Università degli studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Magistrale in

Scienze Statistiche

Relazione Finale

Metodi di analisi del Microbioma: Progetti Microbioma Americano e Italiano

Relatrice: Professoressa Chiara Romualdi
Dipartimento di Biologia

Co-relatore: Dottor Andrea Telatin
BMR Genomics

Laureando: Matteo Calgaro
Matricola: 1131045

Anno Accademico 2017/2018

Indice

Elenco delle tabelle

Elenco delle figure

Sommario

Il corpo umano è popolato da una miriade di microorganismi, il cosiddetto Microbiota. Studiando il corredo genetico di questi microrganismi è possibile ottenere informazioni utili in campo medico ed epidemiologico.
Esistono studi sul microbioma a livello globale: in Italia, attraverso uno stage nell'azienda responsabile del Progetto Microbioma Italiano è stato possibile, a partire dai campioni biologici raccolti fino ad oggi, mettere a punto dei metodi di analisi, per rispondere ad alcune domande biologiche.
Il punto di partenza è il campione biologico, dal quale viene estratto e sequenziato il DNA batterico. Questo procedimento fornisce la tabella delle *Operational Taxonomic Unit* (OTU) avente le unità tassonomiche nelle righe e i campioni biologici nelle colonne. Si tratta di un dato generalmente caratterizzato da un'elevata variabilità e sparsità. Per ogni campione biologico sono disponibili un insieme di metadati relativi alle caratteristiche del partecipante allo studio.
La fallacia insita nel processo di sequenziamento, rende necessaria una fase preliminare di filtraggio e normalizzazione della tabella delle OTU. Delle numerose normalizzazioni presenti in letteratura, viene presentata una rassegna delle più importanti. Partendo dagli approcci più intuitivi (*Total Sum Scaling* e Rarefazione), sviluppati storicamente nelle prime fasi della ricerca circa l'analisi del microbioma, si passa a metodi più recenti (*Cumulative Sum Scaling*), o provenienti dall'analisi RNA-Seq (*Trimmed Mean of M-values* e *Relative Log Expression*), in grado di sfruttare le caratteristiche peculiari del dato di microbioma.
Per individuare le unità tassonomiche differenzialmente abbondanti tra gruppi sperimentali, si ipotizzano, per i valori di conteggio normalizzati, specifici

modelli statistici. In questo elaborato vengono approfondite la modellazione Binomiale Negativa (NB) e il modello mistura Normale *Zero-Inflated* (ZIG) in quanto le più utilizzate in letteratura.
Dopo questa fase preliminare attuata per acquisire il *know-how* necessario per gestire un'analisi su dati reali vengono trattati due casi studio: il primo relativo ai dati dell'*American Gut Project* e il secondo relativo ai dati del Progetto Microbioma Italiano. Gli strumenti di normalizzazione e inferenza sono stati messi alla prova per testare l'abbondanza differenziale del microbioma in gruppi di individui con abitudini alimentari diverse. È stato individuato un segnale biologico moderato per i dati americani che sta ad indicare una differenza tra i microbiomi di vegetariani e di onnivori; nei dati italiani invece, non sono state trovate differenze significative tra i microbiomi di vegetariani e i microbiomi di individui con dieta italiana standard. Da un punto di vista statistico, tra i risultati spicca la difficoltà di adattare una *pipeline* di analisi predefinita a questa tipologia di dato, aprendo la strada per l'implementazione di tecniche che consentano all'utente una valutazione critica del procedimento di normalizzazione e inferenza.

Capitolo 1

Introduzione

1.1 Il Microbioma

L'essere umano è colonizzato da una moltitudine di microorganismi: alcuni studi affermano che per ogni cellula umana siano presenti fino a 3 cellule batteriche (Gill et al. 2006).
L'insieme di tutti i microorganismi presenti nel nostro corpo, dai tessuti alle mucose, costituisce il cosiddetto **microbiota**. Per **microbioma** invece, si intende il loro corredo genetico (Willey, Sherwood e Woolverton 2014).
Studi sul microbioma umano hanno rivelato come questo sia soggetto ad una forte variabilità sia tra individui, sia nello stesso individuo in tempi diversi. La composizione della comunità microbica infatti, è influenzata dalla dieta, dall'ambiente, dalla genetica dell'individuo ospitante, dall'etnia e molti altri fattori. L'analisi dei dati relativi al microbioma sta rivoluzionando l'epidemiologia, consente infatti di caratterizzare la comunità microbica di una popolazione sana e intraprendere di conseguenza nuove terapie per trarre beneficio dell'interazione tra microbiota e rispettivo ospite. In questo capitolo viene presentata una panoramica degli studi relativi al microbioma condotti a livello mondiale fino ad arrivare allo stato dell'arte in Italia. In particolare, l'attenzione viene posta nella parte di Microbioma relativa ai microorganismi che popolano l'intestino umano.

1.2 Studi sul Microbioma

Il primo studio sul microbioma, il Progetto Microbioma Umano, può essere visto come l'estensione logica e concettuale del Progetto Genoma Umano. Con l'obiettivo di identificare e caratterizzare i microorganismi che abitano l'uomo, si è cercato di chiarire come il microbioma umano, e le sue variazioni, sia associato allo stato di salute o patologico degli individui. L'iniziativa è partita negli Stati Uniti da parte del National Institute of Health nel 2007, con un progetto quinquennale avente un budget di 115 milioni di dollari. Esiste un consorzio internazionale per lo studio del microbioma che ha come scopo la creazione di un database complessivo che permetta un'analisi a livello globale (*Human Microbiome Project* 2017).
Nel contesto italiano è nato nel 2014 il Progetto Microbioma Italiano, un progetto di ricerca il cui disegno sperimentale si ispira al modello *open access* utilizzato anche dall'*American Gut Project*. In altre parole si tratta di un progetto di *citizen science*, ovvero un tipo di studio che non è finanziato da una struttura privata o da un ente governativo, ma che si basa sul contributo dei singoli partecipanti. A differenza di altri protocolli di ricerca che selezionano un numero limitato di partecipanti a partire da una serie di criteri specifici, il Progetto Microbioma Italiano permette a chiunque di partecipare e anzi, incoraggia alla partecipazione il maggior numero di persone possibili. Lo scopo del progetto è quello di compilare un atlante delle specie batteriche caratteristiche della popolazione italiana che sono state selezionate dal nostro peculiare stile di vita (*Progetto Microbioma Italiano, Citizen Science* 2017). Gli svantaggi di un disegno sperimentale così concepito tuttavia non sono pochi:

- i canali utilizzati per pubblicizzare l'evento, come ad esempio i *social media* oppure gli eventi organizzati nelle piazze italiane, non riescono a raggiungere una popolazione di riferimento definita in modo rigoroso;
- persone con risorse economiche limitate avranno difficoltà a partecipare;
- una predisposizione a particolari patologie può portare alcuni individui ad essere più sensibili al tema rispetto ad altri individui in salute.

È opportuno tenere conto di possibili distorsioni in fase di analisi e interpretazione dei risultati, dovute alla difficoltà nel reperire un campione adeguato. Tuttavia, grazie alle continue scoperte nel settore e alla veloce diffusione delle informazioni, è plausibile che nel corso degli anni l'analisi del microbioma diventi comune come l'analisi ematica.

1.3 Risultati fondamentali dell'*American Gut Project*

Vista la giovinezza del Progetto Microbioma Italiano, un ragionevole percorso da seguire nella ricerca è quello di acquisire delle conoscenze da un progetto maggiormente sviluppato come quello americano. Molte informazioni presenti in letteratura infatti, derivano proprio da questo progetto, che fornirà i primi spunti di analisi per il progetto italiano.
Per il progetto americano sono disponibili 4658 i campioni biologici, sequenziati e analizzati, provenienti da intestino, pelle e saliva di 3624 partecipanti. Rispetto ad altri studi, questo risulta essere quello avente il maggior numero di campioni rappresentanti diversi gruppi di persone. Nel Novembre 2014 sono stati pubblicati i risultati preliminari del progetto, eccone alcuni (*American Gut Project* 2017):

- il microbioma cambia con la crescita dell'organismo ospitante;
- l'utilizzo di antibiotici riduce la diversità microbica e crea un ambiente intestinale meno salubre;
- un'ampia varietà di vegetali nella dieta si traduce in una maggiore diversità del microbioma intestinale;
- gli alcolici influenzano il microbioma: almeno un bicchiere a settimana ne garantisce una maggior diversità rispetto a coloro che non ne fanno uso.

Capitolo 2

Metodi di analisi

2.1 Il dato

2.1.1 Dal campione biologico alle sequenze di DNA

Per indagare la popolazione batterica residente in una data area del corpo viene utilizzato il sequenziamento genomico *16S rRNA*. Esso si riferisce ad una tecnica di sequenziamento genetico che va alla ricerca del gene ribosomale *16S*, lungo approssimativamente 1500 paia di basi, contenente 9 regioni variabili intervallate da regioni conservate. Una piccola porzione della sub unità del gene in questione (regioni variabili *V3* e *V4*) è utilizzata come standard per la classificazione e identificazione dei microbi, dato che differisce tra microorganismi e non è presente nel genoma nucleare delle cellule umane (*16S ribosomal RNA* 2017).
Tutto parte da un campione biologico prelevato da mucose, cute, materiale fecale, etc. In seguito il campione viene trattato in modo da estrarre gli acidi nucleici (DNA) e separare il materiale genetico batterico da quello umano. Dopo l'amplificazione della regione di interesse tramite PCR, un sequenziatore ne fornisce la sequenza nucleotidica sotto forma di file *.fastq*, da gestire con mezzi informatici.

2.1.2 Dalle sequenze di DNA alle OTU

Prima di arrivare ad una forma utilizzabile, il dato attraversa una fase di pre-processamento e una fase di classificazione tassonomica.
Nella fase di pre-processamento, il dato si presenta come un file di testo *.fastq* con lo scopo di immagazzinare due tipi di informazione: la sequenza nucleotidica e un indice di qualità per ogni nucleotide letto. La qualità viene espressa come una funzione della probabilità che una certa base sia stata letta correttamente. A partire da questo indicatore è possibile compiere un primo filtraggio dei dati in base alla loro qualità. In seguito vengono rimossi dal testo i *primers* e avviene il *demultiplexing* utilizzando i *barcodes*. I *primers* sono delle sequenze nucleotidiche sintetiche appositamente create per formare i punti di innesco per la replicazione del DNA utilizzati in fase di amplificazione. I *barcodes* invece, sono utilizzati per identificare la provenienza dei campioni inseriti nel sequenziatore: i sequenziatori di ultima generazione infatti, sono in grado di leggere campioni di diversi individui simultaneamente per ottimizzare i tempi. Con il *demultiplexing* quindi, tutte le sequenze provenienti da un certo individuo vengono raggruppate e separate dalle altre. Utilizzando strumenti di analisi di sequenze, vengono individuate tutte le sequenze diverse e queste a loro volta, vengono organizzate in cluster allo scopo di eliminare le cosiddette *chimere*, ossia prodotti ibridi tra più sequenze di materiale genetico che potrebbero essere erroneamente interpretate come nuovi organismi. I cluster prendono il nome di OTU, acronimo per *Operational Taxonomic Unit*. Una OTU rappresenta tutte le sequenze di un campione con un livello di similarità almeno del 97%, tale soglia è decisa dal ricercatore ed è quella generalmente usata dalla comunità scientifica. Maggiore è il livello di similarità richiesto a due sequenze per appartenere allo stesso cluster, maggiore è il numero di OTU distinte individuate nel campione: la scelta dell'appropriato valore soglia è fondamentale per non gonfiare il numero di OTU a causa di errori nel sequenziamento oppure per individuare specie batteriche con sequenze più variabili di altre (Nguyen et al. 2016).
Si giunge infine alla fase di classificazione tassonomica in cui la sequenza rappresentativa di ciascuna OTU viene cercata in un database continuamente aggiornato (solitamente Greengenes) dove alla sequenza corrisponde il "no-

me" dell'essere vivente con tale sequenza. Il "nome" viene fornito in formato tassonomico visto la quantità e diversità dei microrganismi che ci abitano. Ad esempio, il formato tassonomico per lo *Staphilococcus aureus*, batterio che abita sulla nostra pelle, è presente in Tabella 2.1.

Tabella 2.1: Classificazione tassonomica *Staphilococcus aureus*.

	Classificazione tassonomica
Dominio	Bacteria
Phylum	Firmicutes
Classe	Bacilli
Ordine	Bacillales
Famiglia	Staphylococcaceae
Genere	Staphylococcus
Specie	*aureus*

2.1.3 Tabella delle OTU e Metadati

Una volta individuate le OTU, è possibile creare la relativa tabella di contingenza. Per indagare il microbioma infatti, oltre ad un'informazione qualitativa sulla varietà delle specie batteriche presenti, è fondamentale un'informazione quantitativa sulla loro abbondanza nei campioni. Questo tipo di informazione si traduce in una matrice che ha per righe le OTU e per colonne i campioni. Per ogni campione, il conteggio del numero di sequenze che appartengono ad una determinata OTU va a popolare la tabella (Tabella 2.2). Il numero totale di sequenze lette in un campione, invece, prende il nome di *library size* e si calcola come $L_j = \sum_{i=1}^{N} c_{ij}$.
Se non tenute in considerazione, molte caratteristiche di tale matrice possono portare a risultati errati nelle analisi a valle. Innanzitutto, la comunità microbica di ogni campione biologico può essere rappresentata da una numero molto diverso di sequenze:

- per motivi biologici dovuti al particolare campione in analisi;
- per motivi tecnici relativi, ad esempio, alla dimensione delle librerie utilizzate nel processo di sequenziamento.

Tabella 2.2: Esempio di tabella di contingenza per dati di microbioma relativa a M campioni ed N unità tassonomiche comunemente nota come *OTU Table*. c_{ij} rappresenta il numero di sequenze lette per l'unità tassonomica i nel campione j. L_j è la *library size* del campione j.

	Campione					
OTU	1	2	$\cdots$	j	$\cdots$	M
1	c_{11}	c_{12}	$\ldots$	c_{1j}	$\ldots$	c_{1M}
2	c_{21}	c_{22}	$\ldots$	c_{2j}	$\ldots$	c_{2M}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\iddots$	$\vdots$
i	c_{i1}	c_{i2}	$\cdots$	c_{ij}	$\cdots$	c_{iM}
$\vdots$	$\vdots$	$\vdots$	$\iddots$	$\vdots$	$\ddots$	$\vdots$
N	c_{N1}	c_{N2}	$\ldots$	c_{Nj}	$\ldots$	c_{NM}
Totale	L_1	L_2	$\ldots$	L_j	$\ldots$	L_M

In quest'ultimo caso, piuttosto che una vera variabilità biologica, viene mostrata una variabilità frutto dell'efficienza degli strumenti di sequenziamento utilizzati. In secondo luogo, la maggior parte delle tabelle delle OTU sono sparse, il che significa che contengono una percentuale variabile ma elevata di conteggi a zero. Questa sparsità implica che i conteggi di OTU rare siano incerti in quanto: sono al limite della capacità di rilevazione dei sequenziatori quando ci sono molte sequenze per campione (campioni con *library size* elevata), mentre non sono rilevabili quando ci sono poche sequenze per campione (*library size* contenute). Inoltre, il numero totale di sequenze che appartengono ad un certo campione non rispecchia il numero assoluto di microbi presenti, in quanto il campione è solo una frazione dell'ambiente originale (Weiss et al. 2017).

Ad ogni campione sono associate delle altre variabili che formano i metadati. Nel Progetto Microbioma Italiano, ad esempio, queste variabili corrispondono alle risposte di un questionario compilato dai soggetti partecipanti prima di spedire il campione per l'analisi.

Questi sono gli ingredienti di base per l'analisi del microbioma: la profondità di campionamento irregolare, la sparsità delle tabelle delle OTU e il fatto che i ricercatori siano interessati a trarre conclusioni inferenziali sull'abbondan-

za dei microbi nell'ecosistema di riferimento in base alle caratteristiche dei pazienti campionati.

2.2 Normalizzazione

Le caratteristiche problematiche delle tabelle delle OTU possono essere in parte mitigate dalla normalizzazione. I dati subiscono una serie di trasformazioni attraverso l'utilizzo di processi computazionali in modo da consentire un confronto accurato delle statistiche provenienti da campioni diversi. Esistono varie procedure di normalizzazione in letteratura, nate principalmente dall'analisi dei dati *RNA-seq* e in seguito adattate per il dato di microbioma. Se per le analisi *RNA-seq* si tratta di un passo fondamentale da compiere prima dell'analisi di differenziale espressione genica, nei dati di microbioma diventa il passo fondamentale da compiere prima dell'analisi di abbondanza differenziale microbiologica. In questo capitolo viene presentata una panoramica di queste metodologie di normalizzazione da un punto di vista teorico e successivamente se ne valutano le prestazioni. Le assunzioni su cui si basano la maggior parte delle normalizzazioni per dati *RNA-seq* sono sintetizzabili nei seguenti punti:

- esistono pochi geni differenzialmente espressi;
- esiste una simmetria tra sovra e sotto espressi;
- l'espressione differenziale non dipende dalla media del segnale.

La similitudine tra il dato RNA-seq e il dato di microbioma consiste nel paragonare un gene (della tabella dei geni RNA-seq) ad un'unità tassonomica (della tabella delle OTU). Le assunzioni su cui si baseranno le normalizzazioni applicate al dato di microbioma diventeranno quindi:

- esistono un numero contenuto di OTU differenzialmente abbondanti;
- le unità tassonomiche caratterizzate da abbondanza differenziale sono equamente ripartite tra unità tassonomiche sovra e sotto abbondanti
- l'abbondanza differenziale in un campione non dipende dalla media delle sequenze contate di quel campione.

2.2.1 Proporzioni e rarefazione

Storicamente, i primi approcci utilizzati per confrontare diversi campioni biologici consistevano nell'utilizzare delle semplici proporzioni, oppure la cosiddetta rarefazione. Partendo dall'idea più semplice, ossia quella dell'utilizzo delle proporzioni, si consideri la tabella generica delle OTU visibile in Tabella 2.2 in cui ogni valore viene rapportato al numero totale di sequenze appartenenti al campione, cioè la *library size*. Si ottiene una matrice delle abbondanze relative, visibile in Tabella 2.3, in cui ogni colonna, per costruzione, somma ad 1. Il generico elemento di posto i, j quindi è una proporzione definita come $p_{ij} = c_{ij}/L_j$ con $i = 1, 2, \ldots, N$ OTU e $j = 1, 2, \ldots, M$ campioni.

