UNIVERSITÀ
DEGLI STUDI
DI PADOVA

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

CORSO DI LAUREA MAGISTRALE IN
Bioingegneria

"Inferenza del potenziale funzionale del microbiota partendo dal dato 16S"

Relatore: Prof.ssa Barbara Di Camillo

Laureando: Matteo Sprocatti

Correlatore: Dott.ssa Barbara Simionati
Dott.ssa Ilaria Patuzzi

ANNO ACCADEMICO 2022 – 2023

Data di laurea 03/07/2023

Ringraziamenti

Ringrazio specialmente i miei genitori, Giampaolo e Giuseppina, perché nonostante tutto, esami non passati al primo colpo o anche momenti difficili, in cui non credevo al 100% in me stesso mi hanno sempre sostenuto e mi hanno sempre cercato di aiutare nel miglior modo in cui riuscivano. Mi hanno permesso di studiare tranquillamente e hanno sempre trovato parole per spronarmi e per gratificarmi, quindi grazie di cuore davvero. Anche se le parole non bastano per ringraziarvi a sufficienza per quello che avete fatto per me e per quanto mi avete sopportato anche nei momenti di nervosismo.

Ci tengo anche a ringraziare specialmente i miei fratelli, Michele e Andrea, perché a modo loro mi hanno sempre sostenuto e in particolare Michele, durante il periodo di redazione di questa tesi mi è stato vicino in particolare per alcune cose di informatica o anche solo per darmi un parere sulla presentazione o anche su come scrivere certi concetti quindi grazie di cuore anche a voi.

Un ringraziamento speciale va a mia pro zia, Ornella, che purtroppo è venuta a mancare l'anno scorso, ma è sempre stata una mia grandissima fan e mai dimenticherò il suo spronarmi a finire questo mio percorso e nemmeno la sua espressione orgogliosa quando avevo finito la triennale. Mi sarebbe piaciuto fossi qua oggi zia ma purtroppo non è così ma sono convinto che da lassù mi stai guardando, a parole è impossibile spiegare quanto io tenessi a te. Questo traguardo è anche per te!!!

Ringrazio poi i miei nonni, Remigio e Luciana, che mi hanno sempre supportato e hanno sempre avuto parole di sprono e di gratificazione nei miei confronti.

Ringrazio anche i miei zii e i miei cugini per avermi sempre spronato e per essermi sempre stati vicini durante questo percorso, nonostante le difficoltà che ho trovato.

Ringrazio anche la mia ragazza, Anna, che in questi mesi mi ha supportato e sopportato e mi ha sempre ascoltato e confortato quando ne avevo necessità, inoltre mi ha sempre spronato e mi è sempre stata vicino ed era sempre in grado di tranquillizzarmi quando magari in certi momenti vedevo tutto nero e vedevo alcuni problemi come insuperabili. Inoltre ha sempre creduto in me e mano a mano che procedevo con la tesi e ottenevo risultati mi ha sempre dimostrato quanto fosse fiera di me e contenta per me. Grazie amore mio.

Non posso non ringraziare i miei colleghi ossia Elena, Alessia, Giulia, Isacco, Alice, Mara e Rebecca per aver condiviso con me, chi in parte e chi totalmente il percorso di studi. Li ringrazio per essermi stati vicino, per avermi aiutato quando ero in difficoltà, per gli esami preparati insieme, per i momenti in cui ci siamo rilassati insieme dopo un esame o dopo le lezioni. Vi voglio bene.

Ringrazio anche i miei amici di una vita, su cui so di poter contare sempre e so che per me ci saranno sempre sia quando si tratta di gioire con me sia quando si tratta di consolarmi e con cui ho un legame veramente profondo che so che durerà per tutta la vita. Vi voglio bene ragazzi.

Infine ringrazio la professoressa Barbara Di Camillo, la dottoressa Ilaria Patuzzi e la dottoressa Simionati per avermi seguito in questo percorso di tesi, per avermi aiutato quando ero in difficoltà e per la disponibilità dimostrata durante la stesura della tesi.

Indice

Abstract

Il microbiota intestinale, ovvero l'insieme dei microbi presenti all'interno dell'intestino, ha un ruolo fondamentale per quel che riguarda la salute di uomini e animali, al punto che alterazioni del suo equilibrio possono influenzare lo sviluppo di diverse malattie. Negli anni sono state sviluppate diverse tecniche per la caratterizzazione del microbiota; in particolare hanno trovato vasta applicazione quelle basate sul "Next Generation Sequencing" ossia sui sequenziatori genici di nuova generazione. All'interno di questo lavoro di tesi sono stati presi in considerazione due metodi di sequenziamento ovvero il 16S rRNA Amplicon Sequencing e poi il Whole Genome Shotgun Sequencing (WGS).
Il primo metodo, più economico, consente di quantificare seppur in maniera relativa, la presenza delle diverse categorie tassonomiche nel campione; il secondo metodo, circa 10 volte più costoso, consente di quantificare, non solo la presenza di diversi batteri, ma anche la presenza dei diversi geni nei genomi batterici, e quindi la loro potenzialità in termini di funzioni che possono svolgere.
In questa tesi ho valutato diversi metodi per l'inferenza del potenziale funzionale, a partire da dati 16s, per poi comparare i risultati con quelli che si ottengono direttamente da dati WGS su 9 diversi dataset, reali e simulati. I tool che sono stati considerati per questa analisi sono PICRUSt2, PanFP e Tax4Fun2. I confronti poi sono stati effettuati utilizzando diverse metriche come correlazione di Spearman, indice di Bray-Curtis, Balanced Accuracy, Precision, Recall e F_1_score e diverse tecniche quali Principal Coordinate Analysis (PCoA) e Clustering.
Il tool migliore, che permette di ottenere una predizione del profilo funzionale che, partendo dal dato 16S, si avvicina al dato WGS, sembra essere PICRUSt2 per 7 dei 9 dataset presi in considerazione. In particolare, se non a quantificare con esattezza la presenza dei diversi geni, questi tool sembrano una buona risorsa per l'analisi della presenza/assenza degli stessi.

Capitolo 1 - Introduzione

1.1 Il microbiota

L'apparato gastroenterico rappresenta un ecosistema complesso in cui le varie componenti sono in stretta relazione tra loro. Questo apparato costituisce un'unità funzionale che ha la duplice funzione di permettere l'assorbimento di nutrienti e degli elementi necessari al metabolismo umano e di impedire a microorganismi patogeni di entrare nell'organismo umano. Il microbiota intestinale, in particolare, rappresenta l'insieme dei microbi presenti nel tratto gastrointestinale e svolge un ruolo fondamentale. Nel corso degli anni sono state sviluppate diverse tecniche per la caratterizzazione del microbiota intestinale tra cui:

- Le tecniche di coltura standard usate anche in microbiologia che però non permettono di rilevare la maggioranza delle componenti del microbiota, con la conseguenza che il contributo di dette tecniche alla caratterizzazione del microbiota è stato poco rilevante;
- Successivamente sono state utilizzate tecniche che si avvalgono della caratterizzazione di sequenze di RNA ribosomiale (porzione 16S (16S rRNA) tramite PCR), le quali hanno dato un apporto significativo alla scoperta delle conoscenze relative alla composizione del microbiota intestinale;
- Ulteriori tecniche di analisi del microbiota sono la DGGE (Denaturing Gradient Gel Electrophoresis) e la TGGE (Temperature Gradient Gel Electrophoresis) che si basano sull'estrazione, l'amplificazione e la successiva analisi elettroforetica del DNA cellulare;
- Un'altra tecnica è rappresentata dalla FISH tramite la quale il 16S rRNA viene marcato con materiale fluorescente e successivamente identificato;
- La tecnica sviluppata più di recente, denominata Next Generation Sequencing (NGS) si basa su sequenziatori genici di nuova generazione, ovvero differenti rispetto a quelli classici sia per tecnologia utilizzata sia per quantità di dati che sono in grado di collezionare;

La metagenomica infatti prevede l'utilizzo di questi ultimi sequenziatori genici per la caratterizzazione delle specie microbiche nei vari campioni biologici. Tra le ricerche più avanzate va citata anche la metabolomica, che consiste nello studio delle impronte chimiche lasciate da specifici processi cellulari delle cellule dell'ospite e del microbiota intestinale e dall'insieme dei metaboliti che un organismo produce. La biologia sistemica sta cercando di realizzare una possibile integrazione tra la metagenomica, la proteomica, la trascrittomica e le informazioni metabolomiche per poter

arrivare ad avere una visione, il più completa possibile, degli organismi viventi e, nello specifico, del microbiota intestinale per capirne meglio le funzionalità.

Il microbiota intestinale è costituito da circa 10^4 microorganismi, in numero 10-20 volte maggiore rispetto alle cellule dell'organismo umano, e rappresenta un patrimonio genetico totale di 3.3 milioni di geni. Esso è composto prevalentemente da batteri oltre che da virus, Archaea e Eukaryota quali funghi, Amoebozoa, protozoi e altri. Nonostante siano in minoranza alcuni studi hanno evidenziato che virus e funghi hanno un ruolo chiave nella patogenesi delle diverse patologie. I fattori che possono alterare il microbiota intestinale sono vari e comprendono sia le condizioni fisiologiche (per esempio: età, dieta, profilo genetico) sia condizioni patologiche (tra cui infezioni, interventi chirurgici, utilizzo di determinati farmaci) che possono portare a conseguenze cliniche ossia problemi di salute.

Esistono tre differenti tipologie di microbiota intestinale:

- Il microbiota intestinale batterico (gut bacteriome) composto da più di mille specie di batteri. Questo, dopo la caratterizzazione, risulta composto da quattro phyla batterici ossia, più specificatamente, da Firmicutes, Bacteroides, Proteobacteria e Actinobacteria per un totale di 11 divisioni. I due phyla batterici maggiormente rappresentati sono quella dei Firmicutes e quella dei Cytophaga-Flavobacterium-Bacteroides. I Firmicutes possono essere suddivisi in due principali gruppi ossia il gruppo Clostridium cluster XIVa o Clostridium coccoides e il gruppo Clostridium cluster IV o Clostridium leptum. Il primo di questi due gruppi (Clostridium cluster XIVa o Clostridium coccoides) include membri dei generi Butyrivibrio, Clostridium, Caprococcus, Dorea, Eubacterium, Lachnospira, Roseburia, Ruminococcus ed è ben rappresentato in termini quantitativi all'interno del lume intestinale. Esso ha la peculiarità di avere al suo interno numerosi batteri capaci di produrre acidi grassi a catena corta, in particolare l'acido butirrico che rappresenta una fonte di sostentamento fondamentale per le cellule epiteliali del colon. Il secondo gruppo invece è costituito da specie facenti capo ai generi Anaerofilum, Clostridium, Eubacterium, Ruminococcus ed è caratterizzato da numerosi batteri anaerobi, estremofili, produttori di SCFA, tra cui in particolare il Faecalibacterium prausnitzii che ha importanti proprietà antiinfiammatorie, tali per cui una riduzione della sua quantità può portare a delle malattie croniche intestinali. Il secondo gruppo ossia quello del Cythopaga-Flavobacterium-Bacteroides (CFB) è costituito in maggioranza dal phylum dei Bacteroides, gram-negativi e anaerobi obbligati. Infine, ulteriori divisioni batteriche, rappresentate in minor misura, sono: Actinobacteria, Cyanobacteria, Deferribacteres, Deinococcus/Thermus, Fusobacteria, Proteobacteria, Spirochates, VadinBE97 e Verrucomicrobia.

- Il microbiota intestinale virale (gut virome o viroma intestinale), il quale è costituito da virus, che nonostante vengano considerati prevalentemente come organismi patogeni dalle recenti metodiche di analisi metagenomica è stato dimostrato che all'interno dell'intestino umano ci sono prevalentemente batteriofagi. Essi hanno un ruolo nello sviluppo e nell'omeostasi del gut microbiota. All'interno dell'intestino si trova una grande quantità di virioni che però a differenza dei batteri sono caratterizzati da una grande variabilità tra diversi soggetti. Questa variabilità è legata alle differenze individuali nella colonizzazione batterica dell'intestino. Il viroma intestinale è prevalentemente composto da fagi, virus che vanno a infettare determinate specie batteriche. Per la prima volta il gut virome è stato descritto attraverso il Whole Genome Shotgun Sequencing di particelle simili al virus (VLP) estratto da campioni fecali. La famiglia di virus maggiormente rappresentata è quella dei Siphoviridae seguita da quella dei Podoviridae. Recenti analisi del viroma intestinale di pazienti con retticolite ulcerosa e con malattia di Chron hanno evidenziato delle anomalie, caratterizzate da una diminuita diversità nel microbiota batterico. Le suddette patologie, inoltre, sono state associate a un aumento dei batteriofagi Caudovirales. Questi dati hanno dimostrato come un incremento dei virioni non sia secondario rispetto all'alterazione del microbiota batterico ma al contrario se ci sono alterazioni del viroma intestinale queste possono contribuire alla disbiosi batterica e all'infiammazione intestinale.
- Il microbiota intestinale fungino (gut mycobiome). Pur non sapendo come i funghi siano rappresentati a livello del gut microbiota nell'uomo, si può dire che essi siano rilevabili a diverse concentrazioni per ciascun tratto con un gradiente crescente dalla bocca all'ano. Nonostante essi possano essere patogeni per altri organismi, all'interno del tratto gastrointestinale essi agiscono come commensali a basse concentrazioni. Nell'intestino sano le specie dominanti sono Wallemia, Trichocomaceae, Rhodotorula, Saccharomycetaceae, Pleosporaceae, Agaricaceae, Metschnikowiaceae, Cystofilobasidiaceae, Ascomycota, Amphisphaeiacea. Gut mycobiome può interagire direttamente o indirettamente con il sistema immunitario dell'ospite tanto che le interazioni tra esso e il sistema immunitario possono portare al riacutizzarsi di patologie gastrointestinali.

Come scritto in precedenza, il microbiota svolge alcune funzioni molto importanti per il mantenimento dell'omeostasi all'interno del nostro corpo e per questo motivo viene considerato un componente essenziale del nostro organismo tanto che esso può essere definito come il sesto apparato del corpo umano. Numerosi processi, infatti, possono essere completati soltanto grazie al gut microbiota, i cui geni sono in grado di codificare una serie di enzimi e di altre proteine coinvolti in differenti funzioni, tra cui quella di barriera, quella metabolica e quella immunitaria. Il gut microbiota

è considerato a tutti gli effetti un componente della barriera gastrointestinale insieme alle cellule epiteliali, alle giunzioni intercellulari e allo strato di muco che riveste le cellule stesse, al sistema immune locale, al flusso sanguigno, ai sistemi endocrini e neuroenterico e agli enzimi digestivi. Esso è inoltre molto utile per proteggere l'intestino dai patogeni esterni. Gli SCFA (acidi grassi a catena corta) svolgono una funzione trofica nei confronti dei colonciti nonché una rilevante azione antiinfiammatoria. I prodotti di derivazione microbica servono anche a stimolare l'espressione delle giunzioni intercellulari e a regolare il metabolismo del muco intestinale. Il gut microbiota svolge, infatti, anche una funzione metabolica in quanto ha un ruolo fondamentale nel metabolismo delle sostanze nutritive all'interno del nostro organismo ed è considerato come un vero e proprio "bioreattore energetico", capace di trarre energia a partire dai nutrienti ingeriti tramite l'alimentazione. I componenti di quest'ultimo si trovano nel colon e sono utili per la metabolizzazione degli alimenti ingeriti nell'intestino tenue, più specificatamente di fibre, polisaccaridi e oligosaccaridi di origine vegetale. Tramite il microbiota è anche possibile metabolizzare i carboidrati tramite un processo anaerobico che permette di ottenere gli acidi grassi a catena corta ovvero gli SCFA, costituiti da acetato, derivato della fermentazione dei carboidrati dalla maggior parte delle specie batteriche del colon come Bacteroides, Clostridium, Bifidobacterium, Ruminococcus, Eubacterium, butirrato, prodotto dal lavoro di Eubacterium, Coprococcus, E. rectale-Roseburia e Faecalibacterium e propionato, prodotto grazie ai Bacteroides, Propionibacterium e Veillonella. Gli SCFA sono assorbiti dai colonciti e vengono inviati a vari organi tra cui ad esempio il fegato, il sistema muscolare e l'encefalo. Il propionato e il butirrato hanno una funzione trofica nei confronti dei colonciti (in particolare per la sintesi delle membrane cellulari) mentre propionato e acetato sono impiegati dall'organismo rispettivamente per glucogenesi e lipogenesi epatica. Gut microbiota ha un ruolo più rilevante nel metabolismo dei carboidrati rispetto al ruolo che svolge nel metabolismo proteico. Le proteine arrivano in quantità ridotte al colon e con una parte di esse vengono prodotti peptidi assimilati al microbiota, con un'altra gli SCFA e in piccola parte vengono prodotti metaboliti tossici come fenoli, indoli, ammonio e derivati.

Il metabolismo dei lipidi può essere modulato dal microbiota, il quale può deconiugare e idrolissare i sali biliari, essenziali per l'assorbimento dei lipidi. Il microbiota influenza anche l'attività della lipoproteinlipasi (LPL), utile a regolare il rilascio degli acidi grassi dalle lipoproteine tramite downregolazione del Fast-Induced Adipocyte Factor (FLAF), inibitore della produzione di LPL. Gut microbiota inoltre modula la lipogenesi tramite la regolazione dei livelli insulinemici e ha anche un ruolo nel metabolismo vitaminico, grazie alla capacità di sintesi di vitamina K e di quasi tutte quelle del gruppo B.

Infine, il gut microbiota svolge anche una funzione immunitaria in quanto esso colonizza il nostro organismo fin dalla nascita, tramite il parto, nonostante il feto sia sterile. Questa contiguità stimola la maturazione del sistema immune. Le strutture molecolari associate ai microbi (MAMPs) sono particelle microbiche identificate da recettori, i quali si trovano sugli enterociti e sulle cellule dendritiche dell'intestino come ad esempio i Toll-Like Receptors con in seguito la maturazione delle cellule e delle strutture necessarie per svolgere la funzione immunitaria. In base al tipo di specie considerata si associa la stimolazione di diverse sottopopolazioni linfocitarie per esempio: i batteri filamentosi segmentati favoriscono la differenziazione dei linfociti T helper 17, mentre i Clostridia portano alla progressione della differenziazione dei linfociti T regolatori. [1]

1.2 Il sequenziamento

Il sequenziamento rappresenta una tecnica che permette la lettura del codice genetico attraverso l'identificazione della successione di 4 basi azotate, nello specifico Adenina, Citosina, Timina e Guanina, ciò che consente di caratterizzare il DNA e i geni. Il DNA contiene tutte le informazioni genetiche di un individuo ed esse sono alla base dello sviluppo dello stesso. Essere in grado, quindi, di determinare la sequenza è utile per comprendere perché e come vivono gli organismi. La conoscenza del genoma quindi è utile in vari campi tra cui la biologia e la medicina, in quanto la scoperta di tecniche di sequenziamento ha permesso di accelerare la ricerca. Infatti, il sequenziamento permette di diagnosticare malattie ereditarie e con il genoma degli agenti patogeni si possono sviluppare medicine contro le malattie contagiose. L'esistenza dei geni era stata ipotizzata dal biologo Gregor Mendel nel 1866, che aveva notato i geni all'interno di incroci di piante o animali con tratti distintivi unici, che poi apparivano in rapporti definiti anche nella generazione successiva. Inizialmente non era chiaro come così poche basi azotate potessero essere sede di tutte le informazioni genetiche ma dopo la scoperta della struttura a doppia elica del DNA, fatta da Watson e Crick nel 1953, si realizzò che l'informazione risiedeva nella successione delle 4 basi. Nel corso degli anni c'è stata anche la significativa riduzione del costo dell'operazione di sequenziamento del DNA, che è passato da circa 100 milioni di dollari nel 2001 ai circa 1000 dollari odierni, ha permesso di non relegare questa tecnologia esclusivamente alla ricerca scientifica ma di utilizzarla anche in altri settori. Il sequenziamento, in medicina, viene utilizzato in campo diagnostico e terapeutico, tanto che in alcuni casi si può addirittura parlare di medicina personalizzata. In altri campi, come quello forense, il sequenziamento viene utilizzato per ottenere un tracciamento dei profili dei criminali; in quello alimentare, si può realizzare il tracciamento del DNA per evitare eventuali truffe e riconoscere la presenza di ingredienti non consentiti o di origine non propria; in campo animale e vegetale, il

sequenziamento ha lo scopo di caratterizzare le variazioni genetiche responsabili di determinate caratteristiche morfologiche in modo tale da poter selezionare le specie più vantaggiose in termini produttivi o di adattamento a uno specifico ambiente. La realizzazione del sequenziamento delle comunità microbiche presenti in un campione si dimostra utile anche per identificare tutti i microorganismi presenti in diverse matrici quali suolo, alimenti o in campioni biologici estratti da un animale. Il sequenziamento si è poi evoluto nel corso degli anni perché partendo dallo studio dei geni (genomica) si è arrivati fino a sviluppare altre tecniche tra cui:

- un'analisi di tipo funzionale per valutare l'espressione di trascritti differenziali (trascrittomica);
- una valutazione delle modifiche di elementi regolativi, tipo lo small-non-coding RNA, di metilazioni differenziali del DNA e modifiche di istoni, utili alla regolazione dell'espressione genica (epigenomica);
- un'analisi delle differenti comunità microbiche (metagenomica).

Si sono anche sviluppati algoritmi in grado di analizzare i dati e interpretare i risultati attraverso la cosiddetta bioinformatica. Tramite l'analisi metagenomica, argomento di cui si parlerà in questo elaborato in riferimento a due diverse tecniche, ossia il 16S rRNA amplicon sequencing e il Whole Genome Shotgun sequencing (WGS), si possono valutare la microflora e la microfauna del suolo e la loro interazione con l'ambiente e con la biodiversità animale e vegetale. Inoltre l'analisi metagenomica permette di poter definire la composizione batterica del rumine e il possibile contributo del microbiota nella salute animale e di determinare le caratteristiche funzionali e organolettiche di un alimento. Con la caratterizzazione genomica e trascrittomica delle specie di interesse, invece è possibile capire le basi biochimiche, fisiologiche e molecolari dell'adattamento della pianta alle variazioni ambientali.

Questa continua evoluzione del sequenziamento ha permesso alla comunità scientifica di avere una quantità sempre più consistente di informazioni a disposizione.[2]

1.3 Due diversi metodi di sequenziamento: 16S rRNA Amplicon Sequencing e Whole Genome Shotgun sequencing (WGS)

L'analisi della composizione del microbiota negli esseri umani e negli animali sta diventando una componente importante nella scoperta delle malattie. Per esempio, se prendiamo in considerazione un tratto dell'intestino umano, possiamo constatare come i cambiamenti all'interno della sua composizione abbiano una correlazione con alcune patologie, quali, ad esempio, la depressione, l'autismo, alcune allergie e alcune malattie infiammatorie croniche. Alla luce di quanto sopra, lo scopo di questa tesi è quello di individuare quali siano gli strumenti bioinformatici più adatti per inferire il profilo funzionale partendo dal dato ottenuto tramite 16S amplicon sequencing, dopo aver fatto alcune considerazioni preliminari riguardo lo stesso e il Whole Genome Shotgun sequencing e dopo un'analisi comparativa tra diversi strumenti bioinformatici. Questo confronto verrà realizzato sfruttando l'output ricavato da questi strumenti e paragonandolo con quello che si ottiene tramite il WGS (gold standard).

Fino ad oggi sono state implementate due tecniche per l'analisi della comunità microbiotica ossia il 16S rRNA amplicon sequencing e il whole genome shotgun sequencing (WGS). Per l'utilizzo di queste tecniche è necessario utilizzare un campione di DNA che poi verrà poi sequenziato in due modi diversi a seconda della tecnica applicata.

Il 16S rRNA sequencing si basa sull'utilizzo della Polymerase Chain Reaction (PCR) per l'amplificazione di una specifica regione nel gene 16S utilizzando dei primer degenerati e questo permette di sequenziare la regione e riconoscere i geni altamente conservati. la PCR rappresenta una delle tecniche più utilizzate per il sequenziamento anche se non si può utilizzare in ogni situazione.

Per utilizzare la PCR è necessario conoscere la sequenza genomica e affinché avvenga la reazione di polimerizzazione è necessario utilizzare due primer ossia delle corte sequenze nucleotidiche complementari alla sequenza che si vuole amplificare. Queste due sequenze si legano a inizio e fine filamento che si vuole amplificare, in seguito le due eliche di DNA vengono separate mediante calore. Ogni elica costituisce così un bersaglio per un primer che viene riconosciuto dalla TAQ polimerasi (una DNA polimerasi che lavora ad alte temperature) e essa poi sintetizza un filamento complementare alla sequenza bersaglio. Il principale limite di questa tecnica consiste nel fatto che l'annotazione si basa su una presunta associazione tra il gene e il taxon, definita Operational Taxonomic Unit (OTU). Gli OTU sono analizzati a livello di genere perciò l'analisi può essere meno precisa a livello della specie. La classificazione dei taxa, ottenuta tramite il 16S, sarà più efficiente se si considera quel determinato livello mentre si otterranno risultati meno interessanti e con una minore accuratezza a livello della specie.

Il sequenziamento 16S è una tecnica valida e ben definita che permette di ricavare una quantità sufficiente di informazioni riguardo la comunità microbiotica partendo da un numero piccolo di sequenze per campione.
Il WGS, invece, è una tecnica che si basa sull'utilizzo di primer scelti in maniera casuale per andare a sequenziare delle regioni del genoma ed è più accurato nell'analisi al livello della specie, in quanto i taxa possono essere definiti in maniera più accurata a quel livello.

1.4 Principali differenze tra 16S e WGS

Una differenza significativa tra le due tecniche sopra menzionate riguarda il procedimento utilizzato per effettuare il sequenziamento: nel WGS, detto procedimento si basa su un approccio di allineamento tramite una mappatura indipendente delle singole read sul database metagenomico di riferimento, mentre la tecnica 16S si basa su un algoritmo di de-replicazione che raggruppa sequenze simili prima di effettuare l'assegnazione di esse a un taxon. Detto algoritmo si avvantaggia sia dell'analisi k-mer, utile per identificare velocemente i migliori candidati all'interno del database di riferimento, sia dell'allineamento a coppie realizzato per tutta la lunghezza delle sequenze, utile per identificare il miglior matching tra quelli offerti dai vari candidati selezionati in precedenza. La tecnica 16S si serve di differenti metodi - QIIME2 e MICCA- per effettuare l'assegnazione delle sequenze a un taxon mentre il WGS utilizza Metaphlan2, il quale permette di settare una soglia di abbondanza per scegliere quali taxon riportare all'interno della tabella finale, cosa che tramite i metodi usati all'interno del 16S non avviene.
Un'altra differenza tra le due tecniche è la possibilità fornita dal WGS di sequenziare batteri, funghi, virus e altri microorganismi, mentre il 16S permette di identificare solo i batteri.
Il WGS si differenzia dal 16s anche perché, oltre al sequenziamento, esaminato in precedenza, permette di ottenere una predizione del profilo funzionale dei dati sequenziati mentre tramite il sequenziamento 16S questa possibilità non è inclusa, infatti per ottenere questa predizione si utilizzano le tabelle ASV, ottenute con esso, come dati in input per i tool considerati in questo lavoro di tesi. Una scelta tra le due tecniche 16S e WGS, dipenderà, in ultima analisi, dalla natura dello studio, in particolare:

- Il 16S è più adatto quando si ha un gran numero di campioni grazie al suo basso costo per campione ma ha una risoluzione funzionale e tassonomica limitata rispetto al WGS, anche se è caratterizzato da un workflow ben delineato

- Il WGS offre la possibilità di avere una migliore risoluzione grazie alla capacità più elevata di trovare le specie batteriche oltre a quelle dei virus, avendo di contro un workflow per l'analisi non ben determinato;
- Il WGS consente di andare a sequenziare delle regioni più ampie del genoma mentre il 16S permette di farlo solo per singole regioni;
- Le due tecniche utilizzano anche database di riferimento diversi e l'abbondanza delle specie ricavata con il 16S è evidentemente inferiore a quella ottenuta con il WGS. La tabella sottostante evidenzia, grazie all'utilizzo di tre metriche diverse di valutazione, che, considerando diversi modelli di machine learning, il WGS risulta migliore rispetto al 16S. [3]

	WGS			16S		
Model	% Accuracy	Log Loss	AUC	% Accuracy	Log Loss	AUC
Random Forest	80	0.626	0.790	79	0.602	0.811
Bagging	75	0.737	0.655	47	0.838	0.439
Decision Tree	50	17.269	0.550	68	10.906	0.633
SVC	50	0.705	0.460	58	0.679	0.500
LDA	50	7.131	0.580	59	6.307	0.583
Gradient Boosting	55	1.506	0.690	79	1.020	0.844

Figura 1: tabella con il confronto tra WGS e 16S rispetto a tre diverse metriche (immagine presa da: https://biodatamining.biomedcentral.com/articles/10.1186/s13040-021-00270-x#additional-information)

1.5 Il dato metatassonomico e metagenomico

Prima di definire una tabella ASV, ovvero l'output che si otterrà tramite il 16S, è necessario definire cos'è un'ASV ovvero una Amplicon Sequence Variant. Le Amplicon Sequence Variant rappresentano dei cluster (raggruppamenti) di varianti di sequenze simili tra loro, ognuno di essi aventi lo scopo di rappresentare l'unità tassonomica di una specie batterica o di un gene in relazione a una soglia di similarità tra sequenze. Generalmente questa soglia di similarità viene posta al valore del 97%. Le tabelle ASV quindi contengono il numero di sequenze che vengono osservate per ogni ASV in ogni campione. Le colonne abitualmente rappresentano i campioni mentre le righe rappresentano il genere o la specie specifica per un ASV.

Capitolo 2 - Inferenza del potenziale funzionale partendo dal dato 16S

2.1 Stato dell'arte

Questo capitolo illustra e descrive il funzionamento di alcuni strumenti bioinformatici che permettono di ottenere l'inferenza del potenziale funzionale partendo dalla tabella di abbondanza degli ASV ricavata tramite il 16S Amplicon Sequencing. Questi strumenti sono dei pacchetti che utilizzano diversi linguaggi di programmazione tra cui R e Python. Ogni strumento è caratterizzato da una pipeline differente in quanto gli step necessari per arrivare al profilo funzionale del dato fornito in ingresso sono diversi sia in numero che in svolgimento, per esempio il primo che verrà considerato utilizza un albero filogenetico per raggiugere l'output mentre il secondo si basa su delle matrici di associazione.

I principali strumenti che vengono utilizzati per la predizione del potenziale funzionale, partendo dal dato 16S, sono i tre seguenti ovvero: PICRUSt2, Tax4Fun2 e PanFP, selezionati tra i molti altri strumenti descritti nella letteratura scientifica in quanto sono quelli che sembrano essere più promettenti, dato che permettono di ottenere predizioni funzionali migliori di quelle ottenute con gli altri non citati.

2.2 I software per l'inferenza del potenziale funzionale

2.2.1 PICRUSt2

Il primo strumento preso in considerazione dall'analisi condotta è il PICRUSt2, il quale è caratterizzato da diversi step che permettono di posizionare le sequenze in una struttura filogenetica di riferimento, di utilizzare un ampio database, di fare delle predizioni stringenti e di predire i fenotipi. Per poter costruire la struttura filogenetica di riferimento sono necessari tre strumenti e più precisamente: HMMER, EPA-ng e GAPPA. L'HMMER è utilizzato per posizionare gli ASV (Amplicon Sequence Variant)/OTU(Operational Taxonomic Unit), l'EPA-ng si usa per determinare la loro posizione ottima e infine il GAPPA serve per ottenere il nuovo albero filogenetico che li incorpora nella loro corretta posizione. All'interno di quest'albero filogenetico sono presenti sia il genoma di riferimento sia il dato 16S che viene utilizzato per trovare il numero di copia della famiglia genetica per ogni ASV/OTU.

Dopo sarà possibile ricavare la mappatura del numero di copia sull'albero filogenetico per ottenere il contenuto metagenomico. In seguito, ogni ASV potrà essere corretto tramite il suo numero di copia e poi moltiplicato per le sue predizioni funzionali per ottenere il metagenoma predetto. PICRUSt2 permette di ricavare anche il contributo che ogni ASV/OTU fornisce ad ogni funzione predetta e questo consentirà di condurre un'analisi statistica. Infine le abbondanze dei pathway vengono ricavate basandosi sulla mappatura del pathway strutturato. Il database di riferimento di default utilizzato da questo strumento è l'Integrated Microbial Genomes (IMG) database. Per validare queste predizioni sarà necessario confrontare l'inferenza del potenziale funzionale ottenuto per ogni dataset e l'abbondanza ottenuta tramite il WGS. Per realizzare questo confronto si considerano i profili funzionali ottenuti partendo da tre dataset estratti dall'HMP e sei dataset non umani. [4]

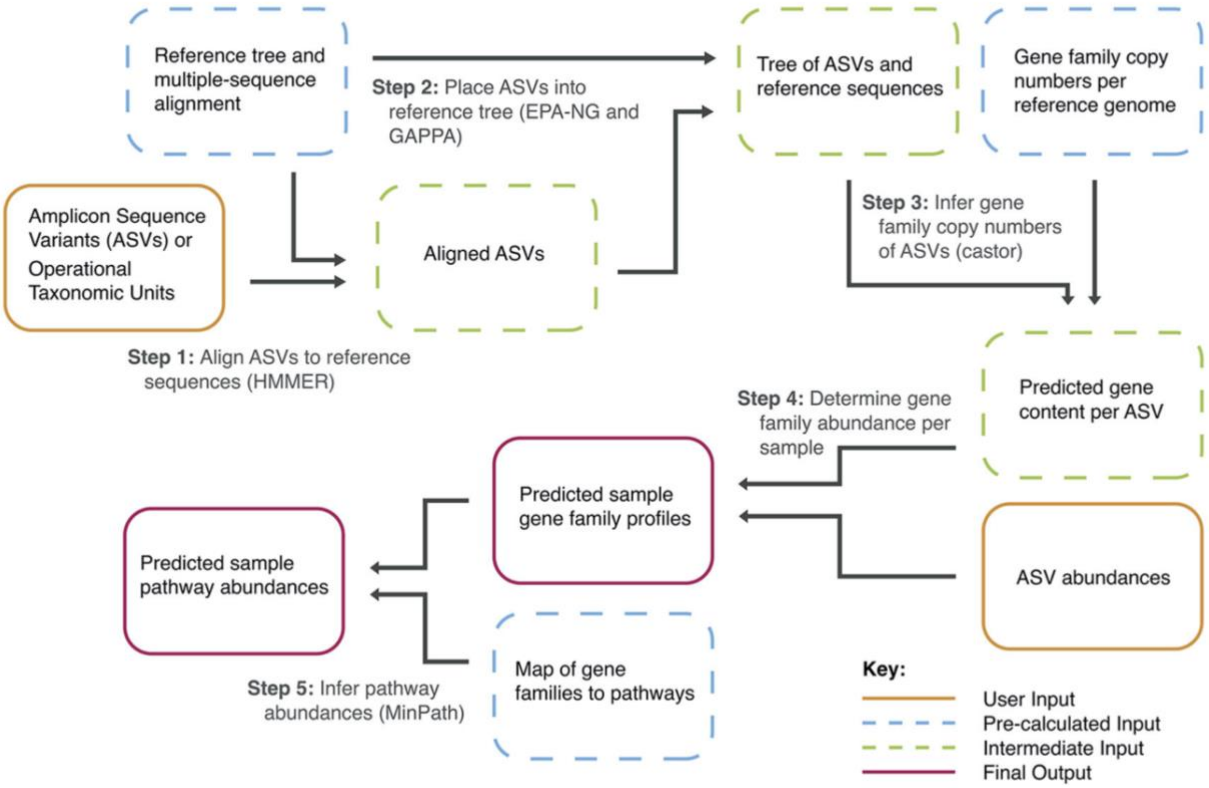


Figura 2: Workflow dello strumento PICRUSt2 (immagine pressa da: https://idp.nature.com/authorize/casa?redirect_uri=https://www.nature.com/articles/s41587-020-0548-6&casa_token=1NhyNpYiuT4AAAAA:HJ4Q6bME85x66OqiKYQxCtb-miESti7P_htt6VNZHzPWRzz5X5PVVnf78PEgVAmN8kesVPwmec1PJvh0Vw)

2.2.2 Tax4Fun2

Tax4Fun2 è un pacchetto R, molto semplice da usare e altamente efficiente a livello di memoria. Tramite questo strumento è possibile considerare dei dati definiti dall'utente e dei dati specifici per ogni habitat. Se i dati, così ottenuti, vengono utilizzati, sarà possibile ricavare risultati migliori mentre la possibilità di generare dei database specifici per ogni habitat permetterà di avere un potere predittivo più elevato. Tax4Fun2 viene utilizzato per ottenere il potenziale funzionale partendo dai dati 16S in input (una tabella di abbondanza degli ASV) sfruttando anche una matrice di associazione. Durante lo step iniziale gli ASV della tabella di abbondanza vengono ricercati all'interno del database di riferimento tramite BLAST utilizzando la pipeline di *runrefblast*, che permette di ricavare il miglior matching tra gli OTU in ingresso e il database di riferimento. Se invece si considera un dataset generato dall'utente allora la ricerca del miglior matching avviene utilizzando quest'ultimo. L'inferenza del profilo funzionale invece si ottiene tramite la pipeline di *makefunctionalprediction,* la tabella ASV viene riassunta basandosi sui risultati della ricerca del miglior matching. Viene successivamente ricavata una matrice di associazione per la tabella ottenuta in precedenza, che conterrà solamente i profili funzionali di riferimento per gli ASV presenti nella tabella. Prendendo in considerazione ogni singolo campione, l'informazione relativa alla sua abbondanza, estratta dalla tabella degli ASV, e la rispettiva informazione funzionale, presente nella matrice di associazione, vengono convertite in un profilo funzionale specifico per ogni campione. Infine i profili predetti vengono schematizzati basandosi sui pathway della Kyoto Encyclopedia of Genes and Genomes (KEGG). Verranno inferiti i profili funzionali solo degli ASV che superano una certa soglia di similarità, ossia quella del 97%. La matrice di associazione viene generata con, all'interno, il profilo funzionale di riferimento per ogni ASV considerato. Le matrici di associazione di riferimento sono la Ref99NR e la Ref100NR e ciascuna delle due consiste in una matrice con sequenze 16S associate al loro profilo di riferimento funzionale. Una volta ottenute, le predizioni tramite Tax4Fun2 verranno validate tramite un confronto con i risultati ottenuti tramite WGS partendo dagli stessi dataset. [5]

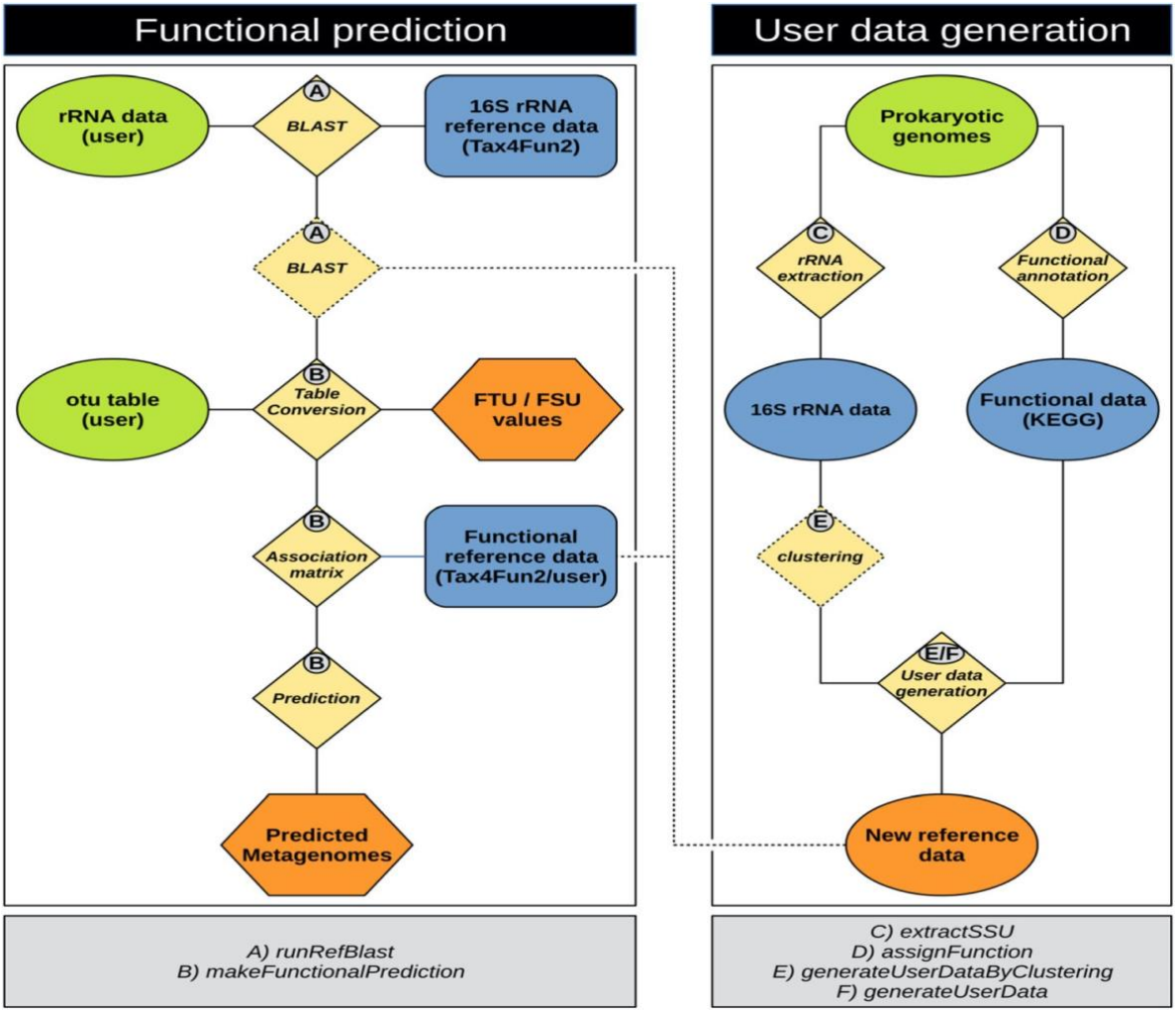


Figura 3: Workflow dello strumento Tax4Fun2 (immagine presa da : https://environmentalmicrobiome.biomedcentral.com/articles/10.1186/s40793-020-00358-7)

2.2.3 PANFP: pangenome-based functional profiles for microbial communities

Questo strumento considera il profilo di abbondanza misurato per gli ASV presi in considerazione e consente di ricavare un profilo funzionale con l'abbondanza che si ritiene corretta per la comunità considerata. Per poter usufruire di PanFP è necessario che nella tabella degli ASV sia presente anche una colonna relativa al lignaggio degli stessi, utile a esplicare le capacità funzionali della comunità considerata tramite la rappresentazione dei geni. Anche il PanFP necessita di un confronto per essere validato, realizzato sempre con le predizioni ottenute a partire dal WGS considerando come input da fornire i medesimi dati. Questo strumento non è limitato dall'utilizzo di un unico set di genomi di riferimento ma permette di integrare tutte le annotazioni funzionali del genoma all'interno della pipeline.

