UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI BIOLOGIA

Corso di Laurea in Biologia Molecolare

ELABORATO DI LAUREA

Alterazione delle interazioni di *cross-feeding* nel microbiota intestinale umano in contesti patologici

Tutor: Prof. Stefano Campanaro
Dipartimento di Biologia

Laureanda: Alessandra Carpanese

ANNO ACCADEMICO 2023/2024

Sommario

L'intestino umano è popolato da centinaia di specie microbiche che nel complesso formano una rete integrata in cui ciascun microrganismo produce output metabolici che fungono da input nutrizionali per altre specie. Queste intricate relazioni trofiche prendono il nome di interazioni di *cross-feeding* le quali sono essenziali per il mantenimento dell'equilibrio dell'intera comunità microbica intestinale. La spiccata interdipendenza può rendere le specie microbiche vulnerabili a fenomeni di estinzione locale con conseguente perdita di partner cruciali e insorgenza di patologie legate ad uno sbilanciamento del microbiota. Tali relazioni però sono difficilmente quantificabili e rimangono scarsamente caratterizzate. In questo lavoro è stato introdotto un innovativo strumento bioinformatico che si è rivelato efficace ossia l'indice *Metabolite Exchange Score* (MES). Utilizzando genomi assemblati da 1661 campioni di metagenoma intestinale, il MES ha permesso di identificare e classificare le interazioni metaboliche significativamente influenzate conseguentemente ad un'alterazione dell'abbondanza relativa dei partner di *cross-feeding* in 10 patologie su 11. Questo approccio rappresenta dunque un potenziale strumento per orientare le future ricerche e sperimentazioni cliniche volte al ripristino delle interazioni tra i microrganismi intestinali e dunque potrebbe essere impiegata come una strategia promettente al fine di sviluppare terapie innovative mirate a ristabilire un sano ecosistema intestinale.

Indice

Stato dell'arte

1.1 Il microbiota intestinale umano

Il corpo umano è colonizzato da circa 10-100 trilioni di microrganismi simbionti che nel complesso costituiscono il microbiota umano. La sua composizione varia da un distretto anatomico ad un altro, dunque è possibile individuare comunità microbiche caratteristiche come quella che risiede sulla cute, nella cavità orale o nel tratto uro-genitale. L'intestino è la sede in cui è concentrata la maggior parte dei microbi, con una densità di circa 10^{13}-10^{14} cellule, di cui il 70% vive nel colon.
In questo compartimento sono comprese approssimativamente più di 1500 specie tra cui virus, archea, lieviti, funghi ma soprattutto batteri, che coesistono in equilibrio dinamico formando una complessa comunità ecologica chiamata microbiota intestinale. Sono stati individuati oltre 50 phyla di cui 6 sono quelli predominanti ossia *Bacteroidetes* e *Firmicutes* seguiti da *Proteobacteria*, *Fusobacteria*, *Tenericutes*, *Actinobacteria* e *Verrucomicrobia* che assieme rappresentano circa il 90% della popolazione microbica totale negli esseri umani.
Tra i funghi sono annoverati *Candida, Saccharomyces, Malassezia e Cladosporium* mentre tra gli Archaea spicca *Methanobrevibacter smithii*, il principale microrganismo metanogeno idrogenotrofo dell'intestino umano. Esso sintetizza metano CH_4 sfruttando substrati derivanti dal metabolismo microbico ossia l'anidride carbonica CO_2 come fonte di carbonio e l'idrogeno molecolare H_2 o il formiato come donatori di elettroni. I microrganismi intestinali svolgono diverse funzioni nei confronti dell'ospite, in particolare coadiuvano la digestione attuando la fermentazione degli alimenti introdotti con la dieta, offrono protezione contro i patogeni che penetrano nel nostro organismo, stimolano e modulano la risposta immunitaria, producono vitamine e altri composti benefici. Recentemente il microbiota intestinale è stato classificato come un vero e proprio "organo vitale" in grado di stabilire assi di collegamento che consentono un *cross-talk* con altri organi periferici regolandone la funzionalità e lo stato di omeostasi. Questi percorsi neurali, endocrini, umorali, immunologici e metabolici si sono rivelati cruciali per il mantenimento della salute dell'ospite infatti alterazioni della microflora intestinale sono state implicate nell'insorgenza di diverse patologie tra cui ipertensione, malattie cardiovascolari, obesità, diabete, malattie infiammatorie intestinali (IBD), cancro, ansia depressione e persino il morbo di Parkinson [1].
I microrganismi intestinali non interagiscono solo con l'ospite ma anche fra di loro, sia in modo intra- che inter-specifico, intessendo una complessa rete di relazioni che influenzano la loro crescita e sopravvivenza, e, in ultima istanza, la composizione della comunità stessa nel tempo e nello spazio. Tali interazioni inoltre inducono risposte sinergiche che spesso non possono essere previste studiando le singole specie in un contesto isolato. I membri di una comunità microbica interagiscono con modalità che possono essere classificate in termini di effetti sulla loro *fitness*: le interazioni negative comprendono la competizione e l'amensalismo mentre le interazioni positive si attuano mediante sfruttamento, commensalismo o mutualismo. Molte di queste dinamiche ecologiche sono mediate da molecole

diffusibili che possono agire come fonti di nutrienti, composti inibitori oppure possono avere una funzione di segnalazione [2].
Tra i meccanismi con cui si instaurano rapporti microbo-microbo sono compresi il *quorum-sensing*, l'inibizione per competizione tramite la produzione di batteriocine, e l'alimentazione crociata o *cross-feeding*. Il *cross-feeding* consiste nello scambio di metaboliti e risorse energetiche tra diverse specie o ceppi microbici e l'insieme di questi scambi metabolici dà origine ad una complessa rete trofica integrata che connette tutti i componenti di una comunità.
Questo meccanismo contribuisce profondamente a plasmare la struttura della comunità microbica intestinale inoltre determina il modo con cui reagisce a perturbazioni dell'ecosistema, di conseguenza influenza direttamente la stabilità della comunità stessa. Oltre a ciò, da esso dipendono anche il metaboloma integrato e le interazioni con l'ospite. L'alimentazione crociata nel microbioma intestinale può essere messa in atto mediante sfruttamento attraverso il quale un consumatore trae vantaggio da un substrato prodotto da un organismo partner, modificando al contempo l'ambiente in modo tale che diventi dannoso per il produttore, ad esempio secernendo un prodotto di scarto tossico. Il *cross-feeding* di tipo commensale si verifica quando un microbo si nutre di metaboliti prodotti da un altro partner senza esercitare alcun impatto su quest'ultimo. Infine, il *cross-feeding* mutualistico si attua quando due specie si nutrono ciascuna di un metabolita prodotto dall'altra, o quando un microbo sfrutta un metabolita sintetizzato da un altro partner e al contempo modifica l'ambiente in modo tale che sia in grado di apportare un vantaggio al produttore. Queste interazioni cooperative sono definite nel complesso con il termine di sintrofia che indica una forma di "metabolismo mutualistico obbligatorio" in cui i partner sfruttano substrati che nessuno dei due potrebbe metabolizzare da solo in un dato insieme di condizioni di crescita. L'amensalismo e la competizione possono anche verificarsi attraverso lo scambio di metaboliti, ma non rappresentano interazioni di alimentazione crociata.
I metaboliti oggetto di *cross-feeding* possono essere classificati in due gruppi: quelli utilizzati direttamente nel metabolismo centrale, tra cui zuccheri e donatori/accettori di elettroni, e quelli che rappresentano nutrienti essenziali come amminoacidi, cofattori e vitamine.
L'alimentazione crociata crea 4 livelli trofici discreti. Il primo è rappresentato dai degradatori primari i quali sono dotati di sistemi enzimatici in grado di idrolizzare polisaccaridi vegetali complessi sfuggiti alla digestione nell'intestino tenue, rilasciando oligo e monosaccaridi accessibili ad altre specie.
I fermentatori primari costituiscono il secondo livello trofico. Essi sono in grado di liberare essi stessi zuccheri più semplici oppure li acquisiscono dai degradatori primari, dopodiché li incanalano attraverso la glicolisi, producendo fosfoenolpiruvato (PEP) che viene utilizzato per generare acidi organici (ad esempio formiato, acetato, succinato) o alcoli (ad esempio 1,2-propandiolo). Scendendo al terzo livello trofico, i fermentatori secondari possono utilizzare questi sottoprodotti in vari percorsi fermentativi o respiratori generando acidi grassi a catena corta (SCFA), tra cui acetato, butirrato, propionato, e gas (CO_2 e H_2). Infine, l'idrogeno molecolare (H_2) prodotto dai fermentatori primari/secondari funge da

donatore di elettroni in molti pathway di crescita anaerobica e in particolare per i batteri solfato-riduttori (SRB), metanogeni e acetogeni (quarto livello trofico).
Comprendere l'entità delle interazioni mutualistiche di cross-feeding nel microbiota intestinale è fondamentale per spiegare alcune delle sue caratteristiche principali ossia la straordinaria diversità di specie che lo compongono e la notevole stabilità nel corso del tempo; infatti gli stessi taxa possono permanere per anni o addirittura per decenni [3]. La struttura della comunità microbica di un individuo comincia a definirsi durante i primi mesi di vita. Successivamente, in età adolescenziale viene raggiunto uno stato di equilibrio dinamico che poi tende a rimanere relativamente costante per tutta la vita dell'ospite. Tuttavia possono intervenire delle perturbazioni esterne considerevoli, come un'infezione da parte di un agente patogeno o un trattamento antibiotico, a seguito delle quali il microbiota intestinale è in grado di ripristinare la sua funzionalità e tornare allo stato originario grazie al fenomeno della resilienza. Dal punto di vista ecologico, questo concetto fa riferimento alla quantità di stress che un ecosistema può tollerare prima che il suo stato omeostatico si sposti verso un nuovo equilibrio; in altre parole è la capacità tampone che permette di tornare alla situazione originaria in risposta ad una perturbazione.
Il nuovo stato raggiunto può svolgere funzioni e servizi ecosistemici diversi, talvolta anche dannosi, nei confronti dell'ospite. Se la resilienza della comunità originale viene meno e il nuovo stato di equilibrio è associato ad un'alterazione della composizione della comunità o ad un'attività microbica aberrante, si parla di disbiosi. Questa condizione porta all'insorgenza di numerosi disturbi patologici, sia sistemici che locali, tra cui la sindrome metabolica e le malattie infiammatorie croniche intestinali come il morbo di Crohn (CD) e la colite ulcerosa.
Esiste una forte correlazione tra biodiversità e stabilità di una comunità, infatti tanto più elevato è il numero di specie che la compongono tanto maggiore è la sua resilienza. Ad ogni modo gioca un ruolo importante anche la ridondanza funzionale ossia la presenza di partner alternativi in grado di produrre gli stessi metaboliti. Ciò consente di sopperire ad un'eventuale mancanza di composti essenziali per la sopravvivenza di specie chiave ed evitare di conseguenza fenomeni di estinzione a catena che porterebbero ad un'erosione della biodiversità nell'ambito dell'ecosistema intestinale. Ripristinare la diversità dei partner microbici che si alimentano in modo crociato rappresenta un obiettivo importante, ma ancora in gran parte inesplorato, per combattere un'ampia gamma di patologie legate a un microbiota intestinale sbilanciato [4].

1.2 Analisi di comunità microbiche

In passato gli studi sul microbiota intestinale venivano condotti con tecniche tradizionali di isolamento e coltivazione, le quali tuttavia erano in grado di identificare solo uno spettro ristretto di microrganismi, pari a circa il 10%-30% delle specie totali presenti in un campione. La coltivazione *in vitro* e l'isolamento in condizioni controllate, rappresentano fasi sperimentali piuttosto difficoltose a causa della complessità, e quindi della scarsa riproducibilità, degli habitat naturali in cui vivono i microrganismi [5]. Un'ulteriore complicazione è data dalla pluralità di interazioni di *cross-feeding* menzionate nel paragrafo precedente.

Con l'avvento di tecnologie molecolari più avanzate, gli ecologi microbici si sono orientati sempre più verso una strategia basata sul *fingerprinting* molecolare.
A partire dal 1977, è stato riconosciuto l'enorme potenziale del gene codificante l'rRNA 16S e dell'*Intergenic Transcribed Spacer* (ITS) per dedurre relazioni filogenetiche, strumenti che in seguito sono stati sfruttati per caratterizzare la composizione filogenetica delle comunità microbiche naturali.
Negli anni '90 hanno fatto la loro comparsa i primi approcci di *fingerprinting* ambientale ad alto rendimento, che comprendono tecniche basate sull'elettroforesi come la *Denaturing Gradient Gel Electrophoresis* (DGGE), il *Terminal Restriction Fragment Length Polymorphism* (T-RFLP) e l'*Automated Ribosomal Intergenic Spacer Analysis* (ARISA).
Più o meno nello stesso periodo, sono stati introdotti i microarray come alternativa agli approcci basati sul *fingerprinting*. Uno svantaggio legato ai microarray sta nel fatto che l'identificazione è limitata alle sequenze note e incluse sull'array.
Sebbene questo aspetto limiti la sua applicazione come strumento di indagine su larga scala, questi strumenti possono comunque rappresentare strumenti preziosi per il biomonitoraggio clinico, industriale o ambientale. Poiché i microarray coprono varie regioni di un genoma, possono essere utilizzati per distinguere specie o ceppi strettamente correlati.
L'avvento delle tecniche di sequenziamento di nuova generazione (NGS) ha rappresentato una svolta nel campo dell'ecologia microbica in quanto ha permesso di caratterizzare in maniera sempre più raffinata e precisa la struttura delle comunità microbiche, sia in termini di ricchezza che di abbondanza relativa di specie.
Gli approcci basati sul sequenziamento *high-throughput* hanno dimostrato che l'ecosistema intestinale è caratterizzato da un'elevatissima complessità e biodiversità, più di quanto si sospettasse in precedenza [6].
Attualmente, gli approcci basati sul sequenziamento del DNA hanno soppiantato tutti gli altri e i più utilizzati per lo studio delle comunità microbiche sono l'analisi della sequenza del gene conservato rDNA 16S e la metagenomica.
Il primo approccio è consolidato per l'identificazione di batteri noti, tuttavia, essendo basato sulla disponibilità di sequenze di riferimento, non consente l'identificazione della maggior parte delle specie batteriche del microbiota intestinale in quanto molte rimangono ancora ignote. In secondo luogo, il sequenziamento 16S fornisce solamente informazioni indirette sulle funzioni dei membri di una comunità, inoltre non permette una risoluzione a livello di ceppo [5,6]. Il secondo approccio mira a catalogare tutti i geni di una comunità tramite il sequenziamento randomico di tutto il DNA estratto da un campione, pertanto fornisce informazioni complementari in merito al ruolo funzionale delle specie.
Questi metodi però non permettono di studiare tutte le interazioni tra i membri di una comunità microbica e questo aspetto risulta fondamentale per comprendere i processi che partecipano alla modulazione della sua struttura nel tempo [5].
Sebbene i rapporti di *cross-feeding* siano centrali nel network del microbiota intestinale, mancano studi quantitativi che rivelino la loro effettiva prevalenza.
Sono disponibili diverse strategie per la ricostruzione delle interazioni di alimentazione crociata. La modellazione ecologica si basa sulla simulazione di microbioti sintetici coltivati all'interno di bioreattori in batch e nella rimozione di

