

CÀLCUL DE LA MIDA MOSTRAL A ESTUDIS METAGENÒMICS

**Esteve Llop Vallverdú**

Màster universitari en Bioinformàtica i bioestadística

TFM- Estadística i bioinformàtica

**Antonio Monleon Getino**

**Alexandre Sánchez Pla**

24/05/2017

[Llicència de Creative Commons](http://creativecommons.org/licenses/by-nc-nd/3.0/es/)  
Aquesta obra està subjecta a una llicència de [Reconeixement-NoComercial-SenseObraDerivada 3.0 Espanya de Creative Commons](http://creativecommons.org/licenses/by-nc-nd/3.0/es/)

**FITXA DEL TREBALL FINAL**

|  |  |
| --- | --- |
| **Títol del treball:** | *CÀLCUL DE LA MIDA MOSTRAL A ESTUDIS METAGENÒMICS* |
| **Nom de l’autor:** | *Esteve Llop Vallverdú* |
| **Nom del consultor/a:** | *Antonio Monleon Getino* |
| **Nom del PRA:** | *Alexandre Sánchez Pla* |
| **Data de lliurament (mm/aaaa):** | *05/2017* |
| **Titulació o programa:** | *Màster universitari en Bioinformàtica i bioestadística* |
| **Àrea del Treball Final:** | *TFM-Estadística i bioinformàtica* |
| **Idioma del treball:** | *Català* |
| **Paraules clau** | *Diversitat microbiana, rarefacció, models bayesians* |
| **Resum del Treball (màxim 250 paraules):** *Amb la finalitat, context d’aplicació, metodologia, resultats i conclusions del treball* | |
| L’estudi de dades metagenòmiques suposa un repte a l’hora de definir quin ha de ser el tamany òptim de les mostres per tal de poder observar la riquesa específica present.  La metodologia aplicada en estudis d’ecologia es basen en diferents mètodes, dels quals destaca l’aplicació de models de rarefacció, per tal d’estimar la riquesa específica presen en una comunitat o hàbitat. Prenent com a base el model clàssic de rarefacció, es planteja definir un model basat en mètodes bayesians per fer l’estimació de la riquesa específica en mostres metagenòmiques, sovint sense espècies conegudes però integrades per unitats taxonòmiques funcionals (OTUs).  S’han simulat matrius de dades a partir de distribucions Diritchlet-multinomials, que presenten unes característiques semblants a les matrius metagenomiques. S’ha aplicat el mètode clàssic de rarefacció i un mètode bayesià de rarefacció desenvolupat recentment. S’han definit diferents escenaris en base als paràmetres de la distribució: riquesa específica, nombre de rèpliques i abundància de les OTUs.  El mètode de rarefacció bayesià ha resultat més eficient per mostres amb una elevada riquesa específica, mentre que el mètode clàssic de rarefacció sembla més adient per mostres amb menys diversitat. Per altra banda, el mètode bayesià ha resultat més sensible a les variacions del nombre de rèpliques i a l’abundància. L’esforç de mostreig també ha resultat respondre a aquests paràmetres, sobretot al nombre de rèpliques sent significativament menor quan hi ha poques rèpliques. | |
| **Abstract (in English, 250 words or less):** | |
| Analysing metagenomic data represents a challenge in terms of selecting the optimal sampling size in order to achieve species richness within microbioms.  Methodology suitable for ecological studies is based on several protocols. Among them, rarefaction has become more popular for estimating species richness in communities or environments. Using classical rarefaction methods, the aim is to define a model based on bayesian methods to estimate species richness from metagenomic samples. Those data lacks taxonomical identification, but they use to be compiled within operative taxonomical units (OTUs).  Several matrices have been generated applying a Diritchlet-multinomial distribution, which generates data with similar distribution described in metagenomic matrices. Classical rarefaction methodology and a newly described bayesian rarefaction method have been applied to simulated data. Additionaly, several scenarios were generated by modifying parameters used to generated Diritchlet-multinomial distributions, namely species richness, number of simulations and OTUs’ abundance.  The bayesian rarefaction method has succeed in estimating more efficiently to highly species-richness complexe samples. On the contrary, classical rarefaction was more suitable for less diverse samples. The bayesian method was more sensitive to changes in distribution parameters, mainly to the number of simulations, as well to abundance. Sampling effort also reacts to parameters, basically to number of simulations. The effort to estimate the 95% richness was lower when the number of simulation was also low. | |

**Índex**

[1. Introducció 1](#_Toc318813117)

[1.1 Context i justificació del Treball 1](#_Toc318813118)

[1.2 Objectius del Treball 1](#_Toc318813119)

[1.3 Enfocament i mètode seguit 1](#_Toc318813120)

[1.4 Planificació del Treball 1](#_Toc318813121)

[1.5 Breu sumari de productes obtinguts 2](#_Toc318813122)

[1.6 Breu descripció dels altres capítols de la memòria 3](#_Toc318813123)

[2. Resta de capítols 4](#_Toc318813124)

[3. Conclusions 14](#_Toc318813125)

[4. Glossari 15](#_Toc318813126)

[5. Bibliografia 16](#_Toc318813127)

[6. Annexos 18](#_Toc318813128)

**Llista de figures**

[Figura 1. Exemple de les corbes generades per la funció *MetagenSample.size.H*. El model genera les corbes de rarefacció clàssica i bayesiana a més del model de predicció. També inclou l’histograma d’abundàncies dels diferents OTUs.](file:///C:\LLOP\UOC\TFM\Treball\Memoria_ELLOP.docx#_Toc484606520)

[Figura 2. Distribució dels valors de l’estimació obtinguda per rarefacció clàssica (dalt), bayesiana (mig), i el nombre màxim predit d’espècies (baix) fetes per als diferents valors d’abundància d’OTUs per a una mostra amb un valor inicial de 50 OTUs](file:///C:\LLOP\UOC\TFM\Treball\Memoria_ELLOP.docx#_Toc484606521)

