ASV a la qual s'havia assignat el gènere *Erysipelatoclostridium* de la família Erysipelatoclostridiaceae en el numerador i per les ASVs a les quals s'havia assignat el gènere *Ruminococcus* de la família Ruminococcaceae, la família Ruminococcaceae, i el gènere *Collinsella* de la família Coriobacteriaceae en el denominador. Per tant els controls tenien una abundància relativa de *Erysipelatoclostridium* més baixa que de *Ruminococcus, Collinsella* i Ruminococcaceae. El valor de l'àrea sota la curva va ser de 0,865. Així, tots els taxons del balanç havien estat identificats com a diferencialment abundants a través de la funció aldex amb filtrat 500.

El diagrama de caixes amb la distribució de les puntuacions del balanç era molt similar al del selbal amb els pacients tenint puntuacions més altes que els controls. Els grups de taxons seleccionats contenien menys taxons que els del selbal malgrat haver construït el model al nivell taxonòmic més baix possible. Per poder fer una comparació entre els resultats dels dos mètodes vaig utilitzar les dades amb les ASVs aglomerades al nivell taxonòmic família com havia fet amb el selbal.

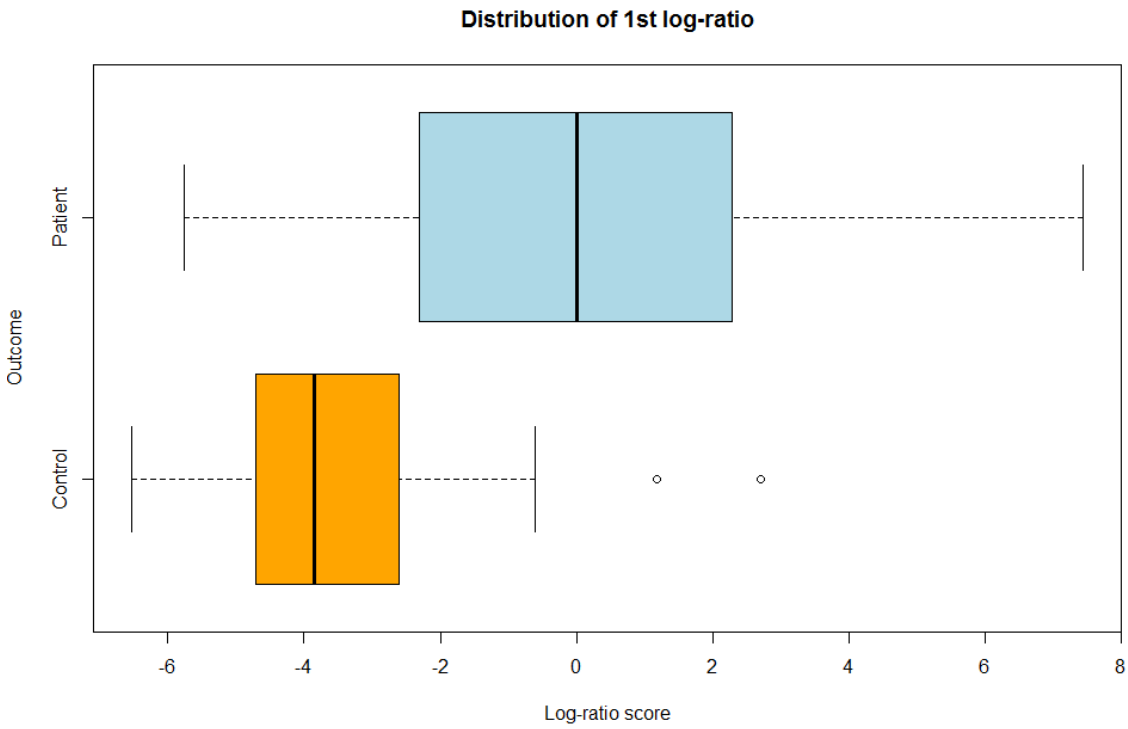

Figura 16. Resultat de construir un model amb CoDaCoRe amb les dades aglomerades a nivell família.

