UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Dipartimento di Ingegneria dell'Informazione

LAUREA MAGISTRALE IN BIOINGEGNERIA

A.A. 2012/2013

METODI IN METAGENOMICA PER L'ANALISI DEL MICROBIOMA: APPLICAZIONE A PAZIENTI AFFETTI DA BRONCOPNEUMOPATIA CRONICO OSTRUTTIVA E DA CANCRO AL COLON

Relatore: Prof.ssa Barbara Di Camillo

Correlatore: Dr. Francesca Finotello

Prof.ssa Luisa Barzon

Laureando: Alessandro Zandonà

Ringraziamenti

Ringrazio i miei genitori, Antonio e Rossella, poiché se ho raggiunto questo importante obiettivo è principalmente grazie a loro. Li ringrazio per avermi incoraggiato e sostenuto nelle mie scelte, per avermi permesso di studiare e per essermi stati sempre vicini. Si dice che nella vita ci vuole sempre un po' di fortuna: io posso dirmi fortunato ad avere dei genitori come loro. GRAZIE DI CUORE.

Un ringraziamento speciale lo dedico alla mia Tania, che con la sua dolcezza non mi ha mai abbandonato, sopportando i miei nervosismi, riuscendo a tranquillizzarmi e mettendoci l'anima per poter fare tutto il possibile per aiutarmi e sostenermi. La ringrazio per aver sempre creduto in me e per avermi dato così l'ennesima conferma che è lei la donna della mia vita. GRAZIE AMORE MIO.

Vorrei ringraziare i miei fratelli, Gianmarco e Giulia, per la pazienza portata nei miei confronti ed i momenti di sane litigate fraterne, così come un grazie lo rivolgo alla nonna Annamaria, che da 25 anni mi sta vicino e che per me è un punto di riferimento. Un altro ringraziamento lo dedico invece allo zio Franco, il mio "fratello" maggiore, per i consigli e le serate-Champions ed un ringraziamento alla zia Giovanna per la pazienza e l'ospitalità. Grazie anche a zia Amelia e a mia cugina Jessica che dal Piemonte sono venute fino a qui per essere presenti in questo momento. Grazie a tutti i miei parenti, per la splendida accoglienza ricevuta ad ogni mia visita.

Ringrazio i miei amici-colleghi, tra tutti Marco, Matteo, Paolo, Giacomo e Stefano, con cui ho condiviso l'intero percorso di studi, i dubbi dell'ultimo minuto prima degli esami, le pause pranzo ed i tanti momenti di divertimento.

Come non ringraziare i miei amici, sui quali da anni posso contare e che so che ci saranno sempre, che sia per una partita di calcio, per un'abbuffata o per una chiacchierata.

Desidero ringraziare la Professoressa Barbara Di Camillo per aver accettato l'incarico di relatrice per la mia tesi, e ringrazio la Dottoressa Francesca Finotello correlatrice dell'università, per la disponibilità durante la stesura del lavoro. Un sentito ringraziamento anche alla Prof.ssa Luisa Barzon, mio seconda correlatrice.

Sommario

Lo studio del microbioma umano, patrimonio genetico delle comunità batteriche presenti all'interno del corpo umano, è da sempre considerato un compito difficile. La complessa struttura in cui sono organizzate le comunità microbiche (definite anche microbiota) rappresenta infatti un ostacolo alla tradizionale coltura in vitro, ed il sequenziamento del microbioma risulta problematico a causa dell'enorme mole di dati da gestire. Ma con lo sviluppo delle recenti tecniche di sequenziamento *high-throughput*, lo studio del microbioma ha registrato progressi notevoli. È così emerso che il microbioma riveste un ruolo centrale, ma ancora ben da definirsi, nello stato di salute dell'uomo, nel suo metabolismo e nell'interazione con i farmaci.

In questo lavoro di tesi si è implementata una pipeline per analizzare il microbiota al fine di evidenziare una relazione tra lo stesso e lo stato di salute dell'ospite. Le analisi perciò sono state effettuate sul microbiota di soggetti con stati di salute differente, in particolare si sono prese in considerazione due patologie: la broncopneumopatia cronica ostruttiva (BPCO) ed il cancro al colon (CRC). La pipeline elaborata prevede in primis il calcolo di tre indici per la quantificazione dell'abbondanza e della distribuzione di generi e OTU in ogni soggetto. Si procede poi con il calcolo di altri tre indici che quantificano le differenze tra microbiota di soggetti con lo stesso stato di salute; in seguito si è ricercato se il microbiota varia in modo statisticamente significativo tra soggetti sani e malati, applicando NPMANOVA e ANOSIM. Infine, si sono individuate le specie che caratterizzano le diverse patologie, mediante l'utilizzo di test di Wilcoxon.

L'analisi mostra che, sia nel caso di studio sulla BPCO sia in quello relativo al CRC, la composizione del microbiota varia in maniera statisticamente significativa tra i soggetti sani e quelli malati. Si noti che ogni step dell'analisi è stato ripetuto più di una volta, in modo da valutare se i risultati fossero robusti. Si sono calcolati infatti tre indici sia per la diversità alfa che per la diversità beta, così come i risultati di NPMANOVA si sono confrontati con quelli di ANOSIM. Non sono mai state registrate discrepanze, il che suggerisce che i risultati possono considerarsi robusti.

Concludendo, l'analisi del microbiota può interpretarsi come un valido contributo sia all'elaborazione di nuovi trattamenti per BPCO e CRC sia come un possibile strumento diagnostico. Grazie alla consultazione delle liste di batteri ottenute con Wilcoxon si può infatti pensare di modificare la composizione del microbiota per ristabilire la situazione tipica dei soggetti sani. Inoltre si è dimostrato che la rilevazione di alterazioni microbiotiche è associata ad uno stato patologico o ad una sua prossima comparsa, e questo può quindi essere la base su cui fondare una diagnosi.

Abstract

The characterization of the human microbiome, defined as the genome of microbial communities living in human body, has been considered a difficult task for a long time. Indeed, it is not always feasible to culture a bacterial species in vitro and, most of all, to capture the complex interactions characterizing bacterial communities. But with the development of the recent *high-throughput* sequencing technologies, the study of human microbiome has gained a significant improvement. It has been demonstrated that the microbiome has an important role in the human health, metabolism and interactions with drugs.

In this thesis, we have implemented a pipeline for characterizing the human microbiota, in order to study the relationship between microbial communities and human health. We have analyzed the microbiota of subjects belonging to different groups, depending on the pathological condition and other covariates; in particular, we have considered two pathologies: chronic obstructive pulmonary disease (COPD) and the colon-rectal cancer (CRC).
Firstly, we have computed three indices to quantify genera and species abundances and their distribution in each subject. In addition, we have computed three other indices in order to quantify the differences between the microbiota of subjects belonging to the same group. We then have tested for statistically significant alterations of microbiota between healthy subjects and COPD or CRC patients, using two statistical methods: NPMANOVA and ANOSIM. Finally, we have investigated which microbial species characterize the different pathologies, using the Wilcoxon test.
The results of our analysis show that, in both BPCO and CRC case studies, the composition of microbiota undergoes statistically significant alterations when pathology occurs.
In conclusion, the pipeline that we implemented for the study of the microbiota can be used to manage and analyze the huge amount of data produced by the *high-throughput* sequencing technologies, and it can be seen as a useful tool for the diagnosis and the treatment of BPCO and CRC. Indeed, the assessment of particular alterations of the microbiota in a specific pathological condition poses the basis for the interpretation of disease mechanisms and for the development of novel treatments.

Indice

1. Introduzione

Il DNA è il componente fondamentale del patrimonio genetico di un organismo, definito genoma, ed è un polimero organico formato da monomeri definiti nucleotidi, disposti lungo due catene. Essi sono costituiti da un gruppo fosfato, una base azotata e lo zucchero deossiribosio. Esistono quattro tipi di nucleotidi, distinti a seconda del tipo di base azotata contenuta: adenina, guanina, timina e citosina. La sequenza delle basi nelle catene nucleotidiche codifica per le informazioni genetiche necessarie alla sintesi di RNA e proteine, molecole indispensabili per il corretto funzionamento della maggior parte degli organismi viventi. Vista l'importanza della sequenza in cui si dispongono i nucleotidi, sono state sviluppate diverse tecniche che consentono di determinare l'ordine delle basi nucleotidiche che compongono il DNA; si parla di tecniche di sequenziamento. Inizialmente esse consentivano di leggere una sequenza alla volta, ma, grazie ad una successiva automazione dei processi, si è reso possibile ricostruire centinaia di sequenze contemporaneamente. Tuttavia si sono recentemente sviluppate tecnologie high-throughput, che consentono di produrre enormi quantità di sequenze ad un costo minore e ad una velocità superiore rispetto alle tecniche precedenti. Si possono infatti ottenere fino a 20 milioni di basi in contemporanea, rendendo tali tecnologie adatte a sequenziare genomi sempre più grandi.
La recente diffusione di queste ultime tecnologie ha apportato notevoli sviluppi nello studio del genoma umano e batterico. Si è così potuto evidenziare l'importanza del *microbiota*, cioè l'insieme delle comunità batteriche che risiedono nel corpo umano, e del *microbioma*, cioè il materiale genetico batterico; diversi studi pongono infatti in risalto una forte relazione tra il microbioma e la fisiologia umana, e quindi il metabolismo, l'interazione con i farmaci e numerose patologie. Al fine di caratterizzare composizione e funzionalità del microbioma e chiarire il suo ruolo nella salute umana, sono nati molteplici studi tra i quali i principali sono lo *Human Microbiome Project (HMP)* ed il *Metagenomics of the Human Intestinal Tract (MetaHIT)*. Di pari passo con i due progetti appena citati, altri studi analizzano il microbioma dei diversi siti corporei e di soggetti affetti da diverse patologie, osservando così il comportamento del microbioma in condizioni di salute differenti.

In questo lavoro di tesi ci si pone l'obiettivo di implementare una pipeline di analisi del microbiota per indagare la relazione tra quest'ultimo e lo stato di salute umana. Si noti che ci si concentra sullo studio del microbiota anziché del microbioma, in quanto il nostro interesse è rivolto a caratterizzare le comunità batteriche in esame in termine di specie e non di genoma. Questa scelta ha inoltre evidenti implicazioni sui costi, in quanto il sequenziamento del genoma di tutti i microrganismi presenti in un sito o tessuto implica costi più elevati di quelli sostenuti

per l'analisi del microbiota; in tal caso infatti basta sequenziare un solo gene (il 16S del DNA ribosomiale) che, essendo specifico delle specie batteriche, ne consente la discriminazione.

In particolare, il primo aspetto affrontato dalla nostra analisi è la quantificazione della diversità delle specie intra-soggetto, in modo da caratterizzare il microbiota e valutarne la variabilità all'interno dei soggetti, indipendentemente dal loro stato di salute.
In secondo luogo, si calcola la variabilità inter-soggetto del microbiota, quantificando le differenze tra le comunità batteriche di soggetti diversi, senza ancora considerare rilevante lo stato di salute al fine dell'analisi.
Infine si quantificano le differenze tra il microbiota di soggetti diversi, con stati di salute diversi, al fine di valutare se la presenza di uno stato patologico influenzi o meno composizione e struttura delle comunità batteriche.
Operativamente, si sono quindi ricercati in letteratura i metodi atti ad analizzare la composizione del microbiota e la distribuzione delle specie al suo interno, e le tecniche che consentono di evidenziare le differenze tra microbiota di soggetti o siti distinti. Successivamente si è implementata la pipeline, la quale viene poi applicata in due diversi casi di studio: nel primo si considerano soggetti affetti da bronco pneumopatia cronica ostruttiva (BPCO), nel secondo soggetti malati di cancro al colon (CRC). Ci si concentra su queste due specifiche patologie visto il loro elevato tasso di incidenza su scala mondiale e la loro gravità. La BPCO è una malattia polmonare progressiva e potenzialmente mortale che secondo l'Organizzazione Mondiale della Sanità diventerà la terza causa di morte nel mondo entro il 2030; per quanto concerne il cancro al colon, in termini di incidenza annuale, è la terza forma tumorale più diffusa al mondo (OMS).
In particolare, la pipeline implementata si applica alle abbondanze relative dei generi batterici presenti nei polmoni di soggetti sani e affetti da BPCO, e le abbondanze dei generi che compongono il microbiota del colon di soggetti sani e di soggetti con CRC.
Nello specifico, il primo passo dell'analisi è il calcolo di tre indici che forniscono un'indicazione quantitativa dell'abbondanza delle specie ed il grado della loro ripartizione all'interno di ogni soggetto (capitolo 4.4). In seguito si calcolano altri tre indici che quantificano le differenze nella composizione del microbiota di soggetti diversi ma con il medesimo stato di salute, definiti indici di diversità beta (capitolo 4.5); dopodiché si ricorre a due test statistici, NPMANOVA (capitolo 4.6) e ANOSIM (capitolo 4.7), per calcolare le differenze statisticamente significative tra la composizione del microbiota di soggetti sani e quello di soggetti malati. Infine si utilizza il test di Wilcoxon per evidenziare quali sono le specie più abbondanti nei diversi stati di salute.

2. Il microbioma ed il microbiota

Il corpo umano offre alle comunità microbiche una molteplicità di siti da colonizzare, basti pensare che al suo interno è stata calcolata in media una quantità di batteri circa dieci volte superiore al numero delle cellule del corpo umano (Savage, 1977) . Grazie alle analisi basate sul DNA ed ai progressi della bioinformatica, è stato possibile caratterizzare tali colonie batteriche. La comunità scientifica ha concordato nel definire *microbiota* per identificare la totalità di organismi microbici presenti in un particolare ambiente, mentre col termine *microbioma* ci si riferisce all'informazione genetica insita nel microbiota stesso.

Nel diciannovesimo secolo, la 'normal flora' poteva essere studiata ricorrendo alla coltivazione in vitro, ma restava difficile riprodurre lo specifico microambiente in cui poter isolare le specie microbiche. In seguito, lo sviluppo di nuovi metodi di analisi ha permesso l'avviarsi di molti progetti di ricerca, i quali con lo strumento del microbioma, mirano a non solo a raccogliere informazioni tassonomiche e funzionali, bensì anche a comprendere l'interazione tra le comunità batteriche ed il corpo umano, la loro influenza sul sistema digestivo, metabolico, sullo sviluppo e sulla fisiologia umana. A tal proposito lo Human Microbiome Project (HMP), ed il corrispondente progetto europeo (MetaHIT), si pone come obiettivo quello di studiare le popolazioni microbiche presenti in differenti siti del corpo umano ed analizzare la possibile relazione tra comunità batteriche e salute umana. Lo HMP è in realtà uno sforzo multidisciplinare ed internazionale, il quale consta quindi di diversi progetti condotti parallelamente in tutto il mondo.

Oltre a quelli in precedenza esposti, lo HMP si prefigge anche i seguenti propositi:

1) Il miglioramento delle tecnologie per isolamento ed analisi degli organismi batterici; a causa del numero ridotto di specie microbiche che possono coltivarsi, non c'è una grande disponibilità di sequenze di DNA su cui basare lo studio, e quindi con lo HMP si propongono nuovi metodi atti ad isolare e coltivare un maggior numero di specie microbiche;
2) Lo sviluppo di nuovi strumenti per l'analisi computazionale, le nuove tecnologie per il sequenziamento, che consentono di esaminare il genoma delle comunità batteriche, producono data set numerosi e complessi, richiedono sempre nuovi strumenti di analisi;
3) Lo sviluppo di un set di sequenze di genoma microbico di riferimento; il fine di tale iniziativa è di sequenziare batteri provenienti o meno da coltivazioni in vitro così come micro organismi non batterici, in modo da ottenere più di 1000 genomi che fungano da riferimento.

4) L'istituzione di un centro per il coordinamento e l'analisi dei dati, il quale gestisca i dati processati e non, coordini le analisi e stabilisca un portale attraverso il quale si possa dare una visibilità internazionale ai progetti e supportare le relazioni internazionali.
5) La creazione di ambienti nelle quali i microbi possano essere coltivati ed il DNA possa essere processato. Inoltre, grazie a tali strutture, le risorse possono essere ampiamente disponibili e accessibili alla comunità scientifica.
6) La valutazione delle implicazioni legali, sociali ed etiche del progetto; infatti deve essere rispettata la privacy dei donatori del microbioma, e devono essere considerati gli aspetti legati al bioterrorismo e i possibili usi forensi dei profili genetici ricavati dal microbioma stesso.

Con la nascita dello Human Microbiome Project da parte dell'NIH americano e del MetaHIT europeo, si è dunque concretizzato il crescente interesse nei confronti del microbioma umano. In occasione dell'HMP sono stati sequenziati 690 campioni prelevati da 15 siti corporei diversi di 300 soggetti, ottenendo 2.3 terabyte di dati relativi all'RNA ribosomiale 16S. Come noto, esso è un componente essenziale della piccola unità dei ribosomi contenente una sequenza specifica per ogni specie batterica ed è quindi usato per l'analisi della composizione di comunità microbiche. L'rRNA 16S viene infatti sequenziato tramite piattaforme NGS e le sequenze simili vengono raggruppate in OTU (*Operational Taxonomic Unit*); esse rappresentano un modo per distinguere le specie e classificare le sequenze nucleotidiche in diversi livelli tassonomici. L'abbondanza delle diverse OTU viene poi stimata sulla base del numero di sequenze corrispondenti.

Le ricerche condotte in questo ambito si basano sul perseguimento di differenti obiettivi, dallo studio della composizione e dalle proprietà funzionali del microbioma alle sue complesse dinamiche e all'interazione dello stesso con l'organismo che lo ospita. In particolare si indaga il ruolo che il microbioma riveste nello stato di salute umana, e a tal proposito si ricercano metodi e strategie atte a manipolare composizione e funzionalità del microbiota, al fine di ottenere benefici sulla salute stessa.

2.1. Il microbiota nei siti anatomici

Una delle prime considerazioni sul microbioma emerse dai progetti sopra citati è la grande *variabilità* nella sua composizione sia tra soggetti diversi, sia all'interno dello stesso soggetto, in siti anatomici e tessuti differenti (Figura 1).

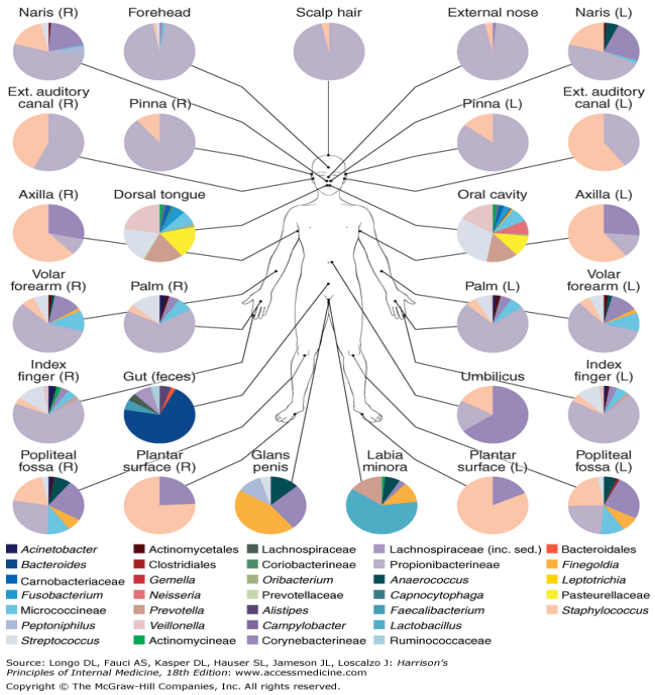


Figura 1: *Differenze del microbioma in base ai siti anatomici*

Si è comunque individuato un insieme di geni comune alla maggior parte dei soggetti, il quale è denominato "*microbiome core*" (Figura 2) ; a partire da questo set di geni, il microbioma si differenzia a causa di diversi fattori: lo stato di salute dell'ospite, la sua dieta, l'ambiente in cui esso vive, il suo genotipo e l'eventuale contatto con altre colonie batteriche (Turnbaugh et al., 2007) .

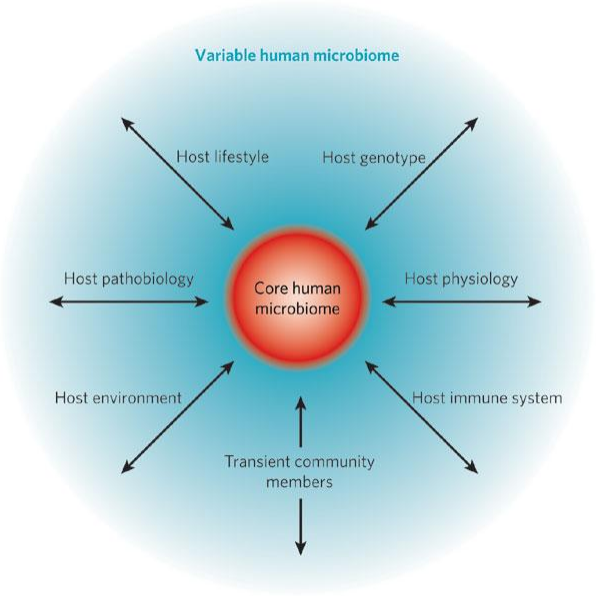


***Figura 2**: Core microbiome (in rosso) e la parte variabile (in blu). I fattori di variabilità sono indicati dalle freccie.*

*[Nature **449**, 804-810]*

Le diversità osservate non si limitano però ai siti anatomici, poiché anche il fattore temporale riveste un ruolo importante; infatti. a composizione e la struttura della comunità microbica può variare. Nel tempo n seguito a perturbazioni, quali per esempio un cambio di dieta, assunzione di antibiotici o infezioni enteriche.

Per quanto concerne la variabilità temporale, uno studio condotto nel 2008 (Dethlefsen, Huse, Sogin, & Relman, 2008) ha dimostrato che, in relazione alla somministrazione di una terapia antibiotica, il microbioma subisce alcune modifiche. Nella fase iniziale del trattamento infatti (monitorato in 18 momenti diversi) si è assistito ad una rapido aumento di omogeneità all'interno della comunità microbica; nel corso invece di una seconda analisi (Dethlefsen & Relman, 2011) effettuata in un momento successivo al trattamento su tre soggetti, si è osservata una composizione batterica con diversità aumentata, ma comunque non uguale a quella presente prima del trattamento stesso.

A sostegno di ciò in figura 6 si riporta un grafico prodotto dallo studio appena citato, in cui si mostrano tre misure di biodiversità nei tre soggetti studiati durante il decorso del tempo. Le tre misure utilizzate sono l'abbondanza di alcune specie prese come riferimento, la diversità filogenetica e l'indice di Shannon. Nello specifico, la prima misura si riferisce all'abbondanza di alcune OTU di riferimento riscontrate nei vari campioni.

Per quanto riguarda la misura di diversità filogenetica, si tratta invece della misura della lunghezza media dei rami dell'albero filogenetico che divide una qualunque coppia di specie.

Infine, la terza misura utilizzata nello studio è l'indice di Shannon, il quale quantifica la diversità di specie microbiche presente in una comunità.
Nei grafici sottostanti si riportano dunque il numero di OTU di riferimento lungo l'asse y di sinistra, mentre su quello di destra si trovano i valori dell'indice di Shannon e quelli della lunghezza dei rami dell'albero filogenetico. Si ricorda che un albero filogenetico è un grafo bidimensionale che mostra le relazioni evolutive tra specie o geni di specie distinte. La lunghezza dei suoi rami è proporzionale ai cambiamenti genetici che intercorrono tra le specie connesse dai rami stessi.
Si nota chiaramente che tutte e tre le misure di diversità variano nel tempo (nel caso specifico si tratta di giorni).

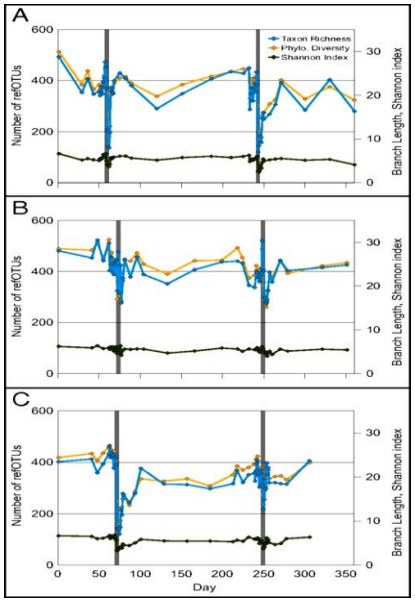


***Figura 3**: Andamento di tre misure di diversità nel tempo.*

Questi studi dimostrano quindi sia la variabilità del microbioma nel tempo, sia l'esistenza del sopracitato *"core"*, che in questo caso è più propriamente detto *"temporal microbiome core"*. Si nota infatti un insieme di geni che rimane invariato nel tempo, le cui dimensioni dipendono però dal sito anatomico in cui esso si trova; per esempio si osserva un "*core"* di dimensioni maggiori nella cavità orale rispetto a quello presente sulla pelle (Caporaso et al., 2011).
In relazione al fattore temporale si può dunque suddividere il microbioma in due categorie: quello transiente, che comprende la parte soggetta a variazioni nel tempo, e quello persistente, che invece rappresenta la parte costante nel tempo.

Altra caratteristica del microbioma è l'*ecosystem resilience*, definita come la capacità di un sistema di rispondere ai disturbi e reagire ai cambiamenti da essi provocati, riorganizzandosi in modo tale da mantenere pressoché inalterate le funzioni, la composizione e la struttura iniziale

del sistema stesso (Walker et al. 2004, 5). È stato anche elaborato un modello per descrivere in maniera schematica il significato della *resilience* (Holling 1973, 1-23), nel quale il sistema è rappresentato come una sfera posizionata su una superficie topografica che indica l'ambiente in cui il sistema sussiste. Sulla suddetta superficie si trovano depressioni denominate "bacini di attrazione", che rappresentano gli stati di equilibrio del sistema, e delle colline ad indicare l'instabilità. Nel modello si considerano ovviamente anche gli eventuali disturbi, i quali alterano la superficie (e dunque l'ambiente in cui vive la comunità) con la conseguente formazione di nuovi bacini che possono influenzare la stabilità del sistema. La sfera che quindi inizialmente si trova all'interno di uno dei bacini di attrazione, può essere attratta da un altro punto più stabile in seguito a perturbazioni. Si può dunque intuire come la *resilience* sia tempo e spazio-dipendente (le modifiche alla superficie possono applicarsi in un punto piuttosto che in un altro, e dilazionate nel tempo piuttosto che istantaneamente).
Basandosi su tale modello, Holling ha delineato 4 importanti caratteristiche riferite alla *resilience*:

- la *panarchia,* cioè l'influenza delle dinamiche del sistema e/o degli altri stati di equilibrio sulla capacità del sistema di recuperare lo stato pre-disturbo;
- la *latitudine,* che indica il massimo disturbo che il sistema può sopportare, oltre il quale esso perde la capacità di recupero;
- la *resistenza* al cambiamento;
- la *precarietà,* cioè la distanza dello stato del sistema dalla soglia oltre la quale esso viene attratto da un "bacino" diverso rispetto a quello in cui si trova.

La *resilience* può inoltre essere quantificata e sono principalmente tre le misure che vengono utilizzate:

- misura dell'elasticità del sistema, che quantifica il tasso di recupero dello stato precedente al disturbo;
- misura del tempo che il sistema impiega per ritornare ad uno stato simile a quello iniziale;
- misura del massimo scostamento dalle condizioni iniziali che il sistema è in grado di contrastare.

Poiché si è finora parlato di *resilience* di un sistema in relazione ad un disturbo che vi agisce, è da precisare che per disturbo si intende un processo interno o esterno che causa un repentino cambiamento nella composizione e/o struttura della comunità considerata, introducendo

eterogeneità e favorendo la proliferazione di alcune specie piuttosto che altre. Tali reazioni nel sistema variano a seconda dell'intensità e della frequenza del disturbo; più precisamente uno stimolo di media intensità consente di introdurre la massima diversità di specie all'interno di una comunità, in quanto più specie hanno la possibilità di colonizzare il sito. Inoltre, se il disturbo si ripete con frequenza costante e ad intensità di volta in volta simile allora la comunità vi si adatta, e la sua struttura e le sue funzioni riflettono la storia di applicazione del disturbo.
Focalizzandoci ora su un particolare ecosistema, quale è il microbioma umano, anch'esso è sottoposto a disturbi, sia biologici (cambio di dieta, uso di particolari detergenti, utilizzo di certi medicinali, ecc) che fisici. Essi compaiono con una frequenza ed un'intensità maggiore rispetto a quelli presenti nella maggior parte degli ecosistemi naturali, e producono cambiamenti significativi e ancora non del tutto conosciuti sulla salute umana, con parziale recupero finale delle condizioni iniziali. Uno studio condotto su dei neonati (McNulty et al. 2011, 106ra106) in cui si è somministrato loro del latte fermentato contenente *Bifidobacterium*, *Lactobacillus*, *Lactococcus* e *Streptococcu*s ha dimostrato che la composizione del microbioma é rimasta pressoché inalterata.

Nonostante l'eterogeneità riscontrata nel microbioma, si è giunti ad una classificazione degli individui basata sulla composizione batterica intestinale, definendo gli *enterotipi* come unità fondamentali di questa classificazione (Arumugam et al. 2011, 174-180). Sulla base dell'analisi di sequenze di DNA provenienti da 39 campioni appartenenti a soggetti di sei diverse nazionalità si sono infatti individuati tre ceppi batterici principali (*Bacteroides, Prevotella* e *Ruminococcus*) i quali sono indipendenti da età, sesso, zona geografica di appartenenza o dieta, ed in base alla prevalenza di uno di essi, ogni individuo viene catalogato in uno dei tre enterotipi. Lo studio condotto dal MetaHIT ha evidenziato come sussista una relazione tra l'enterotipo di appartenenza e le funzioni del soggetto, quali ad esempio la produzione di alcune vitamine, la predisposizione all'obesità e forse anche il gruppo sanguigno (Kau et al. 2011, 327-336).
Si è inoltre evidenziato come il microbiota sia influenzato dallo stile di vita del soggetto, in particolar modo dalla dieta. Tale associazione è stata confermata da uno studio condotto da Wu e altri (Wu et al. 2011, 105-108) su soggetti di età compresa tra i 2 ed i 50 anni, con due diverse abitudini alimentari. Nel caso di dieta ricca di proteine e grassi animali si registra una predominanza dell'enterotipo *Bacteroides*, mentre una dieta composta principalmente da carboidrati comporta una predominanza di *Prevotella*. Si nota però che, al contrario di quanto affermato nei primi studi in questo ambito, non si può effettuare una netta separazione degli enterotipi; si può piuttosto parlare di due tipologie di microbiota definiti "*biome types"* (*Bacteroides-Ruminococcus* e *Prevotella*) intesi come un continuo (gradiente) anziché come

gruppi distinti (Huse et al. 2012, e34242). Nei soggetti è dunque possibile trovare maggior abbondanza di uno dei due *types*, ma ci sarà sempre una loro parziale sovrapposizione.

2.2. Il microbioma durante la crescita

Il microbioma di ciascun individuo è ereditato per la maggior parte dalla madre, attraverso molteplici vie. La rottura del sacco amniotico rappresenta la prima interazione diretta con il microbioma materno, in quanto il bambino entra in contatto con le popolazioni microbiche presenti nella vagina. Successivamente l'allattamento introduce nel bimbo ulteriori organismi, tra tutti i lactobacilli, i quali preparano il tratto gastrointestinale alle specie batteriche che successivamente vi si insedieranno.
Per quanto riguarda invece l'influsso paterno sul microbioma del neonato, si evidenzia un importante contributo a livello di *Helicobacter Pylori*, batterio gram negativo presente nel muco gastrico situato nello stomaco.
Durante la crescita dell'individuo la composizione del microbioma subisce diverse modifiche, come anticipato nel paragrafo 3.1. Per esempio, l'eruzione dei denti provoca una variazione del microbioma orale, così come accade in seguito all'esposizione a microorganismi presenti nell'ambiente o all'assunzione di antibiotici. Altro esempio di variazione del microbioma in relazione all'età è rappresentato dalla modifica del rapporto tra *Bacteroides* e *Firmicutes* a livello intestinale col passare degli anni. Anche a livello di microbiota vaginale si osservano differenze tra quello presente in fase post menopausa e quello del periodo riproduttivo (Cauci et al. 2002, 2147-2152).

In conclusione, sia nella fase post natale sia in quella adulta, il microbioma influenza gli aspetti riproduttivi, cognitivi, metabolici e immunologici dell'individuo; si sostiene quindi che il microbiota sia atto al supporto di alcune funzioni nelle prime fasi di vita mentre nelle ultime esso contribuisca alla morte dell'ospite. Per esempio è emerso che il microbioma possiede un potenziale oncogenico che si esplicita in relazione all'età dell'individuo causando l'aumento della proliferazione cellulare e producendo metaboliti pro-mutageni (per esempio butirrato) (Vanhoutvin et al. 2009, e6759).

2.3. La relazione tra microbioma e lo stato di salute

L'indagine sulla relazione tra microbioma e salute umana non si limita allo studio della composizione e dell'abbondanza delle popolazioni microbiche al momento dello stato di salute e non; essa si estende infatti allo studio delle variazioni del microbioma cui si assiste durante il decorso della malattia della comprensione del ruolo che esso può rivestire nella guarigione o nel compromettere lo stato di salute e di quali siano gli eventuali fattori che influiscono sulle popolazioni batteriche.

Numerosi sono gli esempi che supportano l'ipotesi di una relazione tra microbioma e malattia, e riguardano popolazioni batteriche residenti in vari siti. Un primo riscontro si ha nelle malattie cutanee, nell'ambito delle quali si è verificato, per esempio, che le dermatiti classiche si presentano in regioni della pelle in cui risiedono popolazioni batteriche simili (Grice and Segre 2011, 244-253); così come nei pazienti affetti da ulcere croniche si è evidenziata l'abbondanza di un certo tipo di microrganismi (*Pseudomonadaceae*) rispetto ai soggetti sani. Nel caso invece delle ulcere diabetiche si è riscontrato che l'abbondanza di *Streptococcaceae* aumenta rispetto alla quantità di norma presente in soggetti non diabetici (Price et al. 2009, e6462).

Altri studi hanno dimostrato che il microbioma riveste un ruolo attivo nella genesi dei tumori (Plottel and Blaser 2011, 324-335), in quanto le interazioni tra ospite e microbioma possono stimolare o meno la produzione di cellule neoplastiche. Lo schema di figura 4 illustra tre principali meccanismi che possono portare all'insorgenza di tumori: nella classe A il microbioma interagisce con gli immunociti, nella classe B si ha il contatto diretto tra comunità microbiche e tessuto parenchimale, nella classe C le interazioni locali tra microbioma e tessuto producono effetti su tessuti lontani.

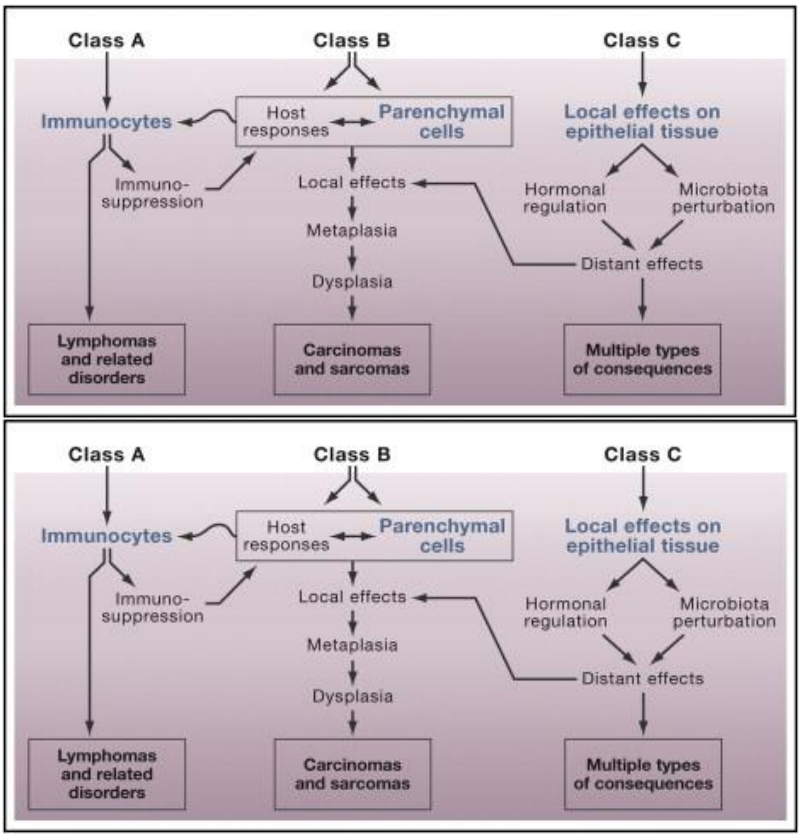


Figura 4. *Le tre classi nelle quali si possono suddividere le interazioni del microbioma-ospite che portano alla formazione di tumori.*

Un esempio della relazione tra microbioma e tumori è rappresentata da *Helicobacter Pylori*, un batterio gram negativo flagellato acidofilo che risiede nello stomaco umano ed è il componente principale del microbiota gastrico. Diversi studi condotti su animali hanno evidenziato che *H. Pylori* è implicato nell'insorgenza dell'adenocarcinoma gastrico (Peek Jr and Blaser 2002, 29); nello specifico la carenza del suddetto batterio coincide con l'aumento dell'incidenza di questa forma tumorale. La perdita di *H. Pylori* infatti comporta alterazioni nella fisiologia del tratto gastro-intestinale (secrezione di acido gastrico, ormoni e immunociti) e cambiamenti nella composizione del microbioma, poiché si altera la proporzione di altre specie microbiche presenti. Questi cambiamenti possono avvenire in un certo punto del tratto gastrointestinale (per esempio nello stomaco) e causare l'insorgenza del tumore in una zona adiacente ma separata (per esempio, nell'esofago); è dunque il caso di un'interazione di classe C tra microbioma e ospite.

Un altro studio evidenzia invece come il linfoma gastrico Mucose-Associated Lymphoid Tissue (MALT) sia causato dalla presenza di *H. Pylori* ; esso infatti è positivo per una certa proteina (*CagA*) e ne permette dunque la traslocazione diretta nei linfociti B. Questo comporta alterazioni nella trasmissione dei segnali intracellulari, bloccando di fatto l'attività di apoptosi

dei linfociti i quali proliferano quindi in maniera incontrollata.

Tra gli studi effettuati sulla ricerca delle correlazioni tra microbioma e stato di salute dell'ospite, è per noi di particolare interesse lo studio relativo alla bronco pneumopatia cronico ostruttiva (BPCO). Come già esposto nel capitolo 2, essa è caratterizzata da un'ostruzione irreversibile delle vie aeree, da ipersecrezione di muco e distruzione dello spazio alveolare (enfisema).
Uno studio condotto da Erb-Downward e colleghi (Erb-Downward et al. 2011, e16384) indaga il ruolo del microbioma polmonare nei soggetti fumatori affetti da BPCO e nei fumatori sani, così da evidenziare l'eventuale relazione tra microbioma e stato di salute dell'ospite. In tale studio si analizzano due tipi di campioni: il fluido ottenuto da lavaggio bronco alveolare (BAL) ed il tessuto espiantato da diverse regioni polmonari di fumatori sani e di fumatori affetti da BPCO. Si è dunque analizzato mediante pirosequenziamento (vedi capitolo 3) il DNA dei batteri presenti nei due tipi di campioni, confrontandolo con quello presente tipicamente nei soggetti non fumatori sani.
Dalle analisi è emerso innanzitutto che il microbioma polmonare dei soggetti affetti da BPCO risulta meno eterogeneo rispetto a quello dei fumatori sani; non è chiaro se ciò sia la conseguenza dell'infiammazione provocata dalla malattia o causa del peggioramento della malattia stessa. La riduzione della diversità misurata all'interno del microbioma in presenza di malattia è confermata anche da altri studi condotti in caso di infiammazione del tratto gastro-intestinale; che confermano l'esistenza di una relazione tra microbioma e stato di salute.
Si osserva inoltre che la presenza di numerosi siti microscopici all'interno dei polmoni comporta l'organizzazione delle comunità batteriche in diverse strutture, anche in regioni adiacenti del polmone. Infine lo studio mette in evidenza che nei fumatori il microbioma presente nei polmoni è significativamente diverso, (sia per tipo che per numerosità delle specie batteriche), dal microbioma della cavità orale.

Ulteriore esempio di come la composizione del microbioma possa portare a conseguenze nello stato di salute è stato riscontrato a livello colon-rettale; si è verificato per l'appunto che sussiste una relazione tra cancro al colon-retto ed il microbiota presente in tale regione. I batteri residenti nel colon possono infatti stimolare una risposta immunitaria esagerata nell'ospite, attraverso le cellule T-helper 17, promuovendo di fatto il tumore (Wu et al. 2009, 1016-1022). Si è inoltre dimostrato che alterazioni del microbiota del colon influenzano l'espressione di alcuni geni coinvolti nel ciclo di regolazione cellulare. Infine altri studi condotti sul microbiota del tessuto colon-rettale affetto da cancro hanno evidenziato differenze con il microbiota presente nell'adiacente tessuto sano. In particolare, nel tessuto malato si è rilevato una significativa abbondanza di batteri appartenenti alla specie *Fusobacterium* rispetto al tessuto sano (Castellari net al., 2012).

