Università degli Studi di Firenze

Scuola di Scienze Matematiche, Fisiche e Naturali

Corso di Laurea in Scienze Biologiche

Analisi bioinformatica di dati da metabarcoding di microbiomi delle vie aeree di pazienti fibrocistici

Bioinformatic analysis of airway microbiome metabarcoding data from Cystic Fibrosis patients

Relatore:
Prof. Marco Bazzicalupo

Candidato:
Chiara Vanni

Correlatore:
Dott. Alessio Mengoni

Anno Accademico 2013/2014

Indice

Capitolo 1

Introduzione

In natura i microrganismi esistono in comunità complesse che contengono molte specie diverse, interagenti tra loro e con il microambiente che occupano (Salipante et al., 2013). Una comunità microbica è costituita da diverse popolazioni microbiche che possono interagire tra loro formando consorzi. In un determinato ambiente i componenti e il numero di cellule microbiche presenti dipendono dalla disponibilità di risorse e dalle condizioni presenti in quel determinato habitat. Il corpo umano non fa eccezione, infatti diversi microrganismi possono coabitare lo stesso organo, o apparato, e/o agiscono in concerto formando biofilm polimicrobici. Lo studio microscopico del corpo umano sano ha dimostrato che le cellule microbiche superano in numero le cellule umane di dieci a uno. Fino a poco tempo fa, però, questa abbondante comunità di microbi associati all'uomo risultava in gran parte non studiata, e la sua influenza sullo sviluppo, la fisiologia, l'immunità e la nutrizione umana rimaneva quasi del tutto sconosciuta. Il Progetto Microbioma Umano (HMP), iniziativa dei National Institutes of Health statunitensi, è stato istituito con lo scopo di generare risorse per ricerche che permettano la caratterizzazione completa del microbiota umano e l'analisi del suo ruolo nella salute e nella malattia. Il microbioma può essere definito come la somma dei geni di tutti i microrganismi presenti in un determinato organismo superiore. Se si presuppone che i prodotti genici di tali microrganismi possano interagire con le cellule umane, in quest'ottica, un essere umano va concepito come un tutt'uno composto da cellule umane e microbiche (Turnbaugh et al., 2007). L'HMP, che si propone dunque di caratterizzare tutti questi geni di origine esterna, si pone come continuazione del Progetto Genoma Umano. Conoscere le funzioni del microbioma comporta un cambiamento di prospettiva: la nostra vita e la nostra salute risultano programmate non solo attraverso le sequenze del nostro DNA, ma dipendono anche dalle variazioni epigenetiche che il microbiota attua sull'espressione dei nostri geni: l'analisi metagenomica, ossia l'utilizzo di tecniche genomiche moderne per lo studio di una comunità microbica nell'ambiente naturale in cui si trova a vivere, si propone di studiare per quali funzioni codifichino tali geni microbici. In questa prospettiva il corpo umano appare costituito da una minoranza di cellule somatiche e da un meta-organismo composto da una miriade di cellule microbiche: siamo un aggregato di geni umani e geni microbici, il nostro metabolismo e quello delle specie che ci abitano si intrecciano, interagiscono ed evolvono parallelamente (Hattori and Taylor, 2009). La distribuzione di tale mondo, invisibile a occhio nudo, ma di importanza strategica riguarda, in modo particolare, la pelle, la bocca, l'esofago, lo stomaco, il colon, la vagina e l'apparato respiratorio; gli organismi che lo compongono possono essere batteri, funghi, protozoi, elminti e virus.
La composizione nelle singole specie microbiche varia molto tra individui diversi e all'interno dello stesso individuo la comunità tipica rappresenta un tratto distintivo e caratterizzante. Da tale nucleo di base individuale varie modifiche continuano a osservarsi nei

diversi stadi della vita o se si instaurano particolari condizioni patologiche.
Il Progetto Microbioma Umano sta cercando di stabilire se all'interno dell'elevata variabilità nelle sequenze del microbioma individuale sia possibile trovare una stabilità di funzione che garantisca un set base di reazioni biochimiche. L'obiettivo del IHMC (International Human Microbiome Consortium, progetto lanciato ufficialmente in occasione della riunione tenutasi a Heidelberg il 15-16 Ottobre 2008) è quello di lavorare sotto un comune insieme di principi e politiche per studiare e comprendere il ruolo del microbioma umano nel mantenimento della salute e il nesso con la malattia e di utilizzare tale conoscenza per migliorare la capacità di prevenire e curare le malattie.
L'utilizzo delle nuove tecnologie di sequenziamento genomico, applicate allo studio del microbiota intestinale e polmonare, potrà aprire la strada a nuove strategie terapeutiche e preventive nei pazienti affetti da malattie croniche, a partire dalla Fibrosi Cistica (FC), malattia genetica grave che coinvolge molti organi ed apparati e in oltre il 95% dei casi evolve in un'insufficienza respiratoria cronica, a causa di infezioni polmonari polimicrobiche. Questo è quanto è emerso dal recente simposio internazionale organizzato dall'Ospedale Pediatrico Bambino Gesù ("The Microbiota and Immunity in Human Diseases – An International Symposium"), che porterà alla creazione di un network italiano di ricerca pediatrica dedicata al "microbioma", l'insieme dei microrganismi presenti nel corpo umano. Questi progressi renderanno possibile lo studio delle funzioni fisiologiche svolte dal pool dei microbi che convivono con il nostro organismo, sia in condizioni di benessere clinico, che in presenza di malattia. Il comprendere come essi interagiscono con il nostro organismo, in particolare con l'apparato immunologico, aprirà la strada a nuove strategie terapeutiche e preventive, anche nei pazienti affetti da malattie croniche.

Questo lavoro di tesi si inserisce in un progetto della fondazione per la ricerca sulla Fibrosi Cistica, con l'obiettivo di indagare sul microbioma delle vie aeree dei pazienti fibrocistici che presentano un severo declino della funzione polmonare e non rispondono alla terapia convenzionale antimicrobica, a cui ha collaborato anche il dipartimento di Biologia dell'Università di Firenze (Department of Biology, University of Florence, Laboratory of Microbial Genetics).

1.1 Fibrosi Cistica

La FC è una malattia monogenetica che si trasmette con meccanismo autosomico recessivo ed è molto frequente nelle popolazioni di discendenza europea (Tobias et al., 2011). Nella popolazione caucasica [1] ne è affetto un neonato ogni 2500-2700 nati vivi, si parla quindi di una malattia rara [2], che tuttavia risulta classificata come una delle malattie genetiche, croniche ed evolutive, più frequenti in grado di accorciare la vita: nel mondo ne sono colpite oltre 70.000 persone. Grazie ai progressi della ricerca e delle cure, i bambini che nascono oggi con questa patologia hanno un'aspettativa media di vita di 40 anni ed oltre, mentre non superavano l'infanzia cinquant'anni fa, quando la malattia è stata scoperta e si è iniziato a curarla (O'Sullivan and Freedman, 2009).
La Fibrosi Cistica è determinata da mutazioni che insorgono nel gene CFTR (*Cystic Fi-*

[1] Caucasico è, in origine, un termine geografico che indica un qualcosa relativo alla regione del Caucaso, fra Europa orientale e Asia occidentale. Esso ha però col tempo acquisito altri significati, specialmente in occidente e negli Stati Uniti, dove è venuto a essere un sinonimo di persona di carnagione chiara, di discendenza europea.

[2] Malattia rara è considerata ogni malattia che ha, nella popolazione generale, una prevalenza inferiore ad una data soglia. L'Unione europea definisce tale soglia allo 0.05% della popolazione, ossia un caso su 2000 abitanti.

brosis Transmembrane Regulator), situato sul cromosoma 7, ad oggi se ne conoscono oltre 1800, la più frequente in tutte le popolazioni è la mutazione DF508 (o F508del). Il gene CFTR codifica per una proteina che regola il passaggio di elettroliti, cloro in particolare, e di acqua, dall'interno all'esterno delle cellule epiteliali, le quali rivestono molti organi del nostro organismo. La mutazione del gene determina la produzione di una proteina CFTR difettosa o addirittura ne impedisce la sintesi, con la conseguenza che le secrezioni risultano povere di acqua, perciò dense e poco scorrevoli; infatti si parla di "muco viscido", da cui il nome passato della patologia: "mucoviscidosi".
La FC è una patologia multiorgano, gli organi che presentano le più importanti conseguenze cliniche sono i bronchi e i polmoni. All'interno dei bronchi il muco tende a ristagnare e questo predispone a infezioni respiratorie. Nell'85-90% dei casi è colpito il pancreas, che è ostruito dalle sue stesse secrezioni e non svolge l'azione normale di riversare nell'intestino gli enzimi per la digestione dei cibi. Altri organi interessati sono l'intestino, il fegato, i dotti deferenti nel maschio. Seppure il grado di coinvolgimento differisca anche notevolmente da persona a persona, la persistenza dell'infezione polmonare, che causa il danneggiamento progressivo del tessuto polmonare, è la maggior causa di morbilità e mortalità nei pazienti fibrocistici. L'alterazione del trasporto degli ioni sodio e cloro a livello delle cellule epiteliali esita in una marcata riduzione di liquido sulla superficie delle vie respiratorie, con la conseguenza di alterare la normale viscosità del muco; la produzione di secrezioni di densità particolarmente aumentata determina quindi un'alterazione del trasporto del muco stesso da parte delle cellule cigliate dell'apparato respiratorio con alterazione della normale clearence batterica. La scarsa eliminazione e la conseguente persistenza dei germi a livello delle vie aeree dà origine ad un continuo stimolo infiammatorio che progressivamente porta all'insorgenza del danno polmonare tipico della FC. Attualmente per monitorare la funzionalità polmonare nella FC e in altre malattie polmonari si utilizza la percentuale del volume espiratorio forzato previsto in 1 secondo %FEV1 (*Forced expiratory volume in the 1st second*) (Taylor-Robinson et al., 2012); valore usato in spirometria, che indica il volume di aria espirata nel corso del primo secondo di una espirazione massima forzata e indica il grado di pervietà delle grandi vie aeree. Diversi studi hanno dimostrato che il valore del FEV1 di alcuni pazienti fibrocistici subisce un grave calo nonostante i trattamenti antibiotici (Sanders et al., 2010). Il tasso di riduzione annuo della percentuale del FEV1 è un valore predittivo del rischio di morbilità e mortalità nella FC. Interpretare il significato delle variazioni nel tempo del valore %FEV1 richiede una più approfondita comprensione della composizione della comunità microbica delle vie aeree e la possibilità di valutare se un agente patogeno sia solo un marker della gravità della malattia o un collaboratore indipendente alla perdita della funzione polmonare. Recenti ricerche hanno rivelato che le infezioni delle vie aeree fibrocistiche sono polimicrobiche e che il microbioma, cioè l'insieme dei microrganismi, dotati di specifico patrimonio genetico, presenti nelle vie aeree dei pazienti fibrocistici, come entità collettiva, può contribuire ai processi fisiopatologici connessi con la malattia cronica delle vie aeree (LiPuma, 2010). La FC è classicamente caratterizzata da un modello di infezione batterica "cronica". I ceppi batterici coinvolti (sia Gram-positivi, come *S. aureus*, che i Gram-negativi non fermentanti, cui appartengono *P. aeruginosa*, *H. influenzae* e *A. xylosoxidans*) acquisiscono caratteristiche di persistenza, spesso si organizzano in biofilm, mostrano una ridotta velocità di divisione cellulare e una spiccata ipermutabilità del loro materiale genetico. Tutto ciò può costituire una spinta evolutiva verso l'instaurarsi di meccanismi di resistenza batterica, sia ereditaria che non ereditaria. È stato ipotizzato che la composizione della comunità batterica possa essere un miglior indicatore della progressione della malattia, rispetto alla presenza di patogeni opportunisti "stand-alone", cioè che stanno da soli, e non dipendono dalla presenza di altre specie batteriche (Rogers et al., 2010). In seguito a trattamenti antibiotici, all'aumen-

tare dell'età del paziente e del declino dell'attività polmonare si osservano cambiamenti nella struttura della comunità batterica delle vie aeree e alcuni generi batterici sembrano svolgere un ruolo diretto nel guidare questi cambiamenti e/o risultano buoni biomarker per l'esacerbazione polmonare [3] (Carmody et al., 2013). Data l'importanza della funzione polmonare per la salute del paziente fibrocistico, per estensione è altrettanto importante capire la complessità del microbiota polmonare fibrocistico e tale conoscenza va considerata come il primo passo nel portare avanti la terapia per i pazienti che presentano un grave declino della funzione polmonare.

1.2 Metodi di studio delle comunità microbiche

1.2.1 Metodi colturali e metodi coltura-indipendenti

In generale i metodi per l'analisi delle comunità microbiche possono essere suddivisi in due grandi sottoinsiemi:
1.Metodi convenzionali: analizzano le caratteristiche fenotipiche dei microrganismi grazie a tecniche basate sulla loro coltivazione, quali l'arricchimento per specifiche attività metaboliche o l'isolamento basato su terreni di coltura selettivi (per esempio resistenza ad antibiotici o capacità degradative di particolari fonti di carbonio).
2.Metodi molecolari: permettono di analizzare le caratteristiche biochimiche e metaboliche o genetiche delle popolazioni costituenti le comunità prese in esame. In particolare, nel caso degli studi genetici, si ricorre all'analisi di appositi geni "marcatori" che consentono di distinguere tassonomicamente i membri delle comunità stesse, o ancora, di evidenziare la presenza di particolari attività metaboliche. In tal caso, è possibile inoltre dividere tutte queste tecniche in due grandi gruppi, ossia: tecniche basate sull'amplificazione PCR e tecniche biochimiche, indipendenti dall'amplificazione PCR.
I metodi di microbiologia tradizionale si basano su caratterizzazioni fenotipiche, quali morfologia, richieste nutrizionali e profili fermentativi. L'identificazione dei microrganismi avviene attraverso l'isolamento di colture pure, seguito da test che analizzano alcune caratteristiche morfo-fisiologiche. In pratica, le cellule microbiche presenti in un determinato ambiente vengono fisicamente separate dalla matrice in cui si trovano e messe a crescere in un terreno artificiale dove si moltiplicano e, mediante la semina su piastre, possono essere contate; successivamente sono caratterizzate per le proprietà metaboliche e l'identità tassonomica. Questi metodi hanno il vantaggio di essere veloci e poco costosi, ma i microrganismi non coltivabili non vengono individuati e spesso si hanno distorsioni dovute alla crescita più veloce o predominante di un microrganismo rispetto ad un altro; inoltre alcuni ceppi possono entrare in competizione con gli altri quando co-coltivati, o può essere presente un numero troppo alto di specie, che impedisce un workup completo (Kirk et al., 2004).
Il basso potere discriminatorio di questi approcci coltura-dipendenti e l'ambiguità di determinati risultati ne invalidano l'impiego in diverse situazioni. Negli ultimi anni sono stati introdotti numerosi metodi molecolari, principalmente indipendenti dalla coltivazione dei microrganismi, per lo studio della diversità delle comunità microbiche ambientali. Questi nuovi metodi hanno affiancato le metodologie più tradizionali, quali la coltivazione su piastra e l'analisi delle caratteristiche metaboliche dei ceppi microbici in coltura, consentendo così di superare i limiti imposti dalla coltivazione dei microrganismi. Tra i metodi molecolari più usati ci sono: il FAME (*Fatty acids methyl esters*) che si basa sulla caratterizzazione

[3] Nella Fibrosi Cistica, più che di "infezioni respiratorie" si usa parlare di "esacerbazioni polmonari", con riferimento agli episodi di esacerbazione, cioè peggioramento di uno stato morboso, di una infezione polmonare cronica.

dei lipidi di membrana dei diversi microrganismi presenti nella comunità microbica; il DGGE (*Elettroforesi su gel a gradiente denaturante*) e il TGGE (*Temperature Gradient Gel Electrophoresis*) che sono tecniche impiegate per la caratterizzazione della struttura e lo studio della dinamica delle popolazioni microbiche; la tecnica FISH di ibridazione, DNA-DNA *in situ* con sonde fluorescenti, il DNA microarray e il T-RFLP (*Terminal restriction fragment length polimorfism*), metodica che si basa sulla digestione dei prodotti di PCR seguita dalla separazione elettroforetica dei frammenti ottenuti. Gli svantaggi di questi metodi riguardano soprattutto i *bias* dovuti alla PCR, quando presente, in alcuni casi la necessità di grandi campioni di DNA e l'individuazione delle sole specie dominanti.
Recentemente, i costi relativamente bassi e la rapidità crescente del sequenziamento del DNA, unitamente ai progressi compiuti negli approcci computazionali per l'analisi di complessi dataset, hanno fatto sì che l'analisi del gene 16S rRNA sia ormai comunemente impiegata ai fini di classificare e determinare rapporti filogenetici tra microrganismi (Petti et al., 2005). Il sequenziamento del DNA ribosomiale può fornire, infatti, una classificazione tassonomica più definitiva, per molti organismi, rispetto a quella risultante da approcci *coltural-based*, richiedendo inoltre minor manodopera.

1.2.2 Metodi di sequenziamento massivo

In generale per sequenziamento si intende l'identificazione della sequenza nucleotidica di un acido nucleico fornito in input alla strumentazione. Il sequenziamento del DNA nello specifico permette di determinare l'ordine con cui le basi si susseguono nelle molecole di DNA contenute nei campioni in esame. I campioni di DNA vengono estratti dalle cellule in analisi e sequenziati per ottenere dei frammenti a sequenza nota, detti reads, di lunghezza variabile in base alla tecnologia del sequenziatore utilizzato.
Sono state ideate diverse strategie per ottenere la sequenza nucleotidica del DNA. I primi metodi, tra cui quello ideato da Allan Maxam e Walter Gilbert nel 1973, erano piuttosto complicati e si basavano su modificazioni chimiche del DNA e sul conseguente taglio in posizioni specifiche. Una svolta si ebbe nel 1977 con la prima pubblicazione di una strategia enzimatica tuttora diffusissima, sviluppata da Frederick Sanger e collaboratori; il cosiddetto metodo dei terminatori di catena, o metodo Sanger (Sanger et al., 1997), che è stato quello di riferimento fino all'avvento dei metodi high throughput dei primi anni 2000 ed è tutt'oggi utilizzato in alcuni ambiti, quali i test di paternità e la biologia forense. Tuttavia, si tratta di un metodo costoso, che richiede molto tempo per sequenziare genomi complessi, e infatti vi si stanno sostituendo le tecnologie NGS (Next-Generation Sequencing) o ultra-high-troughput, che permettono una maggiore mole di dati in output, costi e tempi ridotti, a discapito, tuttavia, della lunghezza delle reads e di una minor accuratezza nella lettura delle basi.
La tecnologia di sequenziamento Sanger standard si limita a sequenziare un singolo gene da un singolo esemplare a ogni esecuzione. Il sequenziamento massivo, al contrario, permette di sequenziare separatamente le singole molecole di DNA e, quindi, accetta miscele di geni, e specie (Kircher and Kelso, 2010). Prima delle tecniche di sequenziamento NGS, i sequenziatori, mediante 4 reazioni distinte, producevano 4 "tracce", attraverso le quali si risaliva alla sequenza del DNA. Nei sequenziatori NGS un'unica reazione produce 4 "tracce" con diverse lunghezze d'onda, ognuna caratteristica di una determinata base (cromatogramma). La capacità throughput delle tecniche di sequenziamento di nuova generazione e il costo ridotto e accessibile ormai da molti centri di ricerca, hanno portato allo sviluppo di questo nuovo campo d'indagine, il quale sfrutta la potenza della tecnica di sequenziamento dei frammenti di DNA ottenuti da popolazioni virali e batteriche, senza però passare da fasi di purificazioni o colture.