El resultat casi no va canviar a nivell de distribució de les puntuacions. Els grups de famílies que definien el balanç van ser Erysipelatoclostridiaceae en el numerador i Akkermansiaceae, Bifidobacteriaceae Coriobacteriaceae, i Muribaculaceae en el denominador. Com en el balanç global del selbal això volia dir que l'abundància relativa de Erysipelatoclostridiaceae era més baixa en els controls que l'abundància mitjana de Akkermansiaceae, Bifidobacteriaceae Coriobacteriaceae, i Muribaculaceae. El valor de l'àrea sota la corba va ser de 0,832, inferior a utilitzant el nivell taxonòmic més baix possible. Així a nivell dels grups de taxons que defineixen el balanç el canvi va ser radical respecte al balanç global de CoDaCoRe anterior. Malgrat no canviar el numerador, el denominador va canviar radicalment, amb la família Ruminococcaceae desapareixent completament (2/3 del denominador anterior) i apareixent les famílies, Akkermansiaceae, Bifidobacteriaceae, i Muribaculaceae, sent Akkermansiaceae i Muribaculaceae també presents en el denominador del balanç global obtingut amb selbal, igual que Erysipelatoclostridiaceae en el numerador. Per tant el balanç global va passar a assemblar-se molt més a l'obtingut amb el selbal, demostrant la similaritat entre els dos mètodes (malgrat que el CoDaCoRe seleccionés un balanç amb grups de taxons més petits), i la importància d'utilitzar les mateixes condicions per fer comparacions. També s'ha de dir que el temps d'execució del selbal va ser més gran que el de CoDaCoRe.

També vaig voler avaluar l'efecte d'utilitzar diferents paràmetres de filtrat sobre els resultats obtinguts amb aquests mètodes. Per fer-ho vaig, com en els casos anteriors, comparar els resultats obtinguts amb les dades sense filtrat i amb les dades amb les ASVs i mostres amb menys de 10, 500 i 1000 lectures totals filtrades. Els resultats obtinguts són resumits en la taula 10 a continuació.

Mètode		**Sense Filtrat**	**Filtrat 10**	**Filtrat 500**	**Filtrat 1000**
Selbal	Numerador	Erysipelatoclostridiaceae, Peptoniphilaceae, Gracilibacteraceae, Synergistaceae, Streptophyta, Selenomonadaceae, Pseudomonadaceae, Atopobiaceae	Erysipelatoclostridiaceae, Peptoniphilaceae, Pseudomonadaceae, Gracilibacteraceae, Clostridiales_Incertae_Sedis_XIII, Morganellaceae	Erysipelatoclostridiaceae, Pseudomonadaceae	Erysipelatoclostridiaceae, Pseudomonadaceae
	Denominador	Coriobacteriaceae, Kiloniellaceae, Puniceicoccaceae, Pasteurellaceae, Hafniaceae, Peptococcaceae_1, Sutterellaceae, Eggerthellaceae, Neisseriaceae, Christensenellaceae, Candidatus_Carsonella, Bacillales_Incertae_Sedis_XI, Porphyromonadaceae, Burkholderiaceae	Coriobacteriaceae, Kiloniellaceae, Peptococcaceae_1, Pasteurellaceae, Sporomusaceae, Hafniaceae, Burkholderiaceae, Neisseriaceae, Christensenellaceae, Candidatus_Carsonella	Coriobacteriaceae, Peptococcaceae_1, Akkermansiaceae, Eggerthellaceae, Muribaculaceae, Fusobacteriaceae, Burkholderiaceae, Bacteroidaceae, Pasteurellaceae	Coriobacteriaceae, Peptococcaceae_1, Akkermansiaceae, Muribaculaceae, Eggerthellaceae, Lactobacillaceae, Enterococcaceae
	AUC	0,962	0,931	0,896	0,889

CoDaCoRe	Numerador	Erysipelatoclostridiaceae, Peptoniphilaceae	Erysipelatoclostridiaceae, Peptoniphilaceae	Erysipelatoclostridiaceae	Erysipelatoclostridiaceae
	Denominador	Akkermansiaceae, Bifidobacteriaceae Veillonellaceae, Sutterellaceae, Coriobacteriaceae, Muribaculaceae, Eggerthellaceae, Pasteurellaceae, Desulfovibrionaceae, Peptococcaceae_1	Rikenellaceae, Akkermansiaceae Porphyromonadaceae, Bifidobacteriaceae, Veillonellaceae, Sutterellaceae, Coriobacteriaceae, Muribaculaceae, Eggerthellaceae, Pasteurellaceae, Desulfovibrionaceae, Peptococcaceae_1	Akkermansiaceae, Bifidobacteriaceae, Coriobacteriaceae, Muribaculaceae, Peptococcaceae_1	Akkermansiaceae, Bifidobacteriaceae, Coriobacteriaceae, Muribaculaceae
	AUC	0,875	0,861	0,844	0,839

Taula 10. Resultat de seleccionar balanços utilitzant diferents paràmetres de filtrat amb selbal i CoDaCoRe.

