Marta Alacid Martín

Anàlisi metagenòmica de la microbiota intestinal en un model de ratolí amb ablació d'un receptor acoblat a proteïnes G

Treball Fi de Grau

dirigit per la Dra. Sonia Fernández- Veledo i supervisat per la Dra. Maria del Mar Rodríguez-Peña

Grau en Enginyeria Biomèdica

UNIVERSITAT ROVIRA I VIRGILI

Tarragona

2022

Resum

La microbiota juga un paper fonamental en la salut d'un individu ja que que contribueix a nombroses funcions metabòliques. La introducció de les tècniques de seqüenciació del gen 16s ARNr han suposat una gran revolució en el coneixement de la composició de la microbiota i de la seva implicació en l'estat de la salut i de malalties de l'ésser humà.

En aquest treball es detallen totes les tècniques, tests estadístics i processos d'anàlisi exploratori de dades utilitzats per caracteritzar la microbiota de models de ratolins controls i de models de ratolins sense un tipus de receptor acoblat a proteïnes G (GPCR). Amb aquesta recerca hem pogut investigar l'efecte que podia tenir l'ablació del GPCR en la microbiota d'un conjunt de ratolins. Per fer aquest estudi, s'ha tingut en compte la composicionalitat de les dades amb les que es treballava, de forma que enlloc d'utilitzar els mètodes tradicionals s'han utilitzat funcions aptes per la naturalesa d'aquestes.

Paraules clau: microbiota, GPCR, 16s ARNr, tests estadístics, composicionalitat de les dades

Índex

1 Introducció

El microbioma es reconeix cada cop més com un component crític en el desenvolupament humà, la salut i la malaltia. És per això que durant els darrers anys s'han realitzat nombrosos estudis destinats a la caracterització de l'estructura, funcionalitat i interacció entre l'intestí i la microbiota de l'hoste, que han revelat una àmplia gama de funcions importants que la microbiota exerceix en la salut humana i animal [1]. Moltes d'aquestes funcions depenen de metabòlits, derivats de la microbiota, que poden actuar al mateix temps tant com a nutrients i com a molècules missatgeres que envien senyals a òrgans distants del cos per configurar la fisiopatologia de l'hoste.

En el cas de l'intestí, el microbioma té nombroses funcionalitats importants com mantenir la integritat de la paret intestinal, evitar el creixement excessiu d'organismes nocius i descompondre i absorbir determinants nutrients [2]. A més, un estudi [3] realitzat per Backhed et al. va demostrar la importància del microbioma en l'obtenció i emmagatzematge de l'energia.

1.1 Metabòlit i el seu receptor

La informació d'aquesta secció s'ha omès per raons de confidencialitat.

1.2 ARN ribosòmic 16S

Per tal de realitzar estudis taxonòmics o de filogènia en els diferents gèneres i espècies bacterianes, l'anàlisi del ribosomal àcid ribonucleic (rRNA) 16S constitueix un gran identificador de bactèries que ofereix una identificació bastant precisa a nivell d'assignació de gènere i espècie. No obstant, com veurem al llarg d'aquest treball, l'alta homologia genètica en alguns gèneres bacterians o un recent canvi en la seva assignació taxonòmica, fa que l'rRNA 16s, en alguns casos, no ens permeti la identificació a nivell d'espècie [5].

L'16s rRNA és un polirribonucleòtid inclòs en la subunitat 30S del ribosoma bacterià de tots els procariotes que està codificat per l'àcid desoxiribonucleic (DNA) ribosòmic rRNA 16S (DNA 16S). Aquest actua com una mena de codi de barres que identifica cadascuna de les bactèries.

El primer microbiòleg que va concebre aquesta macromolècula com a molècula principal identificadora de bactèries va ser el nord-americà Carl Richard Woese. Woese va veure en aquesta macromolècula una sèrie de característiques que van fer entendre la importància del rRNA 16s en la identificació bacteriana [6]:

- Present en totes les bactèries actuals: tots els éssers vius necessiten produir proteïnes per tant tots tenen ARN ribosòmic.
- És completa: donat que la seva mida és relativament llarga (1.500 nucleòtids (nt)) hi ha suficient informació per a analitzar i minimitza les fluctuacions estadístiques.
- En més del 90% dels casos la seqüència m'ajuda a determinar almenys el gènere del bacteri.
- Hi ha trams altament conservadors: fàcil reacció en cadena de la polimerasa (PCR) i fàcil seqüenciació capil·lar.
- Escalable: es poden analitzar mils de seqüències 16S.
- Hi ha una alta disponibilitat de seqüències per comparar ja que la seqüenciació del ADNr 16s és relativament fàcil de seqüenciar.
- La seva estructura i funció han estat constants durant un temps molt perllongat, de manera que les alteracions en la seqüència reflecteixen probablement canvis aleatoris.

La seqüència del gen rRNA 16S presenta aproximadament 1.500 parell de bases (pb) i es compon de 9 zones variables V1-V9 separades per regions conservades (Fig. 1). Utilitzem les regions variables del gen per distingir un tipus de bacteris d'un altre. Una de les regions variables seqüenciades amb més freqüència és la regió V3 - V4 que abasta aproximadament 469 pb [7].

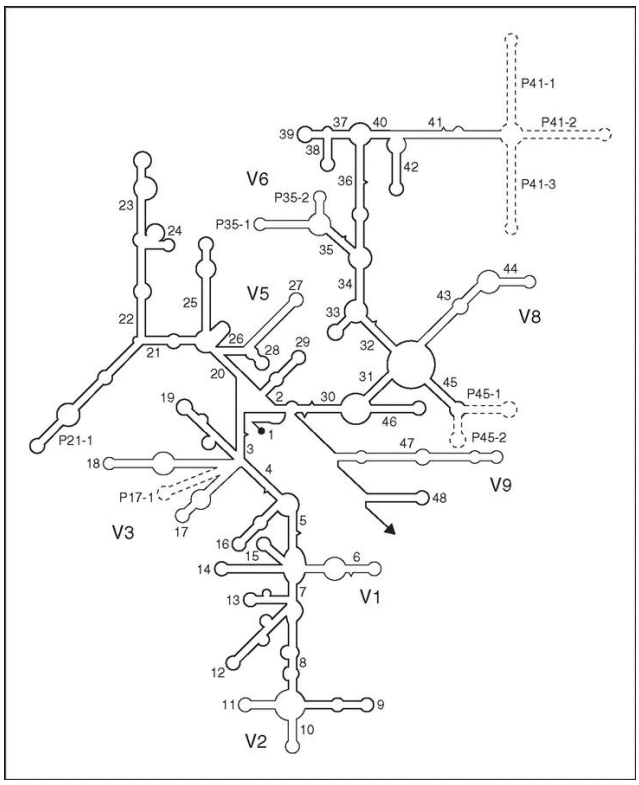

Figura 1. Estructura secundaria del ARNr 16S [8].

En la següent figura (Fig. 2), veiem diferenciades les regions constants; les zones en blau són les regions conservades que podem trobar en qualsevol bacteri, i les regions variables que són les representades amb color taronja.