Il grande sviluppo tecnologico dei sequenziatori è stato associato alla nascita del progetto genoma umano, che aveva come scopo quello del sequenziamento completo del genoma umano. Partendo dall'idea che si diffuse sul finire degli anni '80, il progetto ha coinvolto molti paesi e centri di ricerca internazionali. L'enorme sforzo ha portato alla pubblicazione sulle maggiori riviste scientifiche dei risultati sia del consorzio pubblico che del consorzio privato guidato da Craig Venter che aveva puntato sulla forte automatizzazione del processo utilizzando dei sequenziatori ABI in cui la manodopera richiesta era minima (Venter et al., 2001). Nel 2005, seguendo di pochi anni la conclusione del progetto genoma umano, viene lanciato sul mercato il sequenziatore 454 pyrosequencing della Roche (Li et al., 2012). Questo insieme ad altri due sequenziatori, Illumina e ABI SOLiD, costituiscono le tecniche di seconda generazione, che permettono esperimenti di tipo massivo. Tali techinche di high throughtput hanno la capacità di produrre una quantità di dati impensabili con la tecnica Sanger.

Pirosequenziamento
Ai frammenti di DNA, denaturati per renderli a singolo filamento, sono attaccati due adattatori alle estremità. La fase di amplificazione utilizza delle sfere dal diametro di $28\mu m$, la cui superficie è ricoperta da oligonucleotidi complementari ad uno degli adattatori dei frammenti. Frammenti e sfere sono posti in contatto permettendo la formazione di un legame tra loro. Le sfere con legati i filamenti sono quindi messe, insieme ad enzimi amplificatori, nella fase acquosa di un'emulsione di acqua in olio per far avvenire l'amplificazione clonale tramite PCR (PCR a emulsione o em-PCR) eliminando così la necessità di un clonaggio biologico dei frammenti stampo. Si tratta di un processo ciclico in cui la DNA polimerasi sintetizza il filamento complementare, si denatura la doppia elica e vengono sintetizzati due nuovi filamenti complementari, fino ad ottenere decine di milioni di copie del frammento per ogni sfera. Terminata l'amplificazione, le sfere ricoperte dai nuovi filamenti sono poste sulla piastra di sequenziamento. Nel sequenziatore 454 Roche, la piastra (detta PicoTiterPlate) ha circa 1,5 milioni di pozzetti, il cui diametro ($44\mu m$) permette di accogliere una sola sfera ciascuno. Le sfere vi si depositano per gravità e sono ricoperte da numerose sfere più piccole legate a sulfurilasi, luciferasi, apirasi e luciferina, enzimi e substrati necessari per il processo di sequenziamento. Nel caso della tecnica 454 Roche si parla di pirosequenziamento (Margulies et al., 2005), il quale prevede il completamento iterativo del filamento e la simultanea lettura del segnale emesso dai nucleotidi incorporati. Si tratta di un processo ciclico, in cui un singolo dNTP viene aggiunto di volta in volta alla reazione. Se il dNTP è complementare al filamento stampo, la sua incorporazione causa il rilascio di un pirofosfato che viene prelevato dalla sulfurilasi e convertito in ATP. Quest'ultimo innesca la reazione luminosa della luciferasi e il segnale luminoso emesso viene captato e misurato da un sensore CCD. La quantità di pirofosfati, quindi di ATP, rilasciati e l'intensità della bioluminescenza registrata sono proporzionali al numero di nucleotidi, di uno stesso tipo, incorporati. Misurata l'intensità luminosa è quindi possibile determinare il numero di nucleotidi di uno stesso tipo che si susseguono nella sequenza oggetto di studio. Dopo la registrazione dell'emissione luminosa, i rimanenti nucleotidi liberi vengono degradati e viene fornito il dNTP successivo, proseguendo nella reazione di polimerizzazione. A ogni ciclo è aggiunta una singola specie di nucleotidi (A, T, C, G) e viene portata a saturazione; al ciclo successivo sarà introdotta una delle specie restanti, e così via ripetendo la sequenza di introduzione per centinaia di cicli (Apolloni, 2011/2012).

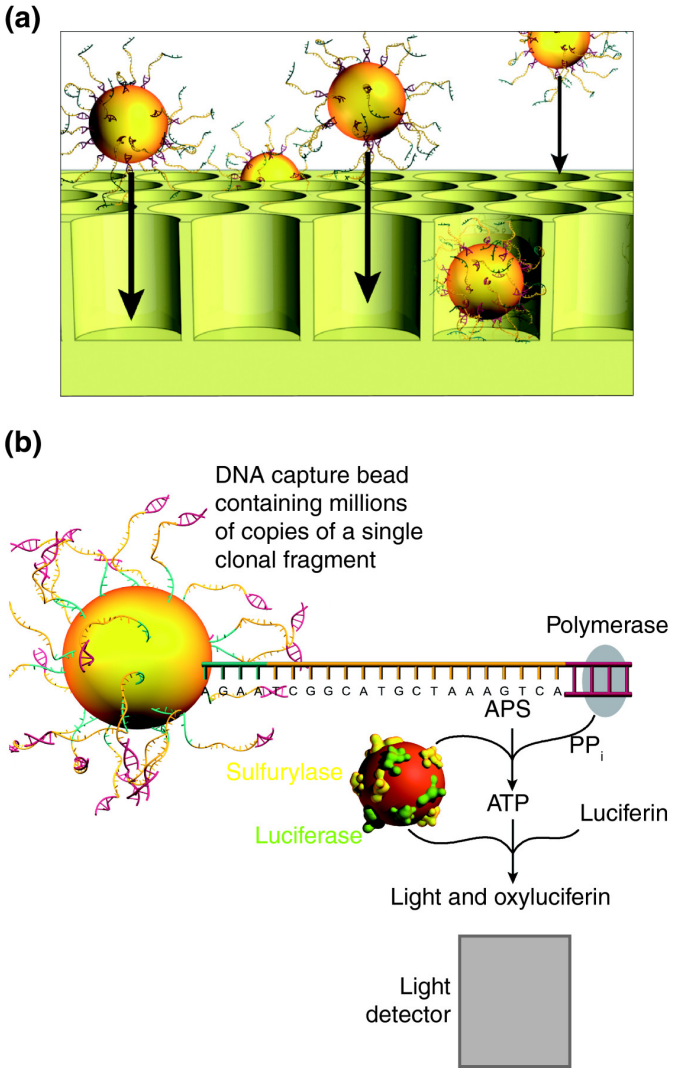


Figura 1.1: Schema della procedura di sequenziamento 454. Nella figura vengono mostrati in maniera sintetica i vari passaggi caratteristici della tecnica di sequenziamento 454 della Roche

Proprietà

Il tasso di errore medio è compreso tra 10^{-3} e 10^{-4}, superiore al tasso di errore del metodo Sanger (generalmente compreso tra 10^{-5} e 10^{-6}). Una rilevante fonte di errore, oltre alle fasi di amplificazione dei frammenti, è la preparazione delle sfere, che possono trasportare copie di sequenze diverse, producendo reads che non rappresentano la sequenza reale. Il sequenziatore ultra massivo 454 GS FLX+ Titanium della Roche permette di sequenziare un milione di sfere per run (corsa), determinando sequenze di lunghezza media fra 400 e 600 bp, che varia in base al numero di cicli di sequenziamento, limitati per il mantenimento dell'efficienza di polimerasi e luciferasi. Il throughput è di 450 Mb e si ha un'accuratezza del 99,9% alla quattrocentesima base e superiore per le basi precedenti. Le più recenti versioni del sequenziatore (GS FLX Titanium XL+) hanno un output di ben 14 G per run in 10 ore e producono reads lunghe fino a 1000 bp con un throughput di 700Mb (http://my454.com/products/gs-flx-system/index.asp).
I recenti progressi nelle tecnologie di sequenziamento del DNA hanno creato opportunità di sequenziamento ad una profondità e ampiezza senza precedenti (Margulies et al., 2005) e il sequenziamento multiplex è emerso come una strategia importante per il sequenziamento parallelo di molti campioni diversi. L'analisi in parallelo delle sequenze di più di

un campione è detta in gergo multiplexing. Per poter procedere al multiplexing, è necessario aggiungere ai frammenti di ciascun campione una sequenza specifica, detta etichetta o sequenza barcode. La sequenza barcode consente di identificare in modo univoco tutti i frammenti di DNA di uno stesso individuo. Le librerie di DNA dei diversi individui così preparate possono poi essere unite in un'unica provetta per la reazione di sequenziamento in multiplexing. Tutte le reads ottenute durante la reazione conterranno anche la sequenze barcode, che viene effettivamente amplificata e sequenziata come le sequenze di DNA a cui è attaccata. Identificando le sequenze barcode è poi possibile procedere al de-multiplexing delle reads, cioè alla loro separazione in librerie di campioni, sulla base dell'appartenenza ai diversi individui. Le reads così separate vengono poi allineate con le sequenze di riferimento dei database fino ad ottenere la sequenza reale di ciascun campione analizzato. Il multiplexing nel sequenziamento di ampliconi, che è ampiamente eseguito per le indagini di diversità di geni 16S rRNA o dei geni funzionali, può essere effettuato sia legando il barcode e gli adattatori di sequenziamento agli ampliconi creati con primer PCR "convenzionali", cioè primer che consistono solo della sequenza modello-specifica, o più semplicemente utilizzando lunghi oligonucleotidi (Fusion Primer) che, oltre ai primer PCR convenzionali, includono già dei tags in 5' con barcode e adattatori di sequenziamento, in modo da eliminare la fase di legatura. Un esempio di questo secondo metodo è l'approccio "One-way Reads", usato in questo progetto, descritto nel 454 Sequencing System-Guideline for Amplicon Experimental Design (454 Life Sciences Corporation). Questo metodo viene usato in esperimenti in cui è più efficiente sequenziare partendo da un solo primer, in modo cioè unidirezionale, massimizzando così la lunghezza delle reads, e in cui non è necessario il livello di accuratezza che si ottiene invece con il sequenziamento bidirezionale.

Primers

Il metodo, chiamato "One-Way Reads", è simile a quello del "Basic design" descritto nella 454 Sequencing System-Guidelines for Amplicon Experimental Design (454 Life Sciences Corporation), ma utilizza il processo di emPCR Lib-L, che consiste in una semplice reazione di PCR, in cui si utilizza il campione di DNA di interesse e una coppia di Fusion Primers, che andranno a legarsi alle "Capture Beads (sfere)" Lib-L, invece che alle "Capture Beads" Lib-A, normalmente utilizzate per le librerie di ampliconi.[Pag. 22] I primer utilizzati per generare la "Amplicon libraries" del progetto sperimentale "One-Way Reads" sono disegnati in modo che l'orientamento della read sia noto; sono composti da tre parti, fuse insieme come mostrato nella Fig. n°(primer usati). La porzione all'estremità 5' è una 30-mer [4], la cui sequenza è dettata dalle esigenze del sistema di sequenziamento 454 e viene usata per ibridare con la "Lib-L" delle DNA Capture Beads e per legare i corrispondenti primers di amplificazione emPCR e primer di sequenziamento. Inoltre la parte 5' deve sempre terminare con una "library key" "TCAG". Esistono due tipi di tali primer, denominati "Primer A" e "Primer B", che consentono il sequenziamento direzionale; tuttavia la reazione "Lib-L" supporta solo il sequenziamento a partire dalla fine del Primer A. La porzione in 3', invece, è progettata per legarsi con una specifica sequenza su entrambi i lati del target d'interesse del campione di DNA, delineando così i margini dell'amplicone che sarà prodotto. In aggiunta vengono poi usati dei MIDs (Multiplex Identifiers), ovvero delle sequenze barcode, per identificare meglio gli ampliconi che protendono da campioni diversi. I MIDs devono essere inseriti subito dopo la "library key". Nel caso degli Adattatori Lib-L i MIDs vengono inseriti solo nel Primer A.

[4]Il numero di fronte al suffisso -mer indica un numero di subunità, se applicato al DNA specifica il numero di basi in una determinata sequenza.

Figura 1.2: Componenti dei Fusion Primers del progetto sperimentale "One Way Reads" per il sequenziamento di ampliconi (454 Life Sciences Corporation).

Output del sequenziamento

Le reads in uscita dal sequenziatore sono sequenze in *Standard Flowgram Format* (SFF), un formato binario che corrisponde ai cromatogrammi. Questi file sono convertiti in vari formati utilizzando gli algoritmi *base caller*, che, data una serie di picchi cromatografici, sono in grado di risalire alla corretta sequenza di DNA che li ha generati ed assegnare ad ogni base una certa "qualità". Il più usato è il software Phred che legge i file in uscita dal sequenziamento del DNA, chiama le basi e assegna un valore di qualità per ogni base chiamata. Il valore di qualità di ogni base è calcolato prendendo in considerazione i principali parametri relativi alla forma ed all'ampiezza dei picchi di intensità luminosa registrati dal CCD del sequenziatore; la qualità così ottenuta è strettamente correlata alla probabilità di ottenere quella determinata base in quel punto del cromatogramma. Il valore di qualità Q, che viene assegnato, è un numero intero associato logaritmicamente a P, cioè alla probabilità che la corrispondente chiamata di base sia incorretta: $Q = -10 \log_{10}(P)$; in questo modo, ad esempio, un Q=20 corrisponde ad una probabilità di errore dell' 1%. I valori di Q variano solitamente tra 0 e 40.
I punteggi Phred sono ormai uno standard de facto per la rappresentazione della qualità delle basi delle reads di sequenziamento, tuttavia un'ambiguità presente in questi tipi di file è il fatto che ogni azienda usa un proprio codice per indicare la qualità, riferendosi al tipo di strumento con cui è stato effettuato il sequenziamento.

Trimming

Il sequenziamento 454 tipicamente produce reads con qualità multo buona nella parte iniziale e via via sempre minore verso la fine. Tipicamente le reads 16S derivano da ampliconi ottenuti mediante PCR da una coppia di primers (FWD e REV). La read può coprire in modo completo o parziale l'amplicone, in entrambi i casi la sua lunghezza è soggetta a variazioni. Nel caso di una copertura completa la lunghezza delle reads varia perchè oscilla quella degli ampliconi, a causa delle regioni ipervariabili del gene (che sono diverse tra i vari microrganismi). Per quanto riguarda la copertura parziale, invece, le lunghezze variano in base a dove viene effettuato il taglio di qualità; poiché la qualità tende a cadere verso la fine della read, le ultime basi tendono ad essere meno affidabili e questo può produrre un allineamento inaffidabile verso la fine delle sequenze. Inoltre una bassa qualità può significare una sequenza diversa da quella reale.
Per ottenere buoni risultati nelle fasi di dereplicazione e clustering, le reads dovrebbero essere globalmente allineabili, e non presentare gap terminali negli allineamenti a coppie

di sequenze strettamente correlate. Se la qualità delle reads è complessivamente buona è possibile mantenere l'intero amplicone (ampliconi *full-length*), in questo caso il trimming è necessario solo se ci sono basi aggiuntive oltre al primer, ad esempio la sequenza di raccordo, cioè l'adattatore. Reads con qualità troppo bassa verso la fine, invece, vengono tagliate ad una lunghezza fissa, o usando una soglia di qualità, che migliora la qualità dei dati e massimizza la quantità di dati conservati.

1.2.3 La metagenomica e il metabarcoding

Con l'analisi metagenomica, che si basa sull'utilizzo di tecniche genomiche moderne, è possibile studiare le comunità microbiche direttamente nell'ambiente in cui vivono, evitando così il problema del prelevamento e della coltivazione in laboratorio. Queste tecniche permettono di determinare la presenza di particolari microrganismi attraverso l'estrazione e il sequenziamento del loro DNA, che può essere poi caratterizzato rivelando la natura del microrganismo che lo conteneva. Dall'analisi delle sequenze di DNA, in particolare quelle che codificano per le sequenze degli RNA ribosomali, è infatti possibile risalire fino alle diverse specie presenti nel campione. Sempre utilizzando il DNA estratto dall'ambiente in cui vive la comunità microbica è inoltre possibile studiare la complessità del microambiente, che corrisponde a quante specie diverse lo abitano, anche senza la necessità di identificare le singole specie.
L'insieme del materiale genetico presente in un campione ambientale, definito metagenoma, è rappresentativo delle specie che lo popolano. Come detto precedentemente anche l'uomo è un ambiente estremamente complesso, contiene circa 1014 cellule batteriche, il cosiddetto "microbioma umano", le quali hanno una profonda influenza sulla fisiologia dell'organismo, sulla nutrizione, e risultano cruciali per la nostra salute. La metagenomica oggi rende possibile la caratterizzazione della composizione e della dinamica di popolazione della comunità microbica umana, e le interazioni cooperative, o antagoniste, con le cellule e i tessuti umani.

In generale lo scopo principale della metagenomica è capire quali sono gli organismi presenti all'interno di un determinato ambiente, qual è la loro relativa abbondanza e quale il loro contributo genetico, individuando la capacità funzionale e l'eterogeneità dei geni, sia intraspecie che interspecie.

Il metabarcoding: una classificazione basata sulla morfologia e l'osservazione degli organismi spesso è risultata difficile e problematica (Tanabe and Toju, 2013), per questo ad essa si è affiancato un metodo alternativo denominato "DNA metabarcoding" (Taberlet et al., 2012). Il principio alla base della metodica deriva dal contributo di Carl Woese il quale ha introdotto l'approccio molecolare come standard per l'identificazione di organismi procarioti: grazie alle prime applicazioni di questi studi su sequenze di geni ribosomali (rRNA) sono stati scoperti gli Archaea; lo studio della variabilità di marcatori molecolari si è poi ampiamente diffuso per le analisi di popolazione, includendo svariati marcatori: rRNA, allozimi, microsatelliti, AFLP [5], ecc.
Il DNA metabarcoding nasce da un'iniziativa di Paul D.N. Hebert della Università di Guelph, Ontario, Canada. L'idea di base del DNA barcoding sta nell'esistenza di un così detto "barcoding gap", per il quale la variazione delle sequenze nucleotidiche all'interno di una specie è minore rispetto a quella che intercorre tra sequenze nucleotidiche di specie diverse. Seppure non rivoluzionario dal punto di vista metodologico, la grande novità del DNA barcoding è la scala di analisi e la standardizzazione del metodo. Per capire l'importanza

[5] AFLP: Amplified fragment length polymorfism, si tratta di un metodo estremamente sensibile per individuare polimorfismi a livello del DNA, descritto per la prima volta da Vos et al. nel 1993.

assunta da questo nuovo approccio basti pensare che esiste il "Barcode of Life", un progetto che è volto a creare un sistema universale, economico e veloce per l'identificazione delle specie accessibile anche dai non specialisti. Il metabarcoding può essere descritto anche come un metodo per valutare la biodiversità che combina due tecnologie: la tassonomia basata sul DNA e le tecniche di sequenziamento massivo. La scelta del tag nucleotidico, utilizzata per l'identificazione delle varie specie, deve avere una serie di caratteristiche:

- la sequenza contenente il tag deve essere distribuita universalmente nel gruppo in esame;
- la regione deve variare con una velocità commisurata alla distanza evolutiva da misurare;
- la regione deve essere specie specifica;
- deve essere standardizzabile;
- deve essere fiancheggiata da sequenze conservate di circa 20 bp per i primers della PCR.