Dimostrata la relazione tra microbioma e salute, uno dei problemi che si sta tuttora affrontando riguarda la causalità della relazione microbioma-malattia. Per risolvere la questione si stanno utilizzando diversi modelli animali, quali ad esempio il ratto. Il modello *gnotobiotic animal* è costituito da animali il cui microbioma è del tutto noto, ma poiché il costo degli animali è alto e si richiede esperienza per il suo utilizzo, tale modello non è molto diffuso. Per quanto riguarda invece il *conventionalized animal*, ci si riferisce ad animali nei quali si inserisce il microbiota intestinale umano per colonizzarne il tratto gastrointestinale. L'ultimo modello è un'evoluzione del precedente in quanto viene trasferito nell'animale il microbiota presente in tutti i tessuti umani.

Si può dunque indagare sugli animali la maggior parte delle patologie umane, monitorando così il ruolo del microbioma nell'insorgere e nell'evolversi della malattia stessa.

Il fine ultimo degli studi in questo ambito è la possibilità di manipolare il microbioma umano in modo tale da diminuire il rischio di alcune malattie o alterare quelle vie metaboliche e immunologiche che risultano essere dannose per la salute.

3. Metodologie per la caratterizzazione del microbioma tramite sequenziamento

Caratterizzare e studiare il microbioma umano significa analizzare il materiale genetico del microbiota ed a tal fine sono due i passi comuni da seguire; il primo step consiste nel processare i campioni biologici ed estrarne il DNA, mentre in un secondo momento esso viene sequenziato, in modo da trovare l'ordine delle basi nucleiche lungo la catena di DNA. Le tecniche per il sequenziamento sono molteplici e nel paragrafo seguente se ne presenta un sommario, con particolare attenzione alle tecnologie definite di "*next generation*".

3.1. Tecnologie di sequenziamento del DNA

Il sequenziamento del DNA, come anticipato, è il processo che consente di trovare l'ordine delle basi nucleiche lungo una catena di DNA; sono disponibili molteplici tecnologie, ma essenzialmente possono essere divise in due categorie: i metodi di "prima generazione" e quelli di "seconda generazione", altresì denominati metodi di "next generation sequencing" (NGS).
Entrambe le tipologie presentano tratti comuni per quanto concerne le fasi che portano alla ricostruzione della sequenza delle basi della catena stampo; il protocollo che si segue per il sequenziamento è infatti composto di 3 fasi, adottate sia dai metodi di prima generazione sia da quelli di seconda: la preparazione della libreria, il sequenziamento, l'acquisizione e l'elaborazione dei segnali prodotti durante la fase precedente.
In primis dunque si procede con la preparazione di una cosiddetta libreria di DNA, che consiste in una collezione di repliche del DNA da analizzare. Successivamente si passa alla fase del sequenziamento vero e proprio, il quale viene eseguito in modo diverso a seconda della tecnica adottata. Infine si acquisiscono e analizzano i segnali prodotti nella seconda fase, anch'essi caratteristici della tecnica adottata.
Per quanto riguarda la preparazione della libreria, il passo basilare è la replicazione della catena di DNA di interesse e questo si realizza mediante PCR (*Polymerase Chain Reaction*). Si tratta di una reazione di amplificazione in vitro di sequenze di DNA, ideata da Mullis e altri nel 1986. Il primo step consiste nel denaturare la catena di DNA da far replicare, in modo tale da ottenere

singole catene; i campioni sono a tal scopo sottoposti a temperature superiori ai 90° C. In seguito si abbassa la temperatura (40-60° C) per consentire a degli oligonucleotidi a singolo filamento (*primer*) di allinearsi con le estremità dei filamenti di DNA stampo; i primer fungono quindi da innesco per la DNA polimerasi (enzima che catalizza la replicazione del DNA), la quale per ogni filamento ne sintetizza uno nuovo (ad una temperatura di 72° C). L'enzima può però procedere solo aggiungendo nucleotidi ad un filamento preesistente, e da ivi si comprende quindi l'importanza dei primer per iniziare la reazione. L'intero processo può ripetersi per molti cicli, al termine di ciascuno dei quali le molecole di DNA raddoppiano.

In Figura 5 si trova uno schema del processo appena descritto.

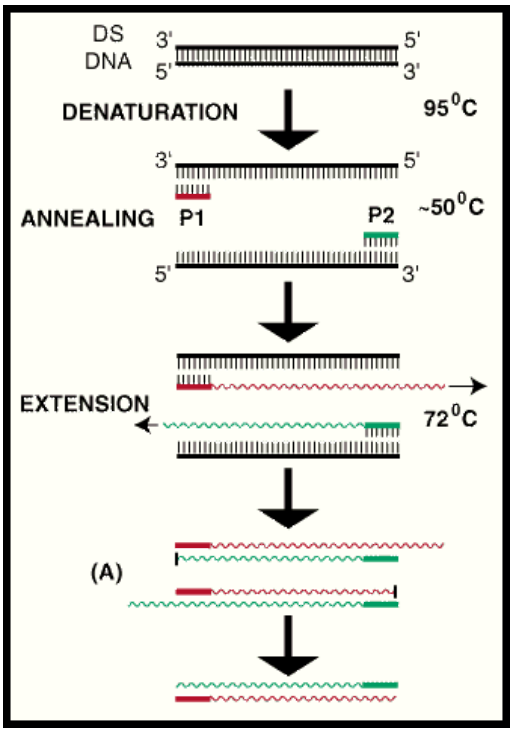


***Figura 5**: Schema delle fasi del processo di PCR*

3.1.1 Metodi di prima generazione

Il metodo Sanger (Sanger 1975, 441) è stata la prima tecnica proposta per l'identificazione delle sequenze amminoacidi che e tuttora la più utilizzata, il cui concetto di base è che la sequenza di

DNA può essere determinata se è possibile calcolare la distanza di ogni tipo di base azotata rispetto ad un'origine nota.

Per quanto riguarda gli aspetti pratici del metodo in questione, il primo passo è rappresentato dalla denaturazione del DNA ed ogni singola catena subisce un processo di purificazione. Successivamente, si procede con la fase di amplificazione al fine di ottenere *ampliconi*, copie "clonali" della catena di DNA da analizzare, definita *template* (Figura 6a). L'amplificazione del DNA consiste nel frammentare lo stesso ed introdurre i pezzi in diversi plasmidi, piccole molecole circolari di DNA presenti nel citoplasma batterico e distinte da quello cromosomiale. Ciascun plasmide contenente un frammento del DNA che si vuole amplificare è inserito in un battere ospite (ad esempio *Escherichia Coli*), il quale viene fatto replicare. Ad ogni sua divisione, anche il plasmide in esso contenuto si duplica ed alla fine del processo si ottiene una colonia batterica in cui ogni elemento contiene una copia del plasmide e dunque del frammento del DNA template. Per ogni reazione di sequenziamento si preleva quindi un plasmide da una specifica colonia, si isola e si procede con una seconda fase di amplificazione, questa volta tramite PCR. Si ottiene così una quantità di materiale genetico che rende possibile effettuare il sequenziamento vero e proprio.
Quindi, dopo aver preparato la libreria di DNA (la collezione di ampliconi generati in precedenza), si realizza una soluzione contenente gli ampliconi, le DNA-polimerasi, una grande quantità di deossinucleotidi e un numero inferiore di dideossinucleotidi marcati con quattro fluorescenti diversi (ddNTP). I deossinucleotidi sono i componenti fondamentali del DNA e sono costituiti dal deossiribosio, uno zucchero, da un gruppo fosfato e da una base azotata (adenina, citosina, guanina o timina). I dideossinucleotidi invece sono in tutto e per tutto uguali ai deossinucleotidi, se non per la presenza del dideossiribosio al posto del ribosio; questo impedisce il legame con altri nucleotidi. Nella soluzione si assiste dunque al sequenziamento vero e proprio, in cui gli

Perciò, quando occasionalmente ed in maniera casuale i ddNTP in soluzione sono inclusi nella catena di DNA che si sta formando sulla base di uno dei frammenti di DNA template, la reazione termina. In tal modo, il metodo di Sanger produce una molteplicità di frammenti di DNA di diversa lunghezza, ciascuno dei quali termina con un particolare nucleotide marcato con un elemento fluorescente o radioattivo. In seguito, i frammenti di DNA sono separati in base alla loro lunghezza tramite elettroforesi su gel, in quanto quelli più corti si muovono più velocemente all'interno dei capillari ripieni di gel e viceversa per i più lunghi. Durante la corsa elettroforetica i frammenti sono eccitati da una sorgente laser, che stimolando dunque gli elementi fluorescenti consentono la formazione di una particolare "traccia". Essa è analizzata da uno specifico software, che la traduce in una sequenza di nucleotidi, definita anche *read*.

Il tipo di base è identificato a partire dal colore del fluorocromo registrato durante la corsa elettroforetica in seguito all'eccitazione laser, in quanto si usano 4 colori di marcatori differenti, uno per ogni tipo di ddNTP. La posizione delle basi è ricostruita invece dall'ordine dei frammenti al termine della corsa elettroforetica; avendo infatti a disposizione frammenti di diversa lunghezza ma appartenenti alla medesima catena di DNA template, il primo a concludere la corsa è quello composto da un solo nucleotide, che per la precisione è la prima base della catena da sequenziare, mentre il frammento più lento corrisponde all'intero template e dunque fornisce l'ultimo nucleotide della read.

Il metodo di sequenziamento di Sanger è consigliato per progetti di scala ridotta, in quanto consente un limitato livello di parallelismo (non molte catene di DNA sequenziate in contemporanea) e poiché i costi (reagenti, strutture bioinformatiche, i sequenziatori a capillari..) sono troppo elevati. Per esempio, per sequenziare 100 geni provenienti da 100 campioni, considerando ogni gene composto da 10 esoni, il costo si stima essere dai 300,000 $ ad oltre 1,000,000 $.

3.1.2 Metodi "next generation"

Queste tecnologie differiscono da quelle di prima generazione non solo nel metodo di sequenziamento ma anche nella preparazione dei template (Figura 6). Infatti il DNA da sequenziare è inizialmente frammentato casualmente e poi ad ogni pezzo vengono legati dei comuni adattatori; si tratta di oligonucleotidi che solitamente consentono il legame con appositi supporti o piattaforme. Successivamente si applica ai frammenti uno dei molteplici approcci (emulsion PCR, bridge PCR…) in modo da ottenere ampliconi raggruppati nello spazio; il risultato è che ogni gli ampliconi di ciascun frammento sono immobilizzati su di un substrato solido (bridge PCR) o sulla superficie di una biglia delle dimensioni del micron.

I principali vantaggi delle tecnologie di seconda generazione sono gli alti livelli di parallelismo (centinaia di milioni di read sequenziale in parallelo) e i costi non elevati per la produzione delle sequenze di DNA. Ma esse presentano anche svantaggi: la lunghezza delle catene sequenziate è ridotta e l'accuratezza non è delle migliori.

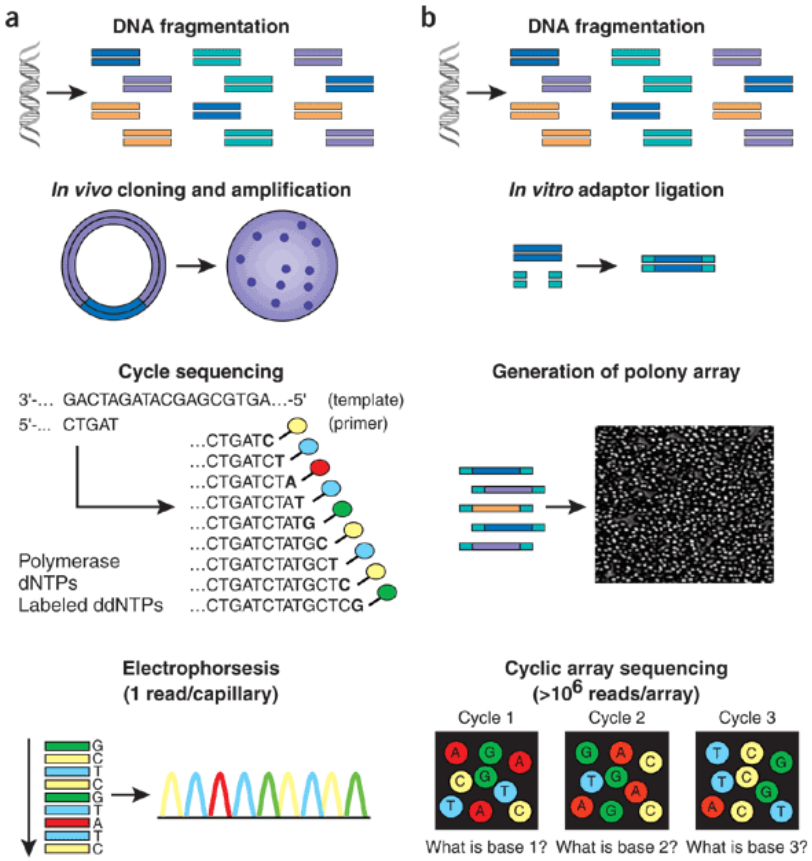


***Figura 6**: Sequenziamento convenzionale (a) rispetto a sequenziamento di seconda generazione (b) [Nature Biotechnology,**26**,10]*

Di seguito dunque saranno presentate 3 tecnologie NGS: pirosequenziamento Roche-454, Illumina e AB Solid.

- *Pirosequenziamento Roche-454*

Questa piattaforma può sequenziare 400-600 megabasi in 10 ore e la lunghezza delle catene sequenziate è di circa 500 nt (fino a 1000 nt).

Come primo passo il DNA è denaturato, ridotto in frammenti ed a questi si aggiungono mediante ligazione dei comuni adattatori; successivamente ogni frammento è catturato sulla superficie di biglie di 28 μm di diametro, le quali sono inglobate in emulsioni di acqua in olio. Gli ampliconi sono generati attraverso emulsion PCR, ogni biglia è caricata assieme a specifici enzimi su di una piattaforma contente vasche di circa 44 μm di diametro, detta PicoTiterPlate (Figura 11c), ed infine si procede con il pirosequenziamento. Tale tecnica si basa sul rilevamento del pirofosfato inorganico (PPi), ottenuto dall'annessione di un nucleotide alla catena in formazione durante la sintesi del DNA. Più precisamente, quando una base complementare è allineata con la catena stampo, si genera un PPi ed esso viene convertito in

ATP mediante la sulfurilasi. L'ATP è poi sfruttato dalla luciferasi, la quale converte la luciferina in ossiluceferina producendo luce; questa reazione può avvenire contemporaneamente nelle vasche separate della piattaforma e questo è un aspetto a favore del parallelismo. Il segnale di luce è quindi rilevato da una camera CCD che consente di identificare mediante coordinate spaziali la vasca da cui proviene il segnale. I risultato sono dunque riportati in un particolare grafico detto pirogramma (Figura 7).

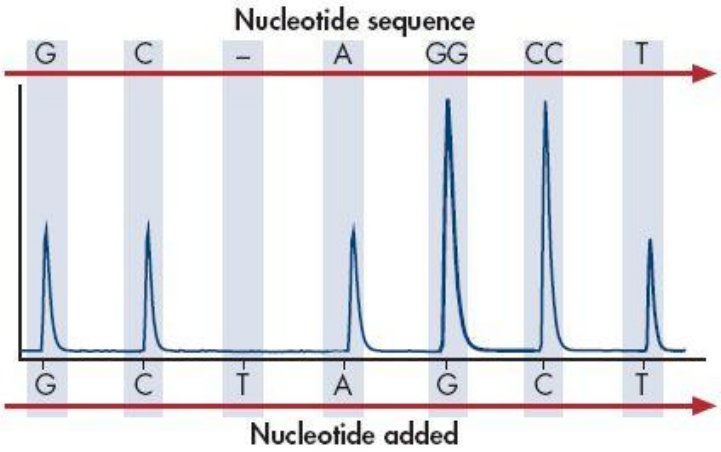


Figura 7: *Esempio di pirogramma. L'ampiezza del segnale è proporzionale al numero di nucleotidi legatisi con il DNA di stampo.*

Un problema tipico della tecnica appena descritta è rappresentato dagli omopolimeri (incorporazioni successive della stessa base), in quanto la loro lunghezza deve essere dedotta dall'ampiezza del segnale misurato, processo affetto da un certo tasso di errore. Un punto forte del pirosequenziamento invece è la lunghezza dei frammenti sequenziati, che varia dalle 200 alle 300 paia di basi.

- *Illumina*

La preparazione della libreria di DNA è simile a quella già esposta per il pirosequenziamento, con la differenza che l'amplificazione non avviene mediante emulsion PCR, bensì attraverso la bridge PCR. Questa tecnica prevede di immobilizzare su di un supporto rigido i due tipi di primer solitamente utilizzati nella PCR, cioè il *forward* ed il *reverse* primer. Il primo funge da innesco per la catena complementare al filamento 3'→5', mentre il secondo permette la ricostruzione della catena complementare al filamento in direzione 5'→3'.
Una volta che un filamento viene amplificato, questo si ripiega in direzione di uno dei primer che sono depositati nelle vicinanze sul supporto, vi si allinea e si assiste quindi ad un'ulteriore amplificazione in direzione opposta di quella con cui si è creata la catena ripiegata. Tale

processo si ripete numerose volte per la stessa catena di DNA, pertanto al termine si ottiene un gruppo (*cluster*) di ampliconi ripiegati su se stessi. Considerando che il DNA da sequenziare è inizialmente sottoposto a frammentazione, ciascuno dei frammenti è amplificato mediante bridge PCR, ragion per cui al termine della preparazione della libreria si trovano sul supporto diversi gruppi di ampliconi, in numero pari a quello di frammenti ottenuti in partenza.
Successivamente si procede con la denaturazione, in modo tale da ottenere degli ampliconi a singola catena e non ripiegati. A questo punto inizia il primo ciclo del vero e proprio sequenziamento: si introducono sulla piattaforma quattro tipi di nucleotidi modificati (ddNTP), ognuno contenente un marcatore fluorescente diverso, a seconda del tipo di base. Si assiste dunque alla replicazione degli ampliconi, con il contributo di un nucleotide per ogni ciclo, in quanto ad ogni incorporazione di nucleotidi nella catena in estensione la reazione si blocca vista la loro natura modificata. Si procede quindi con l'acquisizione di immagini sensibili alla fluorescenza, che permettono ad ogni ciclo di evidenziare che tipo di base sia stata incorporata in ciascun gruppo di ampliconi. Al termine dell'acquisizione il supporto viene lavato e può così iniziare un nuovo ciclo.Ovviamente, a differenza della piattaforma Roche-454, non si incontrano problemi con gli omopolimeri mentre sono frequenti errori di sostituzione. Un vantaggio della piattaforma Illumina è il grado di parallelismo, che permette di generare dalle 18 alle 35 Gb di materiale per ogni ciclo (Metzker, 2010).

- AB SOLiD

La libreria si costruisce come già esposto nel caso della piattaforma Roche-454, per cui il DNA stampo si amplifica mediante emulsion PCR. Una volta che le sfere magnetiche sono fissate ad un supporto solido, si procede con il sequenziamento, nel quale interviene però la DNA-ligasi anziché l'enzima polimerasi. Ad ogni ciclo del processo si liberano sul supporto dei particolari ottameri, i quali sono formati da 2 basi che consentono l'appaiamento con la catena di stampo, 3 basi degenerate e 3 basi universali; gli ottameri sono inoltre marcati con un elemento fluorescente, di colore dipendente dalle prime due basi.
Il sequenziamento inizia con l'appaiamento di un ottamero alla prima coppia di basi della catena stampo adiacenti al primer; dopodiché si acquisisce l'immagine che evidenzia il tipo di fluorocromo e dunque la coppia di basi che si è allineata allo stampo. In seguito si rimuove il marcatore fluorescente scindendo il legame tra la base 5 e la base 6 dell'ottamero, predisponendo l'estremità dello stesso ad una successivo legame con un altro ottamero. Questo si ripete per 10 volte in totale, dopodiché si sposta il primer di una posizione (una base) e si ripete il sequenziamento; il riposizionamento del primer seguito dall'allineamento degli

ottameri alla catena stampo e l'acquisizione delle immagini si ripete per altre 4 volte. In Figura 8a si ripropone una rappresentazione grafica del processo.
Infine, per completare il sequenziamento e ricostruire l'ordine delle basi sulla catena stampo, si confrontano le immagini ottenute durante i vari cicli (Figura 8b).

Per quanto riguarda gli aspetti negativi, AB SOLiD non consente di sequenziare catene molto lunghe (circa 35 paia di basi), e non bisogna dimenticare che l'emulsion PCR, a cui ricorre per creare la libreria, non è una tecnica sempre affidabile. La presente piattaforma dall'altro canto fornisce una grande densità di dati in parallelo, grazie alle biglie da 1μm di diametro (si ricordano i 28 μm nel caso della Roche-454); a parità di dimensione di supporto solido infatti, AB SOLiD consente di utilizzare un maggior numero di sfere e dunque di frammenti di DNA stampo.

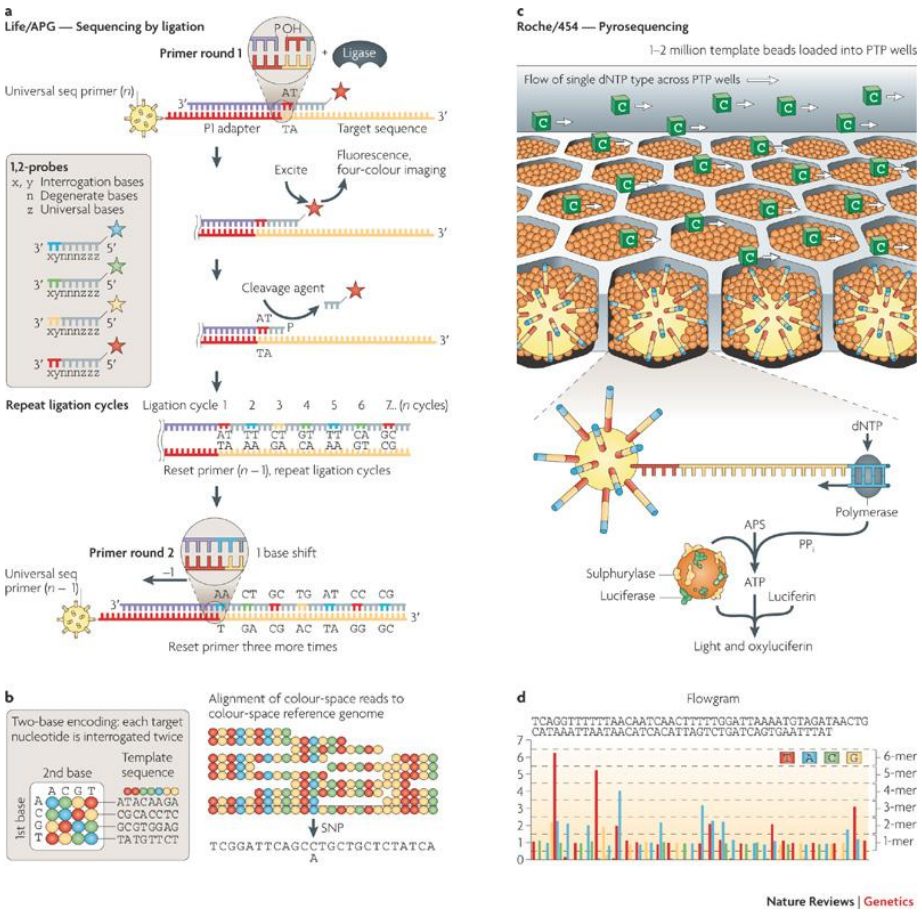


Figura 8*: nel punto a) si espone il procedimento di duplicazione degli ampliconi attraverso gli ottameri e la DNA-ligasi.*

Nel punto b) si rappresenta il passo finale del sequenziamento con AB SOLiD, il confronto tra le immagini ottenute nei diversi cicli. Nel punto c) si riporta la reazione che sta alla base del pirosequenziamento. Nel punto d) si raffigura un pirogramma, risultato ottenuto dalle immagini registrate durante il pirosequenziamento.

3.2 Caratterizzazione in generi e specie

Gli studi del microbioma possono suddividersi principalmente il due categorie: l'approccio metagenomico e l'approccio basato sugli ampliconi marcati.

Nel primo caso si sequenzia l'intero genoma raccolto dal microbiota, cioè da tutti i miceo-organismi presenti in un dato sito, ricorrendo spesso all'alternativa "shotgun"; in tal caso il genoma viene dapprima frammentato in modo casuale ed in seguito si procede con l'amplificazione ed il sequenziamento dei singoli pezzi. Le read ottenute possono essere assemblate per formare sequenze più lunghe dette *contigs*, in modo da ricostruire l'ordine delle catene di DNA originarie, precedenti la frammentazione. L'assemblaggio delle read avviene grazie ad appositi software identificano porzioni di sequenza che esse hanno in comune.

È inoltre di fondamentale importanza nello studio del microbioma l'individuazione dell'appartenenza tassonomica dei componenti delle comunità microbiche. A tal scopo si eseguono confronti tra le read ed un database che cataloga l'associazione tra un certo genoma ed un particolare livello tassonomico.

Il secondo approccio per lo studio del microbioma è invece quello basato sugli ampliconi marcati; in questo caso non si sequenzia l'intero genoma del microbiota, bensì solo alcuni geni considerati filogeneticamente significativi (*markers*). Nel nostro studio, così come nella maggior parte dei casi in questo ambito, si sceglie come *marker* il gene ribosomiale 16S (si veda figura 9); esso è presente in tutti gli organismi viventi e presenta sia regioni ipervariabili, sia regioni costanti. Queste ultime sono comuni a tutti gli organismi e consentono di distinguere il gene 16S dall'intero genoma, mentre le regioni variabili permettono di inferire l'identità tassonomica degli organismi fino a livelli molto bassi (famiglia o genere). Questo è possibile utilizzando dei particolari database, i quali sono più grandi e dettagliati di quelli disponibili per l'approccio metagenomico. Solitamente l'inferenza tassonomica è preceduta dal raggruppamento delle sequenze in base ad una predeterminata soglia di similarità, ottenendo così le OTU (si veda paragrafo 2.1). In pratica, si calcola la distanza tra le sequenze, intesa come distanza genetica e cioè come misura quantitativa della divergenza tra due sequenze, come frazione di mismatch riscontrati tra le sequenze allineate. Avendo a disposizione queste misure è possibile procedere con il clustering.

Il clustering avviene attraverso due step: nel primo si calcola un coefficiente che esprima la similarità tra i dati, nel secondo si rappresenta graficamente l'associazione tra i dati simili (mediante alberi gerarchici o mediante gruppi). Per passare dal primo step al secondo, si deve decidere una regola per raggruppare i dati simili e gli approcci seguiti sono essenzialmente tre: il clustering gerarchico, il *k*-means ed il *two-step*.

Per quanto concerne il cluster gerarchico, si deve stabilire a priori il numero di gruppi che meglio separano i dati e dopodiché essi vengono raggruppati a partire da quelli che presentano maggiore similarità, via via verso i dati con minore similarità. Si può anche procedere nel senso contrario, ma non è detto che i risultati siano gli stessi. La modalità di rappresentazione dei raggruppamenti che di solito viene utilizzata in questo caso è l'albero gerarchico.

Nel clustering *k-means*, si parte da *k* gruppi ed ogni dato viene assegnato a quel gruppo la cui media è alla minor distanza (Euclidea) dal dato considerato; la media dei gruppi viene ricalcolata iterativamente fino a che non sono più possibili nuove assegnazioni o si è raggiunto il numero massimo di iterazioni stabilito. Non si richiede dunque il calcolo della dissimilarità tra i dati, ma tale metodo è più sensibile agli outliers rispetto al precedente.

Il clustering *two-step* è utilizzato nel caso in cui una o più variabili associate ai dati siano categoriche; il raggruppamento dei dati avviene quindi sulla base delle categorie.

Il risultato finale del clustering è che le sequenze che rappresentano diversi generi si sono dunque raggruppate in diverse OTU. Solitamente le OTU con un tasso di diversità entro il 3% sono considerate rappresentative della stessa specie, mentre entro il 5% sono da considerarsi formate da geni provenienti da organismi probabilmente di specie diversa ma dello stesso genere. Il numero delle OTU ottenuto è indice dell'abbondanza e della diversità tassonomica di una certa comunità microbica.

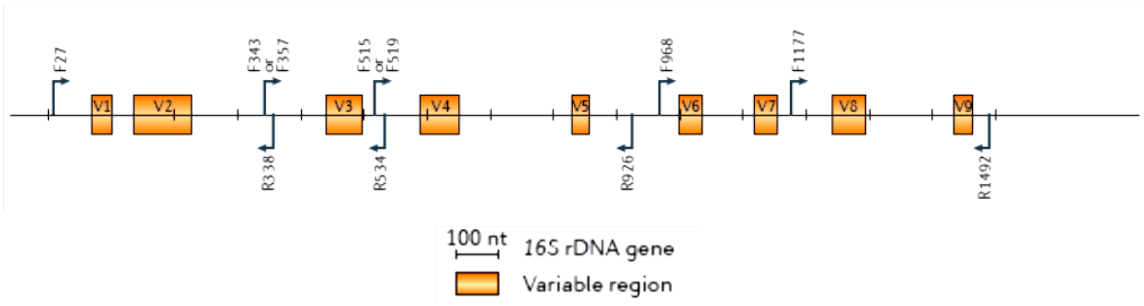


Figura 9. *Parte del gene ribosomiale 16S. Le frecce indicano i primer ed il verso nel quale procedono, mentre i box in arancio indicano le regioni ipervariabili.*

Un approccio molto usato è quello di generare le OTU e selezionare una sequenza da ogni gruppo per poi analizzarla con il classificatore *Ribosomial Database Project* (*RDP*), in modo da associare ogni OTU ad un particolare genere.

Il classificatore RDP è un classificatore Bayesiano naive (tutti gli attributi che descrivono una certa istanza sono tra loro condizionalmente indipendenti, data la categoria a cui appartiene

l'istanza); esso consente di associare in modo veloce ed efficiente le sequenze ad uno dei livelli tassonomici proposti da Bergey (Garrity et al. 2004). La tassonomia di Bergey prevede sequenze di rRNA della piccola sub unità, suddivise in 5.014 specie batteriche e prevede i seguenti livelli tassonomici, dal più basso al più alto: genere, famiglia, ordine, classe, tipo e dominio.

L'algoritmo alla base del classificatore lavora su sottosequenze (word) di 8 basi e la posizione di una word nella sequenza di indagine è ignorata. Sia $W= \{w_1, w_2,\ldots, w_n\}$ l'insieme di tutte le possibili parole, mentre con $n(w_i)$ si indichi il numero di sequenze dal database di riferimento (in questo caso il corpus di Bergey o quello proposto dall'NCBI) che contengono la parola w_i. Sia inoltre $P_i =[n(w_i) + 0,5]/(N + 1)$ la probabilità a priori di trovare una certa parola w_i in una sequenza di rRNA; con N si intende il numero di sequenze del corpus.

Considerando un gene G formato da M sequenze, il numero di sequenze contenenti la parola w_i viene indicato con $m(w_i)$. La probabilità che una sequenza del gene G contenga una certa w_i viene calcolata in questo modo $P(w_i|G) = [m(w_i) + P_i]/(M + 1)$. Allo stesso modo si può calcolare $P(v_i|G)$ con $v_i \in V$, con V insieme di word. Si può dunque stimare la probabilità di osservare nel gene G una sequenza S contente un certo insieme V di parole in questo modo: $P(S|G) = \prod P(v_i|G)$.

Procedendo poi con l'assegnazione secondo il metodo Bayesiano naive, si vuole calcolare la probabilità che una certa sequenza S (appunto la sequenza da classificare poi in uno dei livelli tassonomici) appartenga al gene G; tale probabilità, in accordo con il teorema di Bayes, si indica con P(G|S) e si calcola mediante l'equazione $P(G|S) = P(S|G) \times P(G)/P(S)$, con P(G) la probabilità a priori che una sequenza sia compresa in G e con P(S) la probabilità di osservare la sequenza S in generale tra tutti i geni. Questi ultimi due termini possono però ignorarsi in quanto sono costanti poiché si assume che tutti i generi sono equiprobabili.

Una sequenza viene dunque considerata appartenente ad un certo genere se tale assegnazione massimizza una funzione score.

Si procede in seguito con il bootstrap per stimare degli intervalli di confidenza per ciascuna sequenza. Il bootstrap è un metodo statistico che permette di attribuire una misura di accuratezza alla stima di una statistica. Il bootstrap è un metodo di ricampionamento con reimmissione in quanto a partire da un campione osservato di numerosità pari a *n*, $\mathbf{W}=(w_1,\ldots, w_n)$, si estraggono *B* campioni di numerosità costante pari a *n*. Si costruiscono in tal modo *B bootstrap samples* $W^*_1, W^*_2, \ldots, W^*_B$ ed in ciascuna estrazione bootstrap i dati possono essere estratti più di una volta e ciascun dato ha probabilità $1/n$ di essere estratto. Il vettore $\mathbf{W}^*$ può dunque interpretarsi come versione random di **W**.

Per il campione **W** si considera la statistica θ e per ciascuno dei *B bootstrap samples* si ha la statistica $\theta(w^*_1)$, $\theta(w^*_2)$,…, $\theta(w^*_B)$; tali statistiche sono replicazioni della θ. Quindi considerata la statistica θ ed il suo stimatore $T(w)= \theta_{st}$, si ha che una replicazione di tale stima è data da

$$\theta^*_{st} = \theta(w^*)$$

Si calcola quindi la stima per ogni campione bootstrap, avendo così a disposizione *B* stime di θ, dalle quali calcolare diversi parametri quali per esempio media e varianza bootstrap. Partendo da queste quantità stimate è possibile calcolare gli intervalli di confidenza.

Nel momento in cui, durante la classificazione, si assegnano più di 5 sequenze ad un certo livello tassonomico, si applica un test statistico per valutare la probabilità delle differenze registrate e che hanno portato a tale classificazione. Il P-value si stima dal valore critico Z

$$Z = \frac{\frac{x}{N_1} - \frac{y}{N_2}}{\sqrt{\mu(1-\mu)(\frac{1}{N_1} + \frac{1}{N_2})}}$$

con N_1 e N_2 numero delle sequenze appartenenti rispettivamente alla libreria 1 e alla libreria 2, x e y il numero di sequenze assegnate ad un certo livello T a partire dalle due librerie e μ è uguale a $(x + y)/ (N_1 + N_2)$.

Se invece si hanno meno di 5 sequenze assegnate allora si usa un test che è stato progettato per confrontare i livelli di trascritto.

Per testare il classificatore RDP si può scegliere di applicare l'approccio leave-one-out, in cui di volta in volta si sceglie una sequenza del corpus di Bergey come campione da classificare, tenendo le altre come componenti del training set; procedimento ripetuto per tutte le sequenze del corpus.

Oltre a testare sequenze intere si possono classificare anche sub sequenze formate di diverse dimensioni, scelte a caso.

L'accuratezza del processo di classificazione varia a seconda della lunghezza della sequenza sotto indagine ed in base al livello tassonomico a cui si sta operando. Più precisamente, l'accuratezza è minore in caso di sequenze corte e questo è dovuto allo scarso numero di dati a disposizione su cui operare; a tal proposito, il bootstrap consente di valutare se i dati sono in numero adeguato per una classificazione attendibile. Inoltre è più difficile ottenere risultati accurati nel momento in cui si eseguono assegnazioni a bassi livelli tassonomici (è intuitivamente più facile assegnare una sequenza ad un livello che presenta poche

suddivisioni e quindi meno cluster possibile da scegliere ed in cui inserire la query). Tali considerazioni sono visibili in figura 10.

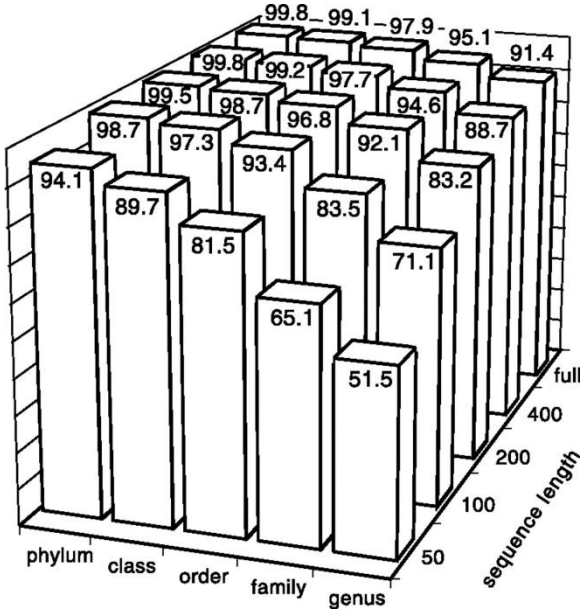


Figura 10: *Accuratezza della classificazione con RDP in base alla dimensione della query e del livello tassonomico.*
I numeri rappresentano la percentuale di classificazioni corrette.

Un'ulteriore considerazione si riferisce alle regioni delle sequenze che si utilizzano per classificare; infatti le regioni ipervariabili del 16S portano a risultati più accurati rispetto alle regioni conservate. Ed è anche per questo che negli studi che coinvolgono l'rRNA 16S, si tengono in considerazione soprattutto le regioni ipervariabili.

4 Metodi di analisi del microbiota umano

Un tipico approccio all'analisi del microbiota umano è il calcolo della sua biodiversità (si veda paragrafo 4.1), al fine di caratterizzare la varietà di specie batteriche presenti.
Si tenga poi conto che dal sequenziamento del microbioma umano si ottengono grandi quantità di dati, si necessita perciò di tecniche atte a ridurre la dimensione dei data set, oltre che ad analizzarli statisticamente, ad esprimerli graficamente in modo da consentirne un'analisi visiva immediata ed interpretare le risposte simultanee di più dati in relazione a certe variabili. A tal fine ci sono molteplici metodi di analisi, che possono catalogarsi in:

- metodi di ordinamento (paragrafo 4.2);
- analisi multivariate basate su test d'ipotesi (paragrafo 4.3).

I metodi appena citati trovano applicazione in diversi ambiti scientifici e dunque si utilizzano su dataset di diversa natura e struttura. In questa sede, si considera però una particolare organizzazione dei dati per coerenza con quelli disponibili per le nostre analisi.
Si hanno indi a disposizione delle matrici sulle cui righe sono disposti dei soggetti, mentre sulle colonne si trovano generi o OTU dei microrganismi componenti il microbioma dei soggetti. I dati in sé sono le abbondanze dei suddetti generi (o OTU) misurate per ciascun individuo.

4.1 Indici di biodiversità

"Biodiversità è uno stato o attributo, di un sito o area, e si riferisce specificatamente alla varietà all''interno e tra organismi viventi, ad assemblaggi di organismi viventi, a comunità biotiche e a processi biotici, naturali o modificati dall'uomo. La biodiversità può essere misurata in termini di diversità o di identità genetica, di numero di specie, di assemblaggio di specie, comunità e processi biotici, di quantità (abbondanza, biomassa, tasso, ecc.) e di struttura di ciascuno di essi; può essere osservata e misurata a qualsiasi scala spaziale, dai micrositi e habitat di piccole dimensioni all'intera biosfera."

(De Long, 1996)

Questa è una delle molteplici definizioni di "biodiversità" presenti in letteratura, e dunque non se ne può ivi fornire una definizione univoca; ma generalizzando per "biodiversità" si può

intendere la varietà di specie che si ritrova in un certo ambiente, e rappresenta dunque l'insieme delle forme di vita che meglio si adattano alle condizioni ambientali e ne descrive la loro variabilità. Inoltre, si utilizza talvolta il termine "biodiversità" per indicare non solo gli organismi viventi, ma anche le loro interazioni con le componenti abiotiche del loro habitat. In conclusione, la biodiversità può riassumersi in due componenti: la varietà delle specie viventi in un certo sito e l'abbondanza relativa di ciascuna di esse, considerando la loro distribuzione all'interno del sito (equidistribuzione).

In particolare la biodiversità si può suddividere in genetica (caratterizzazione del patrimonio genetico all'interno di una specie), specifica (numero e frequenza delle specie in un ecosistema) ed eco sistemica (numero ed estensione di ecosistemi in una data regione), così come si può inoltre distinguere in tipologia alfa, beta e gamma.

Per quanto riguarda la biodiversità alfa, essa definisce la struttura e la complessità di un determinato habitat e con essa si spiega dunque la diversità osservata tra le specie di un dato sito. La biodiversità beta invece esplica le differenze osservate tra specie appartenenti ad habitat distinti. Infine per biodiversità gamma si intende la differenza che sussiste in totale su una vasta area, ed essa deriva da una combinazione delle due precedenti diversità.

Tipo di biodiversità	Definizione
α	Rappresenta la diversità di un habitat o di una comunità. Descrive il numero di specie ed il grado di ripartizione delle abbondanze tra le singole specie di una comunità
β	Rappresenta il grado di cambiamento della diversità specifica tra le comunità distinte presenti in un ecosistema. Descrive le variazioni nella composizione e nell'abbondanza delle specie tra due habitat distinti
$\gamma = \alpha \times \beta$	Rappresenta la diversità specifica totale di una vasta area

***Tabella 1**. Riassunto delle tipologie di biodiversità*

Appositi indici son stati elaborati per quantificare il grado di biodiversità e nel presente lavoro di tesi si sono considerati gli indici relativi alla diversità alfa e beta, tralasciando la diversità gamma in quanto non importante alla luce dei nostri obiettivi.

N.B: Le definizioni finora utilizzate sono nate in ambito ecologico, ma di seguito si adatta la terminologia all'ambito preso in esame dal presente studio; ci si riferirà dunque a specie presenti all'interno di soggetti anziché di siti o habitat.

