Analisi microbioma

Dargenio Elisabetta, Grassi Angela, Marcolin Erica

Introduzione

Il dataset relativo al progetto è stato ottenuto dal pacchetto *curatedMetagenomicData* di Bioconductor che fornisce l'abbondanza di marcatori tassonomici, funzionali e genetici per campioni raccolti da diversi siti del corpo umano (latte, cavità nasale, cavo orale, pelle, feci, vagina). Sono disponibili dati processati relativi al sequenziamento dell'intero metagenoma per migliaia di campioni di microbioma umano, incluso lo "Human Microbiome Project". I dati di microbioma e i metadati associati relativi ai soggetti, alle specie e al sequenziamento sono integrati come oggetti di tipo "Bioconductor ExpressionSet", consentendo analisi dirette. Una funzione di conversione collega i dataset tassonomici al pacchetto *phyloseq*, consentendo di lavorare sui dati utilizzando la classe "phyloseq", molto utile per selezionare il livello (rango) tassonomico a cui analizzare i dati. I dataset metagenomici si presentano come tabelle di conteggi ma in questo caso le feature non sono geni ma microrganismi (batteri e virus) residenti in siti specifici del corpo umano. I dati sulla conta microbica sono rappresentati utilizzando unità tassonomiche operative (OTU). Le unità tassonomiche (dette anche taxa) sono strutturate in una gerarchia di inclusione: dominio ⊃ phylum ⊃ classe ⊃ ordine ⊃ famiglia ⊃ genere ⊃ specie ⊃ ceppo, una struttura ad albero in cui le foglie sono costituite dal ceppo (Strain).
Le caratteristiche principali dei dati di sequenziamento del microbioma sono la sovradispersione e l'elevata inflazione di zeri. La sparsità delle matrici di conteggio può essere spiegata sia da ragioni biologiche che tecniche: alcuni taxa sono davvero rari e si presentano solo in pochi campioni, mentre altri sono poco rappresentati e non possono essere rilevati a causa di una profondità di sequenziamento insufficiente o per altri motivi tecnici, Calgaro *et al.* (2020). Per queste ragioni le analisi di abbondanza differenziale per dati di microbioma risultano complicate.

Dataset

Il dataset considerato nel nostro progetto si riferisce all'articolo di Brito *et al.* (2016), in particolare ad un sottoinsieme dei dati presentati nell'articolo, relativi a microbi residenti nel cavo orale. Si tratta di campioni di saliva da 140 soggetti sani (di cui 53 maschi e 87 femmine) provenienti dalle Isole Fiji. La piattaforma di sequenziamento è la stessa per tutti i campioni, Illumina HiSeq, così come il kit di estrazione utilizzato, "Maxwell_LEV". Purtroppo tra i metadati non sono disponibili possibili variabili confondenti. Per tutti i soggetti in studio troviamo dati completi relativi a "gender" ed età, che consideriamo come variabili di interesse. Lo scopo del lavoro è valutare se nella coorte considerata vi siano microrganismi differenzialmente abbondanti tra maschi e femmine o tra diverse fasce di età.

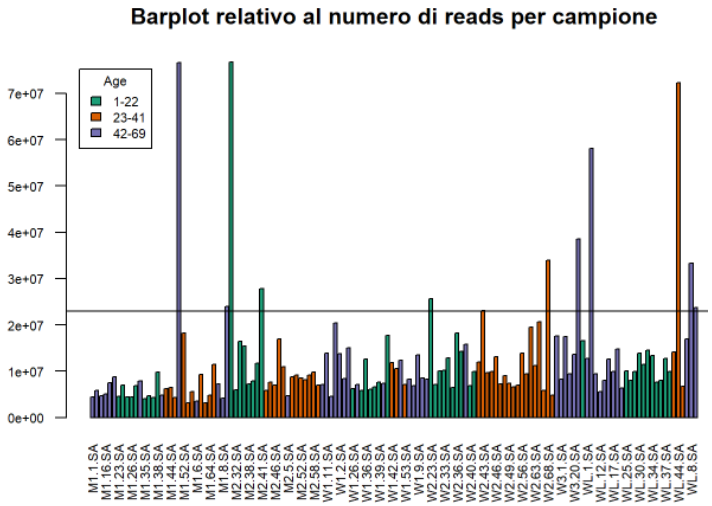
Il dataset è stato salvato dentro un oggetto di classe phyloseq, che comprende: la "Operational Taxanomic Unit (OTU) Table", che contiene i conteggi delle OTU per ogni campione (1864 taxa e 140 samples); la matrice dei Sample Data, contenente i metadati; la Taxonomy Table, contenente la tassonomia predetta per ogni OTU.

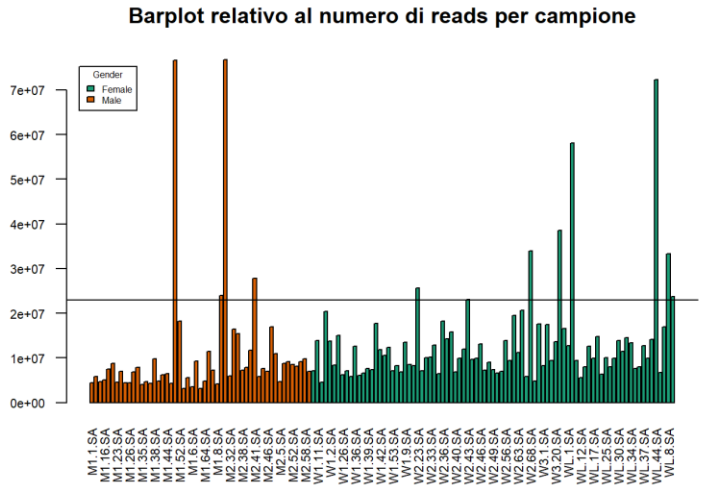
Metodi

Come prima operazione, scegliamo il rango tassonomico, cioè il livello nella gerarchia tassonomica, su cui vogliamo lavorare: Strain (ceppo), l'ultimo livello di aggrezione. Filtriamo i dati sia sui taxa che sui campioni.