Il primo step consiste nella rifinitura del lignaggio tassonomico degli ASV. In questa fase si compiono due attività in modo tale da impiegare totalmente il database composto da procarioti caratterizzati da tassonomia e annotazione funzionale, le quali vengono estratte dall'National Center for Biotechnology Information (NCBI). Quando ci sono delle discrepanze tra il lignaggio tassonomico degli ASV, generato dall'utente, e quello corrispondente estratto dall'NCBI, infatti, questo viene corretto, assegnandogli quello presente all'interno dell'NCBI. Una volta fatto questo, si va a rifinire il lignaggio tassonomico degli ASV partendo dal livello più basso, fino a quando almeno il genoma di un procariote appartenente al taxon non viene identificato. La rifinitura dei lignaggi tassonomici sarà poi utile per l'inferenza funzionale.

Nel secondo step, invece, viene costruita una tabella funzionale prendendo in considerazione anche il lignaggio degli ASV; per ogni ASV, PanFP costruirà una rappresentazione dell'insieme di geni presenti all'interno dell'organismo, mediante un raggruppamento dal dataset di genomi procarioti relativi al lignaggio tassonomico dell'ASV considerato: gli ASV con lo stesso lignaggio, infatti, sono rappresentati dallo stesso insieme di geni. PanFP deriva poi il profilo funzionale dell'insieme di geni. Le funzioni condivise da più organismi sono quelle con una maggiore ricorrenza, utile anche per normalizzare il profilo funzionale tramite un insieme di organismi raggruppati secondo il lignaggio:

$$fq\big(lineage(i), KO(j)\big) = \frac{1}{\mathrm{O(i)}} * occurence(KO(j) in\ pangenome(i))$$

dove il pangenome(i) è l'insieme dei geni per il lignaggio e O(i) il numero di organismi raggruppati per quel lignaggio.

Il terzo step è quello che viene utilizzato per andare a normalizzare l'abbondanza degli OTU considerati in quanto essa riflette il numero totale di occorrenze di organismi assegnati a ogni OTU. Per realizzare questo passaggio si dividerà l'abbondanza degli ASV andando a considerare il numero di copia del gene 16S degli organismi raggruppati secondo il lignaggio tassonomico. In seguito le frequenze degli ASV per un determinato campione vengono normalizzate tramite la dimensione del campione cosicché le frequenze risultanti possano essere comparate tra loro.
Il quarto step servirà per convertire la tabella degli ASV relativi ai campioni in una tabella che contiene sia il lignaggio sia il suo campione relativo. Questa conversione avviene tramite la somma delle frequenze degli ASV con lo stesso lignaggio, seguendo la formula:

$$fq\big(lineage(i), sample(i)\big) = \sum_{OTU(k) has\ lineage} fq(OTU(k), sample(j))$$

Il quinto step permette di fare la conversione della tabella, contenente il lignaggio e il campione corrispondente, in quella contenente la funzione e il campione corrispondente. Si otterrà la tabella con la funzione e il campione andando a combinare il profilo funzionale del lignaggio con i relativi pesi corrispondenti all'abbondanza del lignaggio nel campione:

$fq(KO(i), sample(i)) = \sum_{lineage(k)} fq(lineage(k), sample(j))$*fq(lineage(k), KO(j))

Il sesto e ultimo step serve per valutare l'incertezza nella stima della tabella contenente sia la funzione sia il campione. Per operare questa valutazione si fanno delle assunzioni, tra cui quella per cui ogni profilo di abbondanza di un ASV contribuisce al profilo funzionale in maniera indipendente. Per realizzare questa valutazione viene fatto un bootstrapping che permette di selezionare gli ASV con la relativa abbondanza all'interno dei campioni in maniera randomica. Alcuni ASV possono essere presenti una sola volta o più volte o addirittura non essere presenti nel campione considerato. [6]

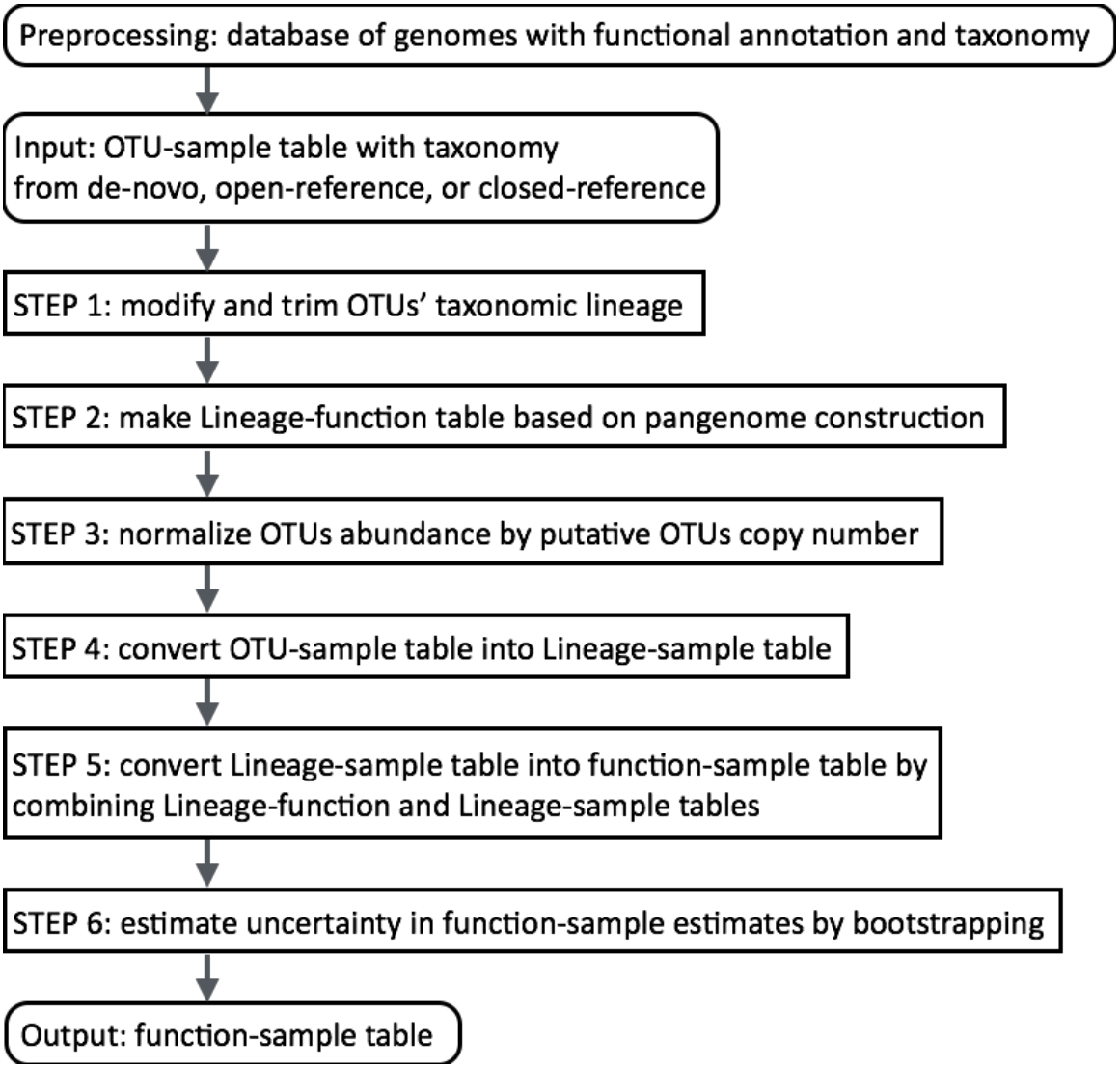


Figura4: Workflow dello strumento PanFP (immagine presa da: https://bmcresnotes.biomedcentral.com/articles/10.1186/s13104-015-1462-8#Sec17)

2.3 Vantaggi e svantaggi dei diversi approcci

Ognuno di questi tre tool è caratterizzato da diversi punti di forza e da alcune limitazioni. Se prendiamo in considerazione il PICRUSt2, esso permette di generare gli output collegati a uno score di confidenza ovvero l'NSTI, di implementare diversi metodi di Hidden State Prediction (HSP), di analizzare anche dati 18s e ITS ed è descritto da una documentazione esaustiva; di contro, però, si potranno ottenere errori nella fase di posizionamento delle sequenze, dovuti a una bassa risoluzione a livello di rRNA. Il Tax4Fun2, invece, sfrutta un algoritmo basato sulla minima similarità tra le sequenze e viene implementato tramite degli script di R efficienti a livello di memoria. Tax4Fun2 è utilizzabile anche con altre piattaforme, genera score di confidenza (FTU e FSU), si riferisce al Kegg Orthologs, in continuo aggiornamento dal 2018, calcola la ridondanza di specifiche funzioni e consente di utilizzare dei database di riferimento specifici per un particolare habitat; di contro non è ancora disponibile per il microbiota degli eucarioti. Infine, PanFP utilizza un profilo funzionale di un insieme di geni, quindi è meno sensibile al trasferimento orizzontale dei geni; di contro, però, non genere nessun tipo di score di confidenza e non è ancora disponibile per il microbiota degli eucarioti.

Strumenti	**Set per la validazione**	**Linguaggio di programmazione**	**Validazione**
PICRUSt2	3 dataset estratti da HMP 6 dataset non umani	Python	Confronto con i risultati del WGS
Tax4Fun2	3 dataset estratti da HMP 6 dataset non umani	R package	Confronto con i risultati del WGS
PanFP	3 dataset estratti da HMP 6 dataset non umani	Python	Confronto con i risultati del WGS

Capitolo 3 - Dataset utilizzati per l'analisi

Per la nostra analisi sono stati considerati 9 dataset differenti, alcuni provenienti da individui umani, altri provenienti dal suolo e anche da mammiferi. Questi dataset, come verrà spiegato in seguito, sono stati estratti da diversi articoli scientifici e si possono suddividere in dataset di origine reale e dataset di origine simulata: i dataset Blueberry, Cameroon, HMP, Indian, Mammal, MammalianGut, Ocean e Primate sono quelli di origine reale mentre il MammalsSimulated, come suggerisce il nome stesso, è l'unico di origine simulata.

3.1 Dataset reali

I dataset che verranno considerati per svolgere l'analisi sono stati estratti dal Sequence Read Archive (SRA), dall'European Nucleotide Archive (ENA), da MG-RAST e dal QIITA database. L'SRA rappresenta l'archivio della National Library of Medicine di dati di sequenziamento e fa parte dell'International Nucleotide Sequence Database Collaboration (INSDC). Questo include l'NCBI SRA, l'European Bioinformatics Institute (EBI) e il DNA Database of Japan (DDBJ). Se ad una qualsiasi delle 3 organizzazioni vengono inviati dei dati questi vengono condivisi immediatamente in maniera automatica a ciascuna delle altre due organizzazioni. L'ENA, invece, è un archivio che permette di avere accesso gratuito a una grande quantità di sequenze di DNA e di RNA annotate. Tramite l'ENA è anche possibile memorizzare informazioni complementari come procedure sperimentali, dettagli dell'assemblaggio di sequenze e altri metadati relativi ai progetti di sequenziamento. MG-RAST, abbreviazione di Metagenomic Rapid Annotations using Subsystems Technology, rappresenta un server di applicazioni web open-source che suggerisce l'analisi filogenetica e funzionale dei metagenomi, oltre ad essere uno dei più grandi depositi di dati metagenomici. I dataset che verranno presentati di seguito sono stati ottenuti da tre differenti articoli scientifici. I dataset di Blueberry, Cameroon, HMP, Indian, Mammal, Ocean e Primate sono stati ricavati dall'articolo "PICRUSt2 for prediction of metagenome functions" di Gavin M. Douglas et al. [4]. Il dataset del MammalianGut è stato tratto dall'articolo "Tax4Fun2: prediction of habitat-specific functional profiles and functional redundancy based on 16S rRNA gene sequences" di Franziska Wemheuer et al. [5]. Per arrivare alla tabella ASV finale ad ogni dataset vengono poi applicate diverse pipeline e diversi criteri di filtraggio in quanto tra essi ci sono differenze tecniche.

3.1.1 Blueberry

I campioni del Blueberry presi in considerazione sono stati estratti dal Sequence Read Archive (SRA) tramite i seguenti numeri di progetto: PRJNA484230 per il shotgun metagenomics (MGS) e PRJNA389786 per il 16S. Questi campioni vengono prelevati dal suolo, all'interno del quale si trovano i semi per la nascita della pianta, la quale genererà poi i mirtilli. I campioni presi in considerazione sono 22 e rappresentano le colonne della tabella ASV, mentre le ASV considerate sono 3333 e rappresentano le righe. Il sequenziamento 16S, per ottenere la tabella ASV, viene realizzato sulla regione V6-V8, utilizzando Deblur e la tecnologia Illumina MiSeq. Tramite il sistema MiSeq è possibile effettuare con un unico strumento la generazione di cluster, l'amplificazione, il sequenziamento e l'analisi dei dati. La potenza di questo sistema fa sì che lo stesso venga utilizzato per generare il 90% del sequenziamento. Il sistema MiSeq è caratterizzato da un flusso di lavoro suddiviso in 4 fasi e precisamente:

- la fase di preparazione delle librerie, in cui vengono aggiunti gli adattatori di sequenziamento ai frammenti di DNA e vengono preparate le librerie per quest'ultimo;
- Il sequenziamento vero e proprio in cui la libreria viene trasferita alla cella a flusso;
- l'analisi dei dati in cui avviene l'elaborazione dei dati, la segnalazione delle variabili genomiche;
- la fase di interpretazione e segnalazione del contesto biologico.

Le librerie per il sequenziamento vengono preparate rapidamente. Dopo aver preparato le librerie si passa ad amplificare, sequenziare e identificare le basi qualitativamente valutate. La qualità dei dati viene garantita tramite la chimica Sequencing By Synthesis (SBS) di Illumina. Questo metodo è basato su terminatori reversibili e permette di sequenziare in contemporanea miliardi di frammenti di DNA, andando a rilevare singole basi finché queste vengono incorporate all'interno dei filamenti di DNA in estensione. Quando vengono aggiunti, i dNTP permettono di acquisire i coloranti fluorescenti, in seguito i dNTP vengono scissi per permettere di incorporare la base successiva. La misurazione dell'intensità del segnale permette di identificare la base durante ogni ciclo e questo permette di ridurre i tassi di errore dei dati non elaborati rispetto ad altre tecnologie. Alla fine si otterrà un sequenziamento base per base molto accurato.

Il WGS utilizza un'altra tecnologia, ossia l'Illumina NextSeq, che si basa sulla piattaforma Illumina. Questa consiste nella preparazione della libreria per il sequenziamento che parte dalla denaturazione della doppia elica di DNA e l'aggiunta di sequenze adapter alle estremità, grazie alle quali si ottiene un filamento stampo pronto per il sequenziamento. Tutte le reazioni avverranno all'interno di una particolare microcella a flusso formata da otto canali indipendenti, su cui ci sono due oligonucleotidi.

Quando i frammenti di DNA vengono inseriti nella cella avviene una ibridazione tra le sequenze adapter e le complementari legate alla microcella. Questi legami si formano in entrambe le estremità dei frammenti, permettendo così di assumere una forma a ponte. Dopo che il frammento è stato immobilizzato, inizia il processo di amplificazione ovvero la DNA polimerasi inizia a sintetizzare il filamento complementare, che viene denaturato e legato alla cella, e si procede così in maniera ciclica fino a ottenere migliaia di cluster di frammenti, questo processo è chiamato bridge PCR. I frammenti di ogni cluster vengono poi sequenziati tramite l'impiego di un oligonucleotide ovvero il primer. Durante ogni sequenziamento saranno necessari la DNA polimerasi e i quattro dNTP legati a molecole fluoroscenti e un terminatore reversibile, utile per bloccare il processo quando sarà finito. L'utilizzo dei 4 dNTP permette la sintesi di più basi in contemporanea e serve un terminatore reversibile che sia in grado di arrestarla, in quanto è una molecola che blocca il processo di sintesi dopo che è avvenuta l'incorporazione di una singola base. Dopo ogni incorporazione, una luce laser eccita il fluoroforo coniugato al dNTP, in modo tale da provocare un'emissione luminosa che ne permetterà l'identificazione. Dopo questo il terminatore viene rimosso e si passa a sequenziare la base successiva. Una volta completato il sequenziamento si utilizza Deblur per ottenere la tabella ASV.
In questo dataset non verrà effettuato un post-processing e non saranno rimossi campioni se non quelli che non sono comuni tra il dato 16S e il dato WGS. La sparsità di questo dataset è 0.76, mentre la standard deviation di questo dataset è pari a 13.255.

3.1.2 Cameroon

I campioni del Cameroon presi in considerazione sono stati estratti dallo European Nucleotide Archive (ENA) tramite i seguenti numeri di progetto: PRJEB27005 per il shotgun metagenomics (WGS), mentre per i dati per il 16S vengono estratti dal MG-RAST, più precisamente dal progetto mgp15238. Il numero di questi campioni è di 69 ma per lo studio realizzato ne vengono considerati 57, che sono quelli comuni tra il dato 16S e il dato WGS, e rappresentano le colonne della tabella ASV. Le ASV sono in totale 4277 ma anche qui ne vengono considerate 4077 e rappresentano le righe della tabella. Questi campioni sono stati estratti da individui provenienti dal Camerun. Il sequenziamento 16S, per ottenere la tabella ASV, viene realizzato sulla regione V5-V6, utilizzando Deblur e la tecnologia Illumina MiSeq. L'MGS utilizza un'altra tecnologia ovvero l'Illumina HiSeq. Sia illumina MiSeq che Hiseq si basano sui principi di Illumina spiegati in precedenza. Una volta completato il sequenziamento, si utilizza Deblur per ottenere la tabella ASV. In questo dataset non verrà effettuato un post-processing e non saranno rimossi campioni se non quelli che non sono comuni

tra il dato 16S e il dato WGS. La sparsità di questo dataset è 0.87 e la standard deviation è pari a 404.279.

3.1.3 HMP

I campioni dell'HMP (Human Microbiome Project) sono stati estratti dal sito: https://www.hmpdacc.org/HMIWGS/healthy/. Il numero di questi campioni è 191 ma ne vengono considerati 137, che sono quelli comuni tra il dato 16S e il dato WGS, e rappresentano le colonne della tabella, mentre le ASV sono 1865 ma ne vengono considerate 1576 e rappresentano le righe della tabella. Il sequenziamento 16S, per ottenere la tabella ASV, viene realizzato sulla regione V4. Viene utilizzata DADA2 perché per il sequenziamento si utilizza la tecnologia Roche 454. Questa tecnologia rappresenta uno dei primi sequenziatori NGS lanciati sul mercato e si basa su una tecnologia di pyrosequencing, chiamata così a causa dei pirofosfati utili per la reazione di sequenziamento. Il Roche 454 è in grado di riconoscere le sequenze in seguito ai segnali luminosi provocati dal distacco di un pirofosfato ogni volta che un nucleotide viene incorporato nel filamento in crescita. I vari frammenti vengono denaturati per ottenere dei filamenti singoli che saranno poi catturati dagli amplification beads, sfere sulla cui superficie sono presenti degli oligonucleotidi, fondamentali per riconoscere e legare gli adattatori, aggiunti alle estremità di ogni frammento. Infatti, quando si parla di DNA library o sequencing library si intende un gruppo di frammenti di DNA preparati tramite l'aggiunta di adattatori. In seguito avverrà un processo cosiddetto di PCR in emulsione (emuslion PCR). Durante il quale gli amplificati vengono posizionati sulla picotiter plate, un supporto sul quale avviene la reazione di sequenziamento vera e propria. Gli amplificati vengono legati a una soluzione contenente un dNTP per volta, reagente per catalizzare la reazione luminosa. Quando il dNTP (dATP, dGTP, dCTP e dTTP) risulta complementare alla prima base da copiare, viene incorporato e così avviene la reazione luminosa causata dal rilascio di pirofosfato. Dopo, il dNTP viene degradato dall'enzima apirasi e lavato via e gli amplificati si legano al dNTP successivo. Le reazioni luminose vengono registrate ed elaborate dalla macchina e il software ricostruisce poi la sequenza. L'aggiunta di un dNTP alla volta viene chiamata interrogazione di singolo nucleotide. Una volta completato il sequenziamento, si applica DADA2 per ottenere la tabella ASV. L'MGS utilizza un'altra tecnologia ovvero l'Illumina HiSeq. Illumina Hiseq, che si basa sui principi di Illumina spiegati in precedenza Alcuni dei campioni vengono esclusi tramite il DADA2 perché caratterizzati da meno di 2000 reads e vengono anche escluse alcune ASV aventi una frequenza minima di 10. La sparsità di questo dataset è 0.98e la standard deviation di questo dataset è 35.364.

3.1.4 Indian

Tutti i campioni dell'Indian presi in considerazione sono stati estratti tutti dall'European Nucleotide Archive (ENA), dal progetto PRJNA397112. Il numero di questi campioni è 108 ed essi rappresentano le colonne della tabella ASV, mentre le ASV sono 2397 ma ne vengono considerate 2237 e rappresentano le righe della tabella. Questi campioni sono stati estratti da individui provenienti dall'India. Il sequenziamento 16S, per ottenere la tabella ASV, viene realizzato sulla regione V3, utilizzando la tecnologia Illumina NextSeq500 e in seguito Deblur. Anche questa tecnica si basa sui principi di Illumina. Una volta completato il sequenziamento, si applica Deblur per ottenere la tabella ASV. L'MGS utilizza la stessa tecnologia. In questo dataset non verrà effettuato un post-processing e non saranno rimossi campioni se non quelli che non sono comuni tra il dato 16S e il dato WGS. La sparsità di questo dataset è 0.91e la standard deviation di questo dataset è 1295.845.

3.1.5 Mammal

I campioni del Mammal sono stati estratti da Short Read Archive (SRA) tramite i numeri di progetto SRP115632 per l'MGS e SRP115643 per il 16S. Il numero di questi campioni è 11 e rappresentano le colonne della tabella ASV, mentre le ASV considerate sono 323 e rappresentano le righe della tabella. Questi campioni sono stati estratti da mammiferi. Il sequenziamento 16S, per ottenere la tabella ASV, viene realizzato sulla regione V6-V8, utilizzando la tecnologia Illumina MiSeq e Deblur. Dopo aver ottenuto le sequenze tramite Illumina MiSeq, si applica Deblur per ottenere la tabella ASV. L'MGS utilizza un'altra tecnologia ovvero l'Illumina HiSeq, entrambe le tecnologie si basano sugli stessi principi di Illumina. La sparsità di questo dataset è 0.75 e la standard deviation di questo dataset è 114.344.

3.1.6 Ocean

I campioni dell'Ocean sono stati estratti da Short Read Archive (SRA) tramite il numero di progetto SRP056891. Il numero di questi campioni è 6 ed essi rappresentano le colonne della tabella ASV, mentre le ASV considerate sono 1148 e rappresentano le righe della tabella. Questi campioni, come dice il nome stesso, sono stati estratti dall'acqua dell'oceano. Il sequenziamento 16S, per ottenere la tabella OTU, viene realizzato sulla regione V4, utilizzando la tecnologia Illumina MiSeq e Deblur. Dopo aver completato il sequenziamento tramite Illumina MiSeq, si applica Deblur per ricavare la tabella ASV. L'MGS utilizza un'altra tecnologia ovvero l'Illumina HiSeq. Entrambe le tecnologie si basano sui principi menzionati in precedenza. In questo dataset non verrà effettuato un post-processing e non saranno rimossi campioni se non quelli che non sono comuni tra il dato 16S e il dato WGS. La sparsità di questo dataset è 0.40 e la standard deviation di questo dataset è 1092.797.

3.1.7 Primate

I campioni del Primate, sia per l'MGS, sia per il 16S, vengono estratti dalla cartella QIITA tramite il codice di accesso 11212. Il numero totale di questi campioni è 154 ma ne vengono considerati 78, che sono quelli comuni tra il dato 16S e il dato WGS, e rappresentano le colonne della tabella OTU, mentre le ASV considerate sono 8251 ma ne vengono considerate 7452 e rappresentano le righe della tabella. Il sequenziamento 16S, per ottenere la tabella ASV, viene realizzato sulla regione V4, utilizzando la tecnologia Illumina MiSeq e Deblur. Una volta completato il sequenziamento, si utilizza Deblur per ottenere la tabella ASV. L'MGS utilizza un'altra tecnologia ovvero l'Illumina HiSeq. Entrambe le tecnologie si basano sugli stessi principi dell'Illumina, di cui si è parlato in precedenza. Questo dataset viene estratto dal QIITA database e anche in questo caso vengono solo esclusi i campioni non comuni tra 16S e WGS. La sparsità di questo dataset è 0.95 e la standard deviation di questo dataset è 26.903

3.1.8 Mammalian Gut

I campioni del Mammalian Gut presi in considerazione sono estratti dal MG-RAST tramite i codici 4,461; vengono considerati i campioni dal 284 al 301, quelli dal 341 al 355, il 357 e il 358, quelli dal 360 al 380 e infine il 383. Il numero di questi campioni è 56 e rappresentano le colonne della tabella ASV mentre il numero di ASV considerate è pari a 28218 e rappresentano le righe della tabella ASV. La sparsità di questo dataset è 0.97 e la standard deviation di questo dataset è 4.205.

3.2 Dataset simulato

L'unico dataset simulato, ossia quello del MammalsSimulated, è stato estratto dall'articolo "MicFunPred: A conserved approach to predict functional profiles from 16S rRNA gene sequence data" di Dattatray S. Mongad et al. [8] In questo articolo si è considerato solo questo dataset simulato poiché agli altri non era possibile applicare il software di PICRUSt2 in quanto non in grado di predire il profilo funzionale pertanto non sarebbe stato utile alla nostra analisi.
Questo dataset è stato simulato a partire dal genoma di tipo Mammal e le regioni variabili V3-V5, da cui si è partiti per generare le tabelle ASV, sono state estratte da sequenze 16S rRNA tramite l'Hidden Markov Model (HMM) implementato nell'V-Xtractor. Ogni variante di queste regioni è stata considerata come una ASV indipendente mentre i profili tassonomici sono stati generati usando la distribuzione randomica di Dirichlet, con un 30/50% di zeri e un 10/20% di sequenze inserite randomicamente in 100 campioni finti. I profili metagenomici, quindi quelli di riferimento ovvero il nostro gold standard (WGS) poi sono stati ottenuti attraverso una moltiplicazione di matrice tra il profilo tassonomico e la tabella con il contenuto dei geni. Questa tabella è stata costruita facendo riferimento a KEGG Orthology (KO), Cluster of genes (COG), TIGRFAM, Enzyme Commission (EC) e Protein Family (Pfam).

3.2.1 MammalsSimulated

Questi campioni sono estratti da mammiferi. Il numero di campioni considerati in questo caso è pari a 101 mentre le ASV considerate in questo caso sono pari a 116. Il numero di campioni rappresenta il numero di colonne della tabella ASV mentre poi il numero delle ASV considerate rappresenta il numero delle righe della tabella. La sparsità di questo dataset è 0.59 e la standard deviation di questo dataset è 758.602.

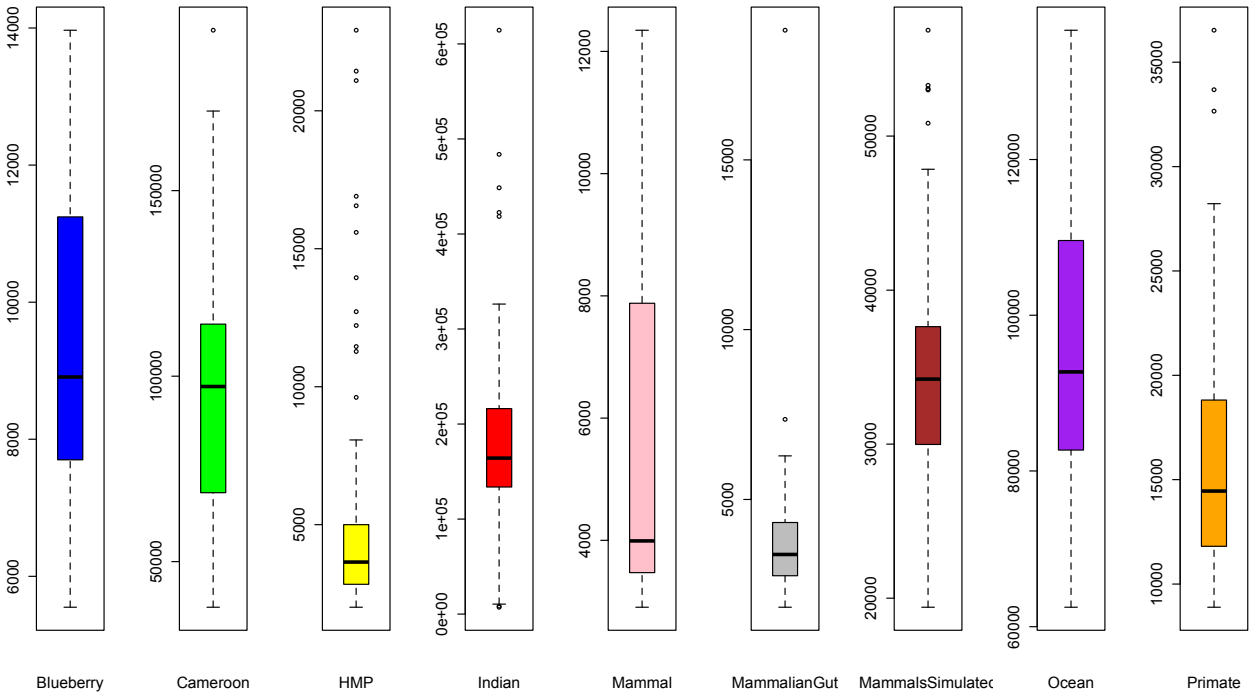


Figura 5: Boxplot per rappresentare la profondità di sequenziamento per colonne per ogni dataset.

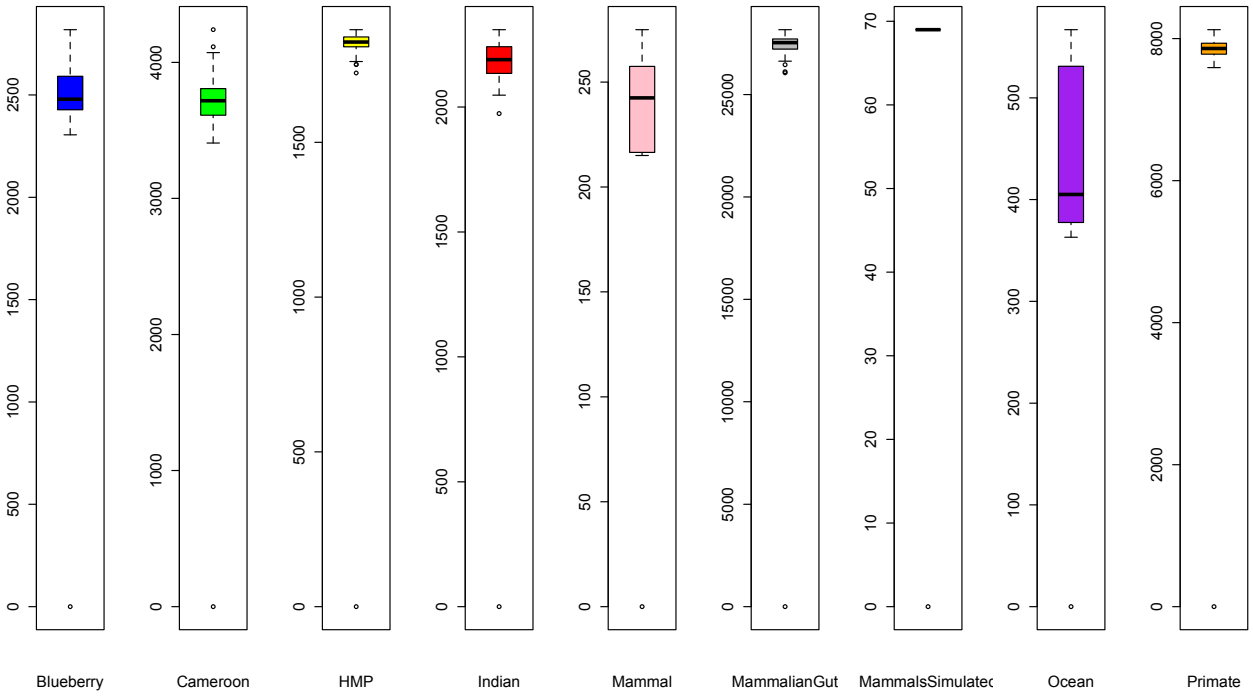


Figura 6: boxplot per rappresentare il numero di zeri per colonne per ogni dataset.

Nelle due figure precedenti è possibile vedere altre due caratteristiche dei dataset considerati ovvero la profondità di sequenziamento per quel che riguarda la prima figura e il numero di zeri nella seconda invece. Queste due caratteristiche sono state ricavate per ogni dataset e all'interno di ognuno sono state calcolate per colonna e poi sono stati rappresentati tramite i boxplot con le mediane che indicano appunto i valori medi del numero di zeri e i valori medi della profondità di sequenziamento per ognuno dei dataset. Si può vedere anche come per la profondità di sequenziamento i dataset MammalianGut e HMP siano caratterizzati da una profondità di sequenziamento più bassa degli altri. Mentre per quel che riguarda il numero di zeri i dataset caratterizzati dal valore più basso sono MammalsSimulated e Mammal.

Capitolo 4 - Metriche utilizzate per il confronto

All'interno di questo capitolo verranno presentate le metriche che sono state selezionate per effettuare il confronto, che si possono suddividere in 2 categorie:

1. Le metriche che considerano i dataframe in termini di presenza/assenza dei KO (geni), dopo aver effettuato una conversione in uni e zeri delle singole celle del dataframe. Esse sono: Balanced Accuracy, Precision, Recall, F_1_score, la distanza di Bray-Curtis binaria. I risultati che saranno ottenuti tramite queste analisi verranno rappresentati, per quel che riguarda le metriche utilizzate per valutare presenza/assenza dei KO all'interno dei campioni, tramite degli istogrammi ottenuti con la funzione ggplot (v. 3.4.1) in R e tramite PCoA ottenuta attraverso il pacchetto ecodist (v. 2.0.9) e clustering gerarchico con la funzione hclust;
2. Le metriche che considerano i dataframe in termini di abbondanza dei KO (geni) ossia considerano i dataframe ottenuti tramite i tool senza andare a fare conversioni. Queste sono rappresentate da Spearman's correlation e distanza di Bray-Curtis. Per rappresentare i risultati ottenuti tramite queste metriche verranno utilizzati i boxplot e poi nuovamente PCoA e clustering.

4.1 Metriche per valutare la capacità di identificare i geni presenti

Le metriche presentate nei seguenti paragrafi sono la Balanced Accuracy, la Precision, la Recall e l'F_1_score. Per calcolare i valori di queste metriche saranno necessari diversi parametri, tra cui il numero dei veri positivi (TP), dei veri negativi (TN), dei falsi positivi (FP) e dei falsi negativi (FN). Per veri positivi e veri negativi si intendono i KO (geni) identificati come presenti all'interno di un campione e la cui abbondanza sarà sostituita con un 1 mentre i veri negativi sono quelli che non sono presenti all'interno del campione e la cui abbondanza sarà sostituita con uno 0; i falsi positivi e i falsi negativi rappresentano i KO (geni) identificati in maniera errata come appartenenti a un campione o come non appartenenti ad esso. Per ricavare questi valori all'interno della nostra analisi sarà necessario andare a verificare la presenza o l'assenza dei geni all'interno della predizione ottenuta tramite i software presentati in precedenza (PanFP, PICRUSt2 e Tax4Fun2) e confrontarla poi con quella dei risultati ottenuti tramite WGS.