singole specie al fine di valutare l'impatto sugli altri membri. Questa procedura permette di ricostruire i rapporti di dipendenza nutrizionale fra i vari microrganismi e di determinare la gerarchia trofica multilivello di una comunità.
Un'altra strategia sfrutta modelli matematici basati su equazioni differenziali ordinarie (ODE) che permettono di quantificare l'entità delle interazioni e l'importanza relativa di alcuni membri rispetto ad altri. Uno dei modelli più semplici basati su ODE è l'equazione generalizzata di Lotka-Volterra (gLV) la quale permette di valutare l'effetto di un microrganismo sul tasso di crescita di un altro dunque analizza le interazioni a coppie e stabilisce se sono positive o negative.
Uno strumento emergente è rappresentato dai modelli metabolici su scala genomica o *Genome-scale metabolic Model* (GSMM). I GSMMs sono modelli che raccolgono e integrano tutte le informazioni metaboliche ottenute da diversi tipi di analisi genomiche e metabolomiche rappresentandole come un network che fornisce una panoramica di tutte le reazioni biochimiche che avvengono in un determinato sistema biologico [7]. Di conseguenza, permettono di prevedere gli effetti di un *knockout* genico o di variazioni dei parametri ambientali. Questi modelli sono attualmente utilizzati in un'ampia gamma di applicazioni, tra cui la progettazione di ceppi per la produzione industriale, la scoperta di molecole ad azione microbicida, lo studio di malattie legate a tratti metabolici e, più recentemente, l'analisi del *cross-feeding* nelle comunità microbiche [8].
Sono disponibili diversi software che consentono di eseguire simulazioni matematiche *in silico* dei *pathways* metabolici di Archaea, batteri ed organismi eucariotici. Questi programmi sfruttano le informazioni depositate in database (es. BiGG) che collezionano geni, reazioni, metaboliti e modelli noti associati a determinati processi biochimici e fisiologici o a specifici organismi, e consentono di definire quantitativamente le relazioni tra genotipo e fenotipo contestualizzando diversi tipi di dati omici come quelli di genomica, trascrittomica, proteomica e metabolomica. L'approccio di ricostruzione *bottom-up* prevede una serie di passaggi: (i) annotazione dei geni con funzioni metaboliche; (ii) identificazione delle reazioni biochimiche in cui sono coinvolti eseguendo una ricerca all'interno di database dedicati come ad esempio KEGG; (iii) assemblaggio di una bozza di rete metabolica; (iv) cura manuale della bozza, operazione molto dispendiosa in termini di tempo. Quest'ultima fase comprende varie procedure di *refinement* tra cui il *gap-filling* che consiste nell'aggiunta delle reazioni mancanti necessarie per generare tutti i precursori di biomassa, la correzione dell'equilibrio e della direzionalità delle reazioni, l'identificazione di cicli futili, e la rimozione di reazioni che non comportano la completa trasformazione dei reagenti in prodotti. Se questi problemi non vengono risolti, possono essere generate previsioni fenotipiche irrealistiche. I GSMMs sono stati utilizzati per stimare gli scambi metabolici tra coppie di specie batteriche per oltre un decennio. Recenti sviluppi nella generazione automatica di GSMMs e la disponibilità di altri modelli curati manualmente per migliaia di microrganismi intestinali, hanno aperto la strada alla creazione di modelli metabolici di intere comunità microbiche.
I progressi metodologici ora consentono di studiare le interazioni tra più specie e addirittura di ricostruire *pathways* metabolici direttamente da dataset metagenomici

su larga scala dunque, stanno divenendo strumenti sempre più potenti per esplorare microbioti complessi [7].
Nei prossimi paragrafi verrà descritto un lavoro scientifico che ha introdotto un nuovo sistema di quantificazione degli scambi metabolici derivati da modelli su scala metagenomica, il quale è stato progettato appositamente per identificare le potenziali interazioni di *cross-feeding* che risultano maggiormente alterate in condizioni patologiche.

Metodologie

2.1 Indagine globale sui metagenomi intestinali e controllo qualità

Indagine globale sui metagenomi. Prima di tutto è stata condotta una ricerca bibliografica al fine di collezionare studi sottoposti a *peer review* riguardanti metagenomi di feci umane con i relativi metadati.
Sono stati selezionati 33 studi pubblicati, relativi a 15 paesi e 11 fenotipi patologici, ed è stata eseguita un'analisi su larga scala di 1661 campioni di metagenoma intestinale di alta qualità, dei quali 871 provenienti da individui sani e 790 da soggetti malati. Sono stati esclusi lavori focalizzati su trattamenti dietetici e farmacologici, sull'esercizio fisico e quelli riguardanti individui di età inferiore a 10 anni. Per ridurre al minimo la variabilità legata alle tecnologie di sequenziamento sono stati inclusi solo gli studi in cui è stato eseguito sequenziamento *paired-end* utilizzando piattaforme Illumina HiSeq o NovaSeq.
La coorte sana includeva individui non affetti da alcuna patologia evidente, esenti da sintomi prodromici e con un indice di massa corporeo (BMI) compreso tra 18,5 e 24,9. Per quanto riguarda la coorte di individui malati, sono stati considerati diversi studi riguardanti 11 patologie diverse e da ciascuno è stato selezionato un massimo di 100 campioni. È stata scelta questa numerosità campionaria al fine di ridurre il dataset ad una dimensione computazionalmente sostenibile ma anche per minimizzare i *batch effects*[1].

Controllo qualità e trimming. Le sequenze grezze sono state scaricate dall'NCBI e sottoposte ad un controllo qualità e ad una procedura di *trimming* con il programma TrimGalore v.0.6.6 utilizzando una soglia di lunghezza minima pari a 80 bp e un punteggio Phred minimo di 25. Il *trimming* è una procedura necessaria al fine di rimuovere sequenze, o parti di esse, con una scarsa qualità o con elevata ambiguità che si trovano soprattutto vicino al sito di *annealing* dei *primers* di sequenziamento e verso le estremità delle sequenze. Questo processo permette anche di rimuovere sequenze non correlate come quelle degli adattatori Illumina o le sequenze di vettori contenute nei cloni di librerie di DNA e ciò risulta importante in quanto, se questi artefatti non vengono rimossi, potrebbero essere introdotte delle distorsioni nell'assemblaggio e di conseguenza nelle analisi a valle.

[1] I *batch effects* sono variazioni nei risultati sperimentali legati a differenze tecniche che si originano quando vengono considerati campioni indipendenti, generati separatamente, in tempi diversi, analizzati con tecniche diverse e manipolati da operatori differenti. Questi effetti si traducono in errori di valutazione che portano a conclusioni fuorvianti non correlate ad alcuna variabilità biologica.

Il *trimming* può essere statico, se le sequenze vengono tagliate tutte nello stesso punto, oppure può essere dinamico se invece vengono tagliate a partire dall'estremità 5' finché non viene raggiunta una certa soglia di qualità.
Successivamente, dopo il controllo di qualità, le *reads* contaminanti derivanti da sequenze umane sono state individuate e rimosse mediante il software Bowtie v.2.3.5 con il quale è stato eseguito un allineamento delle sequenze del metagenoma contro il genoma umano. Si tratta di un tool bioinformatico ultraveloce ed efficiente in termini di memoria che viene utilizzato per allineare le sequenze ottenute dal sequenziamento alle sequenze del database di riferimento.
Dopodiché sono stati eliminati tutti i campioni con meno di 15 milioni di reads al fine di ridurre al minimo l'impatto legato alla variabilità della profondità di sequenziamento tra i vari studi selezionati. A tale scopo è stato utilizzato il programma seqtk v.1.3 grazie al quale è possibile estrarre o rimuovere determinate sequenze da file FASTA o FASTQ. Alla fine è rimasto un set di dati che conteneva 1697 campioni.

2.2 Assemblaggio e classificazione del metagenoma

Assemblaggio de novo. Si tratta di una procedura essenziale nelle analisi di metagenomica, tuttavia risulta complicata a causa di diversi fattori quali il grande volume di dati prodotti, la qualità delle sequenze, l'ineguale rappresentazione dei membri delle comunità microbiche e dunque la disponibilità di una quantità insufficiente di dati per i membri meno rappresentati della comunità, la presenza di microrganismi strettamente correlati aventi genomi simili come ad esempio i ceppi della stessa specie. Questa procedura consiste nell'allineare tutte le *reads* e trovare regioni di *overlap* in modo da ricostruirne l'ordine sequenziale e formare frammenti continui di lunghezza maggiore. Dapprima le *reads* vengono riunite a formare dei *contigs* e questi vengono a loro volta concatenati a formare degli *scaffold* più lunghi. Per l'assemblaggio metagenomico vengono utilizzati gli assemblatori ossia strumenti bioinformatici dedicati con caratteristiche computazionali e prestazioni diverse a seconda del microbioma e dell'obiettivo sperimentale.
In questo studio è stato utilizzato il software MEGAHIT: si tratta di un assemblatore basato su grafi di de Bruijn sviluppato appositamente per eseguire il *de novo assembly* di dati metagenomici complessi e di grandi dimensioni. Questo programma consente di incrementare l'efficienza della procedura in termini di costi e tempo, e quindi di risorse computazionali richieste, inoltre opera un assemblaggio più completo e con una maggiore contiguità rispetto ad altri assemblatori noti. MEGAHIT alla fine fornisce come output una raccolta di assemblaggi sottoforma di contigs, uno per ciascun campione.

Binning e costruzione dei MAGs. Gli studi metagenomici si basano sull'analisi di tutto il DNA presente in una certa matrice e dunque portano all'ottenimento di *reads* relative ad un numero enorme di microrganismi. Il passo successivo è quello di cercare di ricostruire in modo quanto più accurato possibile i genomi rappresentativi di tutti i microrganismi presenti nel campione originario.
Questo può essere fatto mediante la procedura di *binning* che sfrutta algoritmi in grado di stimare la similarità degli *assemblies* in base a determinate caratteristiche

in comune e generano raggruppamenti discreti, denominati *bin*. L'obiettivo ultimo è quello di assegnare ciascuno *scaffold* al genoma, e quindi al gruppo tassonomico di origine. Dunque questo processo è essenziale per definire la biodiversità e la struttura di una comunità microbica, le funzioni di ciascun membro ma anche per identificare nuove specie.
I metodi bioinformatici di *binning* si basano su determinati criteri che includono la composizione nucleotidica delle sequenze in termini di GC *content* e di *codon usage*, la profondità di sequenziamento, il *coverage*, l'affiliazione filogenetica dei geni. Esistono anche metodi ibridi che combinano diversi criteri e altri basati sul *machine learning*. Nella maggior parte dei casi, questo processo risulta molto complesso a causa di una serie di problemi che complicano l'attribuzione delle sequenze ai *bin* corretti. Tra questi fattori si può menzionare l'elevata complessità dei campioni metagenomici in quanto contengono DNA proveniente da più organismi, la generazione di sequenze frammentate che rendono difficile l'assegnazione delle *reads* al *bin* corretto, il *coverage* non uniforme dei diversi genomi in quanto alcuni taxa sono più abbondanti di altri, il sequenziamento parziale di alcuni genomi che pertanto non vengono interamente rappresentati, il trasferimento genico orizzontale, la presenza di sequenze chimeriche risultanti da errori di sequenziamento o di assemblaggio, la variabilità genetica tra ceppi diversi. Pertanto, le tecniche di *binning* rappresentano il "migliore risultato possibile" per l'attribuzione delle *reads* o degli *assemblies* ad un determinato *bin*.
In questo studio, i 1697 campioni sono stati suddivisi in due gruppi per ridurre lo sforzo computazionale richiesto, dopodiché le *reads* filtrate per qualità sono state mappate rispetto a tutti i *contigs* assemblati per ciascun gruppo.
Per eseguire questo passaggio è stato utilizzato l'algoritmo minimap2: si tratta di un programma di allineamento a coppie sviluppato per mappare sequenze di DNA o di mRNA rispetto ad un ampio database di riferimento.
Successivamente è stato utilizzato l'autoencoder variazionale per il *binning* metagenomico (VAMB). Si tratta di un *binner* che utilizza un approccio di apprendimento non supervisionato basato su autoencoder variazionali (VAE)[2].
VAMB integra i dati di coabbondanza e di composizione dei k- mers a partire dai *de novo assemblies* di metagenomica e li raggruppa in *clusters* genomici con elevata coerenza tassonomica e in *bin* campione-specifici. In questo approccio ogni *cluster* corrisponde ad un microrganismo e ogni *bin* all'interno di un determinato *cluster* corrisponde ad una rappresentazione del genoma di quell'organismo per ciascun campione. Questo approccio sfrutta campioni multipli evitando al contempo la formazione di chimere tra campioni diversi.
Secondo gli sviluppatori, le performance di VAMB superano gli attuali approcci di *binning* a campione singolo infatti ha una maggiore capacità di ricostruire genomi quasi completi (NC); inoltre è in grado di clusterizzare in *bins* con elevata accuratezza garantendo una precisa profilazione tassonomica con una risoluzione che a volte arriva a livello di ceppo [9].

[2] I VAE sono modelli generativi di *machine learning* che apprendono rappresentazioni compresse dei dati di *training* originari sottoforma di distribuzioni di probabilità. Dunque apprendono modelli a variabili latenti continue i quali vengono utilizzati per generare nuovi dati creando delle variazioni di tali rappresentazioni apprese evitando al contempo un effetto di sovradattamento.

L'operazione di *binning* dunque è finalizzata alla creazione di *bin* ciascuno dei quali contiene un gruppo di *contigs* con caratteristiche simili e che idealmente sono attribuibili ad uno stesso genoma microbico. In altre parole un *bin* corrisponde ad un *Metagenome-Assembled Genome* (MAG) ossia un genoma microbico ricostruito dai dati metagenomici e che risulta rappresentativo di un determinato taxon.
Completezza e contaminazione dei MAGs risultano due aspetti di fondamentale importanza da tenere in considerazione. Per completezza si intende la presenza di tutti i geni rappresentativi di un determinato microrganismo mentre per contaminazione si intende la presenza di sequenze provenienti da microrganismi diversi. Nella maggior parte dei casi, a causa di limiti associati ai vari passaggi di sequenziamento e di analisi bioinformatica, un MAG non sarà quasi mai privo di contaminazione e non corrisponderà nella sua interezza al genoma microbico da cui deriva. È necessario pertanto stabilire se può essere considerato una buona approssimazione di un determinato genoma oppure se si tratta di un concatenamero di frammenti genomici di diverse origini. [10] Per valutare la qualità dei MAGs ricostruiti, è stato utilizzato il programma CheckM.
Alla fine sono stati ottenuti 55.345 *bin* tra cui 24.369 MAGs di alta qualità con >90% di completezza e <0,05% di contaminazione. Questi MAGs poi sono stati sottoposti ad un processo di de-replicazione che consiste nella rimozione di tutti i MAGs ridondanti e nella selezione del genoma più rappresentativo di ciascun taxon. La de-replicazione è stata eseguita stabilendo una soglia arbitraria di identità nucleotidica media (ANI) ossia una percentuale di nucleotidi condivisi tra due genomi. In tal caso è stata settata una soglia minima del 95%, la quale si ritiene sia quella più appropriata per una risoluzione a livello di specie. A tale scopo è stato utilizzato dRep v.3.0.0, un programma in grado di selezionare il genoma rappresentativo "migliore" in base a più parametri di qualità ossia completezza, contaminazione, eterogeneità a livello di ceppo, statistica N50[3], centralità[4].
La de-replicazione ha prodotto 955 genomi di alta qualità risolti a livello di specie. Il file FASTA contenente gli *assemblies*, è stato poi fornito come input al programma *Genome Taxonomy Database Toolkit* (GTDB-Tk) v.1.5.1 ossia un tool che fornisce una classificazione tassonomica automatizzata e oggettiva di genomi procariotici. La classificazione di un genoma *query* viene eseguita in base ad una combinazione delle seguenti informazioni: la posizione nell'albero filogenetico di riferimento depositato nel *Genome Taxonomy Database*, la divergenza evolutiva relativa (RED) e l'ANI. Questo strumento inoltre permette di visualizzare la prevalenza dei MAGs in tutti i campioni mediante la costruzione di un albero filogenetico visualizzabile con *interactive Tre Of Life* (iTOL), il tool online che consente la visualizzazione, l'annotazione e l'organizzazione di alberi filogenetici.