Selezionando solo il rango 'Strain', il nuovo oggetto phyloseq 'data_strain' è composto da 630 taxa e 140 campioni. Si effettua un filtraggio di questo oggetto in modo da rimuovere i taxa troppo poco espressi, in particolare si tengono solo i taxa con più di 10 conteggi in più di 1 campione. Il nuovo oggetto filtrato 'filtered_strain' è composto da 282 taxa e 140 campioni. Si prova anche un filtraggio sui campioni per rimuovere eventualmente quelli con 0 conteggi, ma nessun campione viene rimosso.

Dalle prime analisi esplorative emerge che alcuni campioni hanno un numero di reads di molto superiore rispetto a tutti gli altri, non sembra che queste anomalie siano correlate con le variabili di interesse, sospettiamo quindi ci sia della variabilità tecnica dei dati. Per questo motivo, non avendo a disposizione nessuna possibile variabile confondente da aggiungere ai modelli, e avendo un numero piuttosto elevato di campioni, decidiamo di rimuovere questi 12 campioni sospetti e procedere con l'analisi sui campioni rimanenti.





Filtriamo nuovamente i taxa che, dopo la rimozione dei campioni anomali, risultano avere 0 reads. La matrice di conteggio finale su cui andiamo a svolgere le successive analisi è in definitiva composta da 256 taxa e 128 campioni.

Per le analisi di abbondanza differenziale dei microrganismi utilizziamo i metodi *edgeR* che adatta un modello binomiale negativo e *limma-voom* che adatta un modello lineare con pesi e sia *edgeR* che *limma-voom* con pesi calcolati tramite binomiale negativa a inflazione di zeri (*zinbFit*).

Dopo aver valutato quattro tipi di normalizzazione, full-quantile, upper-quartile, RLE e TMM, sulla base dei risultati ottenuti, scegliamo di procedere con l'ultima per le analisi successive. Il metodo TMM, in generale, sembra più performante su dataset con un'elevata percentuale di zeri, che è proprio il caso dei dati di metagenomica.

Per l'analisi con *edgeR* la stima della dispersione è stata calcolata tramite la funzione *estimateDisp* che è costituita da tre step: stima di un parametro di dispersione globale, stima della relazione media-dispersione e stima Bayesiana a posteriori del parametro di dispersione per ogni microrganismo, "schiacciato" verso il trend. Adattiamo il modello binomiale negativo con *edgeR* tramite la funzione *glmFit*. Facciamo l'analisi di espressione differenziale usando la funzione *glmLRT* del pacchetto *edgeR*, che utilizza il test di massima verosimiglianza per testare l'abbondanza differenziale dei taxa. Per identificare i taxa differenzialmente abbondanti, usiamo la funzione *topTags*, specificando un cut-off di 0.05 sui p-value aggiustati.

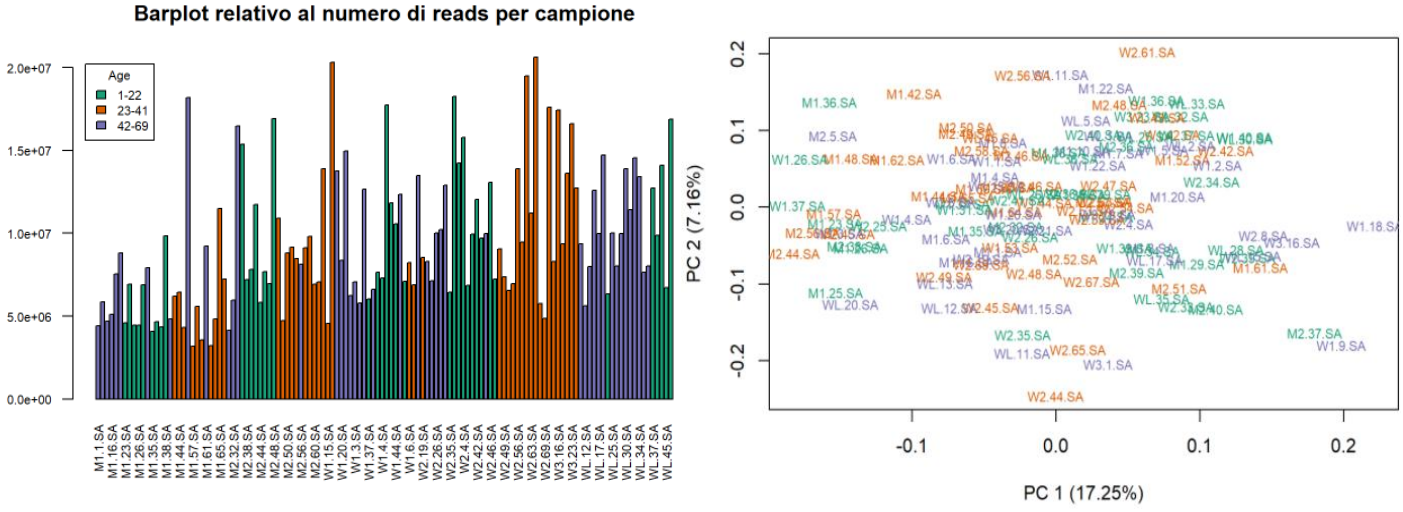
Il metodo *voom*, partendo dal logaritmo delle reads, stima la relazione media-varianza in modo non-parametrico, questa relazione è usata per stimare la varianza di ogni valore di log-cpm e infine l'inverso della varianza (precisione) è usato come peso per ogni osservazione di ogni microrganismo.