4.2 Metodi di ordinamento

I metodi di ordinamento hanno anch'essi finalità descrittive e di sintesi dell'informazione; si suddividono in due categorie principali: i metodi non vincolati e quelli vincolati.
Nel primo caso essi utilizzano le sole informazioni relative alle abbondanze dei generi (o delle OTU), mentre i metodi vincolati prendono in considerazione anche altre variabili ambientali associate al dataset (per esempio sesso, età e stato di salute dei soggetti, informazioni sui campioni prelevati per ottenere il materiale genetico, ecc..).

Il primo metodo di ordinamento non vincolato tipicamente utilizzato su dataset multivariati è la PCA (*Principal Component Analysis*); essa consente di individuare eventuali pattern nei dati, ridurne la dimensione e rappresentarli in grafici preferibilmente bidimensionali in modo da interpretarli più facilmente. Per quanto concerne l'individuazione di pattern nei dati, nel presente lavoro di tesi la PCA consentirà di rappresentare vicini i soggetti con composizione simile del microbiota.
La PCA combina linearmente le variabili presenti nel dataset affinché si possa spiegare il massimo della varianza dei dati di partenza, utilizzando meno variabili. Le nuove variabili sono definite componenti principali. Come per la maggior parte delle tecniche di ordinamento è necessaria l'estrazione di autovalori ed auto vettori da una matrice, nel caso della PCA si considera quella di covarianza.
Si ipotizzi che i dati originali siano disposti in una matrice $\mathbf{X_{NxS}}$, in cui N è il numero di soggetti (sulle righe) ed S il numero di generi o OTU (sulle colonne) in ciascun soggetto. Gli elementi di **X** sono dapprima centrati sulle S colonne, ottenendo una matrice **Y** di egual dimensione:

$$y_{ki} = x_{ki} - \frac{1}{N}\sum_{k}^{N} x_{ki}$$

Si calcola in seguito la matrice **C**, matrice di covarianza dei dati originali, ottenuta moltiplicando **Y** per la sua trasposta **Y'** e dividendo il prodotto per il numero di soggetti N (operazione quest'ultima non strettamente necessaria):

$$\boldsymbol{C} = \frac{1}{N}\boldsymbol{Y'Y}$$

Si estraggono dunque gli autovalori λ_i (i=1,2,…,m) e gli autovettori u_{ki} (k=1,2,…,N; i=1,2,…m) della matrice **S**. Il numero m di autovalori ed auto vettori da estrarre può esser fissato a piacere e corrisponde al numero di componenti principali che si decide di considerare; in molti casi è

sufficiente considerare i primi due o tre autovalori, in ordine decrescente. La percentuale di varianza che la prima componente principale spiega è pari al rapporto tra il primo autovalore e la traccia della matrice **C** e così via.

I risultati della PCA sono poi rappresentati di solito mediante biplot (Jolicoeur P e Mosimann JE 1960, 339-354), in cui negli assi si riportano le prime componenti principali, quelle cioè che consentono di rappresentare la massima varianza dei dati. Nel nostro caso, in cui il dataset è rappresentato da soggetti per ciascuno dei quali sono misurate le abbondanze relative di diversi generi o OTU, i campioni sono rappresentati con dei punti e i generi con delle frecce (Figura 11). La direzione delle frecce indica la direzione della massima variazione dell'abbondanza dei generi (o delle OTU) e la loro lunghezza è in relazione con l'entità della variazione (frecce lunghe sono associate a quei generi che maggiormente contribuiscono alla variazione nel dataset). Inoltre la proiezione di un punto su una freccia indica l'abbondanza di quel genere (o OTU) all'interno del campione.

Bisogna inoltre precisare che il grafico ottenuto dalla PCA può essere interpretato in due modi diversi, a seconda che il metodo sia stato applicato sulla varianza-covarianza tra i dati oppure sulla correlazione tra di essi. Allo stesso modo sono possibili due differenti interpretazioni della relazione tra campioni e generi (o OTU) a seconda che ci si concentri sulle relazioni tra soggetti o su quelle tra generi (o OTU).

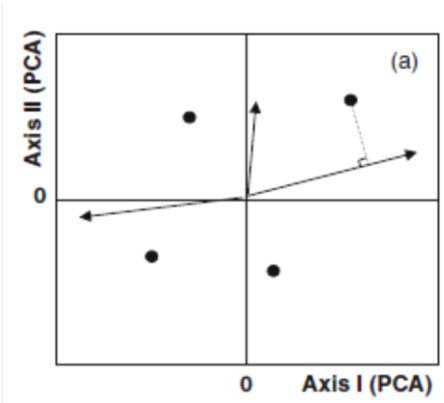


Figura 11. *Diagramma di ordinamento in due dimensioni (biplot). I punti rappresentano i soggetti, le frecce sono i generi (o le OTU).*

I risultati della PCA risultano essere interpretabili più facilmente nel momento in cui gran parte della varianza dei dati sia rappresentata dalle prime (solitamente due o tre) componenti principali.

Un altro metodo di ordinamento non vincolato è la PCoA (*Principal Coordinate Analysis*) che così come la PCA consente di spiegare la maggior parte della varianza dei dati utilizzando meno variabili di quelle originarie. A differenza della PCA però, non si calcolano combinazioni lineari delle variabili di partenza, bensì si ricorre a funzioni complesse dipendenti dal tipo di misura scelta per valutare la dissimilarità tra dati. La PCoA inoltre si applica alle distanze misurate tra coppie di dati, e non alla varianza o alla correlazione tra i dati.
Si calcola infatti dapprima la matrice $\mathbf{D_{NxN}}$ delle distanze tra gli N soggetti e viene trasformata nella matrice **Δ**:

$$\mathbf{\Delta} = -\frac{1}{2}\mathbf{D}$$

Successivamente si centra la matrice **Δ** in modo tale che l'origine del sistema di assi che sarà definito con le nuove variabili si trovi nel centroide degli oggetti; si costruisce così la matrice **P**:

$$c_{ki} = \delta_{ki} - \frac{1}{N}\sum_{p=1}^{N}\delta_{kp} - \frac{1}{N}\sum_{q=1}^{N}\delta_{qi} - \frac{1}{N^2}\sum_{p=1}^{N}\sum_{q=1}^{N}\delta_{pq}$$

in cui nel secondo e terzo termine si identificano le medie di riga e di colonna della matrice Δ, mentre l'ultimo termine rappresenta la media generale di questa stessa matrice.
A questo punto si calcolano gli autovalori λ_i (i=1,2,…,m; $m \leq N-1$) e gli autovettori u_{ki} (k=1,2,…,N; i=1,2,…m) della matrice **P**. Le Coordinate Principali f_{ki} degli oggetti si ottengono moltiplicando gli auto vettori per la radice quadrata dell'autovalore corrispondente:

$$f_{ki} = \sqrt{\lambda_i}\, u_{ki}$$

Come nell'ambito della PCA, è possibile valutare la qualità dell'ordinamento ottenuto per ciascun asse principale sulla base del rapporto tra autovalore corrispondente e la traccia della matrice **P**.
Tuttavia, poichè è possibile che uno o più autovalori siano negativi, Cailliez & Pagès (1976) raccomandano di valutare globalmente la qualità di un ordinamento utilizzando il rapporto:

$$\frac{\sum_{k=1}^{t}\lambda_k + t|\lambda_{min}|}{\sum_{k=1}^{N-1}\lambda_k + (N-1)|\lambda_{min}|}$$

dove t è il numero delle dimensioni dell'ordinamento, N è il numero totali di dimensioni e λ_{min} è l'autovalore negativo di maggior valore assoluto.

Anche in questo caso è possibile costruire un diagramma di ordinamento nel quale i campioni sono rappresentati mediante dei punti e gli assi sono associati alle nuove variabili calcolate dalla PCoA.

Un'ulteriore tecnica di ordinamento non vincolato è l'Analisi delle Corrispondenze (Benzecri et al., 1973), con la quale è possibile rappresentare simultaneamente i punti-soggetto ed i punti-specie, utilizzando coordinate tali da rendere massima la correlazione tra i due insiemi per ogni fattore. Inoltre, il risultato globale non cambia se, ad esempio, le osservazioni relative a due entità tassonomiche la cui separazione è dubbia vengono cumulate o mantenute separate. Allo stesso modo, se un'osservazione è replicata con risultati coerenti, può indifferentemente essere cumulata alla precedente o trattata come una nuova osservazione.

L'Analisi delle Corrispondenze consta essenzialmente di tre fasi: il calcolo di una matrice simmetrica di prodotti scalari, l'estrazione di autovettori ed autovalori della matrice, il calcolo delle coordinate e dei contributi assoluti (contributi delle osservazioni e delle variabili agli assi fattoriali) e relativi (contributi degli assi fattoriali alla descrizione di osservazioni e variabili).

Per ottimizzare le procedure di calcolo, la matrice dei dati $\mathbf{X_{NxS}}$ sarà disposta in modo tale che S≤N; questo però non si riscontra nel caso in cui (come nel nostro studio) si trattino delle liste di specie osservate in un insieme di siti. In tale occasione infatti, il numero di specie (sulle colonne) è superiore a quello delle osservazioni (sulle righe).

Operativamente, la matrice **X** viene trasformata nella matrice **U** dove

$$u_{ki} = \frac{x_{ki}}{\sqrt{x_{k.}x_{.i}}} - \frac{\sqrt{x_{k.}x_{.i}}}{x_{..}}$$

La matrice **U** è dunque composta dagli scarti degli elementi di **X** pesati sulla media geometrica delle somme marginali di riga e di colonna, rispetto alla stessa media geometrica pesata sul totale generale. A partire da **U** si calcola quindi la matrice dei prodotti scalari **S**, di rango ρ, moltiplicando la matrice **U** per la sua trasposta **U'**:

$$\boldsymbol{S = U'U}$$

Si calcolano a questo punto gli autovalori λ_i (i=1,2,…,m; $m \leq S-1$) e gli autovettori v_{kh} (k=1,2,…,S; h=1,2,…m) della matrice **P**. Solitamente è sufficiente calcolare i primi 2 o 3 autovalori e autovettori ai fini dell'analisi.

Si procede quindi con il calcolo delle coordinate delle osservazioni:

$$f_{kh} = \sum_{i=1}^{S} \frac{x_{ki} v_{ih}}{x_{k.} \sqrt{\frac{x_{.i}}{x_{..}}}}$$

Per gli *h* assi fattoriali richiesti. Successivamente si calcolano le coordinate che le variabili assumono nello spazio creato dai precedenti assi:

$$g_{ih} = \sum_{k=1}^{N} \frac{x_{ki} f_{kh}}{x_{.i} \sqrt{\lambda_h}}$$

In seguito, l'ultimo step prevede che si calcolino i contributi assoluti dell'*h*-esimo fattore da parte della *k*-ma osservazione e della *i*-ma variabile

$$ca(f_{kh}) = f_{kh}^2 \frac{x_{k.}}{x_{..} \lambda_h}$$

$$ca(g_{ih}) = g_{ih}^2 \frac{x_{.i}}{x_{..} \lambda_h}$$

Infine, si ricavano i contributi relativi dell'*h*-mo fattore alla *k*-ma osservazione ed all'*i*-ma variabile

$$cr(f_{kh}) = \frac{f_{kh}^2}{\sum_{h=1}^{m} f_{kh}^2}$$

$$cr(g_{ih}) = \frac{g_{ih}^2}{\sum_{h=1}^{m} g_{ih}^2}$$

Per valutare la significatività degli assi fattoriali esistono diversi metodi, il più semplice dei quali effettua il confronto della percentuale di varianza spiegata da ciascuno di essi con quella attesa in base al modello di Mac Arthur.
In figura 12 si mostra un tipico diagramma di ordinamento ottenuto mediante CA.

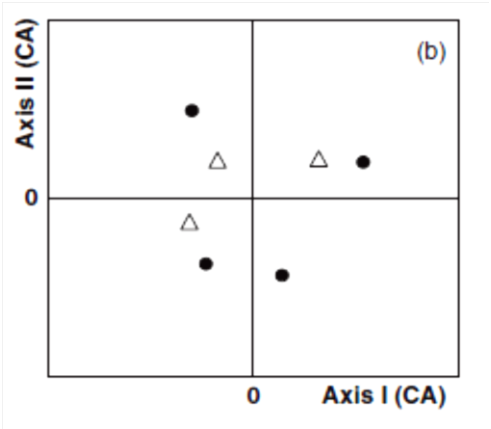


***Figura 12**. Diagramma di ordinamento ottenuto con Analisi delle Corrispondenze. I punti rappresentano sia i generi (o le OTU) sia i soggetti. La distanza tra soggetti e generi indica la probabilità della presenza dei generi nei campioni.*

L'ultima tecnica di ordinamento non vincolata qui esposta è il NMDS (*Nonmetric Multidimensional Scaling*), il quale si basa sulla distanza tra dati. L'algoritmo ordina dapprima le distanze tra oggetti e li riproduce su uno spazio di ordinamento in due dimensioni, evidenziando le distanze ordinate. NMDS differisce dai metodi finora visti per diversi motivi; innanzitutto non si calcolano molti assi per poi rappresentarne solo alcuni, bensì si sceglie a priori il loro numero ed i risultati si adattano allo spazio creato. Inoltre NMDS è un metodo iterativo e quindi si continuano a produrre risultati fino a che non si ottiene una soluzione accettabile o non si raggiunge un numero prefissato di iterazioni; infine NMDS non si basa sul calcolo di autovalori e autovettori.

A livello operativo, come primo passo si calcola la distanza tra tutte le possibili coppie di campioni, scegliendo una qualsiasi misura di distanza; a partire dunque da una matrice $\mathbf{X}_{\mathbf{NxS}}$ se ne ottiene una simmetrica di dimensioni NxN. Dopo aver scelto le *m* dimensioni su cui svolgere l'ordinamento, si costruisce in modo casuale o basandosi sul risultato di un precedente ordinamento la configurazione iniziale dei campioni nello spazio. Successivamente si calcola la distanza (solitamente Euclidea) tra ogni coppia di campioni (distanze basate sull'ordinamento) e se ne esegue la regressione sulle distanze originali, applicando il metodo dei minimi quadrati; si calcolano dunque le distanze previste dal modello adottato. Successivamente, per valutare la bontà di adattamento del modello di regressione, si ricorre alla somma dei quadrati delle differenze tra le distanze basate sull'ordinamento e quelle predette dalla regressione. Di solito, a tal fine si utilizza la funzione di stress di Kruskal:

$$Stress(1) = \sqrt{\sum_{k,j} (d_{kj} - \hat{d}_{kj})^2 \Big/ \sum_{k,j} d_{kj}^2}$$

in cui d_{kj} è la distanza tra il campione k ed il campione j nella loro configurazione generata dall'ordinamento, mentre è la distanza predetta dalla regressione.

Dopo aver calcolato la funzione di stress, si riorganizzano i campioni nello spazio di ordinamento al fine di ottenere un valore più basso della suddetta funzione; si ricalcolano le distanze basate sull'ordinamento, si esegue di nuovo la regressione e si ricalcola la funzione di stress. Tale procedimento viene ripetuto più volte, fino a che la funzione non assume valori più bassi (si è dunque raggiunto un minimo, eventualmente anche locale), oppure fino a che non si raggiunge un numero di iterazioni prestabilito.

Nella rappresentazione grafica dei risultati di NMDS (Figura 13), è possibile riportare sia i soggetti sia i generi (o le OTU) mediante punti; la distanza tra soggetti indica la loro similarità e non corrisponde alle originali distanze calcolate a partire dai dati originali. Inoltre è possibile invertire, riscalare e ruotare gli assi del grafico per una migliore visualizzazione o interpretazione senza però inficiare il risultato dell'ordinamento.

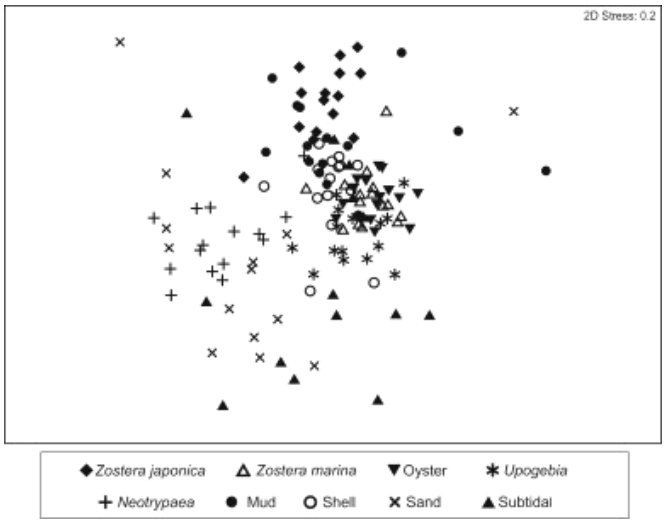


Figura 13. *Esempio di grafico ottenuto con NMDS applicato a distanze di Bray-Curtis tra campioni di macrofauna bentonica provenienti da nove siti a Grays Harbor.*

Tra i metodi di ordinamento vincolati invece, si annoverano la CCA (*Canonical Correspondence Analysis*) e la RDA (*Redundancy Analysis*).

La CCA esamina le correlazioni tra due sottoinsiemi di variabili, quali i generi (o OTU) e le misure dei principali parametri fisico-chimici relative ad un insieme di osservazioni distribuite

nello spazio e/o nel tempo. La matrice dei dati analizzata dall'Analisi delle Correlazioni Canoniche può dunque essere vista come l'insieme delle *N* osservazioni relative a due sottoinsiemi composti rispettivamente da *S* e *E* variabili, con $S \leq E$. Perciò la *k*-ma osservazione può essere rappresentata da due vettori riga **w** e **z**

$$\boldsymbol{w} = (w_{k1}\ w_{k2} \ldots w_{kS})$$
$$\boldsymbol{z} = (z_{k1}\ z_{k2} \ldots z_{kE})$$

in cui **w** rappresenta il sottoinsieme di variabili meno numeroso.

La matrice di covarianza **S** di rango *S*+*E* dell'insieme completo dei dati è perciò ripartibile in due blocchi:

$$\boldsymbol{S} = \frac{1}{N}\begin{pmatrix}\boldsymbol{w}\\ \boldsymbol{z}\end{pmatrix}(\boldsymbol{w}'\ \boldsymbol{z}') = \begin{pmatrix}\boldsymbol{S_{11}} & \boldsymbol{S_{12}}\\ \boldsymbol{S_{21}} & \boldsymbol{S_{22}}\end{pmatrix}$$

in cui $\mathbf{S}_{11}$ è la matrice di rango *S* di covarianza delle variabili del sottoinsieme **w** mentre $\mathbf{S}_{22}$ di rango *E* lo è del sottoinsieme **z**. La $\mathbf{S}_{12}$ è una matrice *SxE* che contiene le covarianze fra i due sottoinsiemi di variabili. Poiché **S** è una matrice simmetrica, $\mathbf{S}_{21}$ è la trasposta di $\mathbf{S}_{12}$.

La CCA, a partire dalla matrice $\boldsymbol{S}$, si pone l'obiettivo di trovare le *S* combinazioni lineari delle variabili **w** e le *S* combinazioni lineari delle variabili **z**

$$u_j = c_{j1}w_1 + c_{j2}w_2 + \cdots + c_{jS}w_S$$
$$v_j = d_{j1}z_1 + d_{j2}z_2 + \cdots + d_{jS}z_S$$

tali da soddisfare le condizioni seguenti:

1. tutte le u_j devono essere indipendenti fra loro;
2. tutte le v_j devono essere indipendenti fra loro;
3. le *p* coppie di combinazioni lineari devono essere tali da rendere massime le *S* correlazioni r_j fra le u_i e le v_i

Si identificano così *u* e *v* come variabili canoniche, mentre le loro correlazioni *r* sono definite correlazioni canoniche.

Il primo step dell'analisi consiste nel calcolo degli autovalori delle due matrici ottenute dai prodotti

$$\boldsymbol{S_{22}^{-1}S_{12}'S_{11}^{-1}S_{12}}$$
$$\boldsymbol{S_{11}^{-1}S_{12}S_{22}^{-1}S_{12}'}$$

Esistono al massimo *S* autovalori non nulli della prima matrice prodotto: tali autovalori sono uguali a quelli non nulli della seconda matrice prodotto. Per ottenere poi i vettori dei coefficienti **c** e **d** si risolvono i due sistemi, di *S* ed *E* equazioni lineari rispettivamente

$$\left(\boldsymbol{S}_{11}^{-1}\boldsymbol{S}_{12}\boldsymbol{S}_{22}^{-1}\boldsymbol{S}_{12}' - \lambda_j\boldsymbol{I}\right)\boldsymbol{c} = 0$$
$$\left(\boldsymbol{S}_{22}^{-1}\boldsymbol{S}_{12}'\boldsymbol{S}_{11}^{-1}\boldsymbol{S}_{12} - \lambda_j\boldsymbol{I}\right)\boldsymbol{d} = 0$$

per ogni λ_j (j=1,2,…,S). Per semplicità si pone

$$c_{j1} = 1 \qquad d_{j1} = 1$$

Le variabili canoniche si ricavano successivamente effettuando un prodotto fra vettori

$$\boldsymbol{u}_j = \boldsymbol{c}'\boldsymbol{w} \quad \boldsymbol{v}_j = \boldsymbol{d}'\boldsymbol{z}$$

Per calcolare la correlazione canonica per ciascuna coppia u_j e v_j

$$r_j = \sqrt{\lambda_j}$$

La correlazione canonica più alta corrisponde al primo autovalore estratto e sulla base di questo è possibile effettuare un test di indipendenza tra i due sottoinsiemi di variabili.
Le variabili canoniche invece possono essere utilizzate per ulteriori analisi, così come possono essere impiegate per un output grafico che mostra la correlazione tra i due sottoinsiemi di variabili e l'ordinamento delle osservazioni in tale ambito.

Un altro metodo di ordinamento vincolato è la RDA che tipicamente consente di determinare le variabili ambientali che permettono di spiegare la massima percentuale di varianza dei dati originali (che si intendono sempre come liste di generi o OTU presenti in diversi soggetti). La RDA può quindi considerarsi come un'estensione della PCA (Rao 1964, 329-358), dove gli assi principali sono in questo caso combinazioni lineari delle variabili ambientali. Inoltre la RDA permette di mostrare la correlazione tra ciascun genere (o OTU) e ciascuna variabile ambientale. Se il data set è composto da una matrice di distanze tra dati, si parla di RDA basata sulla distanza (Legendre & Anderson, 1999), e consente di stabilire quanto le variabili ambientali possano spiegare la variazione tra gli oggetti della matrice. A tal fine si applica innanzitutto una PCoA sulla matrice delle distanze, ottenendo così una matrice con i soggetti sulle righe e le

coordinate della PCoA sulle colonne. Ricorrendo poi ad una classica RDA si mettono in relazione le coordinate principali con le variabili ambientali.
I risultati della RDA possono riprodursi su un triplot in cui i soggetti si raffigurano con dei punti, i generi (o le OTU) si identificano con frecce, mentre le variabili ambientali si rappresentano anch'esse con frecce se si tratta di variabili quantitative, con punti se sono invece qualitative.

4.3 Analisi multivariate basate su test d'ipotesi

Un altro obiettivo nell'analisi dei dati multivariati, oltre che rappresentarli in un diagramma di ordinamento o raggrupparli in base alla loro similarità, è quello di valutare se sussistono differenze tra gruppi di soggetti in base ai valori di abbondanza dei generi (o OTU). Inoltre è di interesse capire se la differenza tra gruppi è maggiore o minore di quella misurata tra i soggetti di uno stesso gruppo.
A tal fine, i metodi utilizzati nel presente studio sono quattro: calcolo degli indici di biodiversità (si vedano paragrafi 4.4 e 4.5), NPMANOVA (si veda paragrafo 4.6), ANOSIM (si veda paragrafo 4.7) ed il test di Wilcoxon (si veda paragrafo 4.8).

4.4 Diversità alfa

Nel caso diversità alfa si annoverano 3 indicatori principali:

- Indice di *Shannon*:

$$H' = -\sum_{i=1}^{S} p_i \times ln\, p_i$$

in cui con p_i si identifica l'abbondanza relativa della specie *i*-esima

$$p_i = \frac{\#\ esemplari\ specie\ i}{\sum_{q=1}^{S} \#esemplari\ specie\ q}$$

mentre S è il numero totale delle specie.

Tale indice valuta l'incidenza quantitativa delle diverse specie all'interno del microbiota del soggetto considerato; il valore minimo assunto dall'indice è 0 ed indica che è presente un'unica specie. Il valore massimo possibile è $-\sum_{i=1}^{S}\frac{1}{S}ln\frac{1}{S}$ ed evidenzia che le abbondanze relative delle specie sono uguali e dunque la biodiversità è massima.

- Indice di *Evenness*:

$$E = \frac{H'}{H_{max}}$$

dove H_{max} è il valore massimo di diversità calcolata utilizzando l'indice di Shannon, nel caso in cui tutte le specie abbiano uguali abbondanze (*evenness* appunto), mentre H' è l'indice di Shannon. Tale indice rappresenta l'eterogeneità di distribuzione delle specie all'interno di un habitat; il suo valore minimo, 0, si calcola quando è presente un'unica specie, mentre il massimo, pari a 1, si ottiene quando tutte le specie sono rappresentate in egual proporzione.

- Indice di *Simpson*:

$$D = \sum_{i=1}^{S} {p_i}^2$$

Può essere calcolato in questo modo oppure nella sua formula complementare, 1-D, che assume valore 0 nel caso di una popolazione perfettamente omogenea e valore 1 nel caso di una popolazione eterogenea al massimo (ogni specie rappresentata nella stessa proporzione). Inoltre è nota un'ulteriore variante, definita come *inverse Simpson*, che come già suggerisce il nome è l'inverso dell'indice D (1/D).

4.5 Diversità beta

Per quanto concerne la biodiversità beta si trovano molteplici indici, dei quali 24 sono riportati nella tabella 2. Per interpretare chiaramente le formule, si osservi poi la figura 14, in cui con *b* e *c* si identifica il numero di specie appartenenti esclusivamente rispettivamente ad uno dei due distinti soggetti a confronto, mentre con *a* si identifica il numero di specie in comune.

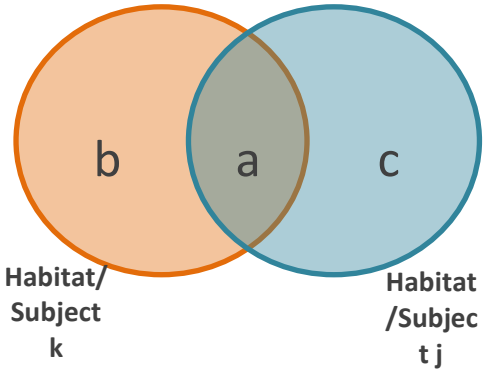


Figura 14. *Rappresentazione mediante insiemi delle specie comuni e non di due soggetti*

Indice di diversità beta	**Formulazione**	**Fonte**
1 β_w	$\frac{a+b+c}{(2a+b+c)/2}$ or $\frac{a+b+c}{(2a+b+c)/2}-1$	Whittaker (1960), Magurran (1988),Southwood & Henderson(2000)
2 β_{-1}	$\frac{a+b+c}{(2a+b+c)/2}-1$	Harrison *et al*. (1992)
3 β_c	$\frac{b+c}{2}$	Cody (1975)
4 β_{wb}	$b+c$	Weiher & Boylen (1994)
5 β_r	$\frac{(a+b+c)^2}{(a+b+c)^2-2bc}$ or $\frac{(a+b+c)^2}{(a+b+c)^2-2bc}-1$	Routledge (1977), Magurran (1988),Southwood & Henderson(2000)
6 β_{rw}	$\log(2a+b+c)-\left(\frac{1}{2a+b+c}2a\log 2\right)-\left(\frac{1}{2a+b+c}((a+b)\log(a+b)+(a+c)\log(a+c))\right)$	Routledge (1977), Wilson &Shmida (1984)

Indice di diversità beta	Formulazione	Fonte
7 β_e	$\exp(\beta_l) - 1$	Routledge (1977)
8 β_{ws}	$\frac{b + c}{2a + b + c}$	Wilson & Shmida (1984)
9 β_{me}	$\frac{b + c}{2a + b + c}$	Mourelle & Ezcurra (1997)
10 β_j	$\frac{a}{a + b + c}$	Jaccard (1912), Magurran (1988),Southwood & Henderson (2000)
11 β_{sor}	$\frac{2a}{2a + b + c}$	Sørensen (1948) basato su Dice (1945); vedi anche Whittaker (1975), Magurran (1988), Southwood & Henderson (2000)
12 β_m	$(2a + b + c)\left(1 - \frac{a}{a + b + c}\right)$	Magurran (1988)
13 β_{-2}	$\frac{\min(b,c)}{\max(b,c) + a}$	Harrison *et al.* (1992)[1]
14 β_{co}	$1 - \frac{a(2a + b + c)}{2(a + b)(a + c)}$	Cody (1993)
15 β_{col}	$\frac{b + c}{a + b + c}$	Colwell & Coddington (1994; 'complementarity' measure), see also Pielou (1984)

Indice di diversità beta	Formulazione	Fonte
16 β_g	$\frac{b+c}{a+b+c}$	Gaston *et al.* (2001)[2]
17 β_{-3}	$\frac{\min(b,c)}{a+b+c}$	Williams (1996a)
18 β_l	$\frac{b+c}{2}$	Lande (1996)
19 β_{wil}	$\frac{bc+1}{((a+b+c)^2-(a+b+c))/2}$	Williams (1996a),1999*Williams et al*. (1999)
20 β_{hk}	$1-\frac{2a}{2a+b+c}$	Harte & Kinzig (1997)[3]
21 β_{rlb}	$\frac{a}{a+c}$	Ruggiero *et al.* (1998)[4]
22 β_{sim}	$\frac{\min(b,c)}{\min(b,c)+a}$	Lennon *et al.* (2001), basato su Simpson (1943)
23 β_{gl}	$\frac{2\lvert b-c\rvert}{2a+b+c}$	Lennon *et al.* (2001)[5]
24 β_z From SAR	$1-\left[\log\left(\frac{2a+b+c}{a+b+c}\right)\Big/\log 2\right]$	Lennon *et al.* (2001), vedi anche Harte & Kinzig (1997) & Appendix

Tabella 2. *S = numero totale delle specie presenti in entrambi i soggetti; $\bar{\alpha}$ = media del numero di specie nei soggetti; α_1= numero totale delle specie in comune; α_2= numero totale delle specie presenti nei due soggetti ma non in comune; αj = numero totale di specie presenti nel soggetto j; α_{max} = Massimo valore di ricchezza delle specie per I due soggetti; N = numero di soggetti; r = numero delle coppie di specie le distribuzioni delle quali si sovrappongono; g = guadagno cumulativo nelle specie; l = perdita*

cumulativa nelle specie; H = range del gradiente del sito; e_i = numero di soggetti comparati nei quali si ritrova la specie i; T = Σe_i = Σaj; C = specie in comune tra due censimenti; Ti = numero totale di specie nel censimento i; r_s = numero di casi in cui non c'è sovrapposizione di specie nel confronto tra due soggetti); SAR = species–area relationship, S=kAz, dove S è il numero delle specie, A è l'area e z *e* k *sono costanti. Il parametro z è una misura di diversità beta basato sul guadagno delle specie.*

[1] $\beta_{-1} = \beta_{-2}$ quando $a = \alpha_{max}$.

[2] Originariamente formulato per cinque, quattro e tre soggetti a confronto, per cui β_g è la percentuale delle "specie di transizione" rispetto al numero totale di specie trovate nella sequenza degli altri soggetti posti a confronto. Le specie di transizione per due soli soggetti a confronto sono qui considerate come *b* e *c*.

[3] Dalla definizione di " turnover = 1 – comunanza (numero di specie in commune diviso per la media del numero di specie nei due soggetti)"

[4] Misura di similarità

[5] Non è da considerarsi come una misura di diversità beta *per se*; è stata originariamente utilizzata per rappresentare le differenze nella ricchezza di specie tra due siti.

Tali indici possono suddividersi in 4 categorie principali:

1. misure di continuità e perdita;
2. misure di gradienti di ricchezza delle specie;
3. misure di continuità;
4. misure di guadagno e perdita.

Gli indici appartenenti alla categoria 1 assumono il valore zero nel momento in cui non sono presenti specie in comune tra i due soggetti (massima diversità), mentre il valore uno corrisponde al caso contrario. Tra gli indici riportati in tabella 1, β_{rlb} appartiene a questo gruppo.

Nella categoria 2 si trovano invece gli indici che dipendono dalla differenza nella ricchezza delle specie tra i due soggetti considerati; si osserva dunque il massimo valore nel momento in cui non ci sono specie in comune ed il numero di specie di un soggetto equivale a quello dell'altro soggetto. β_{gl} appartiene a tale categoria.

Gli indici che invece appartengono alla categoria delle misure di continuità dipendono dal numero di specie in comune tra i siti (qui indicata con *a*). Si osservano dunque valori agli estremi del possibile range (massimo o minimo a seconda dell'entità di *a*) nel caso di grande differenza nella ricchezza delle specie tra i soggetti. All'interno della categoria 3 si individuano inoltre due ulteriori suddivisioni: gli indici che incrementano il loro valore al crescere di *a* e quelli che invece decrescono al crescere di *a* . Nel primo gruppo si trovano gli

indici β_j e β_{sor} i quali si definiscono anche indici di similarità in quanto essi assumono valori bassi in occasione di bassa diversità (e quindi alta similarità). Gli indici β_c, β_w, β_{hk}, β_t, β_m *e* β_z *sono invece appartenenti al secondo gruppo.*

Infine gli indici inseriti nella categoria 4 dipendono sia da a, *sia dalla grandezza relativa di* b *e* c*; rientrano nel gruppo gli indici* β_{co}, β_r, β_l, β_e, β_{rs}, β_{-2}, β_{-3} *e* β_{sim}. *I valori assunti da essi aumentano al decrescere del numero di specie in comune tra i soggetti ma giungono il valore massimo in caso di valori intermedi di* b' *e* c', *intesi come percentuale delle specie presenti esclusivamente nel primo e nel secondo soggetto considerato.*

Tra gli indici presentati si riportano di seguito i tre che sono stati utilizzati nel nostro studio:

- Indice di *Sørensen* o di *Bray-Curtis* (Whittaker 1960, 279-338)

$$\beta_{sor} = \frac{b + c}{2a + b + c}$$

in cui b e c sono il numero di specie appartenenti esclusivamente ad uno dei due distinti soggetti a confronto rispettivamente, mentre con a si identificano il numero di specie in comune (figura 14). L'indice assume valori nel range [0,1] e assume dunque il valor minimo nel caso in cui l'uno e l'altro sito siano perfettamente uguali per quanto riguarda le specie presenti, mentre il massimo della diversità (quando perciò non si hanno specie comuni) è rappresentato dal valore massimo dell'indice.

- Indice di *Jaccard* (Jaccard 2006, 37-50)

$$\beta_{jacc} = \frac{a}{a + b + c}$$

in cui le lettere hanno il medesimo significato presentato per l'indice precedente. Anche in questo caso l'indice assume valori compresi tra 0 e 1, ma con significato opposto all'indice di Sørensen; infatti l'indice di Jaccard è pari a 0 nel caso in cui non ci sono specie in comune e dunque si assiste alla massima diversità, mentre con l'indice pari ad 1 si è in presenza di due soggetti identici per quanto concerne la composizione.

- Indice di *Harte e Kinzig* (Harte and Kinzig 1997, 417-427)

$$\beta_{hk} = 1 - \frac{2a}{2a + b + c}$$

il quale quantifica il grado di diversità con valori da 0 (uguaglianza tra siti) a 1 (massima differenza).

4.6 NPMANOVA: analisi multivariata non-parametrica della varianza

E' di interesse comune di fronte a dataset multivariati ridurne la dimensione, oltre che evidenziare le variabili importanti ed individuarne relazioni e comportamenti. A tal fine si ricorre a particolari test d'ipotesi, tra i quali uno dei più efficaci è la versione non parametrica multivariata dell'analisi della varianza, definita NPMANOVA (*Non Parametric Multivariate ANalysis Of VAriance*). Questo metodo è stato presentato da Anderson (Anderson, 2001) come alternativa alle altre tecniche di analisi multivariata tradizionali, le quali si basano su assunzioni troppo rigide per poter essere applicate ai data set ecologici o ai dati del microbiota. Nello specifico NPMANOVA utilizza una qualsiasi misura di distanza o dissimilarità tra i dati per costruire una statistica che risulta essere la versione non parametrica della statistica F di Fisher (1890-1962). NPMANOVA è dunque l'analogo non parametrico di MANOVA (*ANalysis Of VAriance*).

L'analisi della varianza MANOVA consente di testare la significatività delle differenze tra medie aritmetiche di più gruppi, basandosi sulla varianza dei dati e sulla statistica F, nota anche come distribuzione di Fisher. Fu lo stesso Fisher che nel 1925 (nel *Statistical Methods for Research Workers*) suggerì il metodo per individuare i fattori che sono responsabili della varianza dei dati distribuiti in due o più gruppi e quantificare l'incidenza dei fattori stessi su tale varianza.

Le analisi di tipo MANOVA considerano il valore di F per decidere se accettare l'ipotesi nulla H_0 o l'ipotesi alternativa; la prima afferma che le medie di tutti i gruppi in esame sono uguali, e dunque non sussistono differenze significative tra i gruppi; mentre con l'ipotesi alternativa si conclude che le medie non sono tutte uguali (o almeno una è diversa dalle altre). Fissata una soglia di significatività, se il valore di F è inferiore ad essa, allora viene accettata l'ipotesi nulla; se il valore è invece superiore alla soglia si opta per l'ipotesi alternativa. Nonostante MANOVA e NPMANOVA presentino delle affinità a livello della definizione del test

statistico adottato, esse differiscono per le condizioni di applicabilità ai dati e per la misura di distanza utilizzate. Per quanto riguarda le assunzioni, infatti, MANOVA richiede che i dati siano distribuiti in modo normale, situazione non molto realistica nel caso di dataset biologici ed ecologici. Solitamente infatti le abbondanze relative delle specie presenti in un determinato sito presentano distribuzioni asimmetriche (Gaston et al. 1994, 335-358). Un'ulteriore assunzione dei tradizionali metodi multivariati basati su test statistici è che le distribuzioni delle abbondanze assumano valori continui, mentre si assiste di norma a valori discreti. Inoltre nei test tradizionali si incontrano difficoltà pratiche nel calcolo delle statistiche nel momento in cui le variabili siano in numero superiore rispetto ai campioni. NPMANOVA ed i metodi non parametrici invece risultano più flessibili in quanto sono applicabili anche a dati non modellabili con una distribuzione normale. Le uniche assunzioni della NPMANOVA sono l'indipendenza delle osservazioni e la similarità nella dispersione multivariata dei dati. Inoltre, a differenza di MANOVA, che è basato su distanza Euclidea, NPMANOVA si può avvalere di una qualsiasi misura della distanza tra coppie di osservazioni multivariate (usata per il calcolo delle varianze).

A proposito della varianza, è da evidenziarsi una differenza significativa tra MANOVA e NPMANOVA. Si supponga innanzitutto che i dati multivariati siano suddivisi in diversi gruppi; MANOVA assume che la varianza di ciascuna variabile associata ai dati (dispersione dei valori delle variabili all'interno del gruppo) rimanga uguale in tutti i gruppi e che sia lo stesso per la correlazione tra le variabili nei diversi gruppi.

In figura 15 si mostra un caso di dati descritti da due variabili e suddivisi in due gruppi. Nella figura 15a si evidenziano due gruppi in cui le variabili hanno la medesima varianza, ma correlazione diversa; nella figura 15b invece le variabili dei due gruppi hanno la stessa correlazione ma dispersione diversa all'interno dei gruppi (varianza diversa).

MANOVA è dunque sensibile sia alle differenze nella correlazione tra variabili, sia a differenze nella varianza; NPMANOVA invece risulta essere sensibile alle sole differenze nella varianza.

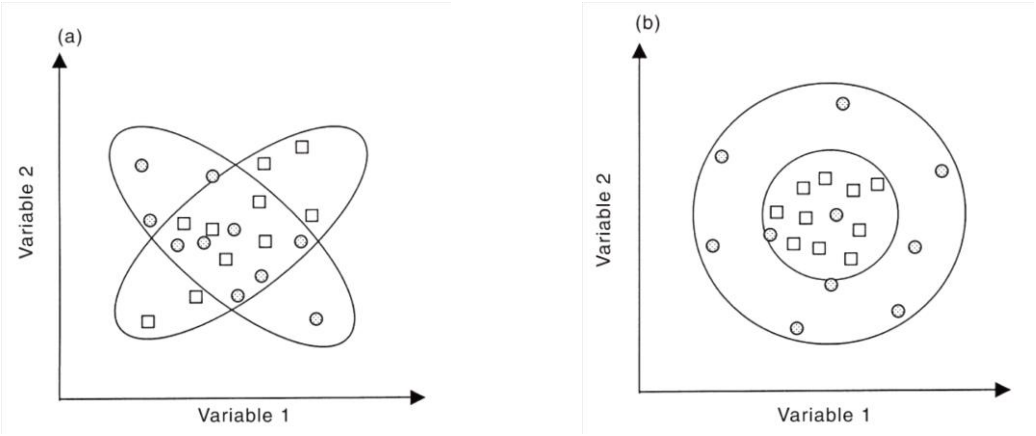


Figura 15. *Dati bivariati organizzati in due gruppi. Nella figura 18a le variabili hanno medesima varianza ma diversa correlazione; nella figura 18b la correlazione tra variabili è la stessa nei due gruppi mentre la varianza è diversa.*

NPMANOVA, come anticipato, parte dal calcolo della distanza tra i dati per giungere all'analisi della loro varianza; il metodo deve infatti confrontare la variabilità interna ai gruppi con quella presenta tra i diversi gruppi. Innanzitutto si calcolano quindi le distanze tra ogni coppia di soggetti e da queste si ricavano le varianze tra (SS_A) ed entro (SS_W) i gruppi e la varianza totale (SS_T). Si supponga che i dati siano inseriti in una matrice, le cui righe presentano i soggetti o gli habitat mentre le colonne specificano le specie ivi presenti e si supponga inoltre che i soggetti siano suddivisi in vari gruppi. Si assume inoltre a titolo di esempio che i gruppi siano *g* e che in ciascun gruppo siano presenti *n* soggetti, per un totale di $N=ng$ soggetti. Si precisa inoltre che il fatto di considerare gruppi formati dal medesimo numero di soggetti è un caso particolare, adottato per semplificare la spiegazione.