Tabella 2.3: Frequenze relative per la tabella delle OTU con M campioni e N unità tassonomiche. p_{ij} rappresenta il rapporto tra il numero di sequenze lette relative al campione j per l'OTU i e L_j, ossia la *library size* del campione j

	Campione					
OTU	1	2	$\cdots$	j	$\cdots$	M
1	p_{11}	p_{12}	$\ldots$	p_{1j}	$\ldots$	p_{1M}
2	p_{21}	p_{22}	$\ldots$	p_{2j}	$\ldots$	p_{2M}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\iddots$	$\vdots$
i	p_{i1}	p_{i2}	$\cdots$	p_{ij}	$\cdots$	p_{iM}
$\vdots$	$\vdots$	$\vdots$	$\iddots$	$\vdots$	$\ddots$	$\vdots$
N	p_{N1}	p_{N2}	$\ldots$	p_{Nj}	$\ldots$	p_{NM}
Totale	1	1	$\ldots$	1	$\ldots$	1

In questo modo, due campioni con *library size* molto diverse diventano confrontabili poiché tutti i valori sono compresi tra 0 e 1 e la loro somma è fissata. Tuttavia, considerando le sole proporzioni, le informazioni relative alle *library size* e di conseguenza alla varianza degli stimatori delle proporzioni vanno perse. In altre parole, una proporzione pari a 0.5 in un campione con 100 di *library size* ha lo stesso valore di una proporzione pari a 0.5 in un campione con 1000 di *library size* anche se la varianza dello stimatore

nel secondo campione è minore rispetto a quella nel primo. A riprova di ciò, un recente studio ha mostrato che questo tipo di normalizzazione può funzionare quando vengono confrontati microbiomi provenienti da ambienti e/o ecosistemi completamente diversi in cui le unità tassonomiche in comune nei diversi ecosistemi sono solo una minoranza (McMurdie e Holmes 2014).
Passando alla rarefazione, questa procedura di normalizzazione consiste generalmente in 3 passi:

1. sia $L_j = \sum_{i=1}^{N} c_{ij}$ la *library size* del j-esimo campione, selezione di un valore soglia minimo di *library size* chiamato livello di rarefazione L_{min};

2. scarto dei campioni con *library size* minore di L_{min};

3. per i campioni rimanenti, sotto-campionamento delle conte per ogni unità tassonomiche senza reinserimento in modo da ottenere $L_j = L_{min}$.

Si nota che il primo passo della procedura prevede una scelta da parte del ricercatore introducendo soggettività nel metodo. La scelta del livello di rarefazione ha ripercussioni su tutto il resto dell'analisi: un livello di rarefazione troppo elevato porta a scartare un grande numero di campioni al secondo passo, mentre un livello troppo permissivo comporta perdita di informazione, tanto maggiore quanto più grande è la differenza tra ogni *library size* e il livello di rarefazione, dovuta al sotto-campionamento nel terzo passo. Solitamente L_{min} è scelta come la più piccola *library size* considerata qualitativamente accettabile. Il sequenziatore, infatti, non fornisce una lettura sempre omogenea: per motivi legati alla tecnologia utilizzata, accade che alcuni campioni biologici vengano letti in modo più superficiale di altri risultando di qualità inferiore. A prescindere da questo, i passi 1 e 3 della procedura di normalizzazione prevedono l'eliminazione, soggettiva al passo 1 e casuale al passo 3, di una parte dell'informazione disponibile.
Le implicazioni di questa normalizzazione sono ben descritte in letteratura (McMurdie e Holmes 2014), ne viene proposto qui in seguito un esempio esplicativo: il livello di rarefazione è fissato a 100 e si vogliono confrontare due campioni contenenti solamente due tipi di microrganismi, $OTU1$ e $OTU2$. Il Campione A ha una *library size* di 100 e il Campione B una *library size* di 1000, Tabella 2.4 (sinistra). Dopo aver eseguito una normalizzazione tramite

rarefazione si ottiene la Tabella 2.4 (destra). Il confronto tra i due campioni può essere fatto tramite un test χ^2 di Pearson oppure, dato che l'esempio riguarda una tabella di contingenza di dimensioni 2x2, il test esatto di Fisher. I rispettivi *p-value* sono riportati in Tabella 2.5. Se prima di essere rarefatti i due campioni risultavano significativamente diversi, una volta ricondotti alla stessa *library size*, essi non sono più distinguibili. Questa perdita di potenza è completamente attribuibile alla riduzione della dimensione del Campione B di 10 volte. La riduzione della numerosità aumenta l'ampiezza degli intervalli di confidenza di ogni proporzione del Campione B rendendole indistinguibili dalle proporzioni del Campione A. Se nell'esempio appena visto la rarefazione porta una riduzione di 10 volte della dimensione del campione originale, nella pratica le riduzioni possono variare anche in misura maggiore. Le *library size* nei dati dell'*American Gut Project* ad esempio, sono molto variabili, passando da un minimo di qualche centinaio di sequenze per campione ad un massimo di circa 95000 sequenze. La riduzione risultante sarebbe dell'ordine di qualche centinaio di volte.

Tabella 2.4: Tabelle delle OTU di esempio con 2 campioni e 2 unità tassonomiche per i dati grezzi (sinistra) e i dati rarefatti (destra).

Dati grezzi		
	Campione	
OTU	*A*	*B*
1	62	500
2	38	500
Totale	100	1000

Dati rarefatti		
	Campione	
OTU	*A*	*B*
1	62	50
2	38	50
Totale	100	100

Tabella 2.5: *p-value* dei Test χ^2 di Pearson e il Test esatto di Fisher F per il confronto dei campioni in Tabella 2.4, prima e dopo la rarefazione

p-value	**Test**	
	χ^2	F
Campioni non normalizzati	0.0290	0.0272
Campioni rarefatti	0.1171	0.1169

Entrambe queste metodologie risultano dunque poco adatte per iniziare l'analisi di abbondanza differenziale di conseguenza ci si potrebbe chiedere come mai continuino ad essere utilizzate. La risposta può essere rintracciata nel modo in cui l'analisi del microbioma si è evoluta nel corso degli ultimi anni. I primi studi, infatti, avevano l'obiettivo di esplorare e descrivere il microbioma presente in campioni spesso provenienti da ambienti/ecosistemi diversi (ad esempio oceano vs. feci) (Gotelli e Colwell 2001). Le procedure di analisi prevedevano di iniziare con la rarefazione che diventò così una presenza costante nei pacchetti di analisi utilizzati dai ricercatori. In seguito, l'evoluzione tecnologica con la conseguente riduzione dei costi di sequenziamento e la possibilità di analizzare diversi campioni in parallelo, aumentò l'accessibilità dei servizi di sequenziamento per rispondere a domande diverse. Oggigiorno, ad esempio, c'è interesse nel determinare abbondanze differenziali tra OTU in determinate classi di campioni di provenienza analoga (ad esempio muco fumatore vs. muco non fumatore) (Charlson et al. 2011). I pacchetti di analisi contengono ancora quegli strumenti ma sta al ricercatore capire quale tipo di normalizzazione sia più appropriata per i dati che sta analizzando: uno strumento nasce per soddisfare un bisogno, se il bisogno muta anche lo strumento può necessitare di qualche cambiamento.

2.2.2 Normalizzazioni basate sui quantili

Nel precedente paragrafo è stata presentata la normalizzazione basata sulle proporzioni in cui ogni elemento p_{ij} della Tabella 2.3 viene riscalato in base alla *library size* L_j. Questa normalizzazione è anche nota con il nome di *Total Sum Scaling* (TSS) e le normalizzazioni basate sui quantili partono esattamente dall'idea di riscalare i dati, come in questa normalizzazione più semplice, ma tenendo in considerazione alcune particolarità degli stessi.
Le *library size* sono molto sensibili a cambiamenti di frequenza per le unità tassonomiche più abbondanti. A prova della precedente affermazione, si immagini un incremento di conteggi per poche OTU già molto popolate: la conseguenza è che le proporzioni delle altre unità tassonomiche, meno ricche di sequenze, diventino ancora più piccole. Se queste proporzioni ridotte vengono confrontate con quelle di altri campioni potrebbero sembrare si-

gnificativamente diverse portando ad un alto numero di falsi positivi e una conseguente perdita di potenza nell'individuare le vere differenze (Soneson e Delorenzi 2013).
Una prima scelta per risolvere il problema è, dopo aver rimosso tutte le OTU con 0 sequenze, quella di calcolare per ogni campione il terzo quartile della distribuzione delle sequenze lette e utilizzarlo come fattore di scala. Il terzo quartile è stato scelto perché considerato alto abbastanza per cadere all'esterno dell'intervallo formato dalle unità tassonomiche con zero sequenze e le unità tassonomiche più rare. Allo stesso tempo però, è anche sufficientemente basso per non essere influenzato dalle OTU caratterizzate da molte o moltissime sequenze, problema presente con la metodologia TSS. *Upper Quartile Normalization* (UQ) è il nome della normalizzazione, chiaro è il riferimento tra il nome e il quantile utilizzato.
Il metodo chiamato *Cumulative Sum Scaling* (CSS), invece, è un'estensione rispetto all'UQ: le conte grezze delle sequenze lette vengono rapportate alla somma cumulata delle sequenze fino ad un determinato quantile determinato usando un approccio guidato dai dati stessi (Paulson et al. 2013). A partire dalla Tabella 2.2 sia q_j^l il quantile l-esimo per il campione j (nel campione j ci sono l unità tassonomiche con $c_{ij} < q_j^l$). Sia $s_j^l = \sum_{i|c_{ij} \leq q_j^l} c_{ij}$ la somma cumulata delle conte per il campione j fino al quantile l-esimo. Le conte normalizzate sono calcolate tramite $\tilde{c}_{ij} = (c_{ij}/s_j^l)F$ dove F è una costante di normalizzazione opportunamente scelta uguale per tutti i campioni. Gli autori del metodo consigliano di scegliere come costante di normalizzazione F la mediana dei fattori di scala s_j^l. A questo punto la scelta di un quantile l appropriato, risulta critica per assicurare che la normalizzazione non introduca artefatti nei dati: la distribuzione delle conte nei campioni dovrebbe essere all'incirca equivalente e indipendente tra campioni fino al quantile scelto sotto l'assunzione che il dato provenga da una distribuzione comune. Il valore specifico scelto risulterà quindi dipendente dal progetto e probabilmente dipenderà dalle caratteristiche sperimentali (preparazione dei campioni, sequenziamento e successiva analisi bioinformatica). Il procedimento adattivo, guidato dai dati stessi, si occupa di determinare un $\hat{l}$ alla luce delle osservazioni appena viste. Dal punto di vista pratico, viene stimato $\hat{l}$ in cui la distribuzione delle conte campione-specifica, devia da una distribuzione di

riferimento opportunamente definita:

1. Viene definita $\overline{q}^l = med_j\{q_j^l\}$, la mediana dell'$l$-esimo quantile tra campioni, come il quantile l-esimo della distribuzione di riferimento;

2. Con $d_l = med_j|q_j^l - \overline{q}^l|$ viene indicata la deviazione assoluta mediana per i quantili campione-specifici attorno al riferimento;

3. Sotto le assunzioni fissate dal metodo, la quantità d_l è stabile per quantili bassi, al contrario, mostra instabilità per quantili elevati.

4. Viene scelto $\hat{l}$ come il più piccolo valore per cui viene rilevata alta instabilità. L'instabilità viene misurata utilizzando le differenze prime penalizzate. Quindi $\hat{l} = min(l|d^{l+1} - d^l > 0.1d^l)$ in cui il valore 0.1, entità della penalizzazione, è stato scelto in modo arbitrario dagli autori del metodo. Nel caso in cui l'instabilità venga rilevata per valori di $l < 0.5$ allora $\hat{l} = 0.5$.

2.2.3 Altre normalizzazioni

Le ultime due normalizzazioni che vengono affrontate in questo elaborato, sono la normalizzazione RLE, acronimo per *Relative Log Expression* e la Normalizzazione TMM acronimo per *Trimmed Mean of the M-values*. Partendo dalla prima, essa consiste nel calcolo di M fattori di scala, uno per ogni campione. Ogni fattore di scala f_j viene calcolato come la mediana dei rapporti tra le conte del campione j-esimo e le conte di un campione di riferimento: la media geometrica tra tutti i campioni. Cioè partendo dalla Tabella 2.2:

1. viene calcolata la media geometrica per ogni unità tassonomica, allo scopo di creare il campione di riferimento, $R_i = \sqrt[M]{\prod_{j=1}^{M} c_{ij}}$ con $i = 1, 2, \ldots, N$;

2. ogni valore di conta viene riscalato sul valore di riferimento e per ogni campione viene calcolata la mediana, $f_j = med(c_{ij}/R_i)$;

3. per fare in modo che i fattori moltiplichino a 1,

$$f_j = \frac{e^{\frac{1}{M}\sum_{j=1}^{M} log(f_j)}}{f_j}$$

L'utilizzo della mediana, che giace nel range delle unità tassonomiche non differenzialmente abbondanti (data l'assunzione di un basso numero di unità tassonomiche differenzialmente abbondanti e la loro simmetria), rende il metodo robusto rispetto ad OTU rare o molto abbondanti. Per quanto riguarda la Normalizzazione *TMM* invece, entrano in gioco le due quantità *Amplitude* e *Magnitude*. Questo tipo di normalizzazione richiede di scegliere un campione di riferimento tra gli M campioni, solitamente questo corrisponde al campione in cui il terzo quartile dei valori di conteggio è il più vicino alla media dei terzi quartili di tutti i campioni. Per ogni coppia di campioni J e R, in cui R è il campione di riferimento, *Amplitude* rappresenta il vettore delle medie delle log-conte delle N unità tassonomiche, mentre *Magnitude* il vettore delle differenze delle log-conte. Quindi, il generico elemento relativo all'i-esima unità tassonomica avrà:

$$A_i = \frac{log(J_i + 1) + log(R_i + 1)}{2}$$
$$M_i = log(J_i + 1) - log(R_i + 1)$$

con J_i e R_i numero di sequenze rispettivamente nei campioni J e R per l'i-esima unità tassonomica. L'insieme dei valori può essere utilizzato per creare il cosiddetto grafico *MA*, in grado di evidenziare differenze sistematiche di abbondanza lungo i livelli medi di quest'ultima. La normalizzazione consiste nel calcolare la media pesata dei valori M opportunamente troncati: in particolare il 30% dei valori più estremi di M e il 5% dei valori più estremi di A viene troncato (Dillies et al. 2013). All'interno di un campione, ogni unità tassonomica segue una distribuzione binomiale con una certa *library-size* e una certa proporzione. Usando il Metodo Delta è possibile calcolare un'approssimazione della varianza del valore M_i che, se invertita, diventa il peso utilizzato per calcolare la media pesata. Indicando con O^* l'insieme di OTU con valori validi di M e A non troncati, la media pesata per la generica coppia di campioni j, r corrisponde a:

$$TMM(j,r) = \frac{\sum_{OTU \in O^*} w_{OTU}(j,r) M_{OTU}(j,r)}{\sum_{OTU \in O^*} w_{OTU}(j,r)}$$

dove i pesi calcolati tramite metodo delta sono:

$$w_{OTU}(j,r) = \left(\frac{L_j - J_{OTU}}{L_j} \frac{J_{OTU}}{(J_{OTU}+1)^2} + \frac{L_r - R_{OTU}}{L_r} \frac{R_{OTU}}{(R_{OTU}+1)^2} \right)^{-1}$$

per ogni $OTU \in O^*$. Tornando alla scala originale si ottiene $f_j = e^{TMM(j,r)}$. Per fare in modo che tutti i fattori moltiplichino a 1,

$$f_j = \frac{e^{\frac{1}{M}\sum_{j=1}^{M} log(f_j)}}{f_j}$$

Grazie al troncamento, il fattore di scala creato da questo tipo di normalizzazione è robusto rispetto alle unità tassonomiche estremamente abbondanti o rare presenti nei campioni confrontati con quello di riferimento.

2.3 Analisi di abbondanza

Una volta calcolati i fattori di scala di una normalizzazione, è possibile procedere con l'analisi vera e propria del dato di conteggio. Esistono principalmente due metodi per confrontare le unità tassonomiche presenti in campioni diversi.
Il primo viene comunemente chiamato *SAD* acronimo per *Species Abundance Distribution*. A livello applicativo, consiste nell'eliminare tutte le unità tassonomiche con zero sequenze e, per ogni campione, ordinare le rimanenti in ordine decrescente di abbondanza. Il metodo si basa su una delle leggi ecologiche più vecchie e universali: ogni comunità ecologica è popolata da un elevato numero di specie rare e solo alcune specie comuni. L'istogramma delle specie presenti infatti, segue un'iperbole e lo studio di queste curve iperboliche può portare a comprendere come una comunità risponda ai cambiamenti provenienti dall'esterno, in relazione al contesto ecologico in cui essa è integrata (McGill et al. 2007). In questo elaborato la metodologia *SAD* non verrà approfondita ulteriormente poiché per l'*American Gut Project* e di riflesso per il Progetto Microbioma Italiano, risulta più interessante indagare la composizione microbica dei campioni piuttosto che la sola struttura.
Il secondo metodo, a differenza del primo, considera per ogni campione anche

la classificazione tassonomica di tutte le OTU, comprese quelle che contengono zero sequenze. Il metodo va a testare la presenza di abbondanza differenziale di ogni unità tassonomica; in altre parole, determina quali OTU siano più o meno presenti, in modo significativo, in un gruppo di campioni confrontati con un altro gruppo. In questo capitolo viene presentata una panoramica di queste metodologie partendo dall'analisi esplorativa ed arrivando ad approcci inferenziali. Per questi ultimi, una particolare attenzione viene posta sulla modellazione del dato di conta OTU per OTU assumendo una distribuzione Binomiale Negativa (NB) prima e una distribuzione mistura Normale *Zero-Inflated* (ZIG) in seguito.

2.3.1 Analisi esplorativa

Un modo semplice per confrontare la composizione batterica di due microbiomi è quella di calcolare alcune misure di diversità nelle popolazioni batteriche osservate. A questo proposito esiste la misura chiamata **alpha-diversità**. Essa si comporta come una statistica riassuntiva di un singolo campione, che conta il numero di unità tassonomiche e la loro abbondanza. Accanto a questa misura, la **beta-diverità**, funziona come un punteggio di similarità tra campioni diversi. Quest'ultima può essere misurata indagando le unità tassonomiche comuni tra campioni, oppure quantificata con uno specifico indice, l'indice di dissimilarità di *Bray-Curtis*:

$$BC_{AB} = \frac{L_A + L_B - 2 \cdot Comuni_{AB}}{L_A + L_B}$$

dove L_A ed L_B sono le *library size* rispettivamente nei campioni A e B, mentre $Comuni_{AB}$ è, una volta individuate le specie in comune nei campioni A e B, la somma dei valori di conta minori tra i 2 campioni. Per esempio, nella Tabella 2.6, gli elementi che compongono l'indice di *Bray-Curtis* sono $L_A = 50$, $L_B = 50$ e $Comuni_{AB} = 10 + 20 = 30$ dato che le OTU comuni sono la 2 e la 3. L'indice di dissimilarità risultante quindi è $BC_{AB} = 0.4$. Per costruzione l'indice è contenuto nell'intervallo $[0, 1]$, dove lo 0 indica che i due campioni hanno la stessa composizione (condividono cioè tutte le specie presenti), mentre 1, al contrario, significa una completa diversità tra i due campioni. A partire da misure di beta-diversità e indici di similarità, è possibile avere una

Tabella 2.6: Calcolo dell'indice di dissimilarità di *Bray-Curtis* per una tabella delle OTU semplificata.

	Campione	
OTU	*A*	*B*
1	0	10
2	10	20
3	30	20
4	20	0
Totale	50	50

panoramica complessiva dei campioni esaminati: *Cluster Analysis*, tecniche di riduzione delle dimensioni come l'analisi delle componenti principali o il *Multi-Dimensional Scaling* sono alcuni esempi.

2.3.2 Inferenza

Se con lo studio delle misure di diversità è possibile avere un'idea descrittiva delle differenze tra campioni, con gli approcci inferenziali è possibile individuare quali siano le unità tassonomiche differenzialmente abbondanti in diversi gruppi di campioni.