[Figura 3 Distribució dels valors de l’estimació obtinguda per rarefacció clàssica (dalt), bayesiana (mig), i el nombre màxim predit d’espècies (baix) fetes per als diferents valors d’abundància d’OTUs per una mostra amb un valor inicial de 100 OTUs.](file:///C:\LLOP\UOC\TFM\Treball\Memoria_ELLOP.docx#_Toc484606522)

[Figura 4. Distribució dels valors de l’estimació obtinguda per rarefacció clàssica (dalt), bayesiana (mig), i el nombre màxim predit d’espècies (baix) segons el nombre de rèpliques fetes per una mostra amb un valor inicial de 250 OTUs](file:///C:\LLOP\UOC\TFM\Treball\Memoria_ELLOP.docx#_Toc484606523)

[Figura 5. Distribució del valor de l’esforç de mostreig per obtenir el 95 de la riquesa específica estimada per rarefacció bayesiana (dalt) o el 95% de la màxima riquesa específica predita (baix) en base al nombre de rèpliques de la mostra amb un nombre inicial de 50 OTUs.](file:///C:\LLOP\UOC\TFM\Treball\Memoria_ELLOP.docx#_Toc484606524)

# 1. Introducció

## 1.1 Context i justificació del Treball

Aquest TFM vol estudiar com es pot definir la mida mostral òptima en estudis metagenòmics perquè sigui representativa d’una població. I així poder discernir entre poblacions, i també esbrinar quina és la diversitat d’organismes presents en una població. En base a dades disponibles tant per part de l’equip de recerca com a la xarxa, s’aplicaran diferents algorismes i tractaments estadístics que permetin assolir els objectius que es plantegen.

El fet de decidir-me per aquest tema de TFM es deu a que bona part de la recerca que faig se situa en l’àmbit de l’ecologia, bàsicament en l’aplicació de la biodiversitat com a indicadora de condicions ambientals. El tema toca aspectes comuns amb la meva experiència prèvia, si més no en el seu plantejament teòric, però varia l’escala a la qual es treballa.

## 1.2 Objectius del Treball

Els objectius generals que es plantegen són:

1. Aplicar algorismes i rutines per calcular la mida mostral òptima per separar poblacions a partir de dades metagenòmiques.

2. Obtenir la màxima biodiversitat possible quan es tracti d’una única població.

## 1.3 Enfocament i mètode seguit

Les estratègies per dur a terme aquest treball s’han basat en desenvolupar simulacions de dades metagenòmiques a partir de distribucions de probabilitat Diritchlet-Multinomial i de barreges d’aquestes amb altres per imitar les diverses circumstàncies que ens podem trobar amb matrius metagenòmiques. Aquestes matrius de dades han estat estudiades aplicant nous algoritmes descrits per l’equip de recerca del tutor del treball. Posteriorment, un cop validats els algorismes amb les simulacions es procedirà a aplicar-los a matrius de dades disponibles a diverses llibreries o facilitades pel tutor.

## 1.4 Planificació del Treball

Els recursos necessaris per dur a terme aquest treball corresponen a eines informàtiques, tant pel que fa a equipament com a programari, per tal de poder fer les tasques previstes. Les tasques previstes han estat:

a) Recerca bibliogràfica

b) Compilació de matrius de dades metagenòmiques

c) Generació de les simulacions

d) Implementació de les rutines i algorismes a les simulacions

e) Validació dels algorismes en les dades metagenòmiques compilades

f) Elaboració de la memòria final

e) Elaboració de la presentació

La temporització establerta per desenvolupar les tasques es va distribuir al llarg del periode docent.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | MARÇ | ABRIL | | MAIG | | JUNY |
|  | 15-31 | 1-15 | 16-30 | 1-15 | 16-31 | 1-21 |
| a) Recerca bibliogràfica |  |  |  |  |  |  |
| b) Compilació de matrius de dades metagenòmiques |  |  |  |  |  |  |
| c) Generació de les simulacions |  |  |  |  |  |  |
| d) Implementació de les rutines i algorismes a les simulacions |  |  |  |  |  |  |
| e) Validació dels algorismes en les dades metagenòmiques compilades |  |  |  |  |  |  |
| f) Elaboració de la memòria final |  |  |  |  |  |  |
| e) Elaboració de la presentació |  |  |  |  |  |  |

Es van definir una sèrie de fites per poder valorar l’acompliment del calendari establert:

Fita 1 – 5 d’abril de 2017, compilació de dades genòmiques i generació de simulacions

Fita 2 – 10 de maig de 2017, implementació de rutines i algorismes a les simulacions i validació en dades metagenòmiques

Fita 3 – 24 de maig de 2017, entrega de la memòria final

Malauradament, la disponibilitat de temps no ha permès assolir l’acompliment de la temporització i ha calgut fer una revisió i simplificació dels objectius d’aquest treball.

## 1.5 Breu sumari de productes obtinguts

Els resultats que s’esperaven amb aquest projecte es van concretar en tres lliurables:

--. Memòria final

--. Producte, és esperable poder redactar un article que inclogui els resultats més rellevants, així com si s’ha pogut desenvolupar un software corresponent dels algorismes aplicats.

--. Presentació virtual

Els canvis en la disponibilitat temporal per dur a terme aquest treball han suposat que no s’hagi plantejat l‘elaboració de l’article proposat.

## 1.6 Breu descripció dels altres capítols de la memòria

La memòria consta de tres parts ben delimitades. La primera part consistirá en la contextualització dels coneixements previs sobre el tema. La segona part de la memòria se centrarà en la metodologia aplicada per a la generació de simulacions de dades metagenòmiques i la implementació de les rutines i algorismes en aquestes simulacions, amb l’objectiu de determinar quina és la mida mostral òptima. Finalment, el tercer capítol recull els resultats obtinguts, a partir de l’aplicació dels diferents algoritmes sobre les simulacions.