Els balanços seleccionats variaven bastant en funció del paràmetre de filtrat utilitzat, amb algunes famílies seleccionades en tots els casos i coincidint en els dos mètodes, com Erysipelatoclostridiaceae al numerador i Akkermansiaceae i Coriobacteriaceae en el denominador. Erysipelatoclostridiaceae i Coriobacteriaceae són les famílies a les quals pertanyen els gèneres identificats com a diferencialment abundants a través de la funció aldex *Erysipelatoclostridium* i *Collinsella*, indicant una certa consistència en els resultats obtinguts amb els tres mètodes. Generalment s'observava una tendència a la disminució del nombre de famílies presents en el balanç a mesura que s'augmentava el paràmetre de filtrat. Cal destacar que els resultats del CoDaCoRe semblaven variar lleugerament amb cada execució de la funció i determinar quins canvis eren provocats per la variació del paràmetre de filtrat i quins eren provocats per les pròpies variacions entre execucions era difícil.

4.2.6. Comparació amb els resultats de l'article de referència

A l'article de referència "Reduced diversity and altered composition of the gut microbiome in individuals with myalgic encephalomyelitis/chronic fatigue syndrome" (Giloteaux et al., 2016) destacaven que hi havia una reducció de la diversitat alfa en els pacients amb síndrome de fatiga crònica respecte els controls. Aquesta diversitat la van mesurar de 3 maneres diferents: mitjançant l'índex Chao1, l'índex Shannon i Phylogenetic Diversity. Els p-valors que van obtenir per amb cada mesura, van ser 0,002, 0,004 i 0,004 respectivament. En la meva anàlisi els resultats de la diversitat alfa van variar molt en funció del paràmetre de filtrat utilitzat, i només van arribar a p-valors tant baixos utilitzant les dades amb totes les mostres i ASVs amb menys de 1000 lectures totals filtrades. Així, assumint que no hi hagués cap altra diferència entre les nostres metodologies, els resultats de l'article semblarien indicar que van utilitzar un paràmetre de filtrat superior a 1000. Malgrat això és evident que hi va haver diferències entre les nostres metodologies de processament, no només perquè que ells van utilitzar el QUIIME i jo l'RStudio, sinó que també perquè ells van agrupar les seqüencies amb un percentatge d'identitat per sobre d'un llindar arbitrari en OTUs i jo vaig utilitzar el mètode DADA2 per deduir les ASVs, possiblement resultant en diferències considerables en les nostres taules.

A continuació estudiaven la diversitat beta de les mostres mesurant les distàncies UniFrac i UniFrac ponderada aplicades a les dades rarificades i representant-les mitjançant un anàlisi dels components principals. No van observar agregació entre cap dels diferents paràmetres que van provar i no sembla que comprovessin si hi havia diferències significatives entre les distàncies mitjançant alguna prova estadística. En els resultats que vaig obtenir es podia observar una certa agregació entre les mostres en funció de si eren controls o pacients en la representació de la ordinació MDS de les distàncies UniFrac no ponderades, però no semblava haver-hi una agregació clara en la representació de les distàncies UniFrac ponderades ni provant les diferents variables disponibles a les metadades. En ambdós casos les diferències entre les distàncies de controls i pacients van resultar ser estadísticament significatives utilitzant el test PERMANOVA.

A nivell d'OTUs diferencialment abundants entre controls i pacients de manera significativa, a l'article de referència en van trobar 40 utilitzant una prova de Wilcoxon, majoritàriament dins el fílum Firmicutes. En aquest sentit els resultats que vaig obtenir també van ser radicalment diferents, i molt variables en funció del paràmetre de filtrat utilitzat. Amb les dades sense rarificar i amb totes les mostres i ASVs amb menys de 1000 lectures totals filtrades, es van identificar només 14 ASVs com a diferencialment abundants de manera significativa, nombre que no és ni la meitat de les 40 de l'article. Part de la gran diferència entre els resultats pot venir donada pel fet que el mètode que jo vaig utilitzar treballava amb log-ratios, ajustant-se a la naturalesa composicional de les dades, i el mètode que ells van utilitzar, no. També és possible que a més utilitzessin un paràmetre de filtrat superior a 1000.