Binomiale Negativa

L'analisi *RNA-Seq* ha portato a comprendere che le variazioni delle conte, per un determinato gene, in **repliche tecniche** (cioè lo stesso campione biologico analizzato ripetutamente) seguono una distribuzione di Poisson (Marioni et al. 2008). Parallelamente, nei dati di microbioma, se le repliche tecniche hanno lo stesso numero di sequenze L_j allora il numero di sequenze per il campione j e l'i-esima unità tassonomica sarà:

$$C_{ij} \sim Poisson(L_j u_i)$$

in cui u_i denota la proporzione dell'OTU i-esima nel campione. Media e varianza per la variabile aleatoria C_{ij} corrispondono alla medesima quantità $\lambda_{ij} = L_j u_i$. I progetti che studiano il microbioma, per poter fare inferenze

riguardanti le popolazioni in esame, utilizzano delle **repliche biologiche**. Queste repliche, a differenza di quelle tecniche, consistono in misure parallele di campioni biologicamente distinti in grado di catturare una componente casuale di variabilità biologica. Se la distribuzione di Poisson assumeva varianza e media equivalenti per i dati di conteggio, con le repliche biologiche la varianza osservata non è più in relazione lineare con la media, un fenomeno conosciuto come sovra-dispersione. Un esempio di tale fenomeno è presente in Figura 2.1 in cui è stata studiata la relazione media-varianza di un insieme di repliche biologiche dell'*American Gut Project*. Questo accade poiché

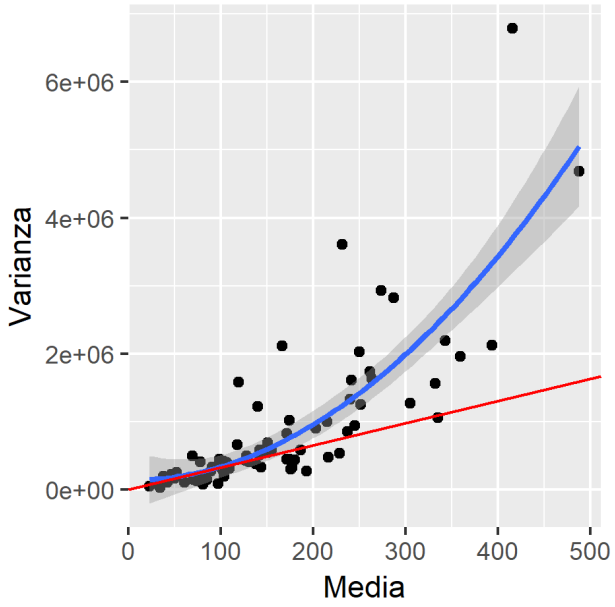

Figura 2.1: Grafico della relazione media-varianza in un insieme di repliche biologiche dell'*American Gut Project*. In rosso è segnata una retta ad indicare una relazione lineare tra le 2 quantità, in blu una relazione quadratica.

le repliche biologiche, oltre ad aggiungere variabilità biologica, sono caratterizzate da *library size* diverse, ulteriore fonte di variabilità. Per risolvere il problema assumiamo che la media della variabile casuale di Poisson, sia essa stessa variabile aleatoria avente una distribuzione Gamma con parametro di

forma r e parametro di scala $p/(1-p)$:

$$\lambda \sim Gamma\left(r, \frac{p}{1-p}\right)$$

Innanzitutto si genera una media casuale, λ, per la Poisson dalla Gamma, e successivamente un valore casuale, c, dalla Poisson(λ). La distribuzione marginale è:

$$\begin{aligned} P(C=c) &= \int_0^\infty Pois(c;\lambda)\cdot\gamma\left(r,\frac{p}{1-p}\right)\,d\lambda \\ &= \int_0^\infty \frac{\lambda^c e^{-\lambda}}{c!}\cdot\frac{\lambda^{r-1}e^{-\lambda\frac{1-p}{p}}}{(\frac{p}{1-p})^r\Gamma(r)}\,d\lambda \\ &= \frac{(1-p)^r}{p^r c!\,\Gamma(r)}\int_0^\infty \lambda^{r+c-1}e^{-\frac{\lambda}{p}}\,d\lambda \\ &= \frac{(1-p)^r}{p^r c!\,\Gamma(r)}p^{r+c}\Gamma(r+c) \\ &= \frac{\Gamma(r+c)}{c!\,\Gamma(r)}p^c(1-p)^r \end{aligned}$$

che corrisponde alla funzione di densità di una variabile casuale con distribuzione Binomiale Negativa di parametri r e p:

$$C \sim NB(r;p)$$

Nella teoria del Calcolo delle Probabilità, la distribuzione Binomiale Negativa è utilizzata per contare il numero di successi in una serie di eventi bernoulliani con probabilità di successo p prima che si verifichino r insuccessi; media e varianza sono:

$$\begin{aligned} E(C) = m &= \frac{pr}{1-p} \\ V(C) &= \frac{pr}{(1-p)^2} \end{aligned}$$

Talvolta viene utilizzata una parametrizzazione diversa che prevede i due parametri: media m e $r = m(1-p)/p$. La marginale diventa quindi:

$$X \sim NB(m;r)$$

con funzione di densità:

$$P(C = c) = \frac{\Gamma(r + c)}{c!\,\Gamma(r)} \left(\frac{r}{r + m}\right) \left(\frac{m}{r + m}\right)^{c}$$

Ricordando che in presenza di repliche biologiche e *library size* diverse è necessario normalizzare la matrice delle OTU, vengono calcolati i fattori di normalizzazione f_j. Ne consegue che le *library size* normalizzate sono calcolate tramite $\tilde{L}_j = L_j/f_j$. Per le repliche biologiche all'interno dello stesso gruppo, come ad esempio gruppo dei vegetariani e gruppo degli onnivori, la proporzione u_i sarà variabile tra i campioni. Quindi riscrivendo la modellazione espressa in precedenza caratterizzando maggiormente le variabili coinvolte si ottiene:

$$\begin{aligned} U_{ij} &\sim Gamma\left(r_i, \frac{p_i}{1 - p_i}\right) \\ C_{ij}|U_{ij} &\sim Poisson(\tilde{L}_j u_i) \\ C_{ij} &\sim NB\left(\tilde{L}_j u_i, \tilde{L}_j u_i \frac{1 - p_i}{p_i}\right) \\ &\sim NB\left(\tilde{L}_j u_i, \frac{1}{\phi_i}\right) \end{aligned}$$

Ricordando dalla seconda parametrizzazione della Binomiale Negativa che $\tilde{L}_j u_i(1 - p_i)/(p_i) = r_i$ e indicando con $\phi_i = 1/r_i$, si introduce il parametro di dispersione per l'i-esima unità tassonomica ottenendo la seguente formulazione per la funzione di densità:

$$P(C_{ij} = c_{ij}; u_i, \phi_i) = \frac{\Gamma(c_{ij} + \frac{1}{\phi_i})}{c_{ij}!\,\Gamma(\frac{1}{\phi_i})} \left(\frac{1}{1 + \phi_i \tilde{L}_j u_i}\right)^{\frac{1}{\phi_i}} \left(\frac{\tilde{L}_j u_i}{\frac{1}{\phi_i} + \tilde{L}_j u_i}\right)^{c_{ij}}$$

con $\Gamma(\cdot)$ funzione gamma, $E(C_{ij}) = \tilde{L}_j u_i$ e $V(C_{ij}) = \tilde{L}_j u_i + \phi_i(\tilde{L}_j u_i)^2$. Da notare che nel caso in cui $\phi_i = 0$, l'unità tassonomica corrispondente segue la distribuzione di Poisson. L'ultimo aspetto di cui tenere conto è il gruppo di cui un campione fa parte, come ad esempio il gruppo dei vegetariani e il gruppo degli onnivori. Questo viene fatto stimando separatamente i valori dei parametri per ognuno dei gruppi presenti. Nella pratica viene introdotto un indice k per le varie condizioni sperimentali, ottenendo per le conte dell'i-esima unità tassonomica, per il j-esimo esperimento e per la condizione k

una Binomiale Negativa così formulata:

$$C_{ij,k} \sim NB\left(\tilde{L}_j u_{i,k}, \tilde{L}_j u_{i,k} \frac{1-p_{i,k}}{p_{i,k}}\right)$$
$$\sim NB\left(\tilde{L}_j u_{i,k}, \frac{1}{\phi_{i,k}}\right)$$

Dunque la relazione media varianza non è più di tipo lineare come nel caso delle repliche tecniche con il dato di conta distribuito secondo una Poisson; bensì la varianza si esprime come la media sommata ad un parametro di dispersione moltiplicato per il quadrato della media stessa:

$$V(C_{ij,k}) = \tilde{L}_j u_{i,k} + \phi_{i,k}(\tilde{L}_j u_{i,k})^2$$

La stima del parametro di disperione ϕ diventa fondamentale nell'individuare le OTU differenzialmente espresse riuscendo a controllare il tasso di falsi positivi. Infatti, molte delle OTU che risultavano differenzialmente abbondanti assumendo una distribuzione di Poisson per il dato di conta, diventano non significative quando i test tengono conto della relazione non lineare media-varianza derivata da $\phi > 0$.
Di seguito vengono illustrate le modalità di stima dei parametri del modello statistico appena descritto. La media delle conte per unità tassonomica e condizione sperimentale $E(C_{ij,k}) = \tilde{L}_j u_{i,k}$ è il prodotto tra $\tilde{L}_j$, la *library size* normalizzata, e $u_{i,k}$, la proporzione dell'unità tassonomica i-esima nei campioni appartenenti alla condizione k. Si noti che la generica stima della proporzione u_i non è influenzata dalla normalizzazione in quanto i fattori di normalizzazione sono di tipo moltiplicativo e quindi:

$$u_i = \frac{\tilde{c}_{ij}}{\tilde{L}_j} = \frac{c_{ij}/f_j}{L_j/f_j} = \frac{c_{ij}}{L_j}$$

In cui $\tilde{c}_{ij}$ rappresentano le conte normalizzate. Quando quest'ultima proporzione diventa specifica della condizione sperimentale k, essa viene stimata tramite la media delle conte normalizzate per l'i-esima unità tassonomica all'interno del gruppo k-esimo. Per quanto riguarda la stima del parametro di dispersione, si tratta di un procedimento più complesso (Robinson e Smyth 2008). In un primo momento viene stimato un parametro di dispersione globale per tutte le unità tassonomiche tramite funzione di verosimiglianza

condizionata e, a partire da questo, tramite stima di massima verosimiglianza vengono calcolati i parametri di dispersione specifici OTU per OTU. Al primo passo, viene utilizzata la verosimiglianza condizionata perché i classici approcci di massima verosimiglianza tendono generalmente a sottostimare la varianza nei casi in cui sia media che varianza vengano stimati dagli stessi dati. Gli autori del metodo hanno preferito questa soluzione dopo aver confrontato le performance di alcuni stimatori per la dispersione cioè:

- modelli di quasi-verosimiglianza e pseudo-verosimiglianza. Questi approcci forniscono statistiche di bontà di adattamento, senza imporre una legge distributiva, per i parametri delle funzioni di varianza in un modello lineare generalizzato.
- Metodi di aggiustamento della verosimiglianza; aggiungendo alla log-verosimiglianza profilo un fattore che tiene conto dell'informazione osservata per la media. In questo caso si parla di massima verosimiglianza condizionata.

Innanzitutto si consideri una singola unità tassonomica i, sia $c_{j,k}$ il valore di conta osservato per la condizione k nel campione j di quella OTU. Ipotizzando un confronto tra due gruppi, come ad esempio onnivori e vegetariani, si ha $k = 1, 2$ e $C_{j,k} \sim NB(L_{j,k}u_k, \phi)$. Per individuare le unità tassonomiche differenzialmente abbondanti tra le 2 condizioni viene effettuato un test per ogni OTU in cui l'ipotesi nulla è $H_0 : u_1 = u_2$ contro l'alternativa $H_1 : u_1 \neq u_2$. La verosimiglianza condizionata per una singola unità tassonomica è ottenuta condizionandosi alle somme delle conte di ciascuna classe, dato che la somma di due variabili casuali NB identicamente distribuite è ancora una NB. Il condizionamento consente di rimuovere il parametro di disturbo u_k. Se le *library size* sono uguali all'interno di ciascuna classe, la log-verosimiglianza condizionata di ϕ per una singola unità tassonomica, data la somma delle conte $S_k = \sum_{j=1}^{M_k} c_{j,k}$ è:

$$l_i(\phi) = \sum_{r=1}^{2} \left[\sum_{j=1}^{M_k} log\Gamma\left(c_{j,k} + \frac{1}{\phi}\right) + log\Gamma\left(\frac{M_k}{\phi}\right) - log\Gamma\left(S_k + \frac{M_k}{\phi}\right) - M_k log\Gamma\left(\frac{1}{\phi}\right) \right]$$

Lo stimatore della dispersione comune massimizza la verosimiglianza comune $l_{comune}(\phi) = \sum_{i=1}^{N} l_i(\phi)$ dove N è il numero di unità tassonomiche.
Le *library size* però non sono tutte uguali: le conte non sono identicamente distribuite e l'argomento condizionante non vale esattamente. Per questo gli autori del metodo utilizzano la massima verosimiglianza condizionata *quantile adjusted*. Dopo il calcolo della media geometrica delle *library size* L^* le conte osservate vengono corrette aumentandole o diminuendole a seconda che la corrispondente libreria sia superiore o inferiore a L^*. Questa trasformazione opera generando delle pseudo-conte, che sono approssimativamente identicamente distribuite, attraverso una normalizzazione quantile dei dati osservati. Le pseudo-conte possono essere inserite nella formula di $l_i(\phi)$ e una volta sommate viene massimizzata la log-verosimiglianza comune rispetto a ϕ.
In genere non è detto che tutte le unità tassonomiche abbiano la stessa dispersione: per questo motivo, a partire dalla dispersione globale si stima la dispersione specifica per OTU ϕ_i che viene stimata tramite massima verosimiglianza pesata (WL), combinando le verosimiglianze individuali con quella comune:

$$WL(\phi_i) = l_i(\phi_i) + wl_{comune}(\phi_i)$$

dove w è il peso dato alla log-verosimiglianza comune. Si noti che definire $w = 0$ equivale ad ottenere delle stime OTU specifiche utilizzando la massima verosimiglianza *quantile adjusted* per ogni OTU. Imponendo invece, w sufficientemente grande, le dispersioni tenderanno a convergere attorno alla dispersione comune. I pesi saranno tanto maggiori quanto più le stime ϕ_i sono diverse (Robinson e Smyth 2008). Per ogni unità tassonomica i, viene eseguito un test per verificare l'abbondanza differenziale tra le due condizioni sperimentali. Si tratta di un test esatto basato sul fatto che la somma di due variabili casuali con distribuzione Binomiale Negativa, identicamente distribuite, hanno ancora distribuzione Binomiale Negativa. Nel metodo presentato, le pseudo-conte godono approssimativamente della proprietà di identica distribuzione. Condizionandosi alla pseudo-somma totale quindi, si può calcolare la probabilità di osservare conte più o meno estreme rispetto a quelle osservate tramite un test a 2 code.

Mistura Normale *Zero-Inflated*

Gli zeri presenti in una tabella delle OTU possono essere di diversa natura e per giustificare l'affermazione bisogna riflettere su come il dato sia stato generato. Come ampiamente spiegato nella sezione 2.1, il sequenziatore è lo strumento attraverso il quale dal campione biologico, opportunamente trattato, è possibile ottenere una serie di valori analizzabili tramite strumenti informatici. Esistono molti tipi di sequenziatori e anche le opzioni di sequenziamento possono cambiare in base alle esigenze della ricerca. Un concetto utile per capire la diversa natura degli zeri presenti nei dati è la cosiddetta **profondità di sequenziamento**. Si tratta del numero di volte in cui ogni base di una sequenza nucleotidica viene letta. Maggiore è il numero di volte in cui una certa base viene letta in una posizione, maggiore è la probabilità che quella base sia effettivamente presente in quel determinato posto. Una maggior profondità si traduce in *library size* maggiori e dunque una lettura più dettagliata del materiale genetico introdotto nel sequenziatore. Proprietà che si rivela fondamentale nell'individuare le specie batteriche più rare visto il basso numero di sequenze appartenenti a queste ultime. Esiste infatti, almeno a livello teorico, una forte correlazione tra il numero di OTU identificate in un campione e la corrispondente profondità di sequenziamento. Tutto questo suggerisce che le misure di abbondanza differenziale siano soggette a distorsioni poiché le unità tassonomiche con zero conte, appartenenti a campioni con basse *library size* sono interpretate come assenti quando in realtà non sono altro che il risultato di una profondità di lettura non sufficiente. Questo modello nasce proprio dalla relazione osservata tra profondità di sequenziamento e numero di OTU identificate. Un esempio tratto dai dati dell'*American Gut Project* e del Progetto Microbioma Italiano è visibile in Figura 2.2. Si noti come la relazione tra le due quantità coinvolte sia, nonostante la dispersione, piuttosto lineare.
Supponendo di avere campioni provenienti da due condizioni, si ipotizza che il dato di conta sia modellato da due popolazioni, ognuna con M_A ed M_B campioni con N unità tassonomiche. Le conte grezze per il campione j e l'unità tassonomica i, indicate con c_{ij} sono trasformate tramite il logaritmo

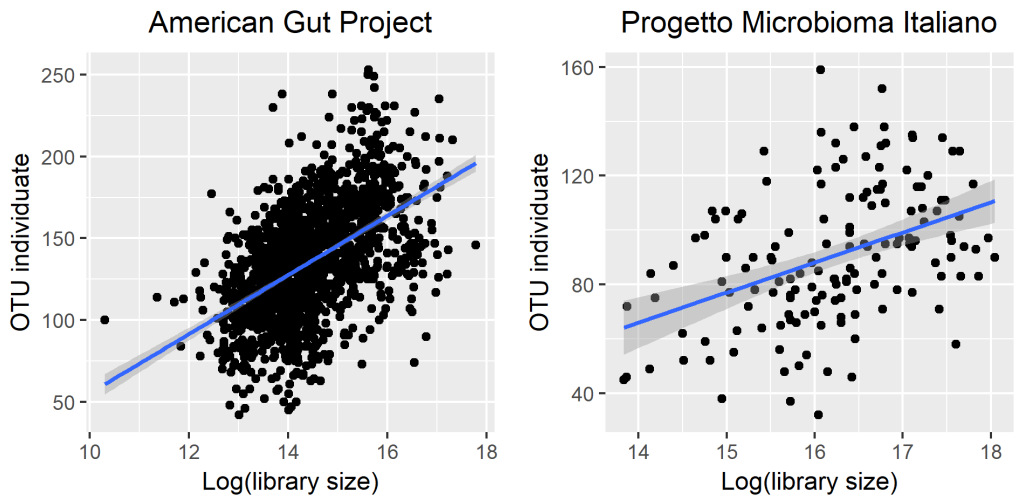

Figura 2.2: Grafico del numero di OTU individuate per *library size* in scala logaritmica in un insieme di repliche biologiche dell'*American Gut Project* (a sinistra) e del Progetto Microbioma Italiano (a destra).