# 2. Resta de capítols

## 2.1.Introducció

A primer vista, la riquesa específica seria el mètode per mesurar la diversitat d’un hàbitat més simple i intuïtiu. La riquesa específica es defineix com el nombre d’espècies que s’observa en un conjunt [1]. El coneixement de la diversitat és un dels elements importants en els estudis d’ecologia. Una de les conseqüències d’això ha estat un especial interès en estimar la riquesa d’espècies, sobretot d’aquells hàbitats amb una elevada diversitat aparent, i sovint rics en grups taxonòmics més mancats d’informació. L’estudi de la diversitat es planteja en base als individus o bé en base a la mostra [2]. Si ens basem en individus, es mesura la riquesa específica, entesa com el nombre d’espècies presents, sovint especificada com a índexos que pretenen compensar l’efecte del motreig en la riquesa. La mesura fonamentada en la mostra ens permet fer l’estimació de la riquesa, amb l‘objectiu de representar la diversitat global d’un hàbitat.

L’estimació de la riquesa específica es pot dur a terme en base a tres possibles aproximacions: corbes d’acumulació d’espècies, aspecte de la distribució de les espècies, altrament coneguts com a mètodes paramètrics, o bé per mètodes no paramètrics [3, 4]. Un dels inconvenients més importants a tenir en compte quan es fa l’estimació de la riquesa específica és la seva dependència respecte de l’esforç de mostreig. S’entén l’esforç de mostreig com el nombre de mostres que cal fer per trobar una nova espècie per a la riquesa. Aquesta informació pot ser important per tal de fer comparacions adequades de la riquesa entre localitats diferents [5].

Enfront de les corbes d’acumulació, que ens permeten inferir sobre la diversitat específica conegut el conjunt d’espècies, les corbes de rarefacció permeten estimar la riquesa d’una mostra a partir de la informació recollida segons el nombre d’espècies que es van observant. Rarefacció requereix que les espècies es distribueixin de manera aleatòria [6], tot i que el més habitual sigui que les espècies tendeixen a agrupar-se. En aquestes circumstàncies la riquesa específica tendeix a ser sobrestimada. Per altra banda, l’estimació pot ser esbiaixada si el mostreig és inadequat o bé les mostres procedeixen de localitats amb distribució de l’abundància molt diferents.

Per tant, un dels factors importants en estudis de diversitat és conèixer la mida de la mostra i quant esforç de mostreig cal fer per tal de poder tenir un estimació el màxim d’acurada. Si estudiem un hàbitat concret la major dificultat rau en saber quantes espècies no han estat observades, i per tant si l’esforç de mostreig ha estat suficient. Si bé aquest punt és fàcilment avaluable en ecosistemes macroscòpics, quan passem a avaluar ecosistemes microbians la metodologia es complica. Per una banda, per la dificultat d’identificar les espècies, fins ara limitat a aquelles espècies capaces de crèixer en medis de cultiu. Però això implicava que no es pogués assolir un coneixement de la diversitat, ja que molts dels organismes presents en aquests hàbitats no són capaços de crèixer en el medis convencials. La solució a aquestes dificultats l’ha aportat en bona part la metagenòmica. La metagenòmica correspon a l’aplicació d’un conjunt de tècniques genòmiques, com ara *whole-genome shotgun* *sequencing* (WGS), *next-generation sequencing* (NGS), o eines basades en microarrays; aixi com un conjunt d’eines bioinformàtiques per poder manipular el volum de dades que generen. Això ha de permetre discernir quin és el conjunt de microorganismes que integren una comunitat microbiana [7, 8, 9, 10].

La unitat taxonòmica utilitzada en metagenòmica difereix del concepte convencial d’espècie. La diferenciació taxonòmica es basa en la seqüència genètica de l’ADN, sobretot ribosòmic, i la semblança amb altres seqüències presents a partir d’un llindar sobre un percentatge de semblança [10]. Això requereix d’un plantejament bioinformàtic prou acurat que permeti fer una diferenciació fiable i una agrupació de les diferents OTUs per poder definir l’abundància de cada una d’elles i poder estimar de manera fidedigna la diversitat del medi estudiat.

La mida mostral segueix sent un problema per resoldre en els estudis metagenòmics. L’elaboració d’inventaris exhaustius de comunitats microbianes resulta impracticable o massa car, i el coneixement de la diversitat present a la microbiota basada en mostrejos metagenòmics requereix d’una estimació molt acurada. Per altra banda, és important saber com una mostra representa de manera fiable la veritable diversitat d’una comunitat.

Determinar la mida mostral és important. Si el nombre de mostres és inadequat, no observarem diferències entre els diferents grups de les mostres, o bé observarem poques OTUs. En canvi, si el nombre de mostres és massa gran, el cost i el temps necessari per fer les anàlisis incrementarà. És important tenir en compte que la mida mostral depèn de la variança que hi hagi entre les mostres.

L’ús de la metagenòmica per estudiar la diversitat de la microbiota condiciona com definim la distribució d’abundàncies de les espècies. Per una banda, cal tenir present que els valors observats acostumen a estar truncats, és a dir que OTUs amb zeros no són observades. Per altra banda, la distribució d’abundàncies pot seguir diferents distribucions. Malauradament, cap dels models clàssics de la distribució d’abundàncies d’espècies (log-sèries, sèries geomètriques, log-normal) són útils amb comunitats microbianes complexes, sobretot en estats variables d’alteració (de pertorbats a no pertorbats) o emprobriment (rics a pobres) [11]. Per tant, un altre punt, amb una dificultat matemàtica intrínseca, té a veure amb quin tipus de distribución de probablitats s’ajusta a les matrius metagenòmiques (espècies, abundàncies, mostres i comunitats), no només per a un mostra senzilla (espècies vs abundància), on seria interessant estudiar la seva distribució d’abundàncies d’espècies. Una possible causa que expliqui aquesta dificultat es deuria a la manca d’homogeneïtat de les comunitats microbianes. Aquestes comunitats microbianes sovint es composen de subpoblacions o comunitats que estableixen una relació genètica particular per a cada individu en el temps i l’espai.