5. Conclusió

En conclusió, la base de dades més adequada per la descàrrega de dades crues de microbioma de manera casual i senzilla és l'ENA, malgrat tenir limitacions a l'hora d'obtenir les metadades associades. El processament de les lectures de la seqüenciació amb R, malgrat la seva infrautilització, no només és possible sinó que també relativament senzill. El problema principal és la gran demanda computacional, especialment problemàtica per estudis amb moltes mostres o per usuaris amb ordinadors poc potents. Això podria explicar part d'aquesta infrautilització.

Els resultats que he obtingut amb l'anàlisi coincideixen amb els de l'article de referència en diversos aspectes com les abundàncies taxonòmiques, la reducció de la diversitat alfa en pacients respecte els controls, i la no agregació de les dades, però divergeixen altament en el resultat de la prova d'abundàncies diferencials i en la significació dels resultats obtinguts. Aquestes divergències poden venir donades pel processament bioinformàtic previ a l'anàlisi estadística (vaig treballar amb ASVs i en l'article treballen amb OTUs) i també per les diferències en el procediment d'anàlisi estadística, sobretot el paràmetre de filtrat. Per poder-ho assegurar, però, s'hauria de processar les dades amb R i analitzar-les amb QUIIME 2 seguint exactament el mateix procediment que el de l'article.

He observat també el gran impacte que té el filtrat sobre els resultats obtinguts, no només a nivell de pèrdua de la informació que aporten els taxons més minoritaris, sinó que també a nivell de grans canvis en la significació dels resultats, sobretot pel que fa a les diferències en la diversitat alfa i en la prova d'abundàncies diferencials, que semblen resultar en diferències més significatives com més gran és el paràmetre de filtrat. L'efecte de la rarefacció sembla ser modest sobre la diversitat alfa però el seu impacte sembla ser més gran sobre la prova d'abundàncies diferencials encara que no tant constant. Pel que fa a la diversitat beta, l'efecte del filtrat sembla ser més gran sobre la representació de les distàncies que sobre les distàncies entre les mostres en sí, i la rarefacció no sembla afectar massa els resultats. Finalment, pel que fa a la selecció de balanços, els resultats demostren clares similituds entre selbal i CoDaCoRe i semblen ser parcialment consistents amb els resultats de la prova d'abundàncies diferencials, malgrat que l'impacte del filtrat sobre ells també és clar, amb l'aparició de menys taxons als balanços com més gran el paràmetre de filtrat.

Els resultats de les anàlisis porten a conclusions similars a les de l'article de referència. Els pacients amb síndrome de la fatiga crònica presentaven una diversitat més reduïda que els controls sans, hi havia diferències significatives entre les distàncies de les mostres de controls i de pacients, i diverses ASVs van ser identificades com a diferencialment abundants, algunes d'elles també apareixent als balanços seleccionats pel selbal i el CoDaCoRe, tot indicant una associació entre la síndrome de la fatiga crònica i una alteració en la composició del microbioma intestinal. Malgrat això, la gran variabilitat de la significació de les diferències entre controls i pacients en funció del paràmetre de filtrat utilitzat fa que aquests resultats no siguin excessivament sòlids. Part d'aquesta falta de significació pot ser explicada pel fet que dins el grup control hi hagués 7 individus amb molèsties gastrointestinals que no estaven indicats a les metadades. Igualment per poder arribar a conclusions sòlides sobre l'associació entre la síndrome de la fatiga crònica i la composició del microbioma intestinal s'hauria de realitzar un estudi amb més individus,

on es comparés el resultat de la seqüenciació shotgun amb el de la seqüenciació d'amplicons, i on hi hagués una avaluació més gran de l'efecte que la metodologia estadística té sobre els resultats.

Per acabar, al llarg del desenvolupament d'aquest treball he pogut apreciar la complexitat computacional i metodològica de l'anàlisi del microbioma i la sensibilitat dels resultats a les diferents tècniques emprades en el processament i anàlisi de les dades. També he pogut observar que algunes de les tècniques més utilitzades no s'adapten a la naturalesa composicional de les dades i/o dificulten la reproductibilitat dels resultats. Això posa de manifest la necessitat de disposar d'uns protocols estandarditzats de processament i anàlisi, amb mètodes que s'adaptin a la naturalesa composicional de les dades, fent així que els resultats obtinguts siguin reproduïbles, significatius, i comparables amb els resultats d'altres estudis.