in base 2 applicando la correzione di continuità:

$$y_{ij} = log_2(c_{ij} + 1)$$

Come nella formulazione della Binomiale Negativa, k indica l'appartenenza di un campione ad un determinato gruppo, in particolare $k(j) = I\{j \in gruppoA\}$. Basandosi sull'osservazione del fatto che il numero di OTU con zero sequenze in un campione dipende dalla rispettiva *library size* L_j, gli autori del modello hanno deciso di modellare i parametri di mistura $\pi_j(L_j)$ come un processo Binomiale (Paulson et al. 2013):

$$log\left(\frac{\pi_j}{1-\pi_j}\right) = \beta_0 + \beta_1 \cdot log(L_j)$$

Una volta definiti i parametri di mistura π_j e la distribuzione delle conte come $f_{conte}(y; \mu, \sigma^2) \sim N(\mu, \sigma^2)$, la funzione di densità per l'unità tassonomica i, del campione j con *library-size* L_j è definita come:

$$f_{zig}(y_{ij}; L_j, \beta, \mu_i, \sigma_i^2) = \pi_j(L_j) \cdot I_{\{0\}}(y_{ij}) + (1 - \pi_j(L_j)) \cdot f_{conte}(y_{ij}; \mu_i, \sigma_i^2)$$

data l'appartenenza al gruppo k per il campione j, la media è definita come:

$$E(Y_{ij}|k(j)) = \pi_j \cdot 0 + (1-\pi_j) \cdot \left(b_{0i} + \eta_i log_2 \left(\frac{s_j^{\hat{l}} + 1}{K} \right) + b_{1i} k(j) \right)$$

In cui b_{0i} rappresenta l'offset relativo all'i-esima unità tassonomica, b_{1i} rappresenta l'incremento relativo nella media delle conte normalizzate dovuto al gruppo di appartenenza. Con il fattore di normalizzazione trasformato $log_2((s_j^{\hat{l}}+1)/K)$, vengono catturati i fattori di normalizzazione OTU-specifici attraverso il parametro η_i. Si noti che $s_j^{\hat{l}}$, derivante dalla normalizzazione *Cumulative Sum Scaling* (sviluppata dagli stessi autori di questo modello distributivo), è definita come la somma cumulata delle conte per il campione j fino al quantile $\hat{l}$-esimo. Tale quantità viene riscalata per F, costante moltiplicativa opportunamente scelta uguale in tutti i campioni: solitamente si tratta della mediana dei fattori di scala $s_j^{\hat{l}}$.
Il set di parametri da stimare risultante è $\theta_{ij} = \{\beta_0, \beta_1, b_{0i}, \eta_i, b_{1i}\}$. Le stime di massima verosimiglianza vengono approssimate usando un algoritmo *EM*. Si ipotizza l'esistenza di una variabile indicatrice latente Δ_{ij} che regola il parametro di mistura:

$$\Delta_{ij} = \begin{cases} 1 & \text{se } y_{ij} = 0 \\ 0 & \text{altrimenti} \end{cases}$$

La log-verosimiglianza del modello con i dati aumentati risulta:

$$\begin{aligned} l(\theta_{ij}, \Delta_{ij}; y_{ij}, L_j) = {} & (1-\Delta_{ij}) log f_{conte}(y_{ij}; \mu_i, \sigma_i^2) \\ & + (1-\Delta_{ij}) log(1-\pi_j(L_j)) + \Delta_{ij} log \pi_j(L_j) \end{aligned}$$

Passo E: stima delle probabilità $z_{ij} = P(\Delta_{ij} = 1)$ date le stime $\hat{\theta}_{ij}$ al passo precedente.

$$\hat{z}_{ij} = \frac{\hat{\pi}_j \cdot I_{\{0\}}(y_{ij})}{\hat{\pi}_j \cdot I_{\{0\}}(y_{ij}) + (1-\hat{\pi}_j) f_{conte}(y_{ij}; \hat{\theta}_{ij})}$$

Notando che per costruzione, $\hat{z}_{ij} = 0 \; \forall \; y_{ij} > 0$.
Passo M: stima dei parametri θ_{ij} date le stime ottenute al passo E. Il parametro di mistura viene stimato utilizzando la stima della variabile latente:

$$\hat{\pi}_j = \sum_{i=1}^{N} \frac{\hat{z}_{ij}}{N}$$

ricordando che N è il numero totale di OTU. Tramite i minimi quadrati vengono poi stimati i parametri β_0 e β_1 del modello logit:

$$log\frac{\hat{\pi}_j}{1-\hat{\pi}_j} = \beta_0 + \beta_1 log(L_j)$$

Infine per calcolare b_{0i},η_i e b_{1i} vengono utilizzati i minimi quadrati pesati, con pesi $1-\hat{z}_{ij}$, notando che solo i campioni con $y_{ij}=0$ hanno pesi minori di 1. Anche le stime delle deviazioni standard sono ottenute usando $1-\hat{z}_{ij}$ come pesi. Per semplificare quanto detto, si immagini di considerare la prima unità tassonomica della tabella delle OTU, il modello di regressione lineare in forma matriciale sarà del tipo:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{M-1} \\ y_M \end{bmatrix} = \begin{bmatrix} 1 & log_2\left(\frac{s_1^{\hat{l}}+1}{F}\right) & 0 \\ 1 & log_2\left(\frac{s_2^{\hat{l}}+1}{F}\right) & 0 \\ \vdots & \vdots & \vdots \\ 1 & log_2\left(\frac{s_{M-1}^{\hat{l}}+1}{F}\right) & 1 \\ 1 & log_2\left(\frac{s_M^{\hat{l}}+1}{F}\right) & 1 \end{bmatrix} \cdot \begin{bmatrix} b_0 \\ \eta \\ b_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_{M-1} \\ \varepsilon_M \end{bmatrix}$$

che ha come forma compatta:

$$Y = X \cdot b + \varepsilon$$

In cui M è il numero totale di campioni ed ε rappresenta il vettore degli errori che segue una distribuzione $N_M(0, \sigma^2 W)$. La matrice $W = diag(1-z_1, 1-z_2, \ldots, 1-z_{M-1}, 1-z_M)$ serve per pesare le osservazioni in base al valore di conteggio osservato: i campioni con valore positivo di conteggio della prima OTU avranno peso pari a 1, mentre i campioni con zero sequenze lette per la prima OTU avranno peso minore di 1. Quindi se per tutti i campioni la prima OTU ha valori positivi, la matrice dei pesi W sarà la matrice identità ed il modello sarà stimato con il metodo dei minimi quadrati classici. Per stimare i parametri del modello si utilizza il metodo dei minimi quadrati pesati, lo stimatore sarà:

$$\hat{b} = (X^T W X)^{-1} X^T W Y$$

Lo stimatore è non distorto e con matrice di covarianza:

$$Var(\hat{b}) = \sigma^2 (X^T W X)^{-1}$$

Lo stimatore non distorto per σ^2 diventa:

$$\hat{\sigma}^2 = \frac{(Y - X\hat{b})^T W (Y - X\hat{b})}{M - 3}$$

In cui $M - 3 = d$ rappresentano i gradi di libertà per il modello stimato.
L'obiettivo dell'analisi inferenziale è quello di individuare quali unità tassonomiche sono differenzialmente abbondanti nelle diverse condizioni sperimentali determinate dall'appartenenza al gruppo $k(j)$. L'ipotesi nulla per rispondere a tale domanda è $H_0 : b_{1i} = 0$ ossia la mancanza di un incremento relativo nella media delle conte normalizzate nei due gruppi. Dalla stima di b_{1i} e del suo standard error, viene costruita una statistica t moderata attraverso il metodo inferenziale Bayesiano empirico, un'approssimazione ad un approccio completamente bayesiano. La particolarità di questo metodo è che, contrariamente all'approccio bayesiano classico in cui la a priori viene fissata ancor prima di osservare i dati, i parametri della *a priori* vengono stimati dai dati stessi.
Vista la necessità di rispondere alla domanda di ricerca per ogni unità tassonomica, si definisce un semplice modello gerarchico in grado di descrivere questa struttura parallela (Smyth 2004). L'informazione *a priori* assunta su σ_i^2 è definita equivalente ad uno stimatore a priori s_0^2 con d_0 gradi di libertà:

$$\frac{1}{\sigma_i^2} \sim \frac{1}{d_0 s_0^2} \chi^2_{d_0}$$

che descrive come la varianza dovrebbe variare tra unità tassonomiche. Per quanto riguarda il coefficiente di interesse b_{1i}, si assume che sia diverso da zero con probabilità:

$$P(b_{1i} \neq 0) = p$$

che rappresenta la proporzione di OTU differenzialmente abbondanti. Per quelle unità tassonomiche in cui il coefficiente è diverso da 0, l'informazione a priori segue una legge distributiva Normale con varianza scalata in base ad un ignoto fattore v_0:

$$b_{1i} | \sigma_i^2, b_{1i} \neq 0 \sim N(0, \sigma_i^2 v_0)$$

La statistica test per il coefficiente b_{1i} creata a partire da questa gerarchia, viene detta moderata poiché la varianza stimata dal modello di regressione

tramite minimi quadrati pesati $\hat{\sigma}_i^2$ viene sostituita con la varianza a posteriori $\tilde{\sigma}_i^2$:

$$\tilde{t}_i = \frac{\hat{b}_{1i}}{\sqrt{\tilde{\sigma}_i^2 V_{i(3,3)}}}$$

In cui V_i è la matrice $(X^T W X)^{-1}$ che moltiplicata per la varianza fornisce la matrice di covarianza per i coefficienti del modello. L'elemento di posizione $(3,3)$ corrisponde al coefficiente di interesse, le posizioni $(1,1)$ e $(2,2)$ invece sono relative rispettivamente ai coefficienti b_{i0} e η_i. La varianza a posteriori, è calcolata come:

$$E(\sigma_i^2 | \hat{\sigma}_i^2) = \tilde{\sigma}_i^2 = \frac{d_0 s_0^2 + d_i \hat{\sigma}_i^2}{d_0 + d_i}$$

$\hat{\sigma}_i^2$ e d_i sono rispettivamente la varianza osservata per l'i-esima OTU e i gradi di libertà ottenuti tramite il modello di regressione, mentre s_0^2 e d_0 sono ottenuti equiparando valori empirici e valori teorici per i primi due momenti di $log_2\hat{\sigma_i^2}$. Dal momento che la distribuzione dei logaritmi delle varianze è più simile ad una distribuzione Normale rispetto a quella delle varianze non trasformate, la stima tramite metodo dei momenti è più efficiente, garantendo momenti finiti per ogni grado di libertà. La statistica $\tilde{t}$ moderata torna ad essere la statistica ordinaria t se $d_0 = 0$, se invece $t_0 = \infty$ diventa proporzionale al coefficiente $\hat{b}_{1i}$. La statistica moderata, sotto l'ipotesi nulla $H_0 : b_1 i = 0$, segue una distribuzione t di Student con $d_0 + d_i$ gradi di libertà. I d_0 gradi di libertà aggiuntivi rispetto a quelli presenti nella distribuzione t ordinaria riflettono l'informazione derivante dal modello gerarchico assunto, in cui viene utilizzato l'insieme di tutte le unità tassonomiche per l'inferenza sulle singole OTU.

2.3.3 Correzione per test multipli

Il problema dei confronti multipli è particolarmente importante nello studio dell'abbondanza. Il livello di significatività di un test statistico corrisponde, in questo ambito, alla probabilità di sbagliare nel respingere l'ipotesi di uguale abbondanza di una certa unità tassonomica nei gruppi sperimentali confrontati; convenzionalmente il livello di significatività viene fissato al 5%. Quindi, se le unità tassonomiche sono ugualmente abbondanti, si corre un

rischio pari al 5% di dichiararle erroneamente differenzialmente abbondanti. Ciò vuol dire che in 20 casi in cui l'ipotesi nulla viene respinta al livello del 5%, ci si aspetta mediamente un falso positivo. Supponendo di avere una matrice delle OTU con 100 unità tassonomiche e l'indipendenza dei test, la probabilità di rifiutare H_0 quando è vera equivale a $1 - 0.95^{100} = 0.994$. Quindi la probabilità di sbagliare almeno una volta è prossima alla certezza, ben lontana da una probabilità di 0.05. Nella realtà possono esistere delle relazioni tra unità tassonomiche, le quali rendono i confronti dipendenti, ma in ogni caso il problema permane.
Esistono diversi metodi per fare in modo di ridurre l'entità dell'errore, chiamati metodi di correzione del *p-value*. La maggior parte di questi si basa proprio sul controllo della quantità chiamata *Family Wise Error Rate* che corrisponde alla probabilità di individuare almeno un falso positivo. Considerando la Tabella 2.7 questa quantità equivale a $FWER = P(V \geq 1)$ Questi

Tabella 2.7: Esito di test multipli di abbondanza differenziale.

	Test		
Realtà	dichiarate non significative	dichiarate significative	**Totale**
H_0	U	V	m_0
H_1	T	S	$m - m_0$
Totale	$m - R$	R	m

metodi di correzione sono ad esempio il metodo di Bonferroni, la correzione di Holm e Holm-Sidak; metodi piuttosto conservativi quando il numero di confronti risulta elevato.
Un altro modo di procedere è correggere usando il *False Discovery Rate* (FDR) cioè la frazione attesa di test significativi che sono veri negativi. La generalizzazione del FDR è il *q-value* che misura la significatività sulla base del FDR. Ad esempio, un *q-value*≤ 0.05 porta ad avere una lista di test significativi con FDR pari al 5% cioè in media, tra tutti i test significativi il 5% sono dei veri negativi. È opportuno ribadire che invece, un *cut-off* basato sul *p-value* dice poco sulle caratteristiche dei test risultati significativi, infatti un *p-value*≤ 0.05 porta ad un FPR del 5%, il quale sta a significare che in media il 5% dei veri negativi sarà identificato come significativo.

Per il calcolo del FDR:

$$V(t) = \#\{p_i \leq t | H_0, i = 1, \ldots, m\}$$

$$R(t) = \#\{p_i \leq t | i = 1, \ldots, m\}$$

$$FDR(t) = \frac{V(t)}{R(t)}$$

dove t è la soglia definita per il *p-value*. Per la stima di $V(t)$ si sfrutta la caratteristica distribuzione uniforme dei *p-value* nei veri negativi: $P(p \leq t|H_0) = t$ da cui segue che $E(V_t) = m_0 t$. Tuttavia m_0 è ignoto e va stimato tramite $\pi_0 = m_0/m$. L'algoritmo di stima del FDR proposto da Benjamini e Hochberg pone $\hat{\pi}_0 = 1$ (Benjamini e Hochberg 1995), da cui risulta:

$$\hat{FDR}(t) = \frac{\hat{\pi}_0 \cdot m \cdot t}{R(t)} = \frac{m \cdot t}{R(t)}$$

Il *q-value* è il minimo FDR che si ottiene quando si definisce significativo un test:

$$\hat{q}(p_i) = \underset{t \geq p_i}{min} \hat{FDR}(t)$$

Capitolo 3

Casi Studio

In questo capitolo vengono applicate le tematiche affrontate in precedenza su dati reali. Gli insiemi di dati coinvolti nell'analisi provengono dall' *American Gut Project* e dal Progetto Microbioma Italiano. Per quanto riguarda il progetto americano, si hanno a disposizione una grande quantità di dati:

- più di 3600 partecipanti;
- diversi campioni per individuo relativi a rilevazioni in distinte parti del corpo;
- un insieme di meta-dati approfondito contenente molte variabili demografiche tra cui le abitudini alimentari giornaliere degli individui.

Questi dati sono pubblici e reperibili al sito del progetto *americangut.org*. I dati del Progetto Microbioma Italiano invece, sono utilizzabili grazie ad un percorso di stage eseguito da Maggio ad Ottobre 2017 nell'azienda BMR Genomics: responsabile italiana del progetto per la ricezione e il sequenziamento dei campioni. L'azienda ha raccolto dal 2014 (inizio del progetto) fino al momento dello stage, circa 200 campioni. Ogni campione è accompagnato da un insieme di meta-dati derivante da un questionario che ogni individuo compila obbligatoriamente prima di spedire il campione all'azienda. L'analisi dei dati nasce da un'esigenza dell'azienda di rispondere ad alcune domande biologiche, come ad esempio, capire in che modo è possibile identificare associazioni, se presenti, tra una condizione patologica, un'abitudine o uno

stile di vita e il relativo microbiota intestinale degli individui. Domande che possono avere una risposta solo attraverso specifiche analisi statistiche. La strada percorribile per rispondere a queste domanda può essere quella di utilizzare l'eredità del Next Generation Sequencing e dei relativi strumenti di analisi di dati RNA-Seq vista la similarità con il dato di microbioma. Lo scopo del capitolo è dunque quello di testare gli strumenti di normalizzazione e inferenza, prima nell'insieme di dati più grande in cui si conoscono già alcuni risultati, e in seguito nei dati del Progetto Microbioma Italiano in modo da evidenziare similarità e differenze rispetto al Microbioma Americano.
L'analisi si svolgerà in modo sequenziale per i due insiemi di dati e sarà costituita delle seguenti parti:

1. rimozione dei campioni di bassa qualità;
2. scelta di una variabile su cui studiare l'abbondanza differenziale;
3. normalizzazioni dei dati: *Cumulative Sum Scaling* (CSS), *Trimmed Mean of M-values* (TMM) e *Relative Log Expression* (RLE)
4. inferenza tramite modello per dati di conteggio con distribuzione mistura Normale *Zero-Inflated* (ZIG) per i dati normalizzati tramite CSS;
5. inferenza tramite modello per dati di conteggio con distribuzione Binomiale Negativa (NB) per i dati normalizzati tramite TMM e RLE.
6. confronto dei risultati per le tre diverse normalizzazioni.

Tutte le analisi sono state eseguite utilizzando il software R, in particolare i pacchetti *MetagenomeSeq* nato appositamente per le analisi del microbioma, ed *edgeR* contenente strumenti per l'analisi di dati RNA-Seq.

3.1 Microbioma Americano

Una volta scaricati i campioni e prodotta la relativa tabella delle OTU, sono stati selezionati per l'analisi solo quelli relativi all'intestino degli individui, tralasciando i campioni provenienti da pelle, saliva e altre mucose. A

partire da questi campioni è stato eseguito un filtraggio di qualità per eliminare dall'analisi quelli in cui i dati erano perlopiù assenti. Il valore soglia per stabilire se un campione merita di rientrare nell'analisi consigliato dalla libreria *MetagenomeSeq* è pari a 1000: tutti i campioni con meno di 1000 sequenze totali (*library size*) vengono eliminati. A questo punto è opportuno scegliere una variabile su cui eseguire le analisi. Dal momento che le relazioni trovate dai primi risultati dell'*American Gut Project* sono relative all'alimentazione e all'utilizzo di antibiotici si è scelto di confrontare il microbioma di individui onnivori e individui vegetariani che non hanno assunto antibiotici nell'ultimo anno. L'insieme di campioni così definito è composto da 39 vegetariani e 856 onnivori con caratteristiche di età, sesso e utilizzo degli antibiotici simili. Il numero di onnivori è circa 22 volte il numero di vegetariani; è opportuno eseguire le analisi successive su un insieme bilanciato per evitare di ottenere delle stime distorte. Per garantire robustezza al metodo inferenziale, data la necessità di bilanciamento, si decide di produrre 500 sottocampionamenti degli onnivori composti da 39 individui casualmente estratti dall'insieme degli 856 onnivori totali. Uno schema che ripercorre in modo semplificato i passaggi da compiere per l'analisi dei dati del microbioma americano è presente in Figura 3.1. Si formano in questo modo 500 ricampionamenti composti dai soliti 39 campioni di vegetariani e dai 39 campioni di onnivori casualmente estratti. Il campionamento casuale avviene solo per il gruppo degli onnivori poiché il gruppo dei vegetariani è già molto ridotto e per non perdere informazione preziosa si preferisce mantenerlo costante in ognuno dei 500 ricampionamenti. Per ogni campione, un secondo filtraggio, stavolta per OTU, è proposto per garantire la presenza di una certa OTU in un numero minimo di campioni. Si è scelto di considerare solo le unità tassonomiche presenti in almeno il 30% campioni. Si noti che avendo a che fare con due gruppi distinti di campioni, è possibile che oltre all'abbondanza differenziale ci sia anche una presenza esclusiva di alcune unità tassonomiche in un gruppo piuttosto che nell'altro. Per questo motivo, il procedimento di filtraggio per OTU viene eseguito nel gruppo degli onnivori e nel gruppo dei vegetariani separatamente. L'esclusiva appartenenza di una OTU ad un gruppo può essere del tutto casuale oppure realmente associata al tipo di dieta. Per ridurre la componente casuale viene studiata, per ognuno dei 500 ricampionamenti, la lista delle unità tassonomi-

che esclusive nel gruppo degli onnivori e nel gruppo dei vegetariani. Una volta ottenute tutte le liste viene stilata la classifica delle 10 OTU più esclusive per ciascuna dieta (Tabella 3.5 e Tabella 3.6). Il genere o la famiglia a cui queste OTU fanno riferimento è accompagnato dalla proporzione tra il numero di campioni in cui la relativa OTU è presente sul totale dei ricampionamenti coinvolti, cioè 500. L'analisi di abbondanza differenziale si esegue solamente per le unità tassonomiche condivise tra i due gruppi, la tabella delle OTU risultante sarà diversa per ognuno dei 500 ricampionamenti ma mediamente composta da 160 unità tassonomiche.