Prenent en consideració la dificultat esmentada anteriorment, contrastarem el mètode clàssic de rarefacció, amb una aproximació aplicant un algorisme bayesià per estimar quantes mostres serien necessàries per assolir un 95% del nombre màxim d’espècies esperades.

## 2.2 Metodologia

La comparació entre rarefacció clàssica i el mètode bayesià s’ha dut a terme en matrius metagenòmiques simulades. Les dades de les matrius han estat generades aplicant una distribució Diritchlet multinomial amb el paquet *LearnBayes* [12]. Aquesta distribució conjuga una distribució Diritchlet i una distribució multinomial, i mostra una semblança amb les matrius metagenòmiques generades a partir de mostrejos reals. Els paràmetres que cal definir són el nombre d’espècies o OTUs (*nsites*), el nombre de rèpliques (*nsimulac*) i l’abundància (*ab.total*), entesa com el nombre d’individus per OTU. S’han generat tres diferents escenaris per a cada paràmetre. Pel que fa al nombre d’OTUs, s’han aplicat valors de 50, 100 i 250, per poder contrastar diferents situacions de riquesa específica. Les diferents situacions per al nombre de rèpliques ha estat 5, 10 i 25. Els tres casos d’abundància han estat 100, 250 i 500. El script presenta el següent format:

*library(Learnbayes)*

*nsites<-xxx*

*nsimulac<-y*

*ab.total<-zzz*

*ppp <- rdirichlet(1, par = rep(1, nsites))*

*ppp*

*matriu <- array(0, dim=c(nsites, nsimulac))*

*for(i in 1:nsimulac){*

*X <- as.vector(rmultinom(1, size =ab.total , prob = ppp))*

*matriu[,i]<-X*

*N <- sum(X)*

*}*

*Matriu*

S’han generat 27 matrius diferents, de cada una d’elles s’ha calculat la seva rarefacció aplicant la funció *specaccum*() del paquet *vegan* [13]. A més, per cada matriu s’ha calculat el nombre estimat d’OTUs i el nombre màxim predit aplicant un algorisme bayesià utilitzant un mètode de MCMC. Aquest algorisme ve definit en la funció *MetagenSample.size.H*(), establerta a partir d’una llibreria descrita per discriminar grups dins de matrius metagenòmiques [14]. La funció inclou en el seu conjunt de comandes tant el model clàssic de rarefacció definit per *specaccum* com el model bayesià. La funció contempla dos possibles situacions, que les probabilitats de presència de les OTUs siguin iguals, o bé que les probabilitats siguin semblants a les observades. Aquesta funció també inclou en el seu script les comandes per calcular la rarefacció de la matriu estudiada.

La funció genera les corbes de rarefacció, així com els valors estimats i màxim predit amb el corresponent esforç de mostreig que caldria fer per assolir el 90, 95 i 99 % de la riquesa estimada o predita. Aquesta estimació l’obté aplicant mètode loess sobre un model logístic.

El script per a un escenari concret tindria la següent rutina de comandes:

*library(R2jags)*

*library(vegan)*

*library(proto)*

*library(nls2)*

*library(ggtern)*

*library(LearnBayes)*

*library(BDSbiost3)*

*## nOTU=50 Repl=5 Abund=100*

*nsites<-50*

*nsimulac<-5*

*ab.total<-100*

*ppp <- rdirichlet(1, par = rep(1, nsites))*

*matriu <- array(0, dim=c(nsites, nsimulac))*

*for(i in 1:nsimulac){*

*X <- as.vector(rmultinom(1, size =ab.total , prob = ppp))*

*matriu[,i]<-X*

*N <- sum(X)*

*}*

*matriu*

*MetagenSample.size.H(matriu, quart.cut=0.3,type=T,model.probability=1)*

En tots els casos s’ha considerat el model on totes les OTUs tenen la mateixa probabilitat d’ésser presents en la mostra.

Els valors obtinguts pel que fa a la rarefacció clàssica i al valor estimat de la riquesa pel mètode bayesià han estat comparats entre ells per veure quin escenari proposat generava una millor predicció. Per a cada valor del número d’OTUs, s’han aplicat anàlisis de la variança (anova) d’un factor per veure quin paràmetre: nombre de rèpliques o abundància, incedeixen més en el nombre estimat d’espècies i l’esforç de mostreig necessari per assolir aquells valors.

Els càlculs han estat desenvolupats amb el paquet estadístic R [15].

## 2.3 Resultats

La funció *MetagenSample.size.H*() genera, a banda dels valors estimats pel que fa a rarefacció i el model bayesià, així com els valors d’esforç mostral (Taula 1), les corbes corresponents pels diferents models. La Figura 1 recull l’exemple de les corbes i l’histograma d’abundància corresponents a l’escenari de 100 OTUs, amb 5 rèpliques i abundàncies de 100 individus per OTU.



Figura . Exemple de les corbes generades per la funció *MetagenSample.size.H*. El model genera les corbes de rarefacció clàssica i bayesiana a més del model de predicció. També inclou l’histograma d’abundàncies dels diferents OTUs.

El número inicial d’OTUs mostra un efecte sobre l’estimació del nombre total. Si el número d’OTUs és baix (50, en el nostre cas), la rarefacció clàssica estima millor la diversitat específica que la rarefacció bayesiana, que la subestima lleugerament. Però quan s’incrementa el nombre d’OTUs identificades, en les situacions on el número inicial d’OTUs és 100 o 250, la rarefacció clàssica subestima la diversitat específica repecte de la rarefacció bayesiana. Els valors estimats són sensiblement menors amb el càlcul realitzat amb el model clàssic que amb el model bayesià.