6. Referències

Berg, G., Rybakova, D., Fischer, D., Cernava, T., Vergès, M., & Charles, T. et al. (2020). Microbiome definition re-visited: old concepts and new challenges. *Microbiome*, *8*(1). doi: 10.1186/s40168-020-00875-0. Retrieved 22/05/2021, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7329523/

Callahan, B. DADA2 ITS Pipeline Workflow (1.8). Benjjneb.github.io. Retrieved 28/05/2021, from https://benjjneb.github.io/dada2/ITS_workflow.html

Callahan, B. DADA2 Pipeline Tutorial (1.16). Benjjneb.github.io. Retrieved 28/05/2021, from https://benjjneb.github.io/dada2/tutorial.html

Callahan, B., Sankaran, K., Fukuyama, J., McMurdie, P., & Holmes, S. (2017). *Workflow for Microbiome Data Analysis: from raw reads to community analyses*. Web.stanford.edu. Retrieved 28/05/2021, from http://web.stanford.edu/class/bios221/MicrobiomeWorkflowII.html.

Calle, M. (2019). Statistical Analysis of Metagenomics Data. *Genomics & Informatics*, *17*(1), e6. https://doi.org/10.5808/gi.2019.17.1.e6. Retrieved 23/05/2021, from https://genominfo.org/journal/view.php?doi=10.5808/GI.2019.17.1.e6

Cao, Q., Sun, X., Rajesh, K., Chalasani, N., Gelow, K., & Katz, B. et al. (2021). Effects of Rare Microbiome Taxa Filtering on Statistical Analysis. *Frontiers In Microbiology*, *11*. doi: 10.3389/fmicb.2020.607325. Retrieved 23/05/2021, from https://www.frontiersin.org/articles/10.3389/fmicb.2020.607325/full

Cavicchioli, R., Ripple, W., Timmis, K., Azam, F., Bakken, L., & Baylis, M. et al. (2019). Scientists' warning to humanity: microorganisms and climate change. *Nature Reviews Microbiology*, *17*(9), 569-586. doi: 10.1038/s41579-019-0222-5. Retrieved 22/05/2021, from https://pubmed.ncbi.nlm.nih.gov/31213707/

Di Rienzi, S., Sharon, I., Wrighton, K., Koren, O., Hug, L., & Thomas, B. et al. (2013). The human gut and groundwater harbor non-photosynthetic bacteria belonging to a new candidate phylum sibling to Cyanobacteria. *Elife*, *2*. https://doi.org/10.7554/elife.01102. Retrieved 28/05/2021, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3787301/

Edgar, R. *Alpha diversity*. Drive5.com. Retrieved 28/05/2021, from https://www.drive5.com/usearch/manual/alpha_diversity.html

Galloway-Peña, J., & Hanson, B. (2020). Tools for Analysis of the Microbiome. *Digestive Diseases And Sciences*, *65*(3), 674-685. https://doi.org/10.1007/s10620-020-06091-y. Retrieved 28/05/2021, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7598837/

Giloteaux, L., Goodrich, J., Walters, W., Levine, S., Ley, R., & Hanson, M. (2016). Reduced diversity and altered composition of the gut microbiome in individuals with myalgic encephalomyelitis/chronic fatigue syndrome. *Microbiome*, *4*(1). doi: 10.1186/s40168-016-0171-4. Retrieved 28/04/2021, from https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-016-0171-4

Gloor, G., Fernandes, A., Macklaim, J., Albert, A., Links, M., & Quinn, T. et al. (2021). ALDEx2. Retrieved 26/05/2021, from https://bioconductor.org/packages/release/bioc/html/ALDEx2.html

Gordon-Rodriguez, E., Quinn, T., & Cunningham, J. Learning Sparse Log-Ratios for High-Throughput Sequencing Data. doi: 10.1101/2021.02.11.430695. Retrieved 22 April 2021, from https://www.biorxiv.org/content/10.1101/2021.02.11.430695v2

Kim, B., Shin, J., Guevarra, R., Lee, J., Kim, D., & Seol, K. et al. (2017). Deciphering Diversity Indices for a Better Understanding of Microbial Communities. Journal Of Microbiology And Biotechnology, 27(12), 2089-2093. https://doi.org/10.4014/jmb.1709.09027. Retrieved 06/03/2021.