Per il calcolo delle varianze si utilizzano le seguenti formule:

$$SS_T = \frac{1}{N}\sum_{k=1}^{N-1}\sum_{j=k+1}^{N} d_{jk}^2$$

in cui con d_{jk} si indica la distanza tra soggetto k=1,…,N e soggetto j=1,…,N.

$$SS_W = \frac{1}{n}\sum_{k=1}^{N-1}\sum_{j=k+1}^{N} d_{jk}^2\,\epsilon_{jk}$$

in cui ϵ_{jk} è posto a 1 quando le osservazioni k e j appartengono allo stesso gruppo, mentre è pari a 0 in caso contrario.

Si può così calcolare il valore della varianza tra gruppi:

$$SS_A = SS_T - SS_W$$

Dopo aver calcolato le varianze si può così ricavare il valore della statistica F:

$$F = \frac{SS_A/(g-1)}{SS_W/(N-g)}$$

In figura 16 si riassume schematicamente il procedimento che porta al calcolo della statistica F, a partire dalla matrice delle distanze tra i dati.

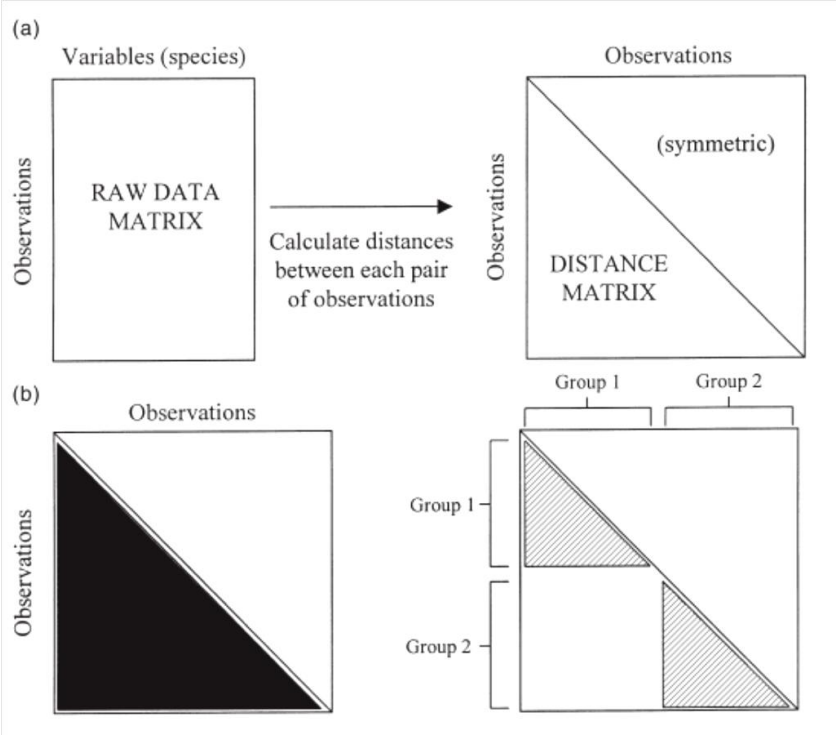


Figura 16. *Nella parte a**)** si mostra il passaggio dalla matrice contenente i soggetti sulle righe e le variabili sulle colonne alla matrice simmetrica contenente le distanze tra le possibili coppie di dati. Nella parte b) si evidenzia invece la varianza totale SS_T (■) , calcolata considerando la metà matrice evidenziata e dividendola per N. Si mostra poi la varianza entro i gruppi SS_W (immagine a destra), ottenuta considerando le parti evidenziate nella matrice e dividendole per n.*

Si nota a questo punto come sia possibile calcolare SS_w e SS_A a partire dalle distanze tra i dati; si tratta di una fondamentale differenza rispetto a MANOVA. Nel metodo tradizionale infatti SS_W è calcolata come somma dei quadrati delle distanze tra dati e media del gruppo al quale

appartengono (centroide del gruppo), mentre SS_A si ottiene calcolando la somma dei quadrati delle distanze tra medie dei gruppi e media totale. In figura 17 si pongono in evidenza le distanze appena esposte, nel caso di dati bidimensionali distribuiti in due gruppi.

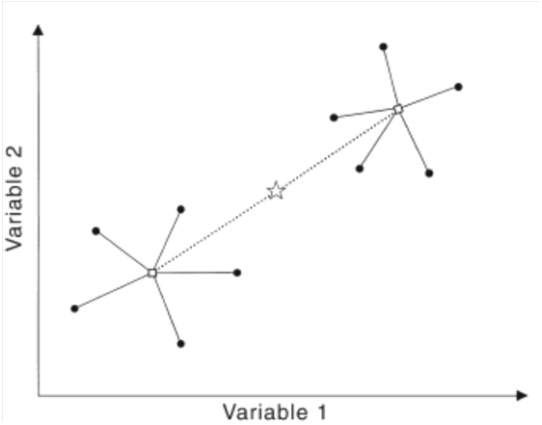


Figura 17. *Distanze utilizzate in MANOVA. (—)Distanze tra punti e centroide del gruppo; (...)distanze tra centroidi dei gruppi e centroide generale; ( )centroide generale, (□)centroide del gruppo, (●) osservazione.*

NPMANOVA invece presenta un metodo alternativo per il calcolo di SS_W e SS_A, in quanto la somma dei quadrati delle distanze tra dati e centroide del gruppo si calcola come somma dei quadrati delle distanze tra dati, divisa per il numero degli stessi (si veda figura 18, nella quale si riporta un caso in due dimensioni). In questo modo si può calcolare la somma di quadrati entro e tra gruppi utilizzando esclusivamente una qualsiasi misura di distanza; si evita dunque di calcolare i centroidi dei gruppi di dati, che nella maggior parte dei casi può risultare problematico. Solamente nel caso di misura Euclidea di distanza, i centroidi dei gruppi possono ricavarsi in modo semplice, calcolando la media di ciascuna variabile su tutte le osservazioni appartenenti al medesimo gruppo.

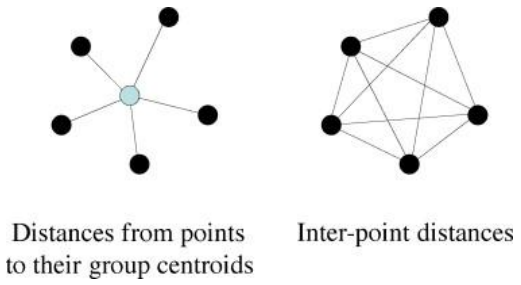


Figura 18. *Nell'ambito di NPMANOVA la somma dei quadrati delle distanze dei dati dal loro centroide (figura a sinistra) è uguale alla somma dei quadrati delle distanze tra i dati (figura di destra), divisa per il numero dei dati stessi.*

Una volta calcolato il valore della statistica F a partire dai dati a disposizione, lo si confronta con la distribuzione della statistica stessa in ipotesi nulla; si ricorda però che i dati non sono distribuiti in modo normale e dunque la statistica F definita secondo questo approccio non presenta la stessa distribuzione su cui si basa il test di Fisher. Si parla perciò di un valore pseudo-F, la cui distribuzione sotto ipotesi nulla si costruisce mediante permutazioni.
Sotto l'ipotesi nulla, infatti, non sono presenti differenze significative tra le medie dei vari gruppi: si può dunque ragionevolmente pensare che, se si scambiassero i soggetti tra i diversi gruppi e si applicasse nuovamente il test F, il risultato non cambierebbe. E' su questo ragionamento che viene costruita la pseudo-statistica F nel caso della NPMANOVA: si scambiano le righe della matrice di partenza (assegnando quindi i soggetti ai diversi gruppi in maniera random) e si calcola ogni volta un nuovo valore di F, indicato con F^{π}. Successivamente il p-value per il valore di F osservato a partire dai dati originali (non permutati) si calcola mediante la seguente formula:

$$P = \frac{(No.di\ F^{\pi} \geq F)}{(No.totale\ di\ F^{\pi})}$$

A questo punto si confronta il valore del P-value con la soglia di significatività prestabilita (solitamente 0.05) e si conclude quale ipotesi scegliere. Se l'ipotesi nulla viene rifiutata, ovvero qualora sussistano delle differenze tra le medie aritmetiche dei vari gruppi, è possibile individuare dove risieda tale differenza utilizzando dei dei test a posteriori. In pratica, si effettuano confronti tra coppie di gruppi e si ricava il p-value con la stessa tecnica delle permutazioni descritta in precedenza, con l'eccezione che esse vengono applicate ai soli dati dei due gruppi di volta in volta comparati.

4.7 ANOSIM: analisi delle similarità

ANOSIM (ANalysis Of SIMilarity), presentato da Clarke (K. R. Clarke, 1993), è anch'esso un metodo per analizzare dati multivariati e ricorre come la NPMANOVA ad un test statistico di verifica d'ipotesi. La presente tecnica non è però imperniata sull'analisi della varianza, bensì su quella della similarità.
Clarke et al. (reference) suggeriscono di effettuare una prima analisi per identificare quali specie caratterizzano la similarità tra soggetti appartenenti allo stesso gruppo o a

gruppi diversi, contribuendo in modo più o meno significativo alla separazione dei gruppi.

Per avere una misura della dissimilarità dei soggetti basata sui dati di abbondanza delle specie, è possibile usare la distanza di Bray-Curtis (Bray and Curtis 1957, 325-349):

$$d_{kj} = \sum_{i=1}^{S} d_{kj}(i)$$

in cui S è il numero totale delle specie, d_{kj} indica la distanza tra il soggetto k-esimo e quello j-esimo, mentre $d_{kj}(i)$ rappresenta il contributo che la specie i-esima apporta alla differenza tra i due soggetti confrontati. In particolare, tale contributo si definisce così:

$$d_{kj}(i) = \frac{100\,|y_{ij} - y_{ik}|}{\sum_{i=1}^{S}(y_{ij} + y_{ik})}$$

dove y_{ij} identifica l'abbondanza della specie i-esima nel soggetto j-esimo e S rappresenta invece il numero totale delle specie presenti.

Si ipotizzi ora per semplicità espositiva che i soggetti siano suddivisi in due gruppi, ciascuno di numerosità *n*; calcolando la media delle distanze d_{kj} ottenute considerando tutte le possibili coppie di soggetti (k,j) con k appartenente al primo gruppo e j al secondo, si ricava la dissimilarità media d_m tra i due gruppi

$$d_m = \frac{1}{n^2} \sum_{k=1}^{N-1} \sum_{j=k+1}^{N} d_{kj}\,\epsilon_{kj}$$

In cui ϵ_{kj} assume valore 1 se i soggetti k e j appartengono a gruppi diversi, mentre è pari a 0 in caso contrario.

Considerando poi $d_{kj}(i)$ e calcolandone la media su tutte le possibili coppie di soggetti (k,j), di cui uno appartiene al primo gruppo e l'altro al secondo, si ottiene d_{mi} il contributo medio che la specie i-esima fornisce alla dissimilarità media d_m.

$$d_{mi} = \frac{1}{n^2} \sum_{k=1}^{N-1} \sum_{j=k+1}^{N} d_{kj}(i)\,\epsilon_{kj}$$

Ad ogni specie si può dunque associare la misura di quanto essa sia influente sulla dissimilarità misurata tra i gruppi di soggetti.

E' possibile inoltre calcolare la standard deviation $SD(d_i)$ dei valori $d_{kj}(i)$. Se d_{mi} è grande e $SD(d_i)$ è piccola, significa che la specie i-esima non solo contribuisce in modo importante alla dissimilarità tra i due gruppi considerati, ma lo fa in modo consistente.

ANOSIM consente inoltre di testare le differenze tra soggetti, in un approccio simile a quello adottato in NPMANOVA, ovvero di testare l'ipotesi nulla "nessuna differenza tra i soggetti appartenenti ai due diversi gruppi testati". Il metodo, infatti, costruisce una statistica considerando la distanza media tra campioni entro lo stesso gruppo (r_W) e la distanza media tra soggetti di gruppi diversi (r_A). Così come in NPMANOVA si utilizzano le varianze entro e tra gruppi per costruire la statistica F, in ANOSIM si usano le distanze r_W e r_A per costruire la statistica R:

$$R = \frac{r_A - r_W}{\left(\frac{M}{2}\right)}$$

con M= N(N-1)/2, in cui N è il numero di soggetti totali. R assume solitamente valori compresi tra 0 e 1, con il minimo ad indicare l'assenza di diversità tra gruppi (ipotesi nulla del test) ed il massimo ad indicare il contrario (ipotesi alternativa). Inoltre un valore di R>0.75 suggerisce una buona separazione tra i gruppi, se R=0.5 i gruppi son in parte sovrapposti, mentre R<0.25 indica una scarsa separazione (K. Clarke & Gorley, 2001) . Anche se poco frequente come situazione, può ottenersi un valore di R negativo; questo significa che la similarità tra gruppi è più alta rispetto a quella riscontrata all'interno dei gruppi.

Infine anche ANOSIM prevede l'utilizzo di permutazioni per costruire la distribuzione di R sotto ipotesi nulla. Una volta scelta la soglia di significatività, la si confronta con il p-value ottenuto a partire dal valore di R e si decide se accettare l'ipotesi nulla oppure quella alternativa.

4.8 Il test di Wilcoxon

Il test di Wilcoxon è un test statistico di verifica d'ipotesi non parametrico, pertanto esso non ipotizza che i dati abbiano una distribuzione precisa. Il test è stato proposto da F. Wilcoxon (Wilcoxon, 1945) e presenta due principali varianti, a seconda che si sia in presenza di dati appaiati (ogni osservazione di un campione è associata/appaiata con una e una sola osservazione di un altro campione) o meno. Nel primo caso infatti si parla di test di Wilcoxon dei ranghi con

segno (che è la variante non parametrica del t-test per dati appaiati), mentre nel caso di campioni indipendenti il test viene definito test di Wilcoxon della somma dei ranghi. In questo caso, con il termine "rango" si definisce la posizione di un dato all'interno di un gruppo o di una serie.

Vista la tipologia dei dati di abbondanza delle specie microbiotiche qui considerati, i quali non sono appaiati, si espone di seguito il test della somma dei ranghi, denominato anche test di Wilcoxon-Mann-Whitney. Nel presente studio il test è utilizzato per identificare le specie le cui abbondanze relative sono significativamente diverse tra soggetti appartenenti a gruppi differenti. In questo modo si possono porre in evidenza gli effetti che le covariate quali il fumo o la malattia hanno sulle specie componenti il microbioma dei soggetti. Trattandosi di un test di verifica d'ipotesi, l'esito del test di Wilcoxon è un p-value, che deve essere confrontato con una soglia di significatività arbitraria per decidere se accettare l'ipotesi nulla o l'ipotesi alternativa. L' equivale a dire che i due campioni sono tratti dalla medesima popolazione, e che quindi per sono caratterizzati dalla stessa siano eguali. L' può essere bilaterale, ovvero se le due distribuzioni sono diverse, o unilaterale, se si testa se una delle due distribuzione è maggiore o minore dell'altra, in senso stocastico.

Per quanto concerne il procedimento alla base del test, i dati dei due gruppi vengono uniti in un'unica serie, nella quale essi sono ordinati in modo crescente (ordinamento in ranghi). In presenza di valori uguali si assegna ad ognuno un rango pari alla media dei ranghi. In seguito si stabilisce la dimensione del gruppo minore (identificata qui con n_1) e quella del gruppo maggiore (n_2); dopodiché si calcola la somma dei ranghi del gruppo meno numeroso, la quale viene chiamata T. Questa, nel caso di ipotesi nulla, tende ad una media attesa μ , la quale si calcola come di seguito:

$$\mu = \frac{n_1(n_1 + n_2 + 1)}{2}$$

Nel caso in cui, invece, sia accettata l'ipotesi alternativa, il valore di T è maggiore o minore della media attesa e può tendere verso un valore minimo (somma dei ranghi minori) oppure verso un massimo (somma dei ranghi maggiori). Per decidere quale ipotesi scegliere, si confronta il valore T ottenuto con la media attesa; per poter poi valutare se esso sia significativamente maggiore o minore della media, è possibile ricorrere ad apposite tabelle se il numero di dati non è considerevole, altrimenti si calcola il p-value.

In sostanza si procede come per il t-test, calcolando il p-value e, se esso risulta essere inferiore alla soglia di significatività si accetta l'ipotesi alternativa, mentre si opta per quella nulla se il p-

value è maggiore.

È inoltre da ricordare che nel caso si effettuino test multipli, come nel presente lavoro di tesi, è consigliato correggere il valore dei p-value ottenuti. Per ogni test infatti si è infatti consapevoli che l'assunzione dell'ipotesi nulla (o di quella alternativa) è affetta da un certo errore, specificato dalla soglia di significatività scelta. Se però si effettuano più test, l'errore si propaga e quindi il risultato finale risente dell'errore commesso in ogni singolo test. Si supponga per esempio di effettuare 20000 test di Wilcoxon; si assume che ogni test è significativo se $p \leq 0.05$ (probabilità del 5% di sbagliare, cioè di dire che la media delle abbondanze relative delle specie di un soggetto è significativamente diversa da quella di un altro soggetto quando in realtà non è così). In questo modo si commette 0.05*20000=1000 errori.
Uno dei metodi che consente di correggere i p-value in modo da limitare l'errore finale, è il *False Discovery Rate* (FDR); il primo passo consiste nell'ordinare in senso crescente i p-value ottenuti dai vari test, dopodiché per ciascun p-value si calcola una quantità così definita:

$$FDR_i = \frac{p_i K_i}{N}$$

In cui p_i è il valore del p-value in questione, K_i è il numero di p-value aventi valori inferiori a quello considerato mentre N è il numero di test effettuati.
Nel momento in cui si deve scegliere quale ipotesi del test accettare, si possono quindi considerare i diversi FDR piuttosto che i p-value.

5 Casi di studio

Sono due i casi di studio presi in considerazione: la broncopneumopatia cronica ostruttiva ed il cancro colon-rettale. Di seguito si riporta dunque, per ciascun caso, la caratterizzazione della malattia, le sue cause ed i trattamenti.

5.1 La broncopneumopatia cronica ostruttiva

5.1.1 Caratterizzazione

La BPCO è una malattia polmonare progressiva e potenzialmente mortale che colpisce circa 65 milioni di persone nel mondo (Dance 2012, 2-3); l'Organizzazione Mondiale della Sanità stima che prima del 2030 la malattia possa diventare la terza causa principale di morte mondiale. La malattia colpisce tipicamente i soggetti di età superiore ai 50 anni ed è in continuo aumento soprattutto tra la popolazione femminile, fino a che nel 2000 negli Stati Uniti il numero di casi diagnosticati nelle donne ha superato quello riscontrato tra gli uomini (Figura 19).

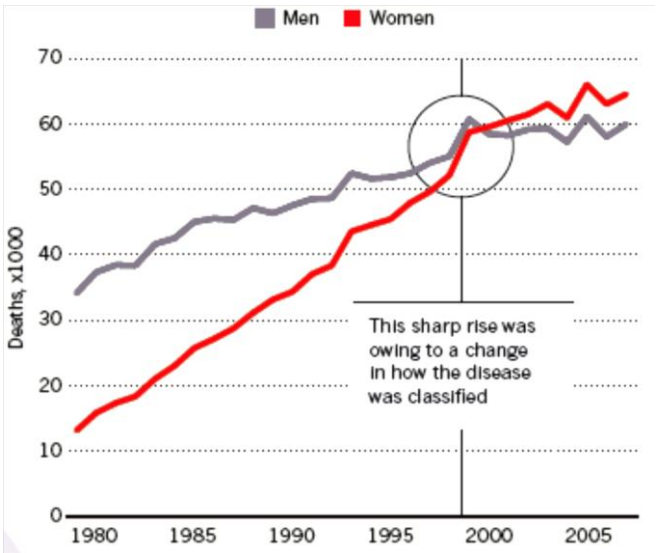


Figura 19. *Numero di morti provocate dalla BPCO tra la popolazione maschile e femminile negli Stati Uniti.*

La BPCO può presentarsi con sintomi simili a quelli dell'enfisema o a quelli della bronchite asmatica cronica, o entrambi contemporaneamente. Nel caso in cui la malattia assuma la forma di enfisema si assiste dunque ad una dilatazione patologica ed irreversibile degli alveoli polmonari con la conseguente lesione del tessuto polmonare; questo riduce la superficie funzionale dei polmoni, compromettendo lo scambio gassoso di ossigeno ed anidride carbonica. Se invece la BPCO si presenta con sintomi tipici della bronchite asmatica cronica, si osserva l'ostruzione delle vie respiratorie a causa di eccessiva produzione di muco causata dall'eccessivo ingrossamento della mucosa dei bronchi. In figura 20 sono evidenziate le differenze che sussistono nei bronchioli e negli alveoli delle persone sane rispetto ai soggetti affetti da BPCO.

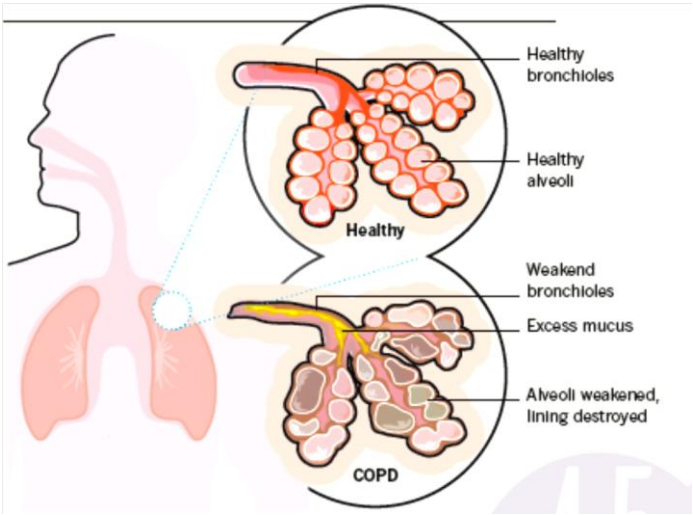


Figura 20. *Differenze a livello di alveoli e bronchioli tra soggetti sani e soggetti affetti da BPCO.*

I soggetti malati soffrono inizialmente di tosse ed espettorazioni mattutine, mentre col progredire della malattia essi vanno incontro a difficoltà respiratorie sempre più accentuate che possono portare alla morte, specialmente in caso di sottovalutazione della malattia.

La mancanza di consapevolezza nei confronti della BPCO è infatti un problema diffuso, basti pensare che il terzo National Health and Nutrition Examination Survey tenutosi nel 2000 ha dimostrato che negli Stati Uniti dei 24 milioni di individui che presentavano sintomi della BPCO, a meno della metà è stata diagnosticata la malattia (Willyard 2012, S8-S9). Ciò è dovuto al fatto che finora non è stata attribuita la giusta importanza alla BPCO rispetto ad altre malattie croniche, quali ad esempio quelle cardiovascolari o il cancro.

Oltre alla mancanza di consapevolezza, un altro problema inerente alla diagnosi della malattia è che essa presenta sintomi che assomigliano a quelli dell'asma e dunque viene spesso confusa con quest'ultima. Sorge così un altro aspetto problematico, che riguarda i metodi e gli strumenti atti alla diagnosi della BPCO; generalmente il medico ricorre infatti all'interpretazione dei

sintomi del paziente, all'analisi della sua anamnesi, e ad uno specifico test per valutare la funzionalità polmonare, chiamato spirometria. Già si è sottolineato come possa essere poco affidabile l'interpretazione dei sintomi, ma è inoltre da evidenziare che anche la spirometria ha i suoi limiti. Questo test consiste nel far soffiare il paziente vigorosamente ed il più a lungo possibile in un boccaglio collegato ad un apparecchio elettronico detto spirometro; si possono così calcolare numerosi parametri che consentono di esprimere una valutazione funzionale di tipo diagnostico. Se però il paziente commette accidentalmente degli errori nell'esecuzione del test (può ad esempio non inspirare a sufficienza prima di espirare nel boccaglio, oppure può non soffiare abbastanza vigorosamente) si può giungere a diagnosi errata; questo accade nel 30-40% dei casi (David Mannino, epidemiologo presso l'Università del Kentucky a Lexington). Inoltre come afferma anche Jan-Willem Lammers, direttore del dipartimento di medicina respiratoria all'University Medical Centre Utrecht, la spirometria consente di valutare l'entità dell'ostruzione delle vie respiratorie ma non di osservare il grado di lesione del tessuto polmonare.

Viste dunque l'inadeguatezza delle tecniche diagnostiche presenti, si stanno ricercando nuovi strumenti per la diagnosi della BPCO. Mannino e Fernando Martinez, specialista dei polmoni afferente all'University of Michigan, stanno per esempio elaborando un device economico e di facile utilizzo che consenta di misurare la massima velocità con cui un soggetto può espirare (*peak flow*). Si è infatti evidenziato che individui che presentano *peak flow* sotto una soglia di normalità possono considerarsi potenziali malati di BPCO.

Un ulteriore strumento diagnostico proposto è la tomografia computerizzata, in quanto lo studio *COPDGene* ha dimostrato che essa può identificare le lesioni polmonari ben prima che ci sia un'ostruzione delle vie aeree e dunque con anticipo rispetto al presentarsi dei primi sintomi effettivamente percepiti dal soggetto.

Inoltre si stanno ricercando dei marcatori naturali per la BPCO, cioè molecole che in base alla loro quantità o presenza nei soggetti permettano di valutare se essi siano a rischio ed eventualmente dedurre lo stadio della malattia. Per esempio, si è visto che nella saliva dei malati di BPCO è presente l'N-acetil-prolina-glicina-prolina, che è un peptide sottoprodotto della rottura del collagene. Non ve n'è invece traccia nei soggetti sani.

5.1.2 Cause della malattia

Il primo fattore di rischio per la BPCO è sicuramente il fumo del tabacco, ma non è di certo l'unico; il solo 15% dei fumatori infatti è affetto dalla suddetta malattia. Non è inoltre ancora del tutto chiaro, il modo in cui le particelle di fumo danneggino i polmoni.

Tra gli altri fattori di rischio, oltre all'inquinamento atmosferico e all'età dei soggetti, è di particolare interesse il fattore genetico (Silverman 2012, S6-S7). Innanzitutto si è osservato che una piccola parte dei soggetti affetti da BPCO presenta una grave carenza di antitripsina α1 (A1ATD), la quale è una glicoproteina in grado di inibire un gran numero di proteasi (enzimi che distruggono le proteine). La deficienza di tale proteina porta alla distruzione del tessuto polmonare, in quanto si è in presenza di un eccesso di proteasi, rispetto ai loro inibitori.
Diversi studi (Cho et al. 2010, 200-202; Cho et al. 2012, 947-957; Wilk et al. 2009, e1000429; Hancock et al. 2009, 45-52) hanno identificato quattro regioni del genoma che possono considerarsi associate alla BPCO.
Le prime due regioni sono vicine ai geni *HHIP* e *FAM13A*, il primo dei quali è componente essenziale per la *Hedgehog pathway* (che regola la differenziazione cellulare e la formazione degli organi durante lo sviluppo embrionale dei vertebrati), mentre il secondo gene ha una funzione tuttora sconosciuta.
Le altre due regioni del genoma invece sono situate nel cromosoma 15 e nel cromosoma 19. In particolare nella regione 19q è presente il gene *CYP2A6* che è coinvolto nel metabolismo della nicotina; nella regione cromosomica 15q25 sono siti inoltre altri geni importanti per diversi componenti dei recettori nicotinici. All'interno di quest'ultima regione è poi presente il gene *IREB2* il quale codifica per una proteina che si pensa sia associata con la suscettibilità alla BPCO.

Si è anche considerata l'ipotesi che la BPCO sia causata da un processo autoimmune patologico, cioè un'alterazione del sistema immunitario che provoca risposte immuni anomale. Si sostiene infatti che il fumo di tabacco alteri le proteine native del soggetto, le quali sono dunque considerate *non self* dal sistema immunitario, che di conseguenza genera una risposta contro di esse. Si è infatti osservato che nei malati di BPCO si riscontra un numero maggiore di anticorpi leganti proteine dell'organismo rispetto alla quantità normalmente presente nei soggetti sani. Per poter dimostrare però che i meccanismi autoimmuni sono fattori determinanti nell'insorgenza della BPCO si devono ancora condurre studi atti a trovare quali anticorpi possono associarsi alla malattia. Non tutti gli anticorpi infatti hanno patogenicità (capacità di causare un danno o l'insorgenza di uno stato di malattia) e per dimostrarla è necessario avere riscontro di evidenti lesioni al tessuto interessato. A tal fine si possono isolare gli anticorpi dai pazienti e valutare i loro effetti sulle cellule umane, mediante coltivazione *in vitro* (Feghali-Bostwick et al. 2008, 156-163). Inoltre si può ricorrere a modelli animali, nei quali riprodurre la malattia e trasferire gli anticorpi provenienti dal paziente. Si tratta dunque di un settore di studio tuttora in evoluzione, ma che potrebbe rivelarsi importante al fine di elaborare nuove terapie.

Infine, diversi studi dimostrano che anche la vitamina D è coinvolta nella patogenesi della

BPCO (Zosky et al. 2011, 1336-1343; Janssens et al. 2010, 215-220). La sua presenza è infatti cruciale per l'attivazione del sistema immunitario, in quanto le cellule T (che riconoscono e distruggono il patogeno) devono legarsi con la vitamina D per poter attivarsi, ed inoltre essa limita la crescita del tessuto muscolare liscio delle vie aeree. Considerando ora che la BPCO causa invece un incremento della massa muscolare dei condotti aerei ed è associata ad un'anomala risposta immunitaria, si nota come vitamina D e BPCO siano associate. Nel 60% dei malati infatti si è riscontrata una carenza di vitamina D, la quale si aggrava col peggiorare della malattia. Tuttavia non è ancora stato provato se la carenza della vitamina possa considerarsi una causa della BPCO ed uno dei principali studi a riguardo è rappresentato da VITAL (*Vitamin D and Omega-3 Trial*), condotto da Gold e Manson. Quello che per ora si può concludere, grazie ad uno studio di Zosky (Zosky et al. 2011, 1336-1343) su due gruppi di topi, è che la carenza di vitamina D in sé non causa grossi problemi a livello polmonare, ma può rivelarsi un'aggravante nel caso di comparsa della BPCO.

5.1.3 Trattamenti

Data la caratteristica di cronicità della BPCO, non si può parlare di cure bensì di trattamenti per alleviarne i sintomi il più a lungo possibile.
Il trattamenti più comuni sono i broncodilatatori e le inalazioni di corticosteroidi, che causano il rilassamento delle vie aeree e riducono le infiammazioni, consentendo un sollievo immediato. Ci sono poi i broncodilatatori "a lungo termine", quali il tiotropio bromuro (Spiriva) e il salmeterolo-fluticasone propionato (Advair), che derivano dalla combinazione di più sostanze e consentono di controllare i sintomi della malattia per molte ore; nello specifico, Advair ha un effetto che perdura per 12 ore, mentre Spiriva dona sollievo per circa 24 ore.
Novartis ha poi sviluppato un'alternativa, il glicopirronio bromuro (Seebri), il quale contrasta l'azione dell'acetilcolina, un neurotrasmettitore che si lega ai recettori muscarinici presenti nel muscolo liscio dei canali aerei e ne stimola la contrazione. Poiché la BPCO stimola la produzione di acetilcolina, che quindi causa la riduzione del flusso di aria tipico dell'asma, Seebri blocca i recettori muscarinici, impedendone il legame con l'acetilcolina. Seebri, così come Spiriva, rientra quindi nella categoria degli antagonisti muscarinici a lungo termine (LAMA, *Long-Acting Muscarinic Antagonist*).
Una seconda categoria è rappresentata dai LABA *(Long-Acting β2-Agonist)*, che non bloccano bensì stimolano i recettori adrenergici β2, i quali consentono il legame a particolari neurotrasmettitori che provocano il rilassamento dei muscoli lisci delle vie

aeree. Dei LABA fanno parte Advair e Relvar, combinazione di un broncodilatatore ed un corticosteroide, prodotto dalla londinese GlaxoSmithKline (GSK).

Nella figura 21 sono riportati i più diffusi trattamenti a lungo termine.

LONG-ACTING, DAILY MEDICATIONS

Do combination therapies really add value to COPD treatments? The pros and cons of long acting therapies in the pipeline.

Name	Main benefit	Type of drug	Active components	Delivery method	Doses per day	Adverse effects	Stage
Advair	Improves lung function for a period of time	Long-acting β2-agonist (LABA) and corticos-teroid	Fluticasone/ salmeterol	Dry powder inhaler	2	Increased risk of non-fatal pneumonia	Available
Spiriva	Improves lung function for a period of time	Long-acting muscarinic antagonist (LAMA)	Tiotropium bromide	Dry powder inhaler	1	Hives, rash, swelling and dry mouth	Available
PT003	Efficient delivery, improved lung function	LAMA + LABA (two molecules)	Glycopyrrolate and formoterol (LAMA + LABA)	Metered dose inhaler (MDI)	2	Headache, dry mouth and coughing	Phase II com-pleted and Phase III to begin 2013
NVA237	Improves breathing in a matter of minutes	LAMA	Glycopyrronium bromide (LAMA)	Dry powder inhaler	1	Headache, dry mouth and coughing	Phase III completed and approval sought in Europe
Relvar	Once daily instead of twice	LABA + inhaled corticosteroid	Vilanterol and fluticasone furoate	Dry powder inhaler	1	Fatal pneumonia reported	Phase III completed and approval sought in USA and Europe
MABA	Dual action molecule improves lung function	LAMA + LABA	MABA	Most likely dry powder inhaler	Unknown	Unknown	Phase II

Figura 21. *Trattamenti della BPCO a lungo termine, con relativi aspetti a favore e non.*

Si è inoltre dimostrato che combinando trattamenti diversi si possono ottenere risultati migliori. Per esempio si è visto che i corticosteroidi presentano effetti antinfiammatori molto più marcati se utilizzati unitamente ai LABA, in quanto questi ultimi facilitano l'ingresso degli steroidi nel nucleo cellulare grazie alla stimolazione dei recettori adrenergici. Si è inoltre ricorsi all'utilizzo combinato di un LAMA e di un LABA contemporaneamente, in modo da agire sia sulle terminazioni nervose che stimolano il rilassamento dei muscoli delle vie aeree, sia bloccando quelle che ne promuoverebbero la contrazione. In alternativa si può somministrare un LAMA ed un LABA non simultaneamente, bensì uno dopo l'altro, ma con effetti molto simili a quelli ottenuti con la somministrazione in contemporanea; in ogni caso l'utilizzo simultaneo porta ad effetti più marcati. Infine, si sta elaborando un nuovo farmaco, prodotto da GSK, che è composto da un'unica molecola avente le proprietà sia di un LAMA che di un LABA; tale farmaco prende il nome di MABA (*muscarinic antagonist–β2 agonist*).

Nonostante i farmaci sopra presentati offrano prospettive interessanti, non se n'è finora trovato alcuno che conduca ad un miglioramento della funzionalità polmonare, afferma Cates, ricercatore alla St George's University di Londra. Per ora dunque, uno dei "trattamenti" più consigliati rimane quello di smettere di fumare.

5.2 Il tumore colon-rettale

5.2.1 Caratterizzazione

Con il termine carcinoma del colon-retto (CRC) si individuano in realtà diversi tipi di neoplasia che possono riguardare differenti sedi, quali il retto, il colon prossimale, il colon discendente, il colon traverso o il sigma (Figura 22). Il CRC è caratterizzato dalla proliferazione incontrollata delle cellule della mucosa che riveste le sedi appena citate.

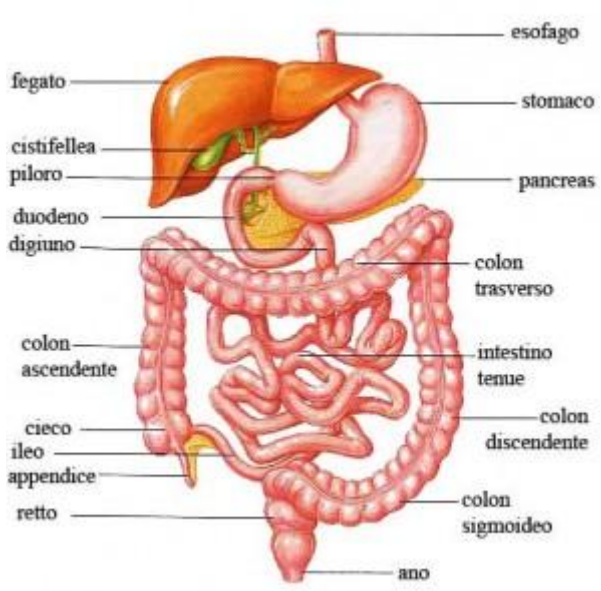


Figura 22. *Schema dell'apparato digerente umano.*

Il CRC è una delle più frequenti cause di morte per neoplasia al mondo, contando oltre un milione di nuovi casi e circa 610.00 morti ogni anno (Organizzazione Mondiale della Sanità, 2011). I paesi più colpiti sono quelli più economicamente avanzati (Nord America ed Europa occidentale in primis) nei quali il cancro del colon-retto rappresenta il terzo tumore maligno per incidenza e mortalità, ma il tasso di incidenza del CRC è in aumento anche nei paesi in via di sviluppo (Zhu et al. 2011, 119-127). È stato inoltre stimato che a 5 anni dalla diagnosi la sopravvivenza è del 90% se la malattia è diagnosticata precocemente, cioè quando è ancora confinata alla parete dell'intestino, del 68% se la neoplasia è estesa a livello loco-regionale, ovvero con coinvolgimento linfonodale, e solo del 10% se sono presenti metastasi a distanza (Jemal et al. 2009, 43-66).

L'incidenza del CRC è abbastanza rara prima dei 40 anni, mentre è sempre più frequente a partire dai 60 anni, raggiungendo il picco massimo verso gli 80 anni (AIRC,

2008). Il tumore è più diffuso tra il sesso maschile soprattutto a livello rettale, dove il rapporto maschi/femmine è di 2:1; non vi sono invece significative differenze relativamente alla localizzazione colica. In generale comunque il tasso di incidenza del CRC è del 25-30% a livello rettale, del 15-20% a livello del colon prossimale, del 12% a livello del trasverso, del 20% nel colon discendente e del 15-20% a livello del sigma.
La sintomatologia del tumore al colon-retto è molto variabile e condizionata da diversi fattori quali la localizzazione del tumore, la sua estensione, la sua stadiazione e la presenza eventuale di ostruzioni o emorragie; ciò fa sì che le manifestazioni del CRC siano sovente sovrapponibili a quelle di molte altre malattie addominali o intestinali e per questo trascurate. Nel caso di tumore al colon destro le lesioni sono di grandi dimensioni e tendono a sanguinare con facilità, quindi uno dei primi sintomi può essere un'anemia causata dalla perdita di sangue mescolato alle feci, minima ma continua. Di conseguenza lo stato anemico può esser causa di stanchezza, perdita di peso, dispnea e sensazione di affanno anche per sforzi di lieve entità.
Nel caso in cui il tumore si sviluppi invece nella parte sinistra del colon si hanno solitamente sintomi più precoci; il tumore tende a penetrare nei vari strati delle pareti dell'intestino causando perdite di sangue ripetute, miste a feci e spesso anche a muco. Questo è solitamente il primo sintomo a manifestarsi, seguito da stitichezza alternata a diarrea, alterazione del calibro delle feci, bruciore a livello anale e fastidio a livello addominale. In un secondo momento possono comparire disturbi urinari, dolori crampi formi e si può percepire una massa palpabile.
In generale, nelle fasi più avanzate dei tumori del colon-retto si assiste a forte dimagrimento, perdita dell'appetito, nausea, vomito ed un evidente decadimento delle condizioni fisiche.

5.2.2 Cause del tumore

Le cause del CRC rimangono sostanzialmente sconosciute, ma sono stati identificati diversi fattori di rischio, suddivisi in 3 categorie:

- *Fattori genetici*: è possibile ereditare il rischio di ammalarsi di tumore del colon-retto se nella famiglia d'origine si sono manifestate alcune malattie che predispongono alla formazione di tumori intestinali. La più importante tra queste è la poliposi adenomatosa familiare (PAF), che è caratterizzata dalla proliferazione abnorme delle strutture ghiandolari

della mucosa del colon; si tratta del più frequente tumore benigno del grosso intestino, che però presenta una certa tendenza alla trasformazione maligna. Tra le altre malattie ereditarie si ricorda la sindrome di Lynch (sviluppo di carcinoma al colon e possibili neoplasie a livello dell'endometrio, dello stomaco, del tratto urinario, dei dotti biliari), la sindrome di Gardner (poliposi intestinale, anomalie della dentizione, anomalia nello sviluppo delle ossa craniche) e la sindrome di Turcot (variante della FAP con comparsa di tumori cerebrali) . Queste malattie sono trasmesse da genitori portatori di specifiche alterazioni genetiche, che possono anche non dar luogo ad alcun sintomo. Inoltre si stima che il rischio di sviluppare un tumore del colon aumenti di 2 o 3 volte se parenti di primo grado sono stati a loro volta affetti da CRC.