[3] La statistica N50 è una misura della qualità dell'assemblaggio in termini di contiguità. È la lunghezza del *contig* più corto che, assieme a quelli più lunghi, è compreso nel 50% della lunghezza totale dell'assemblaggio.
[4]La centralità è una misura di quanto un genoma sia simile a tutti gli altri genomi nel suo *cluster* e dunque fornisce una stima dell'abbondanza relativa dei diversi taxa.

GTDB-Tk implementa l'algoritmo Prodigal, che esegue operazioni di *ORF calling* identificando i geni codificanti proteine, e l'algoritmo HMMER che è in grado di allineare un set di marcatori appartenenti a batteri e Archaea e di assegnare i genomi al dominio con la più alta proporzione di marcatori in comune. Infine i genomi vengono collocati negli alberi filogenetici di riferimento. Parallelamente è stata calcolata l'abbondanza relativa delle varie specie mappando le reads sui 955 MAGs risolti a livello di specie grazie al software KMA v.1.3.1.

2.3 Modellazione metabolica su scala genomica e metagenomica

Ricostruzione dei GSMMs. Per ciascun MAG specie-specifico è stato elaborato un GSMM con il software CarveMe v.1.5 il quale adotta un approccio di ricostruzione *top-down*. A differenza degli altri programmi, CarveMe ricostruisce un modello universale che funge da templato il quale, dopo essere stato curato manualmente, risulta pronto per la simulazione. Successivamente, per ogni nuova ricostruzione, il modello universale viene convertito in un modello specifico riferito ad un determinato organismo mediante un processo chiamato "*carving*" che consiste nella rimozione di tutte quelle reazioni e metaboliti che non sono caratteristici del sistema biologico in esame. CarveMe ha una serie di vantaggi: innanzitutto non richiede lunghe procedure di *refining* manuale e permette di parallelizzare la ricostruzione di più specie simultaneamente. Oltre a ciò, l'approccio *top-down* è in grado di dedurre le capacità di assorbimento/secrezione di un organismo in modo contesto-indipendente e quindi basandosi esclusivamente su dati genetici. Questo aspetto rappresenta sia un vantaggio che uno svantaggio di CarveMe in quanto permette di semplificare la ricostruzione ed eseguire modellazioni metaboliche di microrganismi non coltivabili in terreni definiti, ma al contempo questa maggiore automazione è associata a prestazioni inferiori rispetto ai modelli curati manualmente [8].
Gli approcci *bottom-up*, sebbene più laboriosi, permettono di aggiungere iterativamente nuove reazioni o componenti durante il processo di cura manuale a seconda delle condizioni ambientali che vengono specificate dall'operatore, come ad esempio la composizione del terreno di coltura, permettendo così di ottenere ricostruzioni più accurate.
Infine, CarveMe consente di automatizzare la creazione di modelli estesi ad intere comunità microbiche riunendo modelli di singole specie
L'output corrisponde ad un file *Systems Biology Markaup Language* (SBML) che elenca in modo strutturato tutti i componenti della rete biochimica e le loro funzioni. Questo file può essere importato in qualsiasi strumento di simulazione per eseguire diversi tipi di modellazione matematica.

Flux Balance Analysis. In questo studio è stato utilizzato MICOM v.0.26, un software open source in grado di integrare i tassi di crescita delle diverse specie microbiche e simulare i flussi metabolici tra i membri di una comunità microbica tenendo conto delle loro abbondanze relative. In particolare, questo tool bioinformatico stima gli scambi metabolici eseguendo la cosiddetta *Flux Balance Analysis* (FBA): si tratta di un approccio matematico utilizzato per studiare i

pathways biochimici su scala genomica per ogni singolo taxon microbico a partire dal MAG più rappresentativo.
Questa analisi permette di fare delle previsioni sul tasso di crescita dei microrganismi ma anche sul tasso di produzione e di consumo di un metabolita.
Il primo passo della FBA è quello di rappresentare matematicamente le reazioni metaboliche sotto forma di modello numerico matriciale, definito matrice stechiometrica o matrice S, che riporta i coefficienti stechiometrici di ciascuna reazione. In particolare ogni riga rappresenta un metabolita e ogni colonna rappresenta una reazione. I substrati hanno segno negativo mentre i prodotti hanno segno positivo. Sono riportate inoltre altre colonne aggiuntive: nella prima sono riportate le reazioni di biomassa che simulano i metaboliti consumati durante la produzione di biomassa, nelle altre invece sono descritte le reazioni di scambio che rappresentano il flusso dei metaboliti dentro e fuori la cellula. L'importazione di un composto nell'ambiente intracellulare è rappresentata con un flusso negativo mentre il passaggio di un composto nell'ambiente extracellulare è un flusso positivo. In secondo luogo è necessario fissare dei vincoli in modo da definire uno spazio delle soluzioni differenziali definito; tra questi sono rilevanti il rispetto della stechiometria di reazione, la presenza di un pool di metaboliti intracellulari bilanciati (ipotesi di stato stazionario) e l'irreversibilità del flusso. Tutti questi vincoli devono rispettare le condizioni enzimatiche, termodinamiche e ambientali del contesto oggetto di studio.
Allo stato stazionario, vale la legge di conservazione della massa per cui il flusso attraverso ogni reazione rimane costante nel tempo perciò vale che $\mathbf{Sv} = \mathbf{0}$. Questa condizione definisce un sistema di equazioni lineari e poiché i modelli più complessi comprendono numerose reazioni, in genere esiste più di una possibile soluzione per ciascuna equazione.
Per prevedere il tasso di crescita massimo è necessario definire una funzione obiettivo $Z= c_x*v_x + c_y*v_y \ldots..c_n*v_n$ in cui c è un vettore di pesi che indica quanto ciascuna reazione v contribuisce all'obiettivo.
Infine viene utilizzata la programmazione lineare per identificare una distribuzione di flussi che massimizza o minimizza la funzione obiettivo all'interno dello spazio dei flussi ammissibili definito dai vincoli imposti dalle equazioni di bilancio di massa e dai limiti di reazione [11].
In questo lavoro i flussi sono stati stimati eseguendo una *parsimonious Flux Balance Analysis* (pFBA) che è una variante della FBA standard. Con questo approccio vengono selezionati i geni implicati nel metabolismo cellulare in base al loro contributo al tasso di crescita ottimale del microrganismo e all'entità del flusso biochimico che interessa i prodotti genici. La pFBA in pratica mira alla massimizzazione della produzione di biomassa riducendo al minimo il flusso totale attraverso una rete metabolica perciò permette di definire il percorso metabolico più parsimonioso. Il concetto alla base della pFBA è che le cellule con tassi di crescita più rapidi vengono selezionate positivamente per più generazioni perché possiedono un vantaggio in termini di *fìtness* in quanto utilizzano il minor numero e la minore quantità di enzimi. I modelli GSMMs generati con CarveMe contengono relativamente poche fonti di carbonio a causa del fatto che si tratta di un approccio che non richiede assunzioni a priori sulle condizioni del contesto.

I sistemi *bottom-up* invece consentono di testare diverse fonti di carbonio e simulare terreni di crescita complessi. Dunque utilizzando CarveMe si rischia di ottenere bassi tassi di crescita e conseguente instabilità numerica dei membri di una comunità. Infatti per considerare vitale un certo fenotipo il tasso di crescita deve essere almeno pari a 0,01 h^{-1} [8]. Pertanto, i flussi degli elementi presenti nel medium sono stati moltiplicati per un fattore di correzione pari a 600 in modo da calcolare più agevolmente gli scambi metabolici. Non è stata trovata una soluzione ottimale per 36 campioni, che sono stati rimossi dall'analisi, con conseguente riduzione del dataset finale a 1661 record.

2.4 Calcolo del *Metabolite Exchange Score* (MES)

In questo studio, i ricercatori hanno progettato il primo strumento bioinformatico che ha lo scopo di verificare la correlazione tra le interazioni di *cross-feeding* e l'insorgenza di una malattia, denominato *Metabolite Exchange Score* (MES).
Il MES si calcola come il prodotto del numero di taxa che potenzialmente consumano e il numero di taxa che potenzialmente producono un certo metabolita, normalizzato considerando il numero totale di taxa coinvolti. In altre parole corrisponde alla media armonica tra potenziali produttori e consumatori di un certo metabolita.

$$MES = 2 \times \frac{P \times C}{P + C}$$

È probabile che i metaboliti con MES elevati rappresentino componenti chiave nella catena alimentare microbica mentre quelli con un MES pari a zero non vengono né prodotti né consumati da alcun membro della comunità oppure vengono sintetizzati o consumati ma non scambiati tra i partners. Di conseguenza, confrontando i MES di ciascun metabolita tra microbioti di individui sani e malati, è possibile classificare e identificare i metaboliti più colpiti dalla perdita di partners di *cross-feeding* e le interazioni metaboliche chiave per il mantenimento dello stato di salute.
Una volta che a ciascun metabolita viene assegnata una priorità calcolando i MES, è possibile integrare le informazioni relative alle abbondanze dei taxa e ai loro flussi metabolici stimati per poi ottenere un consorzio di specie che agiscono come principali produttori o consumatori di determinati metaboliti.
Il MES si basa sul presupposto che un individuo in cui uno o più composti sono prodotti e consumati da numerosi membri del microbiota avrà una ridondanza funzionale più elevata e questa in genere è una caratteristica degli ecosistemi sani.
Per identificare i metaboliti significativamente alterati a causa di una perdita di partners di *cross-feeding* è stato utilizzato il test di Kruskal-Wallis[5], unitamente alla correzione di Bonferroni[6], confrontando la popolazione sana con quella fenotipicamente malata. Sono stati inclusi solo i metaboliti presenti in almeno 50 individui, di cui 15 malati, mentre acqua e ossigeno sono stati esclusi dalle analisi.

[5] Il test di Kruskal-Wallis è un test statistico non parametrico utilizzato per verificare l'uguaglianza delle mediane di diversi gruppi e quindi per stabilire se tali gruppi provengono da una stessa popolazione.
[6] La correzione di Bonferroni è un metodo statistico per ridurre la probabilità di commettere errori di tipo I nell'ambito di test di ipotesi multipli e dunque per minimizzare la probabilità di errore per confronto multiplo. Prevede di dividere il livello di confidenza α per il numero di test eseguiti. Ciò garantisce che la probabilità di commettere un errore di tipo I in tutti i test della famiglia rimanga al livello alfa desiderato o al di sotto.

2.5 Analisi statistica finalizzata alla valutazione degli effetti della diversità di specie sulle dinamiche dei flussi metabolici

Il risultato dell'analisi sulla diversità tassonomica ottenuto con il programma KMA, è stato utilizzato per calcolare l'indice di Shannon e determinare la ricchezza di specie in ciascun campione. Questi parametri sono stati poi utilizzati per quantificare la diversità α che esprime l'eterogeneità osservata in termini di numero e di abbondanza relativa di specie all'interno di uno stesso ambiente/campione.
Dopodiché è stato eseguito un confronto tra microbioti sani e malati utilizzando il test di Wilcoxon[7] corretto con il metodo Holm[8] per tener conto dei confronti multipli. I dati ottenuti sulla ricchezza delle specie sono stati successivamente utilizzati come misura della diversità di specie per analisi a valle. In particolare è stata eseguita un'analisi di regressione lineare tramite R per determinare la correlazione tra produttori o consumatori e la ricchezza di specie. La significatività statistica delle differenze tra le pendenze è stata corretta con il metodo Bonferroni.

2.6 Interazioni nutrizionali associate al morbo di Crohn

Validazione del MES. Gli autori hanno selezionato dal set di dati lo studio caso-controllo più corposo sul morbo di Crohn (CD) al fine di dimostrare la validità e l'applicabilità del MES per l'identificazione di promettenti target terapeutici.
Un totale di 84 campioni ha superato il controllo di qualità e sono stati inclusi nelle analisi, tra i quali 46 pazienti con CD e 38 controlli sani.
Uno dei metaboliti maggiormente implicati nelle IBD è l'idrogeno solforato (H_2S). Sebbene sia un importante mediatore gassoso con funzioni fisiologiche molto importanti, come la vasodilatazione e la risposta cardiaca a fenomeni ischemici, alterazioni dei suoi livelli sono implicati nell'insorgenza di quadri infiammatori cronici come quelli riscontrabili nelle IBD. L'H_2S può essere il risultato di una produzione endogena o un prodotto del metabolismo microbico, soprattutto da parte dei gruppi batterici appartenenti ai generi *Fusobacterium, Clostridium, Escherichia, Salmonella, Klebsiella, Streptococcus, Desulfovibri* ed *Enterobacter*. Le cellule epiteliali intestinali sono esposte ad entrambe le fonti, motivo per cui l'intestino deve essere efficiente nel regolare la concentrazione di H_2S [12].

Identificazione di potenziali target microbici coinvolti nel CD. Sono stati stimati i flussi di H_2S pesati per ciascuna specie microbica all'interno della coorte malata e di quella sana moltiplicando il flusso stimato di H_2S, espresso in millimoli all'ora per grammo di peso secco, per le abbondanze relative. In questo modo è stata calcolata la quota di H_2S prodotta o consumata da parte di ciascun taxon microbico. Dopodiché sono state valutate le differenze tra la diversità dei produttori e dei consumatori di H_2S, il rapporto tra produttori e consumatori e i loro flussi reciproci. La significatività di queste misure poi è stata valutata con il test di Kruskal-Wallis unitamente alla correzione di Bonferroni.

Analisi Hidden Markov Model (HMM). Per comprendere le basi genetiche della produzione e del consumo di H_2S, è stata condotta un'indagine utilizzando

[7] Il test Wilcoxon è un test non parametrico che viene utilizzato per confrontare due campioni accoppiati e per valutare se esiste una differenza statisticamente significativa tra di loro.
[8] Il metodo Holm è una modifica della correzione di Bonferroni meno conservativa e più potente.

HMMER v.3.3.2. Si tratta di un software che implementa gli *Hidden Markov Models* (HMM) ossia modelli statistici utilizzati per descrivere l'evoluzione di una serie di eventi osservabili che dipendono da fattori impliciti non direttamente apprezzabili, chiamati "stati". Un certo numero di stati nascosti forma una catena di Markov e il verificarsi di ciascuno stato genera un evento osservabile la cui probabilità di verificarsi viene calcolata sulla base dello stato sottostante.
Questi modelli sono sfruttati in diversi settori, dal riconoscimento vocale alla comunicazione digitale, tuttavia sono estremamente potenti anche nella rappresentazione di sequenze biologiche, ad esempio nella predizione della struttura genica e proteica, nella dinamica molecolare, negli allineamenti multipli e a coppie, nell'identificazione di RNA non codificanti ecc.
HMMER è un software dedicato per la ricerca di domini proteici conservati e sequenze omologhe di DNA all'interno di un dataset di riferimento. In particolare utilizza come "profilo" un allineamento multiplo che, nel caso delle proteine, è costituito da membri appartenenti ad una famiglia proteica. Dopodiché viene fornita come input una sequenza *query*, che in questo caso è rappresentata da un MAG, la quale viene confrontata con ciascuna delle sequenze appartenenti al profilo. HMMER infine attribuisce un punteggio per ciascun allineamento a coppie e maggiore è il suo valore e tanto più elevata è la probabilità che la *query* abbia una relazione di omologia con le sequenze del profilo.
In questo studio sono stati considerati 74 geni coinvolti nel ciclo del H_2S per i quali è stata ricercata un'omologia all'interno dei MAGs ricostruiti.
In seguito è stato utilizzato un modello lineare per verificare che questi geni fossero distribuiti in modo differenziale tra individui sani e affetti da CD. Le analisi sono state eseguite considerando sia l'abbondanza dei MAGs, ottenuta moltiplicando la conta dei geni per l'abbondanza delle specie, che la loro prevalenza in termini di presenza/assenza di specie. Quest'ultimo parametro è tanto più informativo quanto più i taxa rari sono responsabili di un'elevata produzione e consumo di H_2S.