Tutte queste caratteristiche sono state trovate in vari tags che vengono usati in maniera specifica per i diversi organismi: per i batteri sono state scelte le regioni ipervariabili del gene 16S rRNA. Quest'ultimo è altamente conservato tra Archaea e Bacteria e l'assegnazione tassonomica è possibile grazie alla presenza di nove regioni ipervariabili (V1-V9), che contengono una diversità di sequenza sufficiente per permettere la classificazione dei microbi. Inoltre, poiché queste regioni sono fiancheggiate da regioni conservate, è possibile effettuare una PCR con primers universali. Infine il gene 16S rRNA è piuttosto corto, 1542 nucleotidi, e questo rende il suo sequenziamento veloce e poco costoso. Dato che nessuna singola regione ipervariabile, presa da sola, è sufficientemente diversificata per differenziare tra tutte le specie batteriche, solitamente si ricorre al targeting di due o tre regioni ipervariabili in un'unica read, rispettando i limiti sulla lunghezza delle reads, imposti dalle proprietà del sequenziatore (Salipante et al., 2013).

In un esperimento di metabarcoding inizialmente si effettua la raccolta del campione che può provenire dal suolo, dalle acque, ma anche dall'uomo (nel nostro caso si tratta di espettorato). Si procede con l'estrazione del DNA dal campione e l'amplificazione di un determinato gene, tassonomicamente informativo, attraverso la reazione di PCR con primer universali. Il prodotto della PCR, detto amplicone, viene quindi sequenziato attraverso le varie tecniche di sequenziamento massivo (NGS). Infine si hanno le analisi post-sequencing volte ad identificare i taxa e i geni funzionali presenti. Poiché ogni individuo di ogni specie ha contribuito con molti segmenti/frammenti di DNA, ognuno dei quali è stato poi copiato molte volte con la PCR, il set di dati di output deve essere ridotto utilizzando programmi che consentono di raggruppare le sequenze in unità tassonomiche operative, o OTUs. Infine, viene scelta una sequenza rappresentativa da ogni OTU e viene assegnata una tassonomia mediante il confronto con sequenze depositate in database specifici, quali Ribosomal Database Project, SILVA o Greengenes. Si possono, in conclusione, stimare i parametri di biodiversità, mediante, ad esempio, gli indicatori ecologici quali l'indice di Simpson o di Shannon ed effettuare ulteriori analisi statistiche.

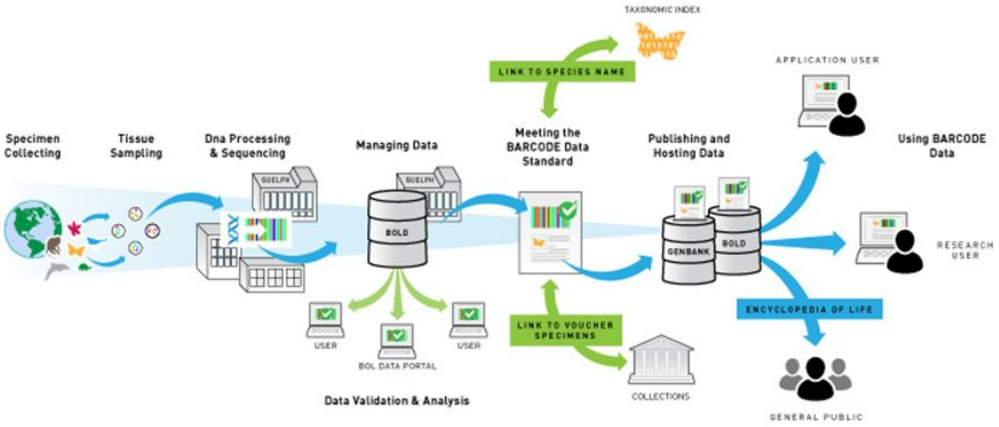


Figura 1.3: La figura mostra i vari step necessari allo sviluppo di un'analisi di metabarcoding.

Il costo e il tempo impiegati sono ridotti rispetto ad una classificazione secondo l'osservazione del microrganismo, tuttavia questo approccio presenta ancora una serie di limitazioni dovute a diversi tipi di errori. Il primo errore è riconducibile alla necessità dell'amplificazione tramite PCR. L'uso di polimerasi proofreading può diminuire errori di sequenziamentoi. Un' altra grossa limitazione è la necessità di database di alta qualità e nel tentativo di risolvere queste difficoltà, nasce il progetto "Barcode of Life" (Taberlet et al., 2012). Ma la maggiore limitazione che viene attribuita al metodo è il fatto di basarsi su di un solo carattere, una sola regione del DNA, e di conseguenza non considerare l'intero contesto, portando così a possibili errori. Un problema sorge poi dalla quantità di dati che devono essere gestiti una volta che il sequenziamento è stato concluso, si parla infatti di decine o migliaia di gigabasi di short reads (Zhang et al., 2011). La fase di assemblaggio delle reads è estremamente lenta e se viene applicato un algoritmo più veloce, questo va a discapito della RAM in maniera proporzionale alla grandezza del genoma da assemblare (Scholz et al., 2012). I laboratori inoltre devono sviluppare sistemi di stoccaggio e management e creare tools informatici per analizzare i dati ottenuti. Una possibile soluzione è l'utilizzo di sistemi in cui l'utente può usare un sistema operativo virtuale per processare i dati utilizzando le competenze messe a disposizione da strutture altamente qualificate.
La tecnologia utilizzata in metabarcoding è l'high-throughput sequencing.

1.2.4 Analisi ecologiche della comunità microbica

Analisi della biodiversità
Per visualizzare se il campionamento è stato effettuato in modo soddisfacente e i campioni sono biologicamente spiegati, si possono costruire delle cosiddette curve di rarefazione, che sono grafici del numero di taxa in funzione del numero di campioni. Nel nostro caso le curve di rarefazione aiutano a stimare se le comunità batteriche sono state campionate correttamente, ovvero se si è ottenuto un numero sufficiente di reads per ogni campione. Una curva con pendenza marcata indica che una grossa frazione di biodiversità non è stata rilevata; se invece la curva raggiunge un plateau, significa che il campionamento è stato esaustivo e un campionamento più intenso produrrebbe solo pochi taxa supplementari ("Ganter Homepage". Tnstate.edu. Retrieved 2013-08-16).
Considerando campioni di N sequenze, la costruzione delle curve di rarefazione (se ne ottiene una per ciascun campione) fornisce il numero atteso di taxa che si accumulano via via

che il numero di sequenze per campione aumenta da 1 a N. Questo si ottiene ricampionando ripetutamente e casualmente l'insieme di tutti i campioni, e riportando in grafico il numero di taxa presenti in ciascun campione, in funzione del numero di sequenze per campione; in questo modo l'insieme di campioni viene progressivamente rarefatto (Gotelli and Colwell, 2001).
Anche in studi di metagenomica si può studiare la biodiversità dei campioni, ma per confrontare i risultati di più campionamenti di microbiomi diversi, si può operare un confronto statistico solo se sono stati usati gli stessi metodi di campionamento ed è stato raccolto un uguale numero di campioni aventi dimensioni tra loro confrontabili. In genere, e nel nostro caso, questa condizione non si verifica; quindi per effettuare un confronto statistico valido è necessario uniformare le dimensioni del campione e questo viene fatto mediante un'analisi di rarefazione.
Per correggere le differenze tra le dimensioni dei campioni non si può semplicemente dividere il numero di taxa presenti per il numero di sequenze campionate, perché questo equivarrebbe a supporre che il numero di taxa diversi aumenta linearmente con il numero di sequenze presenti, che non è sempre vero. Quindi se si hanno n assegnazioni nel campione di dimensioni minori, per l'analisi di rarefazione si procede estraendo dei sub-campioni di n assegnazioni dai campioni di dimensioni maggiori, per evitare distorsioni si ripete il pescaggio un numero sufficiente di volte (noi abbiamo scelto di ripeterlo 1000 volte), e infine si calcola il numero medio di taxa (OTUs) in tali sub-campioni. Questa media può essere paragonata con il numero di taxa effettivamente presenti nel campione con meno assegnazioni e possono essere calcolate una varianza e una deviazione standard per meglio giudicare quanto siano significative le eventuali differenze.

Gli indici di diversità utilizzati sono l'indice di Richness, l'indice di Evenness e l'indice di Shannon-Weaver, che sono stati calcolati sui campioni rarefatti, ripetendo la fase di rarefazione 1000 volte.
Indice di Richness: quantifica quanti *tipi* (o taxa) diversi sono contenuti nel dataset di interesse. La ricchezza in specie (indicata solitamente con S) corrisponde al numero di specie diverse rappresentate in una comunità ecologica, paesaggio o regione, si tratta di un conteggio delle specie, e non tiene conto delle loro abbondanze o delle loro relative distribuzioni di abbondanza. Al contrario, la diversità di specie (biodiversità) prende in considerazione sia la ricchezza di specie (richness), che l'equitabilità delle specie (evenness).
Indice di Evenness: in ecologia l'equitabilità esprime il grado di omogeneità col quale gli individui sono distribuiti nelle varie specie che compongono una comunità. Matematicamente è definito come un indice di diversità che misura quanto un comunità sia equa numericamente.
Indice di Shannon-Weaver: indice molto usato in ecologia, conosciuto anche come indice di Shannon-Wiener o semplicemente indice di Shannon. E' stato introdotto da Claude E. Shannon, "il padre della teoria dell'informazione", per quantificare l'entropia, cioè la quantità di incertezza o di informazione in stringhe di testo (Shannon, 2001).Si usa in statistica nel caso di popolazioni con un numero infinito di elementi. In ecologia è usato come misura della biodiversità e tiene conto sia del numero di specie presenti in ciascun campione che dell'uniformità della distribuzione degli individui all'interno delle varie specie. Si usa in statistica nel caso di popolazioni con un numero infinito di elementi.

Analisi della varianza
L'analisi ANOVA e viene usata in statistica, quando si confrontano contemporaneamente due o più medie. Si tratta di un metodo che fornisce valori e risultati che possono essere testati per determinare se esiste una relazione significativa tra diverse variabili. Il nome

ANOVA deriva dal fatto che, per confrontare due o più medie, di fatto vengono confrontate le varianze. L'analisi ANOVA comprende un insieme di tecniche statistiche, facenti parte della statistica inferenziale, che permettono di confrontare due o più gruppi di dati confrontando la variabilità interna a questi gruppi con la variabilità tra i gruppi.

Analisi multivariata della varianza
L'analisi MANOVA viene applica in statistica, quando si hanno più variabili dipendenti, (a differenza dell'ANOVA che include solo una variabile dipendente). Sui dati è stata eseguita una analisi MANOVA, per ispezionare le differenze nella distribuzione della comunità batterica in relazione alle condizioni cliniche del paziente; in particolare sono stati considerati il volume espiratorio forzato (FEV1) e le condizioni del paziente (stabile o in grave declino). Il metodo MANOVA serve per determinare se le variabili dipendenti vengono significativamente influenzate dalle variazioni delle variabili indipendenti e se ci sono interazioni tra le variabili dipendenti, così come tra quelle indipendenti (Stevens, 2002). MANOVA a differenza dell'analisi univariata ANOVA, utilizza la varianza-covarianza tra le variabili per testare la significatività statistica delle differenze tra le medie; la covarianza viene inclusa perché le variabili sono probabilmente correlate ed è necessario prendere questa correlazione in considerazione quando si esegue il test di significatività. Ovviamente, se dovessimo prendere la stessa misura due volte, allora non avremmo appreso nulla di nuovo. Se, invece, prendiamo una misura correlata, otteniamo alcune nuove informazioni, ma la nuova variabile conterrà anche informazioni ridondanti che saranno espresse dalla covarianza tra le variabili.
L'analisi MANOVA è più efficace quando le variabili dipendenti sono moderatamente correlate (0.4-0.7). Se le variabili dipendenti sono troppo altamente correlate si presume che possano essere misure (diverse) della stessa variabile.

Canonical Component Analysis
La CCA ha per obiettivo lo studio delle relazioni di interdipendenza tra i due gruppi di variabili e permette di individuare due nuovi gruppi di variabili artificiali non correlati al loro interno e massimamente correlati tra loro. Questa analisi può quindi fornire una visione globale dei dati, comprimendo le variabili in 2 o 3 dimensioni, che sono effettivamente in grado di contenere le caratteristiche dominanti dei dati; in questo modo può fornire una migliore comprensione dei sistemi biologici noti o sconosciuti.
In biologia, la CCA è un metodo multivariato per indagare sui rapporti tra le associazioni biologiche delle specie e il loro ambiente. Il metodo è stato introdotto da Harold Hotelling nel 1936 (Hotelling, 1936), ed è stato progettato per estrarre gradienti ambientali sintetici da insiemi di dati ecologici. L'analisi della componente canonica consente di trovare una combinazione lineare tra due vettori al fine di avere la massima correlazione fra loro. In altre parole, se abbiamo due vettori X = (X1, ..., Xn) e Y = (Y1, ..., Ym) di variabili aleatorie, tra cui ci sono correlazioni, allora l'analisi di correlazione canonica troverà combinazioni lineari di Xi e Yj che hanno la massima correlazione tra loro (Hardle and Simar, 2007).

Capitolo 2

Scopo del lavoro

La Fibrosi Cistica (FC) è una malattia monogenetica, caratterizzata da una progressiva perdita della funzione polmonare, a causa di infezioni polimicrobiche; conoscere la composizione del microbioma delle vie aeree può, quindi, contribuire in modo significativo a migliorare il corso e l'esito della malattia. Scopo principale del lavoro di tesi è stato valutare la composizione del microbioma delle vie aeree in un gruppo di pazienti fibrocistici, utilizzando le nuove tecniche di sequenziamento e l'analisi bioinformatica, alla ricerca di patogeni diversi da quelli abitualmente isolati in cultura con i metodi tradizionali, in modo da mettere a punto nuove strategie antinfettive.
Mediante l'analisi di questi dati da metabarcoding si può ottenere un'accurata identificazione dei batteri delle vie aeree dei pazienti fibrocistici, e una migliore comprensione delle implicazioni cliniche di agenti patogeni nuovi e/o emergenti nella popolazione FC e del contributo del microbioma delle vie aeree FC al declino del valore del FEV1.
Tutto questo al fine di ottenere una visione più ampia delle specie microbiche coinvolte nella malattia, di caratterizzare i cambiamenti nella comunità batterica correlati al declino della funzione polmonare e di identificare eventualmente nuovi predittori di tali cambiamenti e biomarcatori per il trattamento e la gestione delle infezioni batteriche nei pazienti con Fibrosi Cistica.

Capitolo 3

Materiali e metodi

3.1 Decrizione dataset

Per questo studio sono stati arruolati, tra Settembre 2012 e Aprile 2013, 78 pazienti, che, in quel periodo, frequentano tre Centri FC italiani: l'Ospedale Pediatrico Bambino Gesù di Roma, il Centro FC dell'Ospedale pediatrico Meyer di Firenze, e l'Ospedale Giannina Gaslini di Genova. A tutti i pazienti era stata diagnosticata la FC secondo linee guida pubblicate (Doring et al., 2004). I pazienti sono stati trattati seguendo le disposizioni per il monitoraggio polmonare con almeno quattro controlli microbiologici all'anno.
I pazienti sono stati ritenuti idonei se potevano essere classificati come clinicamente stabili, privi di esacerbazione polmonare, nelle 4 settimane precedenti e durante il campionamento, ed e.v. antibiotico o terapia orale, sempre nelle 4 settimane precedenti al prelievo di espettorato. Il tasso annuo di declino del valore percentuale di FEV1 è stato usato per suddividere i pazienti. Il tasso di declino della funzione polmonare è stato determinato utilizzando la migliore percentuale del valore predetto di FEV1 (% FEV1) di ogni paziente nell'ultimo anno. Poiché il test di funzionalità polmonare non può essere generalmente eseguito con successo fino a quando i bambini non raggiungono i 6 anni di età, sono stati arruolati solo pazienti CF con età superiore a 6 anni. Inoltre sono stati esclusi dallo studio pazienti che presentavano *Burkholderia cepacia* isolata dall'espettorato negli anni precedenti.
Al fine di ottenere una più profonda conoscenza della composizione della comunità batterica delle vie aeree dei pazienti fibrocistici e per capire meglio le conseguenze cliniche di nuovi e/o emergenti patogeni nella popolazione FC, i campioni di DNA di 72 pazienti sono stati esaminati mediante il pirosequenziamento 454 delle regioni ipervariabili V3, V4, V5 del gene 16s rRNA. Sono stati esaminati due gruppi di pazienti fibrocistici: 37 "severe decline" (SD) pazienti fibrocistici con un grave declino della funzione polmonare, che mostrano un tasso di riduzione del FEV1 $> 5\%$ nell'ultimo anno, che non rispondono alla terapia antimicrobica e risultano liberi da esacerbazione e 35 pazienti, classificati invece come "stabili" (S), che non mostrano cambiamenti nella funzione polmonare e presentano un tasso di declino del FEV1 pari al valore medio di 1,44% (e non superiore a 1,5%). All'interno di entrambi i gruppi, sono stati esaminati tre sotto-gruppi di pazienti fibrocistici, in base al loro stato clinico espresso dal valore percentuale del FEV1: gruppo I, pazienti con funzione polmonare normale o malattia polmonare lieve (FEV1 $\geq 70\%$); gruppo II, pazienti con malattia polmonare moderata ($40\% \leq$ FEV1 $< 70\%$); gruppo III, pazienti con una grave malattia polmonare (FEV1 $< 40\%$). Il numero di campioni di ogni sottogruppo è riportato nella Tabella 3.1.