L’anàlisi de variances pels diferents escenaris mostra que l’abundància d’individus i el nombre de rèpliques tenen un efecte sobre els valors estimats de rarefacció clàssica i bayesiana com també en l’esforç. Els resultats de les diferents anàlisis es troben recollits a l’annex 6.1.

L’abundància té un efecte diferenciat segons del nombre inicial d’OTUs de l’escenari triat. Quan el nombre d’OTUs inicial és 50, l’efecte de l’abundància no és estadísticament significatiu, però s’observa que l’increment en l’abundància suposa una subestimació per la rarefacció bayesiana (Figura 2). Si es pren 100 com a valor inicial d’OTUs, l’efecte de l’abundància és significatiu estadísticament, però manté el mateix patró que amb un valor inicial d’OTUs més baix. L’increment en l’abundància porta a una subestimació de la riquesa específica de la mostra (Figura 3). Aquest mateix patró s’observa, pel que fa a l’efecte de l’abundància en l’estimació de riquesa específica, amb un valor inicial de 250 OTUs.

Pel que fa l’esforç de mostreig, l’abundància no mostra cap diferència siginificativa que faci variar l’esforç, d’acord amb les anoves realitzades sobre els resultats (annexe 6.1). En els diferents escenaris establerts, l’increment en l’abundància també suposava un lleuger increment en l’esforç de mostreig.

Taula . Valors de l’estimació del nombre d’OTUS aplicant la rarefacció clàssica i bayesiana, així com el valor de l’esforç mostral necessari per assolir el 95% de l’estimació bayesiana (esforç mostral1) i el 95% de valor màxim predit pel model (esforç mostral2).

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Nombre d’OTUs | Rèpliques | Abundància | Rarefacció | Estimació bayesiana | Predicció màxima | esforç mostral1 | esforç mostral2 |
| 50 | 5 | 100 | 48 | 48 | 54,1 | 2,82 | 10,07 |
| 250 | 48 | 43 | 43 | 4,11 | 4,91 |
| 500 | 50 | 46 | 48,3 | 3,93 | 7,9 |
| 10 | 100 | 47 | 48,8 | 51 | 3,7 | 7,1 |
| 250 | 49 | 47,1 | 49,8 | 3,53 | 7,65 |
| 500 | 50 | 43,9 | 44 | 4,5 | 4,66 |
| 25 | 100 | 50 | 50 | 51,5 | 6,2 | 9,84 |
| 250 | 49 | 47,4 | 47,4 | 5,97 | 6,06 |
| 500 | 50 | 45,1 | 45,1 | 9,3 | 6,24 |
| 100 | 5 | 100 | 88 | 100 | 114 | 3,06 | 11,67 |
| 250 | 90 | 92 | 99,8 | 3,54 | 8,19 |
| 500 | 95 | 86 | 91,4 | 2,53 | 5,73 |
| 10 | 100 | 88 | 100,1 | 106,8 | 3,86 | 9 |
| 250 | 98 | 99,2 | 103,4 | 5,56 | 10,25 |
| 500 | 97 | 91,1 | 94,8 | 3,8 | 7,01 |
| 25 | 100 | 97 | 103,4 | 103,4 | 7,19 | 6,72 |
| 250 | 99 | 100,1 | 102,8 | 6,05 | 9,39 |
| 500 | 97 | 93,8 | 94,5 | 6,15 | 7,04 |
| 250 | 5 | 100 | 165 | 248 | 281,3 | 2,88 | 10,63 |
| 250 | 209 | 250 | 279 | 3,29 | 10,92 |
| 500 | 224 | 238 | 262 | 3,15 | 9,52 |
| 10 | 100 | 207 | 251,7 | 267 | 3,65 | 8,08 |
| 250 | 227 | 250 | 267,1 | 3,77 | 8,92 |
| 500 | 243 | 247 | 257,1 | 5 | 9,11 |
| 25 | 100 | 227 | 259 | 259 | 7,36 | 6,59 |
| 250 | 236 | 255,3 | 257,3 | 6,3 | 7,23 |
| 500 | 247 | 248,8 | 254,1 | 5,74 | 8,2 |

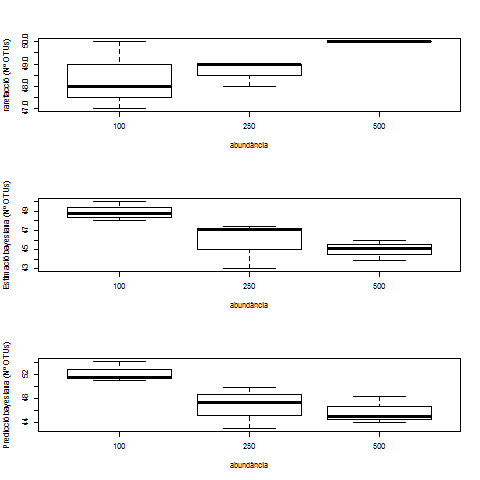


Figura . Distribució dels valors de l’estimació obtinguda per rarefacció clàssica (dalt), bayesiana (mig), i el nombre màxim predit d’espècies (baix) fetes per als diferents valors d’abundància d’OTUs per a una mostra amb un valor inicial de 50 OTUs

El nombre de rèpliques mostra un efecte més interessant que l’abundància sobre les estimacions obtingudes pel model de rarefacció clàssica i pel model de rarefacció bayesiana. Aquest efecte és significatiu pels casos on el nombre inicial d’OTUs és mitjà o gran. Quan inicialment tenim pocs OTUs (50), el nombre de rèpliques no afecta significativament al valor d‘estimació, ja sigui pel model clàssic o pel model bayesià de rarefacció. En canvi quan el nombre d’OTUs s’incrementa, un augment en el nombre de rèpliques afavoreix que s’assoleixin valors superiors en l’estimació de la riquesa específica, ja sigui aplicant el model clàssic o el model bayesià de rarefacció (Figura 4). El patró mostrat per l’escenari amb un valor inicial de 100 OTUs el segueixen tant el cas de 50 com el de 250 OTUs.