Risultati

Eseguiamo prima l'analisi per trovare eventuali differenze nell'abbondanza dei microrganismi residenti nel cavo orale tra le tre classi di età e successivamente tra maschi e femmine.

Analisi per classi di età

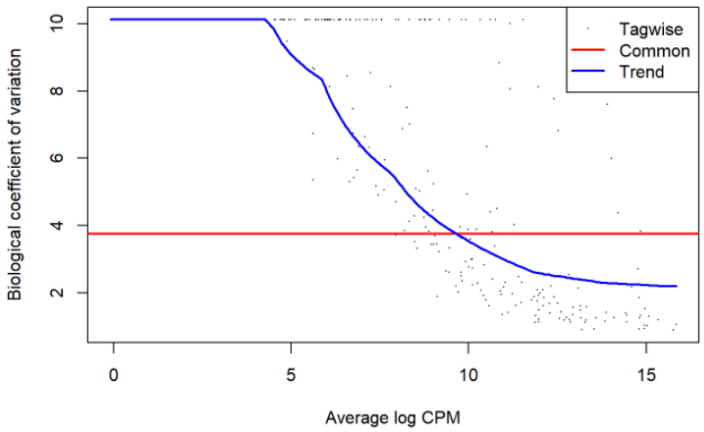
Ricodifichiamo la variabile 'Age' in una variabile categoriale con 3 classi: Giovani (age<=22), Adulti (22<age<=41) e Anziani (age>41).



L'analisi esplorativa dei dati è stata eseguita mediante: barplot relativo al numero di reads per ogni campione e rappresentazione PCA sulle prime due componenti principali.

Dalla PCA non si nota nessun tipo di raggruppamento tra i campioni, e nello specifico, le prime due componenti principali non sembrano riuscire a separare nessuna delle classe di età in esame. Queste prime analisi sembrano indicare che l'età non sia una variabile significativamente legata alla variabilità biologica del fenomeno in studio.

Valutiamo anche, tramite il seguente grafico, l'andamento della varianza rispetto all'espressione media dei taxa, e ci rendiamo conto che per molti dei microrganismi che stiamo analizzando, la varianza è alta sia per valori di espressione media bassi, che per valori alti. Questo aspetto sarà da tenere in considerazione durante l'interpretazione finale dei risultati ottenuti.

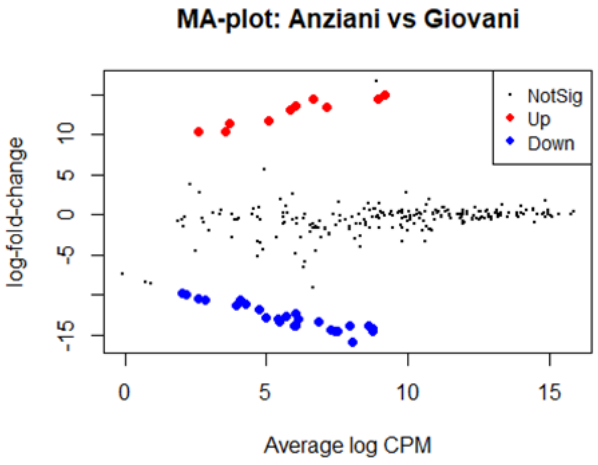


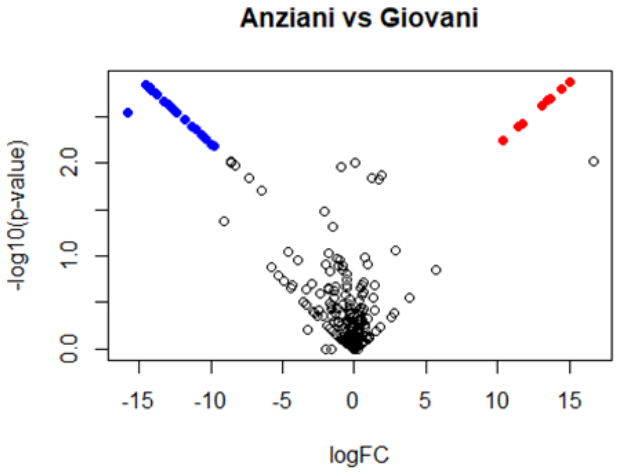
edgeR

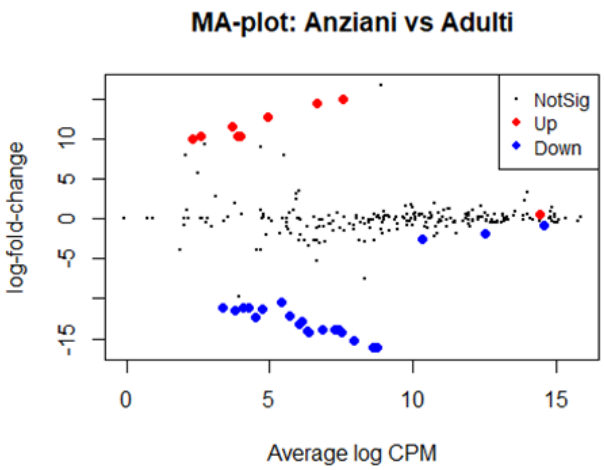
Nel confronto tra le tre classi di età risultano differenzialmente abbondanti 27 microrganismi.