- *Fattori nutrizionali*: diversi studi dimostrano che una dieta povera di fibre, ricca di grassi saturi e ad alto contenuto di calorie è associata ad una maggior predisposizione ai tumori intestinali. Infatti le fibre svolgono un ruolo protettivo data la capacità di legare i grassi e gli acidi biliari; viceversa, il metabolismo dei grassi saturi provoca l'aumento degli acidi biliari, considerato come una delle cause promotrici della cancerogenesi. L'obesità è dunque un fattore di rischio strettamente collegato al CRC umano.
- *Fattori non ereditari*: l'età per esempio è un fattore di rischio importante (incidenza 10 volte superiore tra i soggetti con più di 60 anni, rispetto a quelli con età inferiore ai 40), così come è da tener conto che certe categorie lavorative sembrano essere più a rischio di sviluppo del CRC (metalmeccanici, lavoratori del settore del legno e del cuoio) a causa dell'esposizione a determinati agenti. Inoltre il rischio di CRC aumenta negli individui con un pregresso tumore al colon o affetti in passato da poliposi adenomatose. Inoltre un aumentato rischio si associa alla colite ulcerosa e al morbo di Chron; entrambe sono malattie infiammatorie croniche, la prima causa erosioni ed ulcerazioni della mucosa del colon con tendenza al sanguinamento, mentre il morbo di Chron può coinvolgere qualsiasi tratto del tubo digerente e porta alla formazione di granulomi sotto lo spessore delle sue pareti.

Recenti studi eseguiti su feci o mucosa intestinale hanno inoltre dimostrato differenze tra microbioma di soggetti affetti da CRC e soggetti sani, ma non è ancora del tutto chiaro se le alterazioni nella composizione del microbiota siano precedenti l'insorgenza di CRC oppure ne siano una conseguenza. Si è però osservato che i fattori che inducono lo sviluppo del CRC influenzano anche la composizione microbica dell'intestino; questo potrebbe suggerire che l'alterazione a livello di microbioma è da considerarsi dunque come una causa piuttosto che una conseguenza del CRC. Per esempio, si è detto che i soggetti anziani sono più a rischio di sviluppare il tumore rispetto ai soggetti giovani; uno studio ha evidenziato una riduzione del

numero complessivo di microbi negli anziani in relazione ai giovani, nonché ad una diminuzione di *Firmicutes* e ad un incremento di *Bacteroidestes* nei primi rispetto ai secondi (Mäkivuokko et al. 2010, 227). Inoltre è noto che l'obesità sia un fattore di rischio per il CRC umano e numerosi studi hanno verificato l'associazione tra una alterazione dei phyla dominanti di batteri nell'intestino e aumento del peso corporeo, sia nell'uomo che nei modelli animali (Kalliomäki et al. 2008, 534-538).

Una volta chiarita dunque la relazione causa-effetto tra microbioma e CRC, l'analisi metagenomica del microbioma può rivelarsi uno strumento utile per la diagnosi, nonché il trattamento del tumore.

5.2.3 Trattamenti

La terapia più diffusa è la chirurgia: sulla base della posizione del tumore si procederà con un intervento parziale o, nei casi più gravi, con la totale asportazione del tratto di colon interessato o del retto. La chirurgia del colon si distingue in chirurgia d'elezione e di urgenza. Nel primo caso si effettua la resezione del tratto colico interessato dal tumore e vi si ricorre nell'80% dei casi, mentre la chirurgia d'urgenza viene praticata nel restante 20% dei casi per perforazione o occlusione intestinale.
La radioterapia poi ha una funzione soltanto palliativa nel tumore al colon mentre in quello del retto può avere anche uno scopo curativo, in quanto permette in fase preoperatoria di ridurre la massa tumorale e renderla quindi più facilmente asportabile chirurgicamente. Si è inoltre dimostrato che la radioterapia è in grado di diminuire le ricadute locali.
Infine, la chemioterapia svolge un ruolo fondamentale nella malattia avanzata non operabile, ma non solo. Recentemente sono stati intrapresi diversi studi per valutare l'efficacia di un trattamento chemioterapico cosiddetto adiuvante, cioè effettuato dopo l'intervento chirurgico per diminuire il rischio di ricaduta (come avviene già per il tumore della mammella): i primi dati a disposizione sono positivi. Sono positivi anche gli studi sulla terapia neoadiuvante, cioè effettuata prima dell'intervento per ridurre la dimensione del tumore e facilitare il compito del chirurgo.

6 Dati

Qui di seguito si presentano i dataset originali dei due casi di studio, i quali sono stati dapprima sottoposti ad una fase di preprocessing (si veda capitolo 7) e sui quali successivamente si sono applicate le tecniche esposte nel capitolo 4.

Innanzitutto si prendono in considerazione i dati relativi al caso di studio sulla BPCO.
I dati di partenza sono le abbondanze assolute delle OTU presenti nel microbioma umano di 14 soggetti. Questi dati sono il risultato della classificazione delle reads mediante RDP (si veda capitolo 3) , classificatore che consente di raggruppare le sequenze nucleotidiche sulla base di un certo livello di similarità ed associarle così ad un determinato livello tassonomico.
I campioni sui quali eseguire il sequenziamento sono stati raccolti dall'espettorato indotto di 13 soggetti, mentre in un solo caso si è lavorato sul materiale ottenuto dal lavaggio bronco alveolare (BAL). Di questi 13 soggetti, 6 sono affetti da BPCO, 4 sono fumatori sani, mentre 3 sono non fumatori sani.
Per alcuni soggetti inoltre si è ripetuto l'espettorato due volte, a distanza di mesi; in totale si hanno a disposizione quindi 22 misure.

Di seguito si presentano le sigle usate per identificare le diverse tipologie di soggetti:

- BAL: soggetto al quale si è praticato lavaggio bronco alveolare;
- BP_F: soggetto fumatore affetto da BPCO;
- BP_NF: soggetto non fumatore affetto da BPCO (non disponibile);
- HC_F: soggetto fumatore sano;
- HC_NF: soggetto non fumatore sano.

Il dataset originario è dunque composto da 22 osservazioni, per ciascuna delle quali si sono ricavate 5950 OTU.
Per quanto riguarda invece lo studio relativo al tumore colon-rettale, si hanno a disposizione 56 campioni totali di tessuto proveniente da 32 soggetti. Di questi 8 sono individui sani, dai quali si è estratto un campione di tessuto a testa, mentre gli altri 24 soggetti presentano lesioni di diverso tipo a livello del colon. Per ciascuno di essi si sono estratti due campioni, uno di tessuto lesionato ed un altro di tessuto sano adiacente al precedente. Dei 24 soggetti con lesioni, 8 presentano polipi displastici, 8 sono affetti da polipi non displastici e 8 sono malati di cancro.

Di seguito si presentano le sigle usate per identificare le diverse tipologie di tessuto:

- HC: tessuto sano da soggetti sani;
- TB: tessuto malato da soggetti con polipi non displastici
- TB_C: tessuto sano da soggetti con polipi non displastici
- TD: tessuto malato da soggetti con polipi displastici
- TD_C: tessuto sano da soggetti con polipi displastici
- TM: tessuto malato da soggetti con cancro
- TM_C: tessuto sano da soggetti con cancro

Da ciascun campione di tessuto si sono ricavate le abbondanze assolute delle OTU componenti il rispettivo microbiota. Si hanno così a disposizione 56 osservazioni, per ognuna delle quali si sono misurate le abbondanze di 11877 OTU.

7 Analisi dei dati

7.1 Preprocessing

Il primo step nell'analisi dei dati di entrambi i casi di studio è stato il raggruppamento delle OTU appartenenti allo stesso genere, in modo da ottenere un ulteriore dataset, nel quale si specificano i generi dei batteri componenti il microbiota di ciascun campione. Si sono così ottenuti:

- 83 generi per ciascun soggetto nel caso di BPCO;
- 672 generi per ciascun campione nel caso del tumore colon-rettale.

In seguito, sia nel caso delle OTU sia nel caso dei generi, i dati sono stati normalizzati; in pratica si è calcolata l'abbondanza totale di generi o OTU per ogni soggetto, dopodiché si è calcolato il rapporto tra i counts assoluti e l'abbondanza totale. A tal proposito, si riportano di seguito dei grafici a barre, nei quali si mostrano i generi (Figura 23 e 25) e le OTU (Figura 25 e 26) più abbondanti all'interno dei soggetti/campioni; sulle ascisse si evidenziano i soggetti/campioni di tessuto mentre l'altezza delle barre indica l'abbondanza relativa di un certo genere o OTU. Le figure 23 e 24 sono riferite al caso di studio della BPCO, mentre le figure 25 e 26 si riferiscono al dataset dello studio sul tumore colon-rettale.

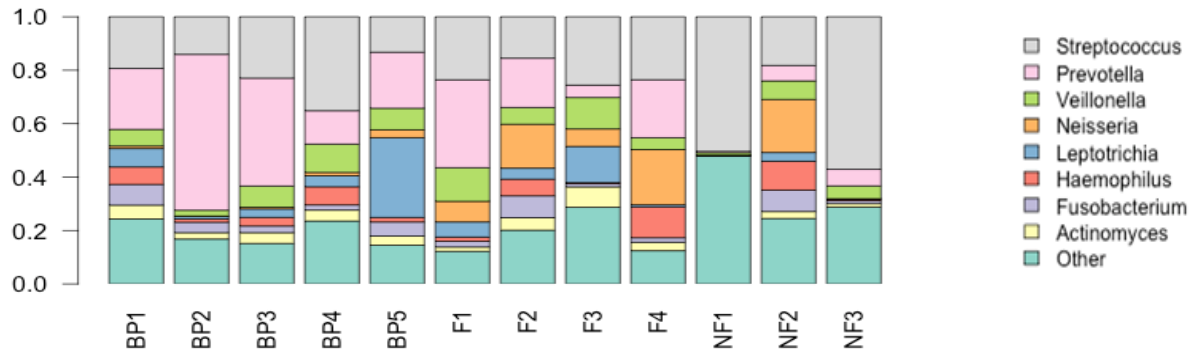


Figura 23. *Caso di studio: BPCO. Generi più abbondanti presenti nei soggetti (fumatori malati (BP), fumatori sani (F) e non fumatori sani (NF)). Sulle ordinate i valori di abbondanza relativa.*

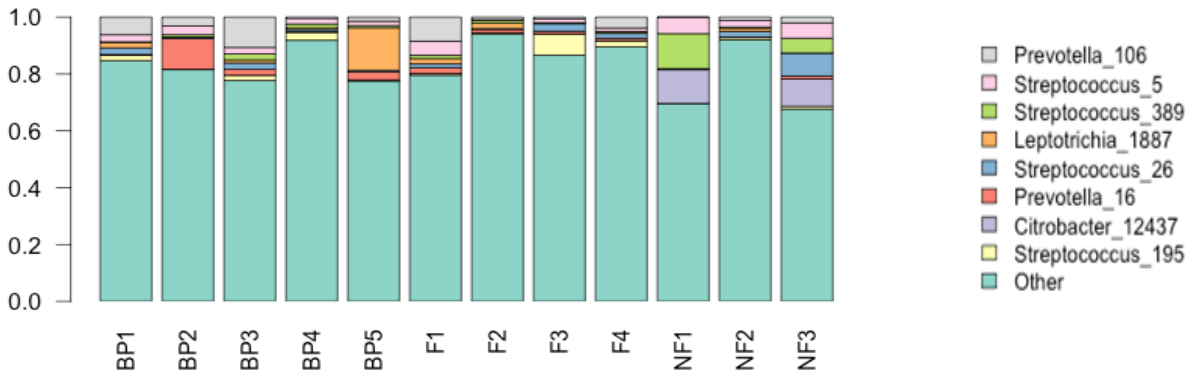


Figura 24. *Caso di studio: BPCO. OTU più abbondanti presenti nei soggetti (fumatori malati (BP), fumatori sani (F) e non fumatori sani (NF)). Sulle ordinate i valori di abbondanza relativa.*

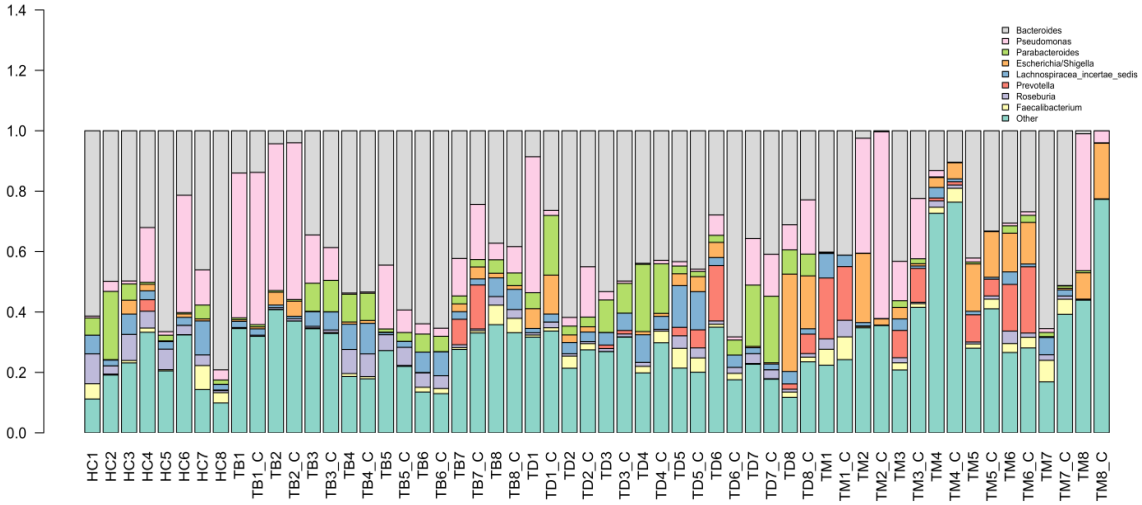


Figura 25. *Caso di studio: tumore colon-rettale. Generi più abbondanti presenti nei soggetti (soggetti sani (HC), tessuto sano di soggetti con polipi non displastici (TB_C), tessuto malato di soggetti con polipi non displastici (TB), tessuto sano di soggetti con polipi displastici (TD_C), tessuto malato di soggetti con polipi displastici (TD), tessuto sano di soggetti con resez. cancro (TM_C) , tessuto malato di soggetti con resez. cancro (TM)). Sulle ordinate i valori di abbondanza relativa.*

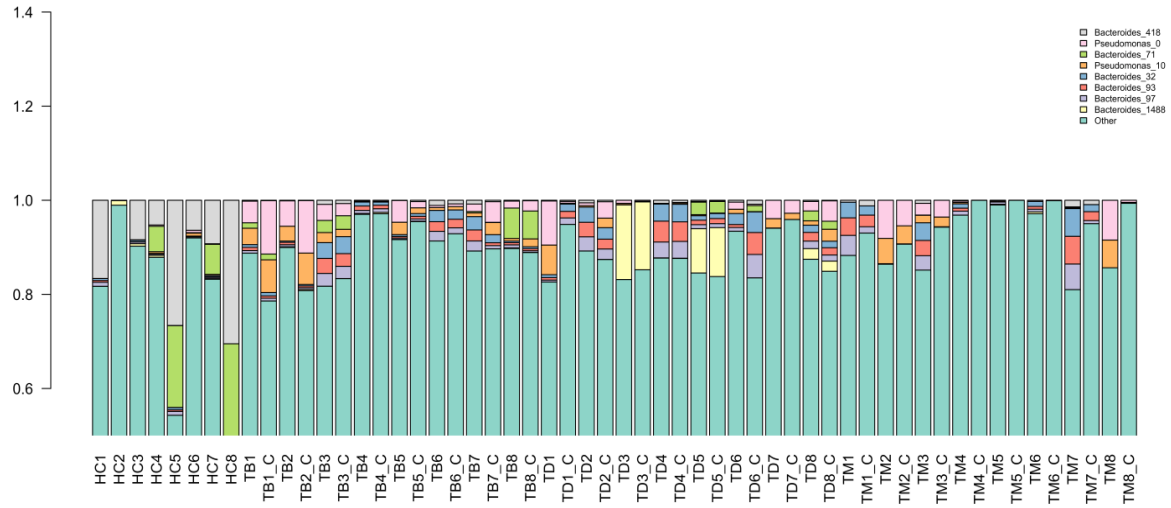


Figura 26. *Caso di studio: tumore colon-rettale. OTU più abbondanti presenti nei soggetti (soggetti sani (HC), tessuto sano di soggetti con polipi non displastici (TB_C), tessuto malato di soggetti con polipi non displastici (TB), tessuto sano di soggetti con polipi displastici (TD_C), tessuto malato di soggetti con polipi displastici (TD), tessuto sano di soggetti con resez. cancro (TM_C) , tessuto malato di soggetti con resez. cancro (TM)). Sulle ordinate i valori di abbondanza relativa.*

Un ulteriore passo di preprocessing è stata la riduzione della dimensione del dataset. Nel caso dei dati relativi alla BPCO così come in quello del tumore al colon si è diminuito il numero delle OTU e dei generi. Si sono infatti escluse quelle OTU e quei generi che in uno o più soggetti presentano meno di 10 sequenze; si sono così ottenuti i seguenti risultati:

- nel caso di studio relativo alla BPCO si è passati da 5950 OTU iniziali a 845, mentre il numero di generi è sceso da 366 ad 83;
- nel caso di studio relativo al tumore colon-rettale si è passati da 11877 OTU iniziali a1802, mentre il numero di generi è sceso da 672 a 212.

Successivamente, solo nel caso del dataset inerente alla BPCO, si è ridotto anche il numero di osservazioni, escludendo:

- BP6 e una delle due repliche di BP3 e BP2 in quanto affetti da possibili infezioni (si veda figura 28);
- BAL (soggetto sottoposto a lavaggio broncoalveolare) poiché i dati si raccolgono da una tecnica non comparabile all'espettorato indotto, a cui si è ricorsi per tutti gli altri soggetti;
- Si sceglie solamente una delle due repliche per i soggetti per cui sono disponibili (si opta per la replica con la maggiore profondità di sequenziamento)

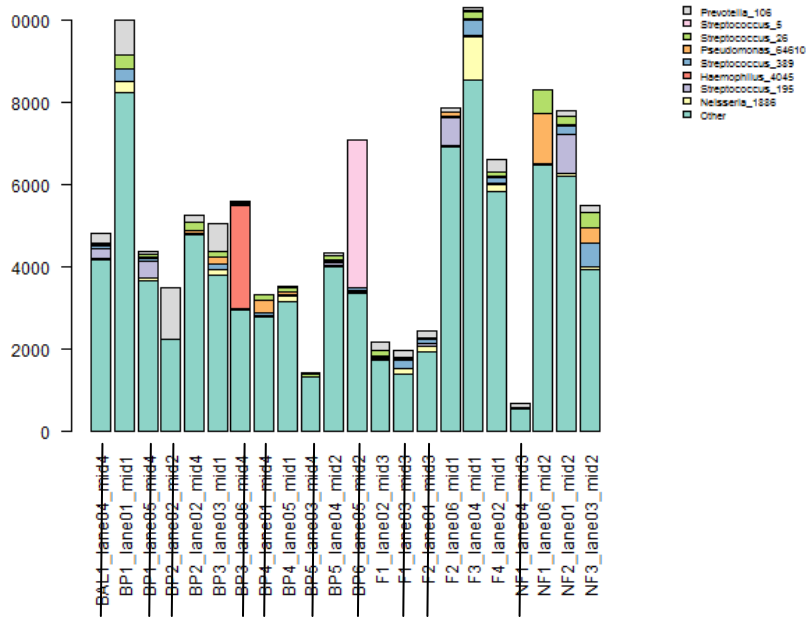


Figura 27. *Distribuzione delle 8 OTU più abbondanti all'interno dei singoli soggetti; le restanti sono catalogate come "other" (verde)*

Infine si è organizzato ciascun dataset in forma matriciale: sulle righe a seconda del caso di studio si trovano i soggetti (fumatori malati, fumatori sani e non fumatori sani) o le tipologie di tessuto (tessuto sano da soggetti sani, tessuto sano e malato da soggetti con polipi displastici, tessuto malato e sano da soggetti con polipi non displastici, e tessuto sano e malato da soggetti affetti da cancro), mentre sulle colonne sono riportate le abbondanze relative dei generi o delle OTU.

La dimensione delle matrici è:

- 12x83 nel caso delle abbondanze relative dei diversi generi e 12x845 nel caso dei dati suddivisi per OTU, riferendosi allo studio sulla BPCO;
- 56x212 nel caso delle abbondanze relative dei diversi generi e 56x1802 nel caso dei dati suddivisi per OTU, riferendosi allo studio sulla BPCO.

7.2 Indici di biodiversità

Il calcolo degli indici è stato effettuato in R, utilizzando funzioni implementate nel pacchetto **vegan**.

Innanzitutto si sono calcolati gli indici di diversità alfa (si veda paragrafo 4.4).
A tal proposito si è utilizzata la funzione *diversity()*, che consente di stimare gli indici di Shannon, di Simpson (e la sua formulazione inversa 1-D) e di Evenness.
La funzione richiede come argomenti in ingresso una matrice, in cui siano presentati i campioni/siti sulle righe e le specie (o, in questo caso generi/OTU) sulle colonne, e l'indice che si desidera calcolare. Per la computazione dell'indice di Shannon è richiesta la base del logaritmo da utilizzare, per la quale è stata scelta *e*.
L'output della funzione è un vettore contente i valori assunti dall'indice selezionato, e di dimensioni pari al numero di soggetti analizzati. Si ottiene quindi un valore per ciascun soggetto.

Successivamente si sono calcolati gli indici di diversità beta (si veda paragrafo 4.5), dapprima tra i soggetti appartenenti alla stessa categoria e poi tra i soggetti delle differenti categorie.

- Indice di Sørensen

A tal proposito si è ricorsi alla funzione *vegdist()*, la quale consente di calcolare indici di dissimilarità (o distanza) a partire da dati quantitativi. In input essa prevede una matrice, le cui colonne identificano le specie, mentre le righe presentano i siti/campioni nei quali le specie son presenti. Altro argomento richiesto dalla funzione è l'indice di dissimilarità da utilizzare (quali la distanza Euclidea, di Manhattan, di Canberra o di Bray-Curtis). Nella presente analisi, per il calcolo dell'indice di Sørensen, si è ricorsi alla distanza di Bray-Curtis in versione binaria:

$$d_{bc} = \frac{b + c - 2a}{b + c}$$

in cui b e c rappresentano il numero delle specie rispettivamente del primo e del secondo sito posto a confronto, mentre a identifica le specie in comune (vedi Fig.14 del paragrafo 4.6).

Come output, *vegdist()* restituisce un oggetto della classe "*dist*": una matrice triangolare contenente i valori dell'indice di distanza calcolato per ogni possibile coppia di soggetti (e quindi di righe della matrice di partenza).

- Indice di Jaccard e indice di Harte-Kinzig

La funzione utilizzata per il calcolo di questi indici è *betadiver()*. Come input riceve una matrice che descrive per ciascun campione (righe) la quantità ed il tipo di specie presenti in esso (colonne). Inoltre, si richiede di specificare l'indice di diversità da calcolare, tenendo conto che alcuni tra di essi forniscono un'indicazione di similarità anziché di dissimilarità (come nel caso dell'indice di Jaccard, di cui si discute di seguito).

Anche in questo caso, l'output della funzione è un oggetto della classe "*dist*" contenente i valori dell'indice di distanza calcolato per ogni possibile coppia di campioni.

Come anticipato, si è calcolata la diversità beta tra i soggetti appartenenti ad uno stesso gruppo (individuato sulla base di criteri che dipendono dal caso di studio), utilizzando tre indici differenti: Sørensen, Jaccard, Harte-Kinzig .

In seguito si è valutata invece la diversità tra due gruppi alla volta. In input alle funzioni dunque si sono passate di volta in volta apposite sottomatrici della matrice di partenza (sia nel caso delle abbondanze dei generi sia di quelle delle OTU), corrispondenti ai gruppi di interesse. Poiché l'output di queste funzioni è una struttura contente gli indici calcolati per ogni combinazione possibile di campioni passati in input, per ottenere un unico valore se ne è calcolata la media.

E' inoltre da precisare che i valori riferiti all'indice di Jaccard, sono in realtà il risultato di 1-β_{jacc}, in modo tale che i valori siano confrontabili con quelli ottenuti dagli altri indici: lo 0 identifica diversità nulla e viceversa l'1 diversità massima.

7.3 Analisi della diversità con NPMANOVA

Il metodo NPMANOVA è stato applicato utilizzando la funzione *adonis()*, implementata nel pacchetto R **vegan**. Il primo argomento che questa funzione richiede in ingresso è la formula di modello, in cui si specifichi la matrice di dati o l'oggetto di classe *"dist"* da essa ottenuto (la matrice di dissimilarità calcolata tra i dati) e le variabili o i fattori che indicano i gruppi da confrontare. Per applicare NPMANOVA si passa infatti attraverso il fit del modello lineare corrispondente, in cui i dati da analizzare sono posti in relazione a dei fattori che identificano la suddivisione dei dati stessi in gruppi. Nello specifico caso di NPMANOVA ad una via (in quanto ogni dato è classificato solo in base al gruppo al quale appartiene), il modello lineare è il seguente:

$$Z_{rs}=\mu+\alpha_r+\varepsilon_{rs}$$

In cui Z identifica i dati, *r* indica il gruppo al quale il dato appartiene, mentre *s* è la posizione del dato all'interno di un certo gruppo; μ è la media generale di tutti i dati che definisce la dimensione dell'esperimento, α quantifica l'effetto dovuto all'appartenenza al gruppo i-esimo, mentre ε è un fattore casuale.

Poiché con NPMANOVA si analizza l'influenza di una o più covariate sui dati, esse devono essere evidenziate a monte dell'analisi; per fare ciò si deve selezionare un apposito criterio per suddividere i dati in gruppi. In tal modo, a seconda di come i dati sono raggruppati, si evidenziano certe covariate rispetto ad altre. I fattori che si specificano quindi in input alla funzione *adonis()* permettono di definire quali sono le covariate scelte. Nello studio relativo alla BPCO le due principali covariate che si son utilizzate sono il fumo (evidenziata raggruppando i soggetti in base a chi è fumatore e non fumatore), e la malattia (raggruppando i soggetti in base alla presenza o meno della malattia). Si effettuano inoltre ulteriori confronti: tra fumatori malati e fumatori sani, tra fumatori sani e non fumatori sani, e tre fumatori malati e non fumatori sani.

Nel caso di studio relativo al CRC invece le covariate sono lo stato di salute dei soggetti dai quali si prelevano i campioni di tessuto, e le condizioni in cui si presentano i vari tessuti (sani o lesionati).

La funzione *adonis()* poi prevede in input la struttura in cui son stati definiti i suddetti fattori, il numero di permutazioni da eseguire per costruire la statistica ed eventualmente il tipo di misura di distanza da calcolare se si passa come primo argomento una matrice di dati anziché di dissimilarità.

La funzione restituisce molteplici output:

- una tabella contente le fonti di variabilità, i gradi di libertà, le somme di quadrati, la media dei quadrati, la statistica F ed il p-value;

- una matrice dei coefficienti del modello specificato in input, le cui righe rappresentano le fonti di variabilità e le colonne indicano le specie;
- una matrice dei coefficienti del modello specificato in input, le cui righe rappresentano le sorgenti di variazione e le colonne indicano i campioni/siti;
- una matrice con la statistica F sotto ipotesi nulla per ciascuna sorgente di variazione;
- la matrice del modello;
- i termini che compongono il modello.

L'approccio NPMANOVA è stata applicata ai dati di partenza, in modo tale da valutare eventuali differenze significative tra i gruppi di soggetti ed indagare l'influenza che i fattori sopra esposti possono rivestire. Si è dunque fornita in ingresso alla funzione *adonis()* la matrice di dati a nostra disposizione (soggetti sulle righe e generi/OTU sulle colonne) e si sono tratte le conclusioni sulla diversità tra i gruppi osservando i p-value restituiti in output.
Lo stesso procedimento è stato applicato sia nel caso di abbondanze relative dei diversi generi, sia nel caso delle abbondanze delle OTU.

7.4 Analisi delle similarità tra soggetti con ANOSIM

Nel presente studio, per l'applicazione di ANOSIM ai dati, si è ricorsi alla funzione *anosim()*, la quale è contenuta nel pacchetto R **vegan**.
In input essa richiede la matrice dei dati originale (con i campioni sulle righe e le specie sulle colonne) oppure una matrice di dissimilarità stimata dai dati (simmetrica, quadrata, contenente le distanze tra le coppie di campioni), eventualmente salvata in un oggetto di tipo *"dist"*. Si deve poi specificare in ingresso quali siano i gruppi da confrontare all'interno del dataset ed il numero di permutazioni da effettuare per costruire la statistica R. Nel caso in cui il primo argomento sia la matrice di dati, si deve inoltre specificare la misura di distanza con la quale costruire la matrice di dissimilarità.
La funzione fornisce molteplici output, tra cui: il valore di R ed il p-value risultato dal test d'ipotesi. Si è dunque dapprima calcolata la matrice di dissimilarità, la quale contiene le distanze di Bray-Curtis tra campioni, a partire dalla matrice contente i campioni stessi sulle righe e le specie sulle colonne; successivamente si son specificati i gruppi da confrontare.

Nel caso di studio sulla BPCO si sono inseriti nel medesimo gruppo i soggetti fumatori (malati e non), mantenendo i non fumatori sani in un gruppo a sé stante. In questo modo il fumo risulta essere l'unico fattore di cui si vuole testare l'influenza sulla natura del raggruppamento dei dati; in seguito si son posti a confronto i soggetti fumatori malati ed i soggetti sani (fumatori e non), considerando in tal caso il ruolo che non più il fumo bensì la malattia riveste nelle eventuali differenze riscontrate tra soggetti.
Successivamente si sono effettuati ulteriori confronti, di cui qui di seguito: fumatori malati / fumatori sani, fumatori malati/ non fumatori sani, fumatori sani/ non fumatori sani.
Per quanto riguarda invece il caso di studio riferito al tumore al colon, si sono confrontate tutte le possibili coppie di campioni, evidenziando così le variazioni del microbiota sia tra soggetti con diversi stati di salute sia tra tessuti di soggetti con la stessa diagnosi.

7.5 Test di Wilcoxon per l'identificazione delle differenze nella composizione microbiotica tra soggetti

Dopo aver verificato la presenza di differenze significative tra i diversi campioni, ci si è posti l'obiettivo di individuare quali specie batteriche siano responsabili di tali differenze. Si è dunque considerata una specie alla volta, sia nel caso del dataset formato dalle abbondanze dei generi sia in quello composto dalle abbondanze delle OTU, e per ciascuna si è confrontata l'abbondanza relativa di quella specie nei diversi gruppi o nei diversi campioni.
Nello specifico, comparando due gruppi alla volta, si è applicato un test di Wilcoxon con ipotesi unilaterali per ogni specie, in modo tale da evidenziare in quale dei due gruppi si trovi un'abbondanza maggiore di una certa specie. Ci si è per questo avvalsi della funzione *wilcox.test ()* presente nel pacchetto **stats**, la quale richiede molteplici argomenti in ingresso, tra i quali:

- i due vettori contenti i dati da confrontare;
- una stringa che specifichi se selezionare ipotesi bilaterali (*"two.sided"*) o unilaterali; in quest ultimo caso, se si opta per un'ipotesi alternativa in cui si affermi che la mediana della prima distribuzione sia significativamente maggiore o uguale alla mediana della seconda, allora la stringa è *"g"* (greater), nel caso contrario è *"l"* (less);
- un valore logico per specificare se i dati siano appaiati (*"T"*) o indipendenti (*"F"*);
- un valore logico per indicare se apportare o meno una correzione di continuità ai dati, nel caso essi non siano appunto continui.

Anche gli output sono molteplici, tra i quali nel nostro caso quello che riveste maggior interesse è il p-value, che si è confrontato con un livello di significatività pari a 0.05.
Per quanto concerne gli input della funzione nel presente studio, un vettore contiene dunque i valori di abbondanza relativa di una certa specie nel primo gruppo, l'altro vettore invece contiene quelli misurati nel secondo gruppo; per quanto riguarda la natura dei dati, essi sono sempre indipendenti e si richiede correzione di continuità.
Il test si è applicato a tutte le specie e selezionando ipotesi bilaterali; in questo caso l'ipotesi nulla afferma che i due gruppi posti a confronto non differiscono significativamente a livello di abbondanza di un certo genere o di una certa OTU. L'ipotesi alternativa asserisce invece il contrario. Dopo aver calcolato i p-value per ciascun confronto, si sono corretti con il metodo FDR, in quanto si è in presenza di test multipli; il livello di significatività con cui si confrontano i p-value è pari a 0.05.
In tal modo, le specie alle quali si è associato un p-value inferiore allo 0.05 presentano abbondanze significativamente diverse nei due gruppi.
Dopo aver svolto i test, si sono ordinati i p-value in ordine crescente, ottenendo dunque le seguenti informazioni:

- i nomi delle specie che presentano abbondanze significativamente diverse nei due gruppi posti a confronto, ordinate in modo tale da avere nelle prime posizioni le specie con un'abbondanza media di molto diversa nei due gruppi, mentre nelle ultime posizioni quelle specie la cui abbondanza media è di poco diversa nei due gruppi;
- per ciascuna specie si evidenzia in quale gruppo tra i due confrontati essa presenta un'abbondanza media maggiore.

8 Risultati e discussione

Nel presente capitolo si riportano i risultati delle analisi effettuate nei due casi di studio, partendo da quello relativo alla BPCO ed a seguire quello inerente al tumore colon-rettale.

8.1 Risultati delle analisi sui dati relativi alla BPCO

8.1.1 Misura della biodiversità intra- e inter-soggetto

Innanzitutto si è calcolata la diversità alfa (calcolata a partire dai valori di abbondanza relativa riferiti a generi e OTU) all'interno di ciascun soggetto, prendendo in considerazione gli indici di Shannon, Evenness e di Simpson, inteso nella forma 1-D, (si veda Paragrafo 4.4) e nelle tabelle 3 ed 4 si riportano i risultati ottenuti.

Si ricorda di seguito il significato del valore minimo e massimo degli indici.
Nel caso di Shannon e Simpson il valore minimo (zero) indica che non ho alcuna specie presente nel soggetto in questione, mentre il valore massimo suggerisce un'elevata abbondanza di specie (in media) all'interno di un dato soggetto. Per quanto riguarda l'indice di Evenness invece, lo zero indica che in quel soggetto si è in presenza di una sola specie (e quindi di massima omogeneità nella distribuzione delle specie), mentre quando l'indice assume il valore massimo (uno), si registra massima eterogeneità nel soggetto e dunque tutte le specie sono presenti in egual quantità.

GENERI												
	BP1	**BP2**	**BP3**	**BP4**	**BP5**	**F1**	**F2**	**F3**	**F4**	**NF1**	**NF2**	**NF3**
Shannon [0-5]	2.620	1.804	2.143	2.506	2.340	2.177	2.570	2.619	2.186	1.477	2.564	1.750
Evenness [0-1]	0.640	0.454	0.530	0.628	0.587	0.554	0.633	0.651	0.553	0.409	0.662	0.452
Simpson [0-1]	0.883	0.635	0.772	0.837	0.836	0.809	0.892	0.880	0.836	0.673	0.886	0.647

***Tabella 3**: Indici di diversità alfa (Shannon, Evenness e Simpson, sulle righe) calcolati per ciascun soggetto, dai valori di abbondanza relativa dei diversi generi. I soggetti, sulle colonne, sono divisi secondo le classi di appartenenza: fumatori malati (BP, in blu), fumatori sani (F, in giallo) e non fumatori sani (NF, in magenta).*

OTU												
	BP1	**BP2**	**BP3**	**BP4**	**BP5**	**F1**	**F2**	**F3**	**F4**	**NF1**	**NF2**	**NF3**
Shannon [0-5]	5.189	4.510	4.954	4.938	4.806	4.735	4.887	4.842	4.586	3.648	4.651	4.398
Evenness [0-1]	0.837	0.766	0.806	0.836	0.789	0.820	0.802	0.794	0.763	0.697	0.772	0.761
Simpson [0-1]	0.988	0.972	0.980	0.986	0.969	0.980	0.984	0.981	0.974	0.949	0.975	0.970

***Tab. 4**: Indici di diversità alfa (Shannon, Evenness e Simpson, sulle righe) calcolati per ciascun soggetto, dai valori di abbondanza relativa delle diverse OTU. I soggetti, sulle colonne, sono divisi secondo le classi di appartenenza: fumatori malati (BP, in blu), fumatori sani (F, in giallo) e non fumatori sani (NF, in magenta).*

Per semplicità d'interpretazione, le tabelle sopra riportate son state ulteriormente rappresentate mediante boxplot, ponendo sull'asse delle ascisse i soggetti mentre sulle ordinate si osservano i valori degli indici di diversità alfa.