4.1.1 Balanced Accuracy

La Balanced Accuracy è una metrica che viene utilizzata per valutare un metodo di classificazione; per il suo calcolo vengono utilizzate due diverse grandezze ossia la Sensitivity (TPR) e la Specificity (TNR). Per quanto riguarda il nostro confronto la classificazione non avviene tra positivi e negativi ma bensì sulla presenza o sull'assenza dei geni all'interno dei risultati ottenuti tramite i software rispetto ai risultati ottenuti tramite WGS. In formula, possiamo esprimere la Balanced Accuracy come:

$$\text{BA} = 0.5 * (TPR + TNR)$$

$$\text{TPR} = \frac{TP}{(TP+FN)} \quad \text{TNR} = \frac{TN}{(TN+FP)}$$

4.1.2 Precision

La Precision è una metrica che serve per indicare l'accuratezza con cui i KO vengono identificati come positivi, appartenenti al campione. Essa è ottenuta tramite un rapporto tra i veri positivi (TP) e il numero totale di elementi etichettati come tali, dato dalla somma tra i veri positivi (TP) e i falsi positivi (FP, etichettati come tali in maniera errata). A questa metrica ci si può riferire anche come Positive Predictive Value (PPV). Se si dovesse ottenere un valore di Precision pari a 1, questo indicherebbe che tutti i risultati sono qualitativamente corretti, ossia se un elemento viene individuato come appartenente a una classe poi appartiene effettivamente alla classe, ma non direbbe nulla riguardo l'aver individuato tutti gli elementi. La Precision viene usata come una misura della qualità, infatti ottenere risultati che implicano una elevata Precision significa ottenere più risultati rilevanti(corretti) rispetto a quelli irrilevanti. In formula, possiamo esprimere la Precision come:

$$\text{Precision} = \frac{TP}{(TP+FP)}$$

4.1.3 Recall

La Recall è una metrica che serve ad individuare le classi positive correttamente predette, ovvero è ottenuta da un rapporto tra il numero di veri positivi e il numero totale di questi ultimi individuati, ottenuto tramite la somma dei veri positivi (TP) e dei falsi negativi (FN, etichettati in maniera errata in quanto avrebbero dovuto essere positivi). A questa metrica ci si può riferire anche come True Positive Rate o Sensitivity (TPR). Un valore di Recall pari a 1 significherebbe che tutti gli elementi rilevanti sono stati individuati, però questo valore non direbbe nulla sul perché vengano individuati anche elementi irrilevanti. La Recall può essere vista come una grandezza complementare all'errore di tipo II ovvero quello relativo ai falsi negativi. Essa è una misura della quantità e se ha un valore elevato significa che l'algoritmo permette di trovare la maggior parte dei risultati nella maniera corretta. All'interno dell'analisi questa metrica è stata calcolata considerando sempre la presenza/assenza dei geni ottenuta facendo l'inferenza tramite i software confrontata con la rispettiva ricavata tramite WGS. In formula, possiamo esprimere la Recall come:

$$\text{Recall} = \frac{TP}{(TP+FN)}$$

4.1.4 F_1_score

L'F_1_score rappresenta la media armonica tra Recall e Precision e anche questa metrica considera la distinzione tra presenza e assenza dei geni invece delle classi positivo e negativo. Anch'essa serve a valutare l'accuratezza della predizione ottenuta tramite i software rispetto ai risultati ottenuti tramite WGS. In formula, possiamo esprimere F_1 come:

$$F_1 = 2* \frac{Recall*Precision}{(Recall+Precision)}$$

4.1.5 Indice di Bray-Curtis

L'indice di Bray-Curtis è un indice di dissimilarità. Può essere utilizzato in ecologia o in biologia per confrontare tra loro due campioni in termini di abbondanza di taxa (phyla, specie, OTU) presenti in ciascuno di essi, in modo tale da valutare la loro dissimilarità. Questo indice è di tipo simmetrico. All'interno di questo paragrafo sarà considerato l'indice di Bray-Curtis binario. In formula, esso si può esprimere come:

$$\mathrm{BC} = \frac{(A+B-2*J)}{(A+B)}$$

All'interno di questa formula le variabili A e B indicano il numero di specie all'interno dei due siti comparati mentre la variabile J indica il numero di specie presente all'interno di questi due siti; i risultati ricavati tramite questa metrica verranno poi utilizzati per realizzare il Principal Coordinate Analysis plot e anche il clustering gerarchico.

4.1.5.1 Principal Coordinate Analysis

La Principal Coordinate Analysis è un metodo per andare a esplorare le similarità o le dissimilarità dei dati. Per farlo si parte da una matrice di dissimilarità e serve per assegnare a ogni elemento una posizione all'interno di uno spazio di dimensionalità ridotta. La PCoA cerca di trovare gli assi principali tramite una matrice. Permette di calcolare una serie di autovalori e autovettori e ce ne sono tanti quanti le righe presenti all'interno della matrice iniziale. Questa tecnica ci permette di individuare differenze di gruppo o tra i singoli elementi e permette anche di ottenere la suddivisione in cluster. Si basa su una matrice di dissimilarità che si ottiene applicando l'indice di Bray-Curtis a diversi campioni presenti all'interno del dataframe preso in considerazione. Questo tipo di analisi può anche essere applicato con altre matrici di dissimilarità ottenute tramite diverse distanze, tra cui ad esempio la distanza euclidea o la distanza Jaccard. Essa assegna a ciascun elemento una posizione in uno spazio a bassa dimensionalità, infatti noi considereremo solamente le due componenti principali per ottenere la rappresentazione grafica dei risultati. La PCoA ha lo scopo di trovare gli assi principali attraverso una matrice e calcola autovettori e autovalori. Ogni autovalore avrà un autovettore corrispondente e il numero di autovettori e autovalori sarà uguale al numero di righe nella matrice iniziale. La classificazione degli autovettori sarà fatta in maniera decrescente e il primo di essi sarà

identificato come l'autovalore dominante o principale. In conclusione, utilizzando la PCoA si possono visualizzare le differenze individuali e/o di gruppo e tramite le differenze individuali si possono mostrare i valori anomali.

4.1.5.2 Clustering

Il clustering o analisi dei gruppi rappresenta un insieme di tecniche di analisi multivariata dei dati finalizzata alla selezione e al raggruppamento di elementi omogenei in un insieme di dati. Questa analisi si basa su misure di dissimilarità tra i dati presi in considerazione, di conseguenza i risultati degli algoritmi di clustering dipendono molto dal tipo di metrica presa in considerazione per andare a valutare la dissimilarità all'interno dei dati. Due diverse filosofie sono alla base del clustering:

- La prima filosofia è rappresentata dal clustering dal basso verso l'alto. Essa prevede che tutti gli elementi vengano considerati, all'inizio, come elementi a sé stanti, mentre in seguito verranno uniti con i cluster più vicini. Questa procedura di unione dei cluster più vicini proseguirà fino ad ottenere un numero prefissato di cluster oppure fino a quando la distanza minima tra i cluster non supererà un certo valore o fino a quando non sarà rispettato un criterio statistico prefissato;
- La seconda filosofia è rappresentata dal clustering che va dall'alto verso il basso. Questa prevede che tutti gli elementi siano inizialmente posti all'interno di un unico cluster, che l'algoritmo inizia a dividere in tanti cluster di dimensioni inferiori. Questa suddivisione del cluster sarà regolata da un criterio che si basa sulla necessità di ottenere gruppi sempre più omogenei. Tutto ciò proseguirà fino a quando non sarà soddisfatta una regola di arresto legata al raggiungimento di un determinato numero di cluster.

Le tecniche di clustering più comunemente usate possono essere classificate tramite due diverse categorizzazioni. La prima categorizzazione prevede la distinzione tra:

- Clustering esclusivo: come esplicitato dal nome stesso ogni elemento potrà essere assegnato ad uno e un solo gruppo, di conseguenza i cluster non possono avere elementi in comune. Questo approccio viene definito "hard clustering";
- Clustering non-esclusivo: come esplicitato dal nome stesso, in questo caso un elemento può appartenere a più cluster con diversi gradi di appartenenza. Questo approccio viene definito come "soft clustering" o "fuzzy clustering".

La seconda categorizzazione invece prevede la distinzione tra:

- Clustering partizionale (detto anche non gerarchico o k-clustering): per andare a definire i cluster viene considerata la distanza dal centroide o dal medioide del cluster ad esempio, in seguito alla definizione del numero di cluster, fatta inizialmente. Questo rappresenta una derivazione dell'algoritmo di clustering k-means;
- Clustering gerarchico: tramite questo approccio viene costruita una gerarchia di partizioni caratterizzata da un numero crescente di gruppi se leggiamo la sua rappresentazione tramite dendrogramma dall'alto verso il basso mentre questo numero sarà decrescente se leggiamo il dendrogramma dal basso verso l'alto. All'interno del dendrogramma sono rappresentati i passi necessari all'accoppiamento/divisione dei gruppi.

Per le analisi svolte in questa tesi si è deciso di utilizzare il clustering gerarchico e, più nello specifico, quello che si basa sull'Average-link proximity.

Clustering gerarchico

Gli algoritmi di questa famiglia non servono per produrre un partizionamento ma una rappresentazione gerarchica tramite un albero. Questi algoritmi si possono suddividere in due classi:

- Agglomerativo: questa classe di algoritmi esplica bene il clustering dal basso vero l'alto;
- Divisivo: questa classe rappresenta bene il clustering dall'alto verso il basso.

In tutti e due i tipi di clustering gerarchico è necessario utilizzare delle metriche per scegliere quali cluster fondere tra loro. Nel caso del clustering agglomerativo, saranno necessarie delle misure di similarità:

- Single-link proximity che permette di ricavare la distanza tra due cluster come distanza minima tra elementi di due cluster diversi e si calcola in questo modo:

$$D(C_i, C_j) = \min\ d(x, y)\ \ \text{x} \in C_i, \text{y} \in C_j$$

- Average-link proximity che permette di ricavare la distanza tra due cluster come la media delle distanze tra i singoli elementi e si calcola in questo modo:

$$D(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{x \in C_i, y \in C_j,} d(x, y)$$

- Complete-link proximity che permette di ricavare la distanza tra due cluster come la distanza massima tra gli elementi appartenenti ai due cluster e si calcola in questo modo:

$$D(C_i, C_j) = \max d(x, y)\ \ \text{x} \in C_i, \text{y} \in C_j$$

- Distanza tra centroidi che permette di ricavare la distanza tra due cluster come quella tra i loro centroidi e si calcola in questo modo:

$$D(C_i, C_j) = d(c_i, c_j)$$

Nel clustering divisivo, invece, sarà necessario individuare il cluster da dividere in due sottogruppi e per questo si utilizzeranno delle misure che permettono di valutare la compattezza del cluster, la sua

densità o la dispersione dei punti presenti all'interno del cluster stesso. Le misure che si utilizzeranno sono:

- Average internal similarity, che serve per valutare la similarità tra gli elementi all'interno di un cluster, come suggerito dal nome: più sono dissimili tra loro, più andranno suddivisi in sottogruppi, e si calcola in questo modo:

$$D(C_i) = \frac{1}{|C_i|(|C_i| - 1)} \sum_{x,y \in C_i, x =! y} d(x, y)$$

- Maximum internal distance, che serve per valutare la distanza massima tra due elementi di un cluster, come suggerito dal nome. Questo valore va anche ad indicare il diametro del cluster e più è basso più il cluster è compatto e si calcola in questo modo:

$$D(C_i) = \max d(x, y)$$

La misurazione che sarà scelta verrà poi applicata alla distanza binaria di Bray-Curtis calcolata con la formula citata in precedenza per poter poi ottenere una clusterizzazione dei campioni, la quale permetterà di vedere se ci sono coppie di campioni individuate correttamente ovvero con un campione di un tool e lo stesso del WGS.

4.2 Metriche per valutare la capacità di identificare l'abbondanza dei geni

Le metriche presentate di seguito differiscono da quelle precedenti in quanto esse non valutano semplicemente la presenza/assenza dei KO ma considerano l'effettiva abbondanza di essi. Infatti, sia la Spearman's Correlation che l'indice di Bray-Curtis non servono per fare considerazioni su valori del tipo 0 (assenza) e 1 (presenza), ma andranno a considerare i veri e propri valori di abbondanza dei geni predetti tramite i software confrontandoli con i corrispondenti ottenuti tramite il WGS. Prima di applicare la funzione che serve per calcolare la distanza di Bray-Curtis, rispetto ai KO tra i vari campioni sarà necessario effettuare una normalizzazione dei dati secondo la proporzione in modo da ottenere un nuovo dataframe in cui la somma di ogni colonna, ovvero dei KO di un singolo campione, sia pari a uno.

4.2.1 Correlazione di Spearman

La Spearman's Correlation serve per misurare il grado di relazione tra due variabili continue e ordinabili. Questo coefficiente viene utilizzato per confrontare l'abbondanza dei KO (geni predetti) all'interno della predizione ottenuta tramite i software con quella ottenuta tramite il WGS.

$$r_s = \frac{\sum_i (r_i - \bar{r})(s_i - \bar{s})}{\sqrt{\sum_i (r_i - \bar{r})^2}\sqrt{\sum_i (s_i - \bar{s})^2}}$$

r_i : indica i ranghi della prima variabile della i-esima osservazione
s_i : indica i ranghi della seconda variabile della i-esima osservazione
$\bar{r}$: indica la media dei ranghi della prima variabile
$\bar{s}$: indica la media dei ranghi della seconda variabile

4.2.2 Indice di Bray-Curtis

L'indice di Bray-Curtis è un indice di dissimilarità, come detto in precedenza e è anche caratterizzato da una formulazione non binaria.
Per calcolarlo è necessario prendere in considerazione solo campioni con la stessa dimensione poiché l'indice di Bray-Curtis si basa sulle abbondanze grezze e non relative delle specie presenti nei campioni considerati. La formula per ricavare questa grandezza è:

$$BC_{jk} = \frac{\sum_{i=1}^{p}(|x_{ij} - x_{ik}|)}{\sum_{i=1}^{p}(x_{ij} + x_{ik})} \; i \in \text{KO (colonne)}, \text{j}, \text{k} \in \text{campioni (righe)}$$

La grandezza x rappresenta l'abbondanza di una specie i nel campione j, per quel che riguarda il primo membro della parentesi al numeratore, mentre la x successiva rappresenta l'abbondanza della specie i nel campione k, per quel che riguarda il secondo membro della parentesi al numeratore, mentre al denominatore viene calcolata la somma delle due grandezze nominate in precedenza. I fattori di sommatoria presenti sia al numeratore che al denominatore stanno a indicare che questo calcolo viene effettuato su tutte le specie presenti all'interno del campione j e del campione k.

4.2.2.1 Principal Coordinate Analysis

La PCA viene utilizzata dopo il calcolo delle metriche non binarie per andare a rappresentare le componenti principali ricavate tramite l'utilizzo delle distanze di Bray-Curtis calcolate con la formula non binaria.

4.2.2.2 Clustering

Il clustering, oltre ad essere usato con le metriche binarie, può anche essere utilizzato con le metriche di abbondanza e anche in questo caso si utilizza il clustering gerarchico con il metodo di average link-proximity applicato alla matrice di dissimilarità ricavata mettendo insieme tutte le distanze di Bray-Curtis calcolate tra i vari campioni rispettivamente ai KO (geni).

Capitolo 5 - Analisi dei risultati ottenuti nella situazione binaria e in quella non binaria

All'interno di questo capitolo saranno presentati i risultati dell'analisi svolta tramite l'utilizzo delle metriche presentate nel capitolo 4. Questi saranno supportati da grafici come i PCoA plot, i dendrogrammi, utilizzati per rappresentare i clustering, degli istogrammi e infine dei boxplot. Saranno commentati prima i risultati ottenuti con le metriche non binarie e, in seguito, quelli ottenuti con le metriche binarie, dopo di che i risultati saranno suddivisi per singoli dataset. Nella prima parte si analizzeranno i risultati ottenuti tramite PCoA plot, basati sull'indice di Bray-Curtis calcolato in precedenza, e a supporto di questi si utilizzerà ciò che è stato ottenuto tramite correlazione di Spearman e i rispettivi boxplot che la rappresentano, il tutto applicato alle abbondanze dei geni, mentre nella seconda parte si analizzeranno i plot della PCoA ottenuti, questa volta, con l'indice di Bray-curtis binario e gli istogrammi ricavati attraverso Balanced Accuracy, Precision, Recall e F_1_score.

5.1 Analisi dei risultati ottenuti tramite le metriche che valutano l'abbondanza dei geni nei dataframe per singolo dataset

Blueberry

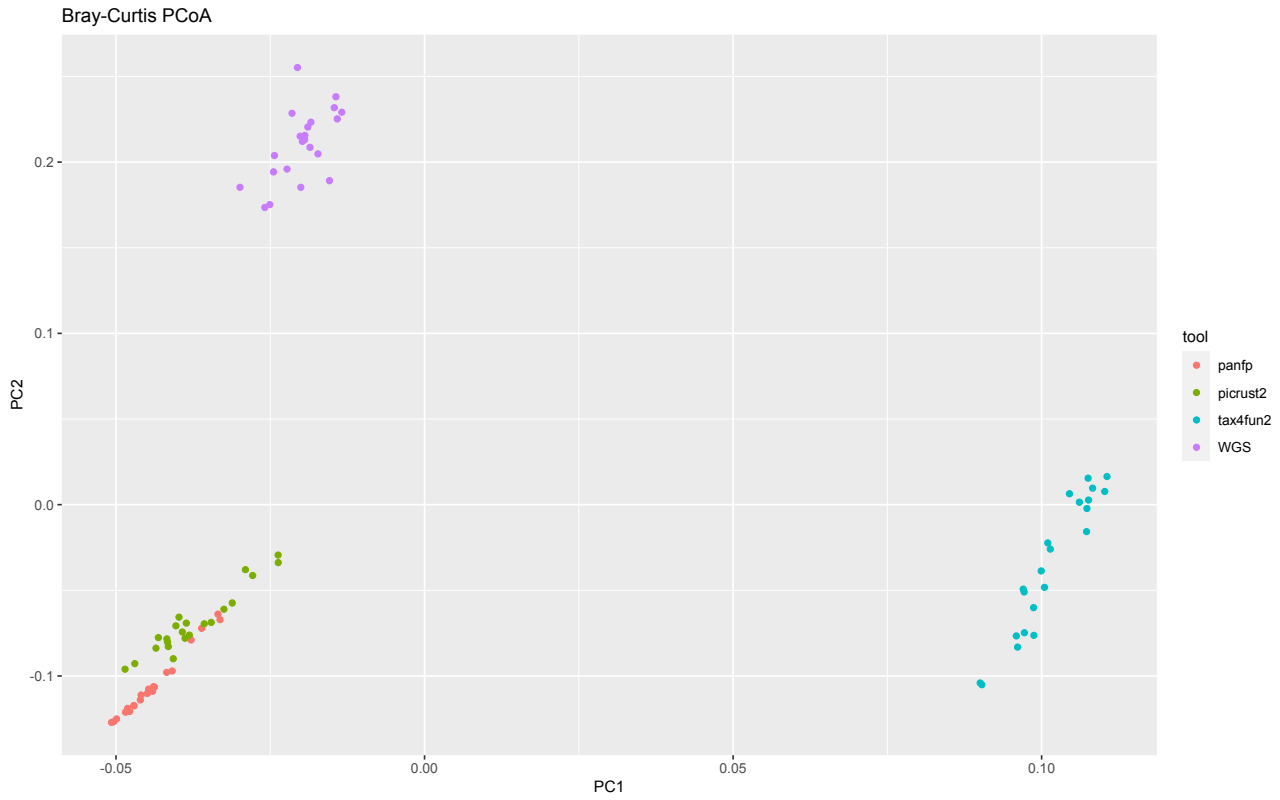


Figura 7: PCoA plot per rappresentare le distanze tra i vari campioni nei diversi tool e nel WGS per il dataset Blueberry, nell'asse x abbiamo la PC1 mentre nell'asse y abbiamo la PC2.

Il primo dataset a essere preso in considerazione è quello del Blueberry e tramite la PCoA, ovvero una riduzione della dimensionalità del dataframe che ci permette di considerare solamente PC1 e PC2, il cui calcolo è basato sulla matrice di dissimilarità di Bray-Curtis, è possibile vedere come per questo dataset i tool migliori siano PanFP e PICRUSt2. Dato che le predizioni dei campioni ricavate tramite questi si avvicinano di più rispetto a quelle del Tax4Fun2 alle stesse ottenute con il WGS. Questo risultato è anche supportato dal boxplot che si ottiene tramite la correlazione di Spearman, tramite il quale si ha conferma che i tool migliori sono gli stessi.

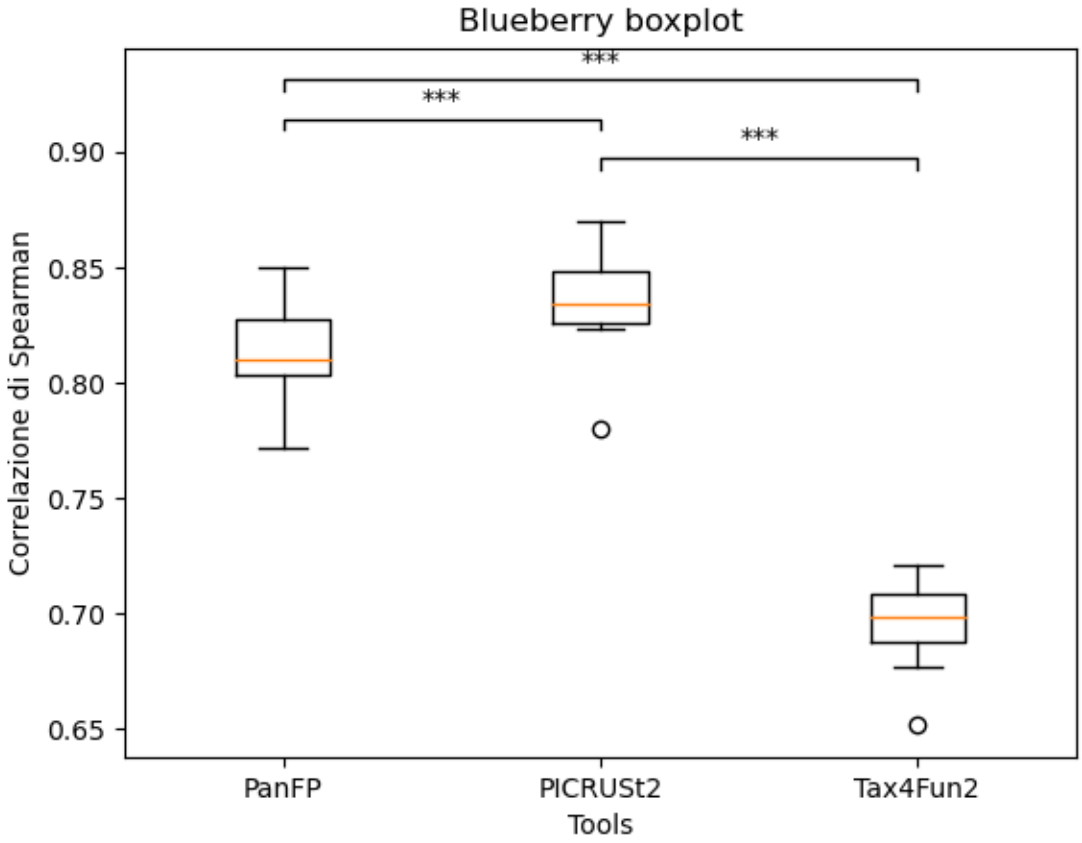


Figura 8: boxplot per il dataset Blueberry, nell'asse x abbiamo la distinzione tra i boxplot dei vari tool mentre nell'asse y abbiamo i livelli di correlazione.

Il grafico di cui sopra, infatti, evidenzia come ci sia poca differenza tra la mediana del boxplot di PanFP e quella di PICRUSt2 e questo rispecchia la vicinanza tra i campioni presenti all'interno del plot della PCoA. Unendo questi due risultati si può concludere che il tool migliore per ottenere delle buone predizioni del profilo funzionale, partendo da un dataset di tipo Blueberry, è il PICRUSt2; infatti guardando i boxplot si vede come i valori del PICRUSt2 siano più alti dei valori degli altri boxplot.

Cameroon

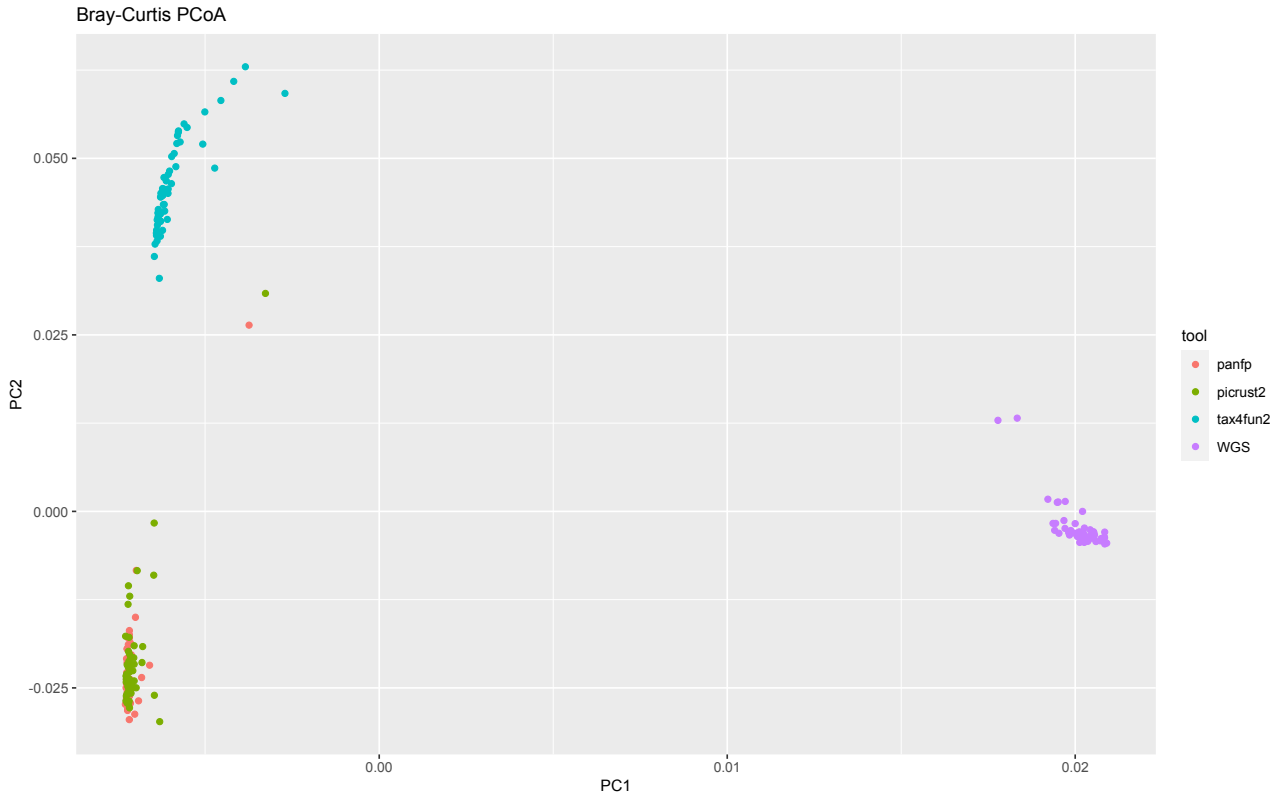


Figura 9: PCoA plot per rappresentare le distanze tra i vari campioni nei diversi tool e nel WGS per il dataset Cameroon, nell'asse x abbiamo la PC1 mentre nell'asse y abbiamo la PC2.

Osservando il plot ottenuto tramite la PCoA per quel che riguarda il dataset Cameroon si può vedere come anche per questo dataset ci sia una sovrapposizione tra le predizioni dei campioni ottenute con PanFP e PICRUSt2, che risultano essere più vicine a quelle ottenute con il WGS; infatti si evince dalla figura come le distanze delle predizioni di PanFP e PICRUSt2 siano minori rispetto alle distanze delle predizioni del Tax4Fun2. Questo, come per il Blueberry, risulterà essere supportato dai boxplot ricavati con la correlazione di Spearman.

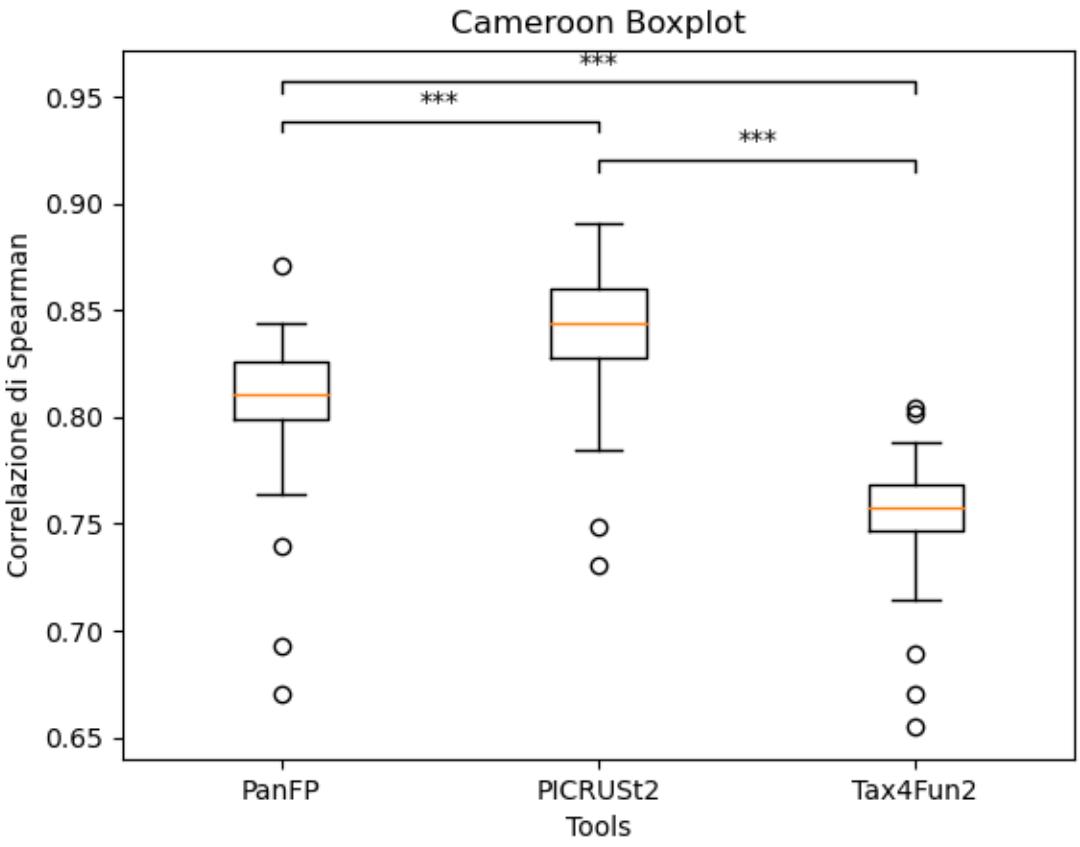


Figura 10: Boxplot per il dataset Cameroon, nell'asse x abbiamo la distinzione tra i boxplot dei vari tool mentre nell'asse y abbiamo i livelli di correlazione.

Anche qui si potrà vedere come i boxplot migliori siano quelli di PICRUSt2 e PanFP e tramite la mediana dei boxplot si potrà concludere che il tool migliore per ottenere predizioni dei profili funzionali per campioni provenienti dal dataset Cameroon sia il PICRUSt2; infatti esso avrà valori dei boxplot migliori rispetto ai corrispondenti degli altri tool e questo dimostra che la correlazione per il PICRUSt2 con il WGS sarà maggiore.

HMP

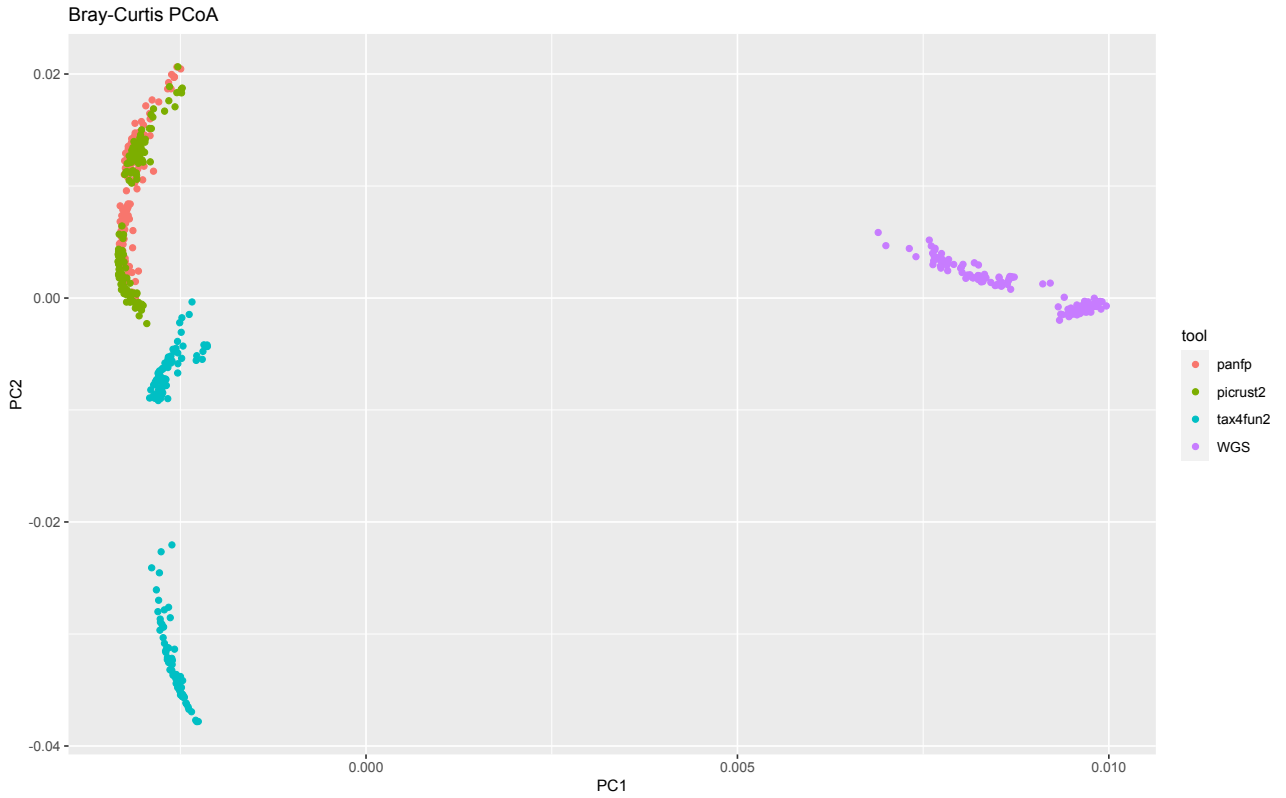


Figura 11: PCoA plot per rappresentare le distanze tra i vari campioni nei diversi tool e nel WGS per il dataset HMP, nell'asse x abbiamo la PC1 mentre nell'asse y abbiamo la PC2.

Anche per quel che riguarda il dataset HMP tramite il PCoA plot si può vedere come i tool le cui predizioni si avvicinino di più ai risultati ottenuti tramite il WGS siano ancora PanFP e PICRUSt2 e come anche in questo caso ci sia una sovrapposizione molto marcata tra le predizioni di uno e quelle dell'altro. Considerando le distanze si nota come le predizioni dei due tool siano più vicine ai risultati del WGS. Tutto ciò viene supportato dai risultati ottenuti tramite la correlazione di Spearman, visibili con i boxplot.

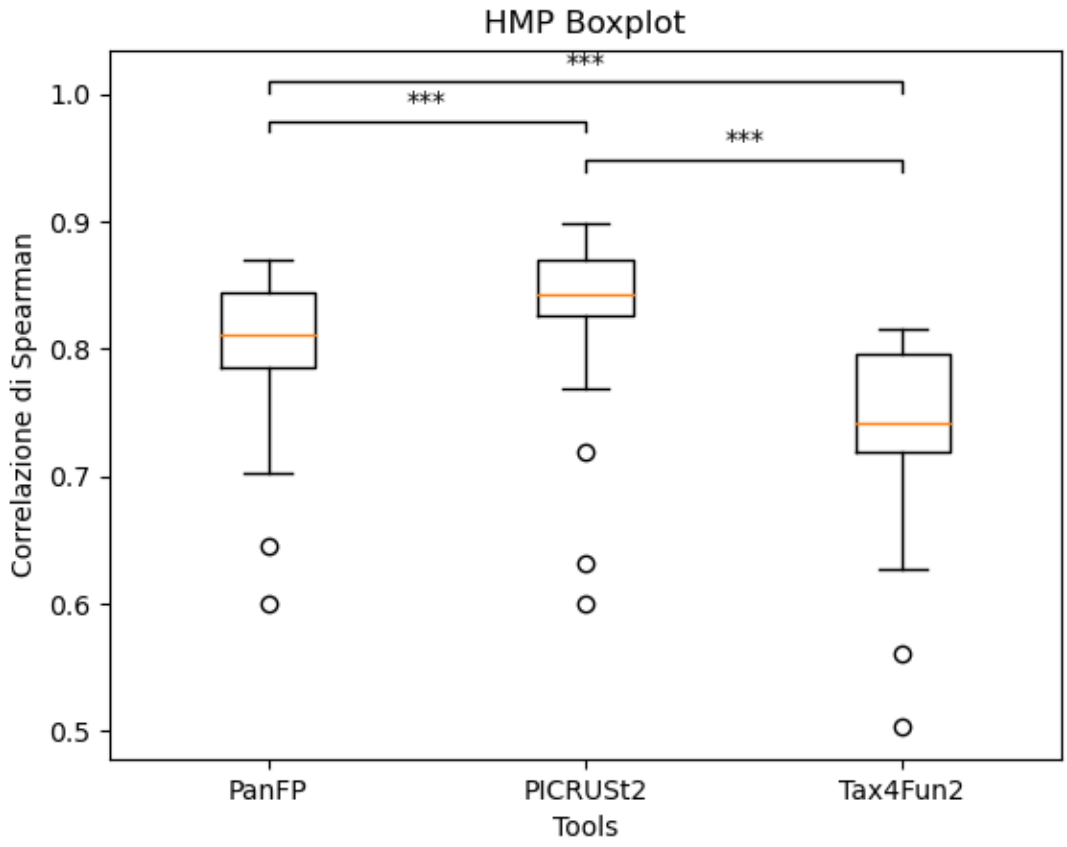


Figura 12: boxplot per il dataset HMP, nell'asse x abbiamo la distinzione tra i boxplot dei vari tool mentre nell'asse y abbiamo i livelli di correlazione.

Dalla figura emerge come i boxplot, ottenuti tramite la correlazione di Spearman per quel che riguarda i tool (PanFP e PICRUSt2) siano molto simili; anche per questo dataset, il tool che risulta essere il migliore, andando a considerare la mediana dei boxplot stessi, è il PICRUSt2; infatti esso avrà valori dei boxplot migliori rispetto ai corrispondenti degli altri tool e questo dimostra che la correlazione per il PICRUSt2 con il WGS sarà maggiore.