L’efecte del nombre de rèpliques sobre l’esforç de mostreig és significatiu a tots els escenaris definits (annexe 6.1), però només afecta a l’esforç de mostreig necessari per fer l’estimació aplicant la rarefacció bayesiana. La relació és directament proporcional, a més rèpliques de les mostres, cal un esforç més gran per assolir el 95% de la riquesa específica estimada (Figura 5). Aquest patró és comú per a tots els escenaris definits, independentment del nombre inicial d’OTUs de la mostra.

L’us de corbes d’acumulació d’espècies ha estat un recurs habitual en estudis d’ecologia convencial, però no és molt habitual fer-ho en estudis de metagenòmica. Valverde & Mellado [16] constaten que l’aplicació de models clàssics d’estimació de la riquesa, com Chao1, en matrius metagenòmiques complexes en termes de diversitat suposa una sobreestimació de la riquesa específica. En el nostre cas la rarefacció clàssica sovint ha suposat una subestimació de la riquesa quan la matriu tendia a ser molt diversa. En canvi l’estimació feta amb el model de rarefacció bayesiana no sobrestimava en excés el valor de riquesa esperat a la mostra (Taula 1).

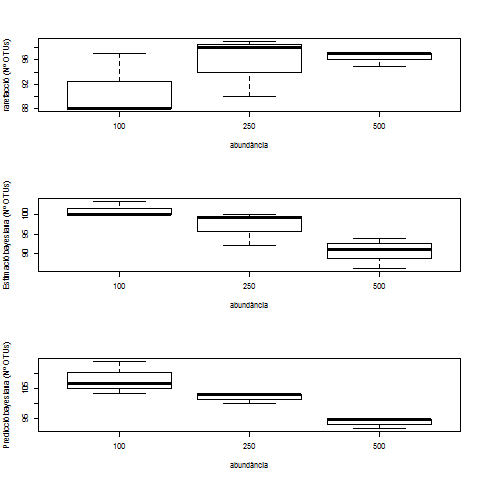


Figura Distribució dels valors de l’estimació obtinguda per rarefacció clàssica (dalt), bayesiana (mig), i el nombre màxim predit d’espècies (baix) fetes per als diferents valors d’abundància d’OTUs per una mostra amb un valor inicial de 100 OTUs.

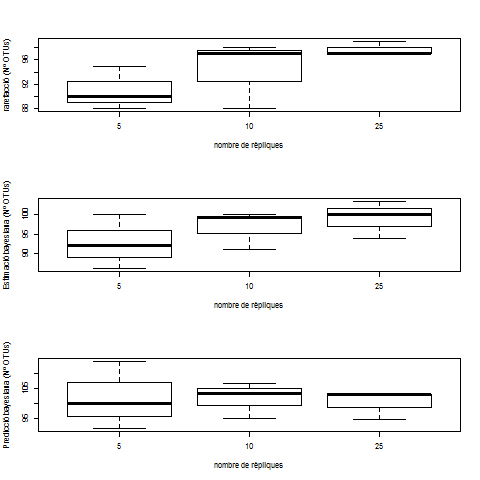


Figura . Distribució dels valors de l’estimació obtinguda per rarefacció clàssica (dalt), bayesiana (mig), i el nombre màxim predit d’espècies (baix) segons el nombre de rèpliques fetes per una mostra amb un valor inicial de 250 OTUs

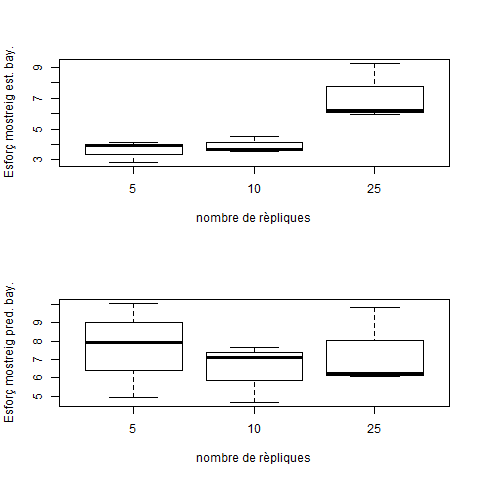


Figura . Distribució del valor de l’esforç de mostreig per obtenir el 95 de la riquesa específica estimada per rarefacció bayesiana (dalt) o el 95% de la màxima riquesa específica predita (baix) en base al nombre de rèpliques de la mostra amb un nombre inicial de 50 OTUs.

# 3. Conclusions

L’aplicació d’un model de rarefacció basat en models bayesians suposa una aproximació més encertada a l’hora d’estimar la riquesa específica d’una mostra si el nombre d’OTUs és gran. En canvi, el model de rarefacció clàssica ha reproduït millor el nombre esperat de tàxons o OTUs d’una mostra estudiada quan aquesta presenta una riquesa específica baixa. Aplicat a matrius metagenòmiques aquest resultat és interessant ja que les dades que s’obtenen en els estudis metagenòmics generalment comporten una gran diversitat d’OTUs, per tant l’ús del model bayesià de rarefacció pot permetre estimar la riquesa específica de manera força viable i amb abundàncies relativament baixes de les OTUs. Per altra banda, el model de rarefacció bayesiana permet assolir unes estimacions molt acurades aplicant nombres baixos de rèpliques. Que a la vegada permeten uns esforços de mostreig prou bons.

Malgrat els bons resultats del treball, els objectius plantejats inicialment incloïen poder establir comparacions entre comunitats, però no ha estat possible. Les dificultats de planificació temporal no han permès dur a terme tots els objectius. No ha estat un problema procedimental o metodològic el que ha impedit assolir els objectius, tot el contrari. La disponibilitat metodològica i el suport tant a nivell conceptual com instrumental per part del tutor han estat un puntal important per poder completar el treball.