3.1.1 Normalizzazione

Il passo successivo consiste nella normalizzazione. Le normalizzazioni da applicare alla tabella delle OTU sono tre: la CSS si pone come obiettivo quello di rendere all'incirca equivalente e indipendente la distribuzione delle conte tra campioni fino al quantile scelto dall'algoritmo di normalizzazione, sotto l'assunzione che il dato provenga da una distribuzione comune, mentre invece, le normalizzazioni TMM e RLE calcolano dei fattori di scala per i campioni attraverso un campione opportunamente calcolato che diventa di riferimento. Le tre normalizzazioni operano in modo diverso l'una dall'altra ma hanno tutte come scopo quello di rendere più confrontabili i campioni biologici. Le normalizzazioni vengono applicate a partire dalla matrice delle conte grezze ma valutate una volta trasformate tramite il logaritmo in base due per rendere la scala dei valori di conteggio più limitata. In particolare, l'elemento di posizione i, j della matrice delle OTU normalizzata diventa: $log_2(c_{ij}+1)$ in cui il valore 1 sommato ad ogni elemento di conteggio serve per fare in modo che l'argomento del logaritmo non sia 0. Per misurare l'efficacia delle normalizzazioni è necessario pensare ad una statistica in grado di misurare la confrontabilità tra campioni prima e dopo ogni normalizzazione. Le statistiche più intuitive sono quelle basate sulla mediana o sullo scarto interquartile del logaritmo delle conte per ogni esperimento: una minor variabilità nella distribuzione campionaria di queste statistiche dovrebbe tradursi in una maggior confrontabilità dei campioni coinvolti. A livello pratico si procede in modo sequenziale:

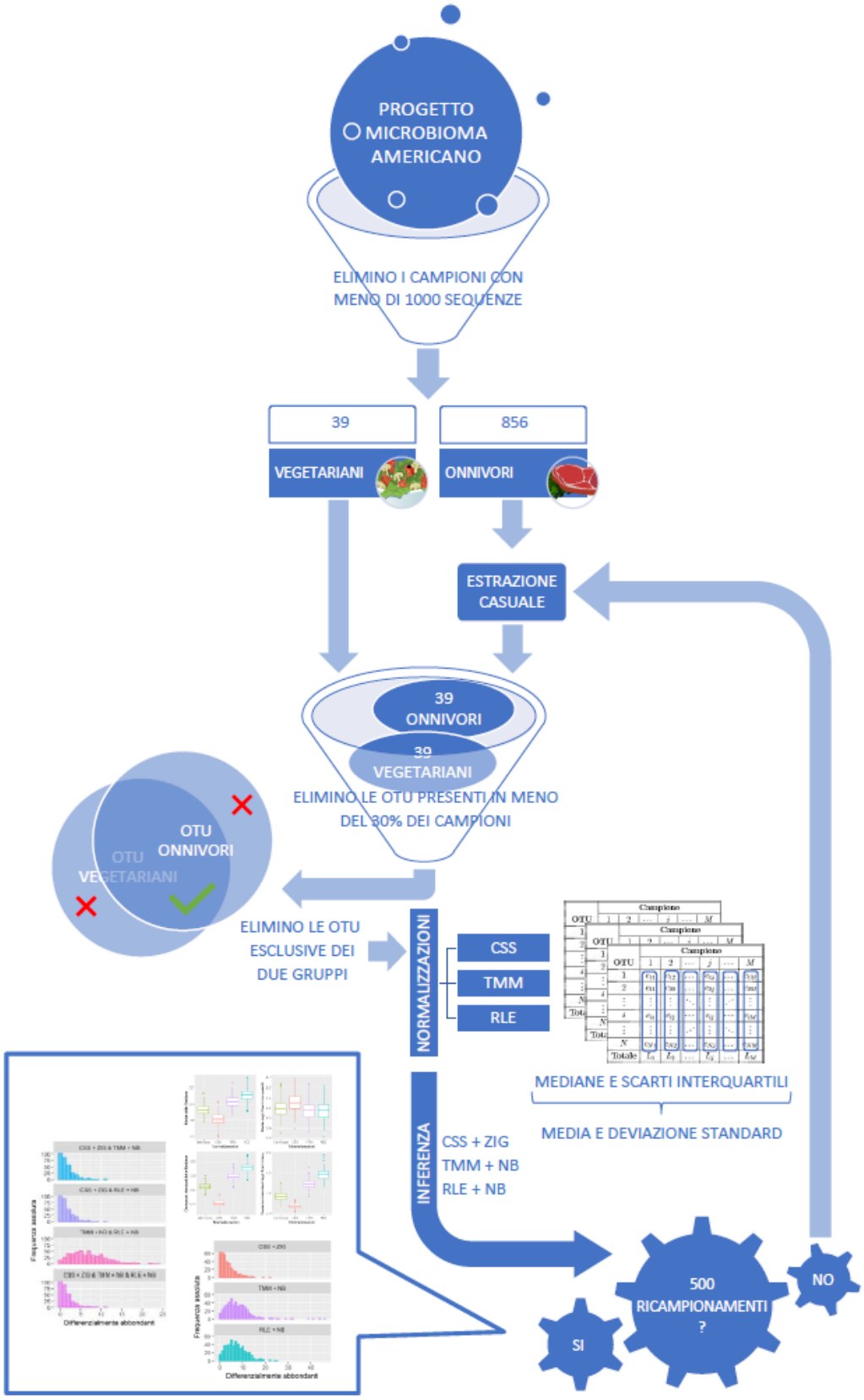

Figura 3.1: Sintesi schematica dell'analisi per i dati *American Gut Project*.

1. considero quattro matrici: dati grezzi, dati normalizzati tramite CSS, TMM e RLE;

2. calcolo mediana e scarto interquartile per ogni campione di ogni matrice;

3. calcolo media e la deviazione standard delle mediane e degli scarti interquartili del punto 2 per ogni matrice;

4. ripeto dal punto 1 fino ad avere 500 ricampionamenti.

Il risultato è visibile in Figura 3.2 per quanto riguarda le medie delle mediane e degli scarti interquartili nei 500 ricampionamenti. Tutte le normalizzazioni, rispetto ai dati grezzi, sono caratterizzate da valori mediamente più bassi ($p\text{-}value < 0.01$ con un test t di Student per dati appaiati) per le medie delle mediane e valori mediamente più alti ($p\text{-}value < 0.01$) per le medie degli scarti interquartili. Osservando la parte bassa della figura Figura 3.2 si nota come le deviazioni standard siano mediamente maggiori per i dati normalizzati tramite TMM o RLE rispetto ai dati grezzi, sia per quanto riguarda le mediane che gli scarti interquartili. La normalizzazione CSS porta a valori della media delle mediane, delle deviazioni standard delle mediane e delle deviazioni standard degli scarti interquartili nettamente diversi rispetto a quelli ottenuti con le altre normalizzazioni.
Questo è un risultato inaspettato per dei dati normalizzati, in quanto un aumento di variabilità nelle distribuzioni esaminate, difficilmente porta ad una maggiore confrontabilità campionaria. Una spiegazione plausibile a questo fenomeno può essere la diversità nel tipo di dato da normalizzare. In ambito RNA-Seq, in cui le normalizzazioni TMM e RLE sono nate, ci sono migliaia di geni per ciascun campione e i valori di conteggio hanno un range di variazione solitamente molto ampio, ma senza l'elevato numero di zeri che caratterizza i dati di microbioma. In questo particolare caso, si ha a che fare con un numero medio di OTU per ricampionamento pari a 160. Un modo per verificare la plausibilità di tale affermazione è quello di variare la soglia di filtraggio attraverso cui un'OTU viene considerata per le analisi (si era scelto di considerare solo le unità tassonomiche presenti in almeno il 30% campioni, cioè 12 su 39). Vengono esplorate due nuove soglie di filtraggio: una soglia più

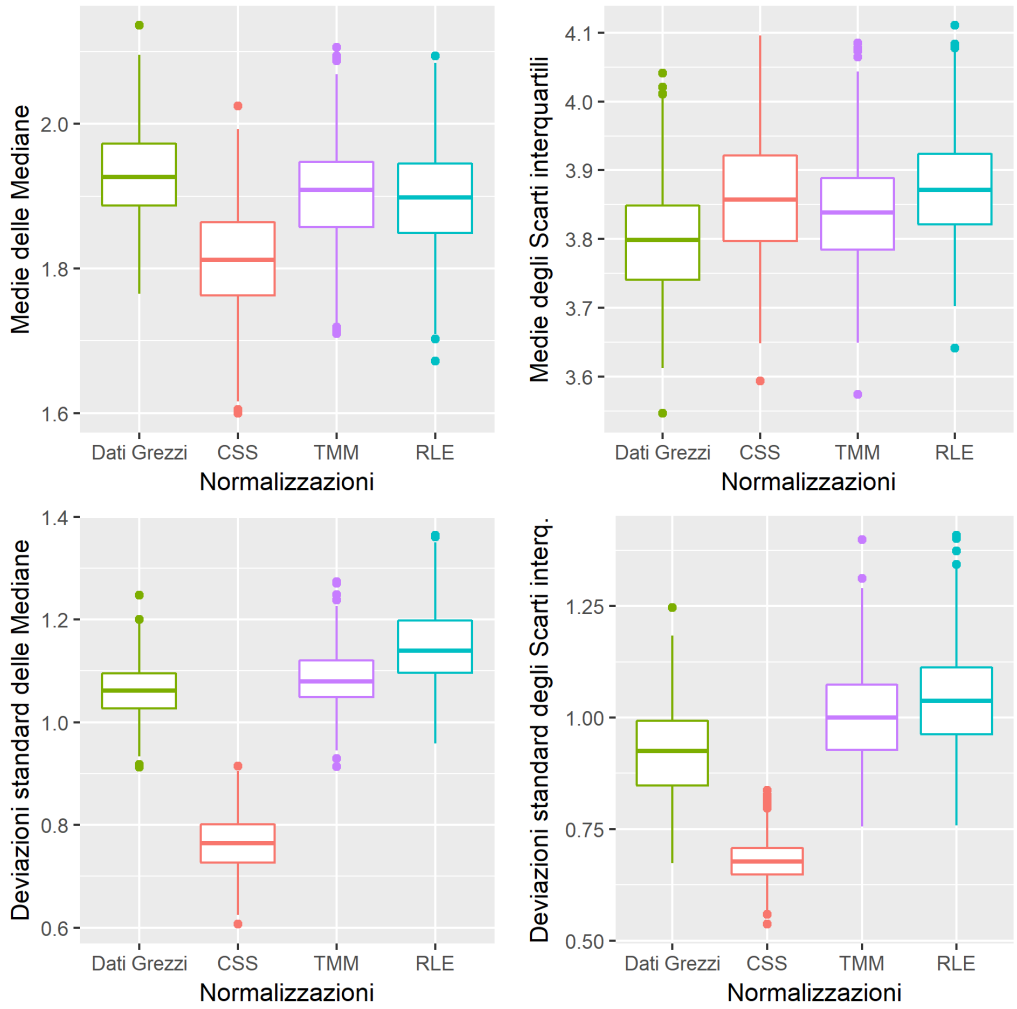

Figura 3.2: Diagrammi a scatola con baffi per le distribuzioni delle medie (in alto) e delle deviazioni standard (in basso) delle mediane campionarie (a sinistra) e degli scarti interquartili campionari (a destra), calcolati per i dati grezzi e le normalizzazioni *Cumulative Sum Scaling* (CSS), *Trimmed Mean of M-values* (TMM) e *Relative Log Expression* (RLE) nei 500 ricampionamenti. Scala di misura log_2.

permissiva in cui un'OTU deve essere presente in almeno il 10% dei campioni, cioè 4 su 39; e una soglia più rigida, in cui un'OTU deve essere presente in almeno il 70% dei campioni, cioè 28 su 39. Per la soglia permissiva, che prevede un aumento del numero medio di OTU per ricampionamento a 235, la media delle mediane e la media degli scarti interquartili con le relative deviazioni standard per i 500 ricampionamenti, nei dati grezzi e nelle diverse normalizzazioni, è visibile in Figura 3.3. Per la soglia più rigida, che porta ad una diminuzione del numero medio di OTU per ricampionamento a 65 (meno rispetto al numero di campioni), le medesime statistiche sono nella Figura 3.4.
Riassumendo:

- la normalizzazione che rende maggiormente confrontabili i campioni sembra essere la CSS grazie alle deviazioni standard minori per entrambe le statistiche;

- la distribuzione delle deviazioni standard delle mediane nelle normalizzazioni TMM e RLE è molto influenzata dalla sparsità della matrice. Nei ricampionamenti con tante OTU e quindi matrici più sparse le deviazioni standard delle mediane risultano mediamente maggiori a quelle nei dati grezzi, anche se di poco. Nei ricampionamenti con poche OTU in cui la matrici contengono pochi zeri, si osservano deviazioni standard delle mediane più basse rispetto a quelle osservate nei dati grezzi.

- il numero di OTU coinvolte ma specialmente la conseguente sparsità della matrice sembrano essere fattori che influenzano il buon funzionamento delle normalizzazioni TMM e RLE.

3.1.2 Inferenza

Una volta ottenute le tabelle delle OTU normalizzate secondo le tre normalizzazioni descritte vengono stimati i modelli per ogni singola unità tassonomica. Seguendo gli approcci standard come sono stati proposti e poi implementati in letteratura, alla normalizzazione CSS segue un modello ZIG, mentre alle normalizzazioni TMM e RLE segue un modello NB. Lo scopo di

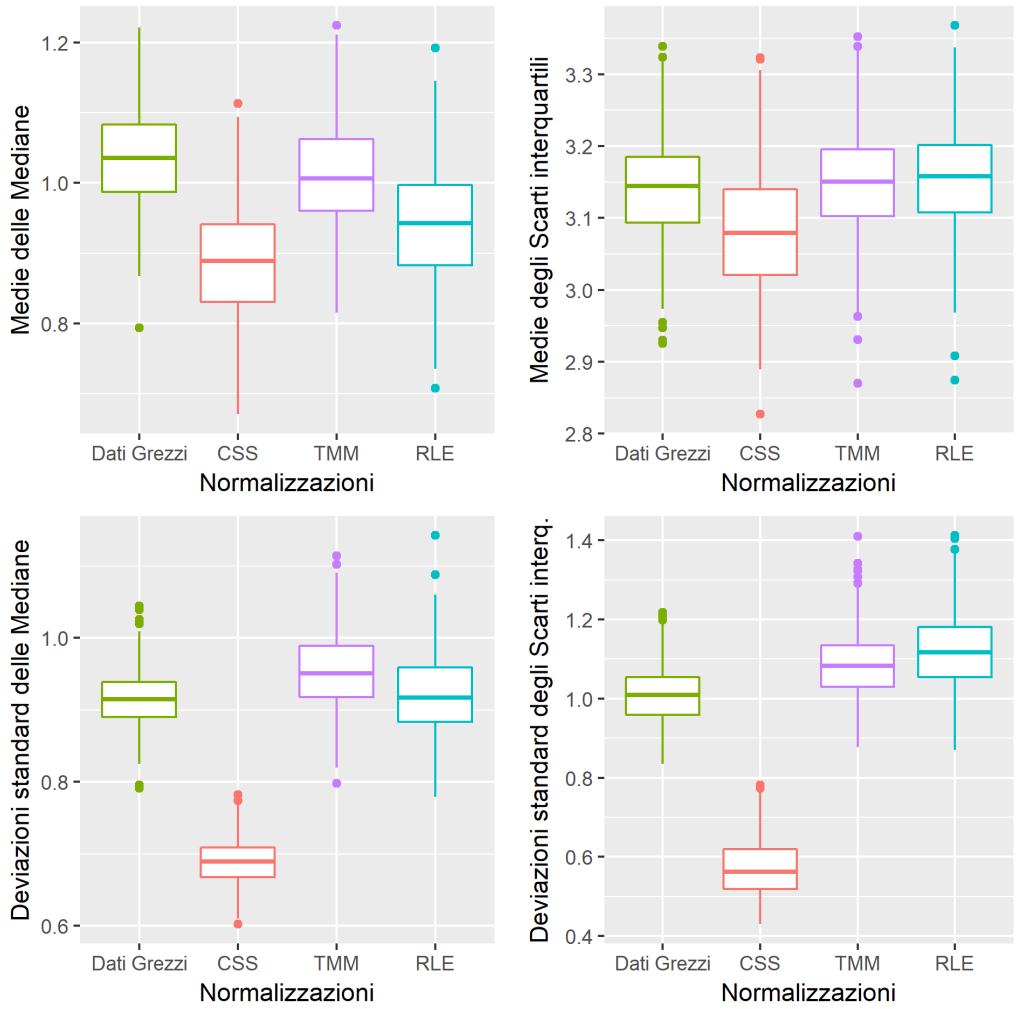

Figura 3.3: Considerando le OTU presenti in almeno il 10% dei campioni. Diagrammi a scatola con baffi per le distribuzioni delle medie (in alto) e delle deviazioni standard (in basso) delle mediane campionarie (a sinistra) e degli scarti interquartili campionari (a destra), calcolati per i dati grezzi e le normalizzazioni *Cumulative Sum Scaling* (CSS), *Trimmed Mean of M-values* (TMM) e *Relative Log Expression* (RLE) nei 500 ricampionamenti. Scala di misura log_2.