Analisi delle componenti principali (PCA). È stato utilizzato il pacchetto mixOmics, un *toolkit* implementato in R, per condurre un'analisi delle componenti principali (PCA). Si tratta di una tecnica di semplificazione dei dati utilizzata in ambito della statistica multivariata che consente di ridurre la dimensionalità dei dati aggregando variabili di partenza. Questo consente di estrarre i fattori che impattano maggiormente sulla variazione dei dati, dunque si limita ad analizzare le variabili principali, ossia quelle con la varianza più elevata. Con questo tipo di analisi i ricercatori hanno determinato e rappresentato graficamente la diversità β, ossia la diversità esistente tra comunità microbiche differenti e quindi tra campioni differenti. In questo modo hanno potuto visualizzare la distanza relativa dei diversi campioni in base alla diversità della composizione microbica.

Analisi Random Forest. La separazione dei dati in gruppi discreti a volte non è così semplice ed evidente. Nel caso in cui la separazione tra gruppi risulti complessa, si utilizza il *Random Forest* (RF) ossia un algoritmo di *machine learning* che combina l'output di più strutture decisionali ad albero per raggiungere un'unica soluzione al problema principale. I singoli alberi decisionali iniziano con una domanda di base e poi procedono ponendosi una serie di domande secondarie che servono per suddividere i dati. Ad ogni quesito viene fornita una risposta e in questo modo vengono creati i cosiddetti nodi decisionali in ciascun albero. Le osservazioni che soddisfano i criteri imposti seguiranno il ramo "sì" e quelle che non lo fanno

seguiranno il percorso alternativo. Dunque le strutture ad albero cercano di trovare il modo migliore per suddividere i dati in gruppi omogenei e sono in genere addestrate attraverso l'algoritmo CART (*Classification and Regression Tree*). In questa analisi il 70% dei campioni è stato selezionato casualmente per il *training* del modello e il restante 30% è stato utilizzato per testare l'efficienza predittiva del modello. In tal modo sono state classificate le specie che spiegavano maggiormente la variazione tra microbioti sani e associati a CD.

Risultati

3.1 La meta-analisi di 1661 microbiomi ha rivelato interazioni metaboliche chiave tra i microrganismi intestinali

Dal processo di de-replicazione sono state rilevate nel complesso 949 specie batteriche e 6 specie di Archaea, che comprendono tutti i phyla microbici dominanti dell'ecosistema intestinale (Fig. 1a). Il mappaggio delle *reads* rispetto ai 955 MAGs ha evidenziato che le comunità microbiche contengono un range di 34-236 specie, con una media di 138. È stato scoperto altresì che 40 specie batteriche e una specie di Archaea sono presenti esclusivamente in individui malati, mentre 59 specie batteriche e una specie di Archaea sono state rilevate solo nei soggetti sani. Il calcolo del MES per ciascun metabolita, inoltre, ha mostrato una notevole variabilità inter-individuo (Fig. 1b).

Sono stati individuati gli scambi metabolici caratterizzati dalla maggiore diversità di partner di *cross-feeding* nella coorte di individui sani. I metaboliti con il MES medio più alto includevano basi azotate come uracile (60,5 ± 17,6) e timina (41,8 ± 21,8), nutrienti essenziali come fosfato (59,9 ± 17,0) e ferro, (40,3 ± 36,9) e zuccheri come glucosio (52,6 ± 22,1) e galattosio (52,3 ± 21,3).

Per identificare i metaboliti maggiormente interessati dalla perdita di partner di *cross-feeding* in diverse condizioni patologiche, sono stati confrontati i MES della coorte sana con quelli relativi a 11 fenotipi di malattia. Dall'analisi è stato osservato che in tutte le patologie è presente una discrepanza significativa rispetto al gruppo di controllo sano, ad eccezione per la schizofrenia.

I metaboliti con MES elevati e noti per la loro importanza nella salute umana, come la vitamina B1 (tiamina) e i precursori degli SCFA (malato, glucosio, galattosio), sono significativamente alterati in più fenotipi patologici. La tiamina è il metabolita con la differenza di MES più elevata tra microbioti sani e malati nella cirrosi e nella spondilite anchilosante, classificandosi al secondo posto dopo H_2S nelle IBD.

Le associazioni tra carenza di tiamina con cirrosi e IBD sono state segnalate in studi precedenti, ma non era mai stato indagato il possibile ruolo del microbiota intestinale dunque questa è la prima evidenza di una mediazione microbica in questi fenotipi patologici. Allo stesso modo, questo studio ha evidenziato per la prima volta un collegamento significativo tra il ribosil nicotinamide di origine microbica e l'artrite reumatoide. I risultati hanno anche confermato correlazioni già segnalate, come l'etanolo nel cancro del colon-retto e H_2S nelle IBD, rafforzando il potenziale del MES come nuovo approccio per identificare relazioni significative malattia-metaboliti microbici. Successivamente i risultati ottenuti sono stati confrontati con quelli dello studio di Zorrilla et al., in cui è stato utilizzato il tool *Species METabolic*

interaction ANAlysis (SMETANA) per quantificare gli scambi metabolici microbici nell'intestino e per rilevare una correlazione con l'intolleranza al glucosio e il diabete di tipo 2 (T2D). Lo studio di Zorrilla et al. ha identificato scambi significativamente diversi per 22 metaboliti, tra cui l'idrogeno solforato (H_2S) e il D-galattosio, che rientrano tra i composti aventi MES significativamente più elevati nei microbioti associati al T2D rispetto ai microbioti sani. È stata rilevata una certa concordanza anche per quanto riguarda i metaboliti scambiati più frequentemente, infatti tre dei sei metaboliti evidenziati da Zorrilla et al., sono tra i primi 15 con MES più elevato nei microbioti sani (L-malato, H_2S e acetaldeide).

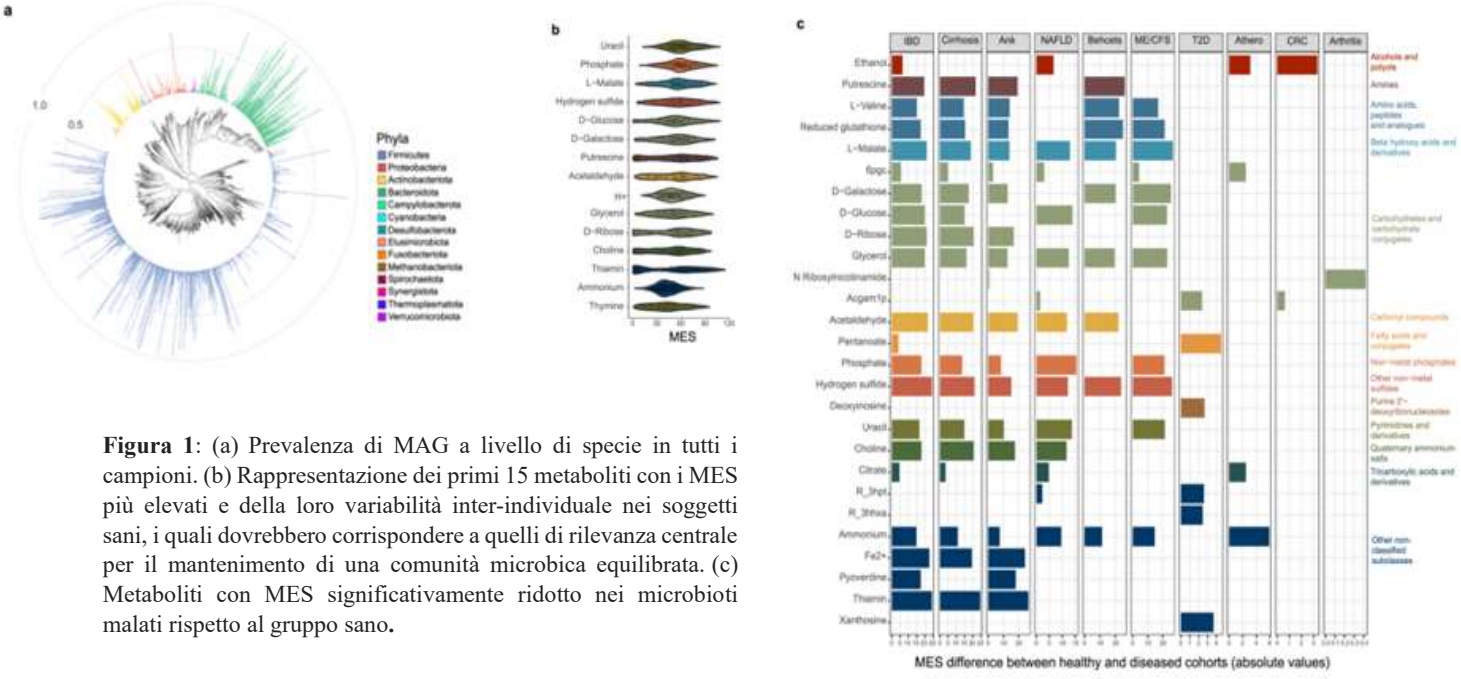


Figura 1: (a) Prevalenza di MAG a livello di specie in tutti i campioni. (b) Rappresentazione dei primi 15 metaboliti con i MES più elevati e della loro variabilità inter-individuale nei soggetti sani, i quali dovrebbero corrispondere a quelli di rilevanza centrale per il mantenimento di una comunità microbica equilibrata. (c) Metaboliti con MES significativamente ridotto nei microbioti malati rispetto al gruppo sano.

3.2 La diversità di specie è correlata in modo diverso con i produttori e i consumatori dei metaboliti scambiati

La diversità delle specie microbiche all'interno della comunità intestinale è comunemente considerata un indicatore dello stato di salute. I microbioti associati a 5 malattie hanno mostrato una riduzione significativa e costante della diversità alfa, come si evince dall'indice di Shannon e dalla ricchezza di specie.
Invece per i microbioti di individui affetti da T2D è stata rilevata una diversità alfa significativamente più elevata rispetto al gruppo di controllo sano.
Le malattie associate ad una bassa diversità di specie, come ad esempio le IBD, hanno evidenziato le massime differenze in termini di MES (Fig. 1c), il che è plausibile dato che il numero di partners che scambiano metaboliti è direttamente correlato al numero di specie presenti nella comunità.

Per comprendere la correlazione tra diversità di specie e il flusso di metaboliti, è stata testata l'ipotesi nulla H_0 secondo cui i produttori e i consumatori di ciascun metabolita sono ugualmente influenzati dalla diversità delle specie. Questa ipotesi implica che il numero di specie produttrici e di specie consumatrici aumentano proporzionalmente all'aumentare della ricchezza di specie e che quindi le interazioni di *cross-feeding* dipenderebbero esclusivamente dal numero di specie che compongono una comunità. È stata eseguita un'analisi di regressione per verificare la presenza di una correlazione tra il numero di specie produttrici/consumatrici di ciascun metabolita con la ricchezza di specie complessiva. Per capire se le differenze fra le pendenze erano statisticamente significative è stato eseguito un t test a due code con la correzione di Bonferroni. L'ipotesi nulla è stata respinta per il 79% dei metaboliti scambiati dai membri del microbiota intestinale (Fig. 2a), infatti la pendenza della correlazione è risultata significativamente maggiore sia per i consumatori (55% dei metaboliti) che per i produttori (24% dei metaboliti). Solo il numero di produttori e di consumatori di glicerolo non hanno mostrato una correlazione significativa con la ricchezza di specie (Fig. 2b-p).

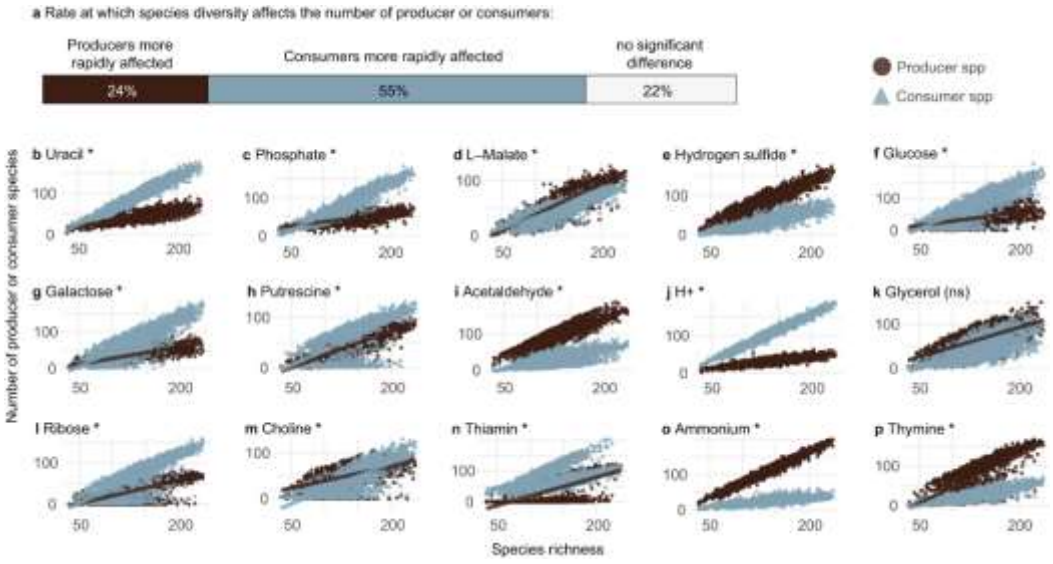


Figura 2: (a) Differenze tra le pendenze delle correlazioni tra ricchezza di specie e produttori o consumatori. (b – p) Rappresentazione della correlazione tra diversità di specie e produttori o consumatori per i primi 15 metaboliti con i MES più elevati nei microbioti sani.

3.3 Il ripristino della rete trofica del microbiota come potenziale strategia terapeutica per il trattamento del morbo di Crohn

Per indagare la potenziale applicabilità dell'approccio analitico basato sul MES nel guidare l'identificazione di promettenti target terapeutici, i ricercatori si sono concentrati sul morbo di Crohn. Il MES, in accordo con le evidenze scientifiche attuali, ha confermato che l'H_2S, è il metabolita più colpito dalla perdita di partner di *cross-feeding* nel microbiota di pazienti affetti da IBD e in particolare da CD.
Mentre la produzione di H_2S da parte del microbiota intestinale è stata oggetto di diversi studi, il consumo di questo gas invece è stato meno caratterizzato.