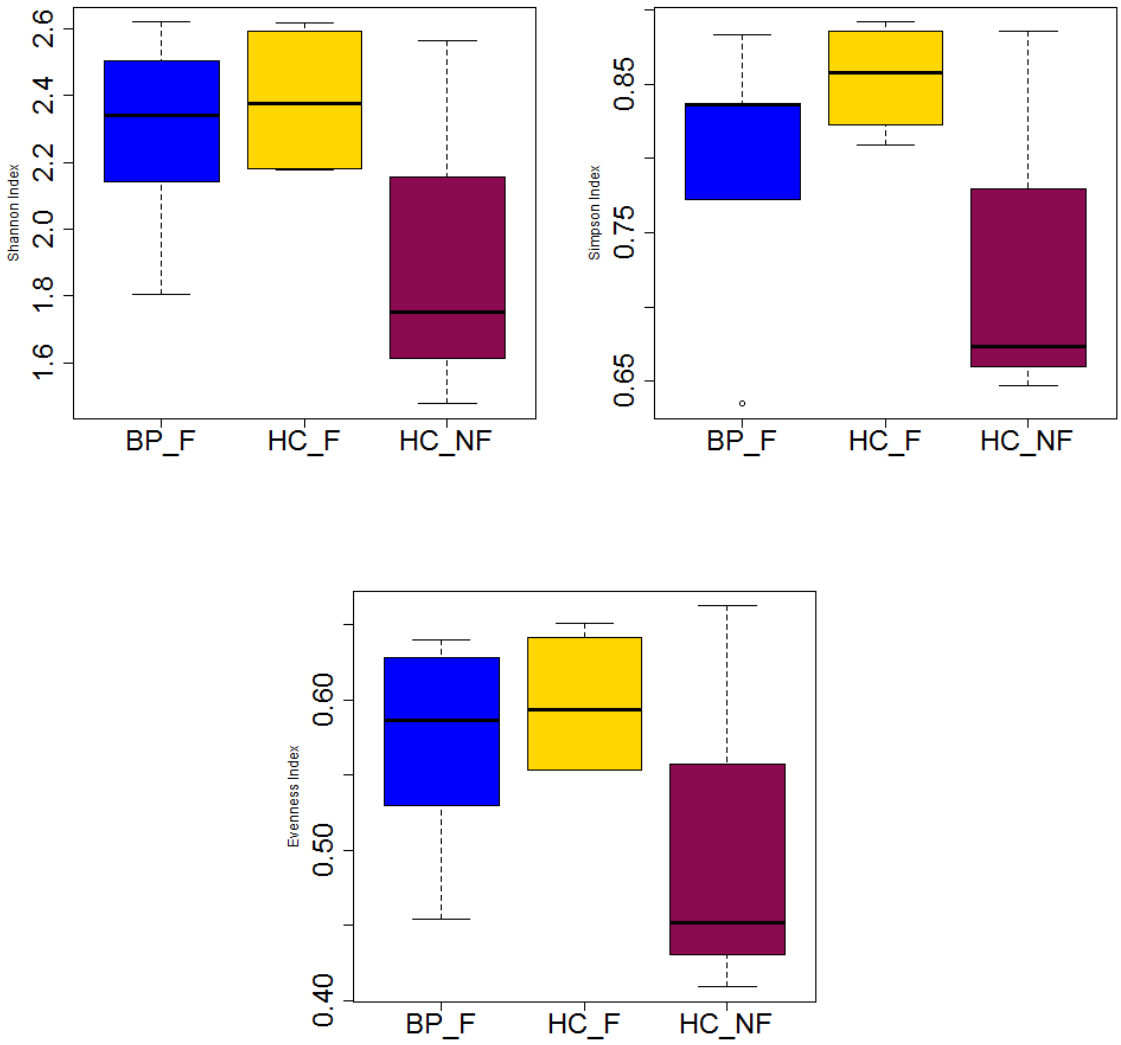


***Figura 28*: .** *Boxplot dei valori degli indici di Shannon, Evenness e Simpson misurati nei fumatori malati (BP_F),nei fumatori sani (HC_F) e nei non fumatori sani (HC_NF), a partire dalle abbondanze relative dei diversi generi.*

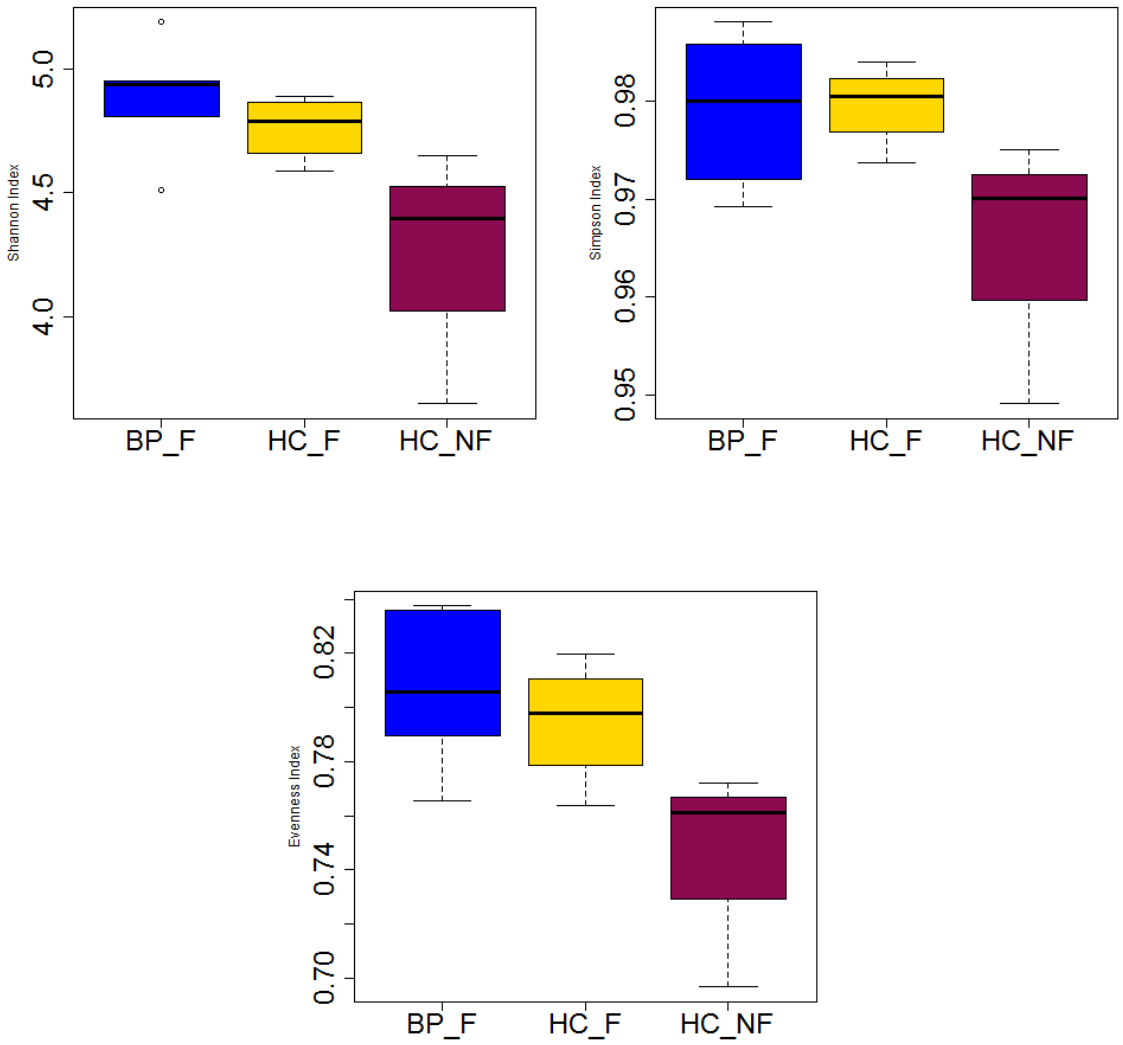
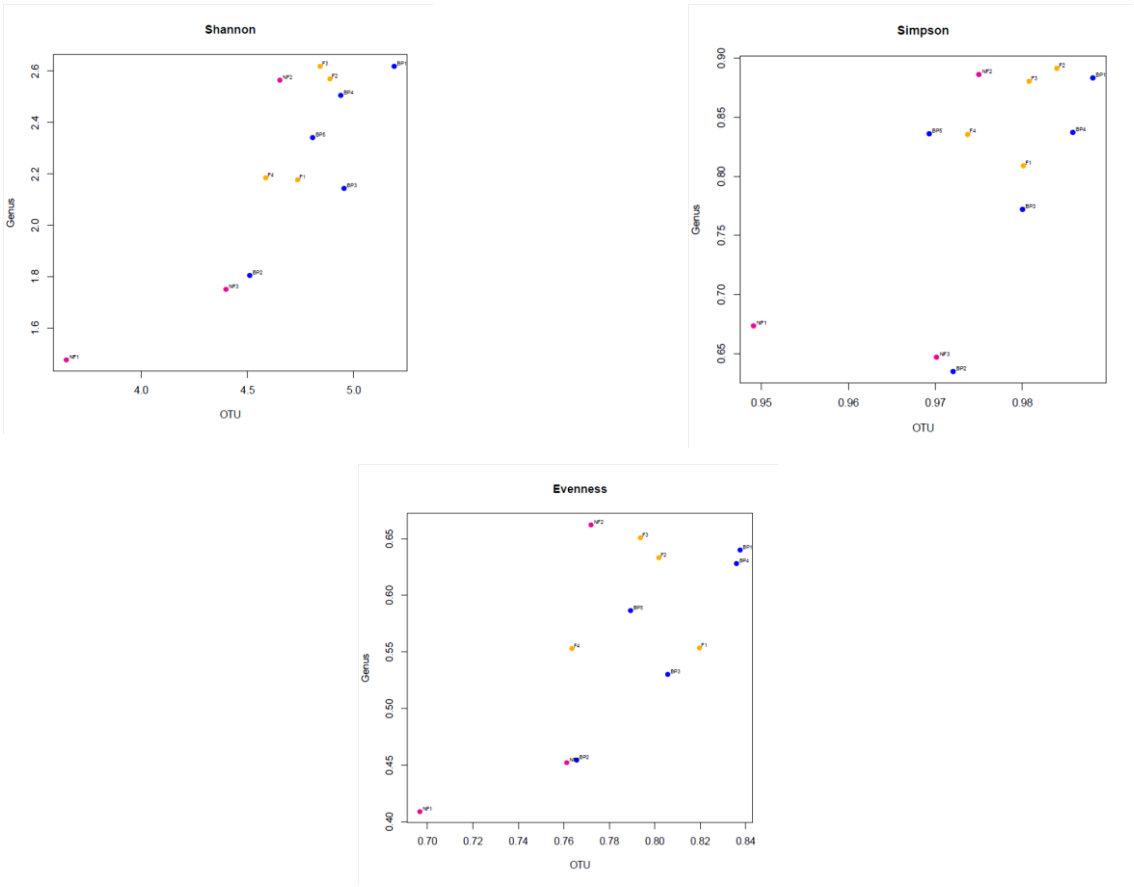


***Figura 29*: .** *Boxplot dei valori degli indici di Shannon, Evenness e Simpson misurati nei fumatori malati (BP_F),nei fumatori sani (HC_F) e nei non fumatori sani (HC_NF), a partire dalle abbondanze relative dei diversi OTU.*

Si nota come in tutti i soggetti la diversità stimata sia maggiore nei dati relativi alle OTU rispetto ai dati relativi ai generi.



(A) **(B)** **(C)**

Figura 30*: rappresentazione dei valori di diversità alfa (Shannon nel grafico* ***A****, Evenness nel grafico* ***B*** *, Simpson nel grafico* ***C****) stimati per i soggetti fumatori malati (BP, in blu), fumatori sani (F, in giallo) e non fumatori sani (NF, in magenta) confrontando i risultati ottenuti a partire dai valori di abbondanza relativa dei diversi generi (sull'asse delle ordinate) e delle diverse OTU (sull'asse delle ascisse).*

Osservando i grafici si nota che tutti e tre gli indici assumono valori più elevati nel caso in cui si parte dai valori di abbondanza delle diverse OTU, anziché dai generi; in entrambi i casi comunque i soggetti che presentano maggiore diversità alfa ed eterogeneità nella distribuzione delle specie sono i fumatori malati ed i fumatori sani. I non fumatori sani presentano invece indici con valori inferiori, perciò in tali soggetti si riscontrano quantità inferiori di specie batteriche rispetto ai fumatori (sani e malati) ed inoltre nei non fumatori sani la varietà dei generi e delle OTU risulta essere minore.

Si evidenziano inoltre due eccezioni, rappresentate da un non fumatore sano (NF2) e da un fumatore malato (BP2); infatti il soggetto identificato con la sigla NF2 presenta valori di diversità alfa che portano a catalogarlo come fumatore, mentre il soggetto BP2 presenta indici con valori simili a quelli solitamente assunti da non fumatori.

In seguito si è calcolata la diversità beta, dapprima tra i soggetti appartenenti alla stessa categoria e poi tra i soggetti delle differenti categorie. I risultati ottenuti a proposito della diversità beta tra soggetti della stessa categoria sono riportati nella tabella 5 per quanto concerne il dataset in cui le abbondanze relative sono riferite ai generi, mentre nella tabella 6 si mostrano i risultati inerenti al dataset di partenza in cui si prendono in considerazione le OTU anziché i generi.

Si ricorda che il valore minimo degli indici (zero) indica che nei soggetti confrontati sono presenti le medesime specie, mentre se gli indici assumono il valore massimo (uno) allora i soggetti posti a confronto non hanno alcuna specie comune. Inoltre i valori riportati per ciascun gruppo, sono le medie dei valori degli indici calcolati confrontando ogni possibile coppia di soggetti all'interno del medesimo gruppo.

GENERI				
	BP_F	**HC_F**	**HC_NF**	**HC_F e HC_NF**
Sørensen [0-1]	0.160	0.214	0.359	0.289
Jaccard [0-1]	0.275	0.350	0.525	0.439
Harte-Kinzig [0-1]	0.160	0.214	0.359	0.289

***Tabella 5**: sulle colonne si riportano i gruppi, i cui soggetti si son posti a confronto per il calcolo della diversità beta, sulle righe sono presenti gli indici (Sørensen, Jaccard nella sua forma inversa e Harte-Kinzig). I dati di partenza sono le abbondanze relative dei diversi generi.*

OTU				
	BP_F	**HC_F**	**HC_NF**	**HC_F e HC_NF**
Sørensen [0-1]	0.363	0.392	0.544	0.289
Jaccard [0-1]	0.531	0.562	0.699	0.439
Harte-Kinzig [0-1]	0.363	0.392	0.544	0.476

***Tabella 6**: sulle colonne si riportano i gruppi, i cui soggetti si son posti a confronto per il calcolo della diversità beta, sulle righe sono presenti gli indici (Sørensen, Jaccard nella sua forma inversa e Harte-Kinzig). I dati di partenza sono le abbondanze relative delle diverse OTU.*

Anche in questo caso si sono riprodotti i risultati in grafici a barre, in cui sulle ascisse sono riportati i gruppi, mentre le ordinate rappresentano i valori degli indici di beta diversità. Si tralasciano i dati relativi ai soggetti sani, in quanto sono semplicemente la media dei valori calcolati nei fumatori sani e nei non fumatori sani.

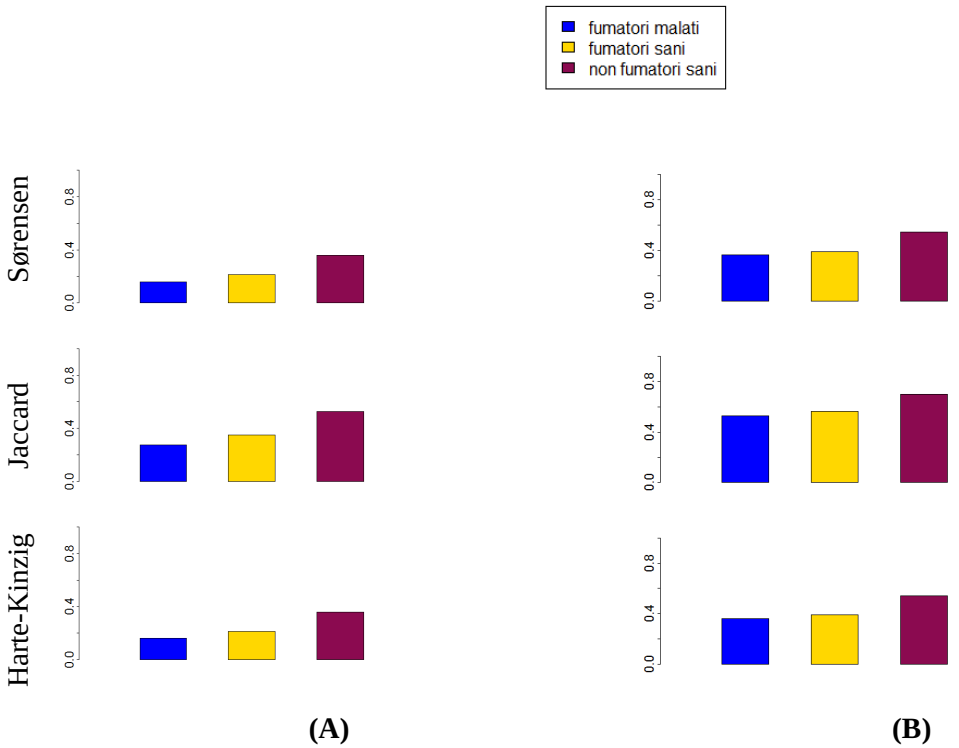


Figura 31: *rappresentazione della media dei valori degli indici di diversità beta calcolati tra soggetti appartenenti ad uno stesso gruppo (fumatori malati, in blu; fumatori sani, in giallo; non fumatori sani, in magenta). I dati di partenza sono le abbondanze relative dei diversi generi (**A**) e delle diverse OTU (**B**).*

Si è successivamente testata la significatività delle differenze tra i gruppi applicando il metodo NPMANOVA (si veda Paragrafo 7.3). Si sono infatti posti a confronto i valori di abbondanza relativa dei diversi generi e delle diverse OTU dei soggetti appartenenti a gruppi diversi; tramite il test si sono dunque comparati i valori di abbondanza relativa media di ciascun gruppo, e osservando i p-value ottenuti si è concluso quali gruppi possono considerarsi significativamente diversi (in base all'abbondanza relativa dei generi o OTU). Di seguito nelle tabelle 7 e 8 si mostrano i p-value ottenuti dai vari test, ricordando che nel caso in cui il p-value assuma valori inferiori allo 0.05, si assume l'ipotesi alternativa e dunque si asserisce che sussistono differenze significative tra i valori medi di abbondanza relativa dei due gruppi posti a confronto. È da specificarsi che un valore del p-value inferiore alla soglia dello 0.05, indica che si ammette una probabilità inferiore al 5% di rifiutare l'ipotesi nulla del test, quando in realtà l'ipotesi nulla è vera. Dunque più basso è il valore del p-value, minore è l'errore commesso nel considerare

diverse le medie delle abbondanze relative di generi o OTU dei due gruppi confrontati, quando in realtà non lo sono.

Di conseguenza, nei risultati sotto riportati, più il p-value è inferiore alla soglia, più aumenta la significatività della differenza tra i gruppi posti a confronto (e dunque la veridicità dei risultati stessi).

GENERI					
	BP_F e HC_F vs HC_NF	**BP_F vs HC_F**	**BP_F vs HC_NF**	**HC_F vs HC_NF**	**BP_F vs HC_F e HC_NF**
p-value (Statistica F)	0.021** (3.900)	0.284 (1.312)	0.034** (3.342)	0.145 (2.369)	0.151 (1.770)

***Tabella 7**: p-value e valore della statistica F (tra parentesi), risultati dall'applicazione di NPMANOVA alle abbondanze relative dei diversi generi, misurate in soggetti appartenenti a gruppi diversi (specificati sulle colonne).*

*(Livelli di significatività del p-value: 0.01 '***' 0.05 '**' 0.1 ' *')*

OTU					
	BPF e HC_F vs HC_NF	**BP_F vs HC_F**	**BP_F vs HC_NF**	**HC_F vs HC_NF**	**BP_F vs HC_F e HC_NF**
p-value (Statistica F)	0.026** (1.776)	0.588 (0.922)	0.053* (1.749)	0.257 (1.297)	0.304 (1.120)

***Tabella 8**: p-value e valore della statistica F (tra parentesi), risultati dall'applicazione di NPMANOVA alle abbondanze relative delle diverse OTU, misurate in soggetti appartenenti a gruppi diversi (specificati sulle colonne).*

*(Livelli di significatività del p-value: 0.01 '***' 0.05 '**' 0.1 ' *')*

I valori della statistica F calcolati con NPMANOVA sono di seguito riportati in un grafico a barre, la cui altezza ne indica il valore, mentre gli asterischi specificano il livello di significatività del p-value associato.

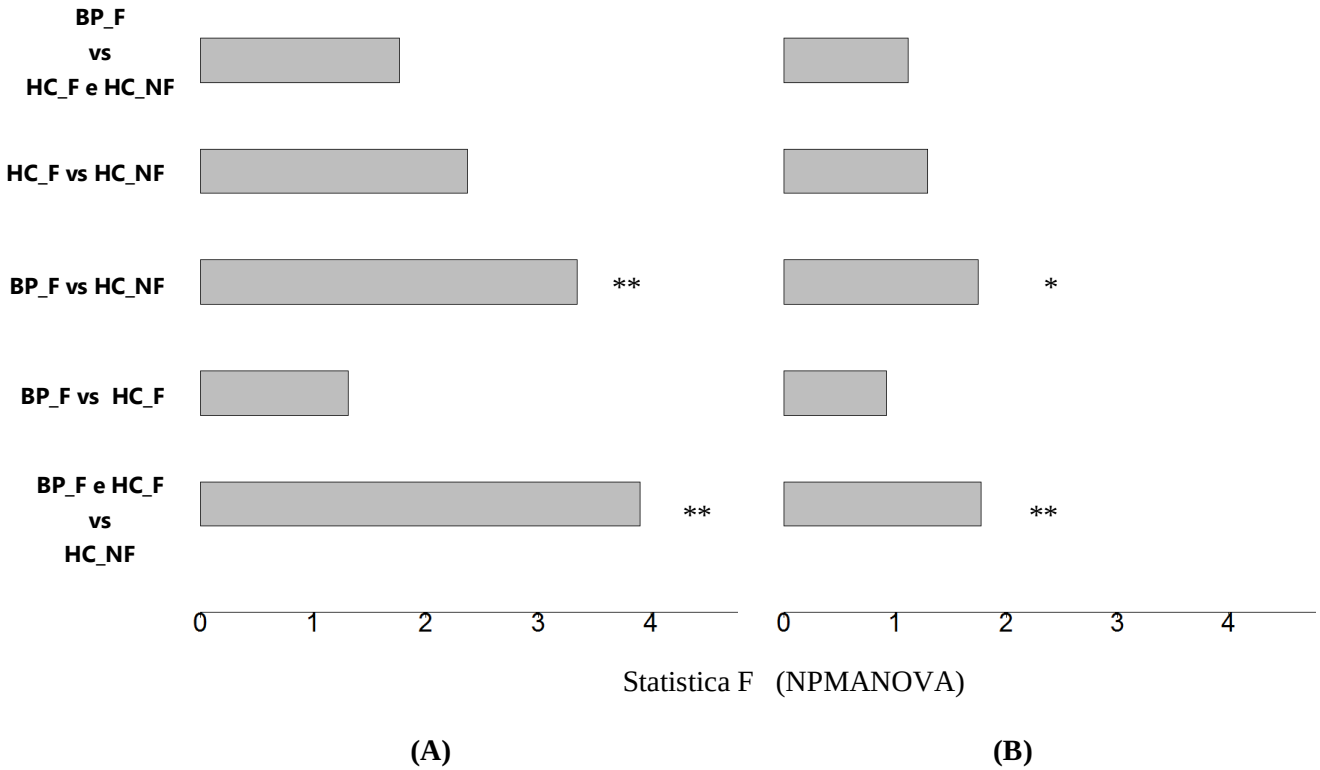


Figura 32: *rappresentazione dei valori della statistica F ottenuti dall'applicazione di NPMANOVA alle abbondanze relative dei diversi generi (**A**) e delle diverse OTU (**B**), misurate in soggetti appartenenti a gruppi diversi.*

*(Livelli di significatività del p-value: 0.01 '***' 0.05 '**' 0.1 ' *')*

Come mostrato nella figura 32, indipendentemente dal fatto che gli indici siano calcolati a partire dalle abbondanze relative dei diversi generi piuttosto che delle diverse OTU, si evidenzia che:

- i valori più bassi di diversità beta si osservano nel confronto tra fumatori malati e fumatori sani, così come tra fumatori malati e soggetti sani (fumatori e non);
- i valori più alti di diversità beta si misurano tra fumatori (malati e non) e non fumatori sani, e tra fumatori malati e non fumatori sani.

Come già notato nel caso della diversità alfa, i valori degli indici calcolati a partire dalle abbondanze relative delle diverse OTU risultano maggiori rispetto a quelli calcolati sui dati che considerano i generi.

Per poter interpretare in maniera adeguata i risultati ottenuti dal calcolo degli indici di diversità beta, è fondamentale considerare i p-value calcolati con NPMANOVA ed osservarne la significatività.
Innanzitutto si osserva che, sia confrontando le abbondanze relative dei generi che quelle delle OTU, solamente in due casi il p-value calcolato con NPMANOVA è inferiore alla soglia dello 0.05 e perciò si può concludere che sussistano differenze significative tra i due gruppi. Nello specifico, in base ai p-value riportati nelle tabelle 7 e 8, i gruppi che possono considerarsi significativamente diversi sono:

- **fumatori (malati e sani) e non fumatori sani**, con relativo p-value pari a 0.021 confrontando le abbondanze relative dei generi e pari allo 0.026 nel caso delle OTU;
- **fumatori malati e non fumatori sani**, con p-value pari a 0.034 considerando le abbondanze relative dei generi e p-value pari a 0.053 considerando le OTU.

Si pone in evidenza che, con l'aumentare del numero di dati su cui si effettua il confronto tra gruppi (è il caso delle abbondanze relative delle OTU rispetto a quelle dei generi) si ottengono risultati più affidabili e nel nostro caso anche p-value più elevati. Di conseguenza ci si focalizza sui risultati ottenuti a partire dalle abbondanze relative delle OTU e si nota che i fumatori (malati e non) ed i non fumatori sani presentano differenze ancor più significative rispetto a quelle osservate tra fumatori malati e non fumatori sani.
Dalle tabelle 7 e 8 si evince inoltre che il p-value con il valore più elevato si osserva in corrispondenza del confronto tra fumatori malati e fumatori sani; considerando che in questo caso si sono calcolati anche i valori più bassi di diversità beta, si può concludere che i soggetti dei due gruppi non presentano significative differenze (se supponessi il contrario ci sarebbe una probabilità di circa 60% di commettere un errore, in quanto il p-value nella tabella 10 è pari a 0.588).
Si denota inoltre che la significatività della differenza tra fumatori malati e non fumatori sani calcolata confrontando le abbondanze relative dei diversi generi diminuisce nel momento in cui si considerano le abbondanze delle OTU (il p-value nel primo caso è inferiore alla soglia dello 0.05, mentre nel secondo caso è di poco superiore). Rimane invece significativa la differenza tra fumatori (malati e sani) e non fumatori sani in entrambi i casi, così come la differenza tra fumatori malati e fumatori sani.
In seguito all'interpretazione dei risultati, si nota perciò che il fumo comporta cambiamenti nell'abbondanza relativa delle specie componenti il microbioma. Per validare ulteriormente questa conclusione, si interpretano i risultati ottenuti mediante ANOSIM, valutando se possono essere confrontabili con quelli sopra riportati.

8.1.2 Analisi delle similarità dei soggetti con ANOSIM

I risultati dell'analisi delle similarità effettuata con il metodo ANOSIM (si veda paragrafo 7.4) sono presentati nelle tabelle 9 e 10, nella prima ci si riferisce all'analisi applicata ai dati di abbondanza relativa dei diversi generi, mentre nella seconda i dati di partenza sono le abbondanze relative delle diverse OTU.

GENERI					
	BP_F e HC_F vs HC_NF	**BP_F vs HC_F**	**BP_F vs HC_NF**	**HC_F vs HC_NF**	**BP_F vs HC_F e HC_ NF**
p-value (Statistica R)	0.014** (0.666)	0.217 (0.106)	0.037** (0.651)	0.113 (0.389)	0.201 (0.076)

***Tabella 9**: p-value e valore della statistica R (tra parentesi), risultati dall'applicazione di ANOSIM alle abbondanze relative dei diversi generi, misurate in soggetti appartenenti a gruppi diversi (specificati sulle colonne).*

*(Livelli di significatività del p-value: 0.01 '***' 0.05 '**' 0.1 ' *')*

OTU					
	BP_F e HC_F vs HC_NF	**BP_F vs HC_F**	**BP_F vs HC_NF**	**HC_F vs HC_NF**	**BP_F vs HC_F e HC_NF**
p-value (Statistica R)	0.039** (0.455)	0.449 (-0.006)	0.034** (0.477)	0.266 (0.167)	0.594 (-0.039)

***Tabella 10**: p-value e valore della statistica R (tra parentesi), risultati dall'applicazione di ANOSIM alle abbondanze relative delle diverse OTU, misurate in soggetti appartenenti a gruppi diversi (specificati sulle colonne).*

*(Livelli di significatività del p-value: 0.01 '***' 0.05 '**' 0.1 ' *')*

Per poter visualizzare in modo più immediato i risultati di tale analisi si è ricorsi ancora una volta a dei grafici a barre (Figura 33), in cui si sono rappresentati i valori della statistica R (altezza delle barre) risultanti da ANOSIM applicata su tutti i possibili gruppi: fumatori (sani e malati)/non fumatori sani, fumatori malati/fumatori sani, fumatori malati/non fumatori sani, fumatori sani/non fumatori sani e fumatori malati/fumatori e non fumatori sani. Gli asterischi

riportati invece sopra le barre identificano il livello di significatività dei p-value associati alla statistica.

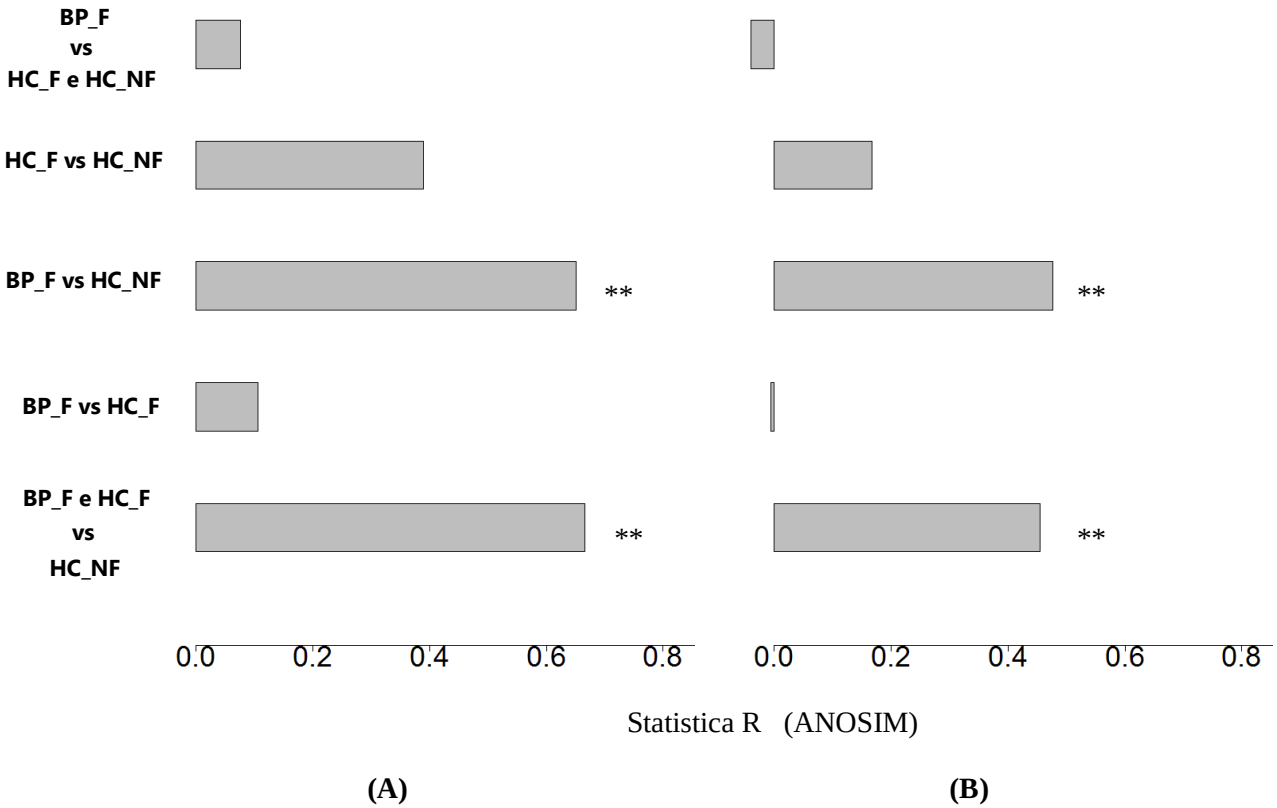


Figura 33: *rappresentazione dei valori della statistica R ottenuti dall'applicazione di ANOSIM alle abbondanze relative dei diversi generi (**A**) e delle diverse OTU (**B**), misurate in soggetti appartenenti a gruppi diversi.*

*(Livelli di significatività del p-value: 0.01 '***' 0.05 '**' 0.1 ' *')*

Si nota che anche ANOSIM, così come NPMANOVA, produce dei p-value inferiori alla soglia dello 0.05 solo nei due confronti già citati in precedenza:

- **fumatori (malati e sani) e non fumatori sani**, con relativo p-value pari a 0.014 confrontando le abbondanze relative dei generi e pari allo 0.039 nel caso delle OTU;
- **fumatori malati e non fumatori sani**, con p-value pari a 0.037 considerando le abbondanze relative dei generi e p-value pari a 0.034 considerando le OTU.

Le conclusioni che possono trarsi in seguito a questi risultati sono le stesse di quelle viste in conseguenza all'analisi NPMANOVA. Si nota infatti una differenza significativa tra fumatori (malati e sani) e non fumatori sani, la quale risulta però in tal caso di poco inferiore a quella riscontrata tra fumatori malati e non fumatori sani. Inoltre i p-value relativi al confronto tra

fumatori malati e fumatori sani risultano essere molto superiori alla soglia di significatività, per la precisione pari a 0.217 nella tabella 9 ed a 0.449 nella tabella 10. Si può dunque concludere che fumatori sani e fumatori malati non presentano differenze significative nella composizione del microbioma.

Inoltre, diversamente da quanto evidenziato nei risultati di NPMANOVA, non c'è una diminuzione di significatività nella differenza tra fumatori malati e non fumatori sani a seconda che si confrontino abbondanze relative dei generi o delle OTU. La significatività delle differenze tra fumatori (malati e sani) e non fumatori sani e quella tra fumatori malati e fumatori sani rimangono invariate in entrambi i casi, come visto anche per NPMANOVA.

Si osserva anche che così come evidenziato nel paragrafo 8.1.1, i p-value ottenuti a partire dall'analisi delle abbondanze relative delle diverse OTU sono maggiori di quelli calcolati a partire dalle abbondanze dei generi.

In base alla concordanza tra risultati dei due approcci NPMANOVA e ANOSIM, visibile in figura 34, si può quindi concludere che il fumo può ritenersi una possibile causa nella variazione della composizione del microbioma umano e un'importante covariabile legata alla malattia (si ricorda la similarità tra soggetti fumatori malati e fumatori sani) .

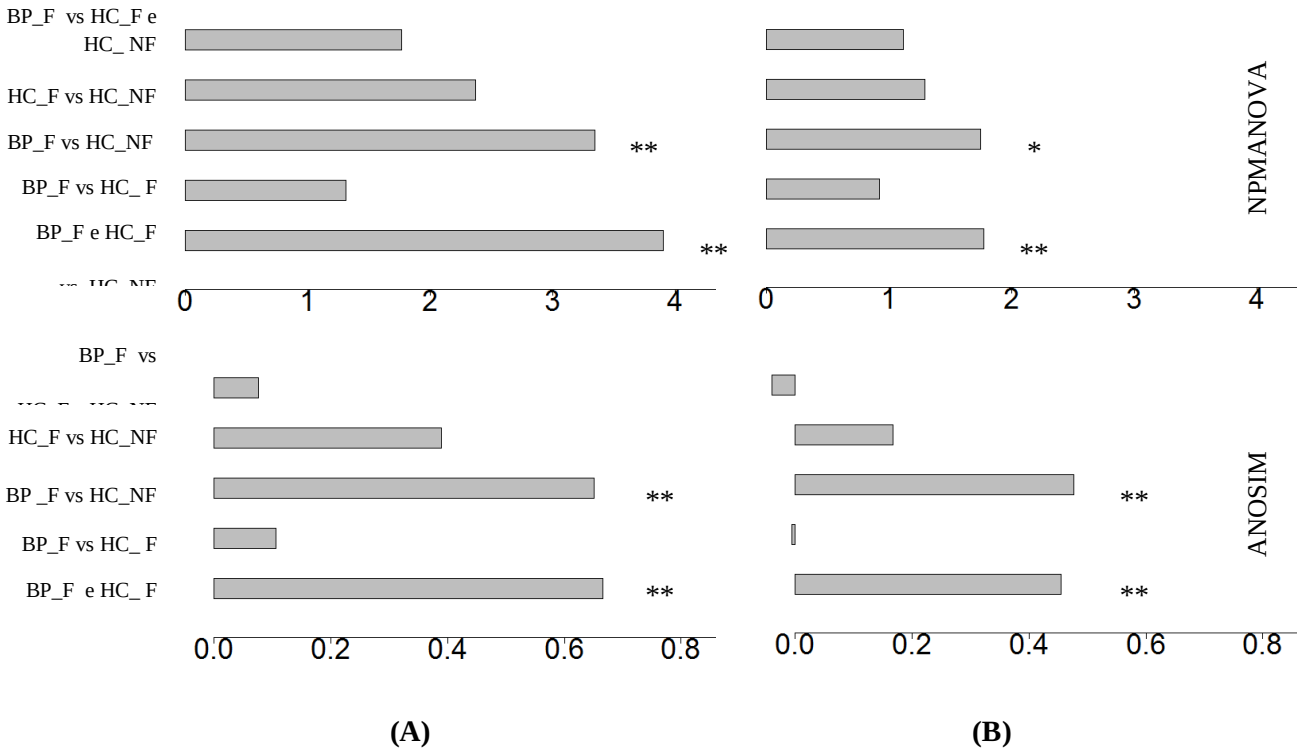


Figura 34. *Confronto delle statistiche ottenute da NPMANOVA (riquadri in alto) e ANOSIM (riquadri in basso) a partire dalle abbondanze relative dei generi (**A**) e delle OTU (**B**). In ascissa si riportano i gruppi posti a confronto.*

8.1.3 Differenze nella composizione microbiotica tra soggetti

Nelle tabelle seguenti si riportano i risultati del test di Wilcoxon applicato alle abbondanze relative dei diversi generi e delle diverse OTU per il confronto tra tutte le possibili coppie di gruppi.

Precisamente, nella prima colonna sono riportati i generi o le OTU le cui abbondanze relative sono diverse tra i due gruppi, avendo imposto un livello di significatività pari a 0.05; si è poi ordinata la lista finale secondo valori di p-value crescenti, i quali sono inseriti nella seconda colonna dopo una correzione FDR. Si mostrano inoltre nella terza e quarta colonna le medie delle abbondanze relative di ciascun genere o OTU all'interno dei gruppi confrontati. Si evidenzia infine il gruppo nel quale un certo genere o OTU presenta maggiore abbondanza media.

Nella tabella 11 si riportano i risultati del confronto tra abbondanze relative dei generi tra fumatori (malati e sani) e non fumatori.

Generi	pvalue	pvalue corretti	BP_F e HC_F	HC_NF
Atopobium	0.0091	0.2606	0.0051	7.00E-04
Hafnia	0.0157	0.2606	0	0.0078
Klebsiella	0.0157	0.2606	0	0.0058
Pectobacterium	0.0157	0.2606	0	0.0008
Tatumella	0.0157	0.2606	0	0..0006
Actinomyces	0.0364	0.3299	0.0402	0.0134
Campylobacter	0.0364	0.3299	0.0139	0.0036
Phocaeicola	0.0364	0.3299	0.0013	0.0001
…	…	…	…	…

***Tabella 11**: Generi le cui abbondanze relative sono diverse (Test di Wilcoxon, livello di significatività 5%) tra fumatori (malati e sani) e non fumatori sani, e relativi p-value. Per ciascuna specie è evidenziato il gruppo in cui essa è più abbondante.*

Nella tabella 12 si presentano i risultati del confronto tra abbondanze relative dei generi tra fumatori malati e fumatori sani.

Generi	pvalue	pvalue corretti	BP_F	HC_F
Neisseria	0.0159	0.8031	0.0116	0.1284
Moryella	0.0317	0.8031	0.0079	0.0022
Roseburia	0.0317	0.8031	0.003	0.0016
Tessaracoccus	0.0442	0.8398	0.0015	0
…	…	…	…	…

***Tabella 12**: Generi le cui abbondanze relative sono diverse (Test di Wilcoxon, livello di significatività 5%) tra fumatori malati e fumatori sani, e relativi p-value. Per ciascuna specie è evidenziato il gruppo in cui essa è più abbondante.*

Nella tabella 13 si mostrano i risultati del confronto tra abbondanze relative dei generi tra fumatori malati e non fumatori sani.

Generi	pvalue	pvalue corretti	BP_F	HC_NF
Atopobium	0.0357	0.4627	0.0051	0.0007
Campylobacter	0.0357	0.4627	0.0092	0.0036
Prevotella	0.0357	0.4627	0.31	0.0416
…	…		…	…

***Tabella 13**: Generi le cui abbondanze relative sono diverse (Test di Wilcoxon, livello di significatività 5%) tra fumatori malati e non fumatori sani, e relativi p-value. Per ciascuna specie è evidenziato il gruppo in cui essa è più abbondante.*

Nella tabella 14 si riportano i risultati del confronto tra abbondanze relative dei generi tra fumatori sani e non fumatori sani.

Generi	pvalue	pvalue corretti	F	NF
Lachnobacterium	0.0497	0.7535	0.0013	0.0001
Atopobium	0.0571	0.7535	0.0052	0.0007
Actinomyces	0.1143	0.7535	0.0427	0.0134
Leptotrichia	0.1143	0.7535	0.0603	0.0112
…	…	…	…	…

***Tabella 14**: Generi le cui abbondanze relative sono diverse (Test di Wilcoxon, livello di significatività 5%) tra fumatori sani e non fumatori sani, e relativi p-value. Per ciascuna specie è evidenziato il gruppo in cui essa è più abbondante.*

Nella tabella 15 si presentano i risultati del confronto tra abbondanze relative dei generi tra fumatori malati e fumatori e non fumatori sani.

Generi	pvalue	pvalue corretti	BP_F	HC_F e HC_NF
Tessaracoccus	0.0091	0.6943	0.0015	0
Moryella	0.0303	0.6943	0.0079	0.0024
Paraprevotella	0.0344	0.6943	0.0036	0.0004
Atopobium	0.048	0.6943	0.0051	0.0032
…	…	…	…	…

***Tabella 15**: Generi le cui abbondanze relative sono diverse (Test di Wilcoxon, livello di significatività 5%) tra fumatori malati e fumatori e non fumatori sani, e relativi p-value. Per ciascuna specie è evidenziato il gruppo in cui essa è più abbondante.*

Da quanto concluso nei paragrafi 8.1.1 e 8.1.2, sono due i casi in cui il microbiota dei soggetti appartenenti ad un gruppo presenta maggiori differenze nella composizione rispetto al microbiota dei soggetti appartenenti ad un altro gruppo. Si considerano quindi in primis questi due casi, osservando i risultati delle tabelle 11 e 13. Innanzitutto, sia confrontando i fumatori (malati e sani) con i non fumatori sani sia i fumatori malati con i non fumatori sani, si nota che nei fumatori indipendentemente dallo stato di salute si ritrova una maggiore abbondanza di *Atopobium* e *Abiotrophium* rispetto ai non fumatori sani. Si potrebbe dunque concludere che la presenza di elevate quantità di questi due generi sia dovuta al fumo, così come si può osservare

che nei fumatori sono assenti quattro generi presenti invece nei non fumatori: *Hafnia*, *Klebsiella*, *Pectobacterium* e *Tatumella*.
Considerando poi la tabella 15, che pone a confronto i fumatori malati ed i soggetti sani (fumatori e non), si può osservare l'influenza causata dalla malattia a livello di abbondanze relative dei generi; si evidenzia in particolare che i soggetti malati presentano una quantità maggiore di *Catonella* e *Atopobium* rispetto ai soggetti sani.
Si denota però che, in base ai valori dei p-value, non si evidenziano differenze significative; questo può esser dovuto al fatto che il dataset a disposizione non è sufficientemente numeroso per un analisi di questo tipo.

Dopo aver applicato il test di Wilcoxon alle abbondanze relative dei generi ci si è concentrati sulle abbondanze delle diverse OTU, ed i risultati del test sono presentati nelle tabelle seguenti.
Nella tabella 16 si osservano i risultati del confronto tra abbondanze relative delle OTU tra fumatori (malati e sani) e non fum1atori sani.

OTU	pvalue	pvalue corretti	BP_F e HC_F	HC_NF
Citrobacter_12427	0.0157	0.6355	0	0.0039
Citrobacter_12437	0.0157	0.6355	0	0.0717
Citrobacter_12455	0.0157	0.6355	0	0.0134
Citrobacter_12496	0.0157	0.6355	0	0.0044
Citrobacter_12507	0.0157	0.6355	0	0.0016
Citrobacter_12535	0.0157	0.6355	0	0.002
Citrobacter_12570	0.0157	0.6355	0	0.003
Citrobacter_12576	0.0157	0.6355	0	0.0017
Citrobacter_12605	0.0157	0.6355	0	0.003
Gemella_12715	0.0157	0.6355	0	0.0005
Citrobacter_13161	0.0157	0.6355	0	0.0143
Citrobacter_13349	0.0157	0.6355	0	0.0023
Citrobacter_13843	0.0157	0.6355	0	0.0012
Citrobacter_13964	0.0157	0.6355	0	0.0032
Citrobacter_14076	0.0157	0.6355	0	0.0008

Streptococcus_15390	0.0157	0.6355	0	0.0005
Prevotella_6	0.0182	0.6355	0.0053	0.0006
Streptococcus_9634	0.0233	0.6355	0	0.0016
Prevotella_855	0.0262	0.6355	0.0012	0
…	…		…	…

***Tabella 16**: OTU le cui abbondanze relative sono diverse (Test di Wilcoxon, livello di significatività 5%) tra fumatori (malati e sani) e non fumatori sani, e relativi p-value. Per ciascuna specie è evidenziato il gruppo in cui essa è più abbondante.*

Nella tabella 17 si riportano i risultati del confronto tra abbondanze relative delle OTU tra fumatori malati e fumatori sani.