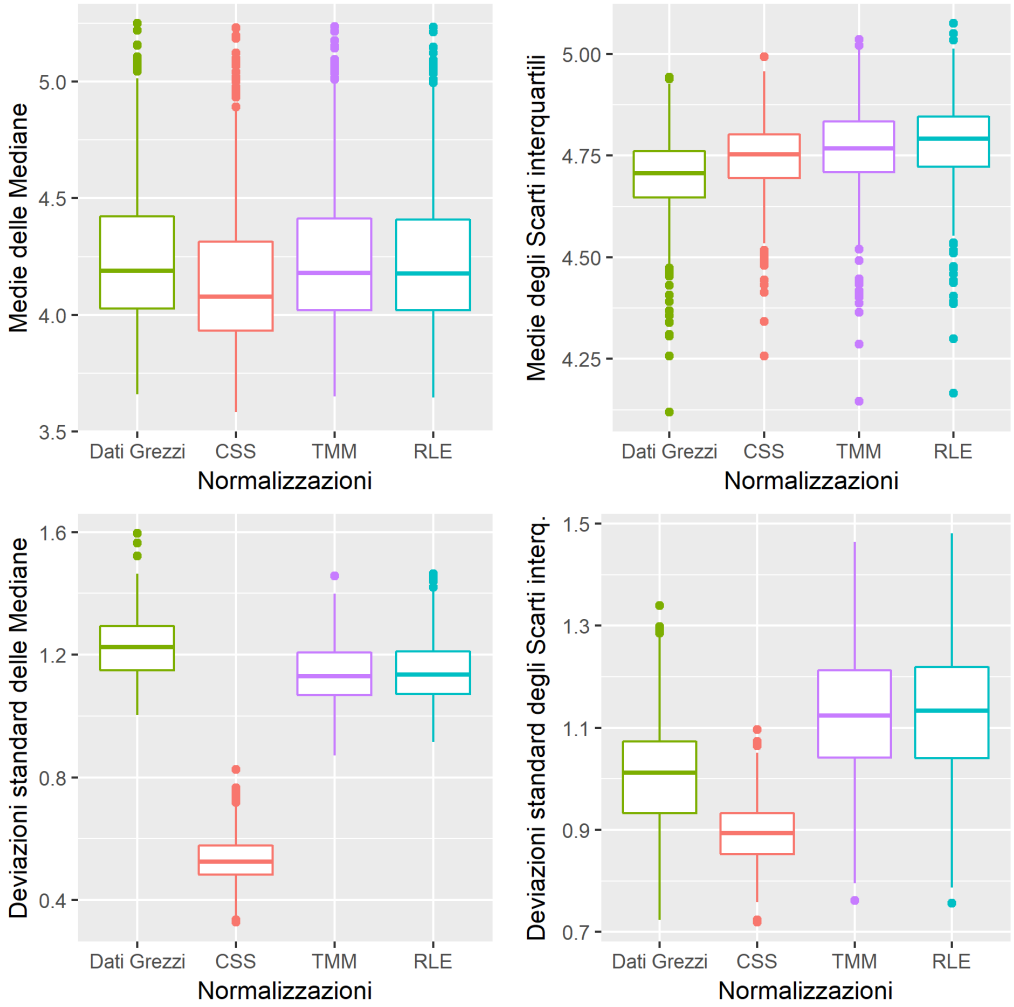

Figura 3.4: Considerando le OTU presenti in almeno il 70% dei campioni. Diagrammi a scatola con baffi per le distribuzioni delle medie (in alto) e delle deviazioni standard (in basso) delle mediane campionarie (a sinistra) e degli scarti interquartili campionari (a destra), calcolati per i dati grezzi e le normalizzazioni *Cumulative Sum Scaling* (CSS), *Trimmed Mean of M-values* (TMM) e *Relative Log Expression* (RLE) nei 500 ricampionamenti. Scala di misura log_2.

questa fase è quello di individuare quali siano le unità tassonomiche differenzialmente abbondanti nei due gruppi di individui ma allo stesso tempo fare luce sul processo inferenziale mantenendo una continuità tra normalizzazione utilizzata e la distribuzione ipotizzata per il dato di conteggio. Ci si aspetta una coerenza nei risultati dei tre percorsi ma allo stesso tempo, conoscendo le loro caratteristiche, non c'è la certezza che tutte le unità tassonomiche individuate come differenzialmente abbondanti da un metodo, vengano individuate da un altro. Vengono riportate per ognuno dei 500 ricampionamenti, tre tabelle contenenti le unità tassonomiche differenzialmente abbondanti per ciascun metodo inferenziale. Per studiare la differenziale abbondanza, si considerano in ciascun percorso inferenziale, i logaritmi dei *Fold-Change*, i *p-value* e i *p-value* aggiustati per *False Discorvery Rate*. In particolare un'unità tassonomica viene definita differenzialmente abbondante se il relativo valore del *p-value* aggiustato risulta minore di 0.1. Visto la lunghezza della nomenclatura per descrivere un'unità tassonomica, queste ultime vengono identificate solo con il genere e/o la famiglia tassonomica; in base al livello di profondità per il quale è stato possibile classificare tassonomicamente quello specifico gruppo di sequenze.
In Tabella 3.1 sono elencate le unità tassonomiche che risultano differenzialmente abbondanti ordinate in base al numero di volte in cui compaiono nei 500 ricampionamenti. Questa prima lista grezza può essere un punto di partenza per un'interpretazione di tipo biologico che non viene affrontata in questo elaborato.
Ulteriori analisi sono presenti in Figura 3.5 in cui sono osservabili tre grafici a barre, uno per ogni percorso inferenziale. Ogni grafico rappresenta la distribuzione nei 500 ricampionamenti del numero di unità tassonomiche differenzialmente abbondanti in un preciso percorso inferenziale. In particolare si nota visivamente come nella normalizzazione CSS con la modellazione ZIG molti campioni abbiano per la maggior parte 0 o al più un paio di unità tassonomiche differenzialmente abbondanti. Diverso è il discorso quando si ha a che fare con le normalizzazioni TMM e RLE con la modellazione NB che sembrano trovare un numero più cospicuo di OTU differenzialmente abbondanti e solo in circa 10 ricampionamenti su 500 nessuna OTU. Con la Figura 3.6 invece, è possibile osservare il numero di OTU differenzialmente abbon-

Tabella 3.1: Proporzione di ricampionamenti in cui le unità tassonomiche elencate, indicate con la famiglia e il genere tassonomico, sono presenti tra le differenzialmente abbondanti indipendentemente dal percorso inferenziale.

Differenzialmente abbondanti		
Famiglia	**Genere**	**Proporzione**
Pasteurellaceae	Haemophilus	0.70
Enterobacteriaceae	Citrobacter	0.65
Ruminococcaceae	Flavonifractor	0.57
Natranaerovirga		0.56
Lachnospiraceae	Fusicatenibacter	0.53
Ruminococcaceae	Pseudoflavonifractor	0.53
Veillonellaceae	Veillonella	0.48
Bacillaceae_1	Bacillus	0.47
Ruminococcaceae	Gemmiger	0.43
Staphylococcaceae	Staphylococcus	0.42

danti individuate da tutte le diverse combinazioni di percorsi inferenziali. Partendo dal terzo grafico, che si presenta in modo diverso rispetto a tutti gli altri, si evidenzia la sovrapposizione tra i metodi inferenziali che utilizzano la distribuzione Binomiale Negativa per modellare i dati di conteggio. Anche se le due normalizzazioni sono diverse, al termine del procedimento inferenziale, vengono individuate un numero di unità tassonomiche in condivisione maggiore di 0 nella maggior parte dei campioni. Le code della distribuzione sono visibilmente pesanti ma complessivamente si tratta di una situazione completamente diversa rispetto a quella osservata nei primi due grafici in cui la moda si trova in 0. Lo stesso discorso vale per l'intersezione tra il percorso inferenziale avente normalizzazione CSS e modellazione ZIG con gli altri percorsi, che in circa il 60% dei campioni rimane vuota, nel 20% contiene solo un'unità, nel 10% ne contiene due e via via crescendo, ma in un numero sempre più piccolo di campioni. Per quanto riguarda quali OTU sono state individuate come differenzialmente abbondanti, vengono presentate una serie di tabelle:

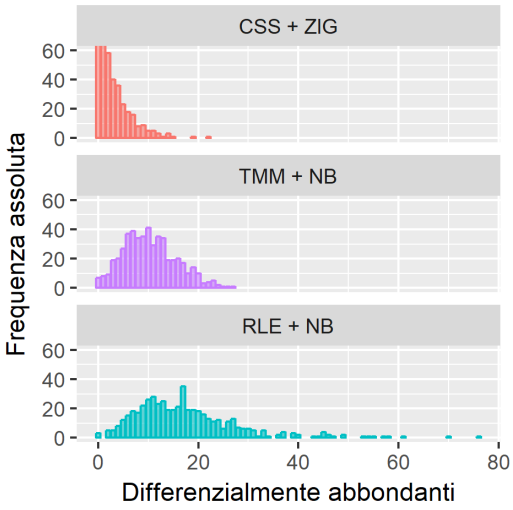

Figura 3.5: Grafici a barre per rappresentare il numero di unità tassonomiche differenzialmente abbondanti per ogni percorso inferenziale. (Viene presentato un ingrandimento del grafico nei valori delle ordinate da 0 a 60 per rendere maggiormente leggibile la parte più informativa del supporto).

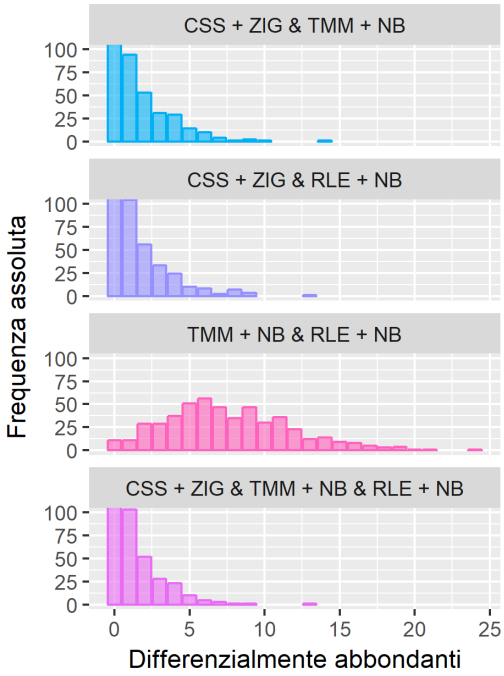

Figura 3.6: Grafici a barre per rappresentare il numero di unità tassonomiche differenzialmente abbondanti condivise in ogni combinazione di percorso inferenziale. (Viene presentato un ingrandimento del grafico nei valori delle ordinate da 0 a 100 per rendere maggiormente leggibile la parte più informativa del supporto).

- la Tabella 3.2 presenta le 5 OTU più individuate nei 500 ricampionamenti solo dal percorso inferenziale CSS con ZIG;
- la Tabella 3.3 presenta le 5 OTU più individuate nei 500 ricampionamenti dai percorsi inferenziali TMM con NB e RLE con NB vista la similitudine rilevata nei due percorsi inferenziali dal terzo grafico di Figura 3.6;
- la Tabella 3.4 presenta le 5 OTU più individuate nei 500 ricampionamenti da tutti i percorsi inferenziali contemporaneamente;
- le Tabelle 3.5 e 3.6 che presentano rispettivamente le 10 OTU esclusive del gruppo dei vegetariani e del gruppo degli onnivori.

Tabella 3.2: Proporzione di ricampionamenti in cui le unità tassonomiche elencate, indicate con la famiglia e il genere tassonomico, sono presenti tra le differenzialmente abbondanti nel percorso inferenziale CSS con ZIG.

CSS+ZIG		
Famiglia	**Genere**	**Proporzione**
Porphyromonadaceae	Butyricimonas	0.15
Veillonellaceae	Propionispora	0.09
Bacillaceae_2	Gracilibacillus	0.06
Sutterellaceae	Parasutterella	0.06
Ruminococcaceae	Ethanoligenens	0.05

Rimane da indagare la bontà di adattamento per ciascun percorso inferenziale. Per fare ciò si studia la distribuzione della statistica RMSE, acronimo per *Rooted Mean Squared Error*. Con questa statistica, calcolata per ogni OTU, è possibile avere un'indicazione in ogni normalizzazione di quanto è appropriato il modello ZIG o il modello NB sulle conte normalizzate. Valori bassi di questa statistica indicano un adattamento migliore del modello ai dati ma non si tratta di un valore assoluto: una volta individuato il percorso inferenziale con i valori più bassi rimane da verificare la presenza di andamenti sistematici nei grafici dei residui standardizzati contro i valori stimati dal modello. I grafici da creare ed osservare sono uno per ogni OTU e non

Tabella 3.3: Proporzione di ricampionamenti in cui le unità tassonomiche elencate, indicate con la famiglia e il genere tassonomico, sono presenti tra le differenzialmente abbondanti nei percorsi inferenziali TMM con NB e RLE con NB simultaneamente.

TMM + NB & RLE + NB		
Famiglia	**Genere**	**Proporzione**
Enterobacteriaceae	Citrobacter	0.31
Pasteurellaceae	Haemophilus	0.28
Lachnospiraceae	Fusicatenibacter	0.27
Vallitalea		0.25
Sphingobacteriaceae	Sphingobacterium	0.25

Tabella 3.4: Proporzione di ricampionamenti in cui le unità tassonomiche elencate, indicate con la famiglia e il genere tassonomico, sono presenti tra le differenzialmente abbondanti in tutti i percorsi inferenziali simultaneamente.

CSS + ZIG & TMM + NB & RLE + NB		
Famiglia	**Genere**	**Proporzione**
Pasteurellaceae	Haemophilus	0.20
Veillonellaceae	Veillonella	0.12
Ruminococcaceae	Pseudoflavonifractor	0.11
Neisseriaceae	Neisseria	0.06
Ruminococcaceae	Flavonifractor	0.05

Tabella 3.5: Proporzione di ricampionamenti in cui le unità tassonomiche elencate, indicate con la famiglia e il genere tassonomico, sono presenti solo nel gruppo dei vegetariani.

Vegetariani		
Famiglia	**Genere**	**Proporzione**
Clostridiaceae_2	Alkaliphilus	1.00
Pasteurellaceae	Actinobacillus	1.00
Rhodospirillaceae	Rhodospirillum	0.87
Lachnospiraceae	Shuttleworthia	0.76
Carnobacteriaceae	Granulicatella	0.74
Methanobacteriaceae	Methanobrevibacter	0.73
Campylobacteraceae	Campylobacter	0.57
Fusobacteriaceae	Fusobacterium	0.53
Erysipelotrichaceae	Holdemanella	0.46
Clostridiaceae_4	Geosporobacter	0.46

Tabella 3.6: Proporzione di ricampionamenti in cui le unità tassonomiche elencate, indicate con la famiglia e il genere tassonomico, sono presenti solo nel gruppo degli onnivori.

Onnivori		
Famiglia	**Genere**	**Proporzione**
Bacteroidaceae	Anaerorhabdus	0.85
Clostridiales_Incertae_Sedis_XIII	Anaerovorax	0.81
Clostridiales_Incertae_Sedis_XI	Ezakiella	0.80
Enterobacteriaceae	Lelliottia	0.75
Natranaerobiaceae	Dethiobacter	0.73
Peptococcaceae_1	Peptococcus	0.72
Enterobacteriaceae	Serratia	0.72
Erysipelotrichaceae	Faecalicoccus	0.65
Porphyromonadaceae	Coprobacter	0.53
Clostridiales_Incertae_Sedis_XI	Murdochiella	0.51

vengono riportati data la loro numerosità. In tutti i percorsi inferenziali sono state osservate unità tassonomiche in cui l'adattamento risultava di scarso livello, nonostante i valori bassi nella statistica RMSE. Tuttavia gli andamenti sistematici più gravi sono stati osservati per il percorso inferenziale CSS con ZIG in cui per alcune OTU si osservava un trend lineare crescente dei residui al crescere dei valori stimati. Per ottenere la statistica RMSE è sufficiente calcolare i residui di Pearson per il modello stimato:

$$r_{ij} = \frac{c_{ij} - \hat{c}_{ij}}{\sqrt{Var(\hat{c}_{ij})}}$$

Con $i = 1, \ldots, N$ numero di OTU, $j = 1, ..., M$ numero di campioni, c_{ij} valore di conteggio normalizzato per l'i-esima OTU e il j-esimo campione e $\hat{c}_{ij}$ il relativo valore di conteggio stimato. Una volta ottenuti i residui, RMSE per l'i-esima OTU è:

$$RMSE_i = \sqrt{\frac{\sum_{j=1}^{M} r_{ij}^2}{M}}$$

Ogni OTU avrà a disposizione tre valori di RMSE, uno per ogni normalizzazione. Di conseguenza, considerando tutte le unità tassonomiche di un ricampionamento si ottengono le distribuzioni degli RMSE relative alle tre normalizzazioni. Si è osservata una sostanziale similarità nelle distribuzioni comune in tutti i 500 ricampionamenti. Quindi, per motivi di semplicità il fenomeno può essere riassunto dalla Figura 3.7 relativa ad uno solo dei 500 ricampionamenti, si considerino solo i primi tre diagrammi a scatola, dai quali è possibile notare valori di RMSE mediamente più bassi per i percorsi inferenziali TMM con NB e RLE con NB rispetto al percorso CSS con ZIP.

3.1.3 Risultati

Partendo dalle normalizzazioni, si è osservata una probabile maggiore confrontabilità dei campioni biologici per i dati normalizzati tramite CSS. Questo non esclude che le normalizzazioni TMM ed RLE non siano adeguate anche se, osservando le distribuzioni delle deviazioni standard delle mediane, è possibile riflettere sulla scarsa adeguatezza di queste due normalizzazioni per il tipo di dato utilizzato e in particolare la sparsità della tabella delle

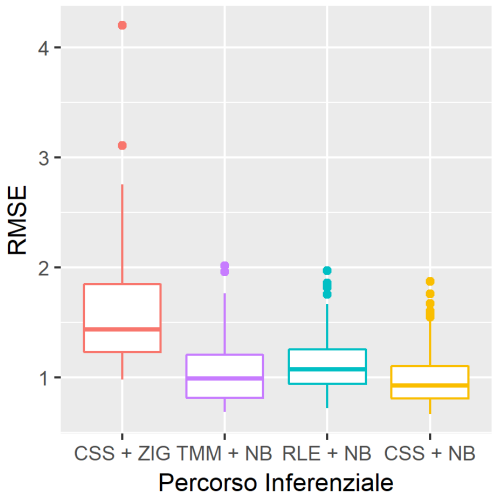

Figura 3.7: Diagramma a scatola con baffi per le statistiche RMSE nei tre percorsi inferenziali CSS con ZIG, TMM con NB, RLE con NB e nel percorso CSS con NB per il primo dei 500 ricampionamenti. Gli altri ricampionamenti presentano valori simili al presente.

OTU. Però, se i dati normalizzati tramite CSS sembrano essere più confrontabili, pare che il modello per i dati di conteggio ZIG non sia il più adatto per spiegarli. Infatti l'accostamento NB ai dati normalizzati tramite TMM e RLE ha portato a risultati mediamente migliori. Nasce spontanea, dopo l'osservazione dei risultati, l'idea di accostare ai dati normalizzati tramite CSS una distribuzione Binomiale Negativa. L'idea pare funzionare se viene utilizzata come metro di giudizio la distribuzione della statistica RMSE visibile in Figura 3.7 nell'ultimo diagramma a scatola. La distribuzione degli RMSE per il dato normalizzato tramite CSS con modello NB per le conte non è significativamente diversa rispetto a quella della TMM con NB e RLE con NB (Test di Kolmogorov-Smirnov, $p\text{-}value > 0.05$). Tuttavia permangono alcune problematiche riguardo agli andamenti sistematici nei grafici dei residui contro i valori stimati. Dall'immagine è possibile pensare che il dato normalizzato, indipendentemente dalla normalizzazione, si adatti meglio ad una distribuzione Binomiale Negativa. Ciò non toglie la sensatezza nel pensare al dato di conteggio come ad una mistura Normale *Zero-Inflated*. Molto probabilmente per i più svariati motivi, che possono derivare sia dalle decisioni applica-

te in fase di filtraggio del dato, sia dall'insieme di dati specifico utilizzato, la componente Normale della mistura non sembra così appropriata. Alcuni sviluppi interessanti potrebbero essere proprio basati sul cambio della legge distributiva per le conte diverse da zero. Modelli mistura Binomiale Negativa *Zero-Inflated*, Poisson *Zero-Inflated* o altri possono essere alcune scelte. In questi ultimi casi un modo di procedere potrebbe essere quello di utilizzare un approccio di simulazione tramite bootstrap parametrico per valutare la sensatezza di una distribuzione piuttosto che un'altra per poi procedere all'implementazione del macchinario inferenziale vero e proprio.