Indian

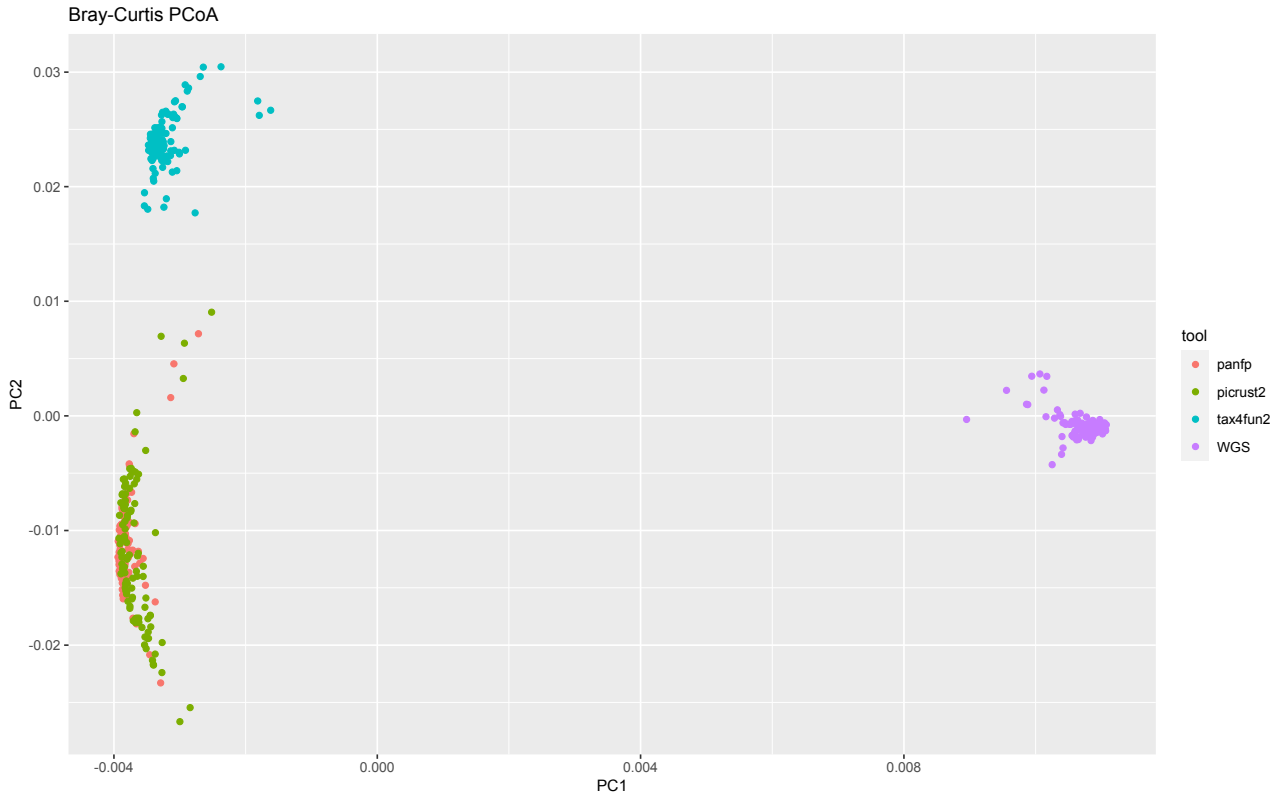


Figura 13: PCoA plot per rappresentare le distanze tra i vari campioni nei diversi tool e nel WGS per il dataset Indian, nell'asse x abbiamo la PC1 mentre nell'asse y abbiamo la PC2.

L'analisi dei risultati ottenuti tramite il PCoA plot per il dataset Indian evidenzia un andamento molto simile agli altri plot della PCoA ovvero le predizioni per i campioni ottenute tramite PanFP e PICRUSt2 sono molto vicine tra loro e ci sono varie sovrapposizioni; esse risultano essere quelle che si avvicinano di più alle predizioni ottenute tramite il WGS, infatti valutando le distanze si vede come per la PC1 e la PC2 esse risultino avere distanza minore dal WGS rispetto alle predizioni di Tax4Fun2. Anche quanto sopra avrà come supporto i risultati ottenuti tramite i boxplot ricavati in seguito al calcolo della correlazione di Spearman.

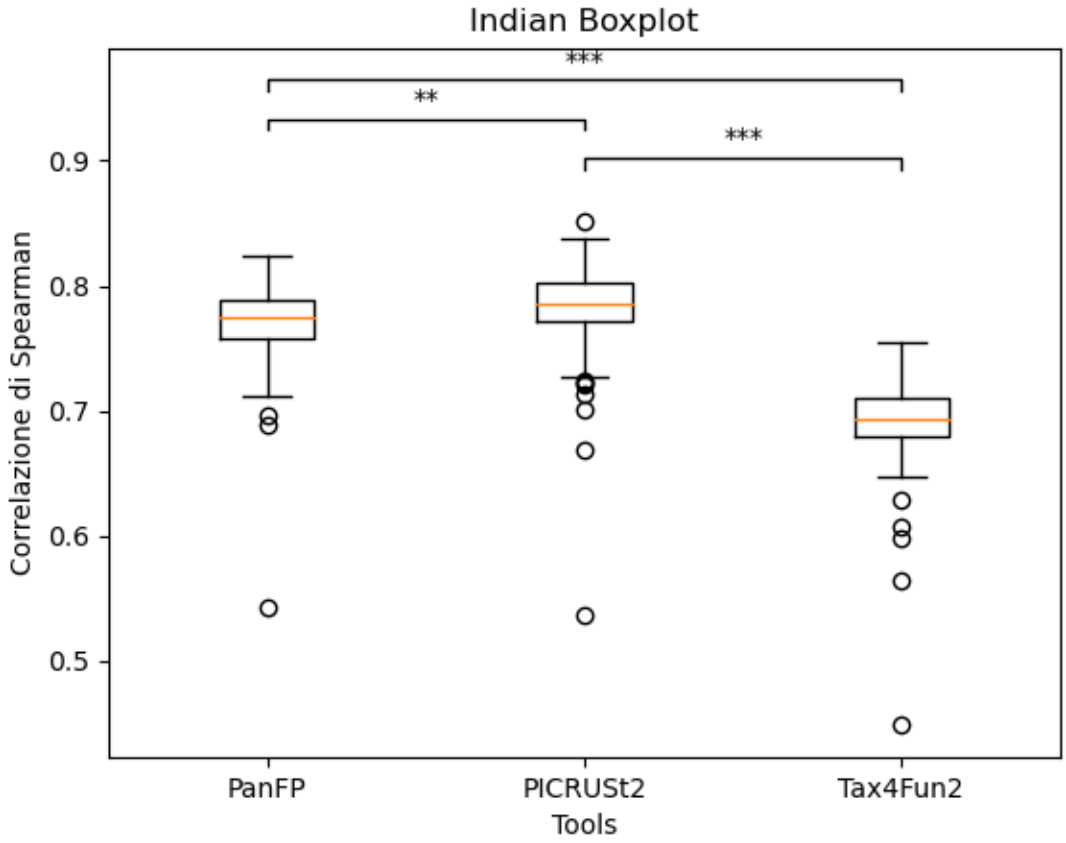


Figura 14: boxplot per il dataset Indian, nell'asse x abbiamo la distinzione tra i boxplot dei vari tool mentre nell'asse y abbiamo i livelli di correlazione.

Anche da questa figura si vede come PanFP e PICRUStt2 permettano di ottenere risultati con una correlazione molto più alta con quelli ottenuti tramite il WGS, se, confrontati con i risultati ricavati utilizzando il Tax4Fun2. Considerando invece la mediana, anche qui, il PICRUSt2 è lo strumento migliore dei tre; infatti esso avrà valori dei boxplot migliori rispetto ai corrispondenti degli altri tool.

Mammal

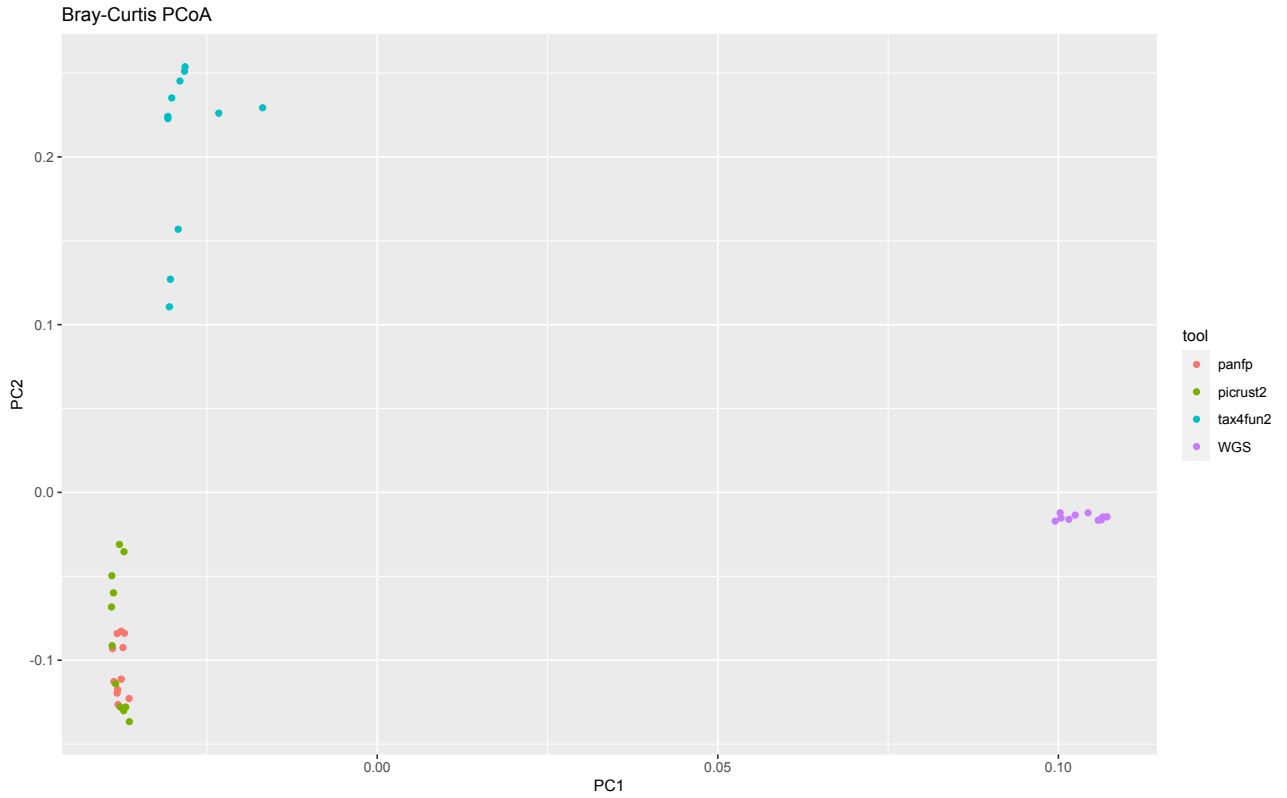


Figura 15: PCoA plot per rappresentare le distanze tra i vari campioni nei diversi tool e nel WGS per il dataset Mammal, nell'asse x abbiamo la PC1 mentre nell'asse y abbiamo la PC2.

Considerando il dataset Mammal, i risultati ottenuti sono simili a quelli ottenuti con la PCoA per gli altri dataset ovvero sono caratterizzati sempre da sovrapposizione e vicinanza per quel che riguarda le predizioni dei campioni ottenute tramite PanFP e PICRUSt2 mentre Tax4Fun2 è posizionato lontano sia da questi ultimi sia dai risultati ottenuti con il WGS. Infatti valutando le distanze si vede come per la PC1 e la PC2 esse risultino avere distanza minore dal WGS rispetto alle predizioni di Tax4Fun2. A supporto di questo c'è ancora il boxplot ricavato tramite la correlazione di Spearman, che come per gli altri dataset ci permetterà di stabilire quale sia il tool migliore.

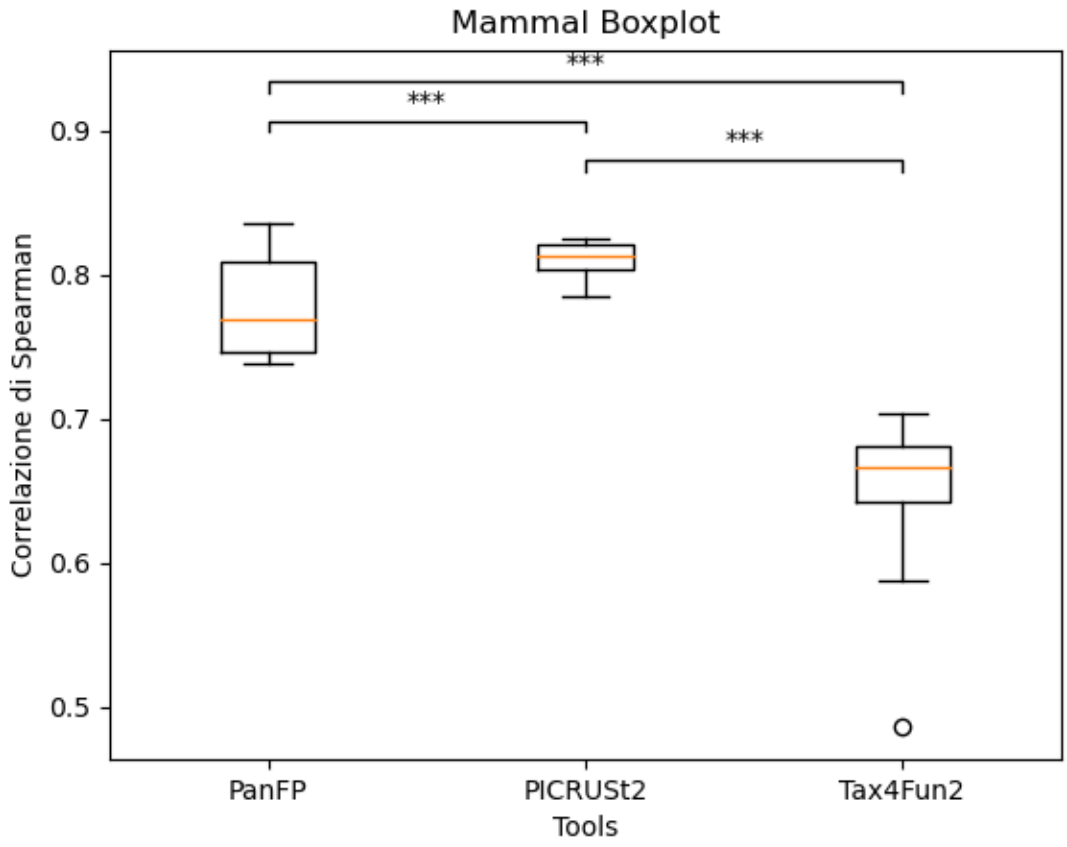


Figura 16: boxplot per il dataset Mammal, nell'asse x abbiamo la distinzione tra i boxplot dei vari tool mentre nell'asse y abbiamo i livelli di correlazione.

Anche in questo boxplot, PanFP e PICRUSt2 sono molto vicini tra loro e permettono di ottenere risultati migliori rispetto al Tax4Fun2 e anche in questa situazione il valore mediano del boxplot relativo a PICRUSt2 è migliore rispetto a quello degli altri; infatti esso avrà valori dei boxplot migliori rispetto ai corrispondenti degli altri tool e questo dimostra che la correlazione per il PICRUSt2 con il WGS sarà maggiore.

MammalianGut

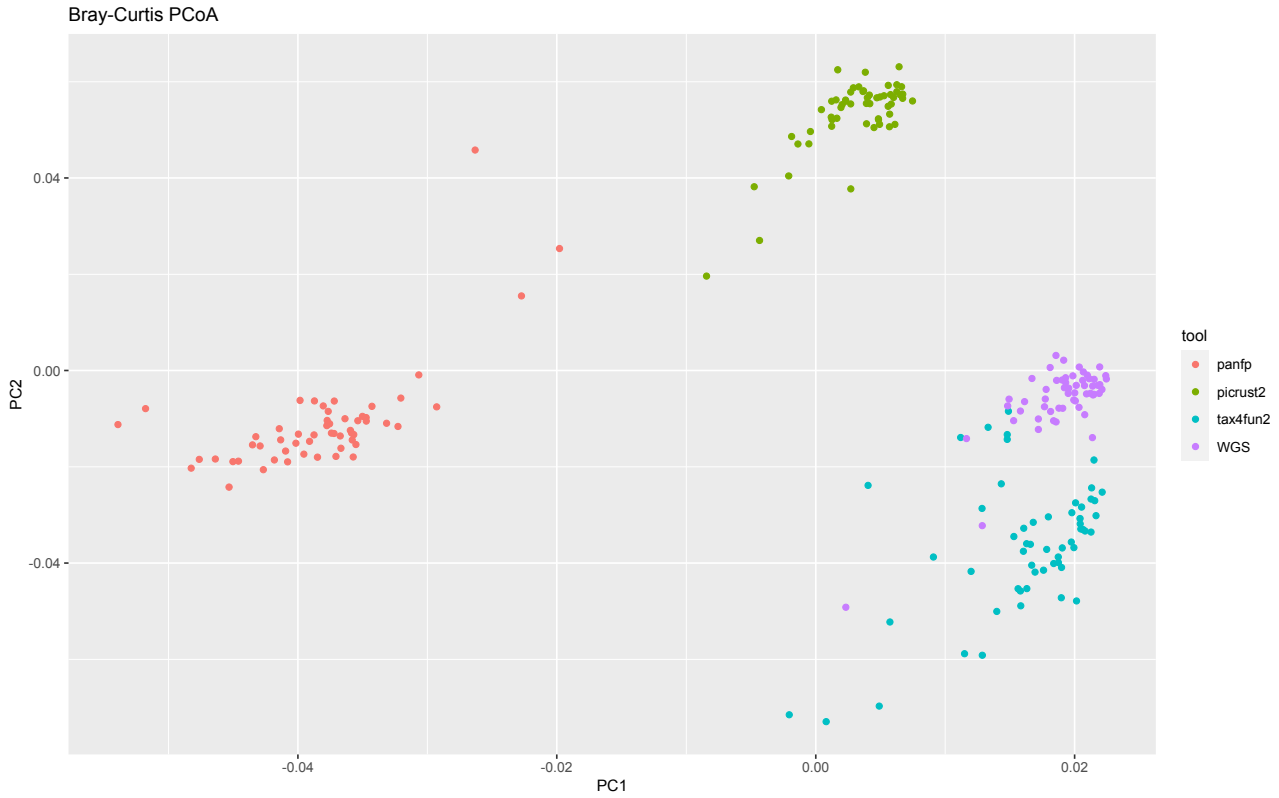


Figura 17: PCoA plot per rappresentare le distanze tra i vari campioni nei diversi tool e nel WGS per il dataset MammalianGut, nell'asse x abbiamo la PC1 mentre nell'asse y abbiamo la PC2.

Il dataset MammalianGut è l'unico che fa eccezione, in quanto, tramite il PCoA plot, si può vedere come le predizioni dei campioni ottenute tramite il Tax4Fun2 siano quelle che più si avvicinano alle rispettive ottenute tramite il WGS mentre i risultati di PanFP e PICRUSt2 in questa situazione sono più distanti oltre a non essere neanche sovrapposti tra loro. Infatti valutando le distanze si vede come per la PC1 e la PC2 esse risultino avere distanza minore delle predizioni di PanFP e di PICRUSt2. Come per ogni dataset, quanto sopra è confermato dai boxplot ottenuti tramite la correlazione di Spearman.

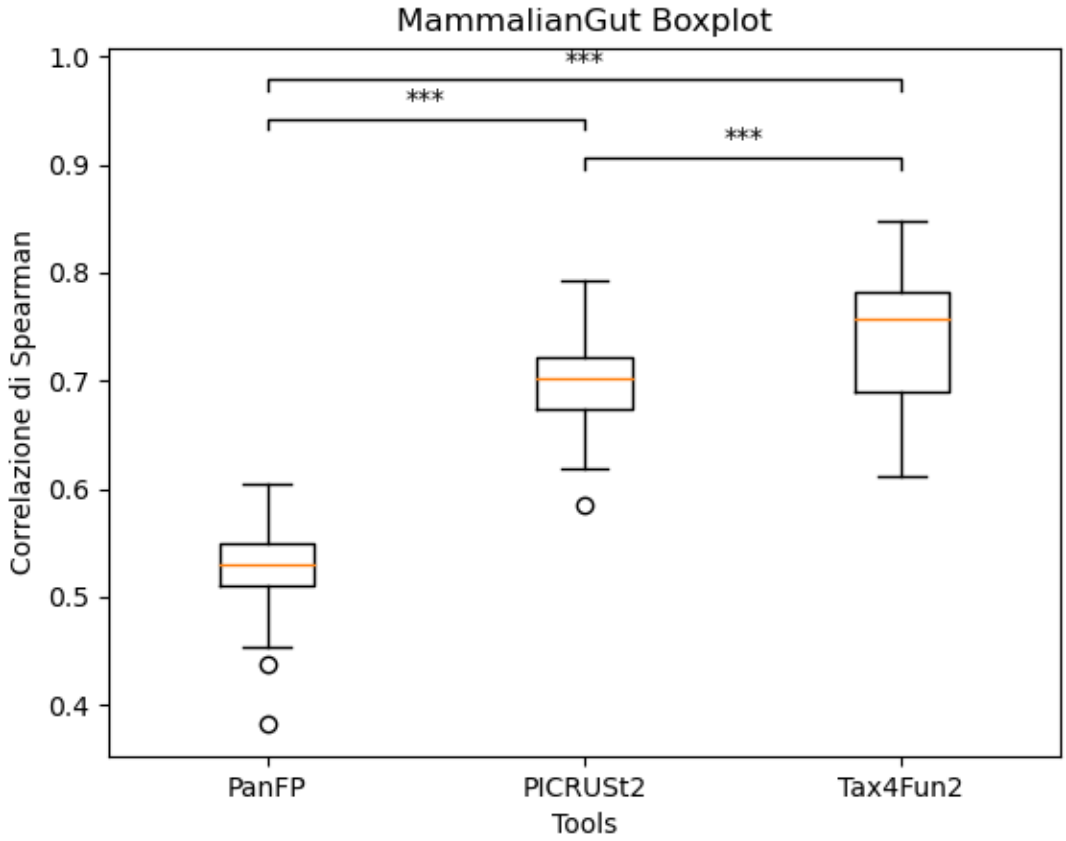


Figura 18: boxplot per il dataset MammalianGut, nell'asse x abbiamo la distinzione tra i boxplot dei vari tool mentre nell'asse y abbiamo i livelli di correlazione.

Da questa figura si può vedere come il boxplot di Tax4Fun2 abbia valori migliori rispetto a quelli degli alti due tool; considerando la sua mediana, infatti, possiamo sostenere che il Tax4Fun2 è il migliore dei tool per questo dataset e questo dimostra che la correlazione per il Tax4Fun2 con il WGS sarà maggiore.

MammalsSimulated

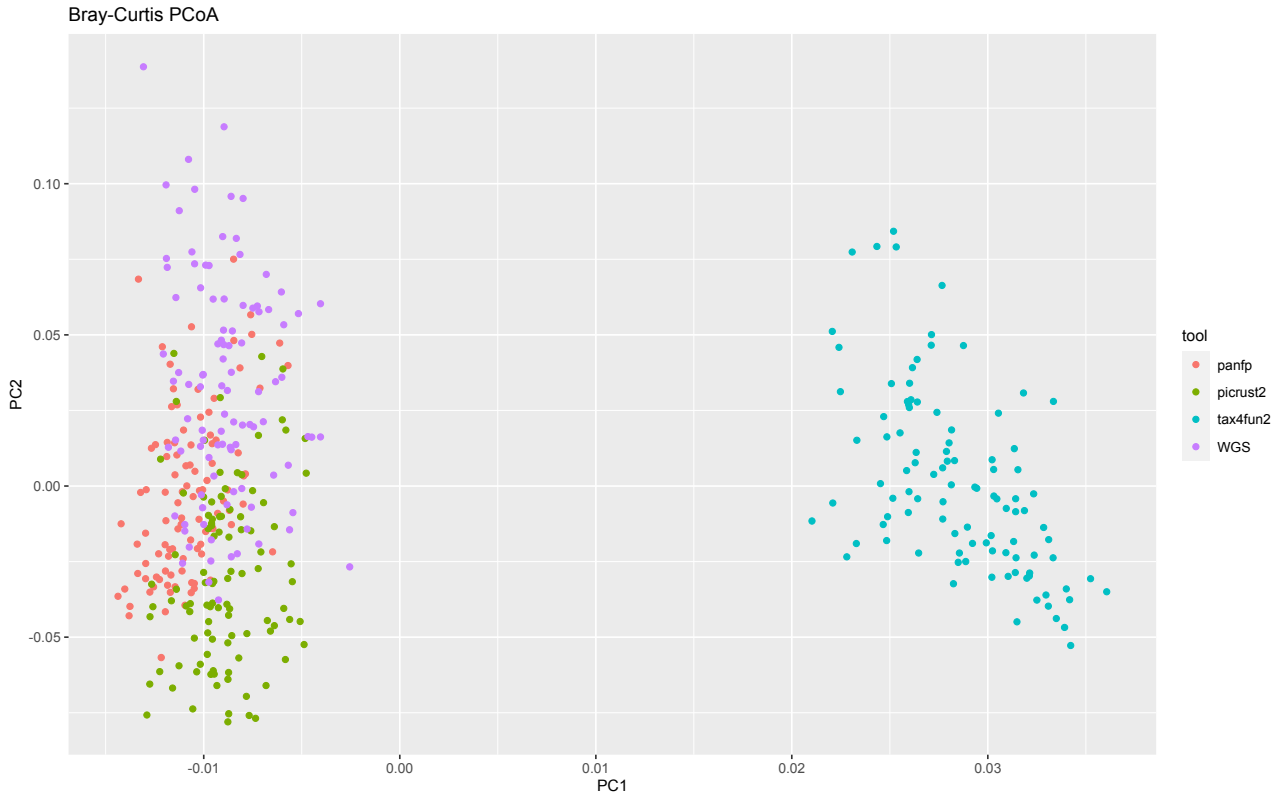


Figura 19: PCoA plot per rappresentare le distanze tra i vari campioni nei diversi tool e nel WGS per il dataset MammalsSimulated, nell'asse x abbiamo la PC1 mentre nell'asse y abbiamo la PC2.

All'interno di questo PCoA plot per il dataset MammalsSimulated viene ripreso il trend secondo il quale PanFP e PICRUSt2 risultano essere i tool che permettono di ottenere le predizioni migliori e che più si avvicinano a quelle ottenute tramite il WGS. Infatti valutando le distanze si vede come per la PC1 e la PC2 esse risultino essere minori rispetto a quelle delle predizioni di Tax4Fun2. Addirittura, qui si può vedere come ci sia sovrapposizione e molta vicinanza oltre che tra le predizioni di PanFP e PICRUSt2, come negli altri plot, anche con quelle del WGS, infatti i tre costituiranno la nuvola di punti presente nella parte sinistra del plot. Tutto questo sarà poi confermato dai boxplot ricavati tramite la correlazione di Spearman.

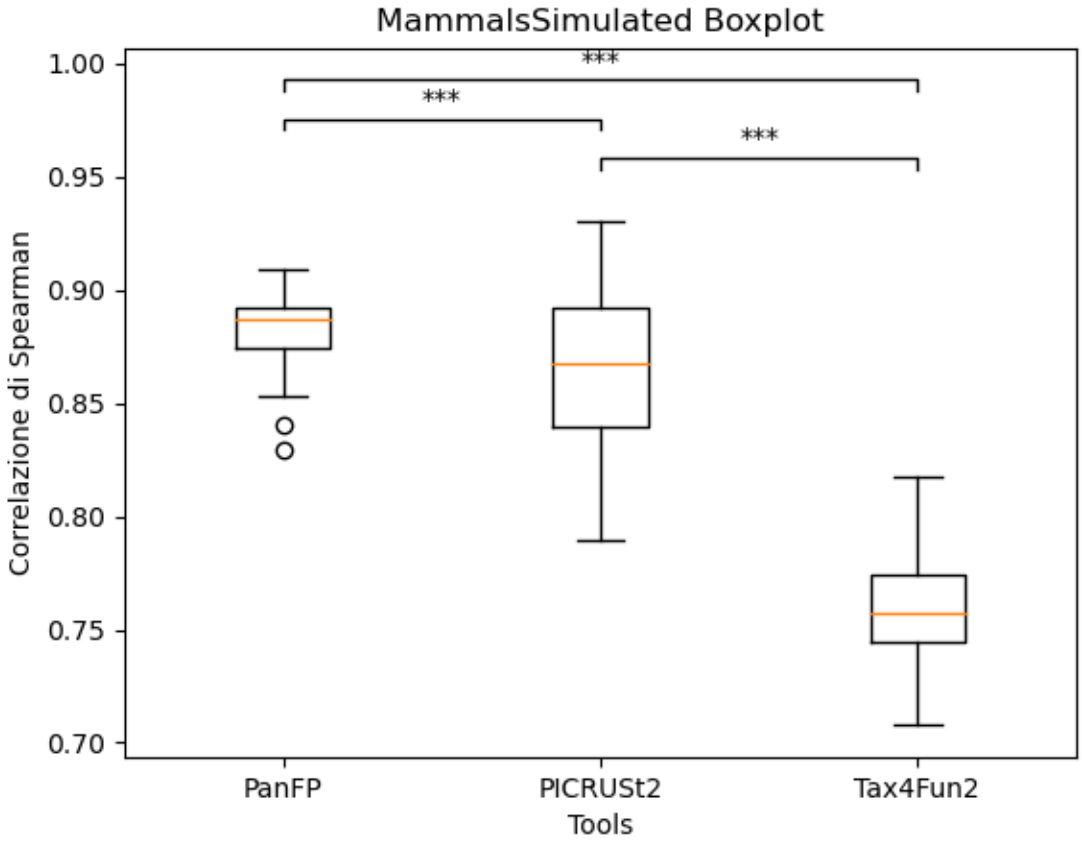


Figura 20: boxplot per il dataset MammalsSimulated, nell'asse x abbiamo la distinzione tra i boxplot dei vari tool mentre nell'asse y abbiamo i livelli di correlazione.

Questo boxplot evidenzia come PanFP e PICRUSt2 abbiano ancora una volta valori di correlazione migliori rispetto a Tax4Fun2, anche se in questo caso il tool con la mediana migliore rispetto agli altri è il PanFP e non il PICRUSt2; infatti esso avrà valori dei boxplot migliori rispetto ai corrispondenti degli altri tool e questo dimostra che la correlazione per il PanFP con il WGS sarà maggiore.

Ocean

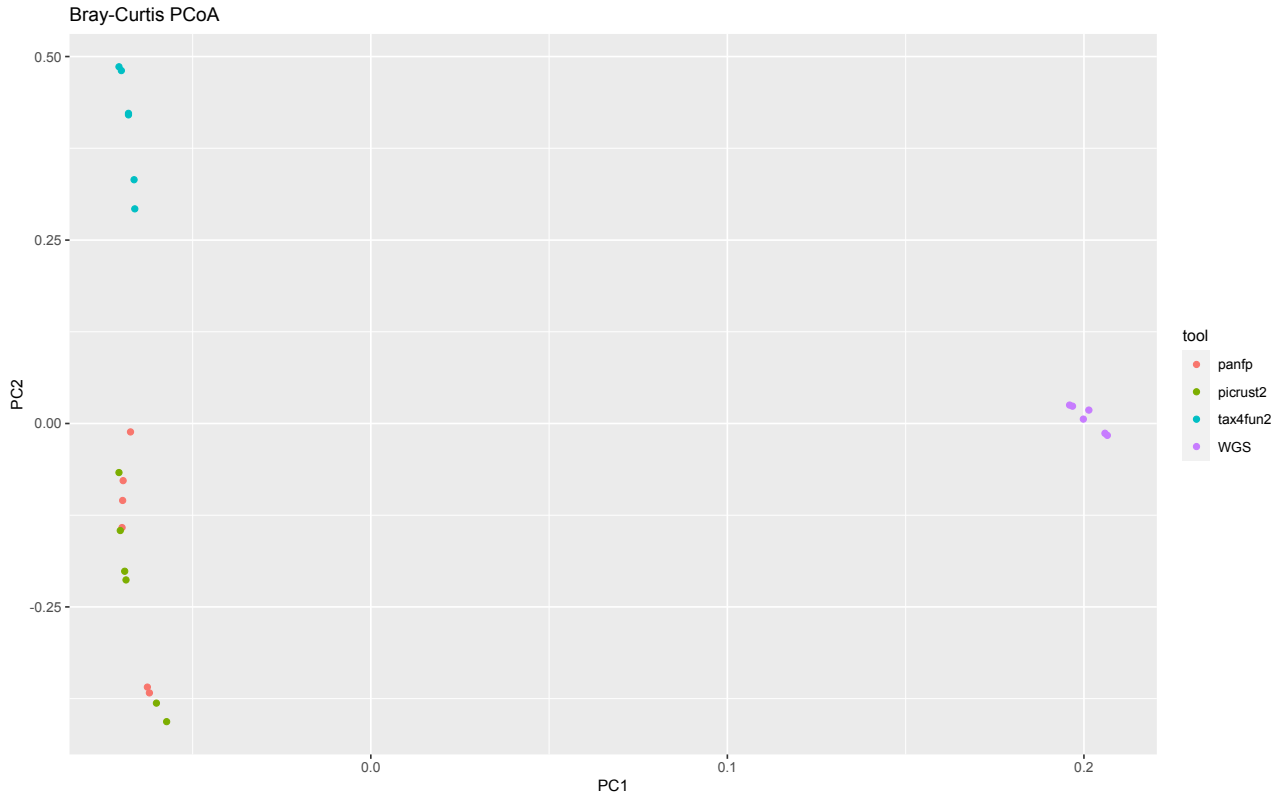


Figura 21: PCoA plot per rappresentare le distanze tra i vari campioni nei diversi tool e nel WGS per il dataset Oceean, nell'asse x abbiamo la PC1 mentre nell'asse y abbiamo la PC2.

Come nei PCoA plot precedenti anche per il dataset Ocean abbiamo sovrapposizione e vicinanza tra le predizioni dei campioni ricavate con PanFP e PICRUSt2 e ancora una volta, come nelle situazioni precedenti, risultano essere i più vicini ai risultati ottenuti tramite il WGS. Infatti valutando le distanze si vede come per la PC1 e la PC2 esse abbiano distanza minore dal WGS rispetto alle predizioni di Tax4Fun2. Anche questi risultati sono supportati dal boxplot.

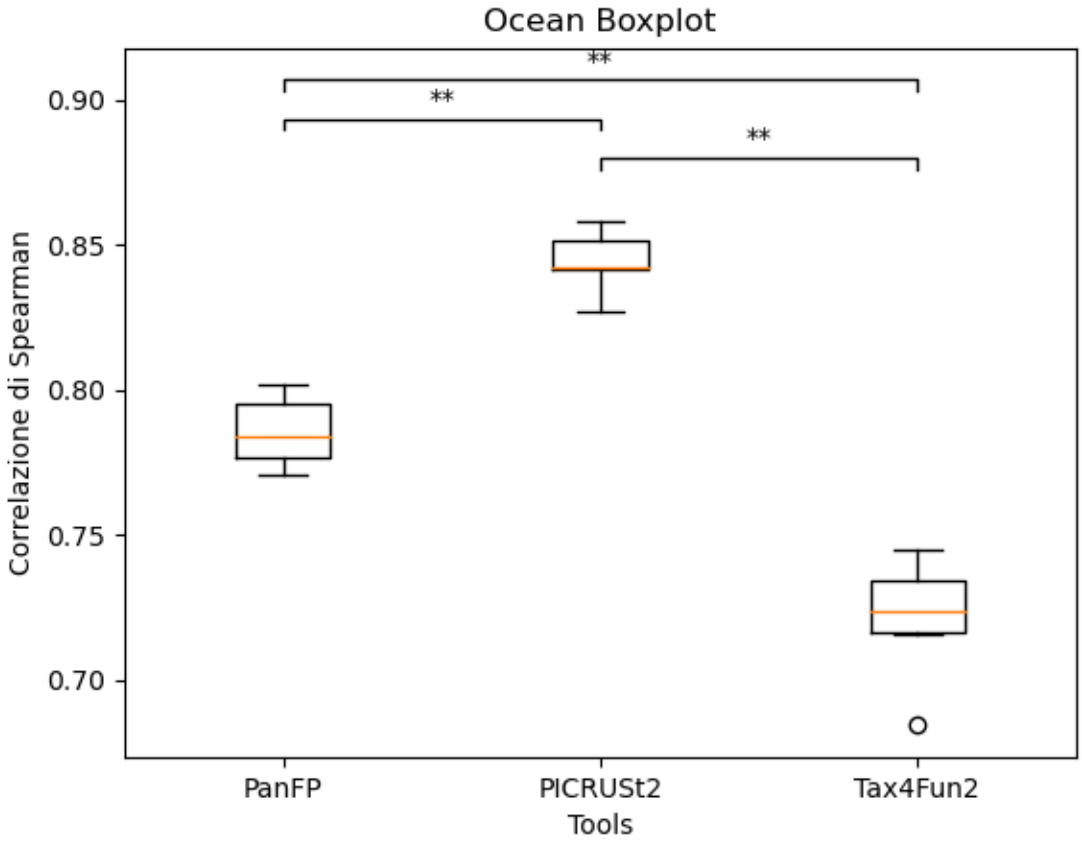


Figura 22: boxplot per il dataset Ocean, nell'asse x abbiamo la distinzione tra i boxplot dei vari tool mentre nell'asse y abbiamo i livelli di correlazione.

Da questo boxplot si vede che in questa situazione c'è una maggiore differenza tra i valori di PanFP e quelli del PICRUSt2 e anche per il dataset Ocean il tool migliore è il PICRUSt2; infatti esso avrà valori dei boxplot migliori rispetto ai corrispondenti degli altri tool e poi avrà anche il valore della mediana più elevato rispetto a PanFP e Tax4Fun2; questo dimostra che la correlazione per il PICRUSt2 con il WGS sarà maggiore.

Primate

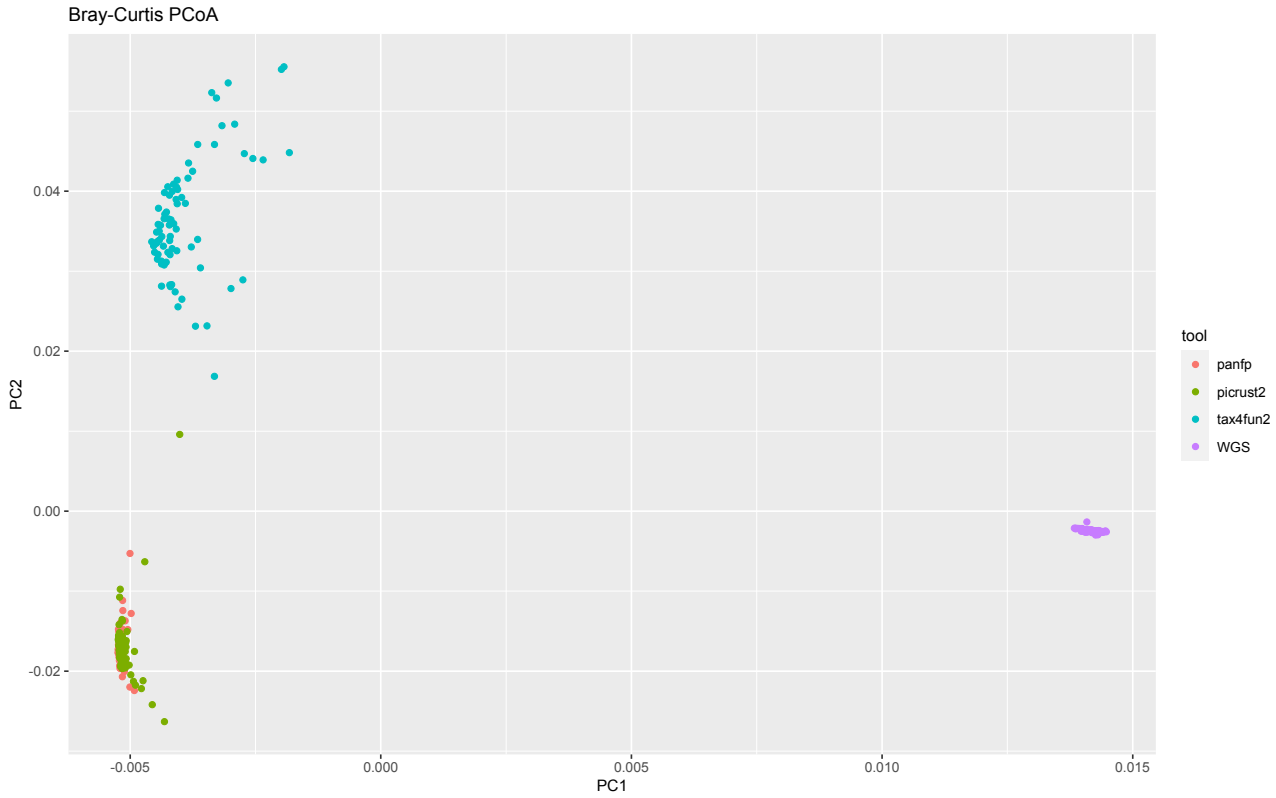


Figura 23: PCoA plot per rappresentare le distanze tra i vari campioni nei diversi tool e nel WGS per il dataset Primate, nell'asse x abbiamo la PC1 mentre nell'asse y abbiamo la PC2.