Un problema important per dur a terme aquest treball, tal i com estava planificat inicialment, ha estat la temporització del mateix TFM. La distribució de les fases de treball al llarg del semestre són molt acotades i requereixen una disponibilitat per de l’alumne dificilment assolibles, sobretot si l’activitat laboral suposa un percentatge elevat del temps. Analitzant la durada de les fases i la dedicació que requereixen, d’acord amb el pla docent tindríem:

PAC1: 15% del temps=56.3 hores, 15 dies d’execució: 3,75 hores/dia

PAC2: 25% del temps=93,75 hores, 21 diesd’execució: 4,5 hores/dia

PAC3: 25% del temps=93,75 hores, 35 dies d’execució: 2,7 hores/dia

Memòria: 20% del temps=75 hores, 14 dies d’execució: 5, hores/dia

Preparació defensa: 5% del temps=18,5 hores, 13 dies d’execució: 1,4 hores/dia

L’organització del temps requereix de més de mitja jornada laboral en alguns periodes d’execució, indistintament del dia de la setmana. L’elecció d’un curs no presencial sovint ve vinculat a la manca de temps per poder assistir a classes presencials. Però sembla que el plantejament del TFM d’aquest màster sigui per qui pugui disposar de molt de temps lliure (treballadors a jornada intensiva, mitja jornada, aturats,...).

El format de la memòria del TFM té un caire tant de treball de marquèting o similar que sembla que no hagi d’ajustar-se a un treball més científic. Això pot estar degut també a la meva experiència professional.

En un futur proper no em veig dedicant-me en aquesta àrea tant específica de la metagenòmica, però ha estat interessant fer-hi una cabussada per conèixer altres aspectes aplicats de la bioestadística.

# 4. Glossari

**Metagenòmica**, estudi de l’ADN genòmic obtingut de microorganismes no cultivats.

**OTU** *Operational Taxonomic Unit*, unitat taxònòmica operacional entesa com a una espècie diferent en termes microbiològics. La diferenciació es fonamenta en el DNA ribosòmic i el llindar de percentatge de similaritat que s’estableix per classificar els microbis en la mateixa OTU o en una OTU diferent.

**Rarefacció**, tècnica per avaluar la riquesa específica obtinguda a partir del mosterig, es basa en la construcció d’una corba de rarefacció, que descriu l’increment en el nombre d’espècies trobades en base al nombre d’individus mostrejats.

# 5. Bibliografia

1 Magurran, A., *Measuring Biological Diversity*, Blackwell Publishing Ltd, Oxford, 2004.

2 Gotelli,N. J. & Colwell, R. K., Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness, Ecological Letters,379-391 / 4, 2001.

3 Colwell, R. K. & Codington, J. A., Estimating terrestrial biodiversity through extrapolation, Philosophical Transactions of the Royal Society of London, 101-118 / 345, 1994.

4 Chazdon, R. L., Colwell, R. K., Denslow, J. S. & Guariguata, M. R., Statistical methods for estimating species richness o woody regeneration in primary and secondary forests of northeastern Costa Rica. In *Forest biodiversity research background and old world case studies* (ed. F. Dallmeier & J. A. Comiskey), pp. 285-309, Parthenon Publishing, Paris, 1998.

5 Gaston, K. J., Species richness: measure and measurement. In *Biodiversity: a biology of numbers and difference* (ed. K. J. Gaston), pp. 77-113, Oxforrd University Press, Oxford, 1996.

6 Krebs, C. J., *Ecological Methodology*, 2nd edn, Harper & Row, New York, 1999.

7 Thomas, T., Gilbert, . & Meyer, F., Metagenomics – a guide from sampling to data analysis, Microbial Informatics and Experimentation, 3 / 2, 2012.

8 Hazen, T. C., Rocha, A. M. & Techtmann, S. M., Advances in monitoring environmental microbes, Current Opinion in Biotechnology, 526-533 / 24, 2013.

9 Chen, K. & Pachter, L., Bioinformatics for whole-genome shotgun sequencing of microbial communities, PloS Computational Biology, 106-112 / 1, 2005.

10 Wooley, J. C., Godzik, A. & Friedberg, I., A primer on metagenomics, PloS Computational Biology, e1000667 / 6, 2010.

11 Lindner, M. S. & Renard, B. Y., Metagenomics abundance estimation and diagnostic testing on species level. Nucleic Acids Research, e10 / 41, 2013.

12 Albert, J., LearnBayes: Functions for Learning Bayesian Inference. R package version 2.15. https://CRAN.R-project.org/package=LearnBayes, 22/05/2017

13 Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H. Szoecs, E. & Wagner, H. vegan: Community Ecology Package. R package version 2.4-3. <https://CRAN.R-project.org/package=vegan>, 22/05/2017.

14-Rodriguez Casado, C. I. & Monleon-Cotino, A., A new R library for discriminating groups based on abundance profile and biodiversity in microbiome metagonimc matrices, International Journal of Scientific & Engineering Research, 243-25 / 7, 2016.

15 R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>, 19/05/2017.

16 Valverde, J. R. & Mellado, R. P. Analysis of metagenomic data containing high biodiversity levels, PLOS ONE, e58118 / 8, 2013.