3.2 Microbioma Italiano

I dati del Progetto Microbioma Italiano consistono in 148 campioni biologici, insieme molto ridotto rispetto alle proporzioni di quello americano. La variabile su cui sviluppare le analisi rimane la dieta, tuttavia il questionario del progetto italiano suddivide le tipologie di diete in modo diverso rispetto a quanto fatto dal progetto americano. In particolare esiste la classe dei vegetariani ma non una classe degli onnivori dì per sé, che tuttavia sembra essere in qualche modo sostituita dalla classe "Dieta italiana standard" ovvero una probabile forma nazionalizzata della dieta mediterranea che può contenere carne oltre ad una grande quantità di frutta e verdura. La percezione individuale di "dieta italiana standard" può essere alquanto soggettiva ma in ogni caso non esclude la carne. Per fare in modo di ottenere un confronto il più possibile esente dall'influenza di variabili confondenti è necessaria una selezione degli individui che non hanno utilizzato antibiotici nel periodo precedente al prelievo ottenendo un pool di campioni composto da 12 vegetariani e 49 diete italiane standard. Lo sbilanciamento rimane evidente ma se per il progetto americano gli onnivori erano 22 volte i vegetariani, per il progetto italiano il rapporto vegetariani-dieta italiana è di circa 1 a 4; per questo si preferisce procedere considerando un unico confronto contenente tutti i campioni dei vegetariani e delle diete italiane simultaneamente per un totale di 61 campioni biologici. Si segue l'iter già visto per l'analisi dei dati del progetto americano:

1. si filtrano i campioni considerando solo le OTU presenti in almeno la metà dei campioni per ogni gruppo cioè in 6 dei 12 campioni per i vegetariani e 25 dei 49 campioni delle diete italiane standard;

2. si considera la matrice delle OTU composta solamente dalle OTU presenti in entrambi i gruppi di campioni ottenendo una matrice con 71 unità tassonomiche e 61 campioni;

A questo punto il dato è pronto per passare alla fase di normalizzazione.

3.2.1 Normalizzazione

Avendo a che fare con un unico confronto, per misurare l'effetto delle normalizzazioni si considerano le quattro matrici: quella dei dati grezzi e delle tre normalizzazioni CSS, TMM e RLE. Per ogni campione di ogni matrice calcolo la mediana e lo scarto interquartile per i valori di conteggio in scala logaritmica. Una minor variabilità in queste statistiche dovrebbe rappresentare una maggior confrontabilità tra campioni. Il risultato è visibile in Figura 3.8, dalla quale è possibile osservare una coerenza con quanto osservato in precedenza in Figura 3.2 relativa ai dati americani. La normalizzazione CSS produce dei campioni caratterizzati dalle mediane dei valori di conteggio in trasformata logaritmica molto più stabili rispetto ai valori delle altre normalizzazioni. Per avere un quadro più chiaro è possibile considerare la Tabella 3.7 dalla quale appare evidente l'uguaglianza delle medie delle mediane per tutte le normalizzazioni (affermazione corroborata anche da un test t di Student per dati appaiati $p\text{-}value > 0.1$) mentre si nota che gli scarti interquartili hanno valori mediamente maggiori rispetto ai dati grezzi per la distribuzione delle mediane nella normalizzazione CSS ($p\text{-}value < 0.006$) e valori simili ai dati grezzi per la normalizzazione TMM e RLE. Tenendo presente che per il progetto italiano è stato utilizzato un filtraggio che considera le OTU presenti in almeno il 50% dei campioni (anzichè il 30% del progetto americano) vista la minore disponibilità di dati, si nota come i valori assunti dalle statistiche nei grafici in Figura 3.8 siano più elevati rispetto a quelli in Figura 3.2. Considerando che la scala di misura è logaritmica si deduce che i campioni del Progetto Microbioma Italiano presentano profondità di lettura e di

conseguenza *library size* molto maggiori rispetto ai campioni dell'*American Gut Project* ad indicare una qualità maggiore del dato.
Concludendo:

- la normalizzazione CSS porta i campioni ad un livello di confrontabilità migliore rispetto alle altre normalizzazioni;

- le normalizzazioni TMM e RLE portano a campioni piuttosto simili ai dati grezzi.

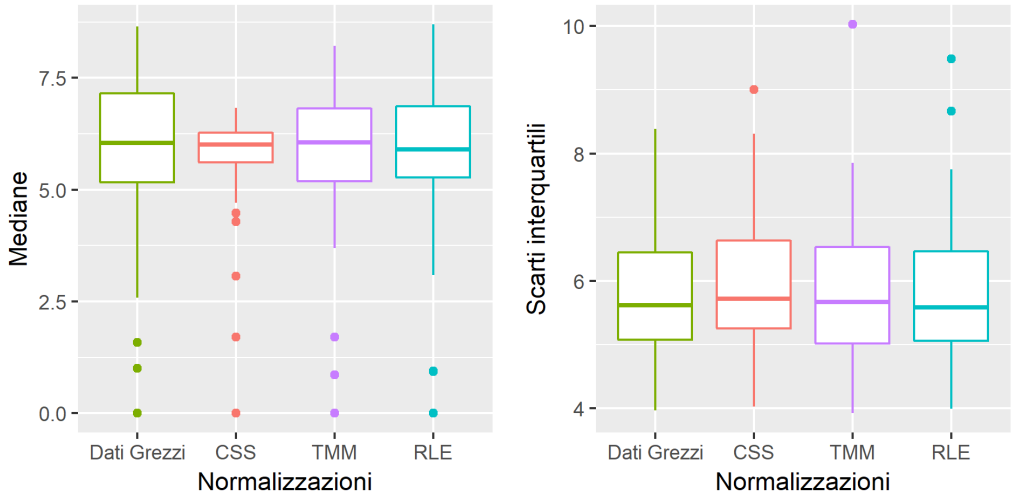

Figura 3.8: Diagrammi a scatola con baffi per le distribuzioni delle mediane (a sinistra) e degli scarti interquartili (a destra), calcolati per i dati grezzi e le normalizzazioni *Cumulative Sum Scaling* (CSS), *Trimmed Mean of M-values* (TMM) e *Relative Log Expression* (RLE) nei 61 campioni. Scala di misura log_2.

Tabella 3.7: Medie e deviazioni standard delle mediane e degli scarti interquartili dei valori di conta in scala logaritmica per i dati grezzi e i dati normalizzati secondo le normalizzazioni *Cumulative Sum Scaling*, *Trimmed Mean of M-values* e *Relative Log Expression* nei 61 campioni.

	Mediane		**Scarti interquartili**	
Normalizzazione	Media	Deviazione standard	Media	Deviazione standard
Dati grezzi	5.80	1.74	5.85	1.01
CSS	5.72	1.10	6.02	1.15
TMM	5.81	1.56	5.91	1.16
RLE	5.80	1.63	5.89	1.11

3.2.2 Inferenza

I percorsi inferenziali sono quelli già visti in precedenza:

- alla matrice delle OTU normalizzata tramite CSS viene proposto un modello ZIG per i valori di conta in scala logaritmica;
- per il dato normalizzato tramite TMM e RLE viene proposta una distribuzione NB per modellare le conte.

I tre percorsi inferenziali non portano all'individuazione di unità tassonomiche differenzialmente abbondanti nei gruppi dei vegetariani rispetto al gruppo delle diete italiane standard. Vengono riportate comunque le tabelle delle OTU che più sembrano avvicinarsi all'abbondanza differenziale senza però arrivare alla significatività statistica per i tre percorsi inferenziali. In particolare vengono tabulate nelle Tabelle 3.8, 3.9 e 3.10 le OTU che in ciascun percorso inferenziale risultano avere un *p-value* non aggiustato inferiore a 0.05.
Si osserva come i *p-value* aggiustati risultino di gran lunga superiori al valore soglia previsto per considerare un'OTU differenzialmente abbondante, ricordando che nei dati del progetto americano la soglia era 0.1. L'ultimo aspetto da tenere in considerazione è la bontà di adattamento valutata tramite RMSE nei tre percorsi inferenziali. Il grafico raffigurante le distribuzioni di queste statistiche è presente in Figura 3.9 nei primi tre diagrammi a scatola. Gli RMSE sono diversi tra il primo percorso inferenziale CSS con ZIG e

Tabella 3.8: *Log-Fold Change*, *p-value* < 0.05 e *p-value* aggiustati con il metodo di correzione del *False Discovery Rate* per le famiglie ed i generi tassonomici nel percorso inferenziale CSS con ZIG.

CSS + ZIP				
Famiglia	**Genere**	***LogFC***	***p-value***	***p-val.* agg.**
Porphyromonadaceae	Parabacteroides	-1.34	0.02	0.64
Lachnospiraceae	Anaerosporobacter	1.28	0.02	0.64
Ruminococcaceae	Sporobacter	0.87	0.05	0.64

Tabella 3.9: *Log-Fold Change*, *p-value* < 0.05 e *p-value* aggiustati con il metodo di correzione del *False Discovery Rate* per le famiglie ed i generi tassonomici nel percorso inferenziale TMM con NB.

TMM + NB				
Famiglia	**Genere**	***LogFC***	***p-value***	***p-val.* agg.**
Actinomycetales	Micrococcineae	2.52	$<$0.01	0.28
Enterobacteriaceae	Shigella	-2.85	0.01	0.32
Lachnospiraceae	Anaerosporobacter	1.71	0.03	0.44
Lactobacillaceae	Lactobacillus	-3.21	0.03	0.44
Peptostreptococcaceae	Romboutsia	-1.44	0.03	0.44
Peptostreptococcaceae	Intestinibacter	-1.88	0.04	0.44

Tabella 3.10: *Log-Fold Change*, *p-value* < 0.05 e *p-value* aggiustati con il metodo di correzione del *False Discovery Rate* per le famiglie ed i generi tassonomici nel percorso inferenziale RLE con NB.

RLE + NB				
Famiglia	**Genere**	***LogFC***	***p-value***	***p-val.* agg.**
Actinomycetales	Micrococcineae	2.51	$<$0.01	0.28
Enterobacteriaceae	Shigella	-2.85	0.01	0.31
Lachnospiraceae	Anaerosporobacter	1.71	0.03	0.46
Lactobacillaceae	Lactobacillus	-3.20	0.03	0.46
Peptostreptococcaceae	Romboutsia	-1.45	0.04	0.46
Peptostreptococcaceae	Intestinibacter	-1.88	0.04	0.46

gli altri percorsi inferenziali: si osserva una distribuzione più schiacciata per il percorso CSS con ZIG e una distribuzione con diversi valori anomali per i percorsi TMM con NB e RLE con NB. Per quanto riguarda i valori non è possibile affermare che gli RMSE del percorso CSS con ZIG siano mediamente maggiori rispetto a quelli degli altri percorsi a tutti i livelli di significatività sensati. Visionando i grafici dei residui standardizzati contro i valori stimati dai modelli, che non vengono riportati data la loro numerosità, si osservano andamenti sistematici specialmente in quelli relativi al percorso CSS con ZIG. In conclusione, l'adattamento migliore sembra appartenere ai percorsi inferenziali con distribuzione Binomiale Negativa per i dati di conteggio, nonostante i diversi valori anomali nelle statistiche RMSE. Viene proposto anche per i dati del progetto italiano il grafico con il quarto percorso inferenziale costituito dai dati normalizzati tramite CSS ma modellati con NB, in Figura 3.9 quarto diagramma a scatola.

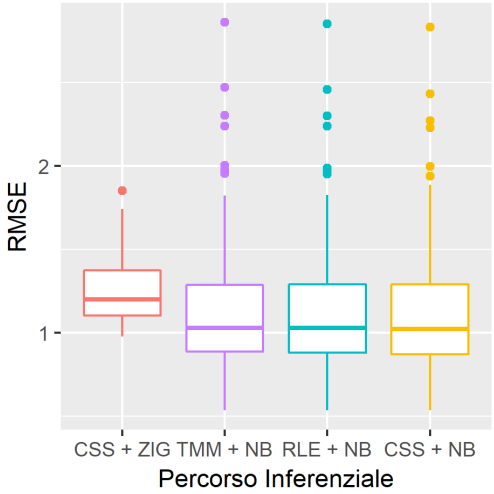

Figura 3.9: Diagramma a scatola con baffi per le statistiche RMSE nei tre percorsi inferenziali CSS con ZIG, TMM con NB, RLE con NB e nel percorso CSS con NB.

3.2.3 Risultati

Da un punto di vista biologico si può accettare il risultato ottenuto pensando alla particolarità della dieta italiana, caratterizzata per la maggior parte da frutta e verdura e solo in minima parte dalla carne. La dieta italiana dunque non risulterebbe così lontana da una dieta esclusivamente vegetariana come invece sembra essere una dieta onnivora americana rispetto ad una dieta vegetariana americana.
Da un punto di vista più metodologico statistico invece, si possono fare un paio di considerazioni. In primo luogo il microbioma dello specifico campione di italiani vegetariani non è così diverso dal microbioma dello specifico campione che ha una dieta italiana standard. Questa ipotesi può acquisire credibilità se si considera un articolo in cui era emerso che vi sono due fattori in grado di influenzare la potenza dei test statistici, aumentandola: la profondità di sequenziamento maggiore (quindi *library size* più grandi) e il numero di repliche maggiore (Liu, Zhou e White 2014). Nel progetto americano sono state individuate un basso numero di OTU differenzialmente abbondanti, ma tale progetto considerava un maggior numero di repliche. L'altra faccia della medaglia è rappresentata dalle caratteristiche dei campioni del progetto americano, in cui i valori di conta e quindi le *library size* sono minori rispetto alle stesse nei campioni italiani. Di conseguenza, se da una parte si toglie, con un numero inferiore di campioni, dall'altra si guadagna, con una maggiore profondità di sequenziamento. Tuttavia lo stesso articolo incoraggia un maggior numero di repliche anziché una maggiore profondità di sequenziamento evidenziando, nel primo caso, il maggiore guadagno di potenza.
In secondo luogo, il numero contenuto di individui che hanno deciso di aderire al progetto italiano, è probabilmente influenzato da particolari stati patologici che sarà opportuno indagare in modo più approfondito. Non valutare le caratteristiche patologiche degli individui può provocare un aumento del rumore nel segnale, impedendo agli strumenti statistici di farsi strada in questa aumentata eterogeneità individuando le informazioni di interesse. Riuscire a tener conto di tutti i possibili confondenti in questo tipo di confronto senza la consultazione di personale specializzato è una visione a dir poco ottimistica. Infatti l'introduzione di una covariata, da considerare come confondente, nel

modello statistico è un'operazione delicata. Una covariata in più richiede una parte dell'informazione totale per essere stimata, togliendola alla variabile su cui si sta lavorando. Informazione che, se utilizzata per rimuovere il rumore che copre il segnale di interesse, è ben spesa ma che, al contrario, se utilizzata per stimare un falso confondente porta a stime inefficienti.
Alla luce di queste congetture, si può pensare di eseguire una breve analisi descrittiva riguardante gli stati patologici di cui sono affetti i partecipanti ad entrambi gli studi. In particolare ha senso considerare il sottogruppo utilizzato per le analisi: gli onnivori o le diete italiane standard e i vegetariani che non hanno fatto uso di antibiotici nei mesi precedenti al prelievo del campione biologico. Il questionario del Progetto Microbioma Italiano ha una sezione dedicata alle patologie di cui è affetto il partecipante e soprattutto una sezione dedicata ai sintomi, patologici e non, che l'individuo presenta al momento del prelievo del campione biologico. Le patologie sono presentate in Figura 3.10, mentre i sintomi al prelievo in Figura 3.11. Le patologie più comuni per i 49 individui con dieta italiana standard sembrano essere l'ansia (8), l'asma (5), la sindrome dell'intestino irritabile (5), l'ipertensione (9), il reflusso gastro-esofageo (9) e la Tiroidite di Hashimoto (7). Solo 21 persone di queste 49 non soffrono di alcuna patologia indicando l'esistenza di persone con diversi stati patologici simultanei. Nei 12 campioni di vegetariani si osserva, per alcune patologie, al massimo una persona malata. In questo gruppo di 12 tuttavia, solo 6 non hanno alcuna patologia. Passando ai sintomi al prelievo, le classi di sintomi più comuni per le 49 diete italiane standard, sono quelle relative al meteorismo (13), alle feci molli (8), e ai rash cutanei (5). Solo 19 individui non presentano alcun sintomo. Nei 12 vegetariani solo 4 individui hanno qualche sintomo al prelievo di cui 2 lamentano feci molli, 1 meteorismo e 1 insonnia. In conclusione, circa la metà degli individui su cui è stata fatta l'analisi del microbioma lamenta sintomi al prelievo o è affetto da patologie più o meno gravi che, in diversi casi, coinvolgono l'apparato digerente.
Per il progetto americano purtroppo la situazione non è così chiara come per il progetto italiano. Non sono disponibili variabili che indicano sintomi al prelievo e le patologie indicate presentano circa la metà di dati mancanti. Vengono riportate in Figura 3.12 i grafici delle patologie per gli onnivori e i

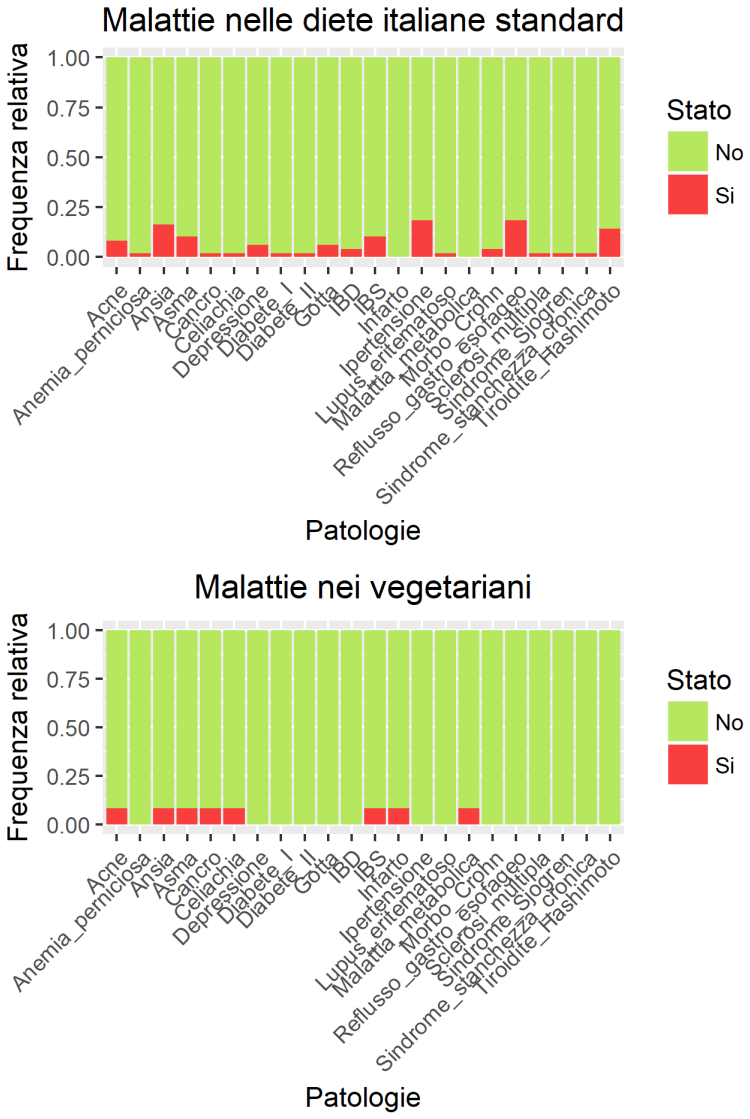

Figura 3.10: Grafico a barre per le malattie di cui sono affette le unità statistiche del Progetto Microbioma Italiano con dieta italiana standard (sopra) o vegetariana (sotto) e che non hanno utilizzato antibiotici nei mesi precedenti al prelievo.