Martín, R., Miquel, S., Langella, P., & Bermúdez-Humarán, L. (2014). The role of metagenomics in understanding the human microbiome in health and disease. *Virulence*, *5*(3), 413-423. doi: 10.4161/viru.27864. Retrieved 22/05/2021, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3979869/

McMurdie, P., & Holmes, S. (2013). phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. Plos ONE, 8(4), e61217. doi: 10.1371/journal.pone.0061217 Retrieved 03/06/2021, from https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0061217

McMurdie, P., & Holmes, S. (2014). Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. *Plos Computational Biology*, *10*(4), e1003531. https://doi.org/10.1371/journal.pcbi.1003531. Retrieved 28/05/2021, from https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003531

Morgan, X., & Huttenhower, C. (2012). Chapter 12: Human Microbiome Analysis. *Plos Computational Biology*, *8*(12), e1002808. https://doi.org/10.1371/journal.pcbi.1002808. Retrieved 23/05/2021, from https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002808

Morton, J., Marotz, C., Washburne, A., Silverman, J., Zaramela, L., & Edlund, A. et al. (2019). Establishing microbial composition measurement standards with reference frames. *Nature Communications*, *10*(1). doi: 10.1038/s41467-019-10656-5. Retrieved 02/03/2021, from https://www.nature.com/articles/s41467-019-10656-5

NIH Human Microbiome Portfolio Analysis Team. (2019). A review of 10 years of human microbiome research activities at the US National Institutes of Health, Fiscal Years 2007-2016. *Microbiome*, *7*(1). doi: 10.1186/s40168-019-0620-y. Retrieved 22/05/2021, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6391833/

Rivera-Pinto, J., Egozcue, J., Pawlowsky-Glahn, V., Paredes, R., Noguera-Julian, M., & Calle, M. (2018). Balances: a New Perspective for Microbiome Analysis. *Msystems*, *3*(4). doi: 10.1128/msystems.00053-18. Retrieved 26/05/2021, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6050633/

Scholz, M. *Metagenomics - Alpha and beta diversity*. Metagenomics.wiki. Retrieved 28/05/2021, from https://www.metagenomics.wiki/pdf/definition/alpha-beta-diversity.

Singh, B., & Trivedi, P. (2017). Microbiome and the future for food and nutrient security. *Microbial Biotechnology*, *10*(1), 50-53. doi: 10.1111/1751-7915.12592. Retrieved 22/05/2021, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5270726/

Stephenson, I. (2015). *What is Beta Diversity?*. Methods Blog. Retrieved 28/05/2021, from https://methodsblog.com/2015/05/27/beta_diversity.

van Nood, E., Vrieze, A., Nieuwdorp, M., Fuentes, S., Zoetendal, E. G., de Vos, W. M., Visser, C. E., Kuijper, E. J., Bartelsman, J. F., Tijssen, J. G., Speelman, P., Dijkgraaf, M. G., & Keller, J. J. (2013). Infusion of Feces for Recurrent Clostridium difficile. *New England Journal Of Medicine*, *368*(22), 2143-2145. doi: 10.1056/nejmc1303919. Retrieved 22/05/2021, from https://www.nejm.org/doi/10.1056/NEJMoa1205037?url_ver=Z39.88-2003&rfr_id=ori%3Arid%3Acrossref.org&rfr_dat=cr_pub++0www.ncbi.nlm.nih.gov

Walshaw, J., Etherington, G., & MacLean, D. (2011). Next-generation sequencing approaches to metagenomics. In D. Marco, *Metagenomics: Current Innovations and Future Trends*. Caister Academic Press. Retrieved 23/05/2021, from https://www.researchgate.net/publication/303753018_Metagenomics_Current_Innovations_and_Future_Trends

Weiss, S., Xu, Z., Peddada, S., Amir, A., Bittinger, K., & Gonzalez, A. et al. (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*, *5*(1). doi: 10.1186/s40168-017-0237-y. Retrieved 23/05/2021, from https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-017-0237-y

Whipps, J., Lewis, K., & Cooke, R. (1988). Mycoparasitism and plant disease control. In M. Burge, *Fungi in biological control systems* (1st ed., p. 176). Manchester and New York: Manchester University Press. Retrieved 22/05/2021, from https://books.google.es/books?hl=ca&lr=&id=qoK7AAAAIAAJ&oi=fnd&pg=PR7&ots=ZFZvuuwUQb&sig=c3wqbgcYEBPHjWYUMI3ChGxHjjU&redir_esc=y#v=onepage&q&f=false

Wickelmaier, F. An Introduction to MDS. Retrieved 26/05/2021, from https://www.hongfeili.com/files/paper100/paper4.pdf

Willis, A. (2019). Rarefaction, Alpha Diversity, and Statistics. *Frontiers In Microbiology*, *10*. https://doi.org/10.3389/fmicb.2019.02407. Retrieved 28/05/2021, from https://www.biorxiv.org/content/biorxiv/early/2017/12/11/231878.full.pdf