Sotto questo aspetto, i risultati ottenuti indicano che l'H_2S consumato dai batteri potrebbe essere incorporato in amminoacidi contenenti zolfo come la cisteina.
È stato scoperto che il microbiota di individui sani comprende un numero maggiore non solo di specie produttrici di H_2S, ma anche di specie in grado di consumarlo (Figura 3a). Dallo studio inoltre emerge che anche la diversità dei potenziali consumatori di H_2S risulta maggiormente influenzata nei pazienti affetti da CD (56% in meno di diversità), rispetto alla diversità dei produttori di H_2S (32% in meno di diversità), di conseguenza il rapporto produttori/consumatori di H_2S è significativamente più elevato negli individui affetti da CD (Figura 3c).
Risultati simili sono stati ottenuti studiando il flusso di H_2S tra i diversi microrganismi. In particolare è stato osservato che il consumo di H_2S nello stato di malattia si è ridotto del 74%, mentre la produzione totale non è stata significativamente influenzata, con conseguente rapporto produzione/consumo di H_2S più elevato negli individui affetti (Fig. 3b, 3d).
L'analisi della presenza e della distribuzione di 46 geni coinvolti nel ciclo dell'H_2S, ha evidenziato che 5 di questi erano significativamente più rappresentati nei microbioti sani (*cysK*, *dcm*, *Fuso_cyst*, *metH* e *metK*). Altri 5 geni invece erano più abbondanti nei microbioti associati a CD (*asrA*, *asrB*, *asrC*, *dmsA* e *dsrC*).
Per identificare le specie chiave associate allo squilibrio di H_2S nel CD, è stato confrontato il contributo di ciascuna specie alla produzione o al consumo totale di H_2S sia nelle coorti sane che in quelle CD.
In particolare per ogni specie, è stato stimato il flusso di H_2S ponderato in base alle abbondanze relative, ed è stata calcolata la differenza tra gli individui sani e CD.
Le specie che mostravano il più alto incremento verso la produzione di H_2S nei pazienti CD includevano membri delle classi Clostridia, Bacteroidia e Bacilli. Tra questi *Enterocloster clostridioformis* (Clostridia) ed *Enterococcus*_B *faecium* (Bacilli) sono stati osservati solo nella coorte CD (Fig. 3e).
Il 45% delle specie dello studio caso-controllo, hanno mostrato capacità sia di produrre che di consumare H_2S secondo le ricostruzioni metaboliche e il ruolo che assumono dipende dal contesto.
Phocaeicola dorei (Bacteroidia) è stata la specie che ha mostrato la differenza più elevata nella produzione di H_2S tra individui sani e CD, nonostante fosse comune in entrambe le coorti. I membri della classe Clostridia invece sono i consumatori di H_2S che hanno evidenziato la maggiore riduzione nel consumo di H_2S nei microbioti CD, tra cui *Roseburia intestinalis, Blautia obeum* e due specie di *Faecalibacterium* (Fig. 3e). I risultati ottenuti da questo approccio innovativo di modellazione metabolica sono stati confrontati con le tradizionali analisi composizionali del microbioma. Comparando i risultati dell'analisi delle componenti principali (PCA), è stato osservato che i microbioti associati al CD formano un cluster distinto. Per identificare le specie che hanno contribuito maggiormente alle differenze tra microbioti sani e malati, è stato utilizzato un classificatore *random forest* che ha assegnato loro un punteggio di importanza. Alcune delle specie identificate con l'analisi RF sono state identificate anche con l'approccio di modellazione metabolica sviluppato dagli autori di questo studio, inclusi i consumatori di H_2S *Roseburia intestinalis*, *Escherichia coli* e *Anaerostipes hadrus* e il produttore di H_2S *Clostridium symbiosum*. Tuttavia le altre 16 specie

delle 20 complessivamente individuate dall'approccio basato sul MES coinvolte maggiormente nello squilibrio tra produzione e consumo di H_2S nel CD, non rientrano tra le prime 30 specie rilevate dall'analisi composizionale (Fig. 3e). Questo indica che il MES è in grado di identificare un maggior numero di correlazioni tra metaboliti microbici e processi patologici rispetto le tradizionali analisi basate sulla composizione delle comunità.

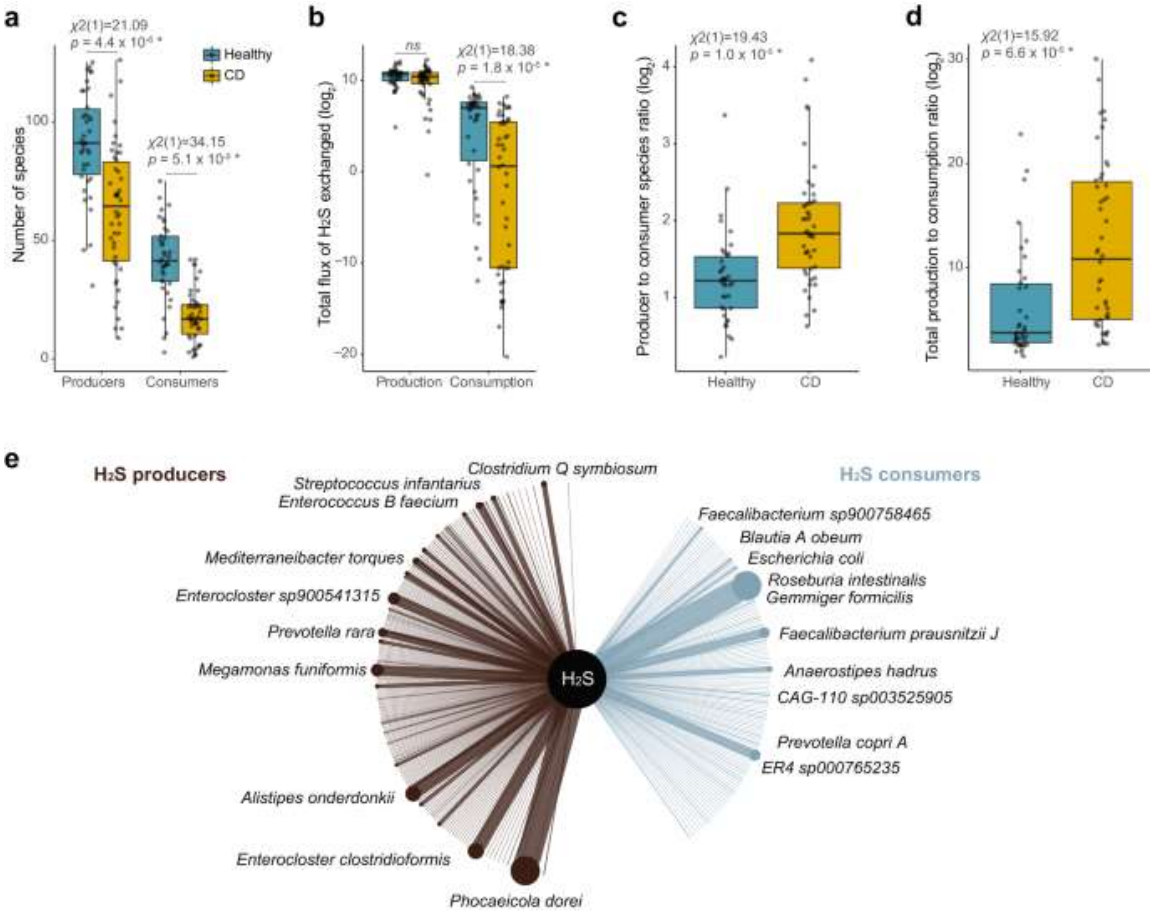


Figura 3: Differenze tra microbioti associati a CD e controlli sani (a) nel numero di specie produttrici e consumatrici di H_2S (b) nel consumo e nella produzione totale di H_2S (c) nel rapporto tra produttori e consumatori di H_2S (d) nel rapporto tra produzione totale e consumo totale di H_2S. (e) Panoramica delle specie microbiche che hanno subito le alterazioni più consistenti nella produzione e nel consumo di H_2S.

Conclusione e discussione

Questo studio ha rivelato un significativo impoverimento delle interazioni di *cross-feeding* nei microbioti intestinali associati a 10 malattie inoltre ha identificato promettenti target terapeutici in uno studio caso-controllo sul morbo di Crohn.
È stato dimostrato che il MES rappresenta un valido strumento analitico in grado di identificare associazioni microbiota-malattia sia note che nuove, fornendo una strategia semplice ed efficace in grado di guidare le sperimentazioni cliniche.
Un esempio di nuova associazione malattia-metabolita scoperta con questo approccio è quella tra artrite reumatoide e ribosil nicotinamide. Questo metabolita è uno dei principali precursori del nicotinamide adenina dinucleotide (NAD^+), che è stato segnalato essere significativamente ridotto negli individui con artrite reumatoide. La somministrazione di ribosil nicotinamide e di altri precursori del NAD^+ porta ad un miglioramento dei sintomi nei pazienti affetti da artrite reumatoide e da altre malattie infiammatorie, neurodegenerative e cardiovascolari.
Questa sarebbe la prima evidenza riportata in letteratura di un ruolo del metabolismo microbico nell'artrite reumatoide.
Questo approccio inoltre ha identificato l'associazione già nota in letteratura tra etanolo e cancro al colon-retto (CRC) infatti si tratta del metabolita più interessato dalla perdita di partners di *cross-feeding* negli individui colpiti da questa patologia. Un consumo di alcol da moderato a pesante è associato ad un rischio più elevato di 1,17-1,44 volte di sviluppare CRC tramite un processo che è in parte mediato dal microbiota intestinale, infatti alcuni batteri metabolizzano l'etanolo producendo l'acetaldeide che è un noto composto cancerogeno.
La capacità di identificare correlazioni significative tra un metabolita e una malattia utilizzando direttamente dati metagenomici è un ulteriore vantaggio del MES. Inoltre è in grado di rivelare un numero maggiore di associazioni rispetto i metodi basati esclusivamente sul confronto della composizione del microbiota tra soggetti sani e malati. È stata osservata anche una certa complementarietà tra l'approccio basato sul MES e i punteggi ottenuti col software SMETANA. I metaboliti identificati come marcatori di diabete di tipo 2 erano tra i metaboliti con MES più elevati nella popolazione sana, il che supporta l'idea che il loro scambio sia una caratteristica importante dei microbioti sani.
Questo studio inoltre ha messo in luce che la dipendenza dei microrganismi dalle interazioni di *cross-feeding* sia influenzata dalla dieta e quindi dalla disponibilità di input nutrizionali nell'ambiente intestinale. Diversi metaboliti con una significativa differenza di MES tra individui sani e malati si trovano negli alimenti (ad esempio, vitamine e zuccheri) e ciò evidenzia l'importanza della dieta come fattore in grado di influenzare gli scambi trofici nell'ambito del microbiota intestinale.
È interessante notare che per molti metaboliti (ad esempio, fosfato, glucosio, galattosio e colina), è stata osservata un'alta percentuale di produttori quando la diversità delle specie è bassa, tuttavia la frazione di consumatori supera la proporzione dei produttori man mano che aumenta la ricchezza delle specie.
Tale effetto potrebbe essere associato ad un deficit di metaboliti disponibili per il consumo e ciò favorirebbe la colonizzazione da parte di specie autosufficienti nella produzione di questi composti in condizioni patologiche. Un'elevata diversità di

specie, d'altro canto, è probabilmente collegata a una maggiore produzione netta di metaboliti da parte della comunità, offrendo maggiori opportunità alle specie consumatrici di prosperare. Questa ipotesi è coerente con due studi recenti che indicano che i microbioti associati alle IBD, che in genere hanno una bassa diversità di specie, sono arricchiti di batteri con genomi che codificano percorsi completi per la sintesi e il metabolismo di aminoacidi essenziali e vitamine (inclusa la tiamina), mentre i microbioti di individui sani sono arricchiti di batteri che per la loro sopravvivenza fanno affidamento sul *cross-feeding* di metaboliti essenziali [13, 14]. Queste evidenze, unitamente ai risultati di questo studio, suggeriscono un'ampia dipendenza dal *cross-feeding* nei microbioti sani. Per quanto riguarda il CD, le evidenze scientifiche suggeriscono che i soggetti affetti hanno una comunità microbica priva di membri in grado di supportare un sano equilibrio di H_2S.
I risultati di questo studio corroborano recenti scoperte che farebbero presupporre che il microbioma dei pazienti con IBD sia particolarmente carente di specie che secernono metaboliti contenenti zolfo, inoltre indicano che le specie consumatrici di H_2S vengono perse in modo preponderante nel CD.
Gli scambi microbici di H_2S possono influenzare l'ospite direttamente attraverso meccanismi come la modulazione del pH luminale, o indirettamente attraverso effetti a cascata sulla composizione del microbioma.
Ad ogni modo, l'accuratezza dell'approccio basato sul MES è limitata dall'uso di ricostruzioni metaboliche automatizzate su scala genomica, le quali sono in grado di rappresentare fenotipi vicini ai modelli curati manualmente ma sono incapaci di prevedere tutti i tratti fenotipici di un certo microrganismo o le reazioni del suo metabolismo secondario, specialmente se si basano su geni e *pathways* biochimici che devono ancora essere caratterizzati. Inoltre è importante tenere presente che solo la frazione di procarioti per i quali vengono ricostruiti MAGs di alta qualità può essere inclusa nei modelli metabolici su scala genomica inoltre le analisi sono state eseguite a livello di specie (95% ANI), il che potrebbe non rilevare differenze di metabolismo tra ceppi diversi. Pertanto esiste un margine per poter perfezionare questo approccio e renderlo più informativo combinando informazioni composizionali a livello di ceppo. Questo consentirà l'identificazione di biomarcatori dello stato di salute individuale e la comprensione dell'ecologia di queste complesse comunità intestinali. Un ulteriore miglioramento delle analisi basate sul MES è possibile attraverso l'incremento della disponibilità di modelli metabolici curati manualmente, l'integrazione di dati "omici" aggiuntivi e di informazioni sulla dieta e sul metabolismo dell'ospite, l'utilizzo di metodi di apprendimento automatico addestrati su dati composizionali per simulare in modo sempre più preciso intere comunità microbiche. I modelli metabolici elaborati da informazioni metagenomiche, abbinati ad una valutazione delle interazioni di *cross-feeding* microbico, aiuteranno ad abbattere uno dei principali ostacoli nello sviluppo di terapie del microbioma: stabilire la priorità su quali specie o metaboliti agire. Concentrandoci sul ripristino di aspetti chiave dell'ecologia intestinale, potremmo essere in grado di introdurre cambiamenti più efficaci e duraturi nel microbioma intestinale umano e mantenere lo stato di salute nel tempo.