OTU	pvalue	pvalue corretti	BP_F	HC_F
Streptococcus_8161	0.0108	1	0	0.0011
Leptotrichia_8658	0.0108	1	0	0.0009
Neisseria_11016	0.0108	1	0	0.0009
Prevotella_457	0.0179	1	0.0009	0.0001
Actinomyces_807	0.0179	1	0.0006	0.0001
Neisseria_6701	0.0179	1	0.0001	0.0011
…	…		…	…

***Tabella 17**: OTU le cui abbondanze relative sono diverse (Test di Wilcoxon, livello di significatività 5%) tra fumatori malati e fumatori sani, e relativi p-value. Per ciascuna specie è evidenziato il gruppo in cui essa è più abbondante.*

Nella tabella 18 si presentano i risultati del confronto tra abbondanze relative delle OTU tra fumatori malati e non fumatori sani.

OTU	pvalue	pvalue corretti	BP_F	HC_NF
Streptococcus_16295	0.0168	0.7313	0	0.0002
Prevotella_457	0.0325	0.7313	0.0009	0
Actinomyces_1958	0.0325	0.7313	0.0005	0
Fusobacterium_2070	0.0325	0.7313	0.0006	0
Prevotella_3005	0.0325	0.7313	0.0007	0
Leptotrichia_3734	0.0325	0.7313	0.0002	0
Prevotella_4487	0.0325	0.7313	0.0009	0
Prevotella_6137	0.0325	0.7313	0.0046	0
Oribacterium_6998	0.0325	0.7313	0.0005	0
Prevotella_6	0.0357	0.7313	0.0056	0.0006
Actinomyces_69	0.0357	0.7313	0.0017	0.0002
Streptococcus_16295	0.0168	0.7313	0	0.0002
...	...		...	...

***Tabella 18**: OTU le cui abbondanze relative sono diverse (Test di Wilcoxon, livello di significatività 5%) tra fumatori malati e non fumatori sani, e relativi p-value. Per ciascuna specie è evidenziato il gruppo in cui essa è più abbondante.*

Nella tabella 19 si osservano i risultati del confronto tra abbondanze relative delle OTU tra fumatori sani e non fumatori sani.

OTU	pvalue	pvalue corretti	HC_F	HC_NF
Megasphaera_1663	0.0436	0.8704	0.0006	0
Streptococcus_8161	0.0436	0.8704	0.0011	0
Leptotrichia_30943	0.0436	0.8704	0.001	0
Prevotella_481	0.0497	0.8704	0.0005	0
Rothia_2111	0.0497	0.8704	0.0001	0.0005
Streptococcus_4000	0.0497	0.8704	0.0001	0.0007
...	...	...	...	...

***Tabella 19**: OTU le cui abbondanze relative sono diverse (Test di Wilcoxon, livello di significatività 5%) tra fumatori sani e non fumatori sani, e relativi p-value. Per ciascuna specie è evidenziato il gruppo in cui essa è più abbondante.*

Nella tabella 20 si presentano i risultati del confronto tra abbondanze relative delle OTU tra fumatori malati e soggetti sani (fumatori e non).

OTU	pvalue	pvalue corretti	BP_F	HC_F e HC_NF
Prevotella_457	0.0032	0.8704	0.0009	0
Leptotrichia_294	0.0091	0.8704	0.001	0
Prevotella_777	0.0091	0.8704	0.001	0
Prevotella_1112	0.0091	0.8704	0.0005	0
Gemella_436	0.0092	0.8704	0.0024	0.0004
Actinomyces_174	0.0101	0.8704	0.0035	0.0006
TM7_genera_incertae_sedis_2066	0.0115	0.8704	0.0008	0.0001
Actinomyces_807	0.0131	0.8704	0.0006	0.0001

Solobacterium_2678	0.0131	0.8704	0.0008	0.0002
Atopobium_2815	0.0147	0.8704	0.0025	0.0005
Hallella_28	0.0177	0.8704	0.0016	0.0004
Prevotella_837	0.0186	0.8704	0.001	0
…	…	…	…	…

***Tabella 20**: OTU le cui abbondanze relative sono diverse (Test di Wilcoxon, livello di significatività 5%) tra fumatori malati e fumatori e non fumatori sani, e relativi p-value. Per ciascuna specie è evidenziato il gruppo in cui essa è più abbondante.*

Osservando la tabella 16, in cui si sono confrontati fumatori (sani e malati) e non fumatori sani, si pone in evidenza l'effetto del fumo sulla composizione del microbiota umano. Valutando dunque la distribuzione delle OTU nei due gruppi si può concludere se il fumo porti alla proliferazione o meno di una certa specie all'interno dei soggetti. Nello specifico si evidenzia che la maggior parte delle OTU riferite al *Citrobacter* sono assenti nei fumatori rispetto ai non fumatori. Inoltre, considerando le OTU *Prevotella* , esse sono presenti in quantità maggiori nei fumatori rispetto ai non fumatori.

Considerando in seguito la tabella 18, si denota che nei fumatori malati le OTU relative a *Prevotella* e *Actinomyces* sono presenti in abbondanze maggiori rispetto ai non fumatori sani, confermando quanto evidenziato nel caso dei generi (tabella 13).

Nella tabella 20 si mostra come si differenzia il microbiota dei soggetti in base allo stato di salute degli stessi; è evidente che i fumatori malati presentano quantità superiori di *Prevotella* e *Actinomyces* rispetto ai soggetti sani (fumatori e non).7

8.2 Risultati delle analisi sui dati relativi al tumore colon-rettale

8.2.1 Misura della biodiversità intra- e inter-soggetto

Innanzitutto si è calcolata la diversità alfa (calcolata a partire dai valori di abbondanza relativa riferiti a generi e OTU) all'interno di ciascun soggetto, prendendo in considerazione gli indici di Shannon, Evenness e di Simpson, inteso nella forma 1-D, e nelle tabelle 21 e 22 si riportano i risultati ottenuti.

Si ricorda di seguito il significato del valore minimo e massimo degli indici.
Nel caso di Shannon e Simpson il valore minimo (zero) indica che si registra una bassa biodiversità nel soggetto in questione e che dunque si hanno poche specie e non distribuite in modo omogeneo. Il valore massimo invece suggerisce un'elevata abbondanza di specie (in media) all'interno di un dato soggetto (alta biodiversità) e le specie presentano uguale abbondanza nei soggetti. Per quanto riguarda l'indice di Evenness invece, lo zero indica che in quel soggetto si è in presenza di una sola specie (e quindi di massima omogeneità nella distribuzione delle specie), mentre quando l'indice assume il valore massimo (uno), si registra massima eterogeneità nel soggetto e dunque tutte le specie sono presenti in egual quantità.

	Generi		
	Shannon	**Evenness**	**Simpson**
HC1	1.686	0.374	0.604
HC2	1.719	0.402	0.69
HC3	2.213	0.481	0.732
HC4	2.667	0.552	0.848
HC5	1.609	0.416	0.548
HC6	2.409	0.561	0.795
HC7	2.161	0.48	0.751
HC8	1.168	0.268	0.37
TB1_C	2.287	0.478	0.721
TB2_C	2.432	0.502	0.719
TB3_C	2.724	0.544	0.82
TB4_C	1.984	0.47	0.689
TB5_C	1.692	0.412	0.622

TB6_C	1.629	0.403	0.56
TB7_C	2.674	0.585	0.866
TB8_C	2.707	0.587	0.828
TB1	2.334	0.492	0.742
TB2	2.538	0.524	0.751
TB3	2.825	0.562	0.84
TB4	2.022	0.474	0.688
TB5	1.998	0.469	0.739
TB6	1.701	0.412	0.579
TB7	2.465	0.541	0.789
TB8	2.768	0.6	0.838
TD1_C	2.661	0.558	0.861
TD2_C	2.456	0.496	0.763
TD3_C	1.935	0.447	0.712
TD4_C	2.322	0.548	0.778
TD5_C	2.257	0.512	0.762
TD6_C	1.591	0.372	0.527
TD7_C	2.11	0.453	0.762
TD8_C	2.795	0.569	0.873
TD1	2.625	0.531	0.778
TD2	1.9	0.411	0.609
TD3	1.899	0.415	0.683
TD4	2.07	0.481	0.746
TD5	2.308	0.551	0.78
TD6	2.499	0.537	0.846
TD7	2.294	0.492	0.803
TD8	2.152	0.454	0.783
TM1_C	2.212	0.545	0.784
TM2_C	1.989	0.441	0.609
TM3_C	2.979	0.641	0.889
TM4_C	2.645	0.627	0.825
TM5_C	2.763	0.69	0.854
TM6_C	2.586	0.603	0.852
TM7_C	2.237	0.532	0.716
TM8_C	2.317	0.53	0.853
TM1	2.201	0.529	0.783
TM2	2.513	0.528	0.796
TM3	2.502	0.501	0.782
TM4	2.621	0.598	0.823
TM5	2.466	0.549	0.785
TM6	2.653	0.604	0.857
TM7	1.702	0.399	0.56
TM8	2.563	0.531	0.776

***Tabella 21**: Indici di diversità alfa (Shannon, Evenness e Simpson, sulle colonne) calcolati per ciascun soggetto, dai valori di abbondanza relativa dei diversi generi. I soggetti, sulle righe, sono divisi secondo le classi di appartenenza: soggetti sani (HC), tessuto sano di soggetti con polipi non displastici (TB_C), tessuto malato di soggetti con polipi non displastici (TB), tessuto sano di soggetti con polipi displastici (TD_C), tessuto malato di soggetti con polipi displastici (TD), tessuto sano di soggetti con resez. cancro (TM_C) , tessuto malato di soggetti con resez. cancro (TM)*

	OTU		
	Shannon	**Evenness**	**Simpson**
HC1	3.972	0.65	0.929
HC2	3.878	0.681	0.951
HC3	4.431	0.722	0.969
HC4	4.616	0.725	0.971
HC5	3.51	0.639	0.891
HC6	3.859	0.71	0.928
HC7	4.507	0.734	0.967
HC8	3.418	0.574	0.863
TB1_C	4.762	0.764	0.973
TB2_C	4.648	0.762	0.972
TB3_C	5.391	0.828	0.99
TB4_C	4.824	0.812	0.981
TB5_C	4.017	0.715	0.951
TB6_C	4.859	0.839	0.984
TB7_C	5.171	0.861	0.989
TB8_C	5.251	0.852	0.989
TB1	4.843	0.789	0.98
TB2	4.821	0.79	0.981
TB3	5.42	0.837	0.99
TB4	4.882	0.832	0.981
TB5	4.222	0.761	0.963
TB6	4.921	0.847	0.984
TB7	5.408	0.887	0.992
TB8	5.222	0.861	0.988
TD1_C	5.344	0.837	0.991
TD2_C	5.428	0.84	0.991
TD3_C	4.248	0.72	0.957
TD4_C	4.804	0.833	0.984
TD5_C	4.836	0.776	0.977
TD6_C	4.696	0.77	0.976
TD7_C	4.781	0.78	0.984
TD8_C	5.326	0.816	0.987

TD1	5.075	0.798	0.979
TD2	5.221	0.826	0.989
TD3	4.325	0.71	0.954
TD4	4.845	0.82	0.984
TD5	4.943	0.824	0.981
TD6	5.125	0.828	0.989
TD7	4.851	0.797	0.985
TD8	4.976	0.765	0.974
TM1_C	5.015	0.855	0.989
TM2_C	4.135	0.795	0.968
TM3_C	4.594	0.855	0.983
TM4_C	4.085	0.795	0.963
TM5_C	3.774	0.874	0.966
TM6_C	4.361	0.85	0.98
TM7_C	4.445	0.806	0.974
TM8_C	3.98	0.76	0.964
TM1	4.766	0.83	0.984
TM2	4.456	0.765	0.97
TM3	5.435	0.83	0.99
TM4	4.329	0.747	0.951
TM5	4.679	0.826	0.982
TM6	5.232	0.854	0.99
TM7	4.658	0.8	0.978
TM8	4.565	0.764	0.974

Tabella 22: *Indici di diversità alfa (Shannon, Evenness e Simpson, sulle colonne) calcolati per ciascun soggetto, dai valori di abbondanza relativa delle diverse OTU. I soggetti, sulle righe, sono divisi secondo le classi di appartenenza: soggetti sani (HC), tessuto sano di soggetti con polipi non displastici (TB_C), tessuto malato di soggetti con polipi non displastici (TB), tessuto sano di soggetti con polipi displastici (TD_C), tessuto malato di soggetti con polipi displastici (TD), tessuto sano di soggetti con resez. cancro (TM_C) , tessuto malato di soggetti con resez. cancro (TM)*

A partire dalle tabelle 21 e 22 si sono inoltre costruiti dei boxplot, per una più immediata visualizzazione dei risultati. Si presentano dapprima i boxplot degli indici calcolati sulle abbondanze relative dei generi e successivamente quelli ottenuti partendo dalle abbondanze delle OTU.

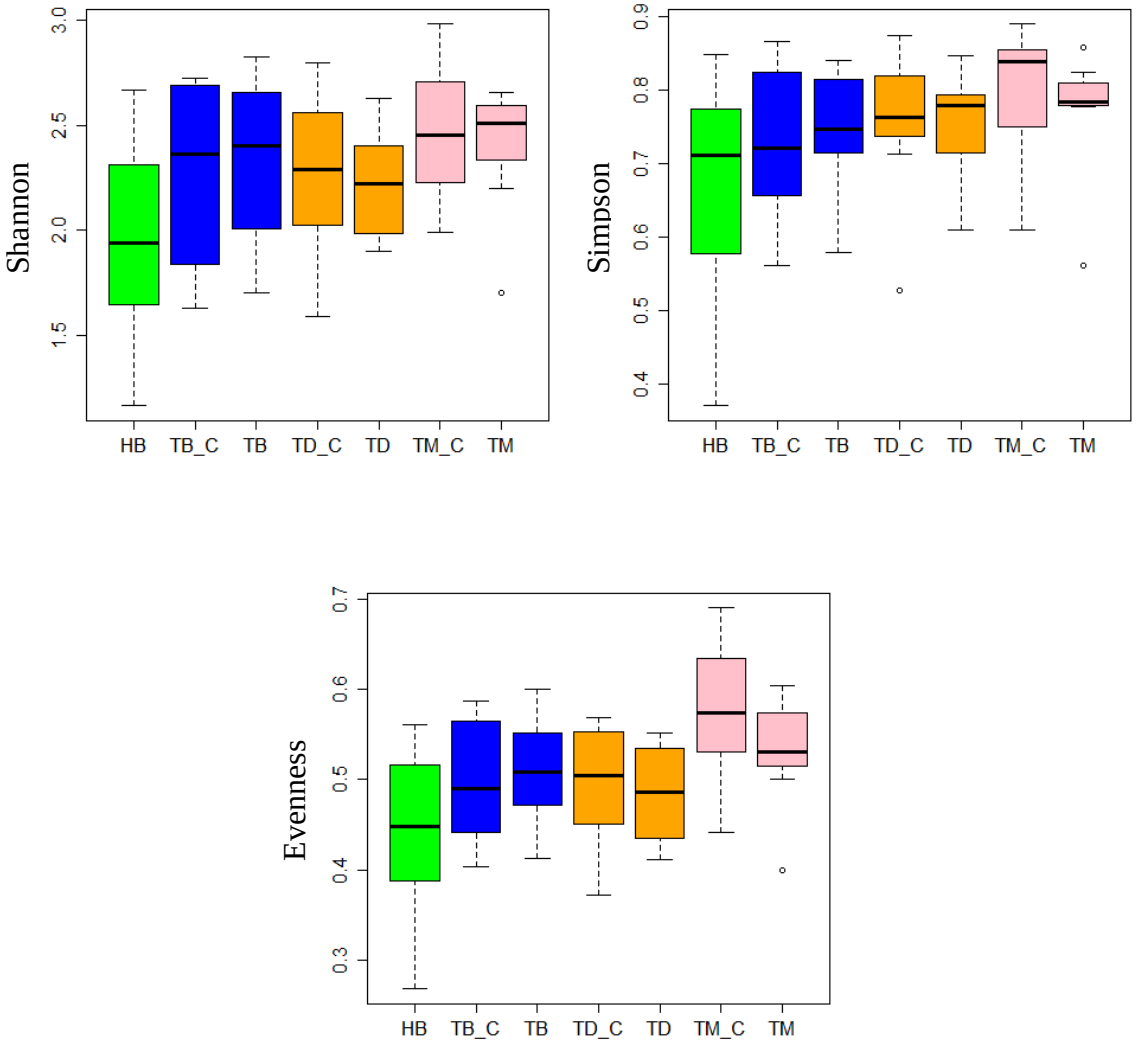


Figura 35. *Boxplot dei valori degli indici di Shannon, Evenness e Simpson dei diversi tipi di tessuto, a partire dalle abbondanze relative dei diversi generi.*

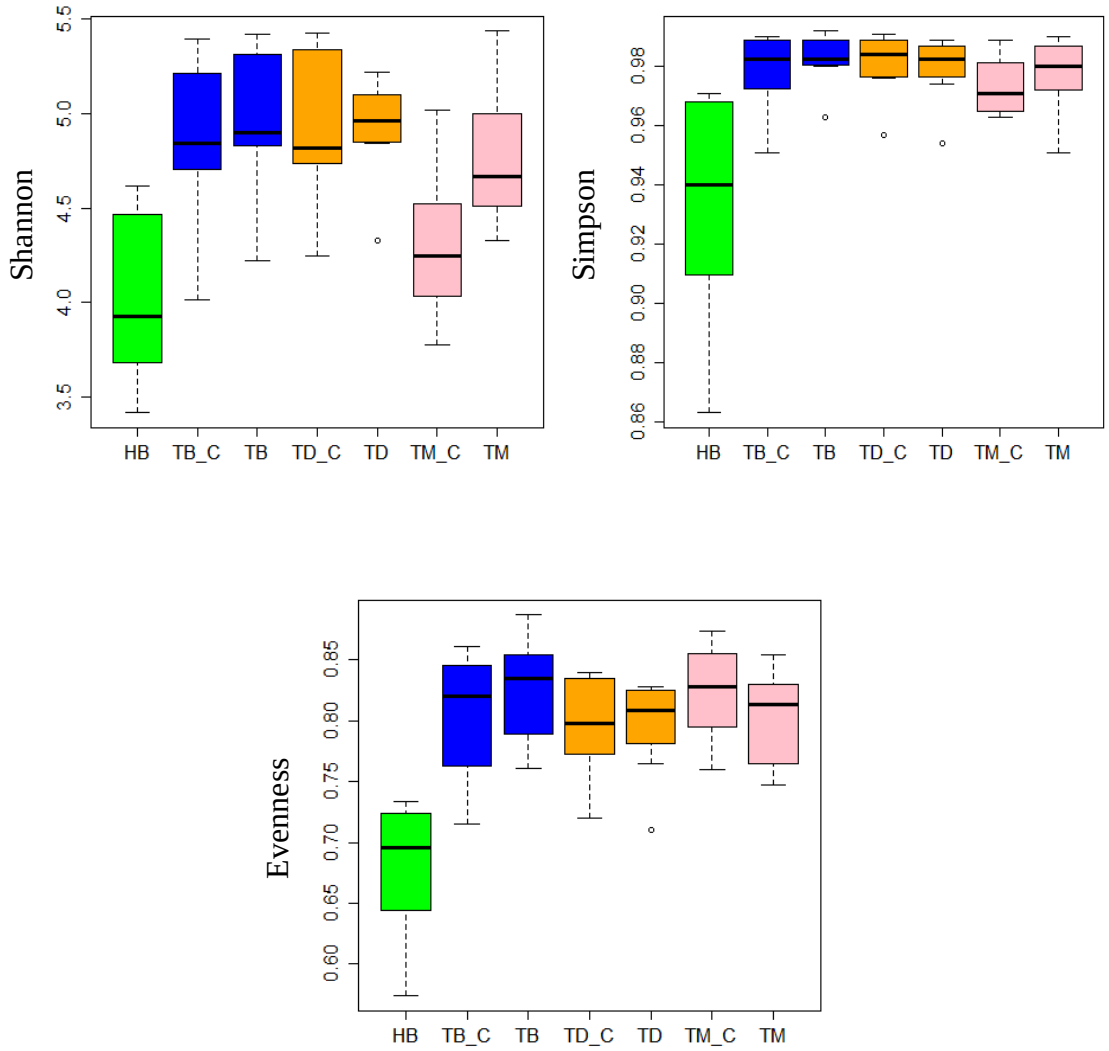


Figura 36. *Boxplot dei valori degli indici di Shannon, Evenness e Simpson dei diversi tipi di tessuto, a partire dalle abbondanze relative delle diverse OTU.*

Si osserva innanzitutto che sia che gli indici siano calcolati a partire dalle abbondanze dei generi che da quelle delle OTU, i valori calcolati risultano essere tutti elevati. Si può quindi concludere che i campioni di tessuto presentano una certa eterogeneità a livello di composizione del microbioma, indipendentemente dallo stato di salute del soggetto dal quale sono prelevati. Solo nel caso di tessuto proveniente da individui sani si riscontra un'eterogeneità minore rispetto agli altri campioni. Al contrario, nel tessuto malato prelevato da soggetti affetti da cancro al colon si registra il valore massimo dell'indice di Evenness e dunque la massima eterogeneità nella composizione del microbioma. Si osserva inoltre che, considerando tessuto sano e tessuto

malato adiacente nel caso di soggetti con polipi (displastici e non), si hanno valori di diversità alfa maggiori nel tessuto malato rispetto a quello sano. Nel caso invece dei tessuti prelevati da soggetti affetti da cancro, si osserva la situazione contraria.

In seguito si è calcolata la diversità beta tra i campioni appartenenti alla stessa categoria, utilizzando tre diversi indici in modo da ottenere una stima robusta della diversità. I risultati sono riportati nella figura 1, in cui l'altezza delle barre rappresenta il valore degli indici, mentre sull'asse delle ascisse si mostrano i campioni considerati. Con il calcolo di tali indici si quantificano le differenze tra le abbondanze relative di generi ed OTU di diversi campioni, appartenenti alla stessa tipologia di tessuto. Si può evidenziare quindi se all'interno del medesimo gruppo, i campioni presentano o meno affinità a livello di microbioma.

Si ricorda che più il valore degli indici è basso, più i campioni confrontati sono simili a livello della composizione del microbioma; in altre parole molte sono le specie comuni riscontrate nei due campioni. Se invece gli indici assumono valore massimo (pari ad uno), allora i campioni non presentano alcuna specie comune e dunque la diversità è massima.

Si precisa inoltre che i valori riportati per ciascun tipo di tessuto, sono le medie dei valori degli indici calcolati confrontando ogni possibile coppia di campioni all'interno del medesimo gruppo.

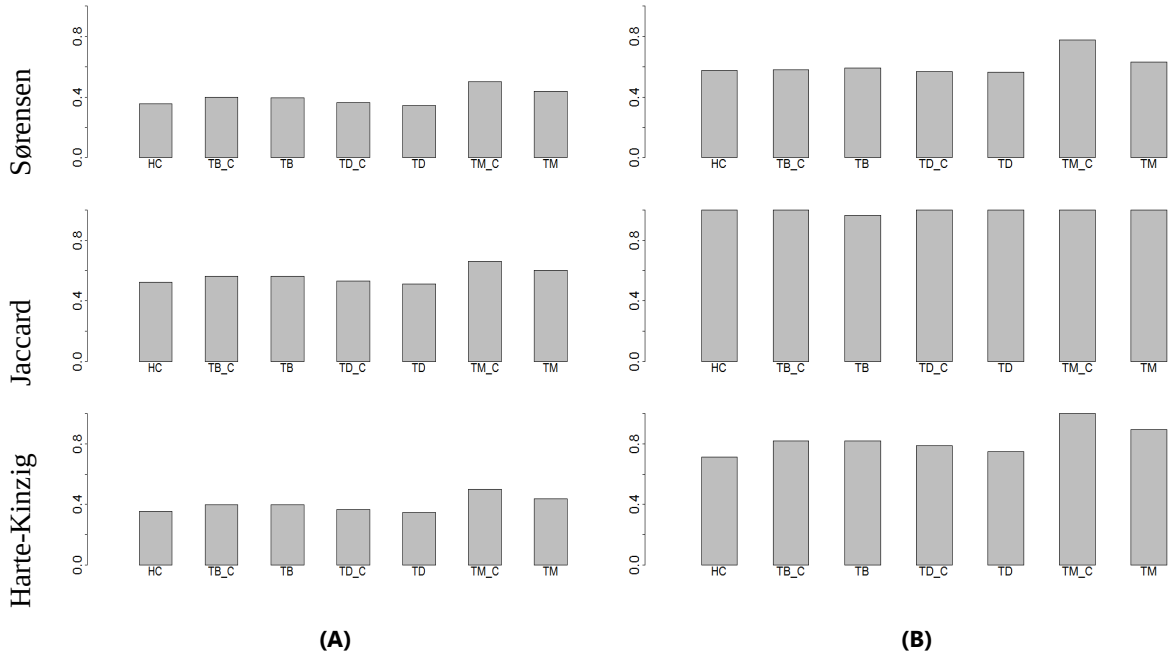


Figura37: *rappresentazione della media dei valori degli indici di diversità beta calcolati tra soggetti appartenenti ad uno stesso gruppo (soggetti sani (HC), tessuto sano di soggetti con polipi non displastici (TB_C), tessuto malato di soggetti con polipi non displastici (TB), tessuto sano di soggetti con polipi displastici (TD_C), tessuto malato di soggetti con polipi displastici (TD), tessuto sano di soggetti con resez. cancro (TM_C) , tessuto malato di soggetti con resez. cancro (TM)). I dati di partenza sono le abbondanze relative dei diversi generi (**A**) e delle diverse OTU (**B**).*

Dai risultati si evince innanzitutto che gli indici assumono valori maggiori quando calcolati a partire dalle abbondanze relative delle OTU; questo è però comprensibile in quanto i dati sui quali misurare la diversità sono più numerosi rispetto a quelli riferiti ai generi.
Inoltre si evidenzia che in media la diversità tra campioni dello stesso gruppo è abbastanza elevata e raggiunge i valori massimi nei campioni di tessuto sano prelevato da soggetti affetti da cancro. Negli stessi soggetti, anche il tessuto malato adiacente a quello sano appena citato porta ad indici con valori elevati, secondi solamente a quelli ricavati dal tessuto malato.
I valori più bassi di diversità beta si ottengono invece confrontando tra loro i soggetti sani.

Una volta confrontati i campioni appartenenti ad una stessa tipologia di tessuto, è di nostro interesse ricercare eventuali differenze nel microbioma di campioni provenienti da tessuti diversi. Ci si pone dunque l'obiettivo di indagare possibili relazioni tra lo stato di salute del tessuto e la composizione del microbioma dello stesso.
Operativamente, si sono considerate tutte le possibili coppie di campioni appartenenti a tessuti diversi e su ciascuna di queste è stata applicata NPMANOVA; si è così valutata la significatività della differenze tra abbondanze relative di generi e OTU dei campioni confrontati. L'ipotesi nulla di NPMANOVA è che le medie dei dati appartenenti ai due gruppi posti a confronto non sono significativamente diverse; osservando quindi i valori dei p-value calcolati da NPMANOVA, si è stabilito quali tessuti presentano abbondanze di generi e OTU significativamente diverse. Si ricordi che un p-value inferiore alla soglia di significatività prescelta (nel nostro caso pari a 0.05), indica che si può accettare con un possibilità d'errore inferiore al 5% l'ipotesi alternativa del test, e cioè che le abbondanze misurate nei due tessuti sono significativamente diverse.
Nella tabella 23 sono riportati tutti i possibili confronti tra tessuti effettuati utilizzando NPMANOVA e, per ciascuno di essi, i relativi valori del p-value e della statistica F (troncati alla terza cifra decimale). Nella parte sopra alla diagonale si presentano i risultati dei confronti tra le abbondanze dei generi, mentre nella parte sotto alla diagonale si osservano i p-value dei confronti tra abbondanze delle OTU.
Sono inoltre evidenziati quei confronti che producono un p-value inferiore allo 0.05, in modo da porre in risalto i tessuti significativamente diversi a livello di microbioma.

		p-value (statistica F) da GENERI						
	Tipi di tessuto	**HC**	**TB**	**TB_C**	**TD**	**TD_C**	**TM**	**TM_C**
p-value (statistica F) da OTU	**HC**		0.223 (1.441)	0.285 (1.192)	0.324 (1.107)	0.275 (1.189)	0.073* (2.342)	**0.009*** (2.876)**
	TB	**<0.001*** (3.25)**		0.991 (0.027)	0.388 (0.875)	0.186 (1.583)	0.248 (1.287)	0.107 (1.187)
	TB_C	**<0.001*** (2.955)**	0.971 (0.31)		0.542 (0.665)	0.238 (1.365)	0.301 (1.132)	0.138 (1.653)
	TD	**<0.001*** (2.764)**	0.384 (1.045)	0.95 (0.804)		0.992 (0.245)	0.235 (1.243)	0.117 (1.669)
	TD_C	**<0.001*** (2.822)**	0.241 (1.186)	0.34 (1.081)	0.993 (0.213)		0.105 (1.935)	**0.02** (2.36)**
	TM	**<0.001*** (3.453)**	0.095* (1.513)	0.147 (1.345)	0.205 (1.212)	0.125 (1.379)		0.859 (0.43)
	TM_C	**<0.001*** (3.896)**	**0.001*** (2.186)**	**0.001*** (2.66)**	**<0.001*** (2.547)**	**<0.001*** (2.719)**	**0.008*** (2.181)**	

***Tabella 23**: p-value e valore della statistica F (tra parentesi), risultati dall'applicazione di NPMANOVA alle abbondanze relative dei diversi generi e OTU, misurate in soggetti appartenenti a gruppi diversi: soggetti sani (HC), tessuto sano di soggetti con polipi non displastici (TB_C), tessuto malato di soggetti con polipi non displastici (TB), tessuto sano di soggetti con polipi displastici (TD_C), tessuto malato di soggetti con polipi displastici (TD), tessuto sano di soggetti con resez. cancro (TM_C) , tessuto malato di soggetti con resez. cancro (TM).*

*(Livelli di significatività del p-value: 0.01 '***' 0.05 '**' 0.1 ' *')*

E' evidente che per alcuni confronti i p-value calcolati da NPMANOVA considerando le abbondanze delle OTU sono più bassi rispetto a quelli ottenuti comparando le abbondanze dei generi; si ricordi che i dati relativi alle OTU presentano una numerosità maggiore di quella dei dati riferiti ai generi e dunque i risultati hanno un più robusto significato statistico.

A proposito di tali discordanze nei valori dei p-value, basti osservare che confrontando le abbondanze dei generi si ottengono p-value inferiori alla soglia dello 0.05 in soli due casi: tessuto sano di soggetti sani a confronto con tessuto sano di soggetti malati di cancro, e tessuto sano di soggetti con polipi displastici a confronto con tessuto sano di soggetti malati di cancro. Applicando invece NPMANOVA alle abbondanze relative delle OTU si ottengono p-value inferiori alla soglia di significatività in undici confronti, tra i quali i due esposti in precedenza.

In tal caso si conclude che le abbondanze relative delle OTU presenti nel tessuto sano di soggetti sani (HC) differiscono in modo significativo da quelle dei tessuti (malati o sani che siano) dei soggetti con polipi displastici (TD e TD_C), di quelli con polipi non displastici (TB e TB_C) e dei soggetti affetti da cancro (TM e TM_C). Inoltre anche le abbondanze delle OTU nel tessuto sano dei pazienti malati di cancro (TM_C) differiscono significativamente da quelle di tutti gli altri tessuti (TB,TB_C,TD,TD_C e HC).

Si noti che le differenze più significative (e quindi i p-value più bassi) si riscontrano nei confronti tra:

- tessuto sano di soggetti sani e tutti gli altri tessuti (p-value inferiori a 0.001);
- tessuto sano di soggetti affetti da cancro e tessuto sano di soggetti con polipi displastici (p-value inferiore a 0.001)
- tessuto sano di soggetti affetti da cancro e tessuto malato di soggetti con polipi displastici (p-value inferiore a 0.001)

Si pone inoltre in evidenza la grande differenza tra p-value calcolato nel confronto tra tessuto sano e tessuto malato di soggetti affetti da cancro, a seconda che si confrontino abbondanze dei generi o delle OTU. Nel primo caso infatti si ottiene un p-value prossimo a 1 e quindi si può stabilire che le abbondanze dei generi sono pressoché uguali nei due tessuti, mentre si afferma il contrario se si considerano le abbondanze delle OTU in quanto il p-value è pari a 0.008 (e quindi inferiore alla soglia dello 0.05).
Un'ultima osservazione riguarda due confronti, quello tra tessuto sano e tessuti malato di soggetti con polipi, sia displastici che non. In tal caso infatti, indipendentemente che si confrontino abbondanze di generi o OTU, si notano p-value con valori molto prossimi a uno, suggerendo una significativa similarità tra i tessuti.
Per rendere più robuste le conclusioni avanzate osservando i risultati di NPMANOVA, si sono ripetuti i confronti utilizzando ANOSIM.

8.2.2 Analisi delle similarità dei campioni con ANOSIM

I risultati dell'analisi delle similarità effettuata con il metodo ANOSIM sono presentati nella tabella 24; nella parte di tabella sopra la diagonale ci si riferisce all'analisi applicata ai dati di abbondanza relativa dei diversi generi, mentre nella parte inferiore i dati di partenza sono le abbondanze relative delle diverse OTU.
Così come visto nel paragrafo precedente, nelle tabelle si riportano i gruppi confrontati ed i valori di p-value e statistica R ottenute da ANOSIM. Ancora una volta tali valori sono stati troncati alla terza cifra decimale e si evidenziano i confronti che producono un p-value inferiore alla soglia di significatività, in questo caso pari a 0.05.

		p-value (statistica R) da GENERI						
	Tipi di tessuto	**HC**	**TB**	**TB_C**	**TD**	**TD_C**	**TM**	**TM_C**
p-value (statistica R) da OTU	**HC**		0.386 (0)	0.46 (-0.022)	0.402 (0.001)	0.362 (0.014)	**0.034** (0.172)**	**0.006** (0.242)**
	TB	**<0.001*** (0.565)**		0.992 (-0.124)	0.689 (-0.056)	0.493 (-0.011)	0.2 (0.051)	0.102 (0.11)
	TB_C	**<0.001*** (0.524)**	0.913 (-0.1)		0.795 (-0.069)	0.495 (-0.004)	0.233 (0.036)	0.154 (0.081)
	TD	**0.001*** (0.516)**	0.358 (0.017)	0.595 (-0.035)		0.98 (-0.113)	0.271 (0.028)	0.09* (0.095)
	TD_C	**<0.001*** (0.511)**	0.29 (0.043)	0.369 (0.017)	0.964 (-0.12)		0.051* (0.134)	**0.01** (0.216)**
	TM	**<0.001*** (0.605)**	0.082* (0.135)	0.081* (0.14)	0.143 (0.085)	0.092* (0.109)		0.815 (-0.066)
	TM_C	**<0.001*** (0.729)**	**<0.001*** (0.479)**	**<0.001*** (0.497)**	**0.001*** (0.518)**	**<0.001*** (0.53)**	**<0.009*** (0.332)**	

***Tabella 24**: p-value e valore della statistica R (tra parentesi), risultati dall'applicazione di ANOSIM alle abbondanze relative dei diversi generi e OTU, misurate in soggetti appartenenti a gruppi diversi: soggetti sani (HC), tessuto sano di soggetti con polipi non displastici (TB_C), tessuto malato di soggetti con polipi non displastici (TB), tessuto sano di soggetti con polipi displastici (TD_C), tessuto malato di soggetti con polipi displastici (TD), tessuto sano di soggetti con resez. cancro (TM_C) , tessuto malato di soggetti con resez. cancro (TM).*

*(Livelli di significatività del p-value: 0.01 '***' 0.05 '**' 0.1 ' *')*

Così come evidenziato nei risultati di NPMANOVA, sono presenti discordanze tra i valori dei p-value ottenuti confrontando le abbondanze relative dei generi e quelli derivanti dai confronti tra le abbondanze delle OTU. Nel primo caso infatti si ottengono p-value inferiori alla soglia di significatività in occasione di tre confronti, mentre nel secondo si evidenziano undici p-value di valore inferiore a 0.05.

Per la precisione, i confronti che evidenziano significative differenze tra abbondanze di generi nei tessuti considerati sono: tessuto sano di soggetti sani e tessuto sano di malati di cancro, tessuto sano di soggetti sani e tessuto malato di soggetti affetti da cancro, tessuto sano di soggetti con polipi displastici e tessuto sano di soggetti malati di cancro.

Per quanto riguarda invece le abbondanze relative delle OTU si notano differenze significative tra tessuto sano di soggetti sani (HC) e tessuti (malati o sani che siano) dei soggetti con polipi displastici (TD e TD_C), di quelli con polipi non displastici (TB e TB_C) e dei soggetti affetti da cancro (TM e TM_C). Inoltre anche le abbondanze delle OTU nel tessuto sano dei pazienti malati di cancro (TM_C) differiscono significativamente da quelle di tutti gli altri tessuti (TB,TB_C,TD,TD_C e HC).

Si osserva inoltre la grande differenza tra p-value calcolato nel confronto tra tessuto sano e tessuto malato di soggetti affetti da cancro, a seconda che si confrontino abbondanze dei generi

o delle OTU. Nel primo caso infatti il p-value prossimo a 1 suggerisce che le abbondanze dei generi sono pressoché uguali nei due tessuti, mentre si afferma il contrario considerando le abbondanze delle OTU, in quanto il p-value è pari a 0.011 (e quindi inferiore alla soglia dello 0.05).
Infine si pone in evidenza che il tessuto sano e quello malato prelevato dai soggetti con polipi displastici possono considerarsi significativamente simili a livello di abbondanze di generi e OTU, e lo stesso si può concludere per il tessuto sano e quello malato dei soggetti con polipi non displastici.
Si nota dunque che i risultati ottenuti mediante NPMANOVA concordano con quelli ricavati con ANOSIM.

8.2.3 Differenze nella composizione microbiotica tra tessuti

Nelle tabelle seguenti si riportano i risultati del test di Wilcoxon applicato alle abbondanze relative dei diversi generi e delle diverse OTU per il confronto tra tutte le possibile coppie di gruppi.

Precisamente, nella prima colonna sono riportati i generi o le OTU le cui abbondanze relative sono significativamente diverse tra i due gruppi, avendo imposto un livello di significatività pari a 0.05; si è poi ordinata la lista finale secondo valori di p-value crescenti (dopo correzione FDR), i quali sono inseriti nella seconda colonna. Si mostrano inoltre nella terza e quarta colonna le medie delle abbondanze relative di ciascun genere o OTU all'interno dei gruppi confrontati. Si evidenzia infine il gruppo nel quale un certo genere o OTU presenta maggiore abbondanza media.

Generi	pvalue	Pvalue corretti	HC	TB
Verrucomicrobium	0.0025	0.3021	0	0.001
Phyllobacterium	0.0037	0.3021	0	0.0021
Proteiniphilum	0.0046	0.3021	0	0.0003
ClostridiumXlVa	0.0104	0.3557	0.0199	0.0041
Crabtreella	0.0128	0.3557	0	0.0003
Varibaculum	0.0128	0.3557	0	0.0004

Flavobacterium	0.0235	0.3557	0.0005	0.0057
Shinella	0.0281	0.3557	0.002	0.013
Pseudobutyrivibrio	0.0287	0.3557	0.0001	0.0018
Serpens	0.0298	0.3557	0	0.0005
Thermofilum	0.0298	0.3557	0	0.0003
Catonella	0.0325	0.3557	0.0001	0.0005
Cerasicoccus	0.0325	0.3557	0	0.0004
Desulfosoma	0.0325	0.3557	0	0.0005
Desulfurispora	0.0325	0.3557	0	0.0001
Halarsenatibacter	0.0325	0.3557	0	0.0002
Halosimplex	0.0325	0.3557	0	0.0003
Howardella	0.0325	0.3557	0	0.0002
Elusimicrobium	0.04	0.394	0.0005	0.0052
Phascolarctobacterium	0.04	0.394	0.0077	0.0016
ClostridiumXVIII	0.0486	0.4559	0.0004	0.0014
...	...	...	...	...