Tabella 3.1: Tabella campioni utilizzati

Variabili	N_campioni
Pazienti	72
Stabili (S)	35
In sostanziale declino (SD)	37
Gruppo I (FEV1 $\geq$ 70%)	26 (14 S e 12 SD)
Gruppo II (40% $\leq$ FEV1 $<$ 70%)	25 (14 S e 11 SD)
Gruppo III (FEV1 $<$ 40%)	21 (8 S e 13 SD)

Trattamento dei campioni

L'analisi della composizione della comunità batterica è stata effettuata su campioni di espettorato spontaneo (spontaneously expectorated sputum, SES), poiché questi rappresentano il metodo più utilizzato nel caso di pazienti fibrocistici (Rogers et al., 2010). I campioni di espettorato sono stati trattati subito con Sputolysin (Calbiochem, La Jolla, CA) per 15 minuti, in accordo con le istruzioni del produttore, e poi divisi equamente in 5 aliquote; tre delle quali sono state congelate e conservate a -85°C, per la successiva spedizione in ghiaccio secco al laboratorio di UO2 (Unità Operativa 2, Dip. di Chimica, Università di Firenze), per l'estrazione di DNA genomico totale e per le indagini molecolari.

Estrazione del DNA. Per il sequenziamento NGS (454 Roche) del gene barcode 16S rRNA, il DNA è stato estratto utilizzando il kit disponibile in commercio, QIAamp DNA Mini Kit, secondo le istruzioni del produttore. Quantità e purezza del DNA estratto sono stati controllati da NanoDrop (NanoDrop Technologies, USA), da Qbit fluorimetro (Invitrogen, Carlsbad, CA, USA) e mediante l'elettroforesi su gel.
Sequenziamento. La preparazione e il sequenziamento delle librerie di ampliconi sono piuttosto flessibili e consentono una vasta gamma di modelli sperimentali. In questo studio è stato utilizzato il seguente flusso di lavoro sperimentale per generare librerie di ampliconi :

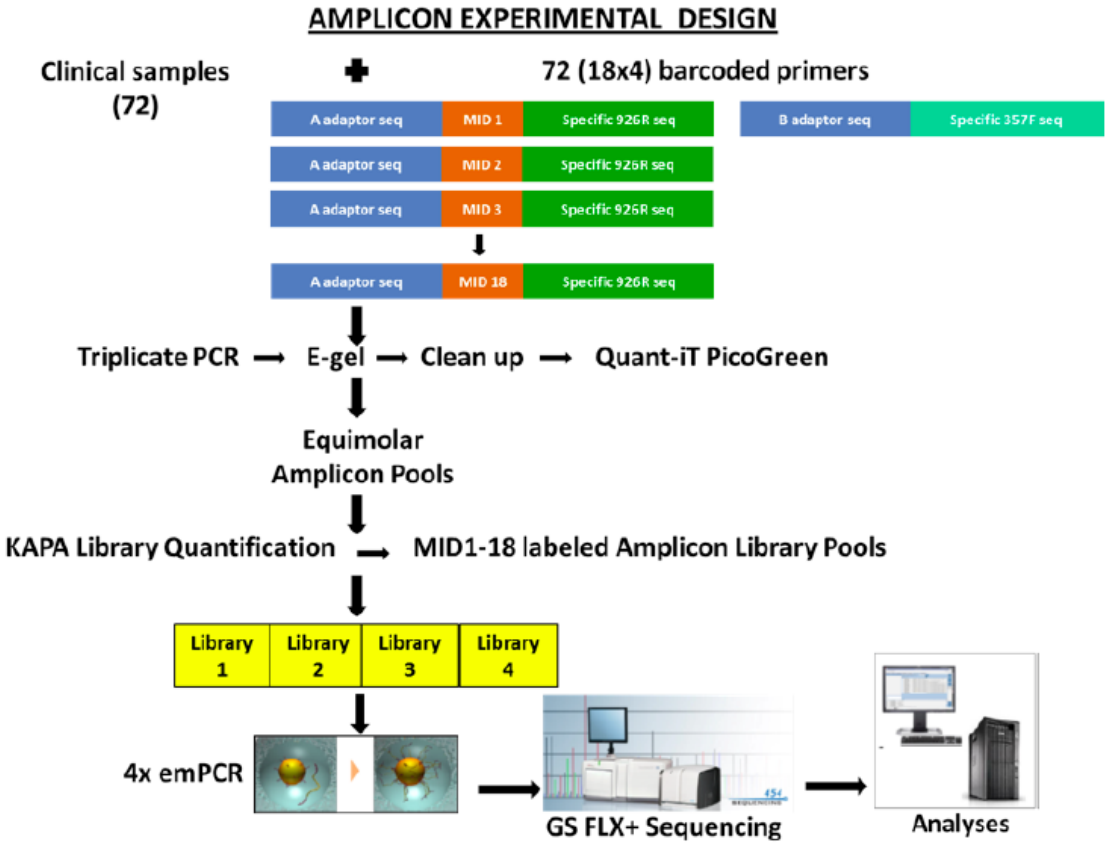


Figura 3.1: Progetto sperimentale per l'analisi degli ampliconi 16S rRNA mediante il pirosequenziamento 454, per lo studio del microbioma delle vie aeree di pazienti fibrocistici.

Dopo l'estrazione del DNA genomico batterico le regioni ipervariabili V3, V4, V5 del gene 16S rRNA sono state amplificate mediante PCR, per ciascuno dei 72 campioni. I campioni sono stati divisi in quattro gruppi, o run, (R1-R2-R3-R4), contenenti ciascuno 18 MIDs diversi (Tabelle 3). Per tutti i campioni è stato usato lo stesso foward-primer, il Fusion Primer 357F (5'-CCTACGGGAGGCAGCAG-3 '), modificato con l'aggiunta dell'adattatore 454 FLX-titanium "B", con sequenza 5'-CCTATCCCCTGTGTGCCTTGGCAGTCTCAG-3 '. Ciascuno dei 18 campioni in ciascun gruppo è stato amplificato utilizzando un reverse-primer universale per il gene batterico 16S rRNA, il Primer Fusion 926R (5'-CCGTCAATTCMTTTRAGT-3'), modificato con l'aggiunta di 18 sequenze uniche, "barcode", di sette nucleotidi ciascuna (MID 1-MID 18) e dell'adattatore 454 FLX-titanium "A" con sequenza 5'-CCATCTCATCCCTGCGTGTCTCCGAC-3' (Tabella 3.2).
Il sequenziamento multiplex permette di sequenziare parallelamente in una singola run più campioni, a ciascuno dei quali è associato un barcode unico, che ne permette la successiva identificazione.
Le sequenze dei barcode e degli adattatori, utilizzate nella presente attività di ricerca corrispondono a quelle riportate dal Consorzio del Progetto microbioma umano (Human Microbiome Project, HMP).

Tabella 3.2: Tabella primers e MID utilizzati

Oligo name	Sequence
B_357FW	5'-CCTATCCCCTGTGTGCCTTGGCAGTC-TCAG-CCTACGGGAGGCAGCAG-3'
A_MID1_926REV	5'-CCATCTCATCCCTGCGTGTCTCCGAC-TCAG-AAGCCGC-CCGTCAATTCMTTTRAGT-3'
A_MID2_926REV	5'-CCATCTCATCCCTGCGTGTCTCCGAC-TCAG-CAAGAAC-CCGTCAATTCMTTTRAGT-3'
A_MID3_926REV	5'-CCATCTCATCCCTGCGTGTCTCCGAC-TCAG-AGTTGGC-CCGTCAATTCMTTTRAGT-3'
A_MID4_926REV	5'-CCATCTCATCCCTGCGTGTCTCCGAC-TCAG-TATCAAC-CCGTCAATTCMTTTRAGT-3'
A_MID5_926REV	5'-CCATCTCATCCCTGCGTGTCTCCGAC-TCAG-AGGCGGC-CCGTCAATTCMTTTRAGT-3'
A_MID6_926REV	5'-CCATCTCATCCCTGCGTGTCTCCGAC-TCAG-CGGTATC-CCGTCAATTCMTTTRAGT-3'
A_MID7_926REV	5'-CCATCTCATCCCTGCGTGTCTCCGAC-TCAG-TGACGAC-CCGTCAATTCMTTTRAGT-3'
A_MID8_926REV	5'-CCATCTCATCCCTGCGTGTCTCCGAC-TCAG-ACAAGGC-CCGTCAATTCMTTTRAGT-3'
A_MID9_926REV	5'-CCATCTCATCCCTGCGTGTCTCCGAC-TCAG-AGACCTC-CCGTCAATTCMTTTRAGT-3'
A_MID10_926REV	5'-CCATCTCATCCCTGCGTGTCTCCGAC-TCAG-ATACCAC-CCGTCAATTCMTTTRAGT-3'
A_MID11_926REV	5'-CCATCTCATCCCTGCGTGTCTCCGAC-TCAG-TCGCGGC-CCGTCAATTCMTTTRAGT-3'
A_MID12_926REV	5'-CCATCTCATCCCTGCGTGTCTCCGAC-TCAG-ATCTTAC-CCGTCAATTCMTTTRAGT-3'
A_MID13_926REV	5'-CCATCTCATCCCTGCGTGTCTCCGAC-TCAG-AACCAGC-CCGTCAATTCMTTTRAGT-3'
A_MID14_926REV	5'-CCATCTCATCCCTGCGTGTCTCCGAC-TCAG-TTCGAGC-CCGTCAATTCMTTTRAGT-3'
A_MID15_926REV	5'-CCATCTCATCCCTGCGTGTCTCCGAC-TCAG-AAGGTGC-CCGTCAATTCMTTTRAGT-3'
A_MID16_926REV	5'-CCATCTCATCCCTGCGTGTCTCCGAC-TCAG-TCTTGGC-CCGTCAATTCMTTTRAGT-3'
A_MID17_926REV	5'-CCATCTCATCCCTGCGTGTCTCCGAC-TCAG-TAATCTC-CCGTCAATTCMTTTRAGT-3'
A_MID18_926REV	5'-CCATCTCATCCCTGCGTGTCTCCGAC-TCAG-TCACCTC-CCGTCAATTCMTTTRAGT-3'

In arancione si hanno le sequenze dei MID, in rosso la "library key", a sinistra di questa si trovano le sequenze degli adattatori "A" e "B" 454-FLX-Titanium, mentre a destra dei MID si hanno le sequenze specifiche per il gene 16S rRNA.

Dataset di sequenze ottenute con il pirosequenziamento

Il dataset ottenuto con il sequenziameno 454 dei campioni (precedentemente descritti), ammonta ad un totale di 201903 sequenze, suddivise tra le varie corse (run) del sequenziatore come mostrato nella Tabella 3.3:

Tabella 3.3: Sequenze ottenute con il pirosequenziamento

Run	Sequenze
1	63751
2	80431
3	57834
4	689

3.2 Analisi bioinformatica

3.2.1 Controllo ed elaborazione delle sequenze

Trimming: I sequenziatori NGS occasionalmente producono reads di scarsa qualità, in particolare si hanno errori nei pressi del sito del primer di sequenziamento e verso la fine delle sequenze più lunghe. Se tali errori non vengono rimossi mediante trimming, possono portare a distorsioni nelle fasi successive di allineamento e analisi delle sequenze, come ad esempio gap (lacune) terminali negli allineamenti cluster.
Le sequenze grezze del nostro dataset sono state trimmate effettuando la rimozione delle basi a basso quality score, utilizzando Streaming-trim 1.0, un algoritmo di trimming in grado di conservare quante più informazioni possibili (Bacci et al., 2014). Inoltre le sequenze filtrate sono state poi assemblate e sottoposte ad un altro controllo di qualità con PANDAseq, programma che assembla le *paired-end* reads molto rapidamente, corregge la maggior parte degli errori e assegna un punteggio di qualità (Masella et al., 2012).

BLAST: Il Basic Logical Alignment Search Tool, o strumento di ricerca di allineamento locale, è una algoritmo usato per comparare le informazioni contenute in sequenze nucleotidiche o proteiche. BLAST permette di confrontare una sequenza di interesse con un database di sequenze note e rende come output il rapporto di similarità tra la query (sequenza di interesse) e la sequenza contenuta nel database di riferimento. Ci siamo serviti di BLAST per vedere quanto dei primers completi si allineava con l'inizio delle sequenze del nostro dataset, le quali contenevano dei contaminanti, costituiti da residui del primer, in particolare dell'adattatore, rimasti in aggiunta al barcode (MID). Nel nostro caso il database è stato costruito con le sequenze dei 18 diversi primer "A', mentre le queries con le sequenze del nostro dataset.
Per creare il database e verificare gli allineamenti abbiamo utilizzato uno script ad hoc in bash (demultiplex.sh). BLAST non mette gap alle estremità, per questo abbiamo stabilito un certo grado di tolleranza (2 nucleotidi) sui limiti che abbiamo dato nello script e un numero di mismatch pari a 4, perchè di default BLAST ne considera già 2, per la presenza di una M e una R nella prorzione di primer modello-specifica per il gene 16S batterico.
In output si ottengono quattro file, corrispondenti alle 4 runs, contenenti solo la frazione di sequenze che produce allineamenti significativi. Nel file di output per ciascun risultato è indicato prima dell'allinemento vero e proprio: *Score*, cioè il punteggio dell'allineamento; *Expect*, che è l'E-value dell'allineamento; *Identities*, cioè la lunghezza dell'allineamento in questione e tra parentesi è indicata la risultante percentuale di identità su quella lunghezza di allineamento; *Gaps*, che indica il numero di gap presenti nell'allineamento; *Strand*, che indica l'orientamento della sequenza query rispetto alla sequenza del database con cui si allinea.

3.2.2 OTU clustering

Una volta assemblate, le sequenze filtrate sono state raggruppate in unità tassonomiche operative, o OTUs, con una soglia di identità del 97%, utilizzando UPARSE, un nuovo metodo di OTU clustering contenuto nel pacchetto di USEARCH [1], pubblicato recentemente su Nature Methods (Edgar, 2013).
Gli ampliconi di geni-marker vengono usati per comprendere la struttura delle comunità microbiche, ma spesso presentano un elevato livello di artefatti dovuti al sequenziamento

[1] USEARCH è uno strumento per l'analisi di sequenze, che offre algoritmi di ricerca e di clustering molto veloci, sviluppato da R.C. Edgar a partire dal 2001.

e all'amplificazione.

Passaggi per il raggruppamento in OTUs
Dereplicazione: consiste nella rimozione delle sequenze duplicate. In input si ha un insieme di reads in formato FASTA, spogliate delle sequenze non biologiche come i barcode. USEARCH supporta una dereplicazione di tipo *full-length*, *prefix* e *substrings*; la prima riguarda sequenze esattamente identiche, delle quali ne viene tenuta una sola, la seconda implica l'eliminazione di una sequenza A dal set se questa risulta essere il "prefisso" di una sequenza B e nella terza una sequenza A viene eliminata se risulta una sottostringa du una sequenza B. Nelle applicazioni del sequenziamento di nuova generazione, la *prefix dereplication* risulta utile nel caso di reads *quality filtered*, che sono spesso troncate all'estremità destra, in corrispondenza della quale i *quality scores* tendono a diminuire. In questa fase abbiamo specificato una lunghezza minima delle sequenze, per essere incluse nell'output, di 100 nucleolidi.

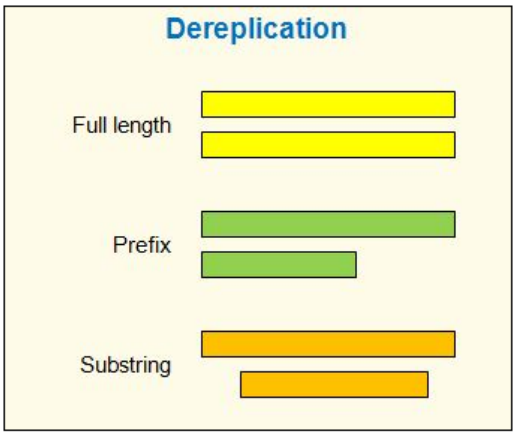


Figura 3.2: Casi in cui è necessario applicare una dereplicazione.

Ordinamento delle reads in base all'abbondanza: mediante il comado `sortbysize` le reads dereplicate vengono ordinate in ordine decrescente di abbondanza. L'ordinamento delle sequenze secondo l'abbondanza viene utilizzato per il clustering nel caso in cui le sequenze più abbondanti rappresentano anche i centroidi migliori. Questo si osserva spesso nel clustering di OTU 16S di reads prodotte mediante sequenziamento di nuova generazione. In tale caso le sequenze più abbondanti sono probabilmente anche le sequenze più accurate biologicamente, mentre le reads rare o i singleton (reads uniche) hanno maggiore probabilità di contenere errori di sequenziamento o dovuti ad artefatti della PCR, come le chimere. In generale le sequenze sono ordinate secondo una dimensione nota, che di solito si riferisce alla dimensione di un cluster, che è specificata da una dimensione di campo = N nell'etichetta della sequenza, dove N è un numero intero. Noi abbiamo usato una dimensione minima per il *cluster* pari a 4.

OTU clustering: il raggruppamento delle reads in unità tassonomiche operative è stato effettuato mediante l'algoritmo UCLUST, che divide un set di sequenze in un cluster, il quale è definito da una sequenza nota come centroide o sequenza rappresentativa. Tale algoritmo effettua un *greedy clustering*, considerando le sequenze più rappresentate come le più attendibili biologicamente. Ogni sequenza nel cluster deve avere un grado di somiglianza con la sequenza centroide maggiore di una determinata soglia d'identità, T. UCLUST è

stato progettato per trovare un insieme di cluster tali che: tutti i centroidi abbiano un grado di somiglianza $< T$ tra loro, e tutte le sequenze membro abbiano un grado di somiglianza con il centroide $\geq T$. Le sequenze vengono elaborate nell'ordine in cui appaiono nel file di input, se la sequenza successiva corrisponde a un centroide esistente, viene assegnata a tale cluster, altrimenti diventa il centroide di un nuovo cluster. Ciò significa che le sequenze devono essere ordinate in modo che i centroidi più appropriati compaiano prima nel file.
Criteri di clusterizzazione: L'obiettivo di UPARSE_OTU è quello di identificare un insieme di sequenze rappresentative OTU, sottoinsieme delle sequenze di input, che soddisfi i seguenti criteri:

1. Tutte le coppie di sequenze OTU dovrebbero avere un'identità di sequenza (per coppia) $< 97\%$;
2. Le sequenze chimeriche devono essere eliminate;
3. Tutte le sequenze di input non-chimeriche devono corrispondere ad almeno un'OTU con un valore d'identità $\geq 97\%$.

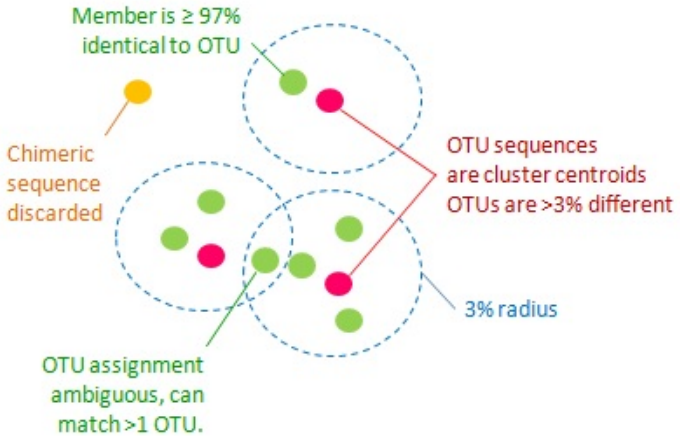


Figura 3.3: Schema dei criceri di clustering applicati dall'algoritmo UPARSE-OTU.