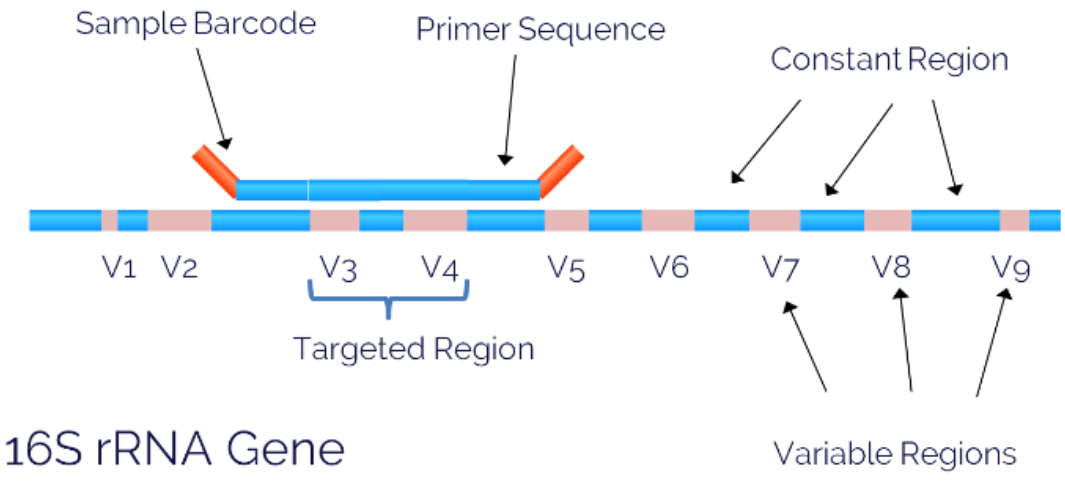

Figura 2. Estructura secundaria del ARNr 16S representada linealment [9].

2 Objectiu

L'objectiu d'aquest treball és avaluar el paper que juga el receptor acoblat a proteïnes G (GPCR) en la regulació de la microbiota intestinal d'un individu. Per veure-ho, hem analitzat la microbiota de dos grups de ratolins: els ratolins controls no modificats genèticament, anomenats *wild type* (WT), i els ratolins modificats genèticament, els *knockout* (KO) del GPCR i que per tant expressen aquest tipus de receptor.

Per tal de determinar les diferències que pot haver-hi en la microbiota dels dos grups, hem dissenyat un programa implementat amb R on hem realitzat diferents tests estadístics.

3 Dades

La informació d'aquesta secció s'ha omès per raons de confidencialitat.

4 Pre-processament de les dades: DADA2

Per poder dur a terme l'estudi de la microbiota és important que les nostres dades d'entrada (*input data*) estiguin ben estructurades i tinguin un alt factor de qualitat. Per garantir això, vam utilitzar el paquet DADA2 que ens ofereix Bioconductor.

El paquet DADA2 infereix variants de seqüències d'amplicons (ASV) exactes a partir de dades de seqüenciació d'amplicons d'alt rendiment, reemplaçant l'enfocament d'agrupament d'OTU (Unitat Taxonòmica Operacional) més gruixut i menys precís.

Aquest mètode tradicional d'OTU assigna unitats taxonòmiques específiques a seqüències basant-se amb el percentatge de semblança entre seqüències de nucleòtids [10]. A diferència d'aquest, DADA2 és més sensible i específic ja que detecta la variació biològica real i permet així una major resolució, millor estimació d'abundàncies relatives, etc. [11].

La canalització DADA2 pren com a entrada fitxers *fastq* demultiplexats i genera les variants de seqüència i les seves abundàncies de mostra després d'eliminar els errors de substitució i quimera.

Per poder utilitzar aquest paquet és imprescindible que les dades d'entrada compleixin les següents característiques [12]:

- Les mostres han d'estar desmultiplexades, és a dir, dividides en fitxers *fastq* individuals per mostra.
- S'han d'haver eliminat els nucleòtids no biològics, per exemple, encebadors, adaptadors, enllaçadors, etc.
- Si es tracta de dades de seqüenciació per parells, els fitxers *fastq* d'avanç i retrocés contenen lectures en el mateix ordre.

El producte final és una taula de variants de la seqüència de l'amplicó (ASV), semblant a la taula tradicional d'OTU de més resolució, que registra el nombre de vegades que es va observar cada variant exacta de la seqüència de l'amplicó a cada mostra. També assignem la taxonomia a les seqüències de sortida, i demostrem com les dades poden ser importades al paquet R phyloseq per a l'anàlisi de dades del microbioma.

Per treballar DADA2, és necessària la seva instal·lació. Aquesta instal·lació la vam dur a terme a través del repositori de paquets de Bioconductor [13]. Bioconductor és un projecte de programari de codi i de desenvolupament obert molt utilitzat per a l'anàlisi i comprensió de les dades procedents de l'experimentació d'alt rendiment en genòmica i biologia molecular [14]. En aquest cas, per tal de fer servir DADA2, és requeria la versió R 4.0.0 i la versió 3.11 de Bioconductor.

En DADA2 és important treballar els *reads* per separat, és a dir, per una banda hem de tenir les copies directes (*forwards*) o R1 i per l'altra les copies inverses (*reverse*) o R2. Per tant, el primer que vam fer va ser carregar els arxius *fastq* de seqüenciació, ordenar-los, separar-los i guardar-los a dos llistes per separat.

4.1 Perfil de qualitat de les mostres

Un cop vam tenir les dades ben organitzades, DADA2 ens va permetre visualitzar el perfil de qualitat de les mostres que volíem mitjançant la funció *plotQualityProfile().* Aquest tipus de funció genera unes gràfiques que mostren la qualitat de la seqüencia al llarg de la lectura d'Illumina.

El *quality score* és una puntuació que mesura la probabilitat de que una base es cridi incorrectament [15]. La puntuació de qualitat de la seqüenciació d'una base determinada, Q, es defineix per l'equació (1) [16]:

$$Q = -10log_{10}(e) \quad (1)$$

On e és la probabilitat estimada que la trucada base sigui incorrecta.

Aquesta Q s'interpreta de la següent forma:

- Les puntuacions Q més altes indiquen una menor probabilitat d'error.
- Les puntuacions Q més baixes poden fer que una part important de les lectures sigui inutilitzable. També poden provocar un augment de les trucades de variants falses positives, donant lloc a conclusions inexactes.

En la següent taula (Taula 1) podem veure la relació que hi ha entre les puntuacions de qualitat i la probabilitat de tenir bases incorrectes:

- Una puntuació de qualitat de 20 (Q20) representa una taxa d'error d'1 entre 100 (és a dir, que cada lectura de seqüenciació de 100 pb pot contenir un error), amb una precisió de trucada corresponent del 99%.
- Quan la qualitat de la seqüenciació arribi a Q30, pràcticament totes les lectures seran perfectes, sense errors ni ambigüitats. És per això que el Q30 es considera un referent de qualitat en la seqüenciació de nova generació (NGS).

Relationship Between Sequencing Quality Score and Base Call Accuracy

Quality Score	Probability of Incorrect Base Call	Inferred Base Call Accuracy
10 (Q10)	1 in 10	90%
20 (Q20)	1 in 100	99%
30 (Q30)	1 in 1000	99.9%

Taula 1. Relació entre *Quality Score* i precisió de les bases [16].