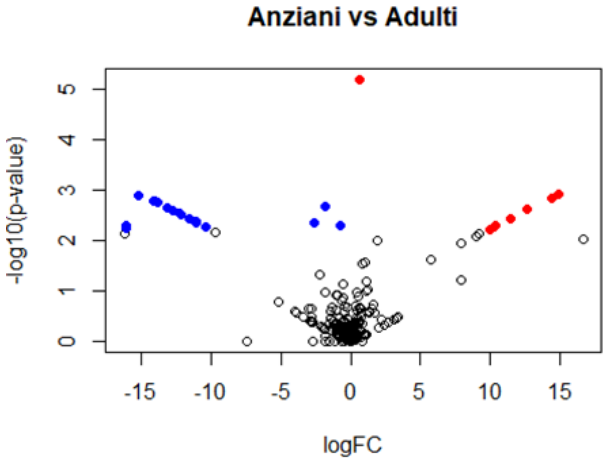
Utilizziamo i contrasti per eseguire i diversi confronti a coppie tra le classi d'età. L'analisi non evidenzia alcun microrganismo differenzialmente abbondante tra "Adulti" e "Giovani", 35 microrganismi nel confronto tra "Anziani" e "Giovani" e 31 microrganismi tra "Anziani" e "Adulti".

Per la visualizzazione dei risultati relativamente ai due contrasti per i quali abbiamo trovato microrganismi differenzialmente abbondanti utilizziamo MA-plot e volcano plot.









Nel contrasto "Anziani" vs "Giovani" si nota un certo bilanciamento nei taxa differenzialmente abbondanti tra le due classi e tutti con log-fold-change elevati.

Nel contrasto "Anziani" vs "Adulti", nella parte centrale del grafico, si notano alcuni taxa identificati come differenzialmente abbondanti che hanno log-fold-change bassi, difficilmente rilevanti dal punto di vista biologico.

limma-voom

Nel confronto tra le tre classi di età non risulta alcun microrganismo differenzialmente abbondante e anche nei contrasti a coppie non si rilevano taxa significativi.

edgeR + pesi ZINB

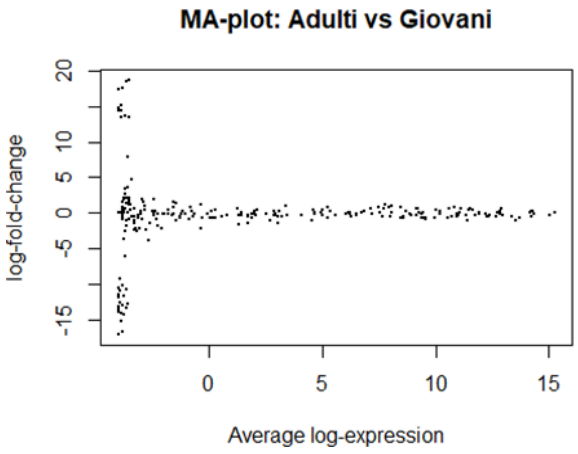
Non viene evidenziata alcuna differenza di abbondanza tra i microrganismi tra le classi d'età.

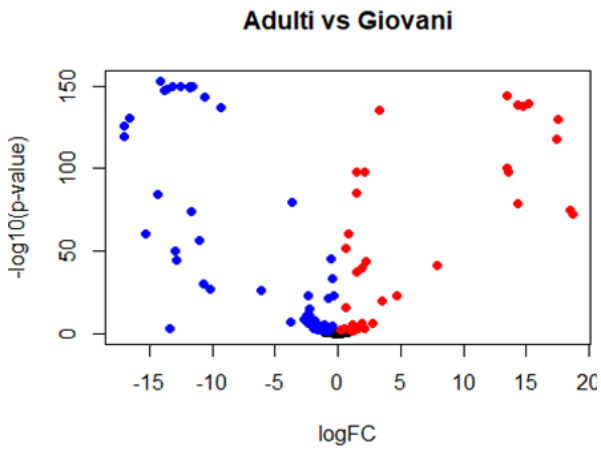
limma-voom + pesi ZINB

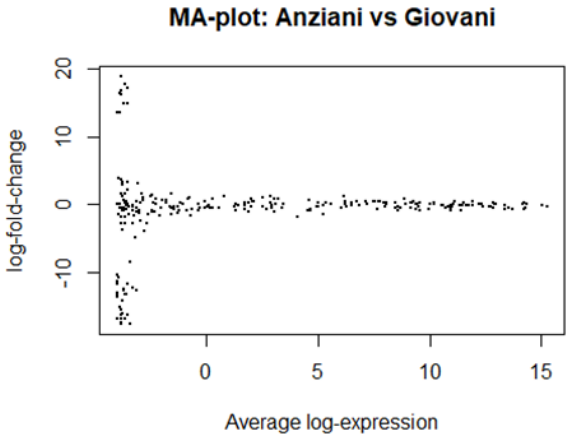
Risultano 68 microrganismi differenzialmente abbondanti tra le tre classi d'età.

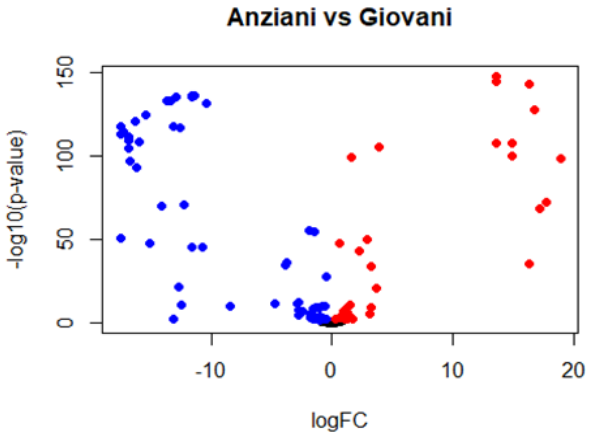
Utilizziamo i contrasti per eseguire i diversi confronti a coppie tra le classi d'età. L'analisi evidenzia 96 microrganismi differenzialmente abbondanti tra "Adulti" e "Giovani", 99 microrganismi nel confronto tra "Anziani" e "Giovani" e 94 microrganismi tra "Anziani" e "Adulti".