# 6. Annexos

## 6.1 Resultats dels ANOVA

test <- read.csv2("C:/LLOP/UOC/TFM/treball/test.csv")  
test$NOTU <- factor(test$NOTU,levels = c(50,100,250),labels = c("50","100","250"))  
test$repl <- factor(test$repl,levels = c(5,10,25),labels = c("5","10","25"))  
test$abund <- factor(test$abund,levels = c(100,250,500),labels = c("100","250","500"))  
  
  
test.50 <- test[1:9,]  
attach(test.50)  
aov.rar50 <- aov(raref~abund+repl)  
aov.bay50 <- aov(StBay~abund+repl)  
aov.pred50 <- aov(PdBay~abund+repl)  
aov.effst50 <- aov(EffBy~abund+repl)  
aov.effpred50 <- aov(EffPd~abund+repl)  
  
anova(aov.rar50)

## Analysis of Variance Table  
##   
## Response: raref  
## Df Sum Sq Mean Sq F value Pr(>F)  
## abund 2 4.6667 2.33333 2.8 0.1736  
## repl 2 2.0000 1.00000 1.2 0.3906  
## Residuals 4 3.3333 0.83333

anova(aov.bay50)

## Analysis of Variance Table  
##   
## Response: StBay  
## Df Sum Sq Mean Sq F value Pr(>F)   
## abund 2 25.7756 12.8878 4.5656 0.09279 .  
## repl 2 5.0422 2.5211 0.8931 0.47789   
## Residuals 4 11.2911 2.8228   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

anova(aov.pred50)

## Analysis of Variance Table  
##   
## Response: PdBay  
## Df Sum Sq Mean Sq F value Pr(>F)  
## abund 2 71.716 35.858 3.6798 0.1240  
## repl 2 0.329 0.164 0.0169 0.9833  
## Residuals 4 38.978 9.744

anova(aov.effst50)

## Analysis of Variance Table  
##   
## Response: EffBy  
## Df Sum Sq Mean Sq F value Pr(>F)   
## abund 2 4.7630 2.3815 2.5977 0.18923   
## repl 2 23.1330 11.5665 12.6164 0.01872 \*  
## Residuals 4 3.6671 0.9168   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

anova(aov.effpred50)

## Analysis of Variance Table  
##   
## Response: EffPd  
## Df Sum Sq Mean Sq F value Pr(>F)  
## abund 2 15.3143 7.6571 2.4963 0.1979  
## repl 2 2.2268 1.1134 0.3630 0.7164  
## Residuals 4 12.2696 3.0674

detach(test.50)  
  
  
test.100 <- test[10:18,]  
attach(test.100)  
aov.rar100 <- aov(raref~abund+repl)  
aov.bay100 <- aov(StBay~abund+repl)  
aov.pred100 <- aov(PdBay~abund+repl)  
aov.effst100 <- aov(EffBy~abund+repl)  
aov.effpred100 <- aov(EffPd~abund+repl)  
  
anova(aov.rar100)

## Analysis of Variance Table  
##   
## Response: raref  
## Df Sum Sq Mean Sq F value Pr(>F)  
## abund 2 50.667 25.333 2.6207 0.1873  
## repl 2 66.667 33.333 3.4483 0.1348  
## Residuals 4 38.667 9.667

anova(aov.bay100)

## Analysis of Variance Table  
##   
## Response: StBay  
## Df Sum Sq Mean Sq F value Pr(>F)   
## abund 2 180.862 90.431 24.904 0.005526 \*\*  
## repl 2 63.762 31.881 8.780 0.034421 \*   
## Residuals 4 14.524 3.631   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

anova(aov.pred100)

## Analysis of Variance Table  
##   
## Response: PdBay  
## Df Sum Sq Mean Sq F value Pr(>F)   
## abund 2 318.18 159.088 9.2487 0.03161 \*  
## repl 2 4.31 2.154 0.1253 0.88560   
## Residuals 4 68.80 17.201   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

anova(aov.effst100)

## Analysis of Variance Table  
##   
## Response: EffBy  
## Df Sum Sq Mean Sq F value Pr(>F)   
## abund 2 1.2075 0.6037 1.1516 0.40272   
## repl 2 17.7850 8.8925 16.9614 0.01113 \*  
## Residuals 4 2.0971 0.5243   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

anova(aov.effpred100)

## Analysis of Variance Table  
##   
## Response: EffPd  
## Df Sum Sq Mean Sq F value Pr(>F)  
## abund 2 13.6565 6.8282 1.9864 0.2517  
## repl 2 1.7861 0.8930 0.2598 0.7833  
## Residuals 4 13.7501 3.4375

detach(test.100)  
  
  
test.250 <- test[19:27,]  
attach(test.250)  
aov.rar250 <- aov(raref~abund+repl)  
aov.bay250 <- aov(StBay~abund+repl)  
aov.pred250 <- aov(PdBay~abund+repl)  
aov.effst250 <- aov(EffBy~abund+repl)  
aov.effpred250 <- aov(EffPd~abund+repl)  
  
anova(aov.rar250)

## Analysis of Variance Table  
##   
## Response: raref  
## Df Sum Sq Mean Sq F value Pr(>F)   
## abund 2 2257.56 1128.78 9.5166 0.03016 \*  
## repl 2 2208.22 1104.11 9.3087 0.03128 \*  
## Residuals 4 474.44 118.61   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

anova(aov.bay250)

## Analysis of Variance Table  
##   
## Response: StBay  
## Df Sum Sq Mean Sq F value Pr(>F)   
## abund 2 121.536 60.768 9.4270 0.03063 \*  
## repl 2 122.562 61.281 9.5067 0.03021 \*  
## Residuals 4 25.784 6.446   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

anova(aov.pred250)

## Analysis of Variance Table  
##   
## Response: PdBay  
## Df Sum Sq Mean Sq F value Pr(>F)   
## abund 2 232.23 116.114 6.7886 0.05179 .  
## repl 2 454.83 227.414 13.2956 0.01710 \*  
## Residuals 4 68.42 17.104   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

anova(aov.effst250)

## Analysis of Variance Table  
##   
## Response: EffBy  
## Df Sum Sq Mean Sq F value Pr(>F)   
## abund 2 0.0624 0.0312 0.050 0.95178   
## repl 2 17.7708 8.8854 14.246 0.01516 \*  
## Residuals 4 2.4949 0.6237   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

anova(aov.effpred250)

## Analysis of Variance Table  
##   
## Response: EffPd  
## Df Sum Sq Mean Sq F value Pr(>F)   
## abund 2 0.6146 0.3073 0.5138 0.63301   
## repl 2 13.6925 6.8462 11.4460 0.02212 \*  
## Residuals 4 2.3925 0.5981   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

detach(test.250)