Els resultats d'aquesta secció s'han omès per raons de confidencialitat.

4.2 Filtrar i retallar les seqüències

Per tal de millorar aquest factor de qualitat vam retallar les seqüències *forward* i *reverse*. Per saber a partir de quin nucleòtid havíem de retallar-les, vam tenir en compte la mida de la regió V3-V4, que, com bé hem dit abans, aquesta regió abasta aproximadament 469 bp. Per tant, la suma dels nucleòtids havia de complir la següent condició (2).

$$nF + nR \cong 469 \quad (2)$$

On nF és el número de nucleòtid a partir del qual retallarem la seqüencia *forward* i el nR és el número de nucleòtid a partir del qual retallarem la seqüencia *reverse*.

Per veure quin era el millor filtre, vàrem provar diferents valors que complien la condició anterior i que, a més, concordaven amb les gràfiques de perfil de qualitat, ja que no tindria

sentit retallar la seqüència a partir d'un nucleòtid on observem que la lectura encara té un factor de qualitat *Q* major a 30.

Els resultats d'aquesta secció s'han omès per raons de confidencialitat.

4.3 Model d'errors

Un cop filtrades les mostres, mitjançant algoritmes d'aprenentatge automàtic (*Machine Learning*), es va crear un model d'error que preveia quines variacions de seqüències podien ser biològiques i quines no, de manera que es podia concloure quantes seqüències reals havia inferit.

Els resultats d'aquesta secció s'han omès per raons de confidencialitat.

4.4 Desreplicació

Per tal de disminuir la redundància i avançar eficientment, vam desreplicar les seqüències. Aquest procés consisteix en combinar totes les lectures de seqüenciació idèntiques en seqüències úniques amb una abundància corresponent igual al nombre de lectures amb aquesta seqüència única [12].

4.5 Alineament seqüències (solapament o assemblatge de pars de lectures)

Un cop vam tenir les seqüències *forward* i *reverse* sense soroll ni redundància, les vam fusionar. Aquesta fusió es realitza alineant les lectures directes (*forward*) eliminades de soroll amb el complement invers de les lectures inverses eliminades de soroll corresponents i després construint les seqüències combinades [12].

L'objectiu de la fusió és convertir un par de lectures en una única que contingui una seqüència i un conjunt de puntuacions de qualitat, per tant, genera un únic fitxer *fastq*.

Per tal de realitzar aquesta fusió (Fig. 3), s'alineen la seqüència de la lectura directa amb el complement invers de la seqüència de la lectura inversa. En la regió de solapament en la que ambdues lectures cobreixen les mateixes bases, s'obté una única lletra i puntuació Q a partir de l'alineat de lletres i puntuacions Q per a cada base. Si la lectura directa i la inversa coincideixen en la trucada de la base, això augmenta la confiança en la base predita, incrementant la puntuació Q. Pel contrari, si les lectures no estan d'acord, això redueix la confiança en la trucada de la base i disminueix la puntuació Q. Les puntuacions Q ajustades per a les coincidències i les discordances es calculen mitjançant l'estadística Bayesiana [17].

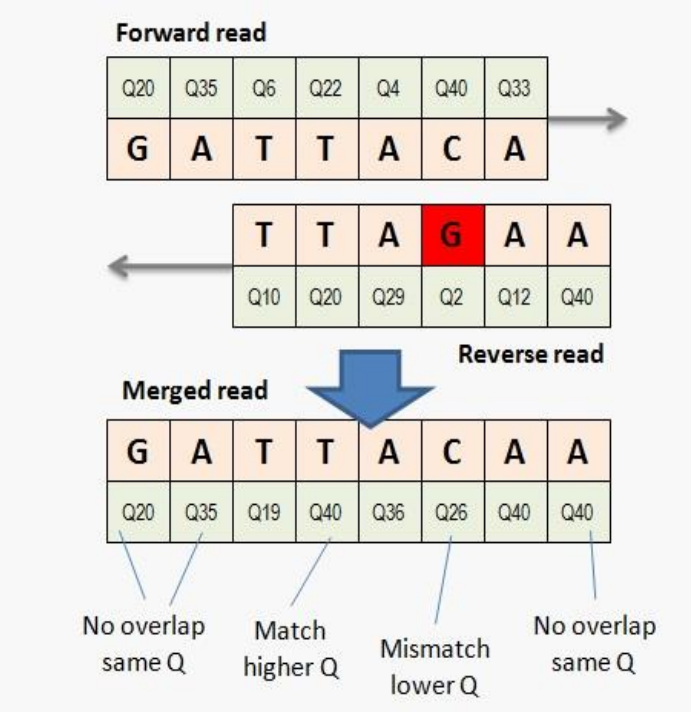

Figura 3. Alineació seqüència directa amb la seva complementària inversa

Aquesta alineació la vam realitzar mitjançant la funció *mergePairs()* la qual combina cada parell de lectures directes i inverses sense soroll, rebutjant els parells que no es superposen prou (s'han de superposar almenys 12 bases) o que continguin massa (>0 per defecte) desajustos a la regió de superposició [18].

Els resultats d'aquesta secció s'han omès per raons de confidencialitat.

4.6 Assignació taxonòmica

Per tal de realitzar la classificació taxonòmica, el primer que vam fer va ser la taula de seqüències. Per fer-ho, vam utilitzar la funció que ens oferia el mateix paquet DADA2 de *makeSequenceTable().* Aquesta funció construeix una taula de seqüències (semblant a una taula OTU) a partir de la llista de mostres proporcionada [19].

Amb aquesta funció es crea una matriu amb files que corresponen a les mostres, i columnes que corresponen a les variants de seqüència.

No obstant això, aquesta taula de seqüències no es podia utilitzar per a realitzar l'assignació taxonòmica ja que, durant els processos anteriors podia haver estat possible l'aparició de seqüències quimèriques, és a dir, seqüències que no són producte real de l'amplificador del gen 16S [20]. Per tant, s'havien d'eliminar aquestes quimeres de les seqüències. Per realitzar-ho, vam utilitzar la funció *removeBimeraDenovo()* que ofereix el paquet DADA2.

Quan ja no teníem seqüències quimèriques, vam obtenir una nova matriu però amb menys ASVs ja que s'havien eliminat aquestes seqüències irreals. A més, vam inspeccionar la distribució de les longituds de les seqüències i vam veure que la majoria de les seqüències tenien unes longituds dins el rang esperat per a aquest amplicó V4 (400- 460).

Un cop obtinguda la taula amb totes les seqüències sense les quimeres, vam assignar la taxonomia a cada seqüència. Per fer-ho, el paquet DADA2 proporciona dos mètodes. El primer mètode, *assignTaxonomy(),* utilitza el mètode de classificació bayesià ingenu de Wang et al.

2007 per assignar taxonomia a diversos rangs (p. ex., regne a gènere) [21]. Aquest mètode pren com a entrada un conjunt de seqüències per classificar-se, i un conjunt d'entrenament de seqüències de referència amb taxonomia coneguda, i produeix assignacions taxonòmiques [12]. El segon mètode, *assignSpecies()*, fa assignacions a nivell d'espècie a les dades 16S.