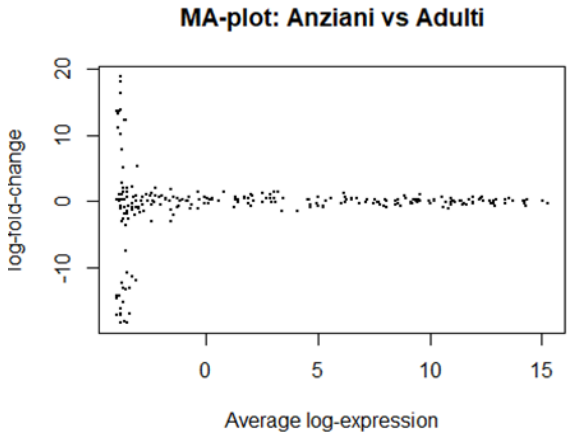
Per la visualizzazione dei risultati utilizziamo MA-plot e volcano plot.

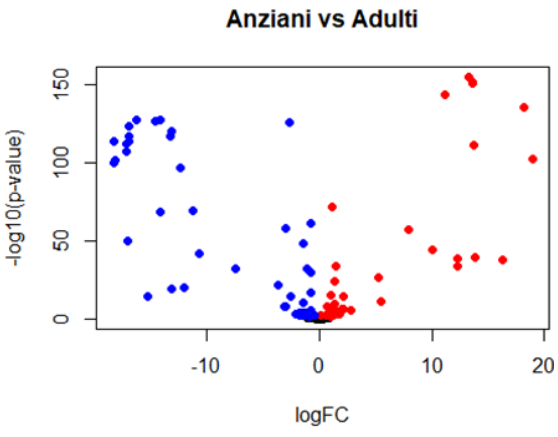




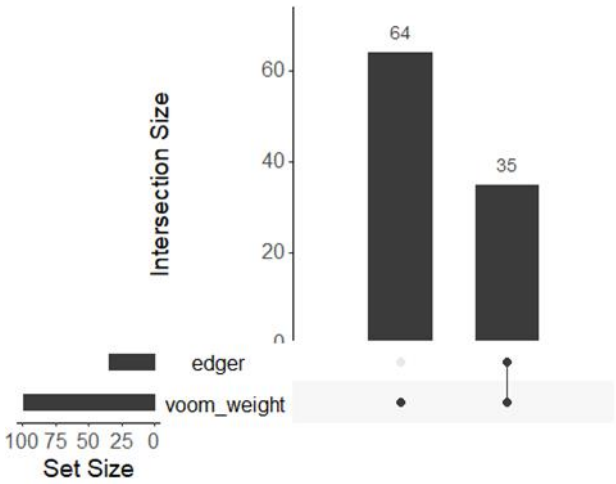


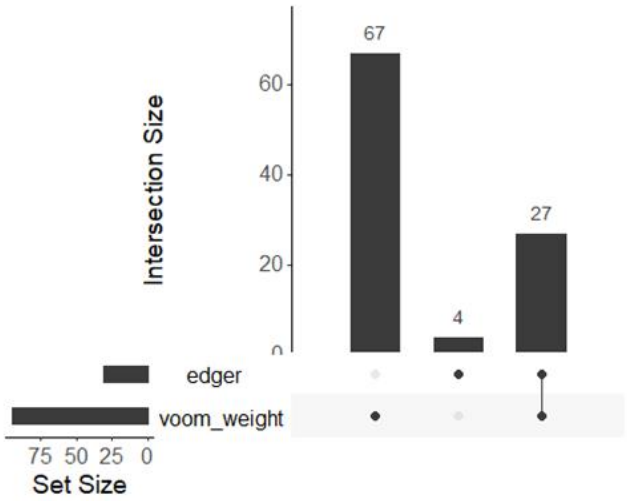






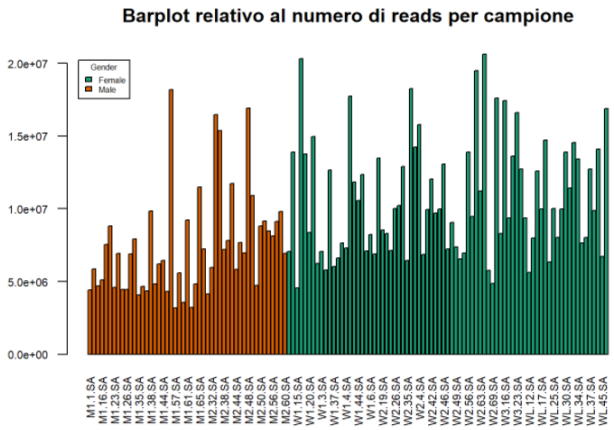
Quello che colpisce è l'elevato numero di taxa differenzialmente abbondanti per ogni contrasto (circa un terzo dei taxa considerati), che ci induce a prendere con cautela il risultato. Dato il numero elevato di taxa significativi, non tutti risultano biologicamente rilevanti, molti di questi infatti hanno log-fold-change molto basso. In tutti e tre i contrasti a coppie, comunque, si nota un certo bilanciamento nei taxa differenzialmente abbondanti tra le classi confrontate.