Bibliografia

[1] Human gut microbiota in health and disease: Unveiling the relationship. Muhammad Afzaal, Farhan Saeed, Yasir Abbas Shah, Muzzamal Hussain, Roshina Rabail, Claudia Terezia Socol, Abdo Hassoun, Mirian Pateiro, José M. Lorenzo, Alexandru Vasile Rusu, e Rana Muhammad Aadil. Frontiers in microbiology, 2022.
[2] Microbiota in health and diseases. Kaijian Hou, Zhuo-Xun Wu, Xuan-Yu Chen, Jing-Quan Wang, Dongya Zhang, Chuanxing Xiao, Dan Zhu, Jagadish B. Koya, Liuya Wei, Jilin Li & Zhe-Sheng Chen. Nature, Signal Transduction and Targeted Therapy, 2022.
[2] A review of computational tools for generating metagenome-assembled genomes from metagenomic sequencing data. Chao Yang, Debajyoti Chowdhury, Zhenmiao Zhang a, William K. Cheung, Aiping Lu, Zhaoxiang Bian, Lu Zhang. Computational and Structural Biotechnology Journal, 2021.
[3] Cross-feeding in the gut microbiome: Ecology and Mechanisms. Elizabeth J. Culp, Andrew L. Goodman. Cell Host Microbe, 2023.
[4] The resilience of the intestinal microbiota influences health and disease. Felix Sommer, Jacqueline Moltzau Anderson, Richa Bharti, Jeroen Raes & Philip Rosenstiel. Nature Reviews Microbiology, 2017.
[5] Application of metagenomics in the human gut microbiome. Wei-Lin Wang, Shao-Yan Xu, Zhi-Gang Ren, Liang Tao, Jian-Wen Jiang, Shu-Sen Zheng. World Journal of Gastroenterology, 2015.
[6] Analysing Microbial Community Composition through Amplicon Sequencing: From Sampling to Hypothesis Testing. Luisa W. Hugerth, Anders F. Andersson. Frontiers in Microbiology, 2017
[7] Modeling approaches for probing cross-feeding interactions in the human gut microbiome. Pedro Saa, Arles Urrutia, Claudia Silva-Andrade, Alberto J. Martín and Daniel Garridoa. Computational and Structural Biotechnology Journal, 2022.
[8] Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. Daniel Machado, Sergej Andrejev, Melanie Tramontano, and Kiran Raosaheb Patil, Nucleic acid research, 2018.
[9] Improved metagenome binning and assembly using deep variational autoencoders. Jakob Nybo Nissen, Joachim Johansen, Rosa Lundbye Allesøe, Casper Kaae Sønderby, Jose Juan Almagro Armenteros, Christopher Heje Grønbech, Lars Juhl Jensen, Henrik Bjørn Nielsen, Thomas Nordahl Petersen, Ole Winther, Simon Rasmussen. Nature Biotechnology, 2021.
[10] Metagenome-assembled genomes: concepts, analogies, and challenges. João C. Setubal. Biophysical Reviews, 2021.
[11] What is flux balance analysis? Jeffrey D Orth, Ines Thiele & Bernhard Ø Palsson. Nature Biotechnology, 2010.
[12] Hydrogen sulfide toxicity in the gut environment: Meta-analysis of sulfate-reducing and lactic acid bacteria in inflammatory processes. Dani Dordević, Simona Jančíková, Monika Vítězová, Ivan Kushkevych. Journal of advanced research, 2021.
[13] Metabolic independence drives gut microbial colonization and resilience in health and disease. Andrea R. Watson, Jessika Füssel, Iva Veseli, Johanna Zaal

DeLongchamp, Marisela Silva, Florian Trigodet, Karen Lolans, Alon Shaiber, Emily Fogarty, Joseph M., Christopher Quince, Michael K. Yu, Arda Söylev, Hilary G. Morrison, Sonny TM Lee, Dina Kao, David T. Rubin, Bana Jabri, Thomas Louie, A. Murat Eren. Genome Biology, 2023.
[14] Microbes with higher metabolic independence are enriched in human gut microbiomes under stress. Iva Veseli, Yiqun T. Chen, Matthew S. Schechter, Chiara Vanni, Emily C. Fogarty, Andrea R. Watson, Bana A. Jabri, Ran Blekhman, Amy D. Willis, Michael K. Yu, Antonio Fernandez-Guerra, Jessika Fussel, A. Murat Eren bioRxiv, 2023.

Altre consultazioni

- Metatranscriptomics-guided genome-scale metabolic modeling of microbial communities. Guido Zampieri, Stefano Campanaro, Claudio Angione and Laura Treu. Elsevier, Cell Reports Methods, 2023
- Genome-Scale Metabolic Modeling Enables In-Depth Understanding of Big Data. Anurag Passi, Juan D. Tibocha-Bonilla, Manish Kumar, Diego Tec-Campos, Karsten Zengler, and Cristal Zuniga. Metabolites, 2022.
- MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Dinghua Li, Chi-Man Liu, Ruibang Luo, Kunihiko Sadakane, Tak-Wah Lam. Bioinformatics, 2015.
- Minimap2: pairwise alignment for nucleotide sequences. Heng Li. Bioinformatics, 2018.

nature communications

Article

https://doi.org/10.1038/s41467-023-42112-w

Disease-specific loss of microbial cross-feeding interactions in the human gut

Received: 14 February 2023

Accepted: 27 September 2023

Published online: 20 October 2023

Check for updates

Vanessa R. Marcelino[1,2,3,4] ✉, Caitlin Welsh[5], Christian Diener[6], Emily L. Gulliver[1,2], Emily L. Rutten[1,2], Remy B. Young[1,2], Edward M. Giles[2,7], Sean M. Gibbons[6,8,9,10], Chris Greening[5] & Samuel C. Forster[1,2] ✉

Many gut microorganisms critical to human health rely on nutrients produced by each other for survival; however, these cross-feeding interactions are still challenging to quantify and remain poorly characterized. Here, we introduce a Metabolite Exchange Score (MES) to quantify those interactions. Using metabolic models of prokaryotic metagenome-assembled genomes from over 1600 individuals, MES allows us to identify and rank metabolic interactions that are significantly affected by a loss of cross-feeding partners in 10 out of 11 diseases. When applied to a Crohn's disease case-control study, our approach identifies a lack of species with the ability to consume hydrogen sulfide as the main distinguishing microbiome feature of disease. We propose that our conceptual framework will help prioritize in-depth analyses, experiments and clinical targets, and that targeting the restoration of microbial cross-feeding interactions is a promising mechanism-informed strategy to reconstruct a healthy gut ecosystem.

The human gut contains hundreds of microbial species forming a complex and interdependent metabolic network. Over half of the metabolites consumed by gut microbes are by-products of microbial metabolism[1] with the waste of one species serving as nutrients for others[2–4]. Species interdependence can render microorganisms vulnerable to local extinction if a partner is lost[5] unless alternative species are available to fill that niche. In this context, having functionally redundant species with the ability to produce or consume the same nutrients is beneficial for the host. While it is generally accepted that high functional redundancy is a characteristic of resilient human gut microbiomes[6–8], the human health impacts of redundancy in metabolic interactions remain largely uncharacterized. Restoring the diversity of cross-feeding microbial partners represents a logical but still largely unexplored rubric to fight a wide range of diseases linked with an unbalanced gut microbiome.

Mechanistic models that simulate microbial metabolism in silico hold the promise to fill our knowledge gap on microbial metabolic interactions[4,9]. Genome-scale metabolic models (GEMs) are based on increasingly comprehensive databases linking genes to biochemical and physiological processes[10,11]. These models have been used to estimate metabolic exchanges between pairs of bacterial species for over a decade[12,13]. Developments in automating the reconstruction of GEMs[14] and the availability of manually-curated GEMs for thousands of gut microorganisms[15,16] have paved the way to build metabolic models for complex microbial communities. Methodological advances now allow modelling interactions between multiple species[17,18], and a recently developed workflow by Zorrilla and colleagues[19] now allows reconstructing metabolic models directly from large-scale metagenome datasets. Studies using community-wide metabolic models have found dozens to hundreds of significantly different metabolic

[1]Department of Molecular and Translational Sciences, Monash University, Clayton, VIC 3168, Australia. [2]Centre for Innate Immunity and Infectious Diseases, Hudson Institute of Medical Research, Clayton, VIC 3168, Australia. [3]Melbourne Integrative Genomics, School of BioSciences, University of Melbourne, Parkville, VIC 3010, Australia. [4]Department of Microbiology and Immunology at the Peter Doherty Institute for Infection and Immunity, University of Melbourne, Parkville, VIC 3010, Australia. [5]Department of Microbiology, Biomedicine Discovery Institute, Clayton, VIC 3800, Australia. [6]Institute for Systems Biology, Seattle, WA 98109, USA. [7]Department of Paediatrics, Monash University, Clayton, VIC 3168, Australia. [8]Department of Bioengineering, University of Washington, Seattle, WA 98195, USA. [9]Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA. [10]eScience Institute, University of Washington, Seattle, WA 98195, USA. ✉e-mail: vrmarcelino@gmail.com; sam.forster@hudson.org.au

exchanges in the gut microbiome associated with type 2 diabetes[19] and in inflammatory bowel disease[20] when compared to healthy controls. A method to rank these metabolic interactions according to an ecology-based framework provides the opportunity to generate targeted hypotheses underlying mechanistic links between the gut microbiome and diseases.

Here, we introduce a metabolite exchange scoring system derived from metagenome-scale metabolic models, designed to identify the potential microbial cross-feeding interactions most affected in disease. We apply our conceptual framework to an integrated dataset of 1661 publicly available stool metagenomes, encompassing 15 countries and 11 disease phenotypes. Our framework identified both known and novel microbiome-disease associations, including a link between colorectal cancer and the microbial metabolism of ethanol, a connection between rheumatoid arthritis with microbially-derived ribosyl nicotinamide, and links between Crohn's disease and specific bacteria that metabolise hydrogen sulfide. The scoring system can help quantify and identify context-dependent disruptions of microbial interactions, which may be targets for microbiome-based medicines.

Results

Potential cross-feeding interactions quantification

To understand the link between cross-feeding interactions and disease, we designed the Metabolite Exchange Score (MES). MES is the product of the diversity of taxa predicted to consume and taxa predicted to produce a given metabolite, normalized by the total number of involved taxa (Fig. 1a and methods). The potential production, consumption and exchange of metabolites by each microbiome member for which MAGs can be reconstructed is estimated through metabolic modelling. As with a centrality measure of a network that defines their most connected nodes, metabolites with high MESs are likely to be key components in the microbial food chain. At the other extreme, metabolites where MES is zero are not produced or not consumed by any member of the community. By comparing MESs for each metabolite across healthy and diseased microbiomes, one can rank and identify the metabolites most affected by the loss of cross-feeding partners (Fig. 1b). Once metabolites have been prioritized with MESs, it is then possible to integrate taxa abundances and their estimated metabolic fluxes to retrieve a consortium of species that act as the main producers or consumers of the targeted metabolites. We propose this approach as a hypothesis generation strategy to guide new discoveries, targeted experiments and clinical trials.

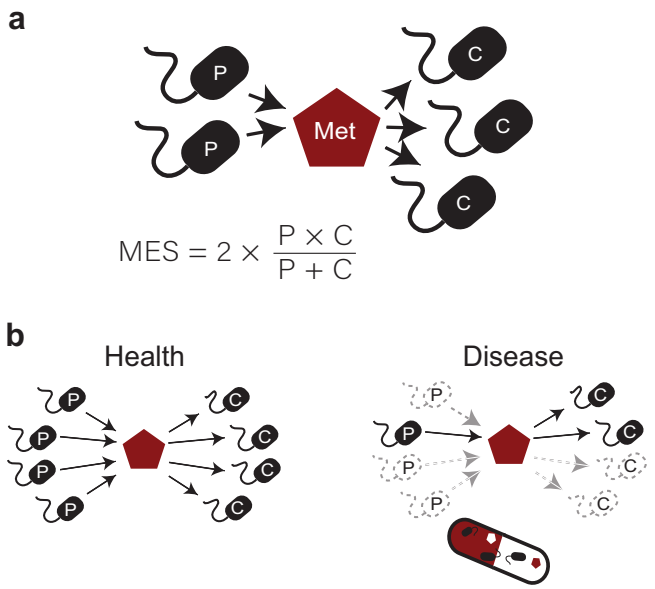


Fig. 1 | Overview of the Metabolite Exchange Score (MES) calculation and application. a MES is the harmonic mean between the number of potential producers (P) and consumers (C) inferred from metagenome-informed metabolic models. **b** Comparative analysis of MES between healthy and diseased cohorts can help identify the species and metabolites required to restore cross-feeding interactions, which may be promising targets of microbiome therapies.

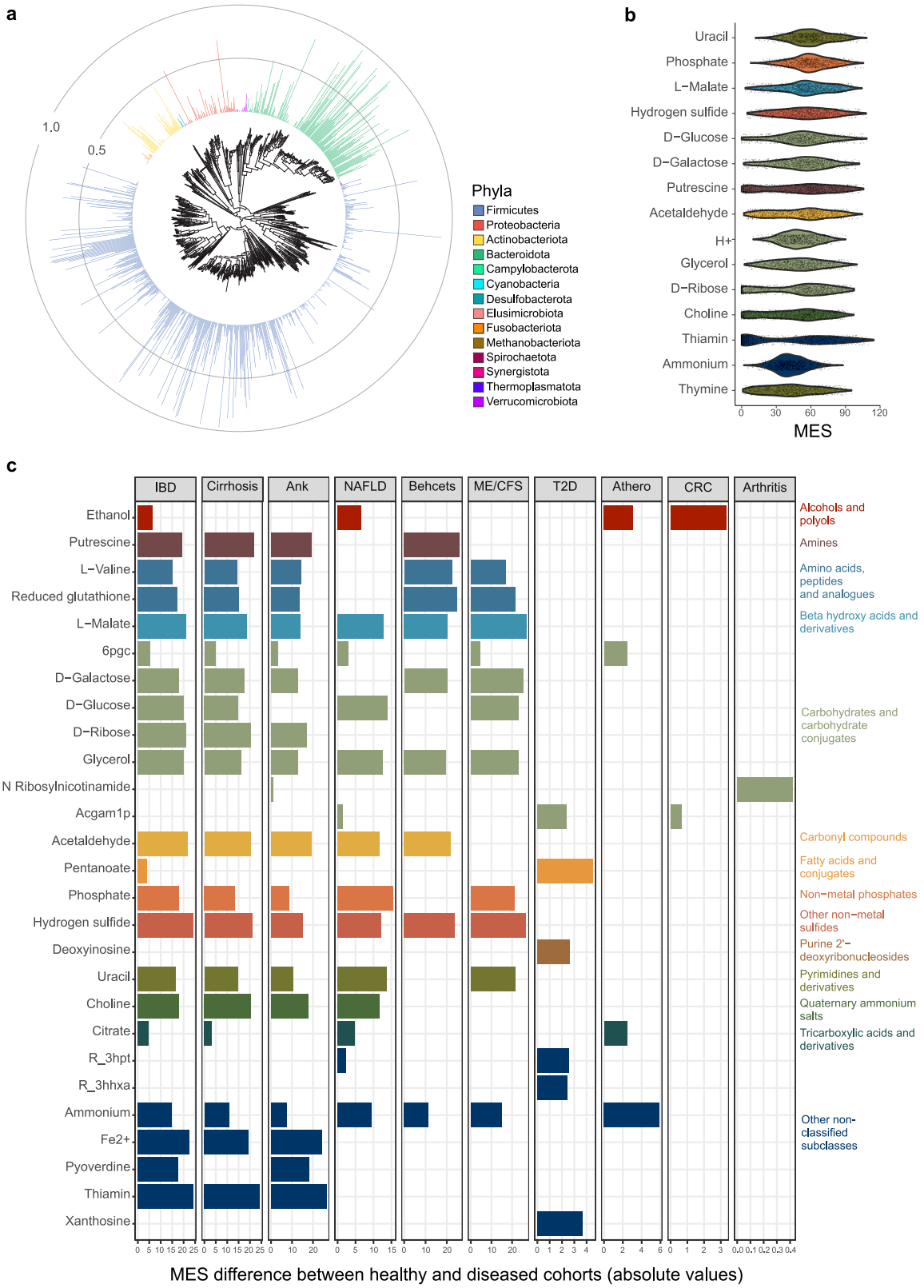
Meta-analysis of 1661 microbiomes reveals key metabolic interactions among gut microorganisms in health and disease

To obtain an overview of the association between cross-feeding interactions and different diseases, we performed a large-scale analysis of 1661 high-quality and deeply sequenced gut metagenome samples, including 871 healthy and 790 diseased individuals from 33 published studies, 15 countries and 11 disease phenotypes (Supplementary Data 1). Integrating studies and countries enabled the assembly of Metagenome-Assembled Genomes (MAGs) for a diverse range of gut microbes and allowed characterization of the baseline MESs in the healthy population. Our healthy cohort was composed of both males and females with a Body Mass Index (BMI) between 18.5 and 24.9 and no reported disease. Samples for which this information was unclear (e.g., disease controls where health status or BMI was not reported) are not included in our dataset (see Methods for details). Within-sample sequence assembly[21], metagenome co-binning[22] and quality control[23] resulted in 55,345 bins, including 24,369 high-quality MAGs with >90% completeness and <0.05% contamination. We selected one representative MAG per species, defined at 95% Average Nucleotide Identity (ANI), resulting in 949 bacterial and 6 archaeal species, encompassing all dominant microbial phyla found in the gut (Fig. 2a, Supplementary Data 2). The presence and abundance of these species were determined by mapping sequence reads against the 955 MAGs. Forty bacterial and one archaeal species were exclusively found in diseased individuals (Supplementary Data 3a), while healthy individuals harboured 59 bacterial and one archaeal species that were not observed in any diseased individual (Supplementary Data 3b). Identifying species in metagenome samples remains a challenge, and it is likely that our MAG-based approach misses rare components of the gut microbiome despite the large dataset used here for co-binning. To infer metabolic exchanges between microbes, we reconstructed Genome-Scale Models (GEMs)[14] for the 955 MAGs, built community-scale metabolic models for each individual based on the species-level abundances using MICOM[18], and calculated MES using custom scripts[24]. Our modelled communities contained an average of 138 species (min = 34, max = 236 species).