Tabella 25. *Generi le cui abbondanze relative sono diverse (Test di Wilcoxon, livello di significatività 5%)*

Generi	**pvalue**	**Pvalue corretti**	**HC**	**TD**
Phascolarctobacterium	0.0035	0.3324	0.0077	0.0003
Verrucomicrobium	0.005	0.3324	0	0.0005
Anaerophaga	0.0128	0.3324	0.0002	0
Eggerthella	0.0128	0.3324	0	0.0004
Oleispira	0.0128	0.3324	0	0.0001

Selenihalanaerobacter	0.0128	0.3324	0	0.0001
ClostridiumXVIII	0.0129	0.3324	0.0004	0.0078
Odoribacter	0.0135	0.3324	0.0089	0.0015
Veillonella	0.0177	0.3874	0.0001	0.0026
Fusobacterium	0.024	0.4573	0.0002	0.002
Caldivirga	0.0298	0.4573	0.0008	0
Campylobacter	0.0325	0.4573	0	0.0001
Citrobacter	0.0325	0.4573	0	0.0002
Phocaeicola	0.0325	0.4573	0	0.0003
Barnesiella	0.0405	0.5161	0.0032	0.001
Phyllobacterium	0.0496	0.5161	0	0.0005
Roseburia	0.0499	0.5161	0.051	0.0165
...	...	...	...	...

Tabella 25. *Generi le cui abbondanze relative sono diverse (Test di Wilcoxon, livello di significatività 5%)*

Generi	**pvalue**	**pvalue corretti**	**HC**	**TM**
Citrobacter	0.0015	0.2362	0	0.0013
Salmonella	0.0037	0.2362	0.0001	0.0085
Lactobacillus	0.0046	0.2362	0	0.0057
ClostridiumXlVa	0.0047	0.2362	0.0199	0.0034
Bilophila	0.006	0.2412	0.0012	0.0001
Serpens	0.0092	0.2573	0	0.0004
Desulfosoma	0.0128	0.2573	0	0.0017

Enterococcus	0.0128	0.2573	0	0.0497
Staphylococcus	0.0128	0.2573	0	0.001
Verrucomicrobium	0.0128	0.2573	0	0.0021
Phascolarctobacterium	0.0165	0.2897	0.0077	0.0004
Roseburia	0.0207	0.2897	0.051	0.0176
Enterobacter	0.0219	0.2897	0.0005	0.006
Propionibacterium	0.0246	0.2897	0.0001	0.0065
Escherichia/Shigella	0.0281	0.2897	0.0113	0.0845
Flavonifractor	0.0281	0.2897	0.0021	0.0009
Parabacteroides	0.0499	0.2897	0.0534	0.01
...	...	...	...	...

Tabella 26. *Generi le cui abbondanze relative sono diverse (Test di Wilcoxon, livello di significatività 5%)*

Generi	**pvalue**	**pvalue corretti**	**TB**	**TB_C**
Proteiniphilum	0.0323	1	0.0003	0.0001
Pseudobutyrivibrio	0.1119	1	0.0018	0.0003
Butyrivibrio	0.1501	1	0.0004	0.0009
Wautersia	0.1709	1	0	0.0002
Azomonas	0.2074	1	0.0006	0.0002
Dysgonomonas	0.2143	1	0.0001	0
Anaerostipes	0.2149	1	0.0009	0.0008
Nubsella	0.2697	1	0	0.0001

Methyloversatilis	0.2786	1	0.0193	0.0088
...	...	...	...	...

Tabella 27. *Generi le cui abbondanze relative sono diverse (Test di Wilcoxon, livello di significatività 5%)*

Generi	pvalue	pvalue corretti	TB	TD
Proteiniphilum	0.0128	0.7169	0.0003	0
Raoultella	0.0128	0.7169	0	0.0005
Enterobacter	0.0263	0.7169	0.0002	0.0023
Haemophilus	0.0289	0.7169	0.0016	0.0027
Azomonas	0.0298	0.7169	0.0006	0
Zavarzinella	0.0298	0.7169	0.0005	0
Desulfonispora	0.0325	0.7169	0	0.0001
Desulfurispora	0.0325	0.7169	0.0001	0
Elusimicrobium	0.0356	0.7169	0.0052	0.0005
Thermofilum	0.0402	0.7169	0.0003	0
Holdemania	0.0425	0.7169	0.0001	0.0007
Phascolarctobacterium	0.0428	0.7169	0.0016	0.0003
Terrimonas	0.0486	0.7298	0.0006	0.0001
...	...	...	...	...

Tabella 28. *Generi le cui abbondanze relative sono diverse (Test di Wilcoxon, livello di significatività 5%)*

Generi	pvalue	pvalue corretti	TB	TM
Caulobacter	0.0037	0.3708	0.0017	0
Methyloversatilis	0.0047	0.3708	0.0193	0.0009
Phyllobacterium	0.007	0.3708	0.0021	0.0002
Citrobacter	0.0072	0.3708	0	0.0013
Lactobacillus	0.0159	0.4871	0.0001	0.0057
Variovorax	0.0159	0.4871	0.0012	0.0001
Enterobacter	0.0199	0.4871	0.0002	0.006
Hydrotalea	0.0246	0.4871	0.0023	0.0001
Clostridiumsensustricto	0.0298	0.4871	0	0.0004
Gemella	0.0325	0.4871	0	0.0004
Kluyvera	0.0325	0.4871	0	0.0008
Leptotrichia	0.0325	0.4871	0	0.0025
Escherichia/Shigella	0.0379	0.4871	0.0125	0.0845
Prevotella	0.038	0.4871	0.0107	0.0682
Proteiniphilum	0.038	0.4871	0.0003	0.0001
Shinella	0.0391	0.4871	0.013	0.0043
Staphylococcus	0.0402	0.4871	0	0.001
Parabacteroides	0.0499	0.5427	0.0422	0.01
...	...	...	...	...

Tabella 29. *Generi le cui abbondanze relative sono diverse (Test di Wilcoxon, livello di significatività 5%)*

Generi	pvalue	pvalue corretti	TD	TD_C
Campylobacter	0.0325	1	0.0001	0
Pelospora	0.0764	1	0	0
Delftia	0.1299	1	0.0003	0.0001
Pasteuria	0.142	1	0	0.0001
...	...	...	...	...

Tabella 30. *Generi le cui abbondanze relative sono diverse (Test di Wilcoxon, livello di significatività 5%)*

Generi	pvalue	pvalue corretti	TD	TM
Parabacteroides	0.0003	0.0612	0.0931	0.01
Barnesiella	0.009	0.612	0.001	0.0058
ClostridiumXVIII	0.009	0.612	0.0078	0.0005
Staphylococcus	0.0128	0.6308	0	0.001
Citrobacter	0.0219	0.6308	0.0002	0.0013
Lactobacillus	0.0246	0.6308	0.0001	0.0057
Leptotrichia	0.0325	0.6308	0	0.0025
Pelospora	0.0325	0.6308	0	0.0001
Methyloversatilis	0.0376	0.6308	0.0056	0.0009
Propionibacterium	0.0376	0.6308	0.0001	0.0065
Variovorax	0.038	0.6308	0.0012	0.0001
Desulfovibrio	0.0402	0.6308	0.0013	0.0001
...	...	...	...	...

Tabella 31. *Generi le cui abbondanze relative sono diverse (Test di Wilcoxon, livello di significatività 5%)*

Generi	pvalue	pvalue corretti	TM	TM_C
ClostridiumXlVa	0.0024	0.4824	0.0034	0.0004
Erysipelotrichaceae_incertae_sedis	0.0128	0.7807	0.0021	0.0001
Murdochiella	0.0128	0.7807	0.0014	0
Akkermansia	0.0249	0.7807	0.0016	0.0006
Sulfurisphaera	0.0323	0.7807	0.0008	0.0001
Isobaculum	0.0325	0.7807	0.0001	0
Lachnospiracea_incertae_sedis	0.0379	0.7807	0.0345	0.0116
Leclercia	0.0425	0.7807	0.0001	0.0156
…	…	…	…	…

Tabella 32. *Generi le cui abbondanze relative sono diverse (Test di Wilcoxon, livello di significatività 5%)*

A conferma di quanto già evidenziato nei paragrafi 8.2.1 e 8.2.2, si osserva che il confronto tra tessuto sano prelevato da soggetto sano e tessuto malato di tumore è l'unico a presentare maggiori differenze tra le abbondanze dei generi che compongono il loro microbiota. In particolare, il tessuto malato presenta generi del tutto assenti nel tessuto sano, quali *Citrobacter*, *Lactobacillus*, *Desulfosoma*, *Enterococcus*, *Serpens*, *Staphylococcus*, *Verrucomicrobium* e *Aeromonas.* Nel tessuto sano abbondano invece rispetto al malato i generi *Roseburia* e *Bilophila*.

Si evidenzia inoltre che, confrontando i tessuti appartenenti a soggetti aventi la stessa diagnosi (tabelle 27,30,32), non si riscontrano differenze tra le abbondanze dei generi batterici del loro microbiota (i p-value del test di Wilcoxon, dopo correzione, presentano valori pari o prossimi ad uno); si confermano dunque i risultati ottenuti con ANOSIM e NPMANOVA.

OTU

OTU	pvalue	pvalue corretti	HC	TB
Schlesneria_52	0.0004	0.2486	0	0.0115
Alistipes_7153	0.0004	0.2486	0.0016	0

Bacteroides_93	0.0006	0.2486	0.0016	0.0131
Pseudomonas_160	0.001	0.2486	0.0001	0.0022
Schlesneria_689	0.001	0.2486	0.0075	0.0001
Pseudomonas_246	0.0011	0.2486	0.0563	0.0018
Pseudomonas_212	0.0037	0.3749	0.0001	0.0228
Bacteroides_4631	0.0037	0.3749	0.0055	0.0001
Pseudomonas_0	0.0038	0.3749	0.0013	0.0271
Pseudomonas_20	0.0038	0.3749	0.0013	0.0134
Achromobacter_1	0.0046	0.3749	0	0.0008
Pseudomonas_422	0.0046	0.3749	0	0.0004
Shinella_510	0.0046	0.3749	0	0.0004
Methyloversatilis_605	0.0046	0.3749	0	0.0009
ClostridiumXlVa_817	0.0046	0.3749	0.0009	0
Pseudomonas_10	0.0047	0.3749	0.0018	0.0168
Bacteroides_122	0.0047	0.3749	0.0017	0.0091
Beijerinckia_62	0.0052	0.3854	0.0001	0.0006
Pseudomonas_233	0.0054	0.3854	0.0137	0.0008
Odoribacter_2285	0.006	0.3947	0.0016	0.0001
Methyloversatilis_48	0.0072	0.3947	0.0004	0.003
Bacteroides_1003	0.0082	0.3947	0.0062	0.0004
Flavobacterium_74	0.009	0.3947	0.0002	0.0024
Flavobacterium_179	0.0092	0.3947	0.0001	0.0015
Methyloversatilis_253	0.0092	0.2486	0	0.0011
Escherichia/Shigella_1353	0.0092	0.2486	0.0034	0

Lachnospiracea_incertae_sedis_2340	0.0092	0.2486	0	0.0013
Lachnospiracea_incertae_sedis_1646	0.01	0.2486	0.0056	0.0002
Bacteroides_418	0.0104	0.2486	0.1287	0.0043
Pseudomonas_163	0.0122	0.2486	0.0002	0.0018
...	...	...	...	...

Tabella 33. *OTU le cui abbondanze relative sono diverse (Test di Wilcoxon, livello di significatività 5%)*

OTU	pvalue	pvalue corretti	HC	TD
Schlesneria_52	0.0004	0.3309	0	0.0042
ClostridiumXVIII_159	0.0006	0.3309	0	0.0012
Schlesneria_689	0.0007	0.3309	0.0075	0.0001
Pseudomonas_246	0.0013	0.3545	0.0563	0.0008
Lachnospiracea_incertae_sedis_631	0.0015	0.3545	0	0.0005
Methyloversatilis_127	0.0018	0.3545	0.0079	0.0001
Pseudomonas_233	0.0019	0.3545	0.0137	0.0011
Lachnospiracea_incertae_sedis_1196	0.0025	0.3545	0	0.0003
Odoribacter_2703	0.0025	0.3545	0.0004	0
Bacteroides_4631	0.0025	0.3545	0.0055	0
Parabacteroides_2409	0.0038	0.4443	0.0004	0.0041
Verrucomicrobium_50	0.0046	0.4443	0	0.0004
Pseudomonas_232	0.0046	0.4443	0	0.0002
Methyloversatilis_605	0.0046	0.4443	0	0.0002
Odoribacter_2285	0.0047	0.4443	0.0016	0.0001
Pseudomonas_0	0.0054	0.4763	0.0013	0.0229

Alistipes_7153	0.006	0.4763	0.0016	0.0002
Stenotrophomonas_463	0.0072	0.4763	0.0033	0.0002
Bacteroides_418	0.0099	0.4763	0.1287	0.0028
Bacteroides_1003	0.0099	0.4763	0.0062	0.0004
Chelativorans_28	0.0128	0.4763	0	0.0003
Flavitalea_138	0.0128	0.4763	0	0.0004
Schlesneria_176	0.0128	0.4763	0	0.0003
...	...	...	...	...

Tabella 34. *OTU le cui abbondanze relative sono diverse (Test di Wilcoxon, livello di significatività 5%)*

OTU	pvalue	pvalue corretti	HC	TM
Schlesneria_52	0.2548	0.2548	0	0.0167
Methyloversatilis_127	0.2548	0.2548	0.0079	0
Schlesneria_176	0.2548	0.2548	0	0.003
Lachnospiracea_incertae_sedis_631	0.2548	0.2548	0	0.0009
Propionibacterium_1227	0.2548	0.2548	0	0.0048
Lachnospiracea_incertae_sedis_1646	0.2548	0.2548	0.0056	0
Bacteroides_1799	0.2548	0.2548	0	0.0071
Alistipes_7153	0.2548	0.2548	0.0016	0.0001
Verrucomicrobium_50	0.3479	0.3479	0	0.0019
ClostridiumXlVa_115	0.3479	0.3479	0.0075	0.0004
Escherichia/Shigella_124	0.3479	0.3479	0.0003	0.0167
Flavitalea_138	0.3479	0.3479	0	0.0004
Bacteroides_226	0.3479	0.3479	0.0008	0
Pseudomonas_233	0.3479	0.3479	0.0137	0.0013

Pseudomonas_246	0.3479	0.3479	0.0563	0.0014
Lachnospiracea_incertae_sedis_310	0.3479	0.3479	0.0008	0
Bacteroides_339	0.3479	0.3479	0.0007	0
Escherichia/Shigella_347	0.3479	0.3479	0.0001	0.0018
Bacteroides_418	0.3479	0.3479	0.1287	0.0036
Stenotrophomonas_463	0.3479	0.3479	0.0033	0.0001
Parabacteroides_508	0.3479	0.3479	0.0048	0.0002
...	...		...	...

Tabella 35. *OTU le cui abbondanze relative sono diverse (Test di Wilcoxon, livello di significatività 5%)*

OTU	p-value	pvalue corretti	TB	TB_C
Escherichia/Shigella_170	0.0128	1	0	0.0003
Butyrivibrio_2513	0.0128	1	0	0.0003
Roseburia_3450	0.0184	1	0.0019	0.0002
Acidovorax_144	0.0298	1	0	0.0005
Methyloversatilis_1202	0.0298	1	0.0035	0
Dechloromonas_419	0.0325	1	0	0.0007
Delftia_4540	0.0325	1	0.0005	0
Pseudomonas_4550	0.0325	1	0.0012	0
Dorea_1868	0.0402	1	0	0.0002
Bacteroides_1769	0.0496	1	0.0005	0.0001
...	...	...	...	...

Tabella 36. *OTU le cui abbondanze relative sono diverse (Test di Wilcoxon, livello di significatività 5%)*

OTU	p-value	pvalue corretti	TB	TD
Escherichia/Shigella_170	0.0015	0.8176	0	0.0027
Parabacteroides_2409	0.0036	0.8176	0.0002	0.0041
ClostridiumXlVa_370	0.0072	0.8176	0.0001	0.0008
Parabacteroides_1079	0.0099	0.8176	0.0004	0.0025
Hydrotalea_72	0.0128	0.8176	0.0006	0
Parabacteroides_1996	0.0128	0.8176	0	0.0014
Raoultella_3549	0.0128	0.8176	0	0.0003
Subdoligranulum_3835	0.0128	0.8176	0	0.0003
Enterobacter_4104	0.0128	0.8176	0	0.0007
Escherichia/Shigella_5816	0.0128	0.8176	0	0.0008
Bacteroides_6942	0.0128	0.8176	0	0.0004
Faecalibacterium_1498	0.0177	0.8176	0.0001	0.0021
Stenotrophomonas_15	0.0238	0.8176	0.0039	0.0016
Elusimicrobium_267	0.0273	0.8176	0.0061	0.0006
Parabacteroides_2391	0.0287	0.8176	0.0001	0.0013
...	...	...	...	...

Tabella 37. *OTU le cui abbondanze relative sono diverse (Test di Wilcoxon, livello di significatività 5%)*

OTU	p-value	pvalue corretti	TB	TM
Escherichia/Shigella_170	0.0046	0.6587	0	0.0021
Lachnospiracea_incertae_sedis_2340	0.0046	0.6587	0.0013	0
Prevotella_4048	0.0046	0.6587	0	0.0055

Enterobacter_4104	0.0046	0.6587	0	0.0007
Methyloversatilis_48	0.0047	0.6587	0.003	0.0003
Methyloversatilis_101	0.0052	0.6587	0.0043	0.0001
Caulobacter_94	0.0092	0.6587	0.0013	0
Methyloversatilis_605	0.0092	0.6587	0.0009	0
Escherichia/Shigella_1459	0.0092	0.6587	0	0.0069
Hydrotalea_72	0.0128	0.6587	0.0006	0
Methyloversatilis_127	0.0128	0.6587	0.0007	0
ClostridiumXVIII_203	0.0128	0.6587	0.0008	0
Methyloversatilis_973	0.0128	0.6587	0.0004	0
Caulobacter_1198	0.0128	0.6587	0.0004	0
Methyloversatilis_1202	0.0128	0.6587	0.0035	0
Variovorax_1223	0.0128	0.6587	0.0008	0
Bacteroides_1966	0.0128	0.6587	0.0001	0.0008
Methyloversatilis_58	0.0131	0.6587	0.0021	0.0003
Parabacteroides_132	0.0131	0.6587	0.0031	0.0002
Parabacteroides_431	0.0136	0.6587	0.0011	0.0001
Escherichia/Shigella_628	0.0136	0.6587	0.0003	0.0018
Methyloversatilis_253	0.0159	0.6587	0.0011	0.0001
...	...	...	...	...

Tabella 38. *OTU le cui abbondanze relative sono diverse (Test di Wilcoxon, livello di significatività 5%)*

OTU	pvalue	pvalue corretti	TD	TD_C
Parabacteroides_4381	0.0136	1	0.0229	0.0147

Blautia_4226	0.0402	1	0.0002	0.0002
Parabacteroides_3122	0.0486	1	0.0005	0.0005
Bacteroides_166	0.0535	1	0.0052	0.0151
Pseudomonas_232	0.055	1	0.0014	0.0023
Pseudomonas_450	0.0562	1	0.0001	0
Pseudomonas_251	0.0642	1	0.0003	0.0001
Blautia_3240	0.0703	1	0	0.0003
...	...	...	...	...

Tabella 39. *OTU le cui abbondanze relative sono diverse (Test di Wilcoxon, livello di significatività 5%)*

OTU	p-value	pvalue corretti	TD	TM
Parabacteroides_1321	0.0004	0.3465	0.0006	0
ClostridiumXVIII_159	0.0006	0.3465	0.0012	0
Parabacteroides_1079	0.0007	0.3465	0.0025	0
Parabacteroides_721	0.001	0.3713	0.0007	0.0001
Parabacteroides_2409	0.0015	0.4208	0.0041	0.0001
Parabacteroides_690	0.0017	0.4208	0.003	0.0002
Parabacteroides_132	0.0024	0.4826	0.0022	0.0002
Parabacteroides_320	0.0026	0.4826	0.0031	0.0002
Parabacteroides_414	0.0035	0.5692	0.0012	0
Parabacteroides_692	0.0043	0.5692	0.0005	0.0001
Parabacteroides_352	0.0046	0.5692	0.0003	0
Parabacteroides_3044	0.0046	0.5692	0.0003	0
Parabacteroides_2508	0.0052	0.594	0.0014	0

Roseburia_73	0.0092	0.6732	0.0007	0
Roseburia_552	0.0092	0.6732	0.0009	0
Lachnospiracea_incertae_sedis_1091	0.0092	0.6732	0	0.0005
Escherichia/Shigella_1459	0.0092	0.6732	0	0.0069
Parabacteroides_1106	0.0122	0.6732	0.0017	0.0001
Hydrotalea_137	0.0128	0.6732	0.0004	0
ClostridiumXVIII_203	0.0128	0.6732	0.0037	0
ClostridiumXVIII_391	0.0128	0.6732	0.0017	0
Parabacteroides_722	0.0128	0.6732	0.0004	0
ClostridiumXVIII_781	0.0128	0.6732	0.0007	0
Parabacteroides_936	0.0128	0.6732	0.0005	0
...	...	...	...	...

Tabella 40. *OTU le cui abbondanze relative sono diverse (Test di Wilcoxon, livello di significatività 5%)*

OTU	p-value	pvalue corretti	TM	TM_C
Lachnospiracea_incertae_sedis_631	0.0015	0.2474	0.0009	0
Bacteroides_1128	0.0015	0.2474	0.0016	0
Bacteroides_1472	0.0025	0.2474	0.0108	0.0001
Roseburia_1559	0.0025	0.2474	0.0026	0
Bacteroides_282	0.0046	0.2474	0.0005	0
ClostridiumXlVa_385	0.0046	0.2474	0.0004	0
Schlesneria_689	0.0046	0.2474	0.0004	0
Escherichia/Shigella_1102	0.0046	0.2474	0.0004	0
Bacteroides_1123	0.0046	0.2474	0.0026	0

Escherichia/Shigella_2055	0.0046	0.2474	0.0009	0
Bacteroides_2933	0.0046	0.2474	0.0006	0
Faecalibacterium_3176	0.0046	0.2474	0.001	0
Bacteroides_8988	0.0046	0.2474	0	0.021
Bacteroides_8994	0.0046	0.2474	0	0.0122
Bacteroides_9006	0.0046	0.2474	0	0.0055
Bacteroides_9012	0.0046	0.2474	0	0.0067
Bacteroides_9041	0.0046	0.2474	0	0.021
Bacteroides_9045	0.0046	0.2474	0	0.0054
Paraprevotella_9425	0.0046	0.2474	0	0.0148
Shimwellia_9859	0.0046	0.2474	0	0.004
Bacteroides_1521	0.005	0.2474	0.0078	0.0001
Bacteroides_1473	0.0071	0.2474	0.0051	0.0002
Bacteroides_1674	0.0071	0.2474	0.0013	0.0001
Bacteroides_1694	0.0071	0.2474	0.0024	0.0002
Lachnospiracea_incertae_sedis_131	0.0092	0.2474	0.0016	0
Bacteroides_410	0.0092	0.2474	0.0038	0
Lachnospiracea_incertae_sedis_1091	0.0092	0.2474	0.0005	0
...	...	...	...	...

Tabella 41. *OTU le cui abbondanze relative sono diverse (Test di Wilcoxon, livello di significatività 5%)*

Dalle tabelle precedenti si osserva che nel confronto tra tessuto sano e tessuto colpito da polipi displastici si evidenziano OTU significativamente più abbondanti in un tessuto rispetto all'altro (p-value inferiori alla soglia dello 0.05). Innanzitutto si nota che nei soggetti sani non sono presenti OTU relative al genere *Schlesneria* e al genere *Lachnospiracea_incertae_sedis*; inoltre

nel tessuto affetto da polipi displastici si hanno quantità significativamente inferiori di OTU afferenti a *Pseudomonas* e a *Bacteroides*. Così come nel confronto effettuato tra i generi, si denota che anche per quanto riguarda le abbondanze delle OTU si evidenziano differenze tra tessuto sano di soggetti sani e tessuto malato di cancro. Ancora una volta, nel tessuto sano sono presenti poche OTU relative ai generi *Schlesneria* e *Bacteroides*.
Per quanto riguarda poi i confronti tra campioni di tessuto sano e lesionato appartenenti a soggetti aventi la stessa diagnosi (Tabelle 36 e 39), non si osservano particolari differenze tra le abbondanze di OTU. Diversamente da quanto riscontrato osservando i generi però, si nota che considerando tessuto sano e tessuto lesionato prelevati da soggetti con tumore al colon emergono OTU che abbondano in un tessuto rispetto all'altro (Tabella 41). Per esempio nel tessuto lesionato dal tumore, adiacente a quello sano, si hanno scarse evidenze di OTU afferenti ai generi *Bacteroides* e *Lachnospiracea_incertae_sedis*.

Si può dunque concludere che, nel momento in cui il tessuto colon-rettale sano subisce lesioni provocate dall'insorgenza di polipi o tumori, il microbiota subisce significative modifiche relativamente alla sua composizione (in base a quanto emerge dai confronti effettuati tra i campioni di tessuto sano di soggetti sani e campioni esportati da soggetti affetti da polipi o tumore). Inoltre si osserva che non solo il microbiota del tessuto lesionato, bensì anche quello del tessuto sano ad esso adiacente subisce le stesse alterazioni (in base ai confronti tra campioni di tessuto prelevati dai soggetti aventi la stessa diagnosi).

9 Conclusioni

In questo lavoro di tesi si è implementata una pipeline di analisi atte a caratterizzare il microbiota di soggetti sani e soggetti affetti da differenti patologie, in modo tale da studiare il comportamento del microbiota in relazione al cambiamento dello stato di salute dell'ospite. Si è così consentito lo sviluppo di uno strumento capace di gestire enormi moli di dati, generati dalle tecnologie di sequenziamento *high-throughput*.

Nello specifico, si è studiata innanzitutto la composizione e l'organizzazione delle comunità batteriche presenti in tessuti di diverso tipo e provenienti da individui in condizioni di salute diverse; in particolare, si sono presi in considerazione soggetti sani, soggetti affetti da bronco pneumopatia cronica ostruttiva, individui con polipi displastici e non a livello colon-rettale, e soggetti affetti da tumore al colon-retto. Per quanto concerne lo studio della composizione del microbiota dei diversi tessuti è stata sufficiente l'osservazione dei dataset, che sono appunto costituiti dalle abbondanze relative delle varie specie batteriche presenti in ciascun soggetto. Invece per caratterizzare l'organizzazione del microbiota, si sono calcolati tre diversi indici di diversità alfa, ottenendo così un'informazione quantitativa sull'eterogeneità delle comunità batteriche; tra i tre indici, quello che fornisce indicazioni migliori, in quanto interpretabili in modo più immediato, è di sicuro l'indice di Evenness.

Nel caso di studio inerente alla BPCO si è osservato che nei non fumatori sani si registra una distribuzione delle specie più omogenea di quella misurata nei fumatori sani e nei fumatori malati; si può dunque concludere che il fumo causa una diminuzione di certe specie piuttosto che un aumento di altre all'interno del microbiota che abita le pareti polmonari. Lo studio dell'eterogeneità del microbiota si è inoltre rivelato utile a fini diagnostici; osservando infatti la distribuzione dei valori degli indici, è stato possibile effettuare una chiara distinzione tra soggetti fumatori (malati e non) e soggetti non fumatori sani. Il calcolo degli indici di diversità alfa ed un'opportuna rappresentazione grafica dei valori ottenuti (si veda Figura 30) può dunque considerarsi un possibile strumento di diagnosi per la BPCO, che come già detto è difficile da diagnosticare con le tecniche attualmente a disposizione. Per confermare la validità di quanto appena affermato si riporta di seguito un episodio emerso durante lo studio. Dopo aver calcolato gli indici di diversità alfa nei soggetti non fumatori sani, si è notato che uno dei tre soggetti (denominato NF2) presentava valori diversi rispetto agli altri due, ma assimilabili a quelli calcolati per i fumatori sani. Dopo opportune indagini è emerso che il soggetto in questione, dichiaratosi inizialmente un non fumatore, aveva in realtà mentito ed era un fumatore.

Nel secondo caso di studio, si sono effettuate le stesse analisi inerenti alla composizione ed all'organizzazione del microbiota prelevato però dal tessuto costituente le pareti del colon di soggetti sani, di soggetti con polipi displastici e non, e di individui affetti da tumore. Ancora una volta si è osservato che le specie batteriche sono distribuite in modo più omogeneo nei tessuti sani rispetto a quelli che risultano alterati, a causa della presenza di polipi o di cancro. È dunque evidente che l'insorgenza di uno stato patologico sia associato alla scomparsa di alcune specie che compongono il microbiota e la comparsa o l'aumento in numerosità di altre.

Constatato che il cambiamento dello stato di salute di un soggetto è in relazione ad un'alterazione del suo microbiota, si sono effettuate delle analisi statistiche al fine di verificare se le differenze tra il microbiota di soggetti con stati di salute diversi siano significative o meno. Vista la natura dei dati a disposizione si è ricorsi a metodi non parametrici, in particolare a NPMANOVA e ANOSIM; il primo metodo si basa sull'analisi della varianza, mentre il secondo sull'analisi della similarità. Si è scelto di utilizzare due metodi affinché la stima della diversità goda di maggiore robustezza.

Analizzando i dati relativi alla BPCO si è evinto che il microbiota dei soggetti fumatori (malati e non) differisce significativamente da quello dei non fumatori, così come i soggetti affetti da BPCO risultano avere un microbiota significativamente diverso da quello dei non fumatori sani. Inoltre le comunità batteriche presenti nei fumatori sani rivelano una grande similarità con quelle presenti nei fumatori affetti da BPCO. Sulla base di quest'ultima osservazione si conferma che il fumo sia una delle cause della malattia, ma, vista la differenza tra microbiota dei soggetti sani rispetto a quello dei malati, si sostiene anche l'efficacia dell'analisi del microbiota come strumento diagnostico.

Applicando poi ANOSIM e NPMANOVA ai dati riferiti al caso di studio del tumore al colon, si osserva che sia il tessuto colpito da tumore, sia quello sano ad esso adiacente, presentano un microbiota significativamente diverso da quello del tessuto prelevato da soggetti sani. In particolare si nota che, in occasione dell'insorgenza di un cancro al colon, nel microbiota si registrano maggiori alterazioni nelle regioni limitrofe al tessuto lesionato rispetto a quelle che si presentano nella zona tumorale. Queste evidenze possono essere interpretate sulla base del modello "drivers-passengers" del CRC (Tjalsma et al., 2012); esso afferma che esistono alcuni batteri intestinali (definiti *drivers*) che causano danni al DNA delle cellule epiteliali del colon e alterazioni del microbiota locale, con conseguente proliferazione di altri batteri (definiti *passenegers*) i quali possono avere proprietà oncogeniche o oncosopressive. Con la comparsa dei *passengers*, i *drivers* scompaiono (Figura 38).

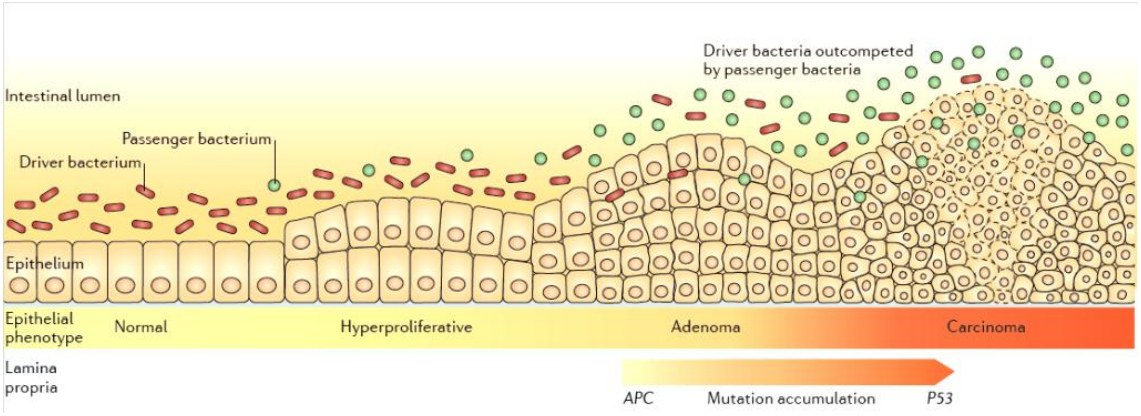


Figura.38. *Rappresentazione della teoria drivers-passengers.*

Il modello appena citato può essere utilizzato per spiegare anche le differenze significative che si registrano tra il microbioma del tessuto affetto da cancro e quello del tessuto sano adiacente.
Inoltre, considerando soggetti con polipi displastici e non displastici, non si sono evidenziate differenze significative a livello di microbiota tra campioni di tessuto diverso ma appartenente a soggetti con il medesimo stato di salute.
Una volta verificato che si riscontrano alterazioni del microbioma in relazione allo stato di salute, si sono identificate le specie batteriche che possono essere associate ad un particolare stato di salute o tipologia di tessuto. A tal fine si è applicato il test di Wilcoxon alle abbondanze relative delle diverse specie batteriche per il confronto tra tutte le possibili coppie di gruppi di soggetti (nel caso della BPCO) o di campioni di tessuto (nel caso del CRC). Si sono così identificate quelle specie che determinano le differenze a livello di microbiota evidenziate nella fase di studio precedente.
Nel momento in cui si dimostri che le alterazioni del microbiota siano conseguenti allo stato patologico, queste ultime analisi possono dunque rivelarsi un utile punto di partenza per la cura delle patologie in questione. Si potrebbe infatti intervenire a livello di comunità batteriche, rimediando agli scompensi creati dalla patologia a livello di composizione del microbiota.

10 Bibliografia

1. Anderson MJ. A new method for non - parametric multivariate analysis of variance. *Austral Ecol*. 2001;26(1):32-46.

2. Anderson MJ. A new method for non - parametric multivariate analysis of variance. *Austral Ecol*. 2001;26(1):32-46.

3. Anderson MJ, Ellingsen KE, McArdle BH. Multivariate dispersion as a measure of beta diversity. *Ecol Lett*. 2006;9(6):683-693.

4. Arumugam M, Raes J, Pelletier E, et al. Enterotypes of the human gut microbiome. *Nature*. 2011;473(7346):174-180.

5. Bray JR, Curtis JT. An ordination of the upland forest communities of southern wisconsin. *Ecol Monogr*. 1957;27(4):325-349.

6. Caporaso JG, Lauber CL, Costello EK, et al. Moving pictures of the human microbiome. *Genome Biol*. 2011;12(5):R50.

7. Cauci S, Driussi S, De Santo D, et al. Prevalence of bacterial vaginosis and vaginal flora changes in peri-and postmenopausal women. *J Clin Microbiol*. 2002;40(6):2147-2152.

8. Cho I, Blaser MJ. The human microbiome: At the interface of health and disease. *Nature Reviews Genetics*. 2012;13(4):260-270.

9. Cho MH, Boutaoui N, Klanderman BJ, et al. Variants in FAM13A are associated with chronic obstructive pulmonary disease. *Nat Genet*. 2010;42(3):200-202.

10. Cho MH, Castaldi PJ, Wan ES, et al. A genome-wide association study of COPD identifies a susceptibility locus on chromosome 19q13. *Hum Mol Genet*. 2012;21(4):947-957.

11. Clarke KR. Non - parametric multivariate analyses of changes in community structure. *Aust J Ecol*. 1993;18(1):117-143.

12. Clarke K, Gorley R. Primer (plymouth routines in multivariate ecological research) v5: User manual. *Tutorial, PRIMER-E Ltd, Plymouth*. 2001;91.

13. Dethlefsen L, Huse S, Sogin ML, Relman DA. The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing. *PLoS biology*. 2008;6(11):e280.

14. Dethlefsen L, Relman DA. Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation. *Proceedings of the National Academy of Sciences*. 2011;108(Supplement 1):4554-4561.

15. Erb-Downward JR, Thompson DL, Han MK, et al. Analysis of the lung microbiome in the "healthy" smoker and in COPD. *PLoS One*. 2011;6(2):e16384.

16. Feghali-Bostwick CA, Gadgil AS, Otterbein LE, et al. Autoantibodies in patients with chronic obstructive pulmonary disease. *American journal of respiratory and critical care medicine*. 2008;177(2):156-163.

17. Ferraro SP, Cole FA. Ecological periodic tables for benthic macrofaunal usage of estuarine habitats in the US pacific northwest. *Estuar Coast Shelf Sci*. 2011;94(1):36-47.

18. Garrity G, Bell J, Lilburn T. *Bergey's Taxonomic Outline*. 2004.

19. Gaston KJ, McArdle BH, Gaston KJ, McArdle BH. The temporal variability of animal abundances: Measures, methods and patterns. *Philosophical Transactions of the Royal Society of London.Series B: Biological Sciences*. 1994;345(1314):335-358.

20. Grice EA, Segre JA. The skin microbiome. *Nature Reviews Microbiology*. 2011;9(4):244-253.

21. Hancock DB, Eijgelsheim M, Wilk JB, et al. Meta-analyses of genome-wide association studies identify multiple loci associated with pulmonary function. *Nat Genet*. 2009;42(1):45-52.

22. Harrison S, Ross SJ, Lawton JH. Beta diversity on geographic gradients in britain. *J Anim Ecol*. 1992:151-158.

23. Harte J, Kinzig AP. On the implications of species-area relationships for endemism, spatial turnover, and food web patterns. *Oikos*. 1997:417-427.

24. Hill MO. Correspondence analysis: A neglected multivariate method. *Applied statistics*. 1974:340-354.

25. Hochberg Y. A sharper bonferroni procedure for multiple tests of significance. *Biometrika*. 1988;75(4):800-802.

26. Holling CS. Resilience and stability of ecological systems. *Annu Rev Ecol Syst*. 1973:1-23.

27. Huse SM, Ye Y, Zhou Y, Fodor AA. A core human microbiome as viewed through 16S rRNA sequence clusters. *PloS one*. 2012;7(6):e34242.

28. Jaccard P. The distribution of the flora in the alpine zone. 1. *New Phytol*. 2006;11(2):37-50.

29. Janssens W, Bouillon R, Claes B, et al. Vitamin D deficiency is highly prevalent in COPD and correlates with variants in the vitamin D-binding gene. *Thorax*. 2010;65(3):215-220.

30. Jeffery IB, Claesson MJ, O'Toole PW, Shanahan F. Categorization of the gut microbiota: Enterotypes or gradients? *Nature Reviews Microbiology*. 2012;10(9):591-592.

31. Jolicoeur P, Mosimann JE. Size and shape variation in the painted turtle. A principal component analysis. *Growth*. 1960;24(4):339-354.

32. Kau AL, Ahern PP, Griffin NW, Goodman AL, Gordon JI. Human nutrition, the gut microbiome and the immune system. *Nature*. 2011;474(7351):327-336.

33. Koleff P, Gaston KJ, Lennon JJ. Measuring beta diversity for presence–absence data. *J Anim Ecol*. 2003;72(3):367-382.

34. Ley RE, Turnbaugh PJ, Klein S, Gordon JI. Microbial ecology: Human gut microbes associated with obesity. *Nature*. 2006;444(7122):1022.

35. McNulty NP, Yatsunenko T, Hsiao A, et al. The impact of a consortium of fermented milk strains on the gut microbiome of gnotobiotic mice and monozygotic twins. *Science translational medicine*. 2011;3(106):106ra106.

36. Parsonnet J, Hansen S, Rodriguez L, et al. Helicobacter pylori infection and gastric lymphoma. *N Engl J Med*. 1994;330(18):1267-1271.

37. Peek Jr RM, Blaser MJ. Helicobacter pyloriand gastrointestinal tract adenocarcinomas. *NATURE REVIEWS| CANCER*. 2002;2:29.

38. Plottel CS, Blaser MJ. Microbiome and malignancy. *Cell Host & Microbe*. 2011;10(4):324-335.

39. Price LB, Liu CM, Melendez JH, et al. Community analysis of chronic wound bacteria using 16S rRNA gene-based pyrosequencing: Impact of diabetes and antibiotics on chronic wound microbiota. *PLoS One*. 2009;4(7):e6462.

40. Rao CR. The use and interpretation of principal component analysis in applied research. *Sankhyā: The Indian Journal of Statistics, Series A*. 1964:329-358.

41. Raymond J, Thiberge JM, Chevalier C, et al. Genetic and transmission analysis of helicobacter pylori strains within a family. *Emerging infectious diseases*. 2004;10(10):1816.

42. Savage DC. Microbial ecology of the gastrointestinal tract. *Annual Reviews in Microbiology*. 1977;31(1):107-133.

43. Simes RJ. An improved bonferroni procedure for multiple tests of significance. *Biometrika*. 1986;73(3):751-754.

44. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. The human microbiome project. *Nature*. 2007;449(7164):804-810.

45. Vanhoutvin SALW, Troost FJ, Hamer HM, et al. Butyrate-induced transcriptional changes in human colonic mucosa. *PloS one*. 2009;4(8):e6759.

46. Walker B, Holling CS, Carpenter SR, Kinzig A. Resilience, adaptability and transformability in social--ecological systems. *Ecology and society*. 2004;9(2):5.

47. Whittaker RH. Vegetation of the siskiyou mountains, oregon and california. *Ecol Monogr*. 1960;30(3):279-338.

48. Wilcoxon F. Individual comparisons by ranking methods. *Biometrics Bulletin*. 1945;1(6):80-83.

49. Wilk JB, Chen T, Gottlieb DJ, et al. A genome-wide association study of pulmonary function measures in the framingham heart study. *PLoS genetics*. 2009;5(3):e1000429.

50. Willyard C. Diagnosis: To catch a killer. *Nature*. 2012;489(7417):S8-S9.

51. Wu GD, Chen J, Hoffmann C, et al. Linking long-term dietary patterns with gut microbial enterotypes. *Science*. 2011;334(6052):105-108.

52. Zosky GR, Berry LJ, Elliot JG, James AL, Gorman S, Hart PH. Vitamin D deficiency causes deficits in lung function and alters lung structure. *American journal of respiratory and critical care medicine*. 2011;183(10):1336-1343.