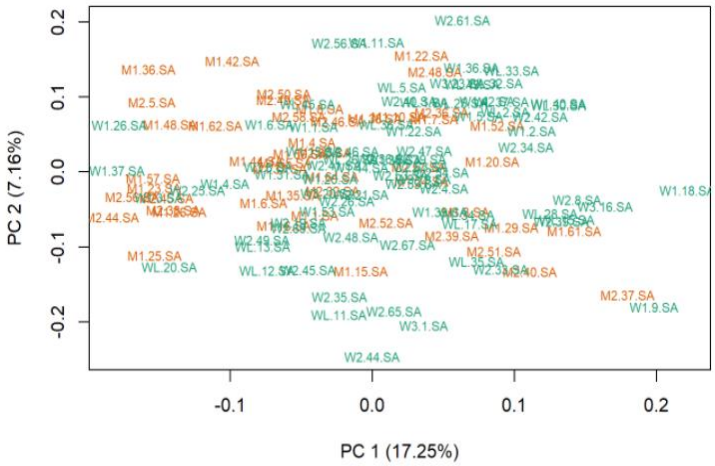




In conclusione quindi, per quanto riguarda il confronto tra “Anziani” e “Giovani”, risultano 35 microrganismi differenzialmente abbondanti in comune tra il metodo *edgeR* e il metodo *voom* con pesi dati dal modello ZINB, mentre per il confronto tra “Anziani” e “Adulti” 27.

Analisi per genere



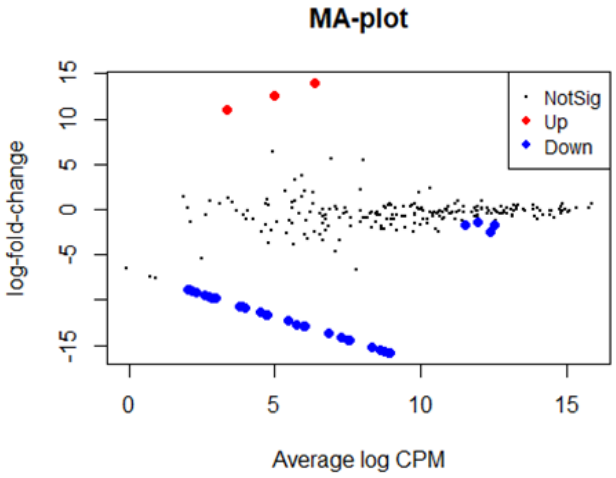


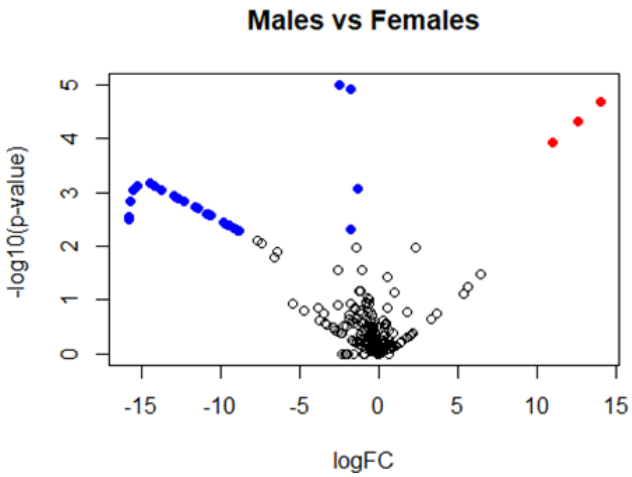
Nel grafico della PCA nessuna delle due componenti principali riesce a separare le due classi della variabile ‘gender’. Queste prime analisi sembrano mostrare che il genere non sia una variabile significativamente legata alla variabilità biologica del fenomeno in studio.

edgeR

I microrganismi differenzialmente abbondanti tra maschi e femmine risultano essere 37.

Osserviamo l’andamento dei taxa differenzialmente abbondanti nel confronto tra maschi e femmine mediante volcano plot e MA-plot.

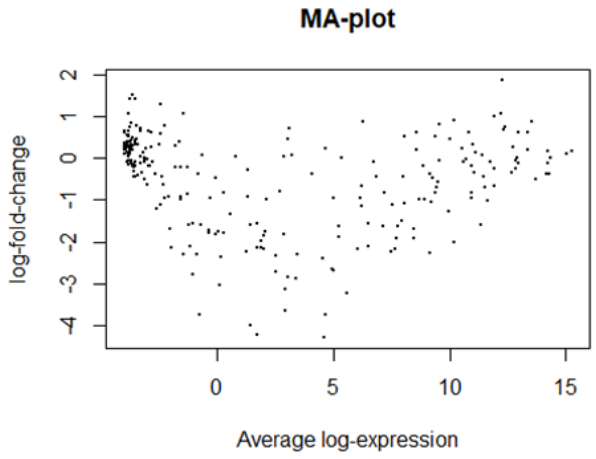


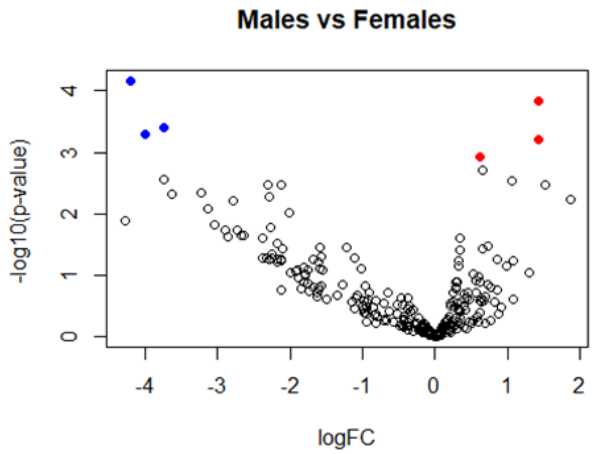


Dai grafici si osserva che non c'è un bilanciamento tra il numero di taxa differenzialmente abbondanti nelle due classi, ma c'è una sovrabbondanza nelle femmine. Quasi tutti i taxa differenzialmente abbondanti hanno log-fold-change elevati, se ne osservano anche 4 con log-fold-change modesti.

limma-voom

Si trovano 6 microrganismi differenzialmente abbondanti tra i due gruppi.