Tant *assignTaxonomy()* com *assignSpecies()* fan assignacions taxonòmiques comparant-les amb una base de dades on hi ha un conjunt de seqüències amb la seva taxonomia coneguda. En el nostre cas, la base de dades que vam utilitzar va ser SILVA, ja que ofereix uns conjunts de dades exhaustius, de qualitat i actualitzats periòdicament de seqüències d'ARN ribosòmic 16S per als bacteris.

Els resultats d'aquesta secció s'han omès per raons de confidencialitat.

4.7 Producte final del DADA2

Finalment, després de tots aquesta passos, vam obtenir diferents taules:

- Taula de variants de la seqüència de l'amplicó (ASV).
- Taula on estan les seqüències amb la seva taxonomia.
- Disseny experimental.

Totes aquestes taules les vam importar al paquet R phyloseq per tal de poder realitzar l'anàlisi d'aquestes.

No obstant, degut que aquestes mostres es van enviar inicialment a un servei de seqüenciació extern, vam utilitzar l'objecte phyloseq que es va obtenir amb el seu preprocessament.

5 Dades composicionals

Per poder tractar correctament el conjunt de dades del microbioma obtingut després de la seqüenciació és important entendre la seva naturalesa i tractar-les com a tal en totes les etapes de l'anàlisi.

Durant molts anys s'han estat utilitzant metodologies com la refracció, distancia Bray-Curtis, anàlisi de les coordenades principals (PCoA), abundància, etc. que no tenien en compte la naturalesa de les dades del microbioma.

Després de diferents estudis s'ha observat que aquestes dades es descriuen com proporcions o probabilitats, és a dir, són dades composicionals. Per això, des de 2017 s'han començat a utilitzar noves tècniques que substitueixen les metodologies tradicionals.

En aquest treball es va pretendre tenir en compte la composicionalitat de les dades i, per tant, realitzar anàlisis adequats per aquest tipus de dades [22].

En la següent figura (Fig. 4) veiem diferents propostes de noves metodologies [23] que substitueixen a les estàndards per tal d'adaptar-se a la naturalesa de les dades.

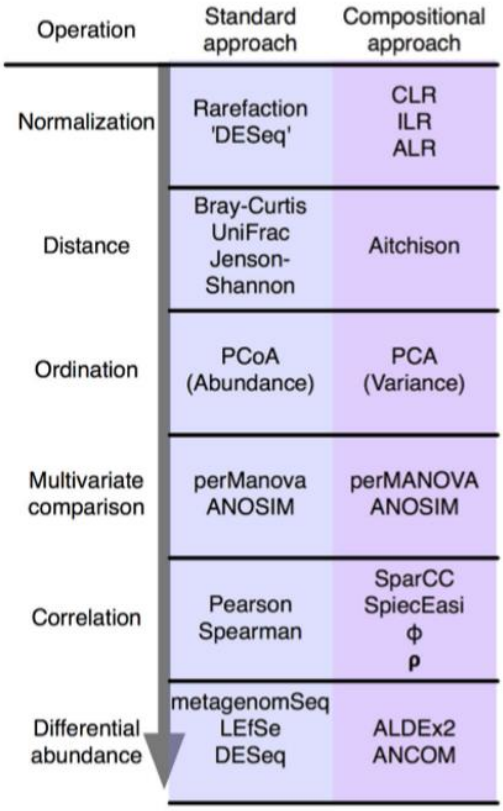

Figura 4. Anàlisi estàndard de les dades i anàlisi proposat per a dades composicionals [23].

6 Filtrar objecte phyloseq

Com hem vist anteriorment, amb DADA2 vam obtenir un objecte phyloseq. No obstant, per poder fer l'anàlisi exploratori de les nostres dades a partir d'aquest objecte phyloseq, és important filtrar-lo.

Durant el procés de filtratge es van eliminar els taxes que no se'ls ho havia assignat un nom, és a dir, aquells taxes que tenien un valor de zero en totes les mostres i a més es va fer una imputació de zeros.

Per realitzar aquesta imputació de zeros vam utilitzar la funció *cmultRepl()*. Aquesta funció està basada en un algoritme que tracta de decidir quin tipus de zero tenim:

- Zero real: aquest tipus de zero significa que realment hi ha una absència d'aquest taxa.
- Zero de l'instrument de mesura: es produeix quan no s'assoleix el límit de detecció que té l'instrument de mesura i per tant, no es pot detectar el taxa.
- Zero de la mostra: es produeix quan apareix el zero a la mostra però no ha caigut a l'instrument de mesura i per tant no s'ha detectat. Aquest és el zero més complicat de detectar.

A més, als zeros que la funció ha etiquetat com a zeros reals, es realitza un *pseudocount*. Aquest procediment consisteix en afegir una petita quantitat als zeros per tal que després no surtin errors.

D'aquesta forma es va obtenir l'objecte *phyloseq* filtrat, sense zeros, amb el qual vam treballar durant tot l'anàlisi d'aquest.

7 Anàlisi exploratori de les dades

Per tal d'explorar les dades d'una manera no supervisada, vam utilitzar el mètode d'ordinació PCoA, a diferents rangs taxonòmics (*phylum*, *family* i *genus*). En aquest cas, atesa la composicionalitat de les dades, era necessari transformar-les amb les transformacions *clr* perquè passessin de residir d'un espai simple a un espai euclidià. D'aquesta forma, amb aquest mètode d'ordinació es veu com les mostres poden formar grups o clústers a partir de la variabilitat que hi ha a les dades.

Tots els resultats d'aquesta secció s'han omès per raons de confidencialitat.

7.1 Transformació *clr*

Abans de començar a fer l'anàlisi exploratori de les dades era important transformar les dades. La elecció d'aquesta transformació depèn del tipus de dades que tenim. En aquest cas, com ja hem explicat anteriorment, les dades que teníem eren composicionals, per tant, era necessari fer una transformació logarítmica [24].

El propòsit de transformar les dades és:

- Propòsit des d'un punt de vista estadístic: es fonamenta en la millora de la normalitat, homogeneïtat de variància i independència.
- Propòsit des d'un punt de vista ecològic: busca que les distàncies ecològiques treballin millor; per tant, redueixen l'efecte de la quantitat total en les unitats de mostreig per a enfocar-se en les quantitats relatives.

La transformació logarítmica s'aplica quan hi ha un alt grau de variació (presència o absència d'espècies), variància està correlacionada positivament amb la mitjana (això ho miraríem a través de l'anàlisi de correlació) i quan a les observacions amb 0 hem d'afegir la unitat (1) al variable. L'objectiu d'aquesta transformació és reduir la sessió de les dades i centrar-les [25] mitjançant l'aplicació de la fórmula (3).

$$clr = \log 10 \left(\frac{x\_r}{g(x\_r)}\right) = log10\, x_r - log10\mu\_r \quad (3)$$

On x_{r} és un únic valor relatiu, g(x_r) és la mitjana geomètrica dels valors relatius de tota la mostra i μ_{r} és la mitjana aritmètica dels valors relatius de tota la mostra.

Per fer aquesta transformació, vam utilitzar la funció *transformCounts()* que ens ofereix R [25]. Aquesta funció transforma les dades d'abundància, mitjançant el mètode, en aquest cas, de transformació logarítmica (*method ="clr"*) per reduir la asimetria de les dades i centrar-les.