Anche in questo PCoA plot si vede come le predizioni di PICRUSt2 e PanFP siano molto vicine e addirittura si sovrappongano e risultino essere, anche per il dataset Primate, le più vicine a quelle ottenute tramite il WGS. Infatti valutando le distanze si vede come per la PC1 e la PC2 esse abbiano distanza minore dal WGS rispetto alle predizioni di Tax4Fun2. Come per gli altri dataset, il boxplot ottenuto con la correlazione di Spearman supporta questi risultati.

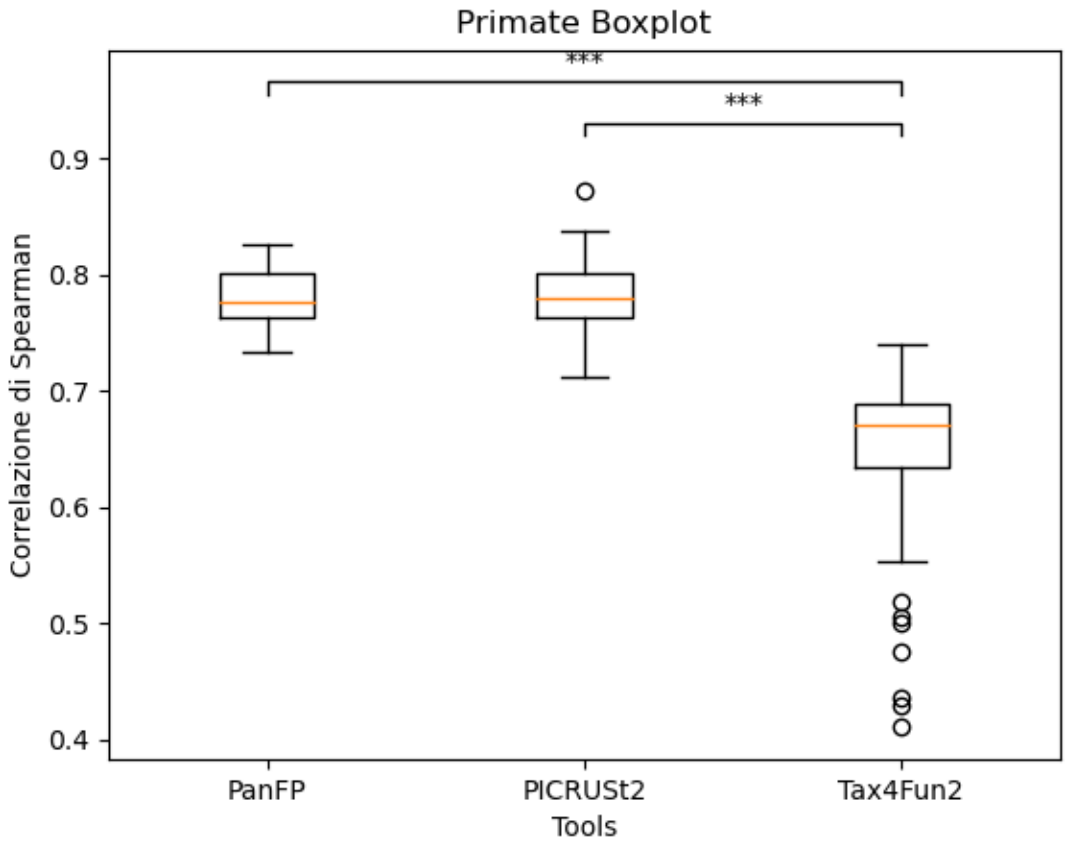


Figura 24: boxplot per il dataset Primate, nell'asse x abbiamo la distinzione tra i boxplot dei vari tool mentre nell'asse y abbiamo i livelli di correlazione.

Questo boxplot permette di vedere, anche per questo dataset, la vicinanza tra i valori di correlazione ottenuti confrontando le predizioni di PanFP e PICRUSt2 con quelle del WGS. Anche in questo caso, il tool migliore è il PICRUSt2 poiché la sua mediana ha un valore più alto rispetto a quella del PanFP; infatti esso avrà valori dei boxplot migliori rispetto ai corrispondenti degli altri tool e questo conferma che la correlazione per il PICRUSt2 con il WGS sarà maggiore.

I risultati ottenuti tramite i plot della PCoA e i boxplot sono, poi, rispecchiati dai dendrogrammi ricavati tramite clustering, sempre considerando la distanza di Bray-curtis come punto di partenza e il metodo Average. Questi dendrogrammi permettono di vedere come ci sia effettivamente distanza tra le predizioni dei campioni ottenuti tramite un tool e i rispettivi risultati ottenuti tramite il WGS, infatti il clustering, a differenza della PCoA, è stato realizzato accoppiando ogni singolo tool con il WGS. Di seguito verranno mostrati i dendrogrammi relativi ai dataset Blueberry, Ocean e Mammal, caratterizzati dal numero di campioni minore, e nel materiale supplementare saranno inseriti i dendrogrammi ottenuti per gli altri dataset.

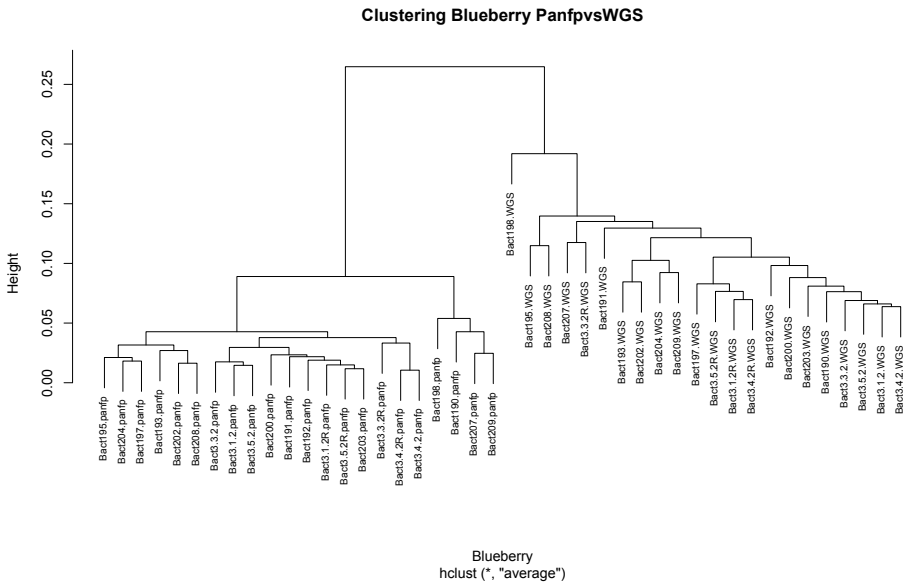


Figura 25: dendrogramma per il clustering gerarchico di PanFPvsWGS per il dataset Blueberry.

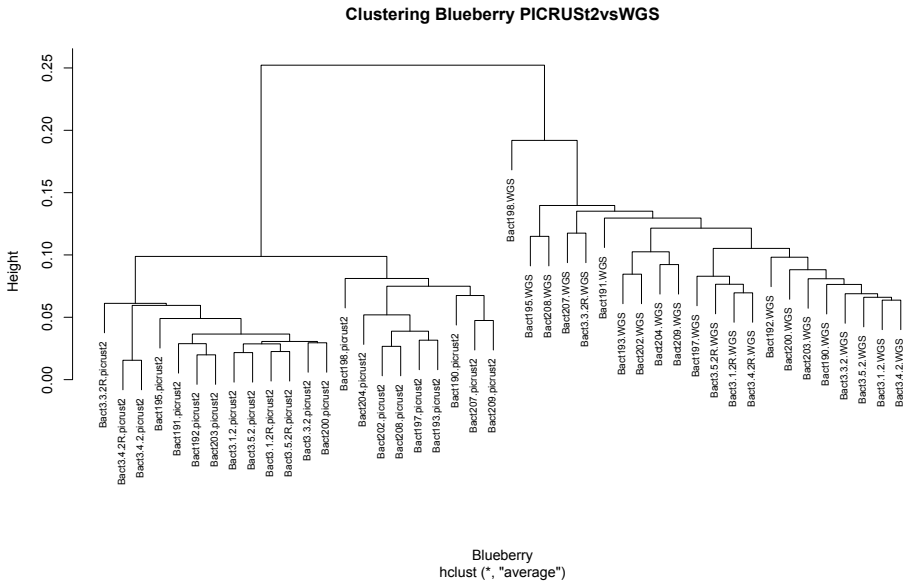


Figura 26: dendrogramma per il clustering gerarchico di PICRUST2vsWGS per il dataset Blueberry.

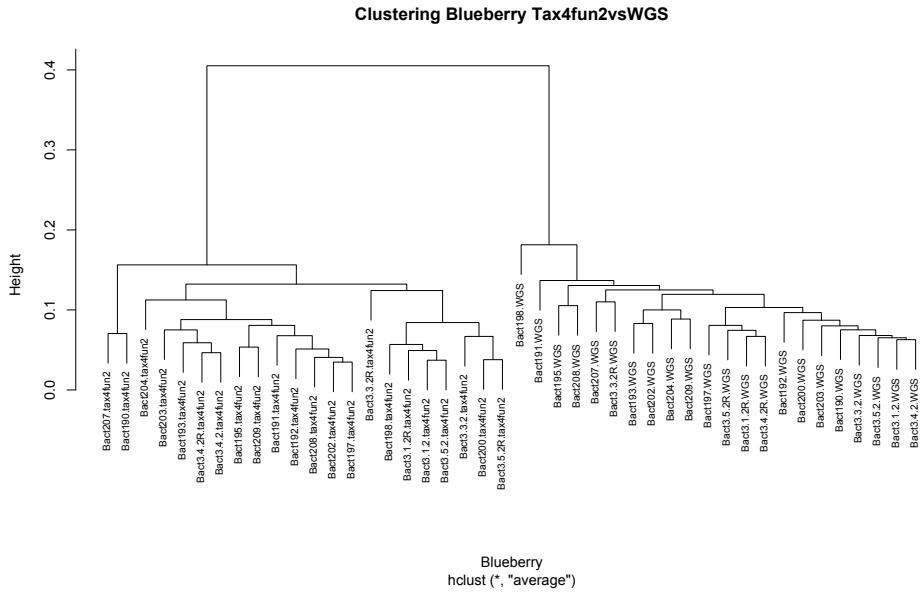


Figura 27: dendrogramma per il clustering gerarchico di Tax4Fun2vsWGS per il dataset Blueberry.

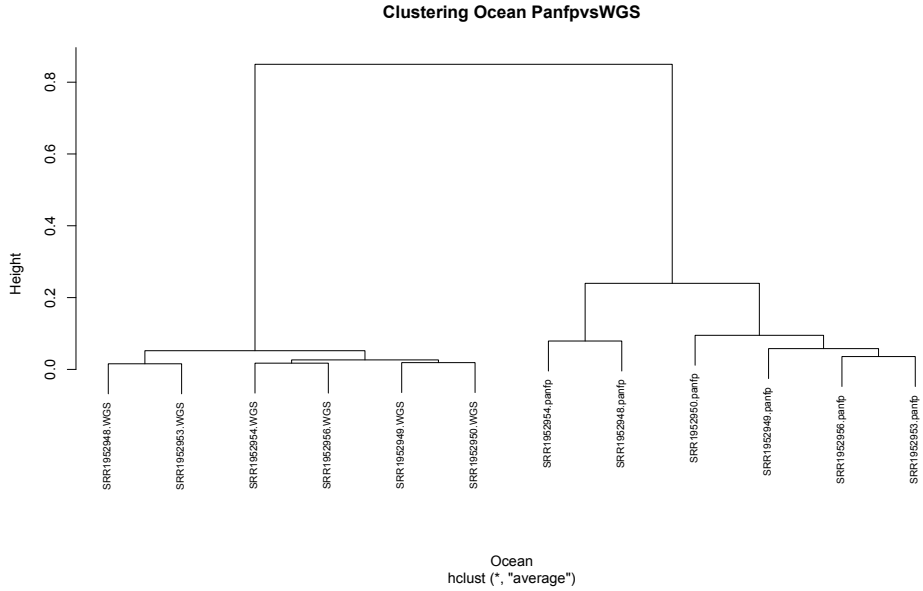


Figura 28: dendrogramma per il clustering gerarchico di PanFPvsWGS per il dataset Ocean.

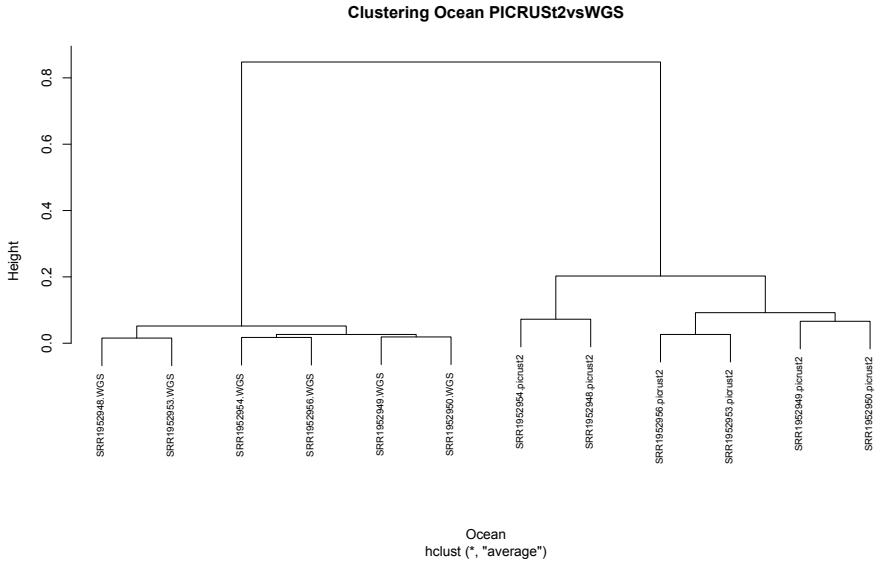


Figura 29: dendrogramma per il clustering gerarchico di PICRUSt2vsWGS per il dataset Ocean.

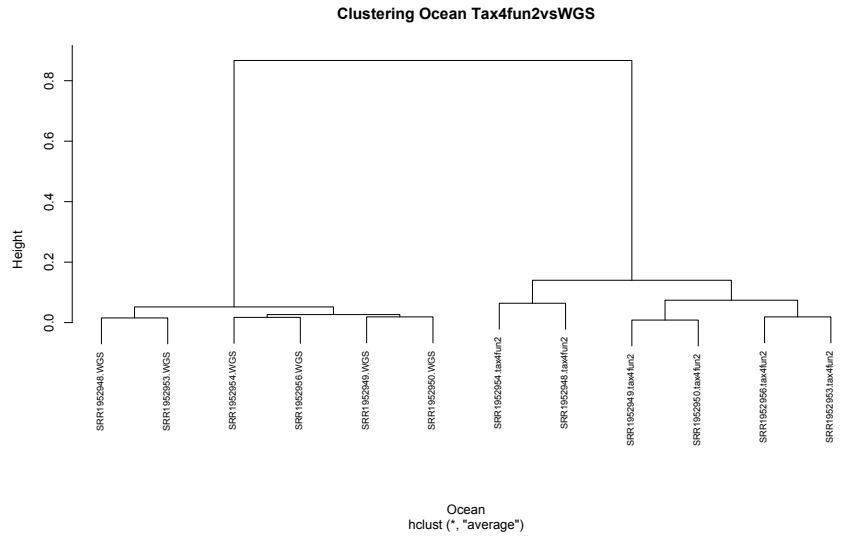


Figura 30: dendrogramma per il clustering gerarchico di Tax4Fun2vsWGS per il dataset Ocean.

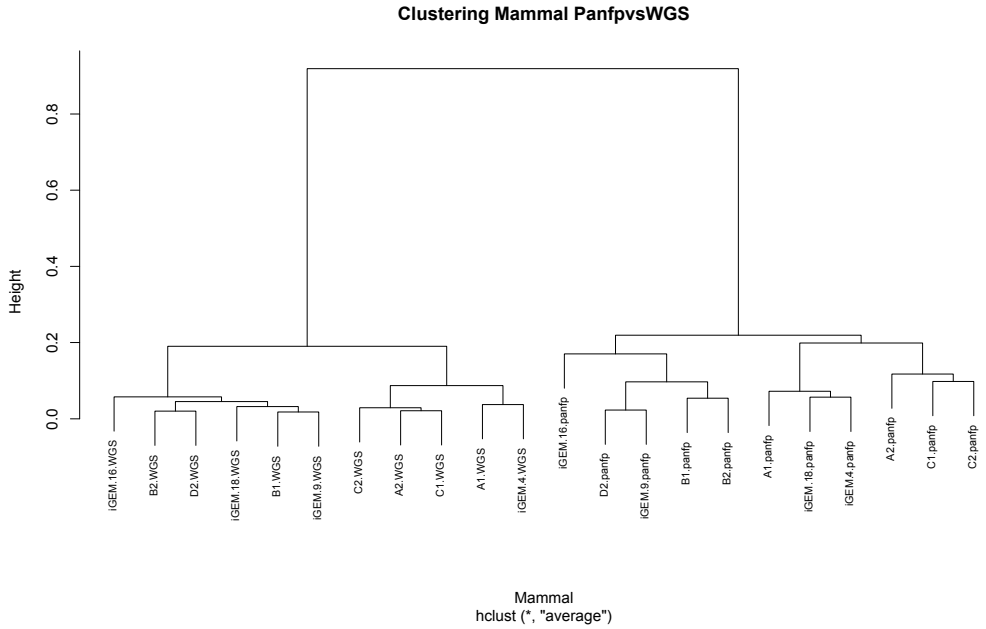


Figura 31: dendrogramma per il clustering gerarchico di PanFPvsWGS per il dataset Mammal.

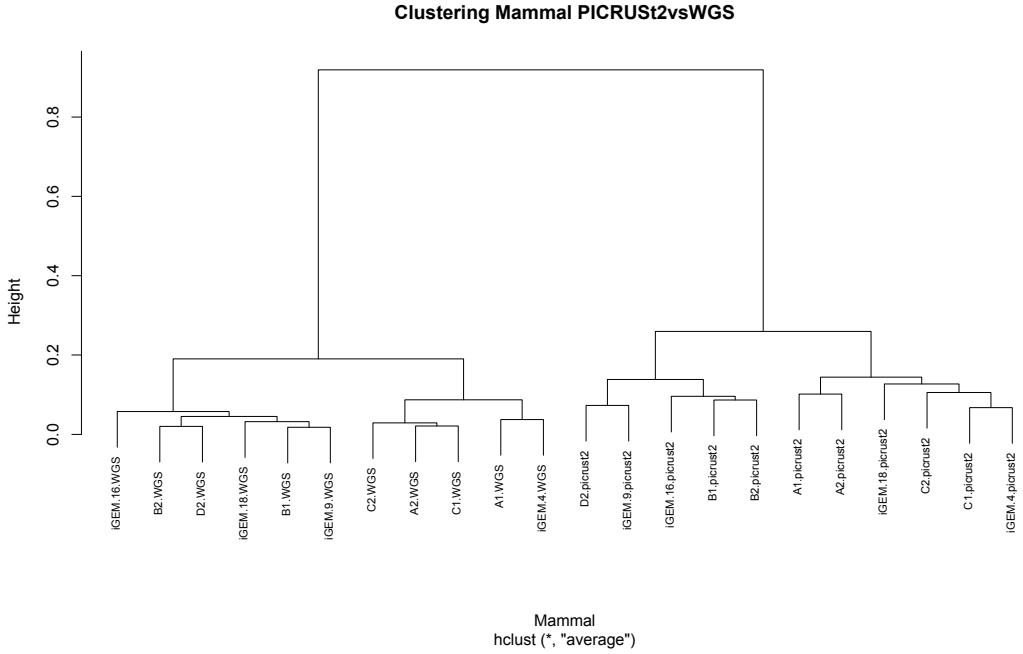


Figura 32: dendrogramma per il clustering gerarchico di PICRUSt2vsWGS per il dataset Mammal.

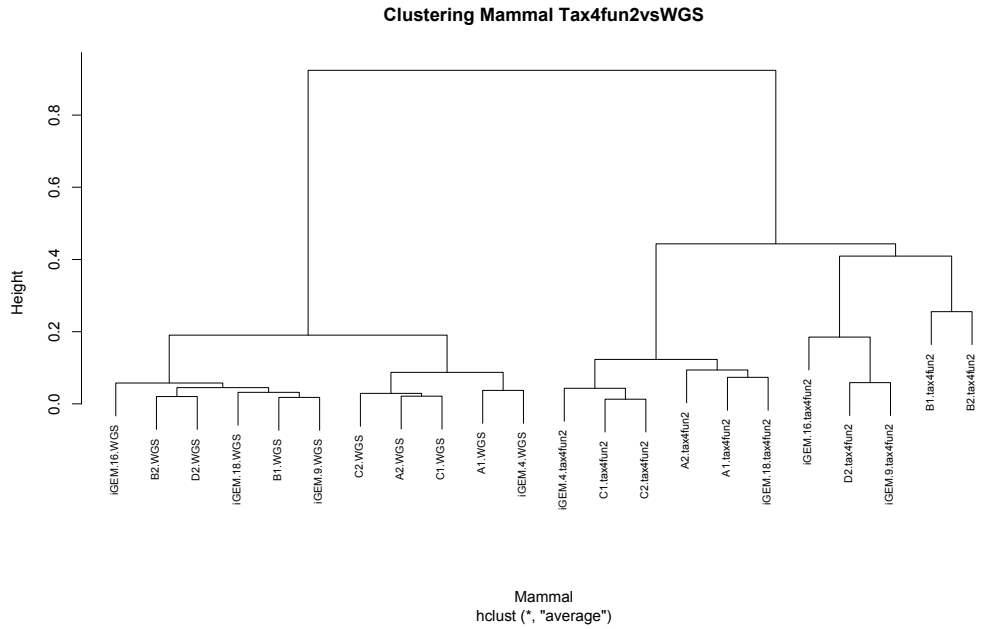


Figura 33: dendrogramma per il clustering gerarchico di Tax4Fun2vsWGS per il dataset Mammal.

L'unico caso in cui si possono vedere delle coppie tra un campione del tool e il campione corrispondente del WGS è quello del MammalianGut, che è anche l'unico dataset per cui il tool migliore risulta essere il Tax4Fun2. Quindi il clustering che permette di vedere queste coppie è quello tra Tax4Fun2 e WGS.

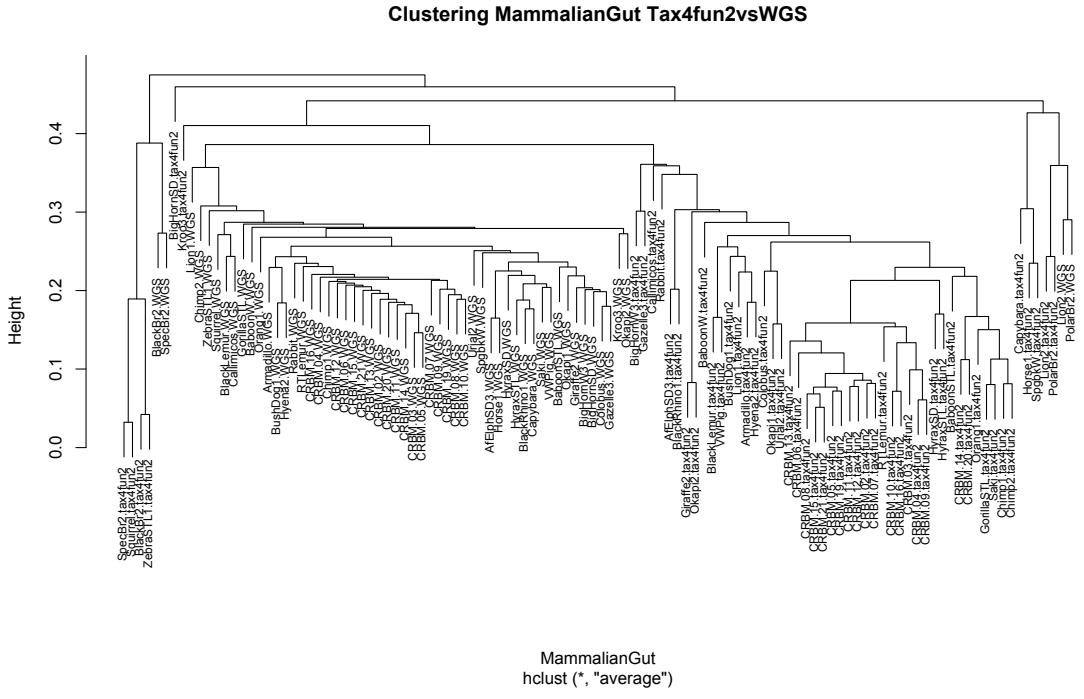


Figura 34: dendrogramma per il clustering gerarchico di Tax4Fun2vsWGS per il dataset MammalianGut.

Nella parte destra della figura, infatti, ci sono 2 coppie contenenti da una parte il campione del tool e dall'altra il campione del WGS e questi due campioni sono Lion2 e PolarB2. Idealmente, dovrebbero risultare sempre coppie composte in questa maniera ma purtroppo non siamo in quel caso e quindi anche a causa delle distanze molto ampie tra campioni è difficile che si ottengano foglie del dendrogramma perfettamente abbinate.

5.2 Analisi dei risultati ottenuti tramite le metriche che vanno a valutare la presenza/assenza dei geni nei dataframe per singolo dataset

Blueberry

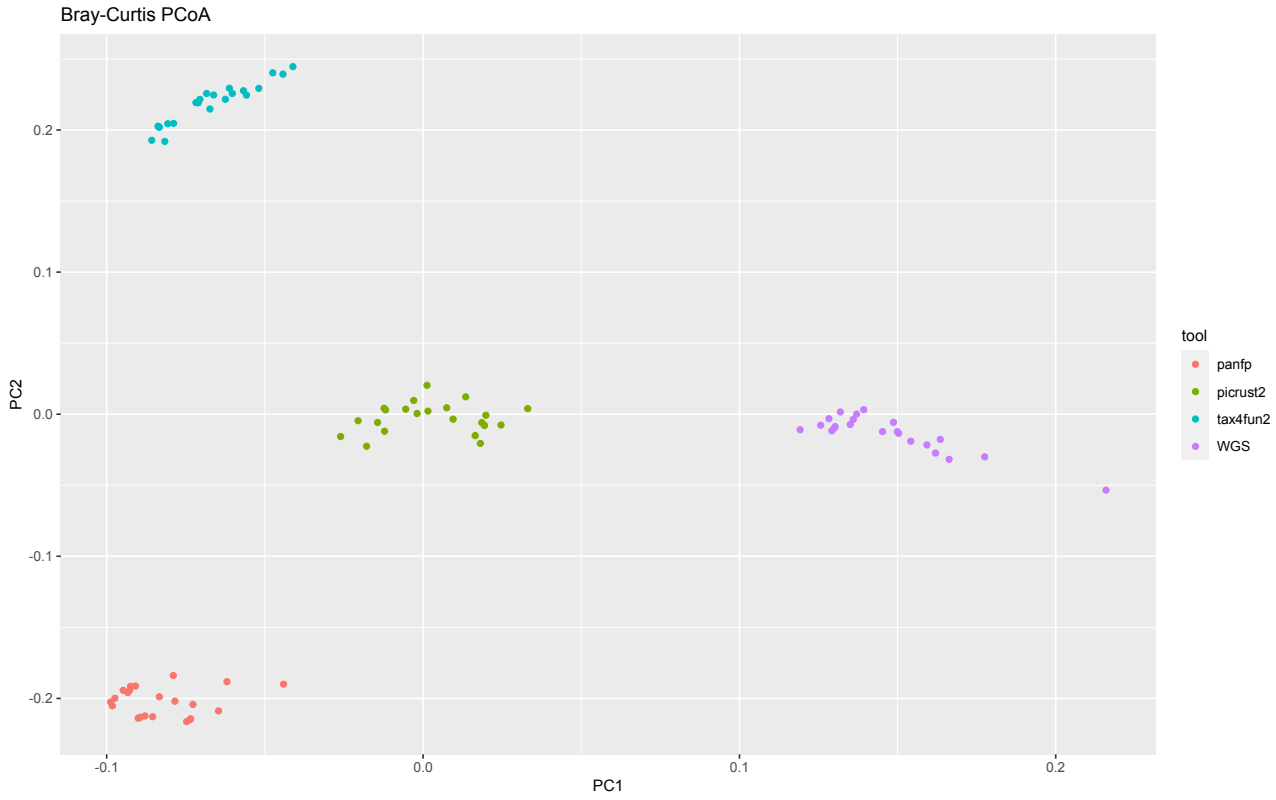


Figura 35: PCoA plot nel caso binario per rappresentare le distanze tra i vari campioni nei diversi tool e nel WGS per il dataset Blueberry, nell'asse x abbiamo la PC1 mentre nell'asse y abbiamo la PC2.

Per il dataset Blueberry, nella situazione in cui si considera l'indice di Bray-Curtis binario per ottenere il PCoA plot, dalla figura rappresentante quest'ultimo si vede come le predizioni ricavate tramite PICRUSt2 siano quelle che si avvicinano di più ai risultati ottenuti con il WGS. Infatti valutando le distanze si vede come per la PC1 e la PC2 esse risultino avere distanza minore dal WGS rispetto alle predizioni di Tax4Fun2 e PanFP.

Cameroon

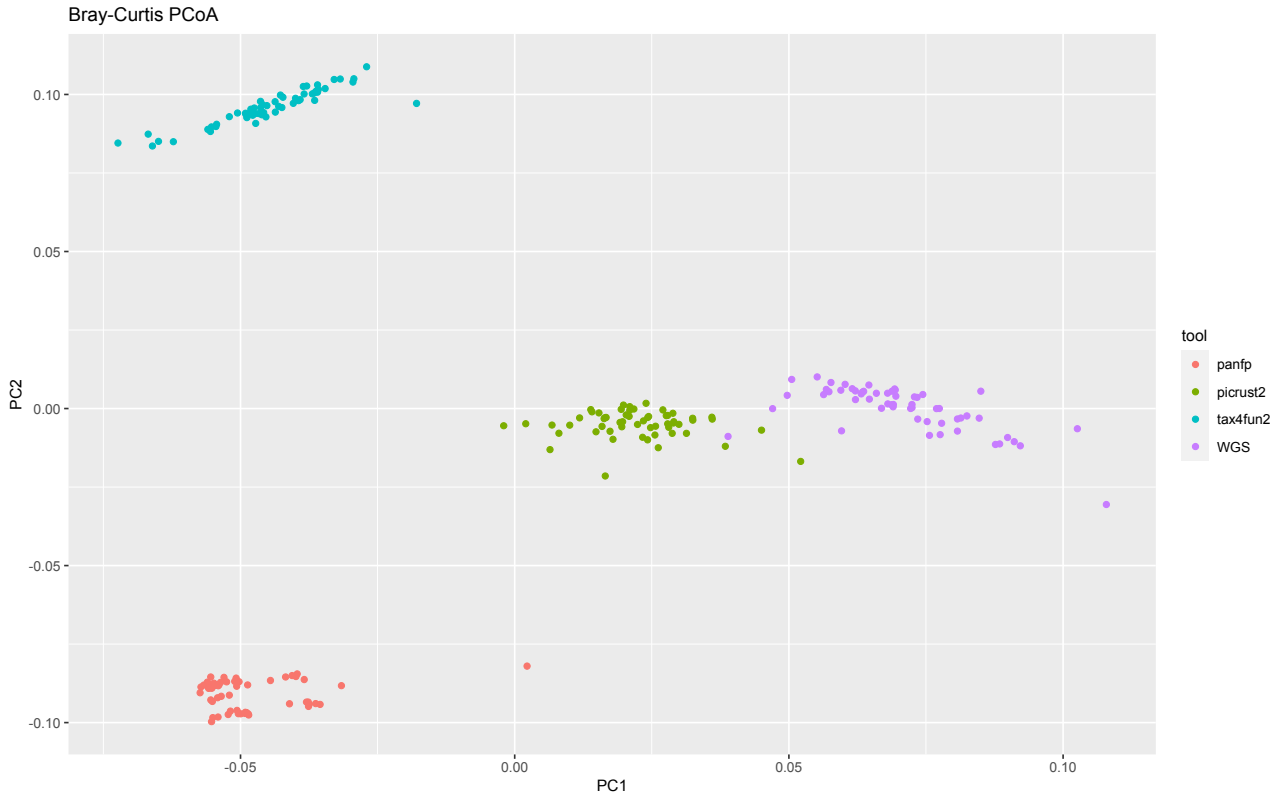


Figura 36: PCoA plot nel caso binario per rappresentare le distanze tra i vari campioni nei diversi tool e nel WGS per il dataset Cameroon, nell'asse x abbiamo la PC1 mentre nell'asse y abbiamo la PC2.

Anche per quel che riguarda il dataset Cameroon, dalla figura emerge come le predizioni ottenute tramite PICRUSt2 siano più vicine a quelle ricavate tramite il WGS rispetto alle stesse prodotte dal Tax4Fun2 e PanFP. Infatti valutando le distanze si vede come per la PC1 e la PC2 esse risultino avere distanza minore dal WGS rispetto alle predizioni di Tax4Fun2 e PanFP.

HMP

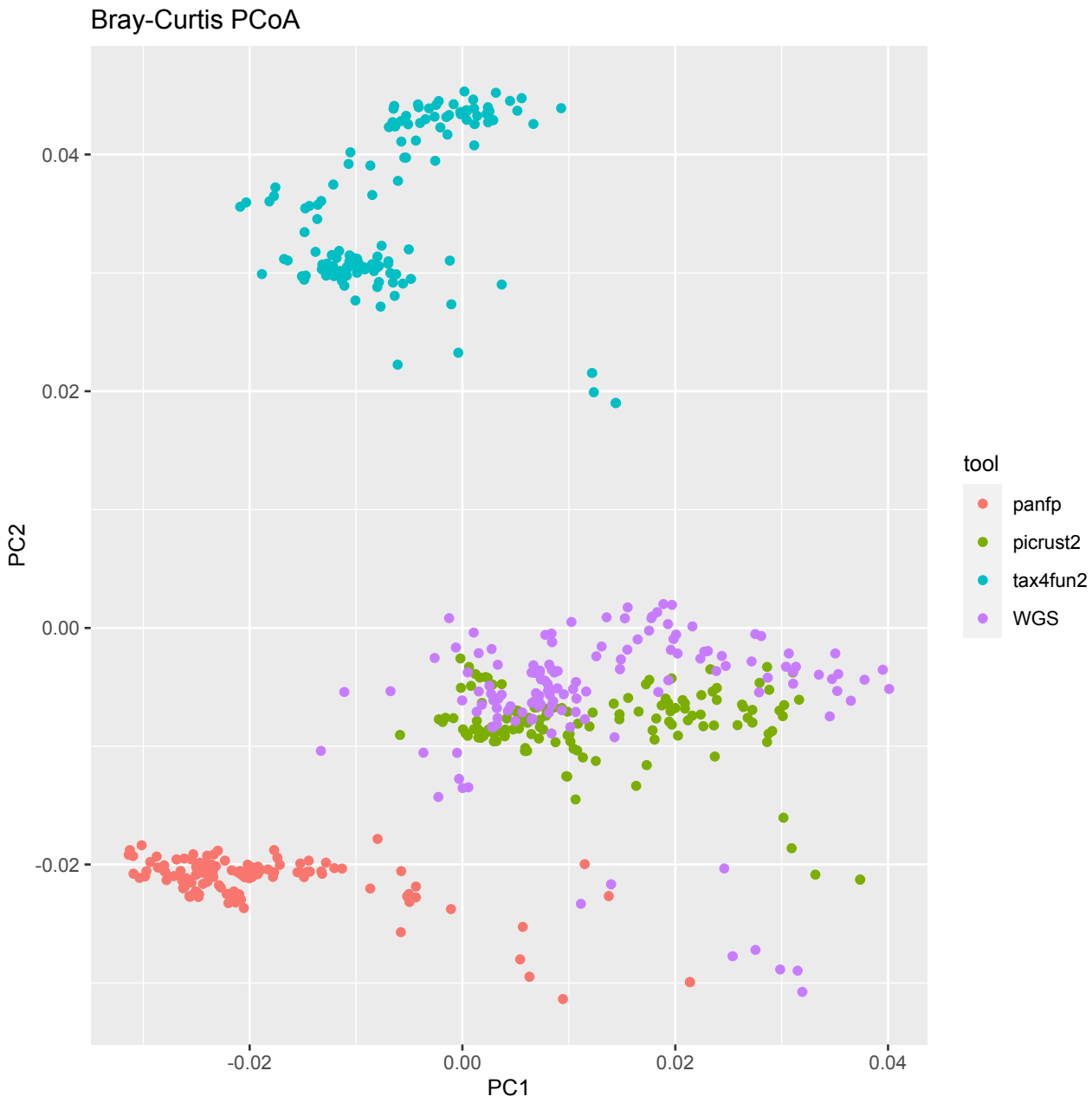


Figura 37: PCoA plot nel caso binario per rappresentare le distanze tra i vari campioni nei diversi tool e nel WGS per il dataset HMP, nell'asse x abbiamo la PC1 mentre nell'asse y abbiamo la PC2.

Per il dataset HMP, il miglior tool per le predizioni resta il PICRUSt2; in quello sopra si può vedere anche come le nuvole di punti del tool e del gold standard si mischiano tra di loro mentre per quel che riguarda PanFP e Tax4Fun2 permettono di ottenere delle predizioni che restano distanti da quelle del WGS, eccetto la predizione di qualche campione ottenuta con il PanFP

Indian

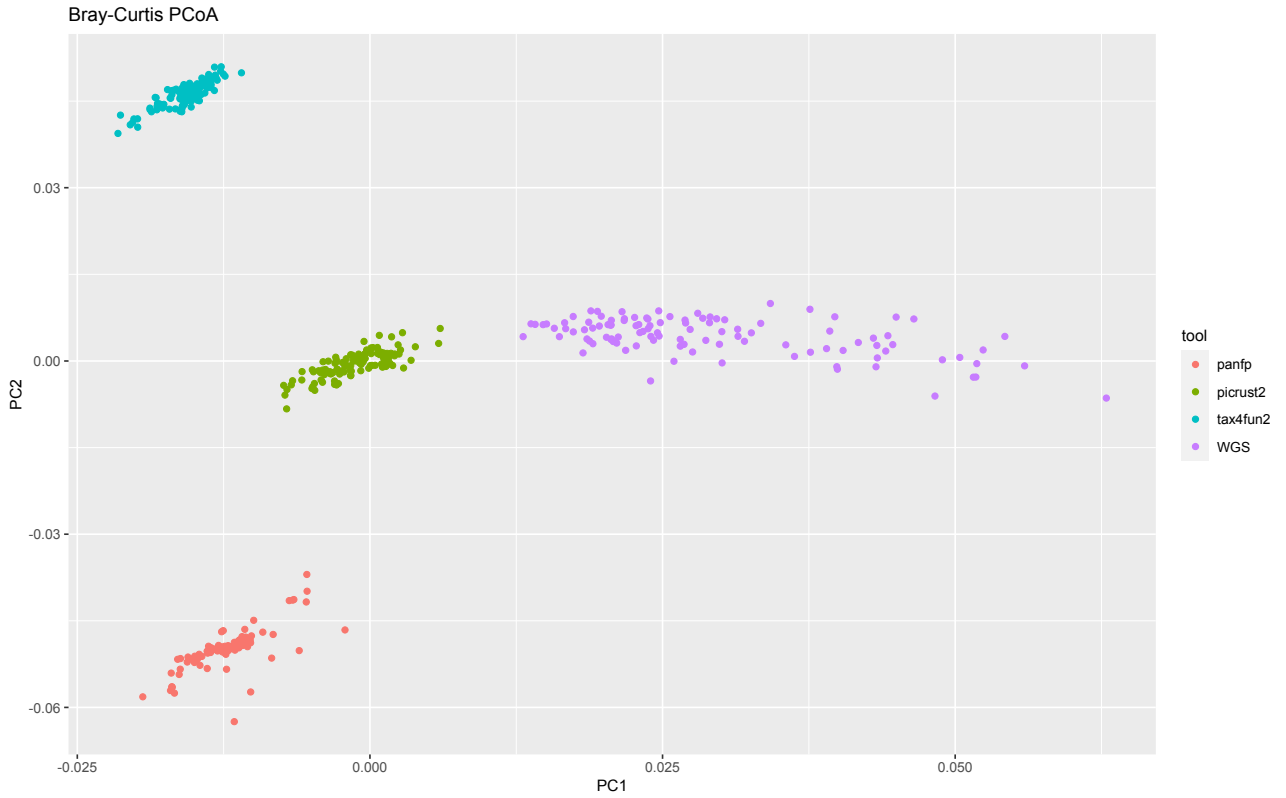


Figura 38: PCoA plot nel caso binario per rappresentare le distanze tra i vari campioni nei diversi tool e nel WGS per il dataset Indian, nell'asse x abbiamo la PC1 mentre nell'asse y abbiamo la PC2.