In questo modo, ci sarà un enorme numero di possibili insiemi di OTUs che soddisfano i criteri di raggruppamento. UPARSE_OTU utilizza un algoritmo greedy per trovare una soluzione biologicamente rilevante. Poichè le reads più abbondanti hanno una probabiltà maggiore di essere le più corrette e di conseguenza di essere vere sequenze biologiche, le sequenze di input vengono considerate in ordine decrescente di abbondanza. Questo significa che i centroidi tendono ad essere selezionati dalle reads più rappresentate, e quindi è più probabile che siano sequenze biologiche attendibili. Le sequenze di input devono essere filtrate in base alla qualità, contenenere nell'etichetta l'annotazione della dimensione, essere globalmente allineabili e non presentare gap terminali. Ogni sequenza in ingresso viene confrontata con il database di OTUs corrente e viene applicato un modello di massima parsimonia della sequenza, mediante UPARSE_REF (Figura 3.4).

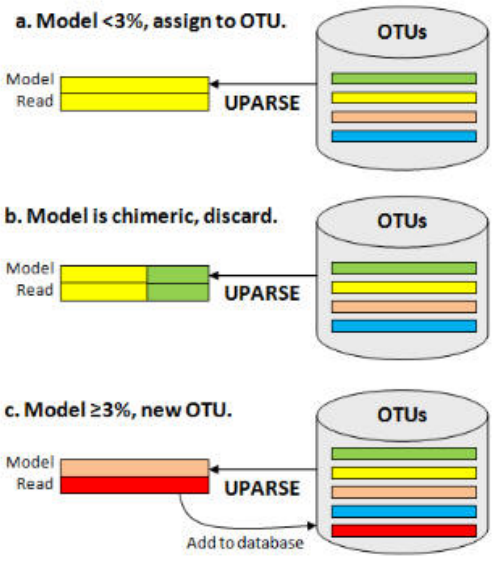


Figura 3.4: Modello di funzionamento dell'algoritmo UPARSE_REF.

A questo punto possono verificarsi tre condizioni:

- a) il modello UPARSE-REF è identico ad una OTU esistente per un valore $\geq 97\%$, quindi è diversa per meno del 3%;
- b) il modello è chimerico, cioè la sequenza in ingresso si allinea metà con una e metà con un'altra sequenza;
- c) il modello è diverso per un valore $> 3\%$ da ogni OTU esistente.

Nel caso a) la sequenza in ingresso diventa membro della OTU, lasciando come centroide la più rappresentata, nel caso b) viene scartata e nel caso c) la sequenza in ingresso viene aggiunta al database e diventa il centroide di una nuova OTU. La soglia di identità al 97% per le sequenze del gene 16S rRNA corrisponde approssimativamente al livello tassonomico di Specie, ed è stata utilizzata per permettere un'analisi ad alta risoluzione della comunità microbica (Konstantinidis and Tiedje, 2007).

Filtraggio delle chimere: le chimere sono sequenze formate da due o più sequenze biologiche unite insieme, che si creano sopratutto durante la PCR. L'algoritmo UPARSE_OTUs scarta le reads con modello chimerico; tuttavia alcune chimere possono non essere riconosciute, soprattutto se derivano da sequenze assenti nelle reads o presenti con una bassissima abbondanza. Per questo è consigliato un ulteriore filtraggio delle chimere, mediante l'algoritmo UCHIME, con cui si ottiene un set filtrato di OTUs.
L'algoritmo agisce cercando un allineamento a 3 vie di una sequenza query con due sequenze madri, A e B, in modo tale che ognuna sia più simile ad una diversa porzione della query. In base all'allineamento viene calcolato un punteggio, quanto più è alto tanto più è forte il segnale chimerico; un cutoff di tale valore, che di default è 0,28, determina se la query è una chimera o no (Edgar et al., 2011).

Labelling OTUs: arrivati a questa fase le etichette delle sequenze OTUs sono costituite dall'etichetta della read originale, modificata con l'aggiunta della dimensione, valore che

indica il numero di reads aventi quella stessa sequenza. Si ritiene quindi utile generare un nuovo set di etichette per le OTUs, ad esempio OTU_1, OTU_2 ... OTU_N dove N è il numero di OTUs.

Tabella di OTUs: dato che le etichette FASTA sono conservate in uscita, il valore di dimensione è quello di input e non si riferisce al numero di sequenze assegnato ad una determinata OTU. Per ottenere il numero di reads contenute in ogni OTU è necessario associare le reads alle OTUs. Il primo passo per creare la tabella è quello di allineare le reads alle OTU, utilizzando le prime come queries e le seconde come database. Se una reads si allinea con più di una OTU, USEARCH la assegna alla OTU con il massimo valore d'identità. Se una reads non trova riscontro con nessuna OTUs, è un singleton o una chimera, o presenta errori nella sequenza per più del 3%, viene scartata; abbiamo utilizzato infatti un citoff di identità del 97%.

3.2.3 Attribuzione tassonomica

L'assegnazione tassonomica è quel processo che permette di analizzare le sequenze in maniera da avere un profilo di diversità tassonomica del campione ambientale. Nel nostro lavoro la classificazione è stata effettuata utilizzando il classificatore online SINA (SILVA INcremental Aligner) sulle sequenze rappresentative ottenute con il clustering. Si tratta di un classificatore dipendente dalla tassonomia, cioè che confronta le sequenze dei geni del 16S rRNA con un database di riferimento come NCBI, GreenGenes o RDP; in questo caso è stato utilizzato il database SILVA, che consente una precisione di allineamento molto elevata (Pruesse et al., 2012). L'algoritmo attribuisce le singole read al rank tassonomico più simile. Ci possono essere però delle ambiguità quando una stessa sequenza può essere assegnata a più di una sequenza di referimento (Ribeca and Valiente, 2011).
SINA consente l'allineamento fino a un massimo di 1000 sequenze, composte da massimo 6000 basi ciascuna, utilizzando le più recenti serie di dati di SILVA, quest'ultimo fornisce una classificazione filogenetica per le piccole e le grandi subunità di rRNA, per Bacteria, Archaea ed Eukarya, nell'*European Nucleotide Archive*. La versione utilizzata è SINA v1.2.11.

3.2.4 Analisi ecologica della comunità microbica

R e vari pacchetti

Per l'analisi delle sequenze e dei dati è stato utilizzato il software R, che permette un'analisi statistica dei dati con appositi script creati ad hoc per il dataset in questione, l'interfaccia grafica Rstudio (R Core Team, 2014) e vari pacchetti disponibili online.
Tra i principali pacchetti usati, per le analisi statistiche, vi è il pacchetto Vegan (*http://vegan.r-forge.r-project.org/*); il quale permette di fare analisi di tipo ecologico e analisi multivariate della comunità ecologica (Oksanen et al., 2013).
L'altro pacchetto di cui ci siamo serviti è ggplot2 (*http://ggplot2.org/*), che è stato utilizzato per generare molte delle figure successive che mostrano in maniera grafica i risultati dell'analisi (Wickham, 2009).

Metodi di ordinazione e analisi statistica

Analisi di rarefazione e stima della biodiversità

Le curve di rarefazione sono state costruite mediante la funzione `rarecurve` di Vegan, che disegna una curva di rarefazione per ogni riga di dati in ingresso; al fine di valutare se le

comunità batteriche sono state campionate in modo soddisfacente. Le curve di rarefazione sono state costruite riportando in grafico il numero di OTUs osservate in funzione dell'intensità di campionamento, cioè del numero di sequenze per ogni campione.
Sulle sequenze OTUs è stata poi eseguita un'analisi di rarefazione al fine di verificare se i campioni condividono un uguale numero di assegnazioni e in caso contrario per uniformare le dimensioni dei campioni in modo da poter effettuare una stima della biodiversità.
Per quando concerne la diversità in specie sono stati utilizzati gli indici di Shannon-Weaver, di Richness e di Evenness, che sono stati calcolati sui campioni rarefatti, ripetendo la fase di rarefazione 1000 volte. Gli indici sono stati calcolati per ogni rarefazione, per i pazienti S e SD suddivisi a seconda dello stato clinico, %FEV1, sono stati calcolati poi i valori medi e la deviazione standard. In aggiunta è stata effettuata un'analisi della varianza (ANOVA), tra gli indici e le condizioni di FEV1, per verificare la significatività dei valori ottenuti.

Indice di Richness: corrisponde al numero di taxa diversi presenti nel campione analizzato.
Indice di Evenness: è stato calcolato utilizzando l'indice di equitabilità J di Pielou, che prende in considerazione la modalità di distribuzione dei singoli individui nelle varie specie:

$$J = \frac{H'}{\log_2 S}$$

o

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

dove: H' è il valore dell'Indice di diversità di Shannon-Weaver S è il numero di specie presenti nella data comunità $H'_{\max}$ è il valore massimo dell'indice di Shannon che si ha quando $H' = \log_2 S$.
L'equitabilità tende a 1 quanto più gli organismi sono distribuiti uniformemente tra le specie. Tende a 0 quanto più si hanno poche specie che dominano numericamente sulle altre.
Indice di Shannon-Weaver:

$$\text{H'} = -\sum_{j=1}^{s} p_j \log p_j$$

Dove p_j è la proporzione di individui appartenenti alle specie j-esima nel dataset di interesse, $\sum_j p_j = 1$, e s è il numero delle specie. La base del logaritmo utilizzato per determinare l'indice di Shannon può essere scelta liberamente. Solitamente si usa il logaritmo in base 2 o 10.

Analisi mediante networks

L'analisi mediante network viene utilizzata per esplorare le proprietà matematiche, statistiche e strutturali di un insieme di elementi (nodi) e le connessioni tra di loro. Come dimostrato in studi recenti, che hanno utilizzato il pirosequenziamento con barcode per esaminare comunità microbiche in un gran numero di campioni (Costello et al., 2009), è ora possibile generare dataset microbici che possono trarre il massimo vantaggio da approcci quali la *Network analysis* e possiamo applicarla anche a comunità molto diverse tra loro, per esplorarne i modelli di co-occorrenza.
La *Network analysis* dei modelli di co-occorrenza dei taxa microbici offre una nuova comprensione della struttura delle comunità microbiche complesse, una conoscenza che integra

ed espande le informazioni fornite dal pacchetto più standard di approcci analitici. In primo luogo, le associazioni inter-taxa, possono contribuire a rivelare gli spazi di nicchia condivisi dai membri della comunità o, simbiosi più dirette tra i membri della comunità.

Correlazione di Spearman L'indice di correlazione per ranghi di Spearman (ρ) è una misura statistica di correlazione non parametrica, usata cioè nel caso in cui non sia possibile assumere che i dati seguano distribuzioni gaussiane. A livello pratico il coefficiente ρ è semplicemente un caso particolare del coefficiente di correlazione di Pearson dove i valori vengono convertiti in ranghi prima di calcolare il coefficiente. Il coefficiente di Spearman valuta quanto bene la relazione tra due variabili può essere descritta utilizzando una funzione monotona, cioè che mantiene l'ordine. Se non ci sono valori di dati ripetuti, una perfetta correlazione Spearman di +1 o -1 verifica quando ciascuna delle variabili è una funzione monotona perfetta dell'altra. Il segno della correlazione indica la direzione di associazione tra due variabili X e Y. Se Y tende ad aumentare quando X aumenta, il coefficiente di correlazione di Spearman è positivo. Se Y tende a diminuire quando X aumenta, il coefficiente di correlazione Spearman è negativo. Una correlazione Spearman zero indica che non vi è alcuna tendenza di Y ad aumentare o diminuire quando X aumenta, ossia le variabili non sono correlate.

Costruzione dei networks. Al fine di generare una rappresentazione *network* dei dati delle comunità batteriche, sono stati calcolati tutti i possibili ranghi di correlazione di Spearman tra le OTUs. Un evento di co-occorrenza tra due OTUs è stato considerato valido se il coefficiente di correlazione di Spearman risultava maggiore di 0,06 e statisticamente significativo se si aveva un p-value minore di 0,01; in questo modo abbiamo ristretto l'analisi alle sole OTUs che mostravano rapporti significativi (Barberan et al., 2011). Sono stati generati cinque diversi networks (S, SD, FEV1 I, II, III), in cui ogni nodo corrisponde a una diversa OTU e ogni collegamento rappresenta la presenza di una correlazione di Spearman, statisticamente significativa tra due OTUs. Tutte le analisi sono state effettuate in R, usando il pacchetto *igraph* per generare i networks (Csardi and Nepusz, 2006), che sono stati poi esplorati e visualizzati con la piattaforma interattiva Gephi (Bastian et al., 2009).

Capitolo 4

Risultati e discussioni

4.1 Elaborazione delle sequenze

Le sequenze grezze, provenienti dal sequenziamento, sono state trimmate, cioè sono state rimosse quelle basi che avevano un quality score non soddisfacente, e private da possibili contaminanti utilizzando BLAST. I risultati sono illustrati nella seguente tabella 4.1 in cui sono riportate le quantità di sequenze per run, pre e post elaborazione.

Tabella 4.1: Risultati dell'elaborazione delle sequenze

Run	Sequenze iniziali	Sequenze post-elaborazione	percentuale sequenze perse
1	63 751	61 694	3%
2	80 431	77 850	3%
3	57 032	54 577	4%
4	689	664	4%

Con i file tabulare ottenuto abbiamo costruito un grafico a barre che mostri la percentuale di sequenze assegnate ad ogni campione nelle varie run. Come si nota nella Figura 4.1 la run 4 contiene un numero di sequenze non confrontabile con gli altri gruppi di campioni, motivo per cui decideremo di non utilizzarla nelle successive analisi.

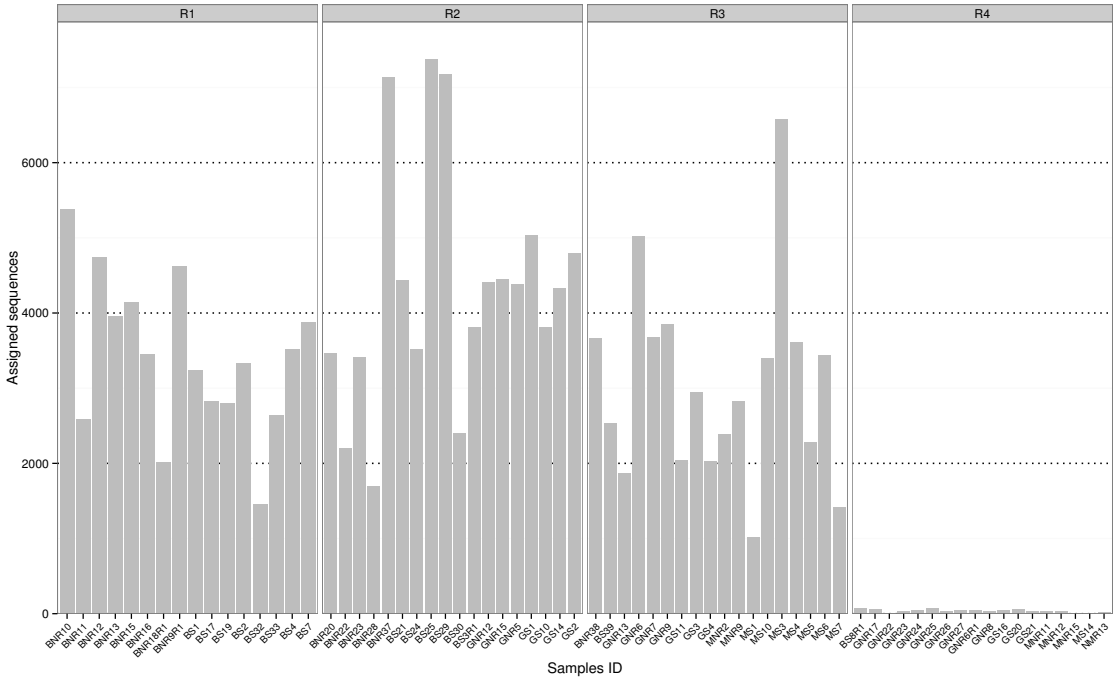


Figura 4.1: Sequenze assegnate ad ogni campione, suddivise per run.

4.2 OTU clustering

Mediante l'OTU clustering si ottengono 44 OTUs diverse. Mappando le reads con le OTUs per creare l' *OTU table* abbiamo ottenuto: una corrispondenza del 84.8%; ovvero su 194785 sequenze ne vengono scartate 29642.
Il numero di assegnazioni OTUs per ogni paziente risulta relativamente alto (in molti casi più di 6000 OTUs per paziente). Tuttavia per 19 campioni le assegnazioni OTUs risultano molto basse (le OTUs assegnate risultano poche), in relazione al basso numero di sequenze presenti in alcuni campioni (Figura 4.1).
In base ai risultati esposti nella figura 4.1 e quelli ottenuti con l'OTU clustering, le analisi successive sono state svolte su un dataset privo del gruppo di campioni R4, in quanto poteva portare ad una distorsione dei risultati.

4.3 Analisi della biodiversità

4.3.1 Curve di rarefazione

Le curve di rarefazione riportate nel grafico 4.2 tracciano il numero di OTUs in funzione del numero di sequenze per ognuno dei 54 campioni. Nonostante il nostro dataset sia formato da campioni con numero di assegnazioni variabile, in quanto anche senza la run 4 si hanno campioni con meno di 1000 sequenze e campioni che ne contengono più 7000, dal grafico delle curve di rarefazione (Figura 4.2) si osserva che tutti arrivano a plateau con un andamento asintotico, ciò indica che le comunità batteriche sono state campionate in modo soddisfacente e che i campioni sono tutti biologicamente spiegati.