7.2 Càlcul distàncies euclidianes

Un cop vam tenir les dades les dades transformades, vam procedir a calcular les distàncies euclidianes.

Una distancia euclidiana entre dos punts A i B es defineix com la longitud del segment que uneix els dos punts. Per tal de calcular-la s'utilitza el Teorema de Pitàgores (4) (Fig. 5).

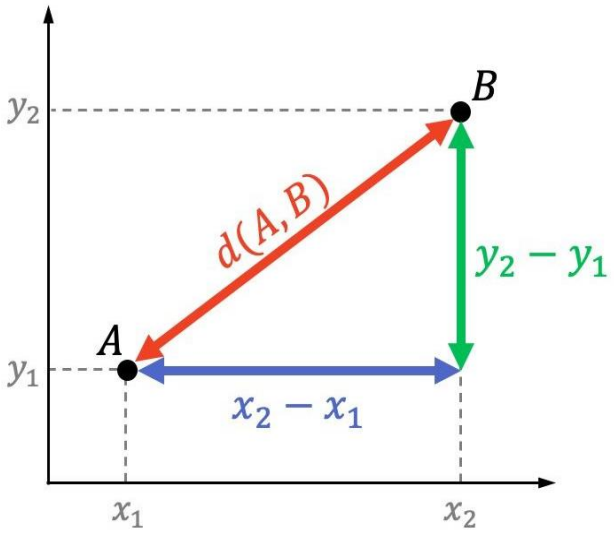

Figura 5. Representació càlcul distància euclidiana entre dos punts a partir de Pitàgores. A és un punt definit per les coordenades (x_1, y_1) i B és un altre punt definit per les coordenades (x_2, y_2) [26]

$$Deuclideana(A,B) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (4)$$

No obstant, en el nostre cas, a través del paquet *vegan* vam poder calcular, de forma automàtica, les distàncies euclidianes que hi havia entre totes les mostres gràcies a la funció *vegidst()* que ofereix aquest paquet. Aquesta funció ens permet introduir el mètode amb el qual volem calcular les distàncies (Bray Curtis, Euclidean) que utilitzarem per realitzar el PCoA i dóna com a resultat una matriu de distàncies.

Un cop calculades les distàncies euclidianes, vam procedir a calcular les coordenades de les mostres mitjançant la funció *pco().* Aquest mètode s'utilitza per explorar i visualitzar semblances o diferències entre les mostres partint d'un matriu d'entrada que fa referència a la matriu de distàncies creada anteriorment i assigna a cada element unes coordenades, és a dir, una ubicació, en un espai euclidià. Finalment, aquestes coordenades les vam guardar a un *dataframe* per tal de visualitzar-les.

7.3 Eigenvalue

Amb aquestes distàncies calculades, vam poder calcular els *eigenvalues*. Els *eigenvalues* [27] fan referència a la direcció que té cada component i representen la magnitud de la informació o variació que té cada component principal (PC). Per exemple, en el cas que l'angle fos de 0 graus respecte la component 1 (PC1) significaria que té la mateixa direcció que la PC1 i per tant explica molt bé aquesta component. La importància dels PC disminueix progressivament i per tant es pot concloure que els dos component principals que representen més informació són el PC1 i PC2.

Per calcular els *eigenvalues* de tots els components vam utilitzar un bucle on, dins d'aquest, es va aplicar la fórmula (5).

$$DEigenvalue = \frac{distància\ euclideana\ d'una\ mostra}{\sum distàncies\ euclideanes\ de\ totes\ les\ mostres} \quad (5)$$

Amb aquest codi vam obtenir una *dataframe* amb tots els eigenvalues de cada PC, que vam poder visualitzar amb la funció *ggplot()* per veure de forma més visual la distribució d'aquestes variació al llarg dels components principals.

7.4 Visualització PCoA

Finalment, amb totes aquestes dades, mitjançant la funció ggplot() i afegint la geometria de punt *(+geom_point())* vam poder visualitzar els PCoA als diferents rangs taxonòmics: *phylum*, *family* i *genus*. Degut que vam utilitzar el mètode de distàncies euclidianes, el PCoA resultant produeix el mateix resultat que produiria el PCA sobre el mateix conjunt de dades.

Amb aquestes gràfiques vam poder observar si les mostres apareixien separades o no per tal de determinar l'existència de diferències entre la microbiota dels dos grups, tant a nivell de *phylum*, *family* i *genus*.

8 Diversitat

A més de l'anàlisi exploratori de dades, en aquest estudi vam estudiar tant la diversitat biològica a nivell local (diversitat alfa), com les diferències entre les comunitats biològiques locals que hi havia la regió (diversitat beta) [28].

Tots els resultats d'aquesta secció s'han omès per raons de confidencialitat.

8.1 Diversitat Alfa

La diversitat alfa busca quantificar la diversitat microbiana que hi ha en cada individu. Per tal de calcular-la, existeixen milers de diversitats ecològiques, però la seva idoneïtat ve determinada segons el seu ús.

Els índex de diversitat combinen la riquesa (nombre d'espècies que hi ha) i abundància d'espècies (nombre d'espècies que hi ha realment) en un únic valor d'uniformitat. Aquests índexs són considerats com mesures de la variància de la distribució de l'abundància de les espècies. Tot i que existeixen molts índex de diversitat (com índex Chao, índex Shannon, Simpson, índex d'uniformitat de Pielou) els índex Simpson i Shannon són els més utilitzats. No obstant, en la majoria d'investigacions relacionades amb l'estudi de les comunitats microbianes [29, 30, 31] s'utilitza l'índex Shannon, també anomenat Shannon-Weaver o Shannon-Wiener.

Les comunitats microbianes que estan dominades per una o poques espècies tindran poca diversitat, és a dir, presentaran poca uniformitat. En canvi, les comunitats que presentin una abundància ben distribuïda entre les diferents espècies, voldrà dir que té una gran diversitat i uniformitat [32].

Aquest índex [33, 34] calculat segons la fórmula (6), es representa com a H', i s'expressa amb valor positiu, que en la majoria dels casos varia entre 0,5 i 5.

$$H' = -\sum_{i=1}^{S} (\frac{q_i}{Q}) \times \ln (\frac{q_i}{Q}) \tag{6}$$

On S és la riquesa d'espècies, q_i és el nombre d'individus de la espècie, i Q, el nombre total d'individus.

En els casos de mostres amb comunitats microbianes que estan dominades per una o poques espècies tindran poca diversitat, és a dir, presentaran poca uniformitat, i aquest índex serà més petit. En canvi, les mostres que presentin una bona distribució d'abundància de les diferents espècies, voldrà dir que té una gran diversitat i uniformitat i per tant, un índex Shannon major [32]. Generalment, els índexs Shannon inferiors a 2 es consideren baixos en diversitat mentre que els majors a 3 són alts en diversitat d'espècies [35].

El paquet *vegan* de R té algunes funcions que ens permeten analitzar la relació espècies-abundància, alguns dels més utilitzats són els models per a la distribució de l'abundància d'espècies (rang d'espècies). Una d'aquestes funcions és *estimateDiversity()* la qual vam utilitzar per calcular la diversitat alfa en el nostre estudi. En aquesta funció vam especificar les mesures de diversitat que volíem calcular, en el nostre cas era l'índex de diversitat Shannon.