Anche per il dataset Indian il tool le cui predizioni si avvicinano di più a quelle del WGS - e quindi il migliore tra i tre è il PICRUSt2, mentre gli altri due sono caratterizzati da predizioni peggiori e quindi più distanti da quelle del WGS. Infatti valutando le distanze si vede come per la PC1 e la PC2 esse risultino avere distanza minore dal WGS rispetto alle predizioni di Tax4Fun2 e PanFP.

Mammal

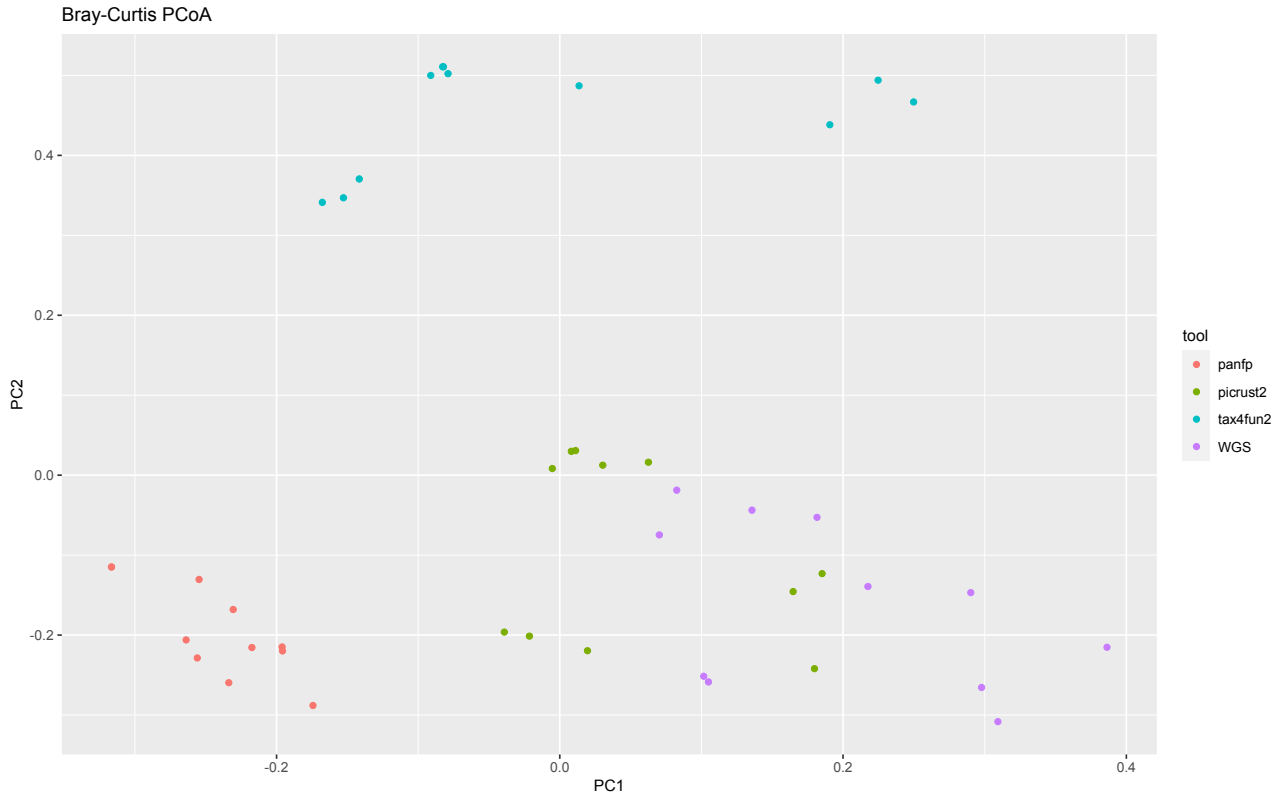


Figura 39: PCoA plot nel caso binario per rappresentare le distanze tra i vari campioni nei diversi tool e nel WGS per il dataset Mammal, nell'asse x abbiamo la PC1 mentre nell'asse y abbiamo la PC2.

Per il dataset Mammal, invece, si verifica una situazione simile a quella dell'HMP, poiché anche per questo dataset il tool migliore è PICRUSt2 e le sue predizioni si mischiano e si avvicinano a quelle ottenute tramite il WGS. Infatti valutando le distanze si vede come per la PC1 e la PC2 esse risultino avere distanza minore dal WGS rispetto alle predizioni di Tax4Fun2 e PanFP.

MammalianGut

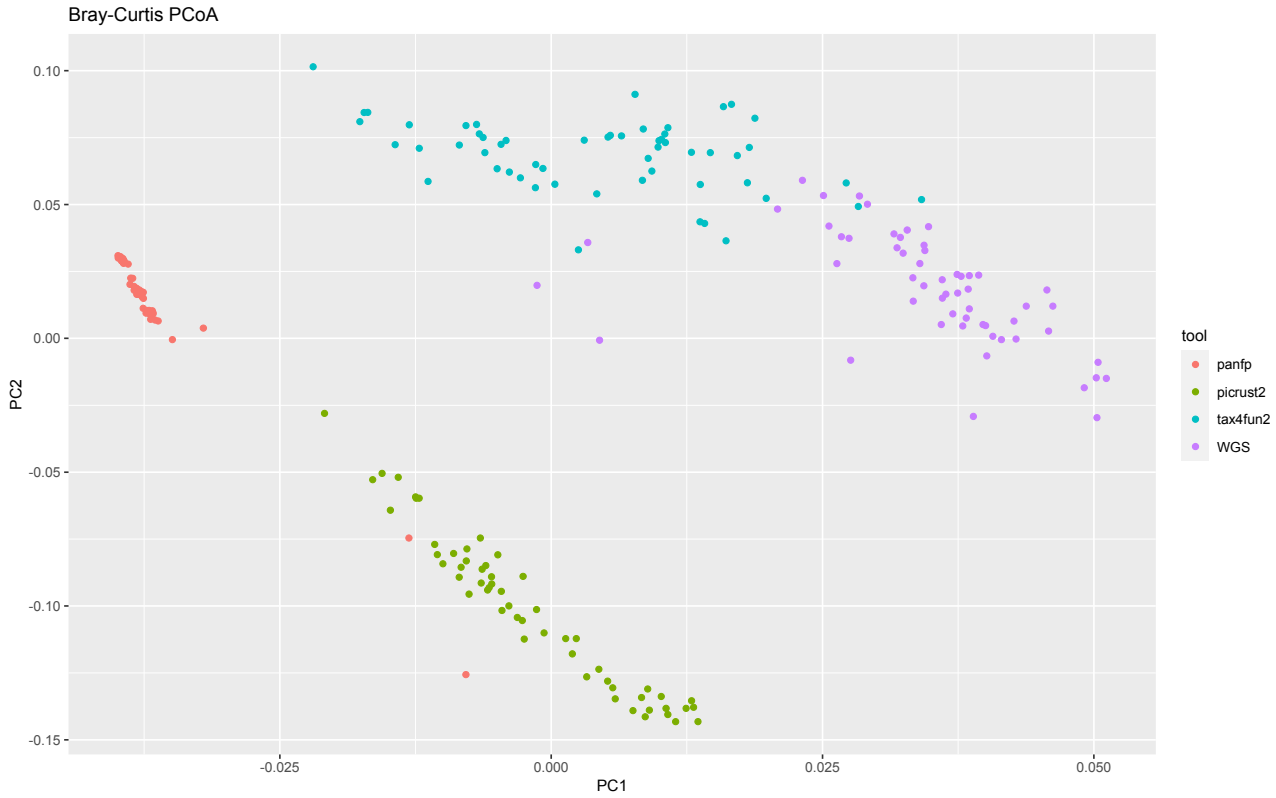


Figura 40: PCoA plot nel caso binario per rappresentare le distanze tra i vari campioni nei diversi tool e nel WGS per il dataset MammalianGut, nell'asse x abbiamo la PC1 mentre nell'asse y abbiamo la PC2.

Nel caso del dataset MammalianGut si ha invece l'unica situazione in cui il tool migliore è differente dal PICRUSt2; dal grafico sopra si vede come le predizioni migliori si ottengano tramite il Tax4Fun2 ed è possibile notare anche la vicinanza significativa di alcune di esse a quelle ottenute tramite WGS. Infatti valutando le distanze si vede come per la PC1 e la PC2 esse risultino avere distanza minore dal WGS rispetto alle predizioni di PICRUSt2 e PanFP.

MammalsSimulated

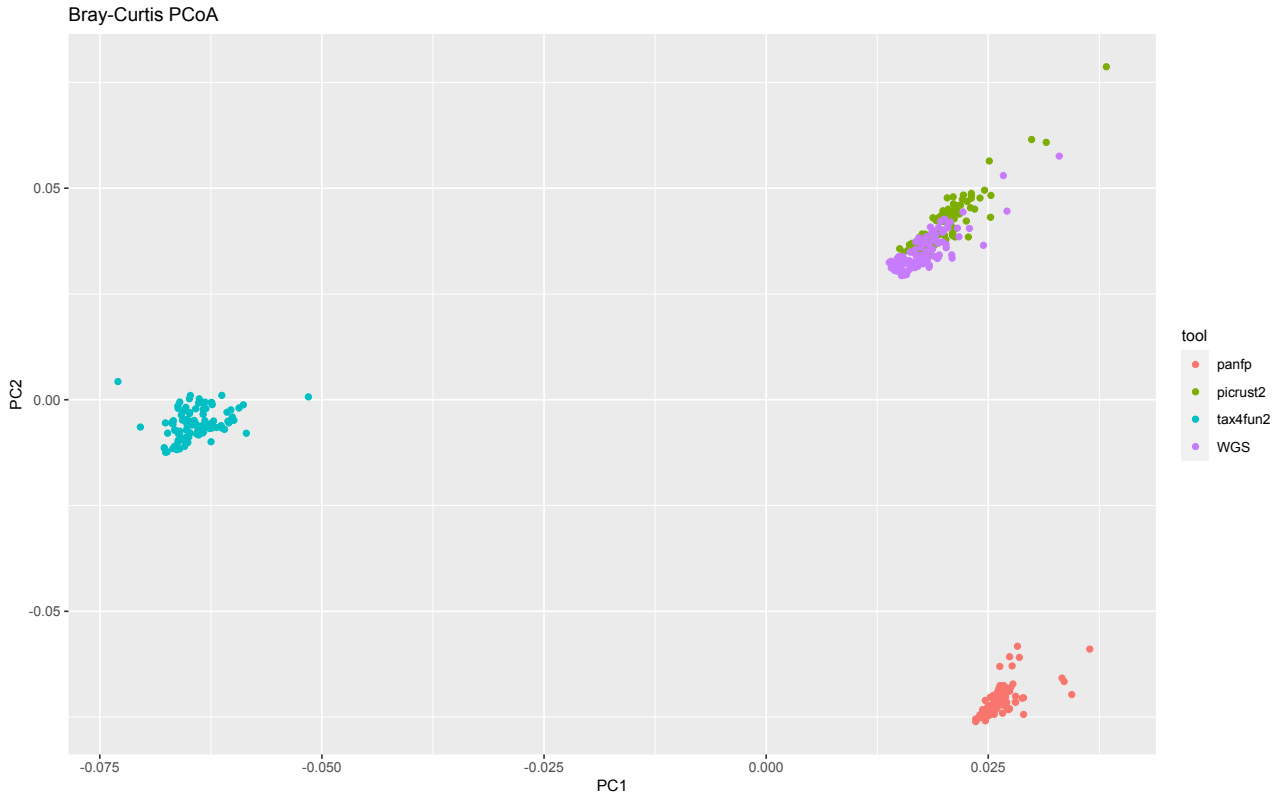


Figura 41: PCoA plot nel caso binario per rappresentare le distanze tra i vari campioni nei diversi tool e nel WGS per il dataset MammalsSimulated, nell'asse x abbiamo la PC1 mentre nell'asse y abbiamo la PC2.

Per il dataset MammalsSimulated si torna a rispettare il trend degli altri dataset, secondo cui il tool migliore è il PICRUSt2: dalla figura sopra si può vedere come ci sia una sovrapposizione quasi totale tra le predizioni dei campioni ottenute tramite il tool e quelle ricavate tramite il WGS, mentre Tax4Fun2 e PanFP permettono di ottenere delle predizioni distanti da quelle del gold standard e peggiori.

Ocean

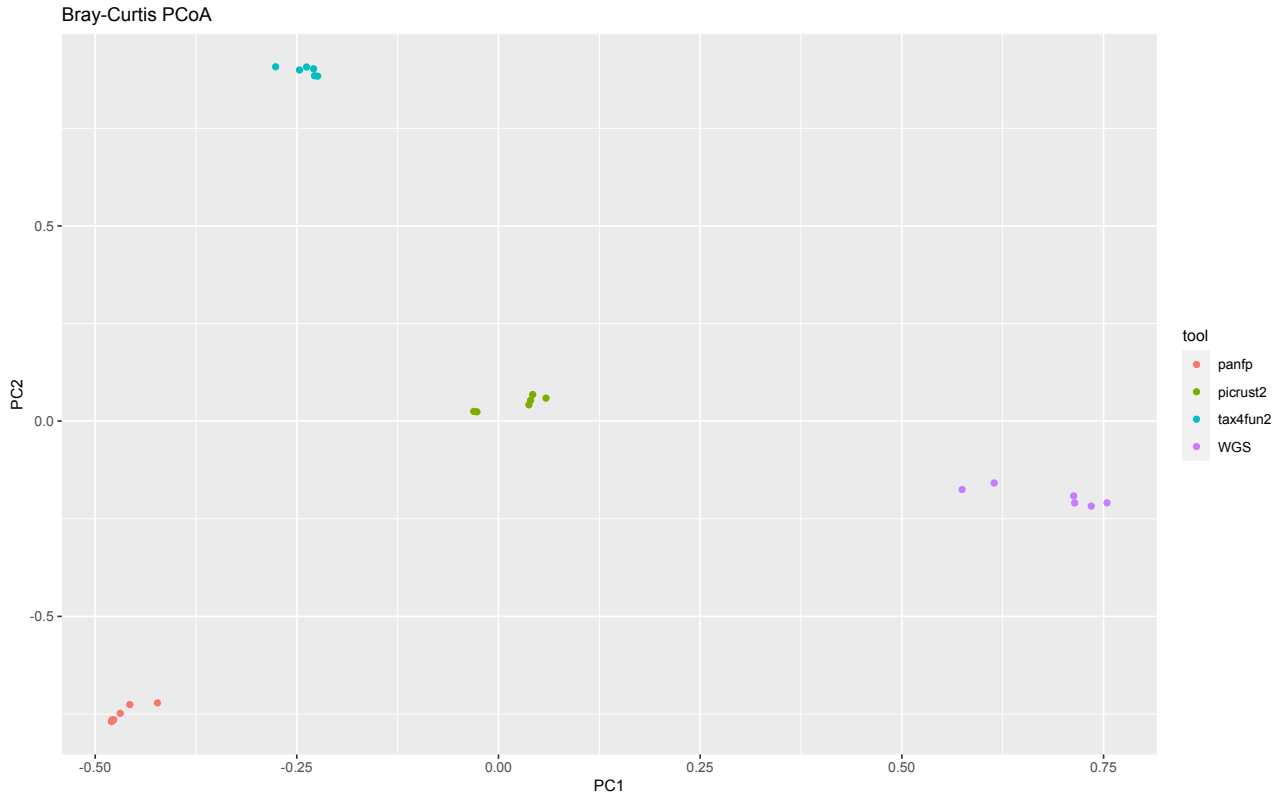


Figura 42: PCoA plot nel caso binario per rappresentare le distanze tra i vari campioni nei diversi tool e nel WGS per il dataset Ocean, nell'asse x abbiamo la PC1 mentre nell'asse y abbiamo la PC2.

Per il dataset Ocean il tool migliore è PICRUSt2 poiché le predizioni dei campioni ottenute tramite esso risultano essere più vicine a quelle del WGS rispetto alle corrispondenti ottenute attraverso gli altri tool. Infatti valutando le distanze si vede come per la PC1 e la PC2 esse risultino avere distanza minore dal WGS rispetto alle predizioni di Tax4Fun2 e PanFP

Primate

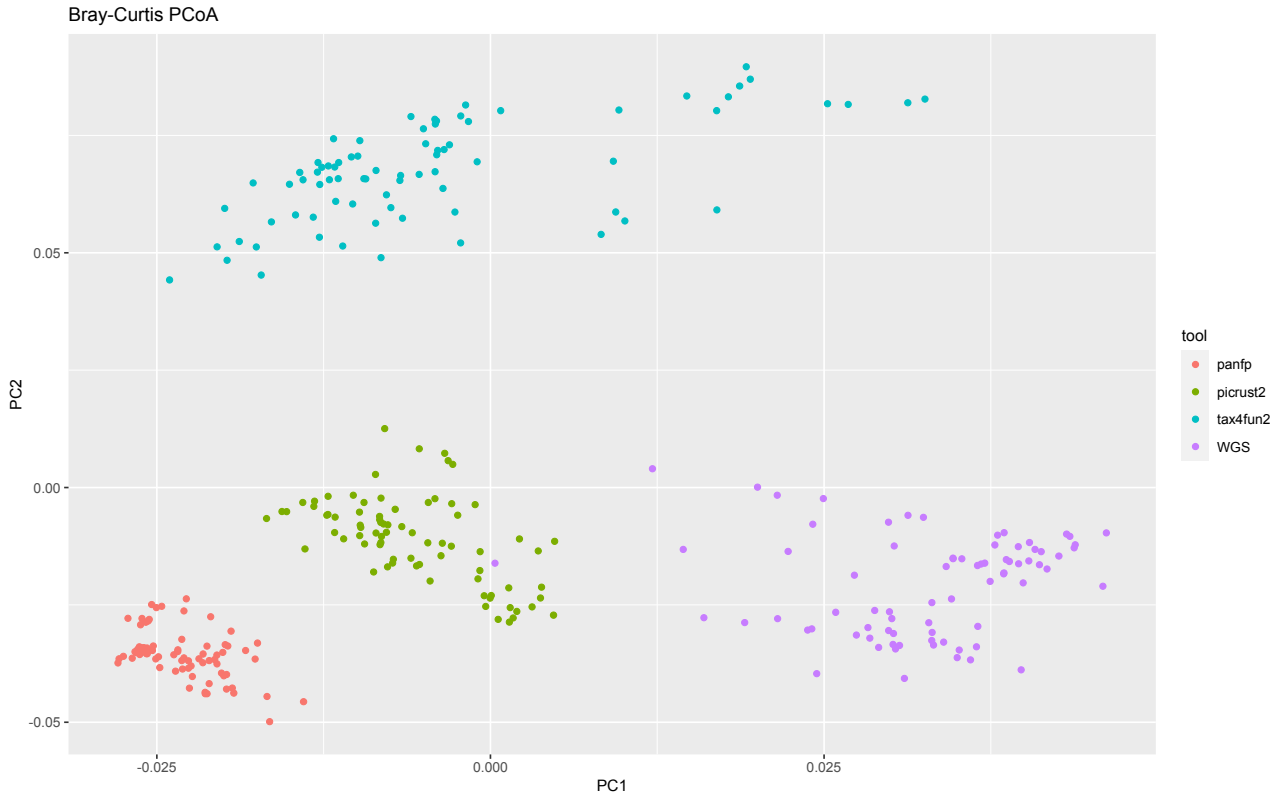


Figura 43: PCoA plot nel caso binario per rappresentare le distanze tra i vari campioni nei diversi tool e nel WGS per il dataset Primate, nell'asse x abbiamo la PC1 mentre nell'asse y abbiamo la PC2.

Per il Primate il tool migliore è sempre PICRUSt2 perché, anche in questo caso, le sue predizioni sono più vicine ai risultati ottenuti tramite il WGS rispetto alle stesse ricavate con gli altri due tool (PanFP e Tax4Fun2). Infatti valutando le distanze si vede come per la PC1 e la PC2 esse risultino avere distanza minore dal WGS rispetto alle predizioni di Tax4Fun2 e PanFP.

A supporto dei risultati di cui sopra, secondo cui, anche nel caso binario, il PICRUSt2 risulta essere il miglior tool per tutti i dataset eccetto per il MammalianGut, si producono gli istogrammi ottenuti per le altre quattro metriche binarie considerate, ossia Balanced Accuracy, Precision, Recall e F_1_score.

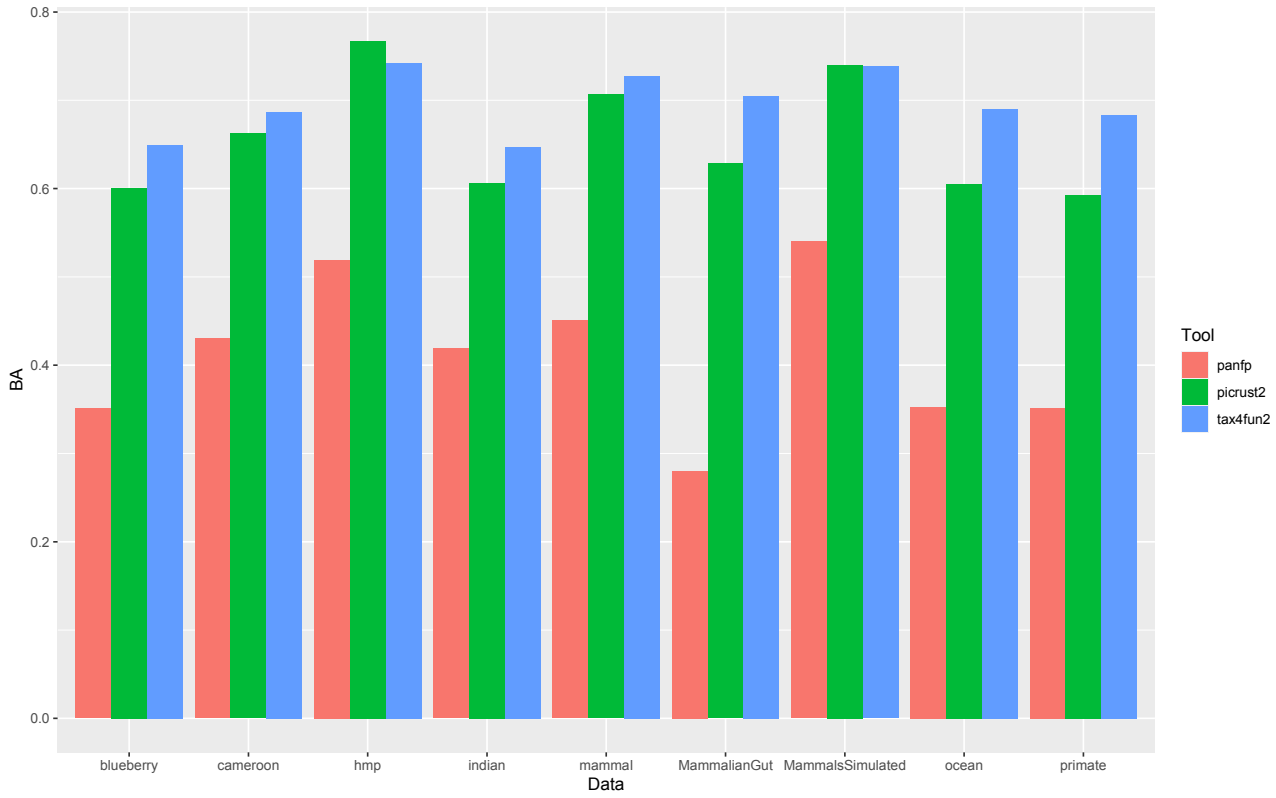


Figura 44: istogramma della Balanced Accuracy per i vari tool raggruppato per dataset, nell'asse x abbiamo la suddivisione per dataset e ogni dataset ha una colonnina per ogni tool mentre nell'asse y abbiamo i valori della balanced accuracy.

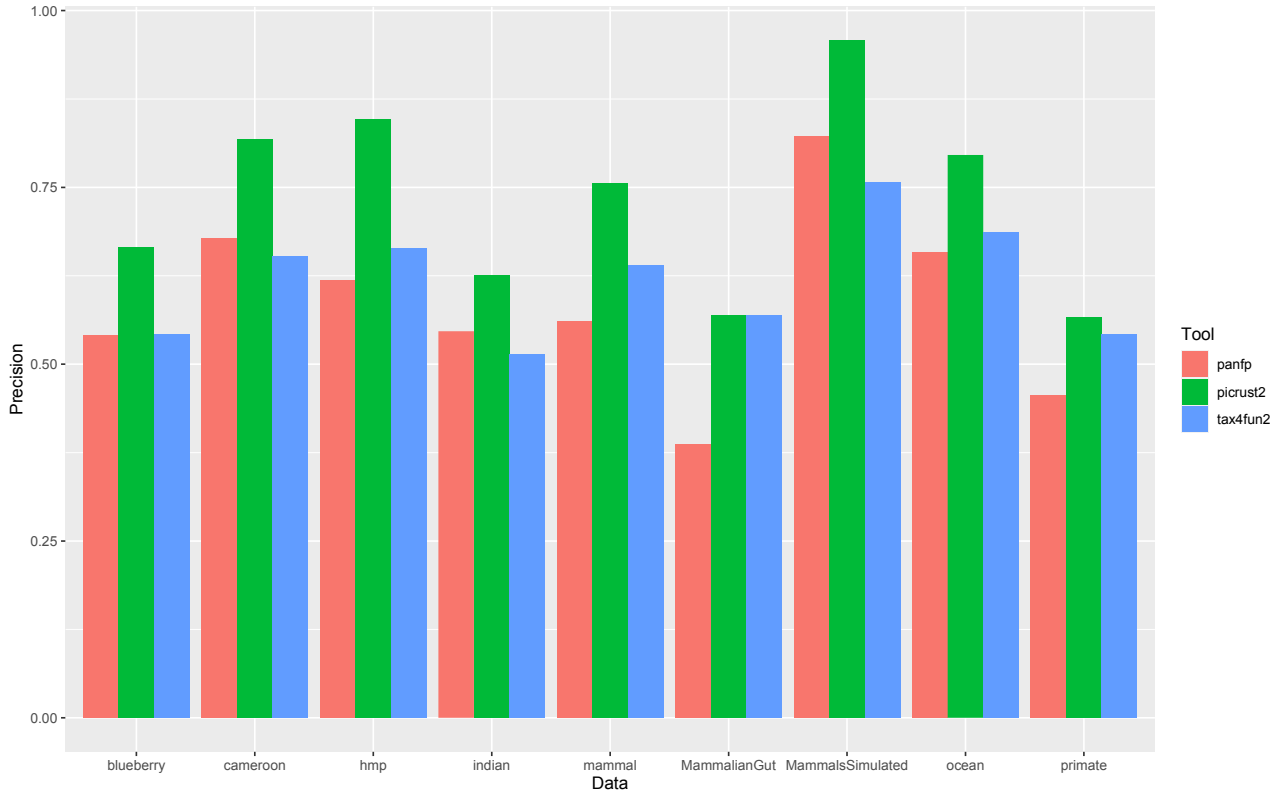


Figura 45: istogramma della Precision per i vari tool raggruppato per dataset infatti nell'asse x abbiamo la suddivisione per dataset e ogni dataset ha una colonnina per ogni tool mentre nell'asse y abbiamo i valori della Precision.

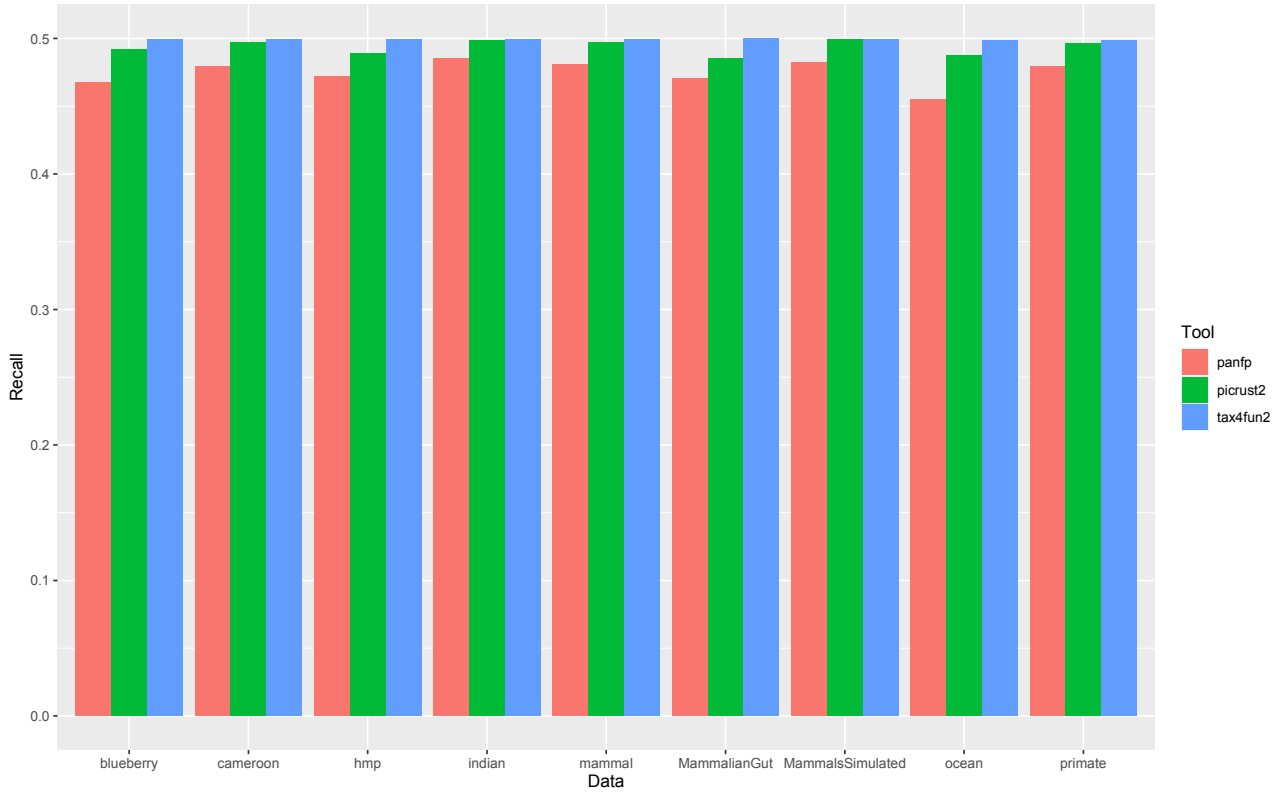


Figura 46: istogramma della Recall per i vari tool raggruppato per dataset infatti nell'asse x abbiamo la suddivisione per dataset e ogni dataset ha una colonnina per ogni tool mentre nell'asse y abbiamo i valori della Recall.

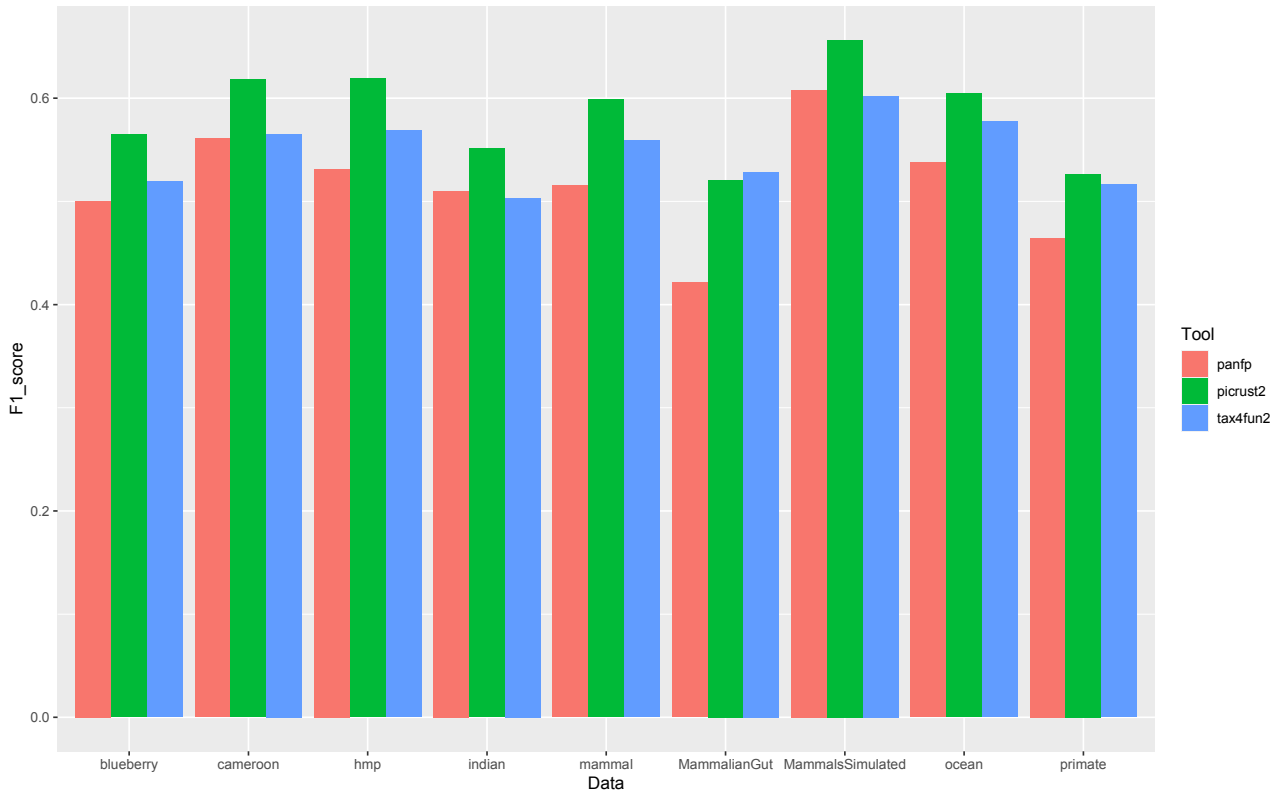


Figura 47: istogramma dell'F_1_score per i vari tool raggruppato per dataset infatti nell'asse x abbiamo la suddivisione per dataset e ogni dataset ha una colonnina per ogni tool mentre nell'asse y abbiamo i valori dell'F_1_score.

Dai 4 grafici sopra riportati emerge come, per quel che riguarda la Balanced Accuracy, lo strumento migliore sia il Tax4Fun2, nel senso che tramite esso si riesce ad ottenere un'accuratezza migliore nel predire la presenza o l'assenza dei geni all'interno dei singoli dataset. Dalla valutazione degli altri tre grafici si vede come PICRUSt2 sia il migliore all'interno dell'istogramma che valuta la Precision. Nella Recall c'è un andamento simile tra PICRUSt2 e Tax4Fun2. Infine, nell'istogramma per valutare F_1_score, ottenuto mettendo insieme Precision e Recall, PICRUSt2 risulta essere il migliore per tutti i dataset eccetto per il MammalianGut, dove il migliore è Tax4Fun2, come emerso dall'analisi dei PCoA plot.

In seguito, anche tramite i dendrogrammi del clustering, si può vedere come ciò che risulta dai plot della PCoA sia rispettato sia nel caso di predizioni che sono distanti tra di loro, e quindi portano a foglie ben divise con da una parte i campioni del tool e dall'altra quelli del WGS, sia nel caso in cui ci sia sovrapposizione. Questo risulta chiaramente tramite i clustering di Blueberry, Ocean e Mammal, che sono quelli con meno campioni, mentre gli altri clustering saranno presenti nel materiale supplementare (1).

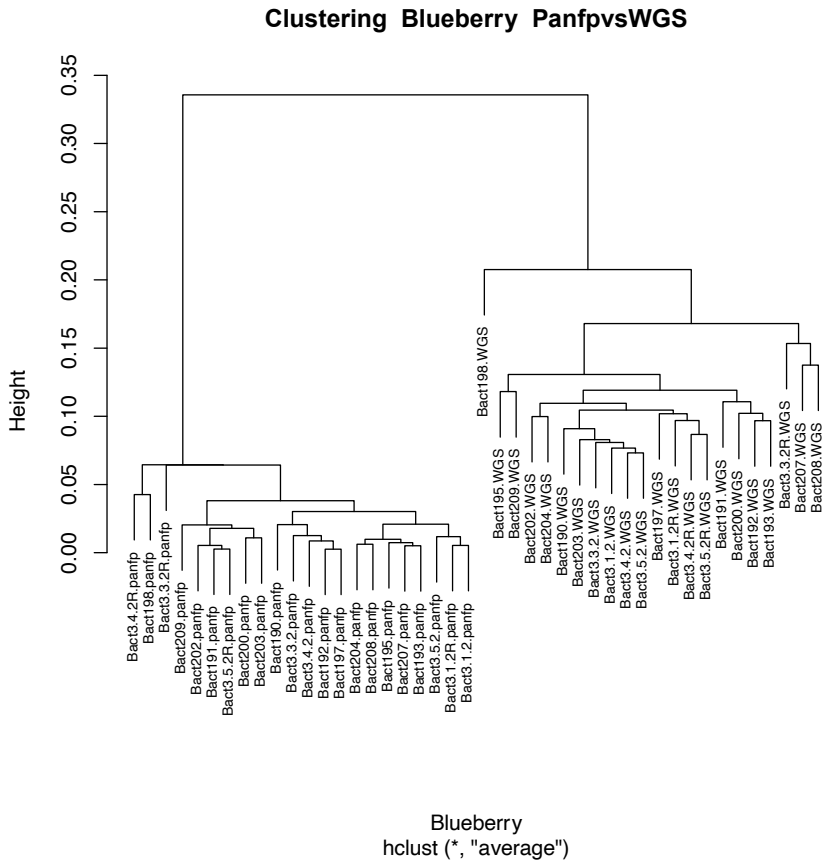


Figura 48: dendrogramma per il clustering gerarchico nel caso binario di PanFPvsWGS per il dataset Blueberry.