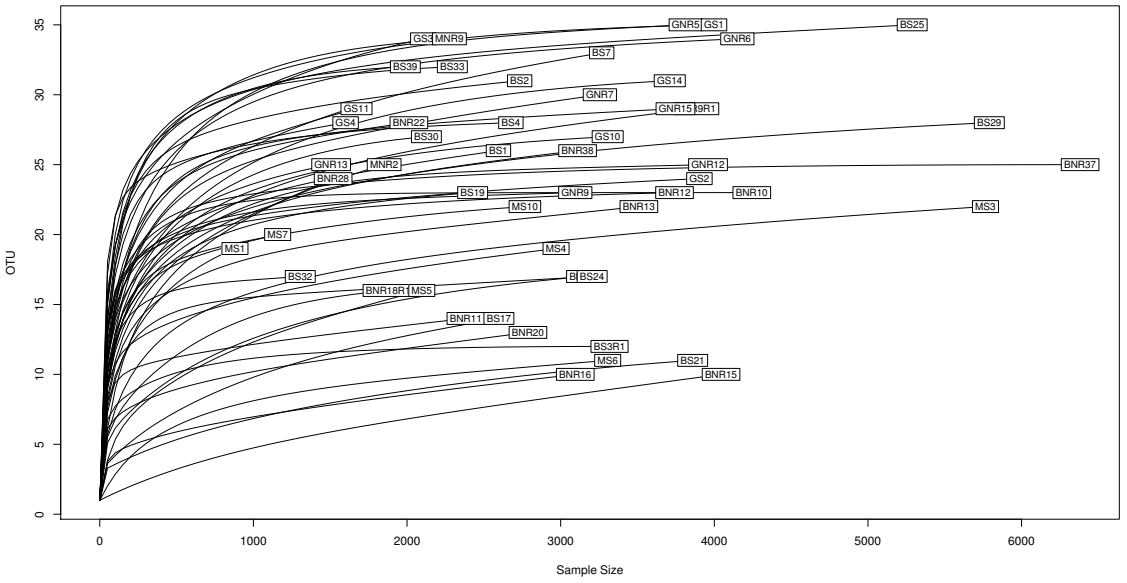


Figura 4.2: Curve di rarefazione

4.3.2 Analisi di rarefazione e stima della biodiversità

La differenza di dimensioni tra i campioni li rende non confrontabili tra loro statisticamente, per questo la tabella delle OTUs è stata rarefatta al campione di dimensioni minori, MS1, che contiene 880 assegnazioni. Mediante uno script ad hoc in R e le funzioni del pacchetto Vegan abbiamo estratto casualmente dal numero di sequenze assegnate a ciascun paziente una serie di sub-campioni di 880 assegnazioni e per evitare di falsare la misura abbiamo ripetuto questa operazione 1000 volte.

Per ogni ciclo di rarefazione sono stati calcolati gli indici di diversità per i vari campioni e ne è stata calcolata la media. I dati sono poi stati suddivisi secondo le condizioni S e SD di ogni gruppo di FEV1 (S e SD I, S e SD II e S e SD III). È stata calcolata la media degli indici per ogni raggruppamento e le relative deviazioni standard. La tabella 4.2 riporta le medie degli indici di diversità utilizzati: ricchezza in specie, indice di Shannon ed equitabilità, e le relative deviazioni standard. La richness mostra una leggera diminuzione dal gruppo I con FEV1$\geq$ 70% al gruppo III con FEV1$<$ 40%. L'indice di ricchezza di specie e l'indice di Shannon maggiori si hanno per i pazienti del gruppo I. (La evenness più alta si ha per i pazienti S del gruppo I, ma non per gli SD del gruppo I).
Nella tabella seguente abbiamo indicato l'indice di richness con S, quello di Shannon con H' e la evenness con J, attenendoci alle notazioni adottate da Vegan.

Tabella 4.2: Indici di diversità

FEV1	Conditions	Means S	σ_S	Means H'	$\sigma_{H'}$	Means J	σ_J
I	S	23.534	6.10	2.201	0.498	0.701	0.120
I	SD	23.420	3.749	1.846	0.483	0.586	0.145
II	S	16.378	6.600	1.310	0.664	0.469	0.210
II	SD	18.181	8.261	1.759	0.598	0.616	0.125
III	S	21.005	6.779	1.971	0.590	0.650	0.145
III	SD	16.285	8.322	1.435	0.796	0.501	0.259

Per meglio valutare la significatività delle differenze nei valori degli indici è stata effettuata un'analisi ANOVA. I risultati (Tabella 4.3) hanno mostrato che sia la richness che l'indice di Shannon si differenziano secondo i gruppi basati sui valori di FEV1, e ne sono quindi influenzati; mentre la evenness risulta soggetta all'effetto dell'interazione tra valore di FEV1 e condizioni S e SD, quindi, in questo caso, i due fattori FEV1 e S e SD non possono essere considerati indipendentemente. A livello di evenness della comunità batterica non si può prescindere da uno dei due fattori.

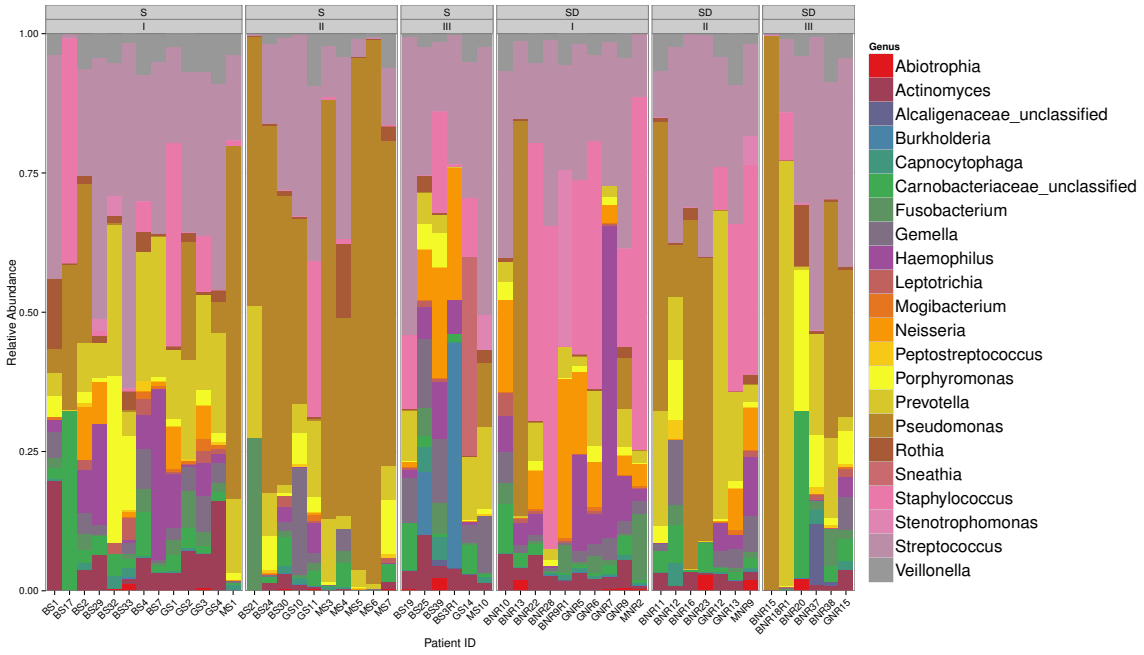
Tabella 4.3: ANOVA indici di diversità

Variabili	p-value Richness	p-value Shannon	p-value Evenness
FEV1	0.01**	0.01**	0.08
Conditions	0.70	0.36	0.37
FEV1:Conditions	0.39	0.05	0.03*

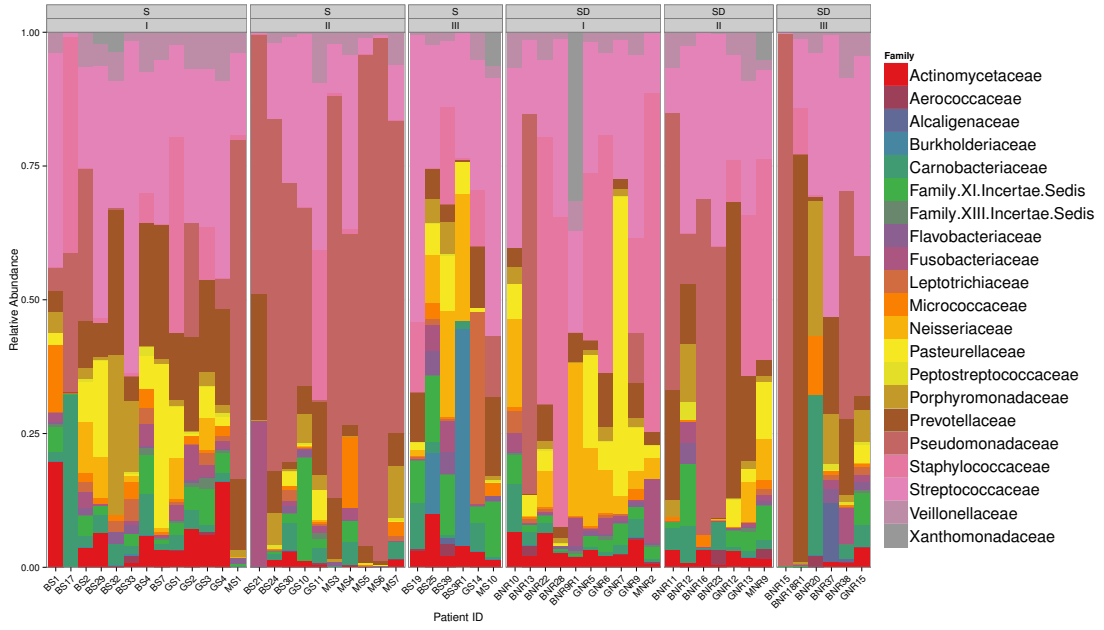
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

4.4 Assegnazione tassonomica

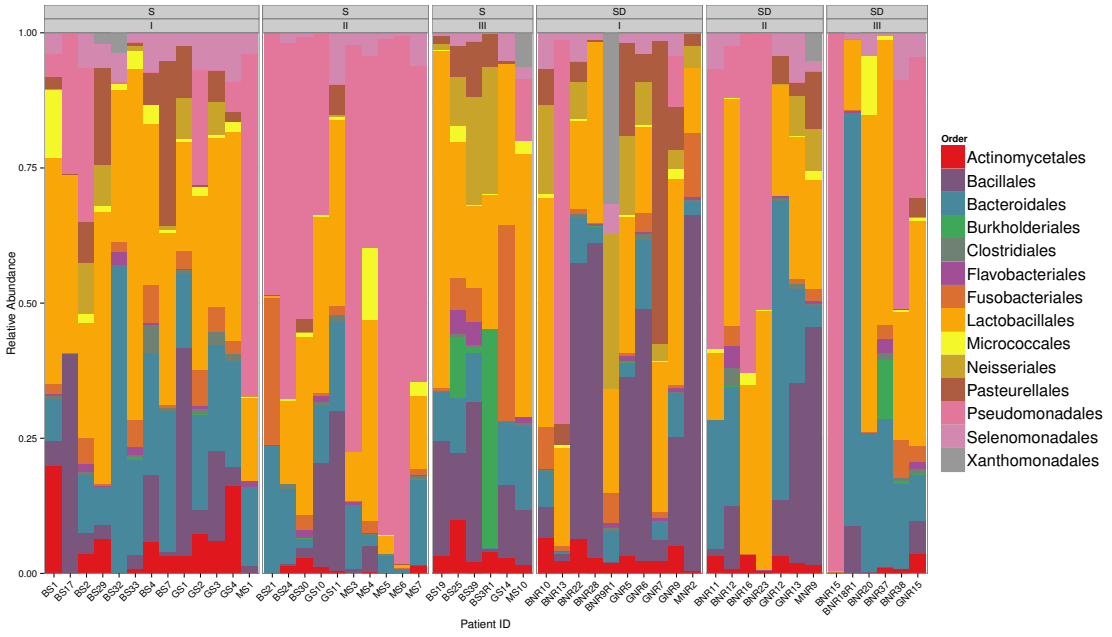
L'output del classificatore SINA è stato sottoposto a varie analisi attraverso l'utilizzo di uno script ad hoc in R. La classificazione è stata eseguita utilizzando un cutoff di similarità di 0.5, come suggerito da SINA. Per ogni campione è stata calcolata l'abbondanza relativa del numero di sequenze assegnate ad uno specifico taxa rispetto al numero di sequenze disponibili. Per ogni livello tassonomico analizzato non sono stati riportati nel grafico i "tipi" sottorappresentati, cioè quelli che presentavano un'abbondanza relativa media tra i campioni inferiore al valore di cutoff di 0.001. Questa operazione è stata ripetuta per cinque livelli tassonomici, dai phyla ai generi, e il risultato è stato esplicitato in altrettanti grafici a barre (Figure 4.3).



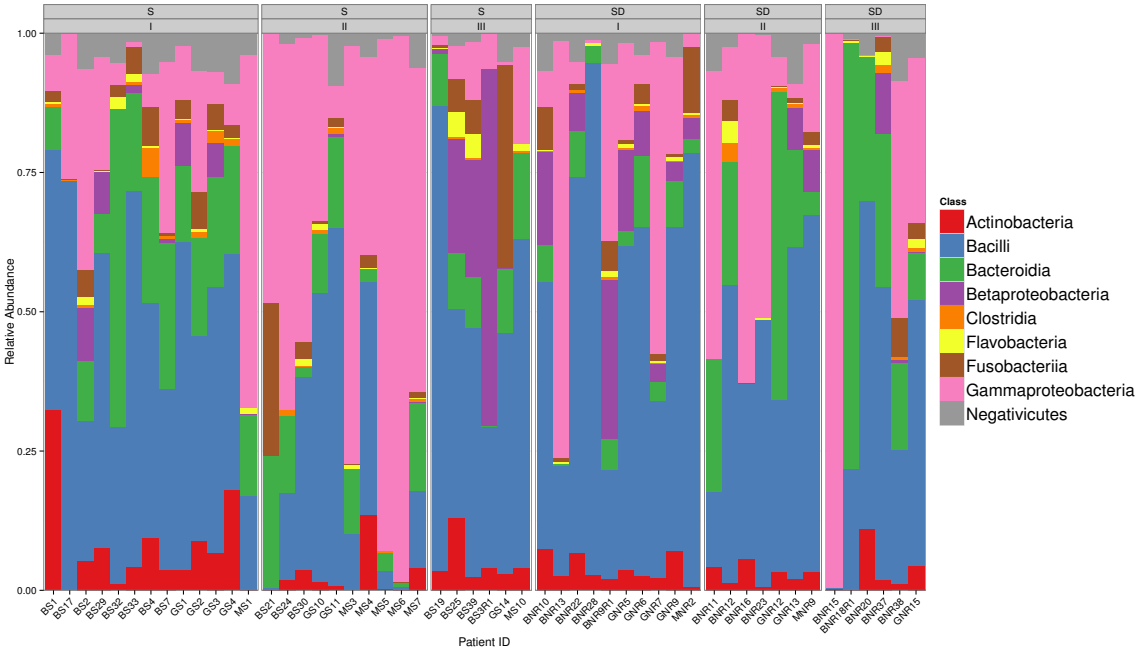
(a) barplot generi



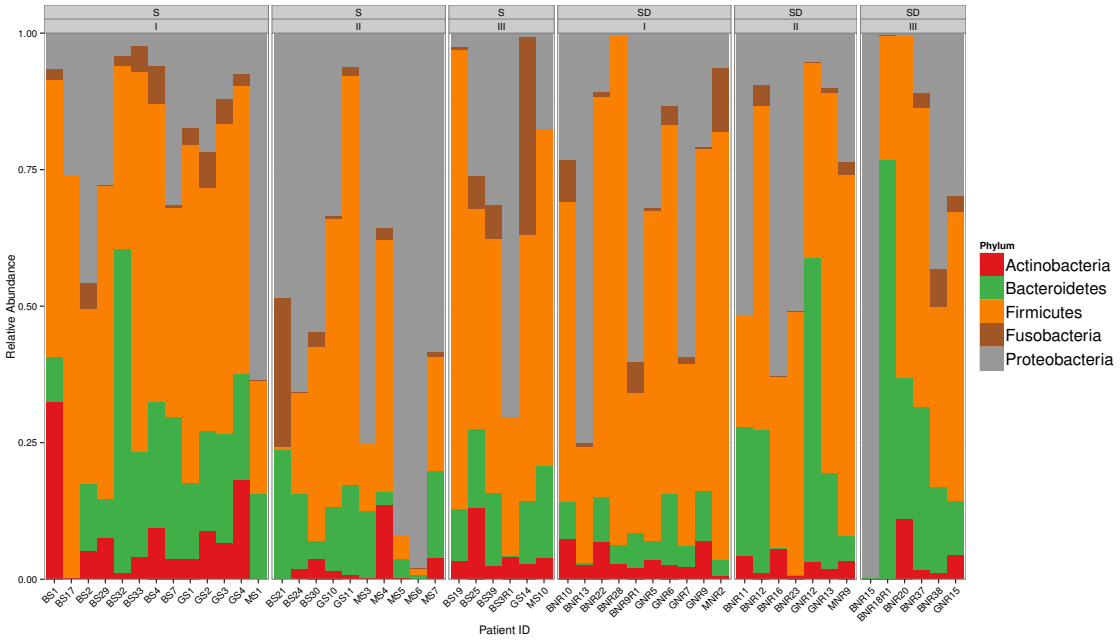
(b) barplot famiglie



(c) barplot ordini



(d) barplot classi



(e) barplot phyla

Figura 4.3: Composizione tassonomica della comunità batterica. I grafici mostrano l'abbondanza relativa dei vari gruppi filogenetici presenti nel microbioma delle vie aeree dei pazienti campionati.

Quello che si osserva dai grafici dei 54 campioni è un'elevata eterogeneità della composizione della comunità microbica, da paziente a paziente. Tuttavia si osservano delle differenze significative nella composizione batterica in particolare tra pazienti S e SD del gruppo III. I pazienti S presentano un microbioma più diversificato rispetto agli SD, nei quali, nello stato clinico di grave malattia polmonare si riscontra un aumento del genere *Pseudomonas* e una diminuzione dei generi *Staphylococcus*, *Stenotrophomonas* e *Neisseria*. Alcune sequenze dei pazienti appartenenti al gruppo S (III) sono state attribuite al genere *Burkholderia*, la cui presenza era uno dei criteri di esclusione dei pazienti dallo studio; averne rilevato la presenza all'interno del microbioma indica che in alcuni casi ceppi di *Burkholderia* posso essere presenti in uno stato vitale, ma non coltivabile.
Per analizzare meglio la composizione dei microbiomi abbiamo riportato in grafico l'abbondanza e la prevalenza, dei generi identificati, nei vari campioni. L'abbondanza indica quante quante sequenze di un determinato campione corrispondenti a quel genere sono contenute in quel campione; mentre la prevalenza è una misura di frequenza ed esprime la percentuale di ogni genere nei vari campioni, cioè il rapporto tra quante sequenze di un determinato genere ci sono in un campione e quante sequenze di quel determinato genere ci sono in totale, considerando tutti i campioni.
In generale sono stati identificati 24 generi diversi, aventi differenti abbondanza e prevalenza nei campioni (Figura 4.4). Alcuni generi risultano presenti in tutti i campioni con una frequenza simile, come ad esempio *Streptococcus*, *Veionella*, *Rothia*; mentre altri risultano molto più variabili in frequenza, come *Pseudomonas* e *Staphylococcus*.