A més, per veure si aquestes diversitats alpha eren estadísticament significatives, vam utilitzar un test no paramètric, el que ens permet no haver de donar per fet la condició de normalitat de les dades.

8.2 Diversitat beta

Al llarg de la historia, en la ecologia s'han utilitzat moltes mesures per estimar la diversitat beta. Tot i així, actualment existeixen molts articles [36] que demostren que l'anàlisi de la diversitat beta mitjançant les distàncies euclidianes és el mètode més fàcil i comú degut la seva senzillesa i alta replicabilitat.

A diferència de la diversitat alfa, l'objectiu de la diversitat beta és demostrar mitjançant un criteri estadístic que la microbiota dels dos grups és estadísticament diferent.

Per tal de comprovar si existia diferència significativa entre les distàncies euclidianes vam utilitzar l'anàlisi multivariant tradicional de la variància (*PERMANOVA*) [32]. Aquest test proposa la hipòtesi nul·la de que la microbiota dels dos grups no són diferents. Per comprovar-ho, el test calcula la distància mínima de tots els punts d'un grup al seu centre i després realitza un canvi de taxes aleatòriament per veure si es pot fer millorar la resolució d'aquesta separació entre els grups, és a dir, per mirar si es podria explicar aquesta separació a l'atzar, simplement canviant taxes d'un lloc a un altre. Aquest procés es repeteix varies vegades i finalment, si menys del 5% dels cops no dona una resolució millor significa que la que hi ha ja és estadísticament significativament i no és deguda a l'atzar.

Per realitzar aquest test, R ens ofereix la funció *adonis().* Aquesta funció té com a entrada (*input*) les dades transformades amb *clr* i la condició a la que pertany cada mostra, és a dir, si és WT o KO.

9 Variabilitat

Quan diferenciem dos grups en un espai geomètric és important entendre que el que defineix un grup és que els punts estiguin junts entre ells. A més de fer un test estadístic per veure si un grup és diferent significativament de l'altre, és apropiat veure la variància de cada grup, és a dir, veure si les distàncies al centre de cada grup són més o menys grans. Si aquestes distàncies unes són molt petites i les altres molt grans pot ser que la diferència que abans hem vist que existia entre els grups, sigui deguda a que la variabilitat d'un grup és molt gran.

Per veure la variabilitat que existeix en el nostre estudi, vam representar mitjançant la funció *boxplot()* les diferents distàncies que hi havia entre els punts d'un grup amb el seu centre.

Un cop visualitzades aquestes distàncies, vam realitzar un test ANOVA on comparaven les distàncies mitjanes al centroide de cada grup. D'aquesta forma vam poder concloure si existia diferencia significativa entre la distància de les mostres amb el seu respectiu *centroide*.

Els resultats d'aquesta secció s'han omès per raons de confidencialitat.

10 Abundància diferencial

Finalment, per obtenir una visió general de la composició de les mostres que tenim, vam visualitzar un gràfic de barres on es representava l'abundància relativa de cada mostra. Aquesta abundància relativa fa referència a la relació entre una o més espècies que habiten en un ecosistema determinat [37] i es calcula amb la següent fórmula (7).

$$Abundancia\ relativa = \frac{Nº\ total\ d'espècies(d'ungrup)}{Nº\ total\ d'espècies(de\ tots\ els\ grups)} \quad (7)$$

Tots els resultats d'aquesta secció s'han omès per raons de confidencialitat.

10.1 Visualització abundància relativa

Per tal de visualitzar l'abundància relativa, primer vam fusionar les espècies que tenien la mateixa taxonomia en un determinat rang taxonòmic, mitjançant la funció *tax_glom()*. Aquesta funció té com entrada, la taula d'abundància ASV que hem generat anteriorment.

Un cop vam tenir aquestes espècies fusionades, amb la funció *ps_melt()*, vam convertir l'objecte *phyloseq* en una taula (*data* frame) per tal de poder fer el gràfic amb *ggplot2*. D'aquesta forma vam poder visualitzar l'abundància relativa de totes les bactèries a diferents rangs taxonòmics (*phylum, family* i *genus*) que hi ha a cada mostra. .

A més, per comparar d'una forma més visual l'abundància relativa dels dos grups, vam considerar interessant visualitzar també l'abundància relativa mitjana dels dos grups.

No obstant aquestes visualitzacions, aquest procés no ens permet determinar si realment un taxó és estadísticament significatiu o no, sinó que només ens permet fer una suposició visual.

10.2 Anàlisi ANCOM

Per analitzar l'abundància diferencial vam utilitzar ANCOM per comparar la composició del microbioma als diferents nivells taxonòmics (*phylum*, *genus* i *family*) als dos grups i veure si les diferències que havíem observat a les gràfiques anteriors eren estadísticament significatives o no.

L'objectiu de l'ANCOM [38] és veure com una mateixa *feature* (quan parlem de feature, en aquest estudi ens referim a un tipus de bacteria) canvia en els dos grups (Fig. 16), és a dir, hem d'arribar a poder representar la *feature* d'una mostra a partir de la mateixa feature de l'altra mostra.

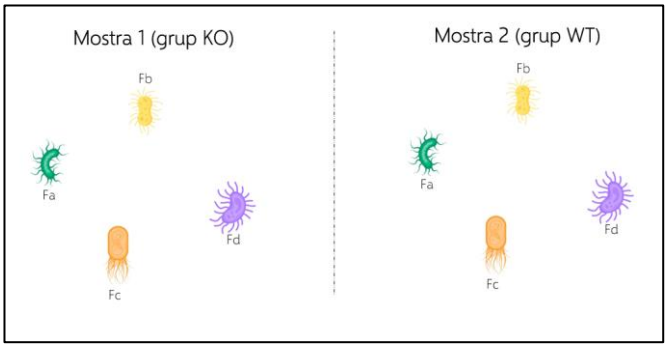

Figura 6. Representació genèrica de les *features* que podria haver en cadascuna de les mostres

Totes aquestes *features* (Fa, Fb...) es relacionen entre elles gràcies a la profunditat de la seqüencia (N) mitjançant la equació (8).

$$Fa + Fb + \cdots + Ff = N \tag{8}$$

Aquesta relació fa que les *features* de cada mostra depenguin de la seva profunditat de seqüenciació (9) (10), per tant no podrem establir una relació directa entre les *features* de dues mostres.

$$Fa = Pa \times N1 \tag{9}$$

$$Fb = Pb \times N2 \tag{10}$$

Per tal de solucionar aquest problema, interpretarem les *features* com una proporció de la seva profunditat de seqüenciació (11).

$$\frac{Fa}{Fb} = \frac{Pa \times N1}{Pb \times N1} = \frac{Pa}{Pb} \tag{11}$$

Tot i així, encara no podem establir una relació (12) entre les *features* ja que aquesta no és commutativa.

$$\frac{Fa}{Fd} \neq \frac{Fd}{Fa} \tag{12}$$

Per resoldre aquest problema, observem la relació que hi ha entre els logaritmes d'aquestes dues (13).

$$\log\left(\frac{Fa}{Fd}\right) = -\log\left(\frac{Fd}{Fa}\right) \tag{13}$$

En aquest punt, el que ens interessa saber és com varia, per exemple, la relació Fa/Fd de la mostra 1 (pertanyent al grup KO) amb la mostra 2 (pertanyent al grup WT).