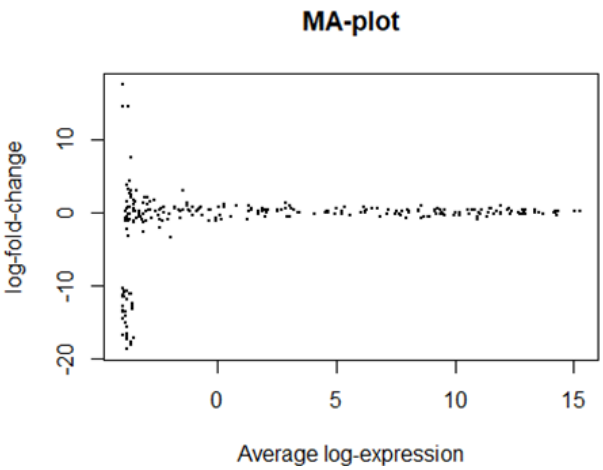
Dal volcano plot risulta che i taxa differenzialmente abbondanti tra i due gruppi sono bilanciati, inoltre si nota che quelli risultati significativi nella classe 'Female' hanno log-fold-change più elevati rispetto all'altra classe.

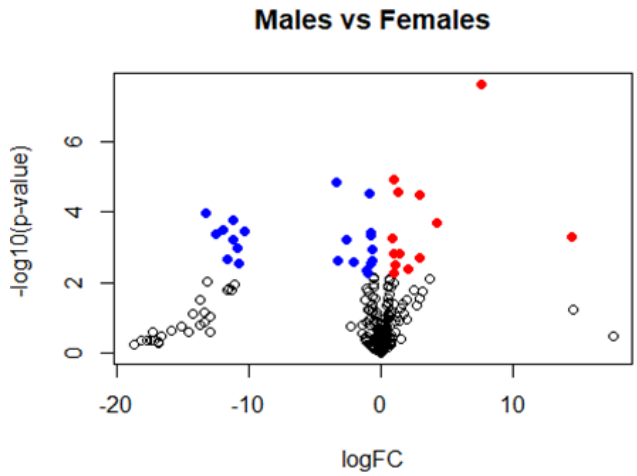
edgeR + pesi ZINB

Nessun microrganismo risulta differenzialmente abbondante tra le due classi.

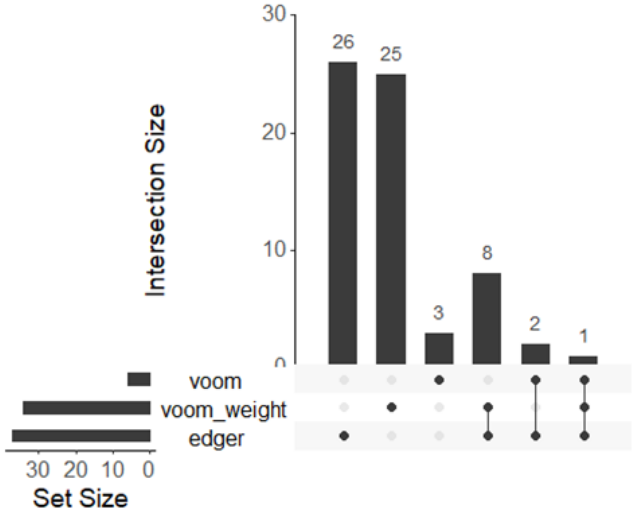
limma + pesi ZINB

Risultano 34 microrganismi differenzialmente abbondanti tra maschi e femmine.





Il numero di taxa differenzialmente abbondanti appare piuttosto bilanciato tra le due classi. Alcuni hanno log-fold-change modesti, mentre altri, soprattutto nelle femmine, hanno valori più elevati.



I tre metodi, *edgeR, limma-voom* e *limma- voom con pesi ZINB*, trovano in comune solo un solo microrganismo differenzialmente abbondante tra uomini e donne. Risultano, inoltre, 8 microrganismi in comune tra *edgeR* e *limma-voom con pesi ZINB* e solamente 2 tra i metodi *limma-voom* ed *edgeR*.

Conclusioni

Il profilo microbico salivare può essere influenzato da diversi fattori: età, sesso, stato di salute dentale, fumo e particolari condizioni cliniche. Il dataset di Brito *et al.* (2016) presenta una coorte piuttosto omogenea, tutti soggetti sani provenienti da villaggi delle isole Fiji, e non sono disponibili metadati relativi a variabili confondenti. Le uniche variabili presenti per tutti i 140 soggetti sono età e sesso.

Vista la numerosità campionaria abbiamo codificato la variabile età come categoriale, suddividendo in terzili ('giovani', 'adulti', 'anziani'), e ci siamo concentrati sui confronti a coppie specificando i tre contrasti. Solo i metodi *edgeR* e *limma voom con pesi ZINB* hanno portato a risultati significativi in 2 e 3 contrasti, rispettivamente. Notiamo che il metodo *limma voom con pesi ZINB* identifica per ogni contrasto circa un terzo dei taxa considerati come differenzialmente abbondanti. Nonostante il risultato sia sospetto, nel seguito proviamo comunque a dare un'interpretazione ai taxa in comune tra i due metodi. Nel confronto Anziani vs Giovani sono stati evidenziati 35 taxa differenzialmente abbondanti, di cui 25 più abbondanti nei Giovani rispetto agli Anziani; mentre nel confronto Anziani vs Adulti 27 taxa, di cui 19 più abbondanti negli Adulti rispetto agli Anziani.
Non avendo trovato in letteratura uno studio con dati comparabili ai nostri (soggetti sani, popolazione non occidentalizzata, età variabile da 1 a 69 anni con IQR 16-45), per discutere i nostri risultati abbiamo utilizzato dati sull'evoluzione del microbioma nella prima infanzia (soggetti sani e meno influenzati dall'ambiente occidentalizzato rispetto agli adulti), facendo riferimento alla review di Xiao J *et al.* (2020). Nella nostra variabile Giovani (age<=22), sono compresi infatti anche