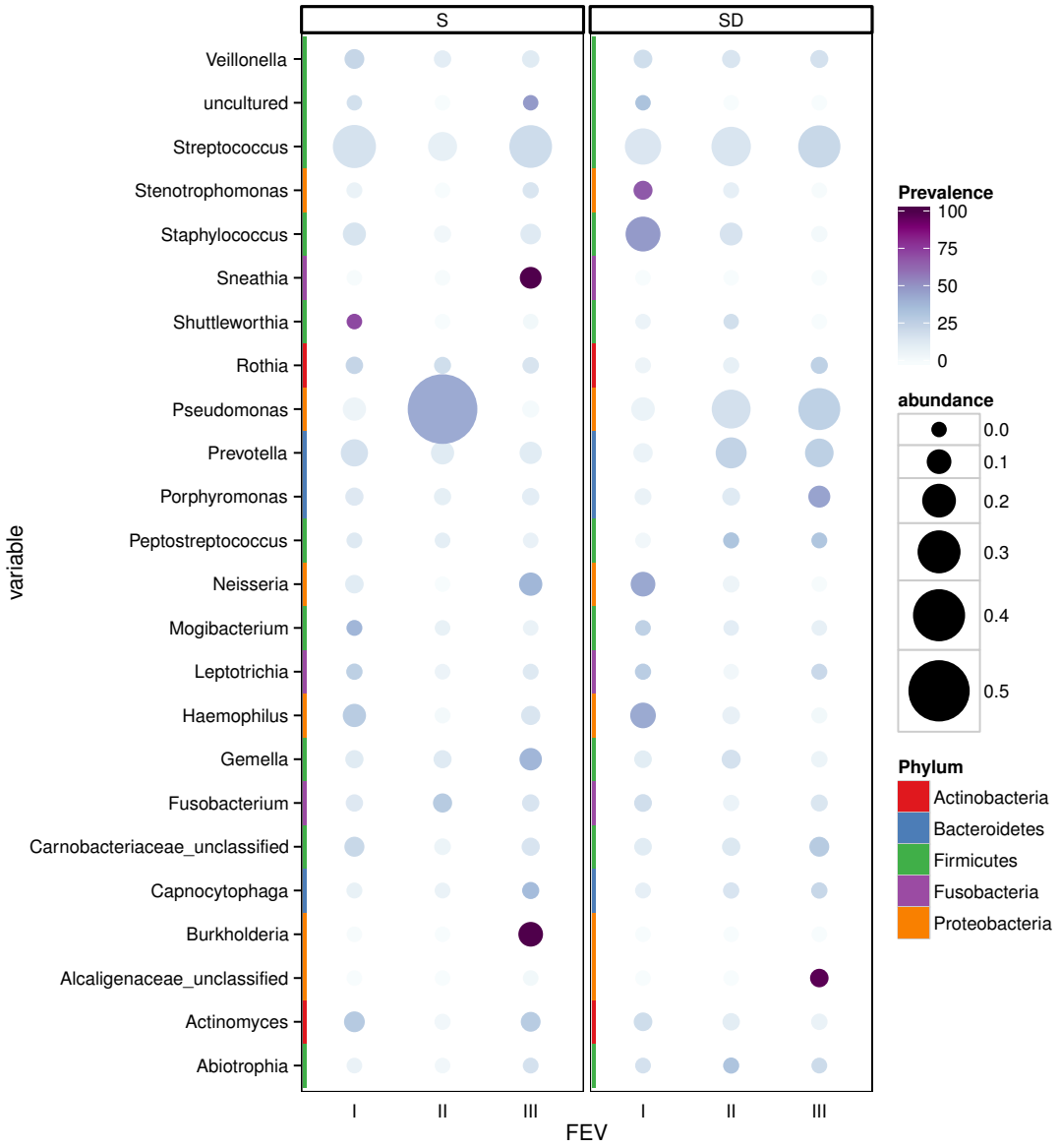


Figura 4.4: Abbondanza e prevalenza dei generi identificati nei vari campioni. Il diametro dei simboli esprime l'abbondanza di un genere dentro un campione. La trasparenza si riferisce alla prevalenza di un genere tra i campioni.

4.5 Analisi statistica multivariata

Al fine di indagare se i diversi valori di FEV1 e le condizioni S e SD sono riflessi dal raggruppamento in OTUs delle comunità batteriche, è stata effettuata una *canonical component analysis*,o CCA. Nel nostro set di dati, il numero di sequenze per campione variava da 880 a 6381 con una media di 3040.5 sequenze per campione. Per correggere le variazioni nella conta totale delle sequenze tra i campioni, l'abbondanza di ogni OTU in un dato campione è stata standardizzata applicando il log (in base 10) ai valori della *OTU table*. Il logaritmo è stato usato per ridurre l'influenza della OTU più dominante (Fodor et al., 2012).
Dai risultati (Figura 4.5) si osserva che i campioni di pazienti sia S che SD, aventi lieve o moderata malattia polmonare (gruppi I e II) sono piuttosto sovrapposti, mentre i microbiomi dei pazienti S ed SD con grave malattia polmonare, appartenenti cioè al gruppo III

(FEV1 <40%), presentano significative differenze. Questa differenziazione delle comunità batteriche in relazione alla condizione S rispetto a quella SD, è stata anche statisticamente supportata dal risultato dell'analisi MANOVA (Tabella 4.4).
Le variabili utilizzate per l'analisi MANOVA sono state le condizioni dei pazienti, stabile (S) e in sostanziale declino (SD), lo stato clinico, espresso dal valore di FEV1 e i campioni. I risultati esposti nella Tabella 4.4 trovano corrispondenza con quelli del grafico 4.5 e indicano che le condizioni, di stabilità della funzione polmonare o di sostanziale declino, influenzano la composizione del microbioma.

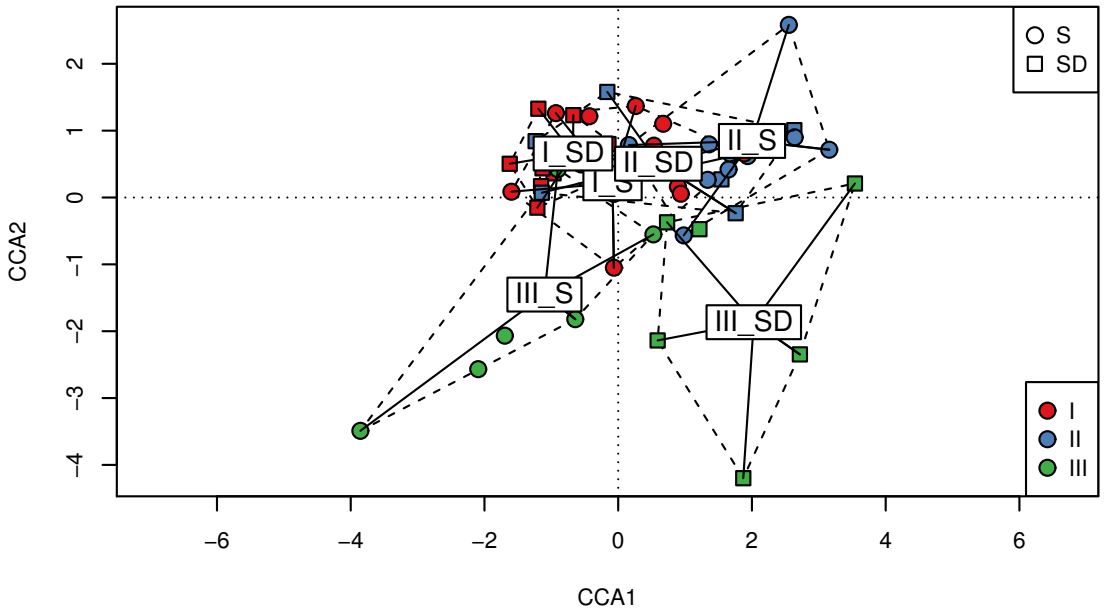


Figura 4.5: La figura è la rappresentazione grafica dell'analisi CCA. I colori sono in base allo stato clinico dei pazienti (gruppi FEV1), mentre la forma è in base alla condizione di stabilità (S) o di sostanziale declino (SD).

Tabella 4.4: Analisi MANOVA

Variabili	p-value(>F)	gradi di libertà	Pillai test	F test	num Df	den Df
FEV1	0.08	2	1.928	2.434	88	8
Conditions	0.04 *	1	0.992	8.822	44	3
FEV1:Conditions	0.35	2	1.872	1.335	88	8
Resuduals	46					

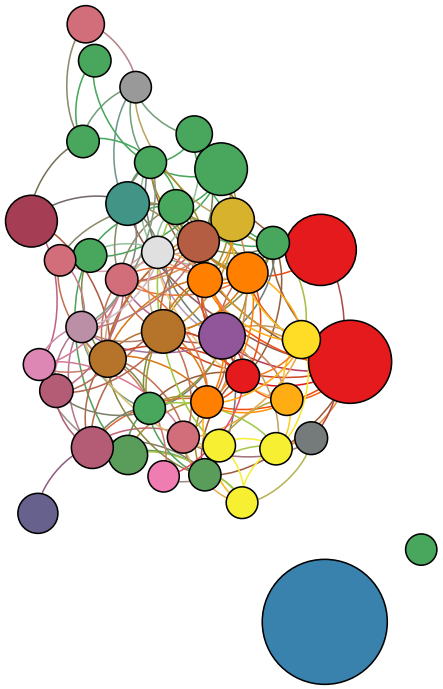
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Nella Tabella 4.4 si osserva che solo il secondo p-value risulta significativo; questo indica che le condizioni, di stabilità della funzione polmonare o di grave declino, sono determinanti per la composizione del microbioma, che varia al loro variare; mentre né lo stato clinico espresso dai valori di FEV1 né un'interazione tra FEV1 e condizioni (S e SD) influiscono sula composizione della comunità microbica. Per la composizione della comunità batterica,

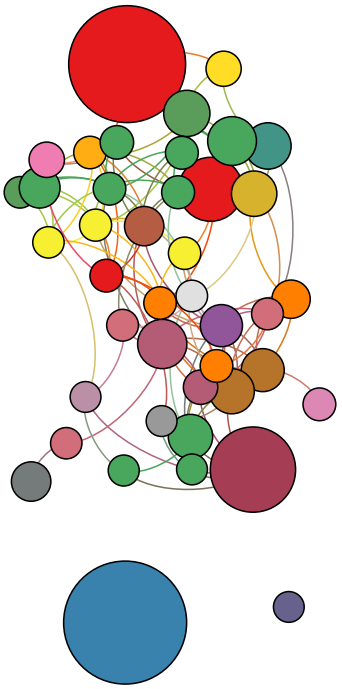
dunque, sono più determinanti le condizioni rispetto allo stato clinico, mente la biodiversità del microbioma è più influenzata da quest'ultimo.

4.6 Networks

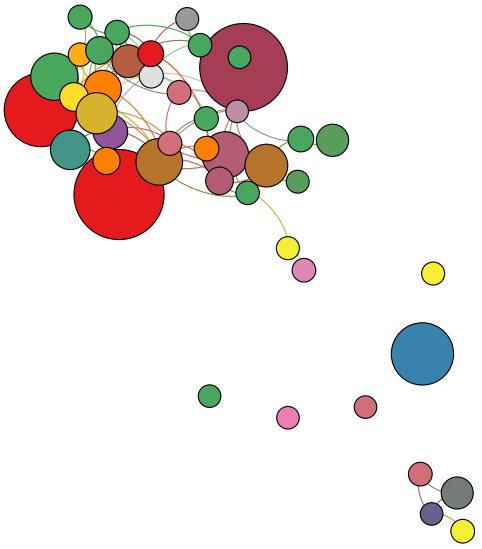
Per controllare ulteriormente quale taxon, o composizione tassonomica, potrebbe essere più legata alla condizione S rispetto alla SD e al raggruppamento secondo FEV1, abbiamo effettuato un'analisi mediante networks, in cui i nodi rappresentano i vari taxa identificati e i collegamenti la presenza di una correlazione nella occorrenza delle OTUs tra i campioni (Barberan et al., 2011).
Documentare le interazioni tra taxa, cioè i modelli di co-occorrenza, di comunità complesse e diverse può aiutare ad accertare i ruoli funzionali o le nicchie ambientali occupate da microrganismi non coltivabili. Utilizzando i dati di sequenza abbiamo esplorato le interazioni dirette e indirette tra i taxa microbici coesistenti nelle vie respiratorie dei pazienti fibrocistici. Come ultima analisi, quindi, abbiamo costruito dei networks che rappresentano i pattern di co-occorrenza (90% di cutoff) tra le OTUs, sulla base delle analisi di correlazione. Una connessione è sinonimo di una forte (ρ di Spearman$\geq$ 0.6) e significativa (p-value$\leq$ 0.05) correlazione. In questi grafici (Figure4.6) la dimensione di ciascun nodo è proporzionale al numero di assegnazioni (sequenze assegnate) e la distanza tra i nodi è inversamente al numero di connessioni (maggiore è il numero di link tra due nodi più sono vicini tra loro). Le diverse OTUs sono colorate secondo la tassonomia (famiglie).
Dai risultati si osserva che la presenza di *Pseudomonas* risulta sempre slegata dalla presenza di qualsiasi altro taxon. È interessante notare che, quando sono state analizzate le proprietà dei networks (Tabella 4.5) sono stati registrati diversi valori di grado medio, con valori più elevati per pazienti S rispetto a quelli dei pazienti SD, sostenendo l'ipotesi di un livello più complesso di relazioni ecologiche nel microbiota delle vie aeree dei pazienti S rispetto al microbiota delle vie respiratorie dei pazienti SD.



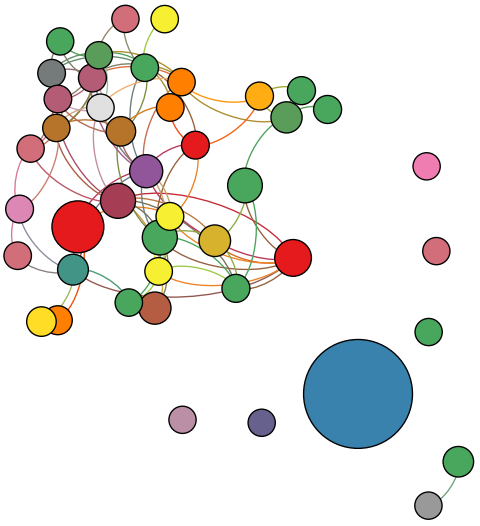
(a) Network pazienti S



(b) Network pazienti SD

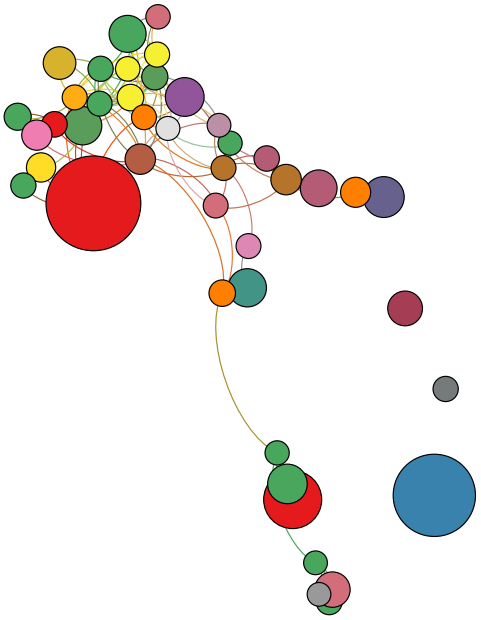


(c) Network pazienti gruppo I (FEV1 $\geq$ 70%)

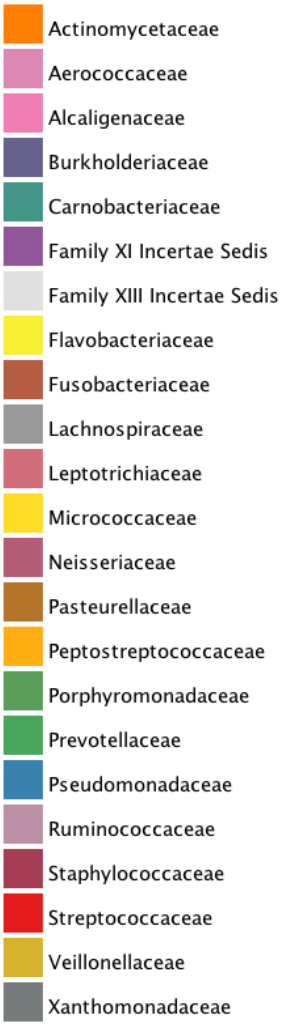


(d) Network pazienti gruppo II (40% $\leq$ FEV1 $<$ 70%)

Alcune proprietà topologiche comunemente utilizzate nell'analisi mediante network sono state calcolate per descrivere il complesso reticolo di inter-relazioni tra OTUs (Newman, 2003) e sono riportate nella Tabella 4.5. Nei networks la distanza media tra tutte le coppie



(e) Network pazienti gruppo III (FEV1 < 40%)



(f) Famiglie

Figura 4.6: Networks dei microbiomi dei pazienti suddivisi in base alle condizioni di stabilità o di sostanziale declino della funzione polmonare e ai valori percentuali di FEV1, indici dello stato clinico. I nodi rappresentano i taxa identificati e sono colorati secondo le famiglie batteriche, mentre i collegamenti indicano la presenza di una correlazione nell'occorrenza delle OTUs tra i campioni

di nodi, o lunghezza media del percorso [1] oscilla tra 2.014 (S) a 2.382 (SD) e tra 2.453 (FEVI) e 3.164(FEVIII). Il diametro, cioè la distanza più lunga all'interno dei vari network è di 4 (S) e 5(SD), 6(FEVI) e 7(FEVIII). Il coefficiente di clustering, cioè una misura del grado in cui i nodi di un grafo tendono a raggrupparsi, oscilla tra 0.432 (SD) e 0.557 (FEVIII). Nel complesso si osserva che il network del microbioma delle vie aeree dei paziente stabili (S) è composto da OTUs molto connesse (4.634 collegamenti per ogni OTU), mentre sono meno connesse le OTUs relative al microbioma dei pazienti SD (3.071collegamenti per OTU). Inoltre in accordo con i valori degli indici di diversità il network dei pazienti S è più biodiverso rispetto a quello degli SD, ad indicare che con l'aumentare della gravità della malattia la diversità diminuisce e la comunità batterica è meno interconnessa.

[1] *Average path length*, cioè il numero medio di passaggi lungo i percorsi più brevi per tutte le possibili coppie di nodi del network. Si tratta di una misura dell'efficienza di trasporto di informazioni o di massa in un network.