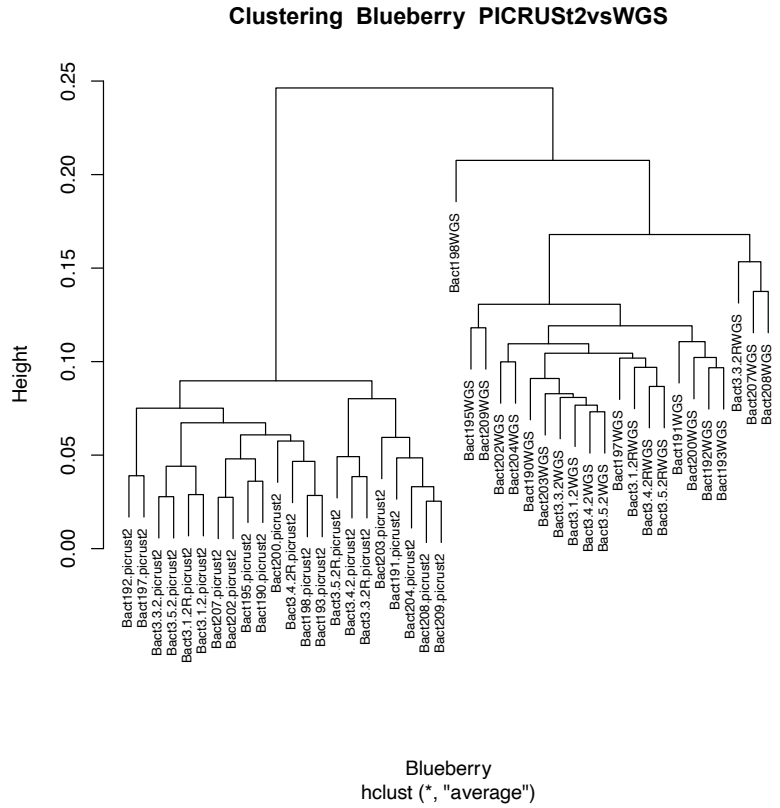


Figura 49: dendrogramma per il clustering gerarchico nel caso binario di PICRUSt2vsWGS per il dataset Blueberry.

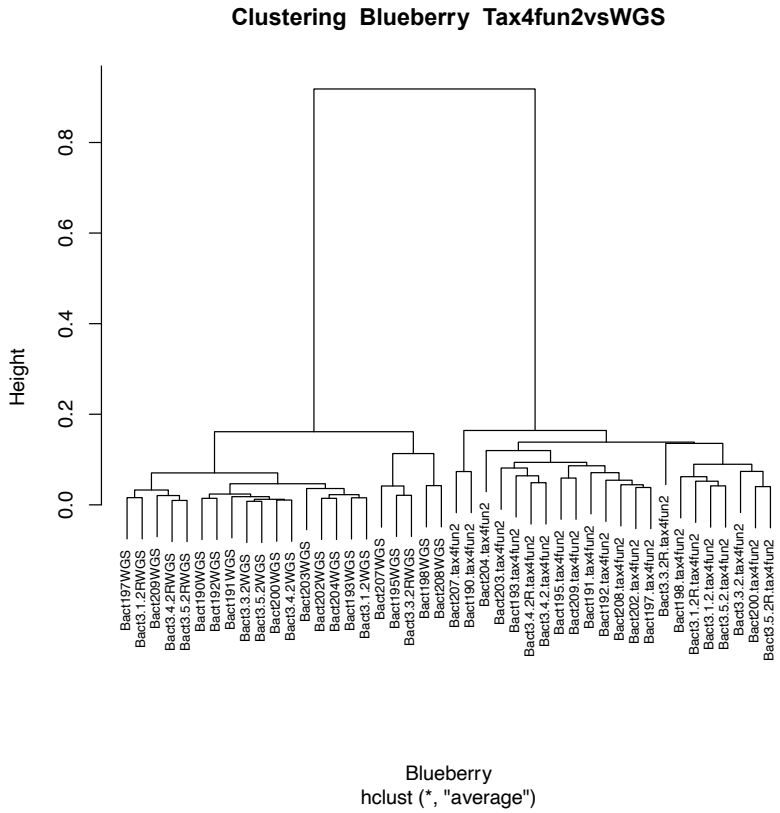


Figura 50: dendrogramma per il clustering gerarchico nel caso binario di Tax4Fun2vsWGS per il dataset Blueberry.

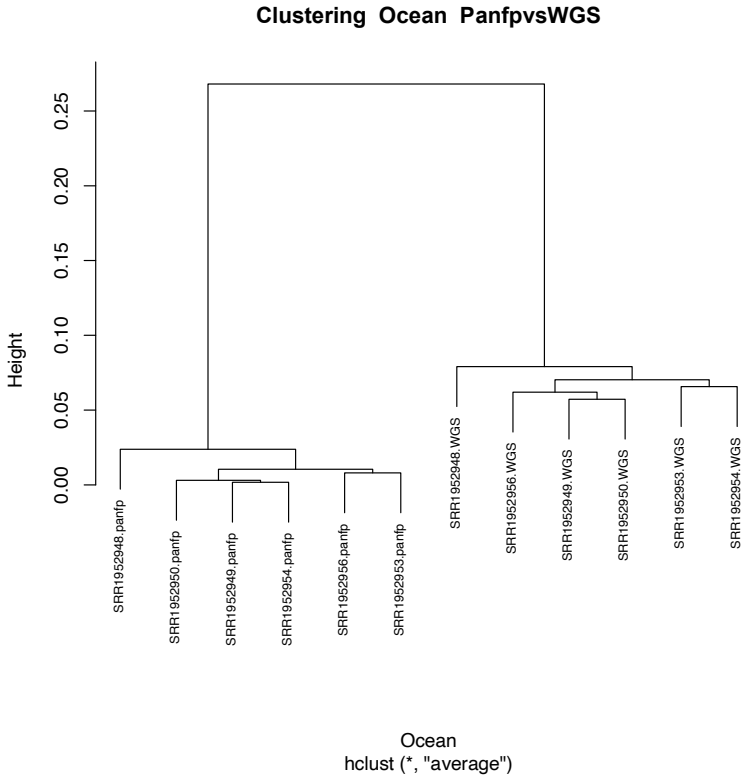


Figura 51: dendrogramma per il clustering gerarchico nel caso binario di PanFPvsWGS per il dataset Ocean.

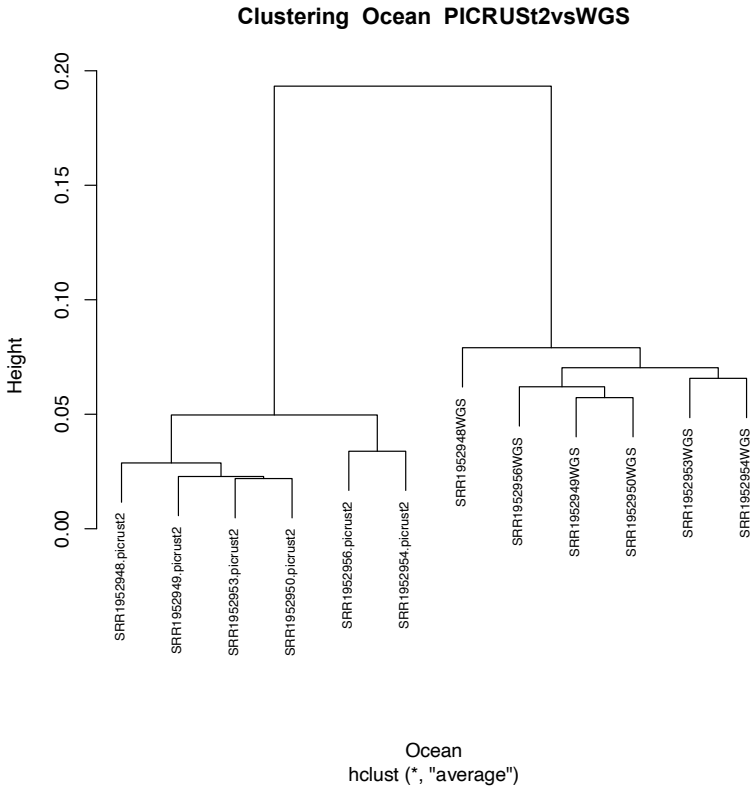


Figura 52: dendrogramma per il clustering gerarchico nel caso binario di PICRUSt2vsWGS per il dataset Ocean.

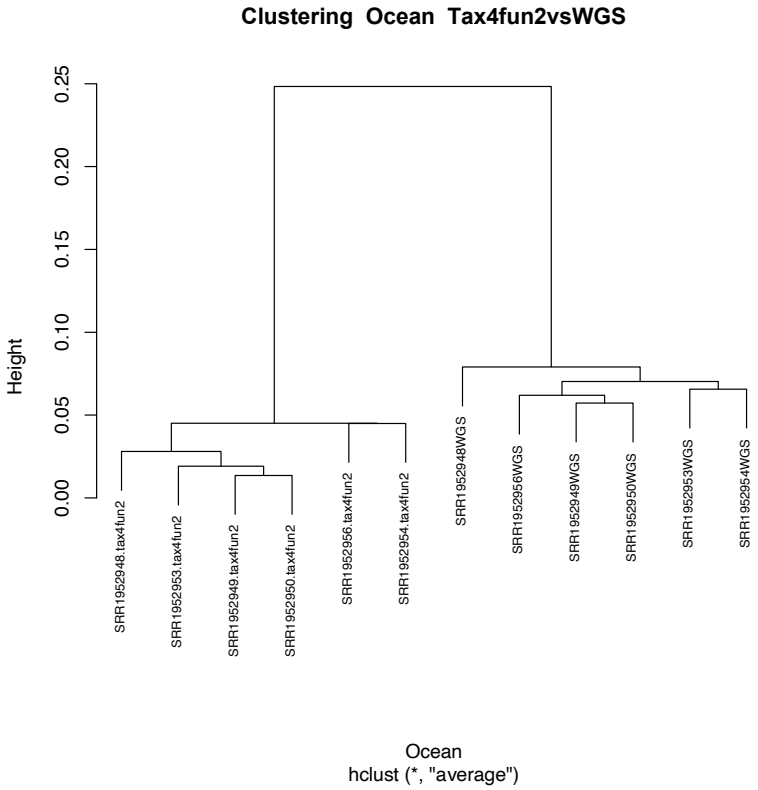


Figura 53: dendrogramma per il clustering gerarchico nel caso binario di Tax4Fun2vsWGS per il dataset Ocean.

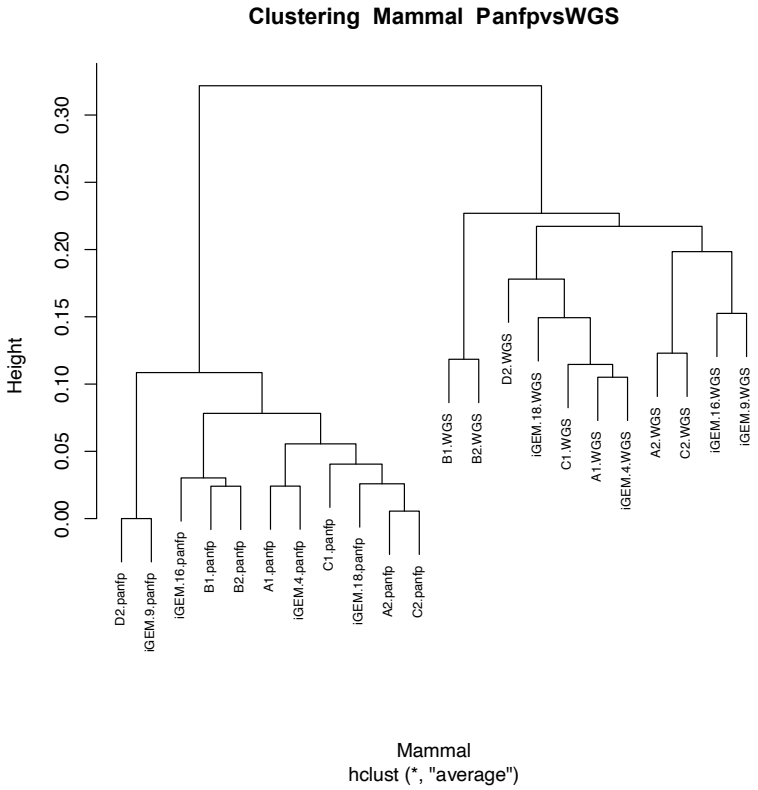


Figura 54: dendrogramma per il clustering gerarchico nel caso binario di PanFPvsWGS per il dataset Mammal.

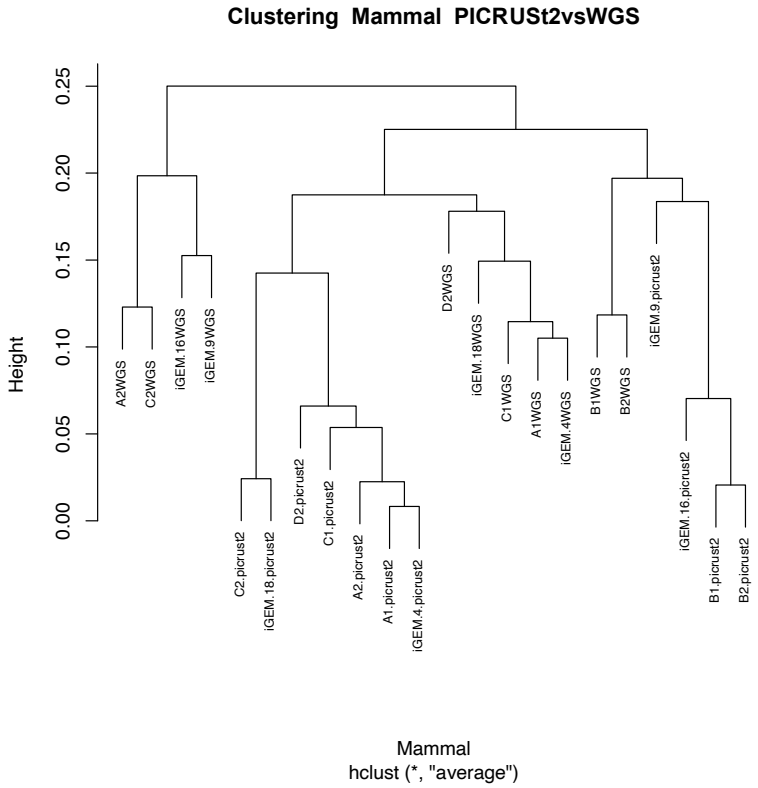


Figura 55: dendrogramma per il clustering gerarchico nel caso binario di PICRUSt2vsWGS per il dataset Mammal.

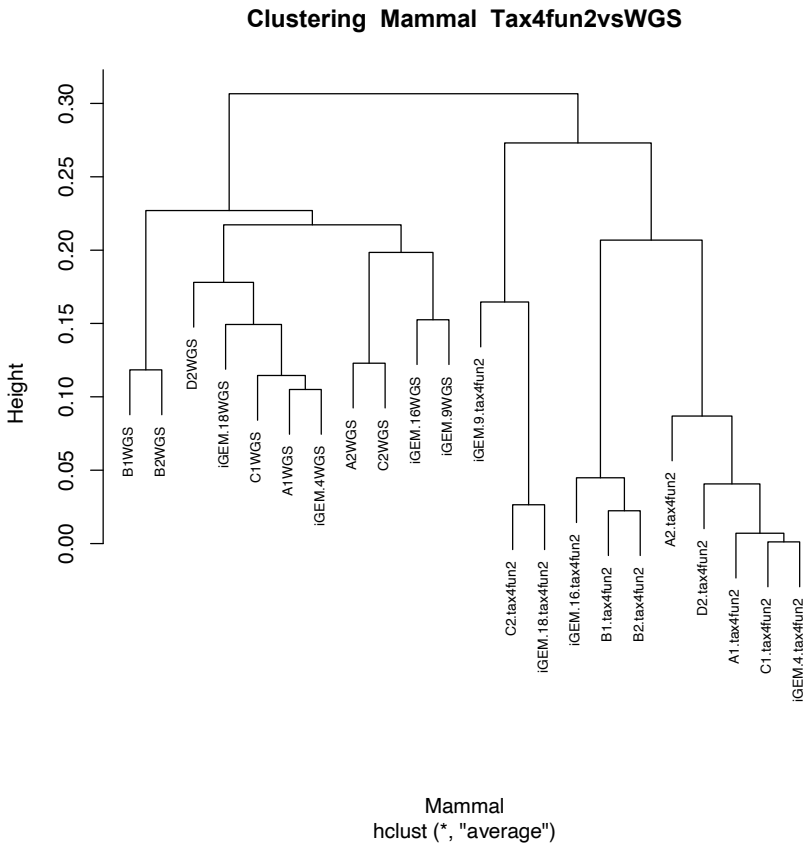


Figura 56: dendrogramma per il clustering gerarchico nel caso binario di Tax4Fun2vsWGS per il dataset Mammal.

Capitolo 6 - Conclusioni

All'interno di questo lavoro di tesi si è partiti dalla descrizione del microbiota e di cosa vuol dire sequenziamento. In seguito sono state analizzate due diverse tecniche di sequenziamento ovvero il 16S e il WGS e si è anche effettuato un confronto tra esse. L'intento con cui si è partiti per fare questo lavoro di tesi era quello di trovare un tool che permettesse partendo dalle tabelle ASV, ottenute dopo aver utilizzato il sequenziamento 16S di inferire il loro potenziale funzionale in modo da ottenere risultati che si avvicinassero il più possibile a quelli del WGS. Sono stati consultati vari articoli scientifici per selezionare i tool da mettere a confronto e sono stati scelti i seguenti ossia PanFP, PICRUSt2 e Tax4Fun2 in quanto erano più promettenti degli altri in quanto permettono di ottenere risultati migliori.

Dopo aver selezionato i tool si sono estratti, dagli stessi articoli scientifici, dei dataset su cui poterli applicare. Si sono selezionati nove dataset provenienti da diversi individui, diversi animali e anche diversi campioni di sottosuolo. Ognuno di questi dataset era caratterizzato da una tabella ASV. utilizzata come input per ognuno dei tre strumenti. Per il PanFP era necessario che a questa tabella fosse aggiunta una colonna con la classificazione tassonomica per ogni ASV. I dataset che sono stati scelti si possono suddividere in dataset reali ossia Blueberry, Cameroon, HMP, Indian, Mammal, MammalianGut, Ocean e Primate e dataset simulato ossia il MammalsSimulated. Sempre tramite i suddetti articoli si sono, in seguito, scelte le metriche tramite cui confrontare i tre diversi tool. Le metriche che sono state scelte sono sia binarie sia non binarie. Le metriche binarie servono per valutare sia la distanza tra la predizione, ottenuta tramite il tool, e l'obiettivo ossia il risultato del WGS sia per confrontare i geni presenti o assenti nella predizione con i medesimi all'interno dei risultati ottenuti con il WGS. Le metriche non binarie, invece, servono per valutare la distanza tra le predizioni ottenute tramite i tool e i risultati ottenuti tramite il WGS e vedere quale tool permette di ricavare predizioni che si avvicinano di più al gold standard (WGS). Le metriche binarie sono: Balanced Accuracy, Precision, Recall e F_1_score, che vengono valutate tramite degli istogrammi, avendo sull'asse x la distinzione per dataset e per ogni dataset una colonna di colore diverso per ogni tool, l'Indice di Bray-Curtis binario tramite il quale si ottengono i plot della PCoA per poter vedere i vari cluster, rappresentanti i tool, e anche le distanze tra le predizioni dei singoli tool e quelle tra i tool e il WGS; con l'indice di Bray-Curtis si ottengono, in seguito, i dendrogrammi che rappresentano i cluster e essi rispecchiano i risultati della PCoA; infatti la situazione ideale ossia quella con le coppie di foglie costituite da un campione del tool e da uno del WGS si realizza solamente in pochi casi. Le metriche non binarie invece sono: correlazione di Spearman e Indice di Bray-Curtis. La prima delle

due permette di ricavare i vari boxplot, aventi lungo l'asse x la distinzione tra i diversi tool mentre nell'asse y i vari valori di correlazione ottenuti con ogni tool e in questo caso le valutazioni sono state fatte basandosi sulla mediana di ogni boxplot.
L'Indice di Bray-Curtis, come detto in precedenza, permette di ottenere i plot della PCoA e i dendrogrammi del clustering; questi sono stati valutati basandosi sulle distanze tra le predizioni dei campioni, ottenute con i tool, e i risultati del WGS e suddette distanze sono rispecchiate dal clustering, dove le foglie che rappresentano i campioni dei tool sono quasi sempre separate da quelle dei campioni del WGS. I plot della PCoA sono utili per poter vedere, in due dimensioni ossia quelle della PC1 e della PC2, le predizioni dei campioni ottenute tramite i diversi tool e quelle ottenute tramite il WGS in modo da poter vedere quale tool permette di ottenere risultati che più si avvicinano a quelli del WGS. I dendrogrammi invece sono caratterizzati da delle foglie che rappresentano le predizioni dei campioni tramite un tool e le stesse tramite un WGS, esse rispecchieranno il plot della PCoA in quanto se al suo interno ci sono coppie di predizioni correttamente abbinate, ci saranno anche nel dendrogramma. Gli istogrammi, invece, permettono di vedere i livelli, tramite le colonne di colore diverso per ogni tool, delle 4 metriche binarie.
Grazie a dette analisi si è raggiunto l'obiettivo iniziale di determinare quale sia il tool migliore per ottenere l'inferenza del profilo funzionale di una tabella ASV in modo che si avvicini il più possibile al profilo funzionale della stessa ottenuto tramite il WGS. Dall'analisi dei risultati si può concludere che il tool migliore per ottenere delle buone predizioni di profili funzionali in vari campioni e all'interno dei vari dataset sia il PICRUSt2. Si evidenziano però, tre eccezioni. La prima riguarda il dataset MammalianGut, per il quale il tool migliore è il Tax4Fun2. La seconda riguarda il dataset MammalsSimulated, per il quale nel caso non binario il tool migliore è PanFP. La terza, invece, riguarda una delle metriche binarie, ossia la Balanced Accuracy per la quale il tool migliore risulta essere ancora il Tax4Fun2. Nei vari dataset, inoltre, si possono notare delle differenze sia per le metriche binarie sia per quelle non binarie infatti alcuni di essi evidenziano come le predizioni dei campioni ottenute tramite PICRUSt2 siano molto vicine e ci siano anche alcune sovrapposizioni.
Per esempio, per MammalsSimulated sia nel caso binario sia nel caso non binario, per HMP e Mammal invece, questo è vero solo nel caso binario. Quanto sopra vale anche per il dataset MammalianGut sia nel caso binario sia in quello non binario ma non per il PICRUSt2 bensì per il Tax4Fun2. In conclusione, in base alle analisi condotte, PICRUSt2 risulta essere migliore di PanFP e di Tax4Fun2 per 7 dei 9 dataset presi in considerazione sia per quel che riguarda le metriche binarie sia per quelle non binarie.

Bibliografia

[1] Manuale di gastroenterologia, edizione 2016-2019.

[2] https://ibba.cnr.it/blog/evoluzione-del-sequenziamento-dalle-origini-ai-giorni-nostri/.

[3] S. Lewis, A. Nash, Q. Li, and T. H. Ahn, 'Comparison of 16S and whole genome dog microbiomes using machine learning', *BioData Min*, vol. 14, no. 1, Dec. 2021, doi: 10.1186/s13040-021-00270-x.

[4] G. M. Douglas *et al.*, 'PICRUSt2 for prediction of metagenome functions', *Nature Biotechnology*, vol. 38, no. 6. Nature Research, pp. 685–688, Jun. 01, 2020. doi: 10.1038/s41587-020-0548-6.

[5] F. Wemheuer *et al.*, 'Tax4Fun2: Prediction of habitat-specific functional profiles and functional redundancy based on 16S rRNA gene sequences', *Environmental Microbiomes*, vol. 15, no. 1, May 2020, doi: 10.1186/s40793-020-00358-7.

[6] S. R. Jun, M. S. Robeson, L. J. Hauser, C. W. Schadt, and A. A. Gorin, 'PanFP: Pangenome-based functional profiles for microbial communities', *BMC Res Notes*, vol. 8, no. 1, Sep. 2015, doi: 10.1186/s13104-015-1462-8.

[7] C. Djemiel, P. A. Maron, S. Terrat, S. Dequiedt, A. Cottin, and L. Ranjard, 'Inferring microbiota functions from taxonomic genes: a review', *GigaScience*, vol. 11. Oxford University Press, 2022. doi: 10.1093/gigascience/giab090.

[8] D. S. Mongad, N. S. Chavan, N. P. Narwade, K. Dixit, Y. S. Shouche, and D. P. Dhotre, 'MicFunPred: A conserved approach to predict functional profiles from 16S rRNA gene sequence data', *Genomics*, vol. 113, no. 6, pp. 3635–3643, Nov. 2021, doi: 10.1016/j.ygeno.2021.08.016.

[9] O. Laroche, X. Pochon, S. A. Wood, and N. Keeley, 'Beyond taxonomy: Validating functional inference approaches in the context of fish-farm impact assessments', *Mol Ecol Resour*, vol. 21, no. 7, pp. 2264–2277, Oct. 2021, doi: 10.1111/1755-0998.13426.

[10] A. Arfken, B. Song, J. S. Bowman, and M. Piehler, 'Denitrification potential of the eastern oyster microbiome using a 16S rRNA gene based metabolic inference approach', *PLoS One*, vol. 12, no. 9, Sep. 2017, doi: 10.1371/journal.pone.0185071.

[11] N. R. Narayan *et al.*, 'Piphillin predicts metagenomic composition and dynamics from DADA2-corrected 16S rDNA sequences', *BMC Genomics*, vol. 21, no. 1, Jan. 2020, doi: 10.1186/s12864-019-6427-1.

[12] G. Jing, Y. Zhang, W. Cui, L. Liu, J. Xu, and X. Su, 'Meta-Apo improves accuracy of 16S-amplicon-based prediction of microbiome function', *BMC Genomics*, vol. 22, no. 1, Dec. 2021, doi: 10.1186/s12864-020-07307-1.

[13] S. Sun, R. B. Jones, and A. A. Fodor, 'Inference-based accuracy of metagenome prediction tools varies across sample types and functional categories', *Microbiome*, vol. 8, no. 1, Apr. 2020, doi: 10.1186/s40168-020-00815-y.

[14] X. Kang, D. M. Deng, W. Crielaard, and B. W. Brandt, 'Reprocessing 16S rRNA Gene Amplicon Sequencing Studies: (Meta)Data Issues, Robustness, and Reproducibility', *Front Cell Infect Microbiol*, vol. 11, Oct. 2021, doi: 10.3389/fcimb.2021.720637.

[15] A. P. Heikema *et al.*, 'Comparison of illumina versus nanopore 16s rRNA gene sequencing of the human nasal microbiota', *Genes (Basel)*, vol. 11, no. 9, pp. 1–17, Sep. 2020, doi: 10.3390/genes11091105.

[16] S. H. Ong *et al.*, 'Species Identification and Profiling of Complex Microbial Communities Using Shotgun Illumina Sequencing of 16S rRNA Amplicon Sequences', *PLoS One*, vol. 8, no. 4, Apr. 2013, doi: 10.1371/journal.pone.0060811.

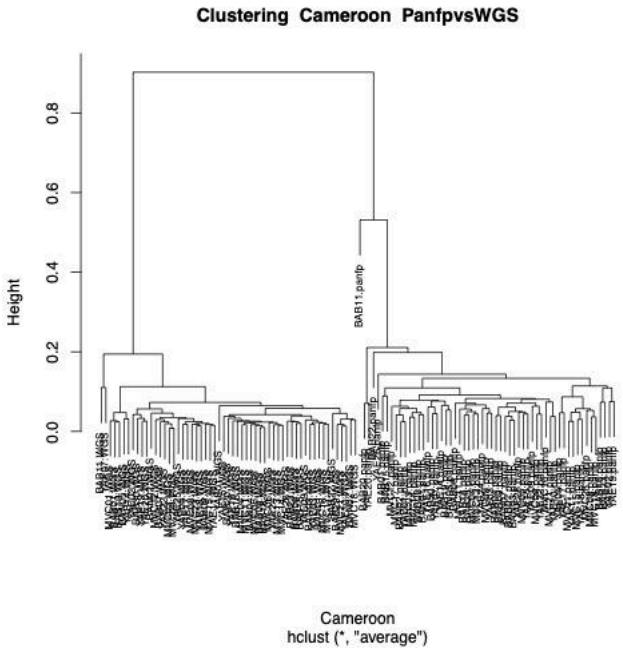
[17] R. Poretsky, L. M. Rodriguez-R, C. Luo, D. Tsementzi, and K. T. Konstantinidis, 'Strengths and limitations of 16S rRNA gene amplicon sequencing in revealing temporal microbial

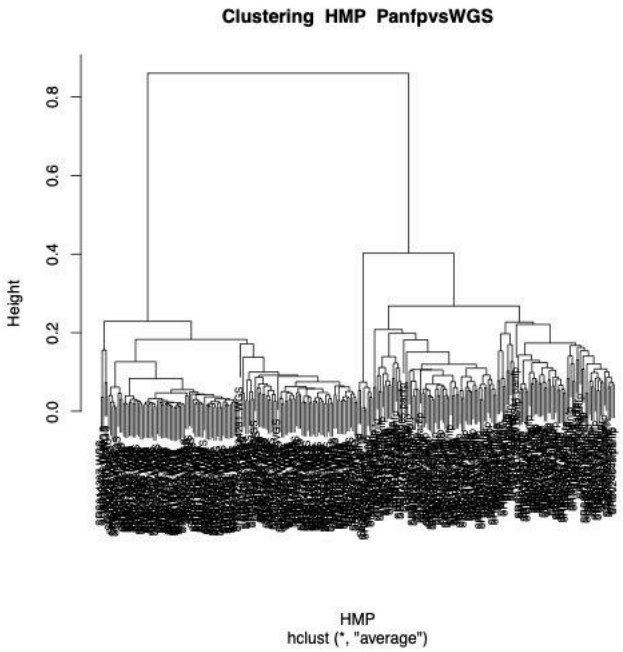
community dynamics', *PLoS One*, vol. 9, no. 4, Apr. 2014, doi: 10.1371/journal.pone.0093827.
[18] R. Ranjan, A. Rani, A. Metwally, H. S. McGee, and D. L. Perkins, 'Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing', *Biochem Biophys Res Commun*, vol. 469, no. 4, pp. 967–977, Jan. 2016, doi: 10.1016/j.bbrc.2015.12.083.
[19] J.-W. Chen *et al.*, 'Functional Proling of Saliva Microbiome is Essential for Oral Cancer Prediction', doi: 10.21203/rs.3.rs-45330/v1.
[20] G. M. Douglas *et al.*, 'PICRUSt2: An improved and extensible approach for metagenome inference', doi: 10.1101/672295.
[21] C. Hoskinson *et al.*, 'Composition and Functional Potential of the Human Mammary Microbiota Prior to and Following Breast Tumor Diagnosis', *mSystems*, vol. 7, no. 3, Jun. 2022, doi: 10.1128/msystems.01489-21.
[22] I. Laudadio, V. Fulci, F. Palone, L. Stronati, S. Cucchiara, and C. Carissimi, 'Quantitative Assessment of Shotgun Metagenomics and 16S rDNA Amplicon Sequencing in the Study of Human Gut Microbiome', *OMICS*, vol. 22, no. 4, pp. 248–254, Apr. 2018, doi: 10.1089/omi.2018.0013.
[23] S. Woloszynek, J. C. Mell, Z. Zhao, G. Simpson, M. P. O'Connor, and G. L. Rosen, 'Exploring thematic structure and predicted functionality of 16S rRNA amplicon data', *PLoS One*, vol. 14, no. 12, Dec. 2019, doi: 10.1371/journal.pone.0219235.
[24] C. Sansupa, S. F. M. Wahdan, S. Hossen, T. Disayathanoowat, T. Wubet, and W. Purahong, 'Can we use functional annotation of prokaryotic taxa (Faprotax) to assign the ecological functions of soil bacteria?', *Applied Sciences (Switzerland)*, vol. 11, no. 2, pp. 1–17, Jan. 2021, doi: 10.3390/app11020688.
[25] D. H. Huson *et al.*, 'MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data', *PLoS Comput Biol*, vol. 12, no. 6, Jun. 2016, doi: 10.1371/journal.pcbi.1004957.
[26] K. D. Brumfield, A. Huq, R. R. Colwell, J. L. Olds, and M. B. Leddy, 'Microbial resolution of whole genome shotgun and 16S amplicon metagenomic sequencing using publicly available NEON data', *PLoS One*, vol. 15, no. 2, Feb. 2020, doi: 10.1371/journal.pone.0228899.
[27] H. H. Caicedo, D. A. Hashimoto, J. C. Caicedo, A. Pentland, and G. P. Pisano, 'Overcoming barriers to early disease intervention', *Nature Biotechnology*, vol. 38, no. 6. Nature Research, pp. 669–673, Jun. 01, 2020. doi: 10.1038/s41587-020-0550-z.

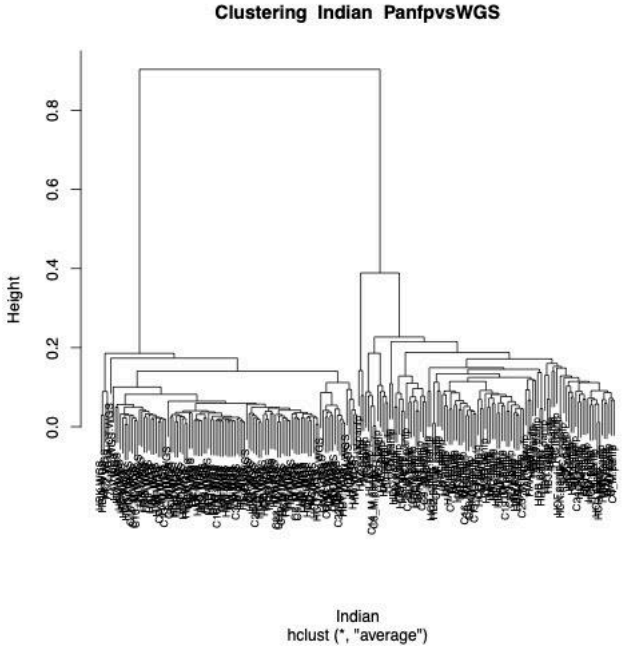
Materiale Supplementare (1)

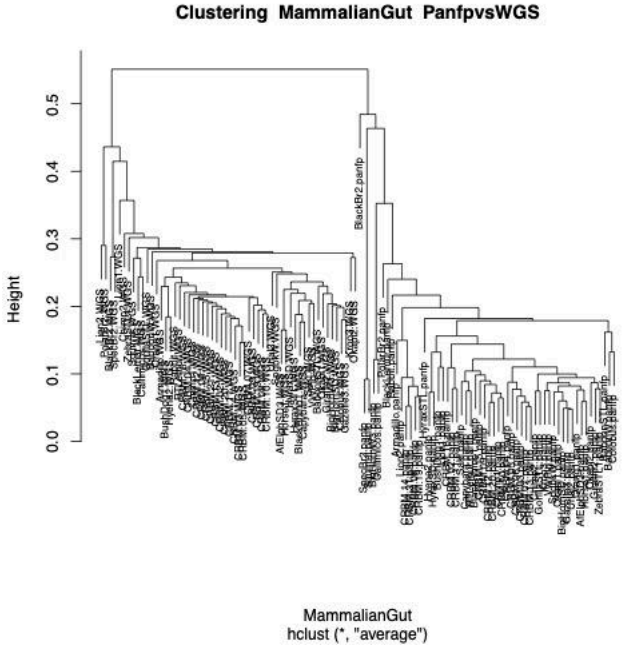
Caso non binario

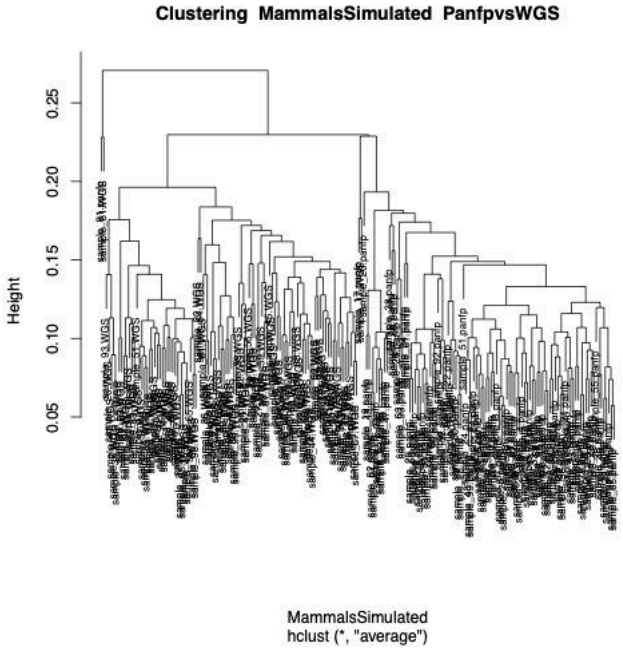
PanFPvsWGS

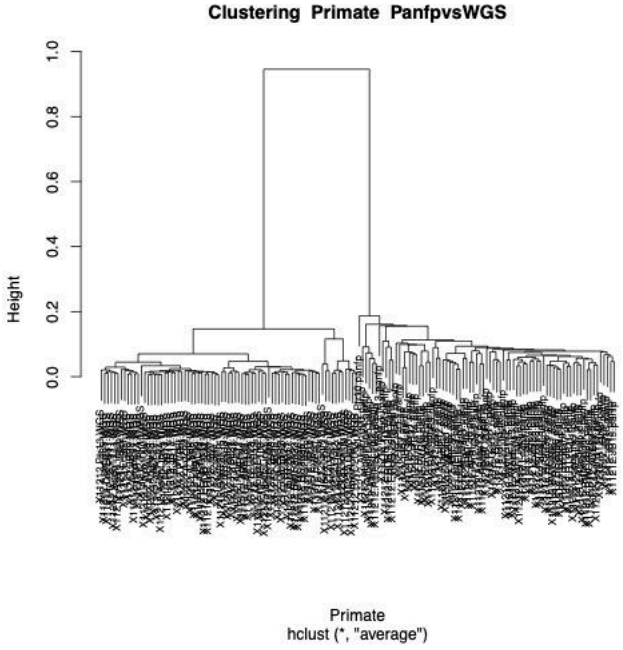




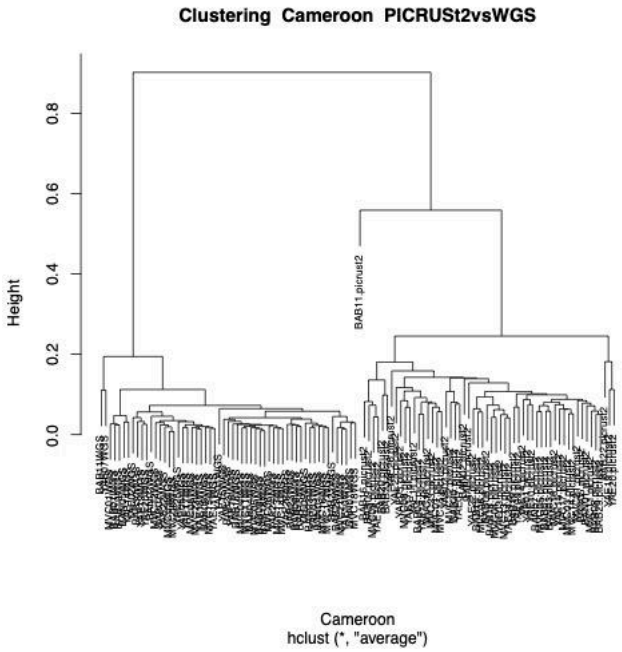

Clustering Indian PanfpvsWGS
Height
0.0
0.2
0.4
0.6
0.8
Indian
hclust (*, "average")