We first sought to identify the metabolic exchanges with the highest diversity of cross-feeding partners in healthy microbiomes by analysing the MESs of each metabolite of the entire healthy group. Metabolites showed a wide variation of MESs between individuals (Fig. 2b, Fig. S1). Metabolites with the highest mean MES included nucleobases such as uracil (MES mean and sd = 60.5 ± 17.6) and thymine (41.8 ± 21.8), essential nutrients such as phosphate (59.9 ± 17.0) and iron (40.3 ± 36.9), and sugars such as glucose (52.6 ± 22.1) and galactose (52.3 ± 21.3).

To identify the metabolites most affected by the loss of cross-feeding partners during disease, we compared MESs between the healthy group and the eleven disease phenotypes. This analysis identified significant loss of cross-feeding partners for specific metabolites in all disease groups except for schizophrenia (Fig. 2c, Fig. S2). Metabolites with high MESs in healthy individuals and known to be important for human health, such as vitamin B1 (thiamin)[25] and precursors of short-chain fatty acids (e.g., malate, glucose, galactose)[26], were significantly affected in multiple disease phenotypes (Kruskal–Wallis' $p < 0.05$/number of tests to correct for multiple comparisons). Thiamin was the metabolite with the highest difference in MESs between healthy and diseased microbiomes in cirrhosis and ankylosing spondylitis, ranking second in Inflammatory Bowel Disease (IBD) (Fig. 2c). Associations between deficiency

https://doi.org/10.1038/s41467-023-42112-w



of thiamine with cirrhosis and IBD have been previously reported[27–29], but to our knowledge, this is the first indication of a possible microbial-mediation of this phenotype. Likewise, this is the first indication of a link between microbially-derived ribosyl nicotinamide and rheumatoid arthritis (Fig. 2c). The results also confirmed previously reported microbially-mediated disease-metabolite associations, such as ethanol in colorectal cancer[30] and hydrogen sulfide in IBD[31,32], reinforcing the potential of our novel approach to identify reasonable relationships.

We next compared our results with the study of Zorrilla and colleagues[19], who used SMETANA[17] to quantify microbial metabolic exchanges in the gut and link those with glucose intolerance and type 2 diabetes (T2D). Their study identified significantly different exchanges for 22 metabolites, including for hydrogen sulfide (H_2S)

https://doi.org/10.1038/s41467-023-42112-w

Fig. 2 | Global analysis reveals most common metabolic exchanges among healthy gut microbes and disease-specific loss of cross-feeding partners. **a** Prevalence of species-level MAGs across all samples. **b** Top 15 metabolites with the highest MESs in healthy individuals, which are expected to be central to sustain a healthy microbial community structure. **c** Metabolites with significantly reduced MES in diseased microbiomes when compared to the healthy group (one-sided Kruskal–Wallis' $p < 0.05$/number of comparisons within each disease category), suggesting significant loss of microbial cross-feeding partners for those metabolites. The panel of metabolites shown here include the top 5 metabolites with the highest MES differences between healthy and diseased groups for each disease (metabolites with increased MES in diseased microbiomes are not included). No significant difference in MES was found in patients with schizophrenia ($n = 87$) after accounting for multiple comparisons. Sample sizes and Bonferroni-corrected p-value thresholds: IBD inflammatory bowel disease ($n = 63$, $p < 1.27 \times 10^{-4}$), liver cirrhosis ($n = 54$, $p < 1.30 \times 10^{-4}$), Ank ankylosing spondylitis ($n = 72$, $p < 1.32 \times 10^{-4}$), NAFLD non-alcoholic fatty liver disease ($n = 71$, $p < 1.25 \times 10^{-4}$), Behcet's disease ($n = 18$, $p < 2.21 \times 10^{-4}$), ME/CSF myalgic encephalomyelitis/chronic fatigue syndrome ($n = 17$, $p < 2.99 \times 10^{-4}$), T2D type 2 diabetes ($n = 32$, $p < 1.37 \times 10^{-4}$), Athero atherosclerosis ($n = 98$, $p < 1.18 \times 10^{-4}$), CRC colorectal cancer ($n = 143$, $p < 1.17 \times 10^{-4}$), Arthritis rheumatoid arthritis ($n = 135$, $p < 1.18 \times 10^{-4}$). Colours in **b**, **c** represent metabolite Sub Classes according to the Human Metabolome Database.

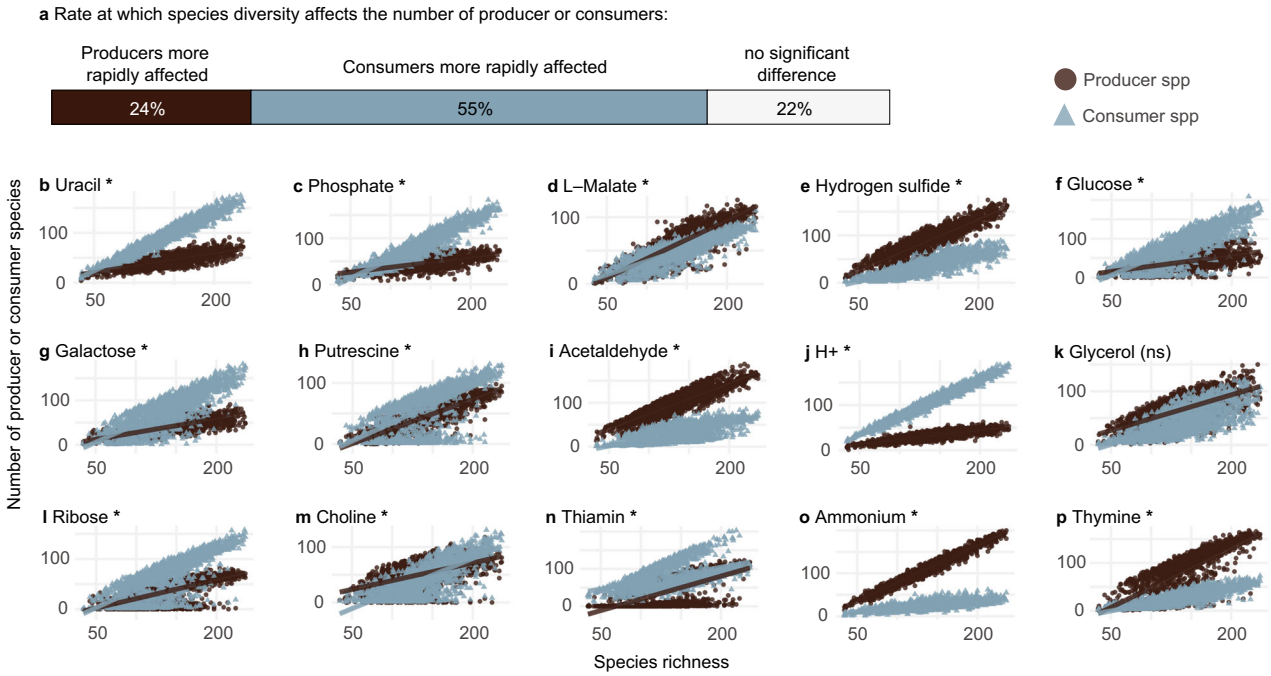


Fig. 3 | Producer to consumer dynamics is affected by species richness for most metabolites. **a** Significant differences between the slopes of the species richness vs producers or consumers correlations were observed for the majority of metabolites, with producers having a steeper slope in 24% of the metabolites, and consumers having a steeper slope in 55% of the metabolites analysed. **b**–**p** Representation of the correlation between species diversity vs producers or consumers for the top 15 metabolites with the highest MESs in healthy microbiomes. Analyses included all samples from our dataset ($n = 1661$, including healthy and diseased cohorts), and only metabolites exchanged within at least 50 microbiomes. Each subplot contains two points for each sample to represent the diversity of producers (brown circles) and consumers (blue triangles). Asterisks indicate a significant p value of the t-test associated with the linear regression model (two-sided) after Bonferroni correction (i.e., $p < 0.00011$).

and D-galactose, which were also identified in our analyses as having significantly higher MESs in T2D-associated microbiomes when compared to healthy microbiomes (Supplementary Data 4). There was also some concordance between our results regarding the metabolites identified as being most frequently exchanged between gut bacteria, with three out of the six metabolites highlighted in Zorrilla et al. (Fig. 3a in ref. 19), being among the top 15 metabolites with the highest MESs in healthy microbiomes (L-malate, H_2S and acetaldehyde).

Species diversity has distinct relationships with producers and consumers of exchanged metabolites

Diversity of microbial species within the gut community is commonly considered a marker of health status. Microbiomes associated with five diseases showed significant and consistent reduction in alpha diversity across indices (Shannon index and species richness), while microbiomes from individuals with type 2 diabetes had a significantly higher alpha diversity when compared with the healthy group (Fig. S3). Diseases associated with low species diversity (e.g., Inflammatory Bowel Disease) showed the highest magnitude in MES differences (Fig. 2c), which is expected given that the number of microbial species exchanging metabolites naturally correlates with the number of species in the community.

To further understand the relationship between diversity and metabolite exchange, we tested the null hypothesis that producers and consumers are equally affected by species diversity. Specifically, we correlated the number of producer or consumer species of each metabolite with species richness to determine statistical differences between the slopes of these correlations for metabolite production and consumption. The null hypothesis (no statistical difference between slopes) implies that the number of producer species and consumer species increases at the same rate as species richness increases. Such results would imply that cross-feeding interactions dependent only on the number of species present in the community. This null hypothesis was rejected for 79% of metabolites exchanged by the gut microbiome (Fig. 3a, Supplementary Data 5), with the slope of the correlation being significantly steeper either for consumers (55% of metabolites) or producers (24% of metabolites). From the metabolites

https://doi.org/10.1038/s41467-023-42112-w

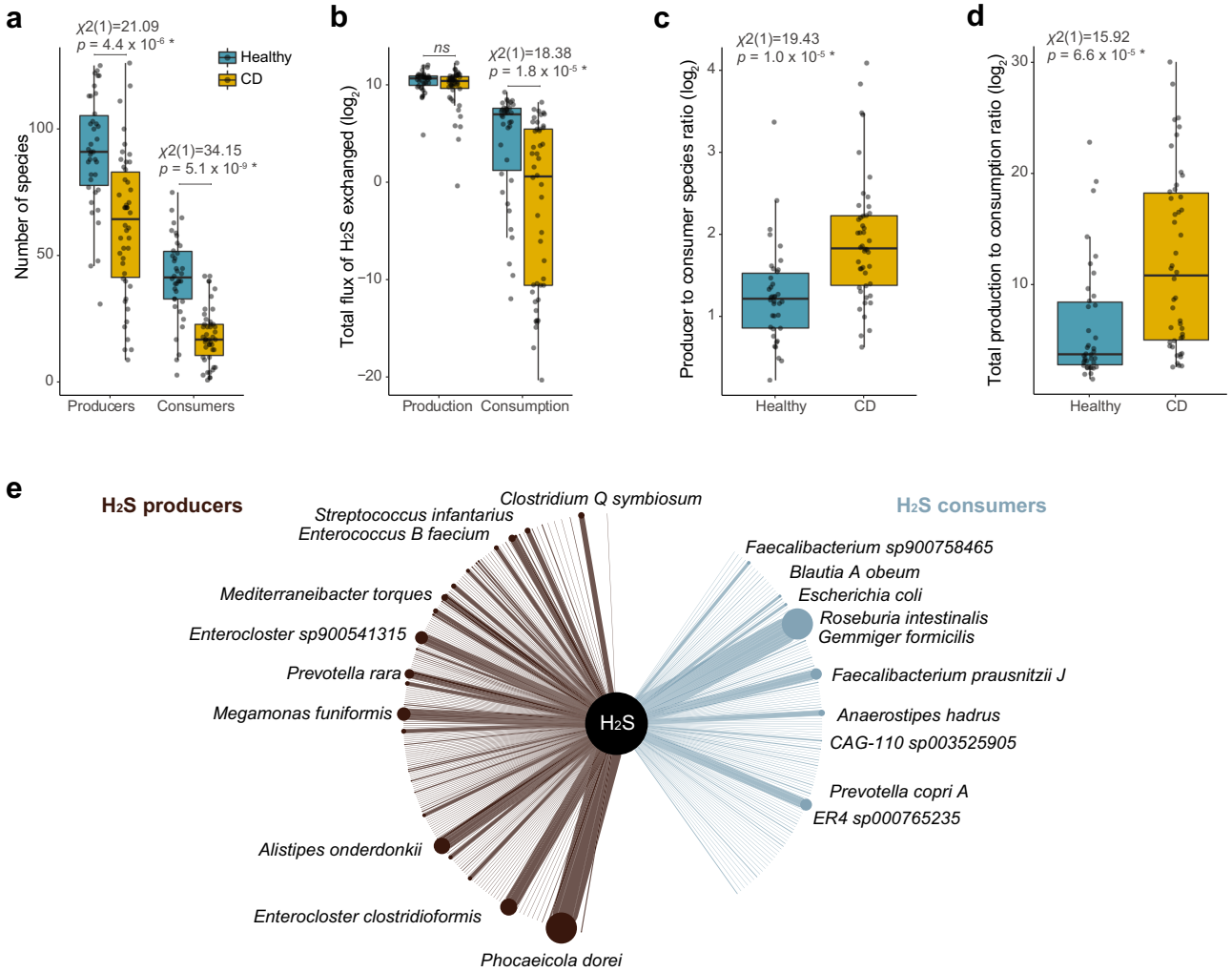


Fig. 4 | Shift in hydrogen sulfide production-consumption equilibrium associated with Crohn's disease. a The number of species with potential to produce or consume H_2S is significantly reduced in microbiomes associated with CD when compared to healthy controls. **b** The total estimated consumption of H_2S is depleted in CD, while production was not significantly affected (fluxes estimated in millimoles per hour per gram of dry weight). A significant increase in the ratio of number of producers to consumers (**c**) and in the total estimated H_2S production to consumption (**d**) was found in microbiomes associated with CD. **e** Species involved in the exchange of H_2S that are most altered in CD, which might be promising targets of microbiome therapy. The network shows the H_2S producers with increased production (brown), and the consumers with reduced H_2S consumption (blue) in CD when compared to healthy controls. The 10 species contributing most to H_2S production or consumption are highlighted. The thickness of the nodes and edges are proportional to the species' weighted flux sum of H_2S within the consumer or producer categories. Statistical tests in **a**–**d** were performed with a one-sided Kruskal-Wallis test, degrees of freedom = 1, $p < 0.05$ were considered significant. Box-plot elements in **a**–**d**: centre line = median; box limits = upper and lower quartiles; whiskers = 1.5× interquartile range; points = samples ($n = 84$ biologically independent samples).

with the highest MESs, only producers and consumers of glycerol showed no significant difference in response to species richness (Fig. 3b–p).