Tabella 4.5: Proprietà networks

Properties	STABLE	SEVERE DECLINE	FEV1 I	FEV1 II	FEV1 III
Nodes	44	42	44	44	43
Edges	204	129	105	103	115
Network diameter	4	5	6	5	7
Connected components	2	3	7	8	4
Average degree	10.091	6.143	4.773	4.682	5.349
Graph density	0.235	0.150	0.111	0.109	0.217
Modularity	0.256	0.385	0.376	0.422	0.366
Avg. clustering coefficient	0.535	0.432	0.517	0.509	0.557
Avg. path length	2.014	2.382	2.453	2.493	3.164
Edges per OTUs	4.634	3.071	2.386	2.341	2.674

Conclusioni

In questo lavoro di tesi abbiamo analizzato, con metodi bioinformatici, dati metabarcoding del microbioma delle vie aeree di pazienti fibrocistici con lo scopo di arrivare ad una migliore comprensione delle implicazioni cliniche di agenti patogeni nuovi e/o emergenti nella popolazione FC e di identificare nuovi bersagli e biomarcatori per il trattamento e la gestione delle infezioni batteriche nei pazienti con Fibrosi Cistica. Quello che è stato osservato è una netta differenza tra i microbiomi dei pazienti S e SD, le maggiori differenze si rilevano in particolare tra i pazienti con malattia polmonare grave (gruppo III), per i quali il genere *Pseudomonas* risulta più abbondante nei pazienti SD(III), mentre i generi *Staphylococcus* e *Stenotrophomonas* sono più abbondanti nei i pazienti S(III). Per quanto riguarda *Pseudomonas* va evidenziato come dall'analisi network la sua presenza non risulta correlata a quella di nessun altro taxon, suggerendo che la sua colonizzazione delle vie aeree sia indipendente dalla preesistente composizione della comunità microbica. In alcuni pazienti appartenenti al gruppo S si è riscontrata la presenza di sequenze attribuite al genere *Burkholderia*, poiché la presenza di tale batterio era uno dei criteri di esclusione dei pazienti dallo studio, la sua presenza all'interno del microbioma indica che in alcuni casi ceppi di *Burkholderia* possono essere presenti in uno stato vitale, ma non coltivabile. Questo dato necessita di ulteriori approfondimenti, anche attraverso uno studio longitudinale, in modo da valutare la comparsa di *Burkholderia* come ceppo coltivabile e la corrispondente condizione clinica di grave malattia polmonare. In generale la comunità microbica dei pazienti stabili risulta molto più interconnessa e biodiversa rispetto a quella dei pazienti in grave declino. Le differenze osservate tra pazienti S e SD in termini di generi batterici e di OTUs potrebbero essere la base per lo sviluppo di tools molecolari per la diagnosi precoce e la prognosi dei pazienti fibrocistici in declino polmonare.

Appendice

Tabella dati dei pazienti.

STUDY ID	FEV1	DNA ID	MID	Conditions	Run
BS1	I	1	M1	S	R1
BS2	I	2	M2	S	R1
BS3	III	3	M3	S	R1
BS4	I	4	M4	S	R1
BS32	I	55	M5	S	R1
BS33	I	56	M6	S	R1
BS7	I	7	M7	S	R1
BS3R2	III	57	M8	S	R1
BNR9R1	I	58	M9	SD	R1
BNR10	I	10	M10	SD	R1
BNR11	II	11	M11	SD	R1
BNR12	II	12	M12	SD	R1
BNR13	I	13	M13	SD	R1
BNR18R1	III	59	M14	SD	R1
BNR15	III	15	M15	SD	R1
BNR16	II	16	M16	SD	R1
BS17	I	17	M17	S	R1
BS19	III	19	M18	S	R1
BNR20	III	20	M1	SD	R2
BS21	II	21	M2	S	R2
BNR22	I	22	M3	SD	R2
BNR23	II	23	M4	SD	R2
BS24	II	24	M5	S	R2
BS25	III	25	M6	S	R2
BNR37	III	60	M7	SD	R2
BS3R1	III	27	M8	S	R2
BNR28	I	28	M9	SD	R2
BS29	I	29	M10	S	R2
BS30	II	30	M11	S	R2
GNR12	II	31	M12	SD	R2
GNR5	I	32	M13	SD	R2
GS10	II	33	M14	S	R2
GNR15	III	34	M15	SD	R2
GS2	I	35	M16	S	R2
GS14	III	36	M17	S	R2
GS1	I	37	M18	S	R2

STUDY ID	FEV1	DNA ID	MID	Conditions	Run
GS4	I	38	M1	S	R3
GS3	I	39	M2	S	R3
GNR6	I	40	M3	SD	R3
GNR7	I	41	M4	SD	R3
GS11	II	42	M5	S	R3
GNR9	I	43	M6	SD	R3
GNR13	II	44	M7	SD	R3
BS39	III	62	M8	S	R3
MS7	II	46	M9	S	R3
MS10	III	47	M10	S	R3
MS4	II	48	M11	S	R3
MS5	II	49	M12	S	R3
MNR2	I	50	M13	SD	R3
MS6	II	51	M14	S	R3
MNR9	II	52	M15	SD	R3
MS3	II	53	M16	S	R3
MS1	I	54	M17	S	R3
BNR38	III	61	M18	SD	R3
BS8R1	II	63	M1	S	R4
GS16	I	64	M2	S	R4
GNR17	I	65	M3	SD	R4
GNR6R1	I	66	M4	SD	R4
MNR15	III	81	M5	SD	R4
GS20	II	68	M6	SD	R4
GS21	II	69	M7	S	R4
GNR22	II	70	M8	SD	R4
GNR23	II	71	M9	SD	R4
GNR24	III	72	M10	SD	R4
GNR25	II	73	M11	SD	R4
GNR26	III	74	M12	SD	R4
GNR8	I	75	M13	SD	R4
GNR27	III	76	M14	SD	R4
MNR11	III	77	M15	SD	R4
MNR12	III	78	M16	SD	R4
MNR13	III	79	M17	SD	R4
MS14	II	80	M18	S	R4

Bibliografia

A. Apolloni. *Confronti di metodi statistici per la misura dell'espressione differenziale di dati di RNA sequencing*. Tesi di laurea, Università degli Studi di Padova, 2011/2012.

G. Bacci, M. Bazzicalupo, A. Benedetti, and A. Mengoni. Streamingtrim 1.0: a java software for dynamic trimming of 16s rrna sequence data from metagenetic studies. *Molecular Ecology Resources*, 14(2):426–434, 2014. ISSN 1755-0998. doi: 10.1111/1755-0998.12187. URL http://dx.doi.org/10.1111/1755-0998.12187.

A. Barberan, S. T. Bates, E. O. Casamayor, and N. Fierer. Using network analysis to explore co-occurrence patterns in soil microbial communities. *The ISME Journal*, 6(2):343–351, 2011. ISSN 1751-7362. doi: 0.1038/ismej.2011.119. URL http://dx.doi.org/10.1038/ismej.2011.119.

M. Bastian, S. Heymann, and M. Jacomy. Gephi: An open source software for exploring and manipulating networks. *International AAAI Conference on Weblogs and Social Media*, 2009. URL http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154.

L. A. Carmody, J. Zhao, P. D. Schloss, J. F. Petrosino, S. Murray, V. B. Young, J. Z. Li, and J. J. LiPuma. Changes in cystic fibrosis airway microbiota at pulmonary exacerbation. *Annals of the American Thoracic Society*, 10(3):179–187, 2013. ISSN 2329-6933. doi: 10.1513/AnnalsATS.201211-107OC. URL http://dx.doi.org/10.1513/AnnalsATS.201211-107OC.

P. J. A. Cock, C. J. Fields, N. Goto, M. L. Heuer, and P. M. Rice. The sanger fastq file format for sequences with quality scores, and the solexa/illumina fastq variants. *Nucleic Acids Research*, 38(6):1767–1771, 2010. doi: 10.1093/nar/gkp1137. URL http://nar.oxfordjournals.org/content/38/6/1767.abstract.

E. K. Costello, C. L. Lauber, M. Hamady, N. Fierer, J. I. Gordon, and R. Knight. Bacterial community variation in human body habitats across space and time. *Science*, 326(5960):1694–1697, 2009. doi: 10.1126/science.1177486. URL http://www.sciencemag.org/content/326/5960/1694.abstract.

G. Csardi and T. Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006. URL http://igraph.org.

G. Doring, N. Hoiby, and C. S. Group. Early intervention and prevention of lung disease in cystic fibrosis: a european consensus. *Journal of Cystic Fibrosis*, 3(2):67–91, 2004. URL http://dx.doi.org/10.1016/j.jcf.2004.03.008.

R. C. Edgar. Uparse: highly accurate otu sequences from microbial amplicon reads. *Nature Methods*, 10(10):996–998, 2013. ISSN 1548-7091. doi: 10.1038/nmeth.2604. URL http://dx.doi.org/10.1038/nmeth.2604.

R. C. Edgar, B. J. Haas, J. C. Clemente, C. Quince, and R. Knight. Uchime improves sensitivity and speed of chimera detection. *Bioinformatics*, 27(16):2194–2200, 2011. doi: 10.1093/bioinformatics/btr381. URL http://bioinformatics.oxfordjournals.org/content/27/16/2194.abstract.

A. A. Fodor, E. R. Klem, D. F. Gilpin, J. S. Elborn, R. C. Boucher, M. M. Tunney, and M. C. Wolfgang. The adult cystic fibrosis airway microbiota is stable over time and infection type, and highly resilient to antibiotic treatment of exacerbations. *PLoS ONE*, 7(9), 09 2012. doi: 10.1371/journal.pone.0045001. URL http://dx.doi.org/10.1371%2Fjournal.pone.0045001.

N. J. Gotelli and R. K. Colwell. Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecology Letters*, 4 (4):379–391, 2001. ISSN 1461-0248. doi: 10.1046/j.1461-0248.2001.00230.x. URL http://dx.doi.org/10.1046/j.1461-0248.2001.00230.x.

W. Hardle and L. Simar. *Applied Multivariate Statistical Analysis*. Springer, 2nd edition, 2007.

M. Hattori and T. D. Taylor. The human intestinal microbiome: A new frontier of human biology. *DNA Research*, 16(1):1–12, 2009. doi: 10.1093/dnares/dsn033. URL http://dnaresearch.oxfordjournals.org/content/16/1/1.abstract.

H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(3-4):321–377, 1936. doi: 10.1093/biomet/28.3-4.321. URL http://biomet.oxfordjournals.org/content/28/3-4/321.short.

M. Kircher and J. Kelso. High-throughput dna sequencing, concepts and limitations. *BioEssays*, 32(6):524–536, 2010. ISSN 1521-1878. doi: 10.1002/bies.200900181. URL http://dx.doi.org/10.1002/bies.200900181.

J. L. Kirk, L. A. Beaudette, M. Hart, P. Moutoglis, J. N. Klironomos, H. Lee, and J. T. Trevors. Methods of studying soil microbial diversity. *Journal of Microbiological Methods*, 58(2):169 – 188, 2004. ISSN 0167-7012. doi: http://dx.doi.org/10.1016/j.mimet.2004.04.006. URL http://www.sciencedirect.com/science/article/pii/S0167701204000983.

K. T. Konstantinidis and J. M. Tiedje. Prokaryotic taxonomy and phylogeny in the genomic era: advancements and challenges ahead. *Current opinion in microbiology*, 10 (5):504,Äì509, October 2007. ISSN 1369-5274. doi: 10.1016/j.mib.2007.08.006. URL http://dx.doi.org/10.1016/j.mib.2007.08.006.

Y. Li, S. Li, N. Hu, Y. He, R. Pong, D. Lin, L. Lu, and M. Law. Comparison of next-generation sequencing systems. *Journal of biomedicine and biotechnology*, 2012. ISSN 1110-7243. doi: 10.1155/2012/251364. URL http://dx.doi.org/10.1155/2012/251364.

J. J. LiPuma. The changing microbial epidemiology in cystic fibrosis. *Clinical Microbiology Reviews*, 23(2):299–323, 2010. doi: 10.1128/CMR.00068-09. URL http://cmr.asm.org/content/23/2/299.abstract.

M. Margulies, M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y.-J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. I. Alenquer, T. P. Jarvie, K. B. Jirage, J.-B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M.

Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley, and J. M. Rothberg. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437:376–380, 2005. ISSN 0027-8424. doi: 10.1038/nature03959. URL `http://dx.doi.org/10.1038/nature03959`.

A. Masella, A. Bartram, J. Truszkowski, D. Brown, and J. Neufeld. Pandaseq: paired-end assembler for illumina sequences. *BMC Bioinformatics*, 13(1):31, 2012. ISSN 1471-2105. doi: 10.1186/1471-2105-13-31. URL `http://www.biomedcentral.com/1471-2105/13/31`.

M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45: 167–256, 2003. doi: 10.1137/S003614450342480.

J. Oksanen, F. G. Blanchet, R. Kindt, P. Legendre, P. R. Minchin, R. B. O'Hara, G. L. Simpson, P. Solymos, M. H. H. Stevens, and H. Wagner. *vegan: Community Ecology Package*, 2013. URL `http://CRAN.R-project.org/package=vegan`. R package version 2.0-10.

B. P. O'Sullivan and S. D. Freedman. Cystic fibrosis. *The Lancet*, 373(9678):1891 – 1904, 2009. ISSN 0140-6736. doi: http://dx.doi.org/10.1016/S0140-6736(09)60327-5. URL `http://www.sciencedirect.com/science/article/pii/S0140673609603275`.

C. A. Petti, C. R. Polage, and P. Schreckenberger. The role of 16s rrna gene sequencing in identification of microorganisms misidentified by conventional methods. *Journal of Clinical Microbiology*, 43(12):6123–6125, 2005. doi: 10.1128/JCM.43.12.6123-6125.2005. URL `http://jcm.asm.org/content/43/12/6123.abstract`.

E. Pruesse, J. Peplies, and F. O. Glöckner. Sina: accurate high throughput multiple sequence alignment of ribosomal rna genes. *Bioinformatics*, 2012. doi: 10.1093/bioinformatics/bts252. URL `http://bioinformatics.oxfordjournals.org/content/early/2012/05/02/bioinformatics.bts252.a`

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. URL `http://www.R-project.org/`.

P. Ribeca and G. Valiente. Computational challenges of sequence classification in microbiomic data. *Briefings in Bioinformatics*, 12(6):614–625, 2011. doi: 10.1093/bib/bbr019. URL `http://bib.oxfordjournals.org/content/12/6/614.abstract`.

G. Rogers, S. Skeleton, D. Serisier, K. D. Bruce, and C. J. van der Gast. Determining cystic fibrosis-affected lung microbiology: Comparison of spontaneous and serially induced sputum samples by use of terminal restriction fragment length polymorphism profiling. *Journal of Clinical Microbiology*, 48(1):78–86, 2010. doi: 10.1128/JCM.01324-09.

S. J. Salipante, D. J. Sengupta, C. Rosenthal, G. Costa, J. Spangler, E. H. Sims, M. A. Jacobs, S. I. Miller, D. R. Hoogestraat, B. T. Cookson, C. McCoy, F. A. Matsen, J. Shendure, C. C. Lee, T. T. Harkins, and N. G. Hoffman. Rapid 16s rrna next-generation sequencing of polymicrobial clinical samples for diagnosis of complex bacterial infections. *PLoS ONE*, 8(5):e65226, 05 2013. doi: 10.1371/journal.pone.0065226. URL `http://dx.doi.org/10.1371%2Fjournal.pone.0065226`.

D. B. Sanders, R. C. L. Bittner, M. Rosenfeld, L. R. Hoffman, G. J. Redding, and C. H. Goss. Failure to recover to baseline pulmonary function after cystic fibrosis pulmonary exacerbation. *American journal of respiratory and critical care medicine*, 182(5):627–632, 2010. ISSN 1535-4970. doi: 10.1164/rccm.200909-1421OC.

F. Sanger, S. Nicklen, and A. R. Coulson. Dna sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12):5463–5467, 1997. ISSN 0027-8424.

M. B. Scholz, C.-C. Lo, and P. S. Chain. Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis. *Current Opinion in Biotechnology*, 23(1):9 – 15, 2012. ISSN 0958-1669. doi: http://dx.doi.org/10.1016/j.copbio.2011.11.013. URL http://www.sciencedirect.com/science/article/pii/S0958166911007245. Analytical biotechnology.

C. E. Shannon. A mathematical theory of communication. *SIGMOBILE Mob. Comput. Commun. Rev.*, 5(1):3–55, 2001. ISSN 1559-1662. doi: 10.1145/584091.584093. URL http://doi.acm.org/10.1145/584091.584093.

C. D. Sibley, M. E. Grinwis, T. R. Field, C. S. Eshaghurshan, M. M. Faria, S. E. Dowd, M. D. Parkins, H. R. Rabin, and M. G. Surette. Culture enriched molecular profiling of the cystic fibrosis airway microbiome. *PLoS ONE*, 6(7):e22702, 07 2011. doi: 10.1371/journal.pone.0022702. URL http://dx.doi.org/10.1371%2Fjournal.pone.0022702.

J. P. Stevens. *Applied Multivariate Statistics for the Social Sciences, Fifth Edition*. Routledge, 5th edition, Feb 2002.

P. Taberlet, E. COISSAC, F. POMPANON, C. BROCHMANN, and E. WILLERSLEV. Towards next-generation biodiversity assessment using dna metabarcoding. *Molecular Ecology*, 21(8):2045–2050, 2012. ISSN 1365-294X. doi: 10.1111/j.1365-294X.2012.05470.x. URL http://dx.doi.org/10.1111/j.1365-294X.2012.05470.x.

A. S. Tanabe and H. Toju. Two new computational methods for universal dna barcoding: A benchmark using barcode sequences of bacteria, archaea, animals, fungi, and land plants. *PLoS ONE*, 8(10):e76910, 10 2013. doi: 10.1371/journal.pone.0076910. URL http://dx.doi.org/10.1371%2Fjournal.pone.0076910.

D. Taylor-Robinson, M. Whitehead, F. Diderichsen, H. Olesen, T. Pressler, R. Smyth, and P. Diggle. Understanding the natural progression in %fev1 decline in patients with cystic fibrosis: a longitudinal study. *Thorax*, 67(10):860–866, 2012. ISSN 0040-6376. doi: 10.1136/thoraxjnl-2011-200953.

E. S. Tobias, M. Connor, and M. Ferguson-Smith. *Essential Medical Genetics*. Wiley-Blackwell, 6th edition, March 2011.

P. J. Turnbaugh, R. E. Ley, M. Hamady, C. M. Fraser-Liggett, R. Knight, and J. I. Gordon. The human microbiome project: exploring the microbial part of ourselves in a changing world. *Nature*, 449(7164):804 – 810, 2007. ISSN 0028-0836. doi: 10.1136/thoraxjnl-2011-200953. URL http://dx.doi.org/10.1038/nature06244.

J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, and R. J. e. a. Mural. The sequence of the human genome. *Science*, 291(5507):1304–1351, 2001. doi: 10.1126/science.1058040. URL http://www.sciencemag.org/content/291/5507/1304.abstract.

H. Wickham. *ggplot2: elegant graphics for data analysis*. Springer New York, 2009. ISBN 978-0-387-98140-6. URL http://had.co.nz/ggplot2/book.

W. Zhang, J. Chen, Y. Yang, Y. Tang, J. Shang, and B. Shen. A practical comparison of <italic>de novo</italic> genome assembly software tools for next-generation sequencing technologies. *PLoS ONE*, 6(3), 03 2011. doi: 10.1371/journal.pone.0017915. URL http://dx.doi.org/10.1371%2Fjournal.pone.0017915.