17 bambini fino a 7 anni. Confrontando il genere dei taxa più abbondanti nei Giovani rispetto agli Anziani con quelli che caratterizzano l'evoluzione del microbioma della prima infanzia riportati nella review, se ne trovano diversi in comune (e.g. Streptococcus, Gemella, Veillonella, Fusobacterium, Lactobacillus, Actinomyces e Leptotrichia). Gli stessi generi di batteri vengono rilevati dalle nostre analisi come più abbondanti anche negli Adulti rispetto agli Anziani. Pur essendoci evidenza in letteratura che i soggetti anziani (età>64) hanno profili microbici diversi rispetto agli adulti, Lira-Junior R *et al.* (2018), non troviamo corrispondenza diretta tra i taxa da noi evidenziati come significativi e quelli pubblicati. Probabilmente in parte è dovuto al fatto che i soggetti del nostro studio sono piuttosto giovani e che anche quelli definiti 'Anziani' nella nostra categorizzazione sono per lo più adulti con età inferiore a 65 anni.
Osserviamo anche tre taxa più abbondanti negli Anziani, sia rispetto ai Giovani che rispetto agli Adulti: *Streptococcus phage Sfi19* (batterio del genere *Streptococcus* che va in direzione opposta rispetto a quanto visto prima), *Fujinami sarcoma virus* (virus del sarcoma di Fujinami) e *Murine osteosarcoma virus* (che però è un virus murino).

Per quanto riguarda l'analisi di abbondanza differenziale rispetto alla variabile 'gender', i metodi testati hanno portato a risultati poco sovrapponibili. Un unico taxon, *Fusobacterium gonidiaformans*, risulta differenzialmente abbondante tra maschi e femmine, in comune tra tre dei quattro metodi. Questo ceppo però si trova generalmente nel tratto urogenitale ed è stato solo occasionalmente rilevato in altre siti. In letteratura, nel microbioma del cavo orale non sono state riportate differenze in questo ceppo legate alla variabile sesso, mentre sono state trovate differenze in *Fusobacterium nucleatum* e *Fusobacterium periodonticum*, Henne *et al.* (2018).

Visti i risultati poco robusti rispetto ai diversi metodi applicati e a volte controversi, riteniamo che le criticità incontrate nell'analisi del dataset – in particolare la presenza di una percentuale molto elevata di conteggi nulli, la sovradispersione, il numero di conteggi molto diversi tra i campioni e la mancanza di variabili confondenti da includere nei modelli – non siano state affrontate in modo ottimale e che i modelli proposti per l'analisi di espressione differenziale non siano stati in grado di evidenziare le reali differenze che potrebbero esserci tra i due gruppi. Potrebbe anche essere che il livello di risoluzione a cui abbiamo lavorato sia troppo fine e si potrebbe considerare di rieseguire l'analisi ad un livello superiore nella gerarchia tassonomica, ad esempio a livello di Specie (codice R per estrarre i dati a livello di Specie riportato nell'allegato). Un'ulteriore analisi potrebbe essere quella di applicare metodi che stimano le variabili confondenti direttamente dai dati, ad esempio l'approccio RUV, il quale si basa sull'assunzione che esistono dei microrganismi, chiamati "controlli negativi", la cui espressione non è associata alla variabile biologica di interesse. In questo caso specifico, non disponendo di controlli negativi, si potrebbero calcolare i "controlli negativi empirici" a partire da un dataset con campioni e condizioni simili a quelle in studio.

Confronto con i risultati dell'articolo originale

I risultati delle nostre analisi non sono direttamente confrontabili con quelli presentati nel lavoro di Brito et al. (2016), in cui il focus è rivolto alla caratterizzazione della funzione e della distribuzione dei geni mobili nel microbioma di due popolazioni molto diverse tra loro. Nell'articolo, infatti, vengono analizzati e confrontati sia il microbioma orale che quello intestinale di abitanti di villaggi nelle Fiji e di abitanti di metropoli nel Nord America.

Il nostro progetto riguardava solo un sottoinsieme dei dati originali relativo al microbioma del cavo orale degli abitanti dei villaggi delle Fiji, con lo scopo di eseguire un'analisi di abbondanza differenziale tra soggetti di età diverse e di sesso diverso.

Pur non potendo fare un confronto diretto, nell'articolo stesso si fa riferimento al fatto che la composizione del microbioma tra villaggi diversi delle Fiji è simile.

References

Calgaro, M., *et al.* (2020) Assessment of statistical methods from single cell, bulk RNA-seq, and metagenomics applied to microbiome data, *Genome Biol* **21**, 191.

Brito IL, *et al.* (2016), Mobile genes in the human microbiome are structured from global to individual scales, *Nature,* **535**(7612):435-439.

Xiao J, *et al.* (2020), Oral microbiome: possible harbinger for children's health, *International Journal of Oral Science*, **12**, 12.

Lira-Junior R, *et al.* (2018), Salivary microbial profiles in relation to age, periodontal, and systemic diseases, *PLoS One*,**13**(3):e0189374.

Henne K., *et al.* (2018), Sex-specific differences in the occurrence of Fusobacterium nucleatum subspecies and Fusobacterium periodonticum in the oral cavity, *Oncotarget*, **9**(29): 20631–20639.