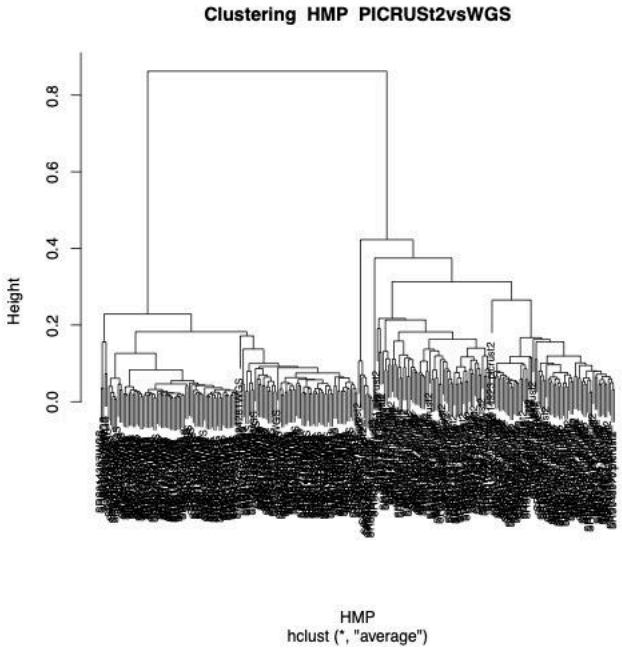

Clustering MammalianGut PanfpvsWGS
Height
0.0
0.1
0.2
0.3
0.4
0.5
MammalianGut
hclust (*, "average")

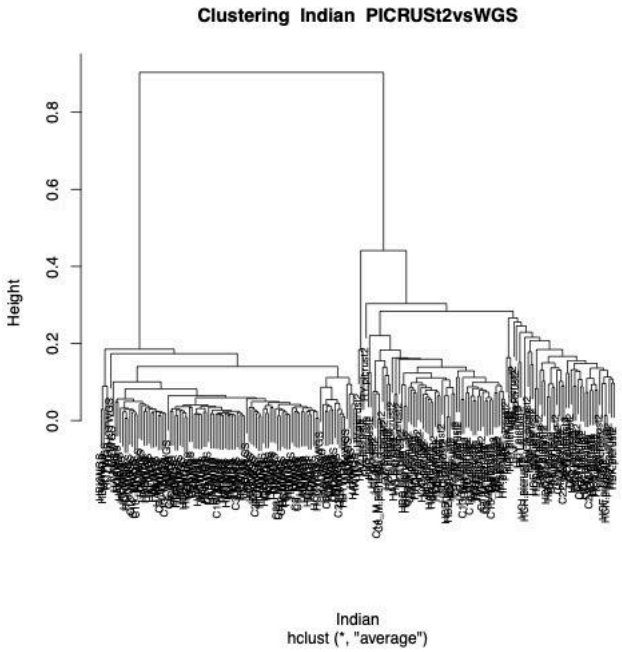

Clustering MammalsSimulated PanfpvsWGS
Height
0.05
0.10
0.15
0.20
0.25
MammalsSimulated
hclust (*, "average")

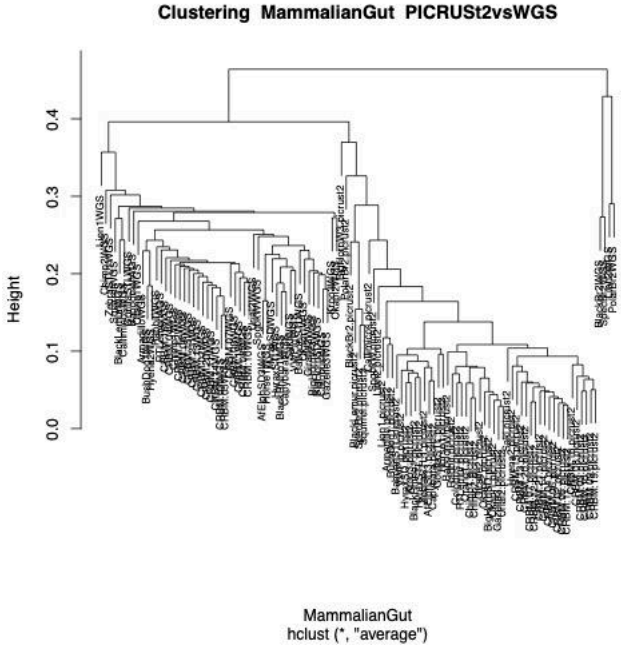

Clustering Primate PanfpvsWGS
Height
0.0
0.2
0.4
0.6
0.8
1.0
Primate
hclust (*, "average")

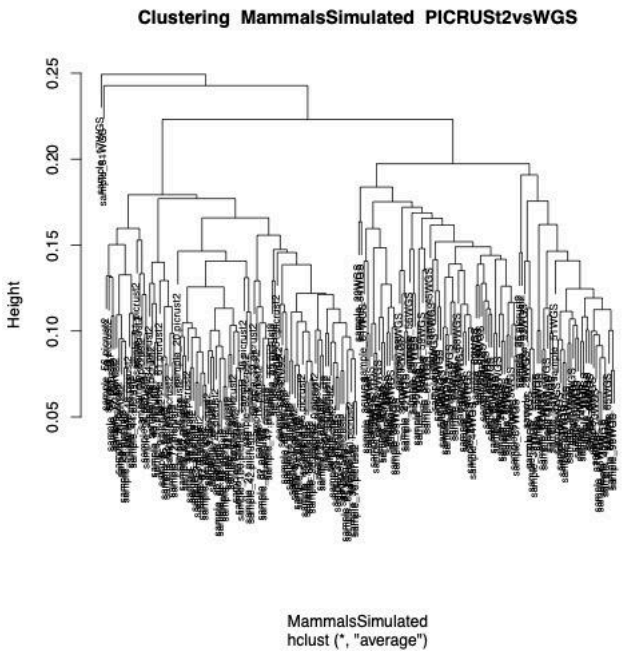
PICRUSt2vsWGS

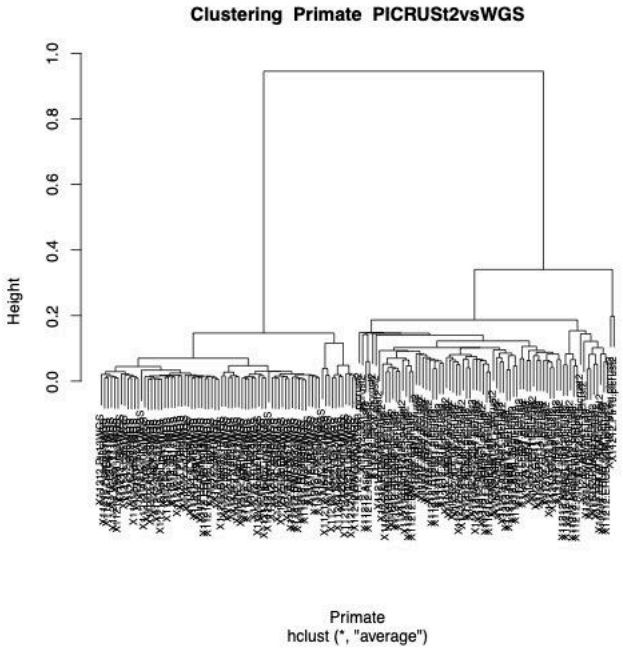




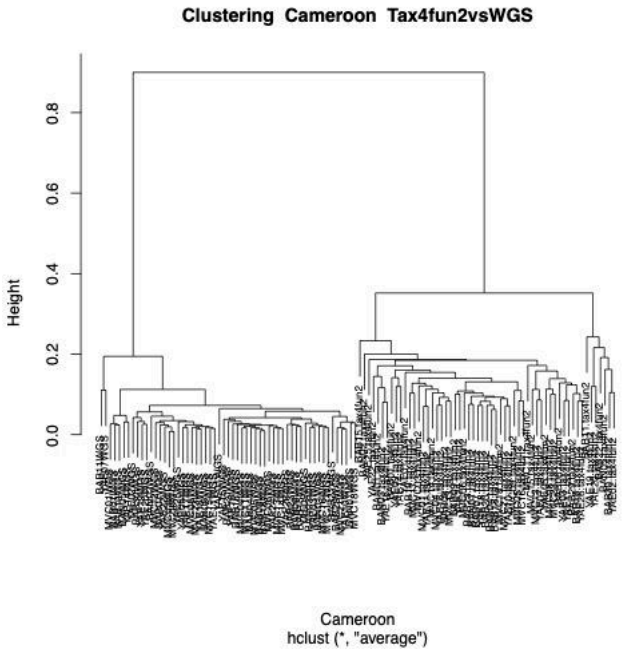

Clustering Indian PICRUSt2vsWGS
Height
0.0
0.2
0.4
0.6
0.8
Indian
hclust (*, "average")

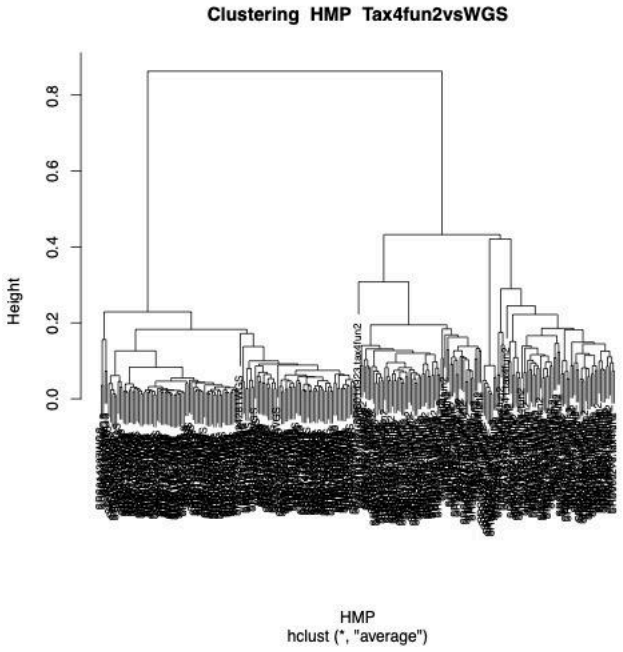

Clustering MammalianGut PICRUSt2vsWGS
Height
0.0
0.1
0.2
0.3
0.4
MammalianGut
hclust (*, "average")

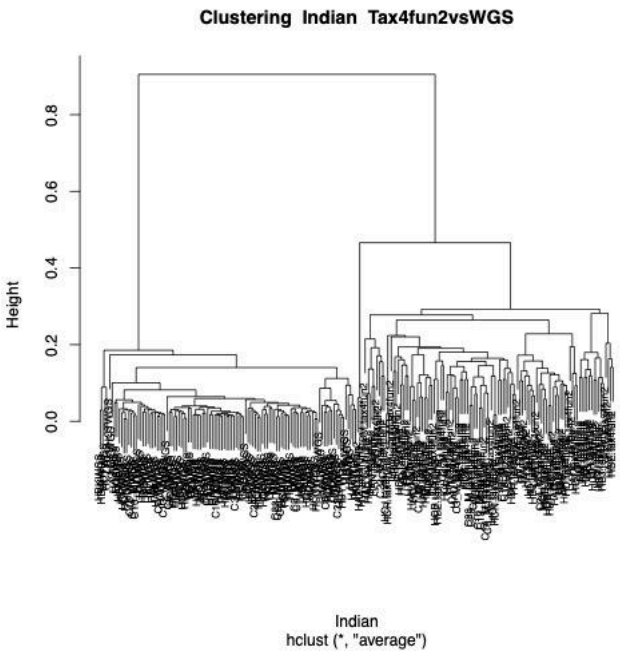

Clustering MammalsSimulated PICRUSt2vsWGS
Height
0.05
0.10
0.15
0.20
0.25
MammalsSimulated
hclust (*, "average")

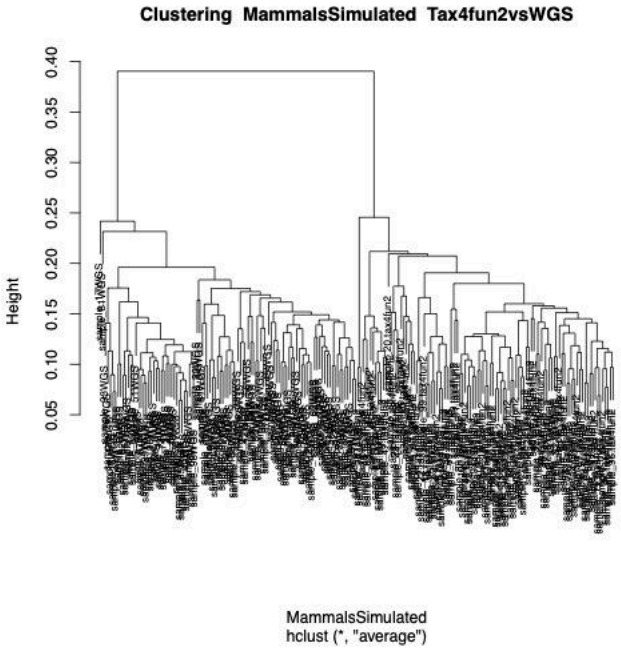

Clustering Primate PICRUSt2vsWGS
Height
0.0
0.2
0.4
0.6
0.8
1.0
Primate
hclust (*, "average")

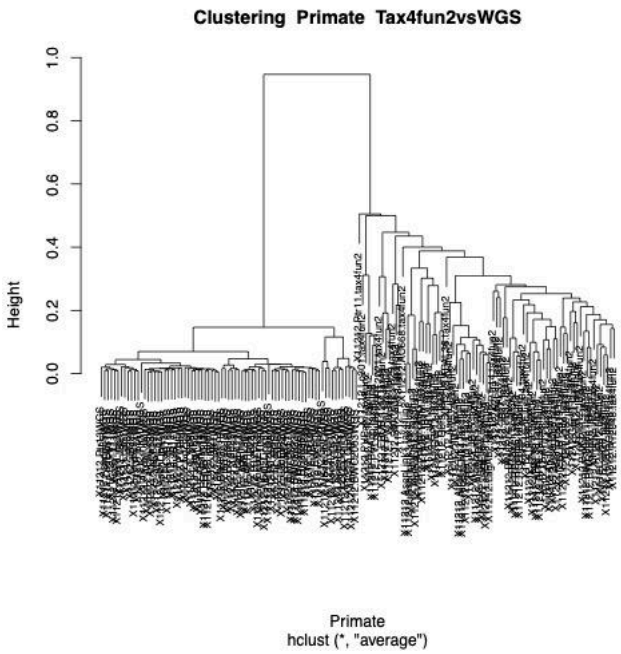
Tax4Fun2vsWGS





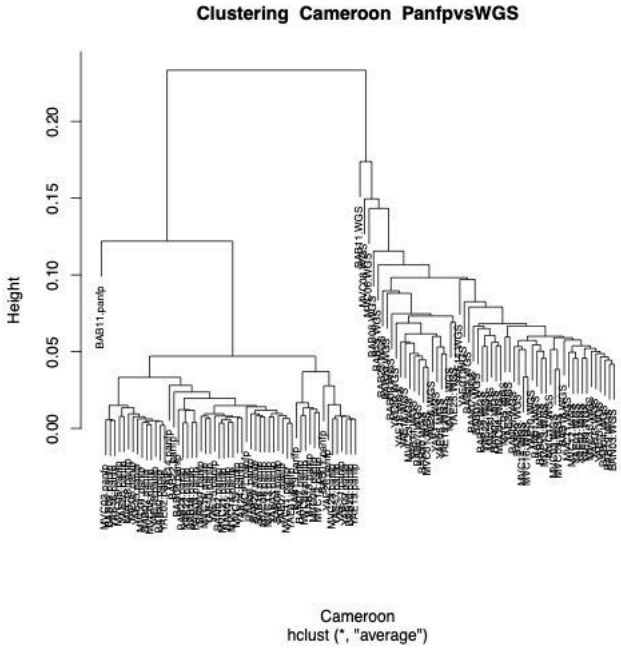

Clustering Indian Tax4fun2vsWGS
Height
0.0
0.2
0.4
0.6
0.8
Indian
hclust (*, "average")

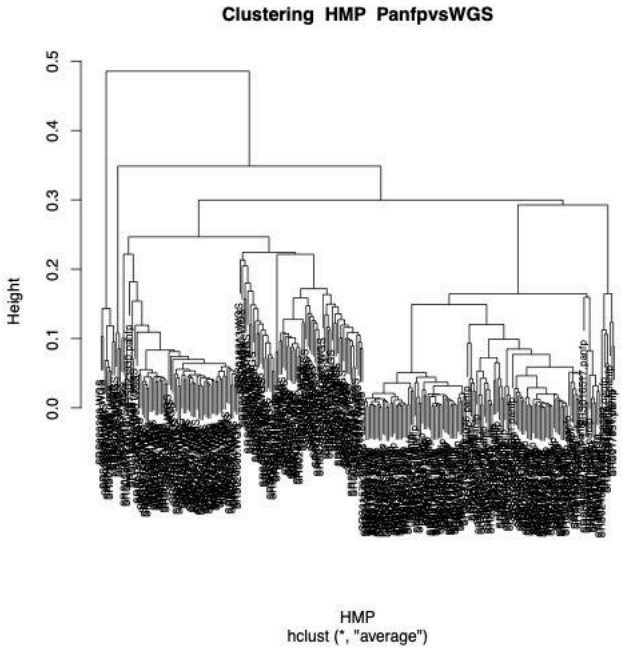

Clustering MammalsSimulated Tax4fun2vsWGS
Height
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
MammalsSimulated
hclust (*, "average")

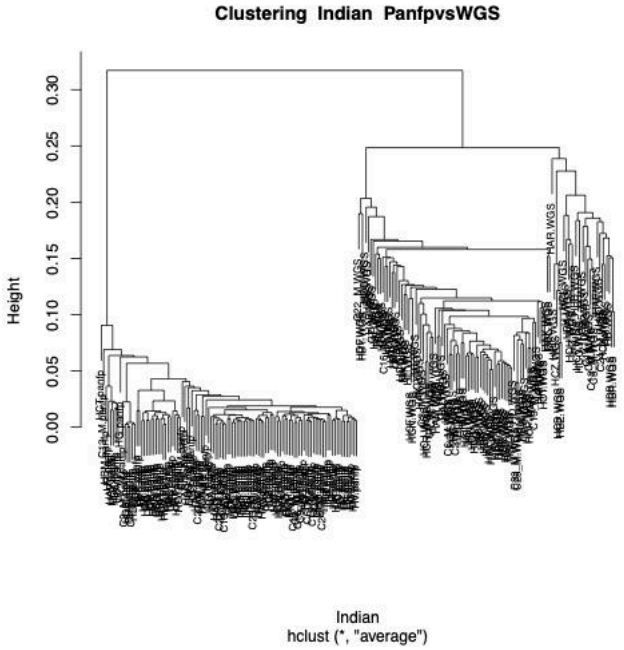

Clustering Primate Tax4fun2vsWGS
Height
0.0
0.2
0.4
0.6
0.8
1.0
Primate
hclust (*, "average")

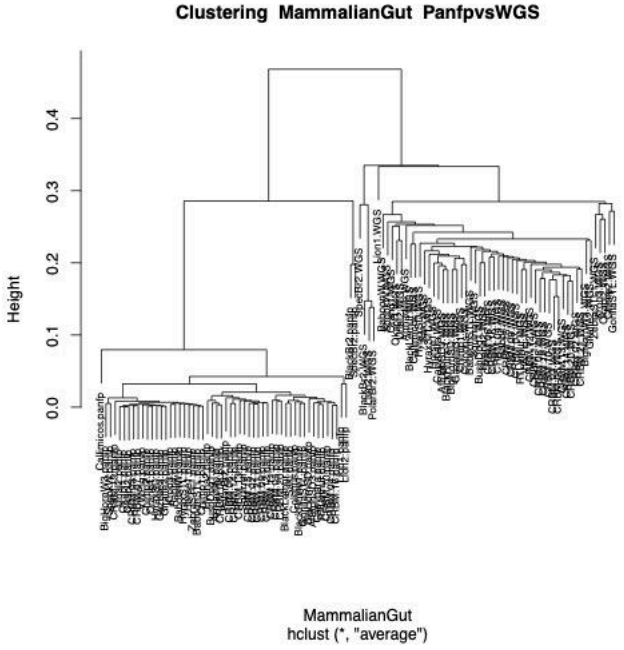
Caso binario

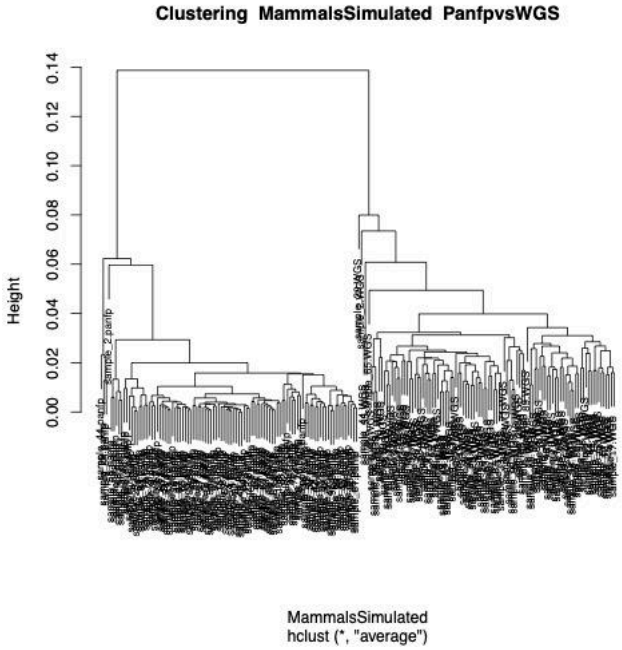
PanFPvsWGS

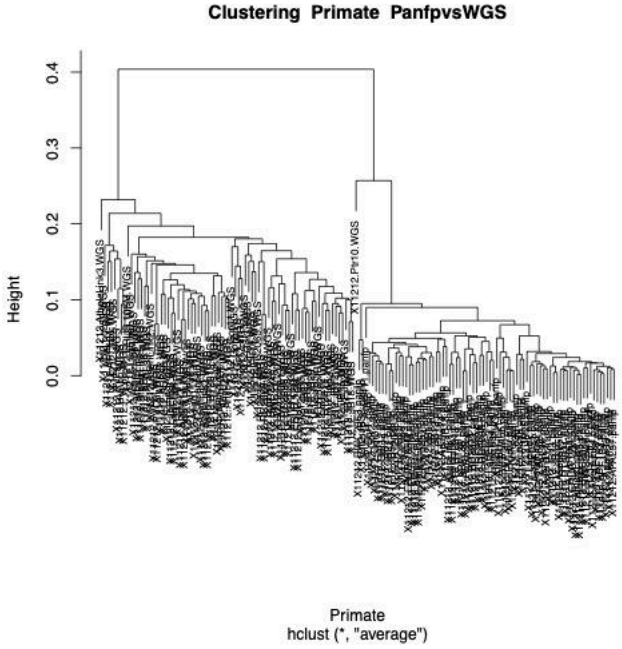




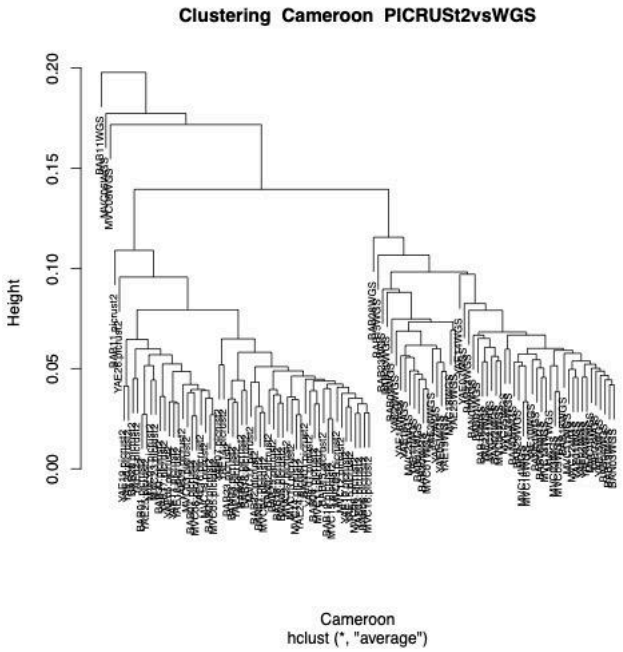

Clustering Indian PanfpvsWGS
Height
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Indian
hclust (*, "average")

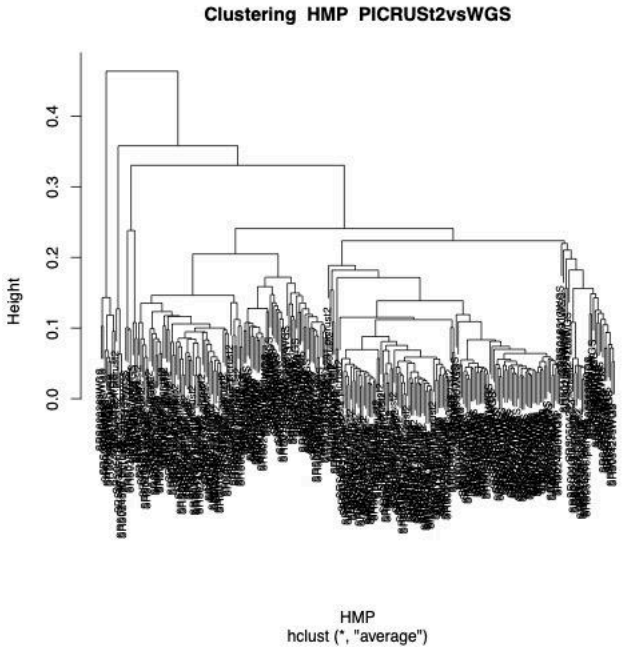

Clustering MammalianGut PanfpvsWGS
Height
0.0
0.1
0.2
0.3
0.4
MammalianGut
hclust (*, "average")

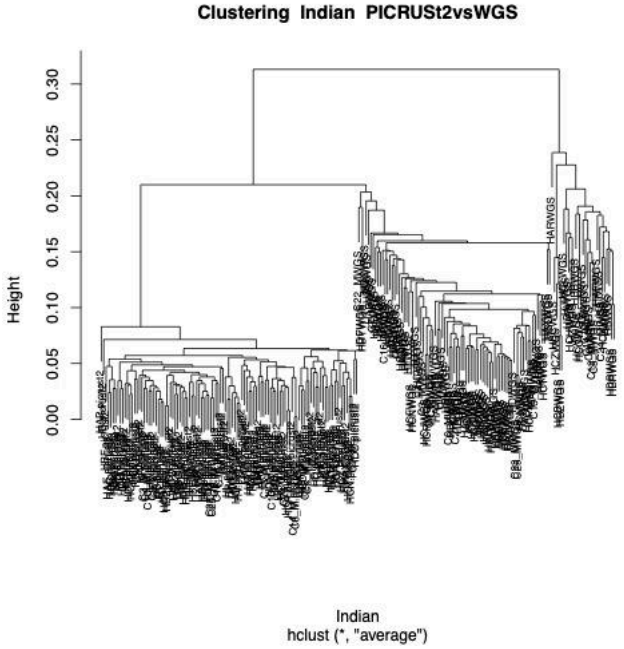

Clustering MammalsSimulated PanfpvsWGS
Height
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
MammalsSimulated
hclust (*, "average")

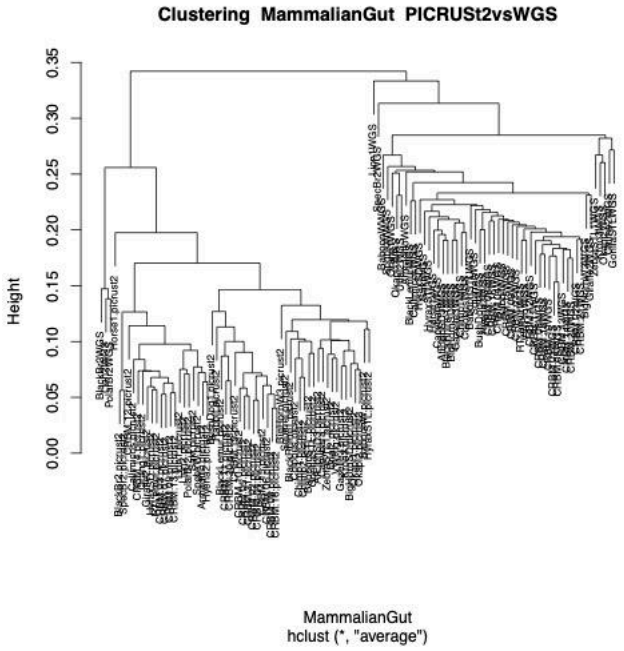

Clustering Primate PanfpvsWGS
Height
0.0
0.1
0.2
0.3
0.4
Primate
hclust (*, "average")

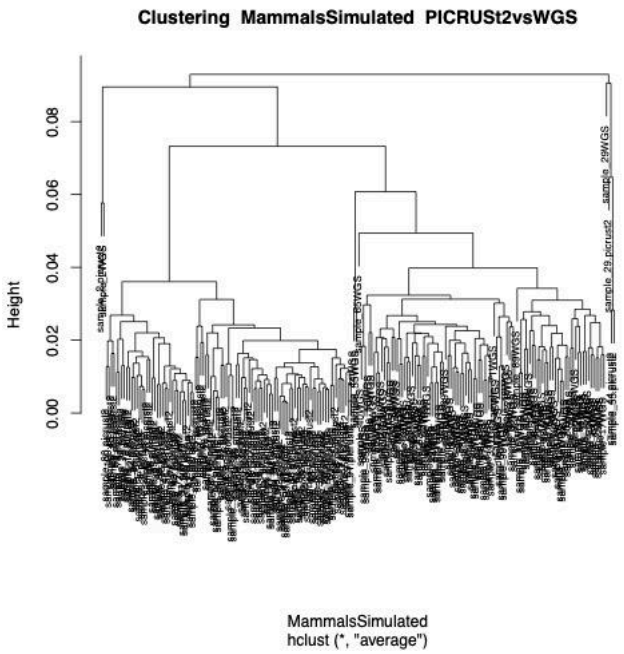
PICRUSt2vsWGS

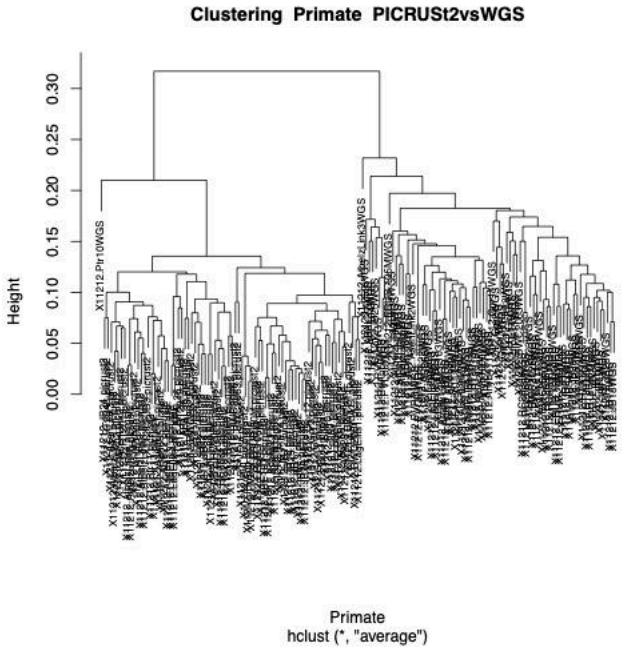




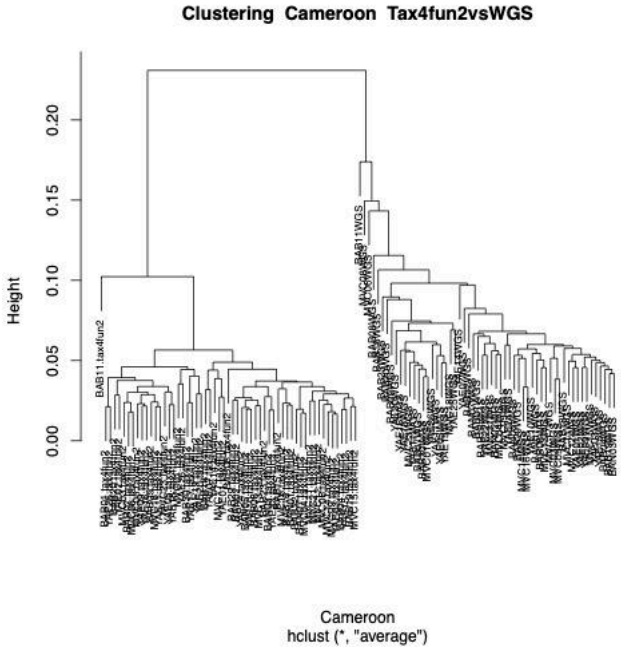

Clustering Indian PICRUSt2vsWGS
Height
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Indian
hclust (*, "average")

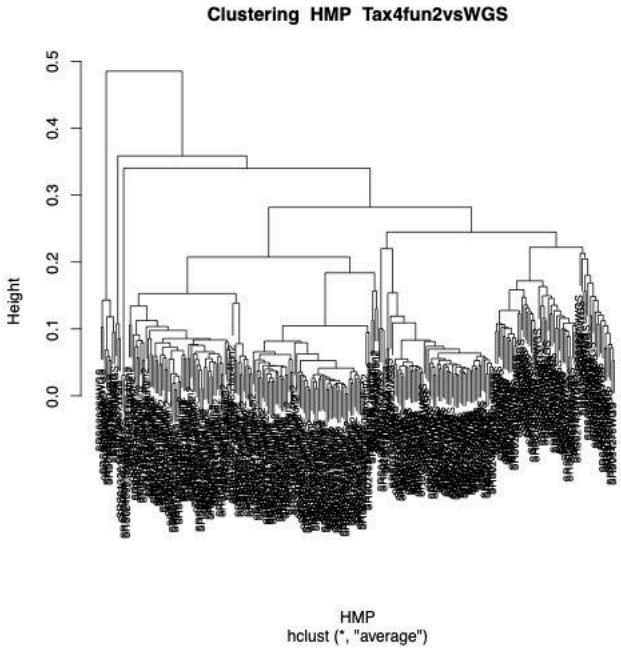

Clustering MammalianGut PICRUSt2vsWGS
Height
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
MammalianGut
hclust (*, "average")

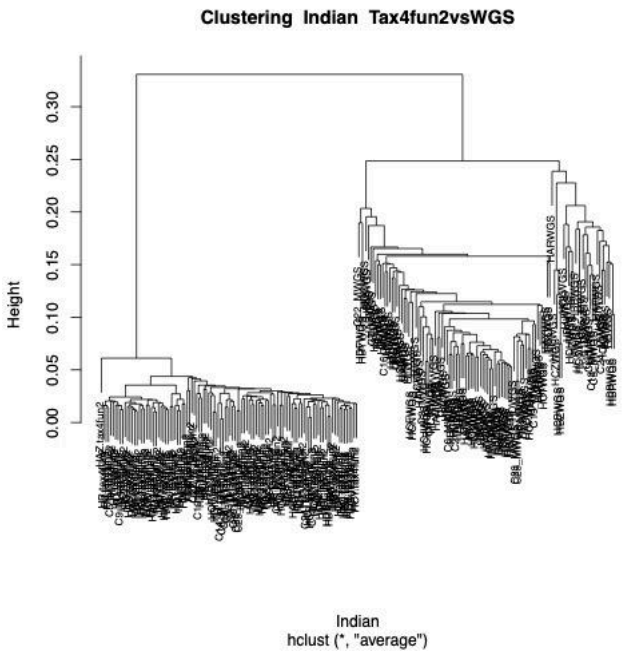

Clustering MammalsSimulated PICRUSt2vsWGS
Height
0.00
0.02
0.04
0.06
0.08
MammalsSimulated
hclust (*, "average")

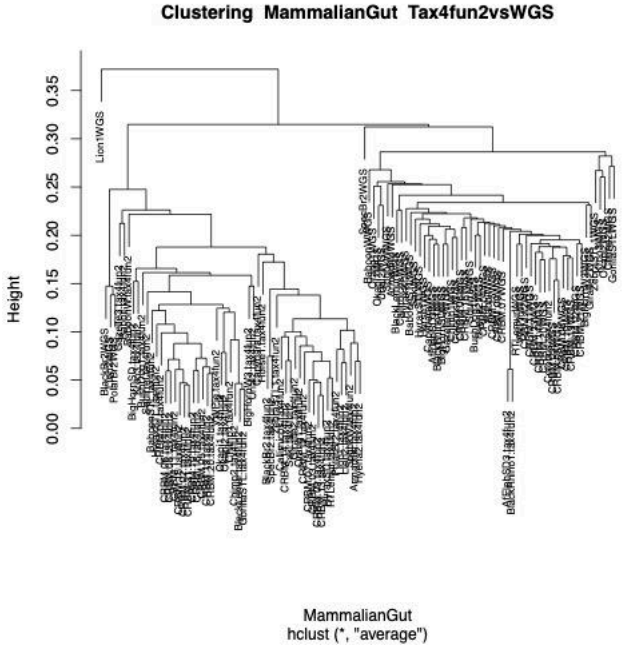

Clustering Primate PICRUSt2vsWGS
Height
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Primate
hclust (*, "average")

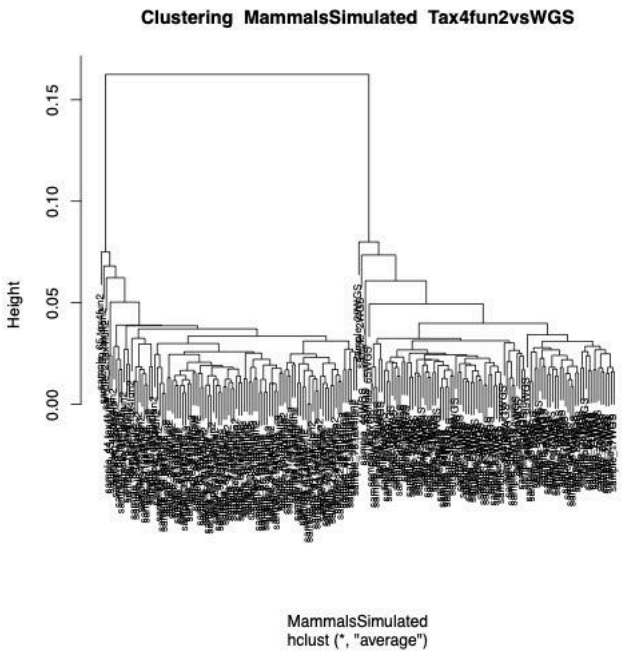
Tax4Fun2vsWGS

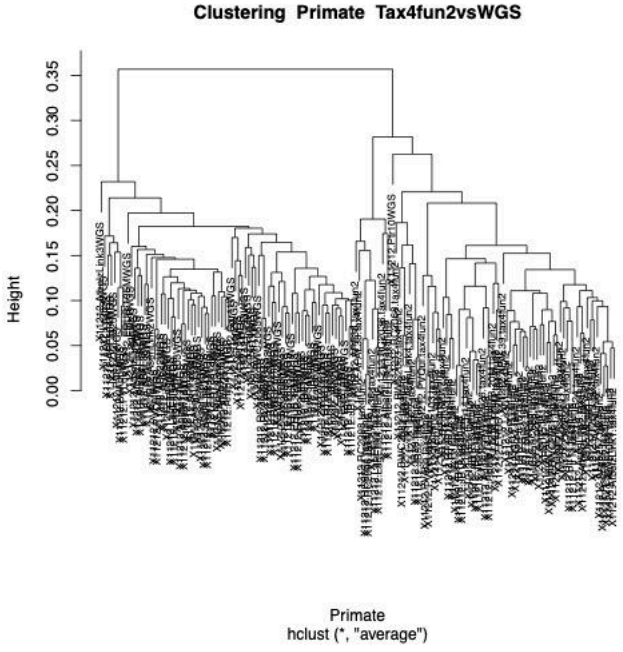





Clustering Indian Tax4fun2vsWGS
Height
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Indian
hclust (*, "average")


Clustering MammalianGut Tax4fun2vsWGS
Height
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
Lion1WGS
MammalianGut
hclust (*, "average")


Clustering MammalsSimulated Tax4fun2vsWGS
Height
0.00
0.05
0.10
0.15
MammalsSimulated
hclust (*, "average")


Clustering Primate Tax4fun2vsWGS
Height
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
Primate
hclust (*, "average")