Veiem que les fórmules (14) (15) defineix la relació dels logaritmes.

$$\log\left(\frac{Fa}{Fd}\right) = \alpha + \beta_1 + \varepsilon_R \tag{14}$$

$$\log\left(\frac{Fa}{Fd}\right) = \alpha + \beta_2 + \varepsilon_R \tag{15}$$

On *Fa*, *Fd* són les features, α és la intercepció global (igual a les dues mostres), ε_R és el residual i β_1 i β_2 és l'efecte de la mostra 1 i 2 respectivament. Per tant, no són iguals.

Per veure si realment una *feature* ha canviat en les dues mostres, hem de comparar els valor β_1 i β_2

- Si són iguals significa que la proporció *Fa/Fd* de la mostra 1 no ha canviat a la mostra 2.

- Si són diferents significa que tenim una proporció significativament diferent. Si la diferència en β_2 en comparació amb β_1 és negativa significarà que o bé *Fd* ha augmentat o *Fa* ha disminuït.

Per determinar quina és la més significativa, ANCOMBC fa totes les comparacions entre totes les *features* de les mostres. D'aquesta forma s'obtenen diferents puntuacions W (W són les connexions significatives que té cada *feature*) que ens faran veure les *features* que realment són diferencialment abundants i en quina direcció.

Per aplicar aquest mètode, R ofereix el paquet ANCOMBC. Aquest paquet ens permet normalitzar les dades d'abundància microbiana observada a causa de fraccions de mostreig desiguals entre les mostres, i identificar els tàxons que són diferencialment abundants respecte a la covariable d'interès (per exemple, grups d'estudi) entre dos grups de mostres múltiples o més. [39]. A més, la funció *ancombc()* està pensada per treballar amb dades composicionals (CoDa) ja que, a diferència de mètodes anteriors com DeSeq [40], introdueix un terme de compensació específic de la mostra en un marc de regressió lineal. Aquest terme serveix com a correcció del biaix, i el marc de regressió lineal a escala logarítmica és semblant a la transformació logarítmica que hauríem d'utilitzar per tractar la composicionalitat de les dades del microbioma [41].

A més, diferents estudis [42] han demostrat que ANCOMBC en permet identificar els falsos positius com a diferentment abundants. Per tant, podem concloure que aquest mètode és una molt bon candidat per realitzar l'anàlisi d'abundància diferencial de les nostres dades.

Aquesta funció té com a paràmetre d'entrada l'objecte phyloseq i com a resultat, diferents paràmetres que ens permetran determinar els tàxons (bactèries) que són diferencialment abundants:

- *Beta*: variable que ens indica com podem explicar un paràmetre a partir de l'altre
- *Se*: errors estàndards.
- *W*: connexions significatives.
- *Result*: ens diu si hi ha diferencia significativa, és a dir, si el taxó és diferencialment abundant (TRUE) o no (FALS).
- *Pvalue*: aquest valor també ens permet determinar si els tàxons són diferencialment abundants o no:
  - pvalue < 0.05 = hi ha diferencia significativa = taxons són diferencialment abundants.
  - Pvalue >=0.05 = no hi ha diferencia significativa = taxons són iguals en els dos grups.
- *Qvalue*: fan referència als valors p ajustats.
- *Prevalence*: ens diu en quantes mostres apareix aquest grup.
- Els resultats d'aquesta secció s'han omès per raons de confidencialitat.

11 Futurs estudis

La informació d'aquesta secció s'ha omès per raons de confidencialitat.

12 Conclusions

El fet d'haver tingut en compte la composicionalitat de les dades al llarg del nostre estudi, ha comportat uns resultats més fiables que els que s'haguessin pogut obtenir mitjançant les tècniques tradicionals.

Aquests resultats van concloure que existia diferència significativa entre la microbiota dels dos grups, el que suggereix que aquest GPCR juga un paper important en la microbiota.

Les conclusions reals s'han omès per raons de confidencialitat.

13 Referències

[1] Dhariwal A, Chong J, Habib S, King IL, Agellon LB, Xia J. MicrobiomeAnalyst: a web-based tool for comprehensive statistical, visual and meta-analysis of microbiome data. Nucleic Acids Res. 2017 Jul 3;45(W1):W180-W188. doi: 10.1093/nar/gkx295. PMID: 28449106; PMCID: PMC5570177.

[2] Fayfman, M., Flint, K. & Srinivasan, S. Obesity, Motility, Diet, and Intestinal Microbiota—Connecting the Dots. Curr Gastroenterol Rep 21, 15 (2019). https://doi.org/10.1007/s11894-019-0680-y

[3] Bäckhed F, Ding H, Wang T, Hooper LV, Koh GY, Nagy A, Semenkovich CF, Gordon JI. The gut microbiota as an environmental factor that regulates fat storage. Proc Natl Acad Sci U S A. 2004 Nov 2;101(44):15718-23. doi: 10.1073/pnas.0407076101. Epub 2004 Oct 25. PMID: 15505215; PMCID: PMC524219.

[4] Wu, Q., Liang, X., Wang, K., Lin, J., Wang, X., Wang, P., … Jiang, C. (2021). Intestinal hypoxia-inducible factor 2α regulates lactate levels to shape the gut microbiome and alter thermogenesis. Cell Metabolism, 33(10), 1988-2003.e7. https://doi.org/10.1016/j.cmet.2021.07.007

[5] Pina Pérez, M. (2011). Mètodes d'identifació bacteriana al laboratori de microbiologia. [Treball fi de grau, Universitat de Valencia]. https://mobiroderic.uv.es/bitstream/handle/10550/75672/M%C3%A8todes%20d%C2%B4identificaci%C3%B3%20microbiana%20-%20Laboratori%20Microbiologia.pdf?sequence=1&isAllowed=yMÈTODES

[6] Hablemos de Biología Molecular. (Setembre, 2021). El gen del rRNA 16S como código de barras genético de bacterias [Video]. Youtube. https://www.youtube.com/watch?v=7qdhRnfEbFA&t=2s

[7] Vargas-Albores, F., Ortiz-Suárez, L. E., Villalpando-Canchola, E., & Martínez-Porchas, M. (2017). Size-variable zone in V3 region of 16S rRNA. RNA Biology, 14(11), 1514–1521. https://doi.org/10.1080/15476286.2017.1317912

[8] Rodicio, M. del R., & Mendoza, M. del C. (2004). [Identification of bacteria through 16S rRNA sequencing: principles, methods and applications in clinical microbiology]. Enfermedades Infecciosas y Microbiologia Clinica, 22(4), 238–45. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15056441

[9] Nelly Sélem Mojica; Diego Garfias Gallegos; Claudia Zirión Martínez; Jesús Abraham Avelar Rivas; Aaron Jaime Espinosa; Abel Lovaco Flores; Tania Vanessa Arellano Fernandez (2022, Jan). Metagenomics for Software Carpentry lesson, Jan 2022. Zenodo. https://doi.org/10.5281/zenodo.4285900