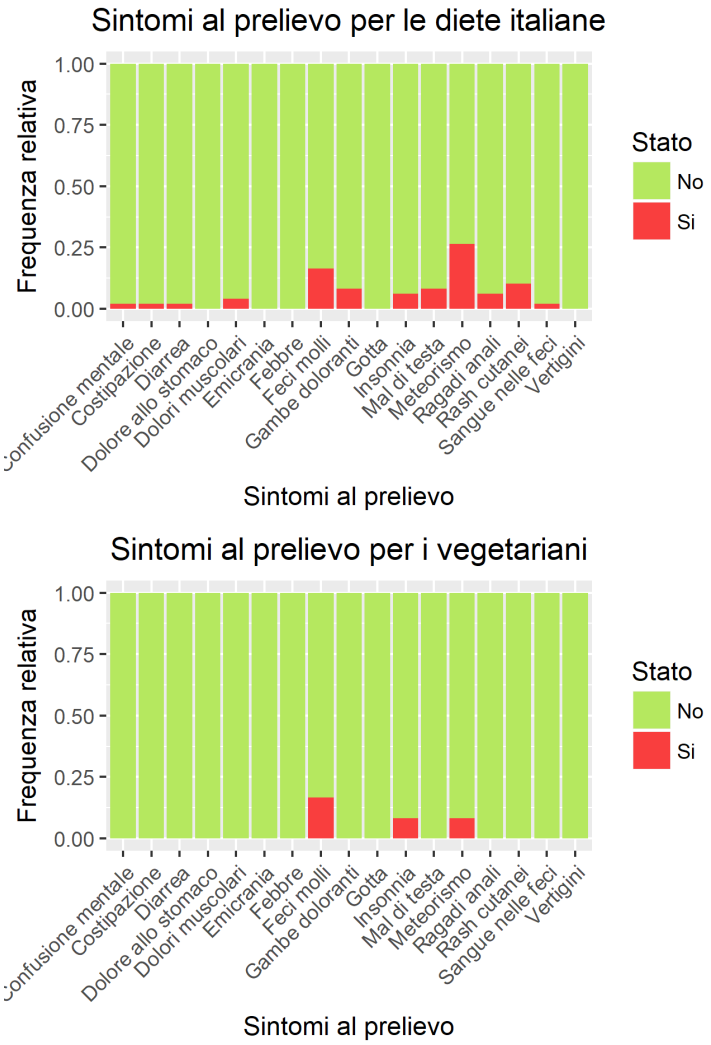

Figura 3.11: Grafico a barre per i sintomi al prelievo di cui sono affette le unità statistiche del Progetto Microbioma Italiano con dieta italiana standard (sopra) o vegetariana (sotto) e che non hanno utilizzato antibiotici nei mesi precedenti al prelievo.

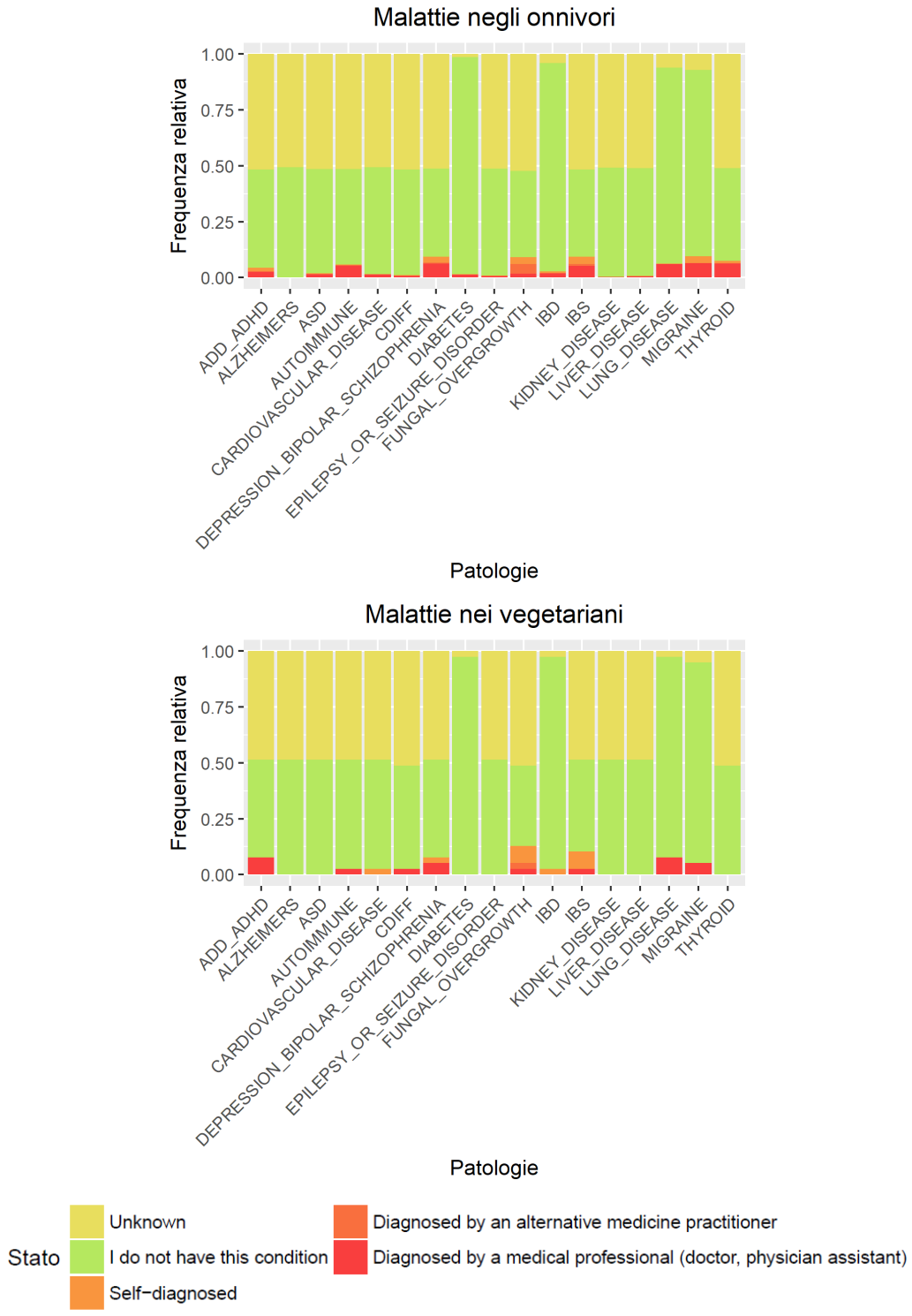

Figura 3.12: Grafico a barre per le malattie di cui sono affette le unità statistiche dell'*American Gut Project* con dieta onnivora (sopra) o vegetariana (sotto) e che non hanno utilizzato antibiotici nell'anno precedente al prelievo.

vegetariani del progetto americano. Salta all'occhio la maggior classificazione delle risposte, una patologia può infatti essere dichiarata dal partecipante allo studio in diversi modi ovvero: auto diagnosticata, diagnosticata da un medico di medicina alternativa o diagnosticata da un medico professionista. Nonostante questo valore informativo della diagnosi non è possibile eseguire un confronto alla pari tra le patologie nei due progetti nazionali vista la mancanza di dati del progetto americano.
Con i dati a disposizione, tuttavia è possibile ottenere nei due insiemi di dati la proporzione di individui che soffre di almeno una patologia legata all'intestino. Nei dati italiani le malattie più importanti in questione sono: la celiachia, il diabete di tipo 1 e 2, IBS, IBD, malattie metaboliche e morbo di Crohn. Anche le altre patologie possono presentare dei sintomi che disturbano l'apparato digerente, ma si preferisce considerare solamente quelle esattamente localizzate in questa parte del corpo. Nell'insieme di dati americano invece, le malattie maggiormente legate all'intestino sono: l'infezione da *Clostridium Difficile* (CDIFF), il diabete, IBD e IBS. Nei dati italiani la percentuale di individui che soffre di almeno una di queste patologie è pari al 18.03%, mentre nei dati americani, senza considerare i valori mancanti, è pari al 26.48%. Considerando che le numerosità campionarie nei due insiemi di dati sono rispettivamente 61 per l'italiano e 423 per l'americano, si esegue un test χ^2 per testare l'uguaglianza delle proporzioni. Non si rifiuta l'ipotesi di uguaglianza in quanto il *p-value* fornito dal test risulta pari a 0.21.
In conclusione l'*American Gut Project* ha individuato una qualche differenza tra onnivori e vegetariani, il Progetto Microbioma Italiano non ha individuato nulla tra vegetariani e diete italiane. In seguito vengono riassunte le possibili motivazioni dei risultati ottenuti:

- l'analisi dei dati del progetto americano coinvolgeva 39 onnivori e 39 vegetariani per ogni confronto, mentre quella del progetto italiano 12 vegetariani e 39 diete italiane standard. Per un totale di 78 campioni contro 61. La potenza dei test è maggiormente influenzata dal numero di repliche rispetto che dalla profondità di sequenziamento.

- non in tutti i 500 ricampionamenti effettuati per l'analisi del progetto americano risultavano differenzialmente abbondanti delle unità tasso-

nomiche dimostrando l'entità del segnale biologico sui campioni coinvolti. Il genere tassonomico Haemophilus e quello Citrobacter sono stati trovati differenzialmente abbondanti rispettivamente nel 70% e 65% dei ricampionamenti indicando un segnale biologico moderato ma non forte.

- in entrambi i progetti circa il 20% delle unità statistiche soffre di almeno una patologia legata all'intestino e, per il progetto italiano si conoscono anche sintomi patologici al prelievo. Questi fattori potrebbero influenzare il microbioma stesso degli individui aumentandone la variabilità da catturare con i modelli.

- una dieta italiana standard è probabilmente molto più simile ad una dieta vegetariana rispetto a quanto non lo sia una dieta onnivora in America.

Capitolo 4

Conclusione

Sono stati fatti molti studi sul microbioma e anche in Italia è partito recentemente il Progetto Microbioma Italiano che vede BMR Genomics come azienda responsabile della parte sperimentale. L'azienda si occupa della raccolta e del sequenziamento dei campioni biologici provenienti dai cittadini italiani che vogliono dare un contributo alla ricerca; ogni campione sequenziato permette al cittadino di conoscere la composizione della propria flora intestinale.
Questo elaborato nasce dalla necessità dell'azienda di mettere a punto dei metodi di analisi adeguati, per rispondere poi ad una serie di domande biologiche, come ad esempio: che differenze si riscontrano nei microbiomi di persone caratterizzate da stili di vita, abitudini alimentari, patologie diverse?
A partire dalla tabella delle OTU, nella prima parte dell'elaborato sono state presentate una serie di metodi di normalizzazione e inferenza. Non tutti i metodi comunemente utilizzati hanno senso se applicati alle tabelle delle OTU in quanto le differenze con i dati di RNA-Seq, da cui sono stati sviluppati diversi metodi, sono principalmente due: il minor numero di OTU rispetto al numero dei geni e la maggiore sparsità della matrice. Queste caratteristiche hanno reso necessario lo sviluppo di alcuni metodi specifici in grado di considerare le particolarità della matrice. La normalizzazione *Cumulative Sum Scaling* (CSS) con la modellazione delle conte in scala logaritmica tramite mistura Normale *Zero-Inflated* (ZIG) è uno di questi metodi presente in let-

teratura.
Nella seconda parte dell'elaborato si è voluto confrontare quest'ultimo percorso di analisi con altri due percorsi più classici, provenienti delle analisi RNA-Seq: normalizzazioni *Trimmed Mean of M-Values* (TMM) e *Relative Log-Expression* (RLE) con il dato di conta modellato secondo una distribuzione Binomiale Negativa (NB). La normalizzazione CSS ha portato i campioni ad un grado maggiore di confrontabilità rispetto a quanto abbiano fatto le altre normalizzazioni. La distribuzione ipotizzata, la ZIG, tuttavia non si è rivelata essere la migliore dal punto di vista della bontà di adattamento rispetto alla NB.
Alla luce dei risultati in entrambi i casi studio, è possibile individuare un problema di fondo nei metodi di analisi applicati: la difficoltà di una distribuzione ad adattarsi ad un dato così eterogeneo. Ecco perché l'utilizzo di *pipeline* di analisi predefinite è sconsigliato visto le caratteristiche dataset specifiche del dato. Una soluzione consiste nell'implementazione di metodi di analisi che permettano maggior flessibilità distributiva ma allo stesso tempo una visione più critica riguardante le performance dei metodi in ogni step dell'analisi.
Il bisogno dell'azienda di studiare il microbioma degli italiani richiede la creazione di uno strumento complesso. Attraverso i campioni analizzati in questo elaborato è stato possibile indagare il microbioma di 12 italiani vegetariani e confrontarlo con 39 italiani aventi dieta italiana standard. Non sono state individuate differenze significative associate con i diversi tipi di dieta. Con l'arrivo sempre più frequente di campioni, si auspica un aumento del livello di conoscenza sul tipo di dato. In particolare, un maggior numero di campioni saranno essenziali per tenere sotto controllo le variabili che probabilmente agiscono come confondenti per lo studio dell'associazione dieta-microbioma. Con la struttura di analisi creata per questa relazione finale, l'azienda avrà a disposizione un punto di partenza per indagare ulteriori associazioni di interesse.

Bibliografia

16S ribosomal RNA (2017). en. URL: https://en.wikipedia.org/w/index.php?title=16S_ribosomal_RNA&oldid=778112338.

American Gut Project (2017). URL: https://americangut.org/our-results-so-far/.

Benjamini, Yoav e Yosef Hochberg (1995). «Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing». In: *Journal of the Royal Statistical Society. Series B (Methodological)* 57.1, pp. 289–300. URL: http://www.jstor.org/stable/2346101.

Charlson, Emily S. et al. (2011). «Disordered Microbial Communities In The Upper Respiratory Tract Of Cigarette Smokers». In: A52. SMOKING AND LUNG DISEASE. American Thoracic Society, A1766. URL: http://www.atsjournals.org/doi/abs/10.1164/ajrccm-conference.2011.183.1_MeetingAbstracts.A1766.

Dillies, Marie-Agnès et al. (2013). «A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis». English. In: *Briefings in bioinformatics* 14.6, pp. 671–683. DOI: 10.1093/bib/bbs046. URL: http://www.ncbi.nlm.nih.gov/pubmed/22988256.

Gill, Steven R. et al. (2006). «Metagenomic Analysis of the Human Distal Gut Microbiome». English. In: *Science* 312.5778, pp. 1355–1359. DOI: 10.1126/science.1124234. URL: http://www.sciencemag.org/cgi/content/abstract/312/5778/1355.

Gotelli, Nicholas J. e Robert K. Colwell (2001). «Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness». English. In: *Ecology Letters* 4.4, pp. 379–391. DOI: 10.1046/j.

1461-0248.2001.00230.x. URL: http://onlinelibrary.wiley.com/doi/10.1046/j.1461-0248.2001.00230.x/abstract.

Human Microbiome Project (2017). URL: https://en.wikipedia.org/wiki/Human_Microbiome_Project.

Liu, Yuwen, Jie Zhou e Kevin P. White (2014). «RNA-seq differential expression studies: more sequence or more replication?» eng. In: *Bioinformatics (Oxford, England)* 30.3, pp. 301–304. DOI: 10.1093/bioinformatics/btt688.

Marioni, John C. et al. (2008). «RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays». English. In: *Genome research* 18.9, pp. 1509–1517. DOI: 10.1101/gr.079558.108. URL: http://www.ncbi.nlm.nih.gov/pubmed/18550803.

McGill, Brian J. et al. (2007). «Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework». English. In: *Ecology Letters* 10.10, pp. 995–1015. DOI: 10.1111/j.1461-0248.2007.01094.x. URL: http://www.narcis.nl/publication/RecordID/oai:pure.rug.nl:publications%2F2faff93c-87d9-4822-80fe-6802a45aeedd.

McMurdie, Paul J. e Susan Holmes (2014). «Waste not, want not: why rarefying microbiome data is inadmissible». English. In: *PLoS computational biology* 10.4, e1003531. DOI: 10.1371/journal.pcbi.1003531. URL: http://www.ncbi.nlm.nih.gov/pubmed/24699258.

Nguyen, Nam phuong et al. (2016). «A perspective on 16S rRNA operational taxonomic unit clustering using sequence similarity». English. In: *NPJ Biofilms and Microbiomes* 2, p. 16004. DOI: 10.1038/npjbiofilms.2016.4. URL: https://search.proquest.com/docview/1782221615.

Paulson, Joseph N. et al. (2013). «Differential abundance analysis for microbial marker-gene surveys». en. In: *Nature Methods* 10.12, pp. 1200–1202. DOI: 10.1038/nmeth.2658. URL: http://www.nature.com/nmeth/journal/v10/n12/full/nmeth.2658.html.

Progetto Microbioma Italiano, Citizen Science (2017). URL: http://progetto-microbiomaitaliano.org/progetto/page-12/.

Robinson, Mark D. e Gordon K. Smyth (2008). «Small-sample estimation of negative binomial dispersion, with applications to SAGE data». en. In:

Biostatistics 9.2, pp. 321–332. DOI: 10.1093/biostatistics/kxm030. URL: https://academic.oup.com/biostatistics/article/9/2/321/353777.

Smyth, Gordon K. (2004). «Linear models and empirical bayes methods for assessing differential expression in microarray experiments». eng. In: *Statistical Applications in Genetics and Molecular Biology* 3, Article3. DOI: 10.2202/1544-6115.1027.

Soneson, Charlotte e Mauro Delorenzi (2013). «A comparison of methods for differential expression analysis of RNA-seq data». English. In: *BMC bioinformatics* 14, p. 91. DOI: 10.1186/1471-2105-14-91. URL: http://www.ncbi.nlm.nih.gov/pubmed/23497356.

Weiss, Sophie et al. (2017). «Normalization and microbial differential abundance strategies depend upon data characteristics». English. In: *Microbiome* 5. DOI: 10.1186/s40168-017-0237-y. URL: https://search.proquest.com/docview/1874317872.

Willey, Joanne, Linda Sherwood e Christopher Woolverton (2014). *Prescott's microbiology*. English, pp. 713–721. ISBN: 9780073402406.