[10] García-Mazcorro, J. F., Garza-González, E., Marroquín-Cardona, A. G., & Tamayo, J. L. (2015, August 1). Caracterización, influencia y manipulación de la microbiota gastrointestinal en salud y enfermedad. Gastroenterologia y Hepatologia. Ediciones Doyma, S.L. https://doi.org/10.1016/j.gastrohep.2015.01.004

[11] Callahan, B. McMurdie, J. Holmes, S (2022). Introduction to dada2. https://www.bioconductor.org/packages/devel/bioc/vignettes/dada2/inst/doc/dada2-intro.html

[12] Callahan, B. DADA2 Pipeline Tutorial 1.16. https://benjjneb.github.io/dada2/tutorial.html

[13] Callahan, B. DADA2 Installation. https://benjjneb.github.io/dada2/dada-installation.html

[14] Schadow. (2006). Bioinformatics and Computational Biology Solutions Using R and Bioconductor. Robert Gentleman, Wolfgang Huber, Vincent J. Carey, Rafael A. Irizarry and Sandrine Dudoit(Ed.), Briefings in Bioinformatics, 8(2), 136–137. https://doi.org/10.1093/bib/bbl020

[15] Thomas, G. W. C., & Hahn, M. W. (2019). Referee: Reference assembly quality scores. Genome Biology and Evolution, 11(5), 1483–1486. https://doi.org/10.1093/gbe/evz088

[16] *Experiments & Protocols: sequencing Quality scores.* Retrieved February 24, 2022, from https://www.illumina.com/science/technology/next-generation-sequencing/plan-experiments/quality-scores.html

[17] Merging paired reads. Retrieved March 1, 2022, from https://drive5.com/usearch/manual7/merge_pair.html

[18] *MergePairs: Merge denoised forward and reverse reads.* (n.d.). Retrieved March 1, 2022, from https://www.rdocumentation.org/packages/dada2/versions/1.0.3/topics/mergePairs

[19] *MakeSequenceTable: Construct a sample-by-sequence observation matrix.* (n.d). Retrieved March 4, 2022, from https://rdrr.io/bioc/dada2/man/makeSequenceTable.html

[20] Gómez, B. Guerrero, A. Metagenómica 16S/18S. Posgrado en Ciencias, CIAD, A.C. Biología Computacional. https://sites.google.com/a/ciad.mx/bioinformatica/home/metagenomica/descontaminacin/quimeras

[21] Taxonomic Assigment. Callahan, B.. https://benjjneb.github.io/dada2/assign.html

[22] Martín-Fernández, J. A., Barceló-Vidal, C., & Pawlowsky-Glahn, V. (1998). Measures of difference for compositional data and hierarchical clustering methods. Proceedings of IAMG, 98, 526–531.

[23] Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., & Egozcue, J. J. (2017, November 15). Microbiome datasets are compositional: And this is not optional. Frontiers in Microbiology. Frontiers Media S.A. https://doi.org/10.3389/fmicb.2017.02224

[24] Oscar Alonso Rodríguez Gracias (2021). Transformación de Datos. [Video]. Youtube. https://www.youtube.com/watch?v=tzbVSgynGBM

[25] *TransformCounts: Transform Counts.* (n.d). Retrieved April 4, 2022, from https://rdrr.io/github/FelixErnst/mia/man/transformCounts.html

[26] *Fórmula de la distancia entre dos puntos (geometría).* (2022). Retrieved April 11, 2022, from https://www.geometriaanalitica.info/formula-de-la-distancia-entre-dos-puntos-geometria-ejemplos-y-ejercicios-resueltos/

[27] Joseph Adewumi. (Mar 26, 2019). Understanding the Role of Eigenvectors and Eigenvalues in PCA Dimensionality Reduction. https://medium.com/@dareyadewumi650/understanding-the-role-of-eigenvectors-and-eigenvalues-in-pca-dimensionality-reduction-10186dad0c5c

[28] Balsega, A., & Gómez-Rodríguez, C. (2019). Diversidad alfa, beta y gamma: ¿cómo medimos diferencias entre comunidades biológicas? Nova Acta Científica Compostelana (Bioloxía), 26, 39–45.

[29] Llanos Villarreal, J. (2013). El estudio de nuestro microcosmos: las comunidades microbianas asociadas al cuerpo humano. Boletín Micológico, 28(2). https://doi.org/10.22370/bolmicol.2013.28.2.878

[30] Zhang, Z., Chen, X., Loh, Y. J., Yang, X., & Zhang, C. (2021). The effect of calorie intake, fasting, and dietary composition on metabolic health and gut microbiota in mice. BMC Biology, 19(1). https://doi.org/10.1186/s12915-021-00987-5

[31] Hill, T. C. J., Walsh, K. A., Harris, J. A., & Moffett, B. F. (2003). Using ecological diversity measures with bacterial communities. FEMS Microbiology Ecology, 43(1), 1–11. https://doi.org/10.1016/S0168-6496(02)00449-X

[32] Xia, Y. Sun, J. Chen, D. (2018). Community Diversity Measures and Calculations. En Jiahua Chen · Ding-Geng Chen (Ed.), Statistical Analysis of Microbiome Data with R (167-190). ICSA Book Series in Statistics

[33] Beisel, J. N., & Moreteau, J. C. (1997). A simple formula for calculating the lower limit of Shannon's diversity index. Ecological Modelling, 99(2–3), 289–292. https://doi.org/10.1016/S0304-3800(97)01954-6

[34] Beisel, J. N., Usseglio-Polatera, P., Bachmann, V., & Moreteau, J. C. (2003). A comparative analysis of evenness index sensitivity. International Review of Hydrobiology, 88(1), 3–15. https://doi.org/10.1002/iroh.200390004

[35] Pla, L. (2006). Biodiversidad: inferencia basada en el índice de shannon y la riqueza. Interciencia, 31(8), 583–590.

[36] Chen, B., He, X., Pan, B., Zou, X., & You, N. (2021). Comparison of beta diversity measures in clustering the high-dimensional microbial data. PLoS ONE, 16(2 February). https://doi.org/10.1371/journal.pone.0246893

[37] EcoVida Sostenible. (2021). Abundancia ecológica, abundancia local y relativa [Video]. Youtube. https://www.youtube.com/watch?v=hAPo5m0C1Nc

[38] QIIME 2. (2021). PD Mice: ANCOM differential abundance testing [Video]. Youtube. https://www.youtube.com/watch?v=A6o2nOnDsJU

[39] *Analysis of compositions of microbiomes with bias correction.* (2003-2022). Retrieved 25 May, 2022, from https://www.bioconductor.org/packages/release/bioc/html/ANCOMBC.html

[40] Love Michael, I. Anders, S. Huber, W. (2022). Analyzing RNA-seq data with DESeq2. https://bioconductor.org/packages/devel/bioc/vignettes/DESeq2/inst/doc/DESeq2.html

[41] Lin, H., & Peddada, S. D. (2020). Analysis of compositions of microbiomes with bias correction. Nature Communications, 11(1). https://doi.org/10.1038/s41467-020-17041-7

[42] Khomich, M., Måge, I., Rud, I., & Berget, I. (2021). Analysing microbiome intervention design studies: Comparison of alternative multivariate statistical methods. PLoS ONE, 16(11 November). https://doi.org/10.1371/journal.pone.0259973