Microbial food web restoration as a potential therapeutic strategy for Crohn's disease

To investigate how the application of MES and our modelling framework may guide the identification of promising therapeutic targets, we focused on Crohn's disease (CD), a form of IBD. We selected a single case-control study[33] with the largest number of samples from healthy and diseased individuals within our quality-controlled dataset to minimize batch effects. In accordance with the global analyses, we found that H_2S – a gas previously implicated in CD and IBD symptoms[31,32,34]—was the metabolite most affected by the loss of cross-feeding microbial partners (twofold reduction, Supplementary Data 6). While H_2S production by the gut microbiome has been the subject of several studies (e.g., refs. 35,36), the consumption of this gas is less characterized, and our modelling results indicate that H_2S consumed by bacteria can be incorporated into sulfur-containing amino acids such as cysteine (Fig. S4).

Focusing on H_2S, we found that the microbiome of healthy individuals contained more species with the potential to produce H_2S, as well as more species with the potential to consume H_2S, than the microbiomes associated with CD (Fig. 4a). Interestingly, the diversity of potential H_2S consumers was more affected in CD patients (56% less diverse on average, Supplementary Data 7) than the diversity of H_2S producers (32% less diverse), resulting in a significantly higher H_2S producer to consumer ratio in individuals affected by CD (Fig. 4c). We observed similar results when investigating the flux of H_2S among microorganisms. The total estimated ability of the microbiome to consume H_2S in the disease state was reduced by 74%, while the total production was not significantly affected, resulting in a higher H_2S production to consumption ratio in CD (Fig. 4b, d, Supplementary Data 7). The excess of H_2S (i.e., H_2S

predicted to be exported to medium) was not significantly different between healthy and diseased subjects (Kruskal–Wallis $\chi^2(1) = 0.0356$, $p = 0.8503$). The indication that H_2S consumers are more affected than H_2S producers in CD stands after correcting for the confounding effects of species diversity, although no significant difference was observed for the flux of H_2S exchanged among microorganisms (Supplementary Data 8).

To better understand the genetic basis of the metabolic modelling results, we investigated the distribution of 46 genes known to be involved in H_2S cycling[36] in the MAGs present in the CD case-control study. We found between one and 23 genes in each MAG (Supplementary Data 9). Five genes involved in H_2S cycling were significantly more prevalent in microbiomes associated with healthy individuals (Supplementary Data 10): *cysK, dcm, Fuso_cyst, metH* and *metK* (linear model, using species diversity as confounder variable and a two-way *t*-test to assess significance, $p < 0.0012$ accounting for multiple comparisons). Another five genes were more prevalent in CD-associated microbiomes: *asrA, asrB, asrC, dmsA* and *dsrC* ($p < 0.0012$), the first four genes also being significantly enriched when accounting for species abundance (Supplementary Data 10).

To identify the key species associated with H_2S imbalance in CD, we compared the contribution of each species to the total H_2S production or consumption in the healthy and CD cohorts. For each species, H_2S flux (weighted by relative abundances) was estimated and the difference of total H_2S weighted flux in healthy and CD individuals calculated. The species showing the highest increase towards H_2S production in CD patients included members of the classes Clostridia, Bacteroidia and Bacilli (Fig. 4e, Supplementary Data 11). *Enterocloster clostridioformis* (Clostridia) and *Enterococcus_B faecium* (Bacilli) were only observed in the CD cohort. Many species (45% of the MAGs from the case-control study) showed an ability to both produce and consume H_2S according to the models, and their role was dependent on their community context. *Phocaeicola dorei* (Bacteroidia) was the species showing the highest difference in predicted H_2S production between healthy and CD individuals despite being common in both cohorts. We found multiple genes related to H_2S metabolism in this species (*cysK, bsh, dcm, Fuso cyst, luxS, metK, sufS*, and two copies of the *malY* and *metH* genes). Members of the Clostridia class were the H_2S consumers showing the highest reduction in H_2S consumption in CD microbiomes, including *Roseburia intestinalis, Blautia_A obeum*, and two *Faecalibacterium* species (*F. prausnitzii_J* and *F.* sp900758465) (Fig. 4e, Supplementary Data 11). The top 5 consumer species had between two and four copies of the cysteine desulfurase (*iscS*) gene, in addition to a range of other genes involved in H_2S metabolism (Supplementary Data 9 and 11).

We next compared the results obtained from our metabolic modelling approach with traditional compositional microbiome analyses. Community beta-diversity was visualized using principal component analysis, showing that microbiomes associated with CD formed a distinct cluster (Fig. S5a). To identify the species that contributed most to these differences we used a random forest (RF) classifier (70% of data used for training, 30% for testing). The out-of-bag error rate of the training dataset was 9.52%, and the accuracy on the test dataset was 100%. The species contributing most to the differences between healthy and CD-associated microbiomes were identified through their importance scores (Fig. S5b). Some of the species identified with the RF analysis were also identified with our metabolic modelling approach, including the H_2S consumers *Roseburia intestinalis, Escherichia coli and Anaerostipes hadrus*, and the H_2S producer *Clostridium_Q symbiosum*. Sixteen out of the 20 species identified by our modelling approach as contributing most to the H_2S production to consumption ratio unbalance in CD (Fig. 4e) were not among the top 30 species selected with this compositional-based analysis.

Discussion

In this work, we introduce a new MES-based conceptual framework and apply it to an integrated dataset of metabolic models for 955 gut species from 1661 publicly available stool metagenomes, encompassing 15 countries and 11 disease phenotypes. This approach revealed a significant depletion of potential cross-feeding interactions in the microbiomes associated with 10 diseases and identified promising therapeutic targets in a case-control Crohn's disease study.

We show that our analytical framework identifies both known and novel microbiome-disease associations, providing a cost-efficient and mechanistically grounded strategy to prioritize experiments and guide clinical trials. One example is the link between rheumatoid arthritis and ribosyl nicotinamide (also known as nicotinamide riboside or NR). This metabolite is one of the main precursors of nicotinamide adenine dinucleotide (NAD^+), which has been reported to be significantly reduced in individuals with rheumatoid arthritis[37]. Administration of NR and other NAD^+ precursors leads to improved clinical outcomes for rheumatoid arthritis patients[37] and for a range of other inflammatory, neurodegenerative and cardiovascular diseases[38]. To our knowledge, this is the first reported evidence for a role of microbial NR metabolism in rheumatoid arthritis. We also identified ethanol as the metabolite most affected by loss of cross-feeding in individuals with Colorectal Cancer (CRC). Moderate to heavy alcohol consumption is associated with a 1.17 – 1.44 higher risk of developing CRC[39] via a process that is at least partially mediated by the microbiome, as gut bacteria metabolise ethanol to produce the carcinogenic acetaldehyde[40]. The capacity to identify these and other coherent metabolite-disease links using exclusively metagenome data is further evidence for the validity and utility of our approach. Some associations observed in our study such as links between *Roseburia intestinalis* and CD could be retrieved using analyses based solely on the composition of the microbiome, but most associations could not (e.g., *Phocaeicola dorei*), with the modelling framework yielding additional insights on the metabolic and ecological processes underlying these associations. We also observed a complementarity between our MES approach and previously proposed methods based on SMETANA scores. Metabolites identified as markers of T2D progression[19] were among the metabolites with highest MESs in the healthy population, supporting the idea that the exchange of these metabolites is an important feature of healthy microbiomes.

The reliance of microbes on cross-feeding is expected to be influenced by the availability of metabolites in the gut environment. Several metabolites with significant MES difference in health and disease are found in food (e.g., vitamins and sugars), highlighting the importance of diet in understanding cross-feeding in the gut microbiome. Interestingly, for many metabolites (e.g., phosphate, glucose, galactose and choline), we observe a high proportion of producers when species diversity is low, but the proportion of consumers overtakes producers as species richness increases (Fig. 3). We speculate that low species richness is associated with a lack of metabolites available for consumption, favouring species that are self-sufficient in producing these metabolites. High species diversity, on the other hand, is likely linked to higher net metabolite production by the community, providing more opportunities for consumer species to thrive. This hypothesis is consistent with two recent studies indicating that microbiomes associated with IBD (which typically have low species diversity) are enriched in bacteria with genomes that encode complete pathways for the synthesis and metabolism of essential amino acids and vitamins (including thiamine), while microbiomes of healthy individuals are enriched with bacteria that are expected to rely on cross-feeding for essential metabolites[41,42]. These studies, together with our results, suggest an extensive reliance on cross-feeding in healthy and diverse microbiomes.

Using CD as a case study, we demonstrated how the modelling framework can help define mechanistically informed hypotheses for

targeted experimental and clinical validation. Our results suggest that CD patients lack microbial community members to support a healthy H_2S balance. This gas is expected to have a protective effect in the gut when present in small amounts, but it disrupts the mucus layer and may cause inflammation when present in larger quantities[43–46]. Our results corroborate recent findings suggesting that the microbiome of IBD patients is particularly deficient in secreting metabolites containing sulfur[20], and additionally indicate that H_2S consumer species are disproportionately lost in CD. Microbial exchanges of H_2S may affect the host directly through mechanisms such as modulating luminal pH[32], or indirectly through cascade effects on microbiome composition.

The accuracy of the modelling framework applied here is limited by the use of automated genome-scale metabolic reconstructions, which represent phenotypes close to manually-curated models[14] but are naturally unable to predict all organism-specific traits or secondary metabolism, especially if those rely on genes and pathways that are yet to be characterized. Automated genome-scale models provide an opportunity for a top-down approach, where large scale analyses like the one performed here can guide a range of more refined hypothesis-driven studies, ideally coupled with experimental validation. Additional refinement can be obtained in future studies handling smaller datasets by manual model curation, integration of additional 'omics data, e.g., ref. 47 and other lines of evidence (e.g., machine learning methods trained on compositional data), and by integrating personalized data on host diet and metabolism[48]. It is also important to note that only the prokaryotic fraction of the microbiomes for which high-quality MAGs were reconstructed could be included in the models and that our analyses were performed at the species level (95% ANI), which may miss strain-level differences in metabolism. Future research applying the MES approach in combination with strain-level compositional information will be highly informative to identify biomarkers of health status and to better understand the ecology of these complex gut communities.

We expect that metagenome-informed metabolic models, coupled with an assessment of microbial cross-feeding interactions, will help alleviate one of the main barriers in the development of microbiome therapies – prioritizing which species or metabolites to target. By focusing on restoring key aspects of the gut ecology, we may be able to introduce more effective and long-lasting changes in the human gut microbiome.

Methods

Global survey of gut metagenomes and quality control

We performed a literature search for peer-reviewed studies with publicly available human stool metagenomes and associated metadata. These included large-scale meta-analyses of gut metagenomes and metadata compilations[49,50]. Studies focusing on dietary interventions, medications, exercise and children (<10 years old) were excluded. For longitudinal studies, only one sample per individual was included in the analyses. To minimize the impact of sequencing technologies, only studies reporting paired-end sequencing using Illumina's HiSeq or NovaSeq platforms were included.

The healthy cohort included individuals reported as not having any evident disease or adverse symptoms[50]. Samples classified as disease controls and where the health status could not be determined were excluded. To avoid ambiguous health/disease status, samples from individuals with colorectal adenoma (non-cancerous tumour) and impaired glucose tolerance (pre-diabetes) were excluded, and only individuals with a Body Mass Index (BMI) between 18.5 and 24.9 were included in the healthy cohort. Samples with less than 15 M PE reads after quality control were excluded to minimize the impact of sequencing depth. A maximum of 100 samples per disease category from each study were used to minimize batch effects and reduce the dataset to a computationally feasible size.

Raw sequence reads were downloaded from NCBI and subject to quality control with TrimGalore v.0.6.6 (Krueger F. http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) using a minimum length threshold of 80 bp and a minimum Phred score of 25. Potential contamination with human sequence reads was removed by mapping the metagenome sequences to the human genome with Bowtie v.2.3.5[51]. To minimize the impact of sequence depth, samples were rarefied to 15 M fragments (30 M PE reads) with seqtk v.1.3 (https://github.com/lh3/seqtk). The quality-controlled dataset contained 1697 samples, which are provided along with their metadata in Supplementary Data 1.

Metagenome assembly and binning

Assembly was performed for individual metagenomes with Megahit v.1.2.9[21]. It has been shown that co-binning multiple samples yields a higher number of high-quality MAGs, but using co-abundance information requires significant computational resources[52]. We, therefore, divided the 1697 samples into two batches (indicated in Supplementary Data 1) and, for each of these batches, followed the steps recommended in the VAMB v.3.0.2[22] workflow. In short, we mapped quality-filtered sequenced reads against all contigs assembled within that batch with minimap2[53], and used VAMB to identify metagenome bins. The snakemake workflow for these steps (adapted from the VAMB github) is available in our Zenodo repository[24]. Completeness and contamination levels of metagenome bins were assessed with CheckM[23]. We retrieved 24,369 bins with >90% completeness and <0.05% contamination. These bins were dereplicated at 95%ANI using drep v.3.0.0[54], which selects the 'best' representative genome based on multiple quality metrics (completeness, contamination, strain heterogeneity, N50, centrality). De-replication resulted in 955 high-quality, species-level (95% ANI) metagenome-assembled genomes. These MAGs were taxonomically classified with GTDBtk v.1.5.1[55] and their species abundances across samples were calculated by mapping sequence reads to MAGs with KMA v.1.3.13[56]. The prevalence of MAGs across all samples was visualized along a tree built with GTDBtk[55] and visualized with iTOL[57].

Genome and metagenome-scale metabolic modelling

Genome-scale metabolic models (GEMs) were reconstructed for each species-level MAG with CarveMe v1.5[14]. GEMs were produced using domain-specific templates for archaea and bacteria, an average European diet[58] as medium for gap filling, and the IBM Cplex solver.

Metabolic exchanges between community members of a microbiome were calculated with MICOM v.0.26[18]. MICOM simulates growth and metabolic exchanges among members of the microbiome while accounting for their differential abundances, and it has been shown to estimate realistic growth rates. Furthermore, MICOM is computationally tractable when it comes to simulating diverse microbial communities (i.e., dozens-to-hundreds of species). Metabolic exchanges were estimated with MICOM's growth workflow, using a 0.5 trade-off parameter, an average European diet as medium, and parsimonious Flux Balance Analysis (pFBA) to identify optimal growth rates and metabolic fluxes. The underlying CarveMe models contain relatively few carbon sources, leading to low growth rates and consequent numerical instability. Therefore, the fluxes of medium items were multiplied by 600 to feasibly calculate metabolic exchanges, and then corrected in the final results. We verified the bacterial growth rates estimated with MICOM for all samples, which were within the expected range (Fig. S6), suggesting that this multiplication step did not induce unrealistic growth. An optimal solution was not found for 36 samples, which were removed from the analysis (identified in Supplementary Data 1), resulting in a final dataset of 1661 samples. A snakemake workflow is provided in the Zenodo repository for reproducibility[24].

Metabolite exchange scores

The underlying rationale to define the Metabolite Exchange Score (MES) is that an individual where metabolites are produced and consumed by multiple members of the microbiome will have a higher functional redundancy than an individual where these metabolites are produced and consumed by fewer species, which is a characteristic of most healthy ecosystems. For homogenized stool-derived metagenomes, which do not capture the patchiness in microbial aggregates found in the gut, high functional redundancy increases the likelihood that most micro-niches are populated by at least one species. The MES weighs the number of microbial species consuming and producing a given metabolite, in a given microbiome sample. MES was defined for each metabolite as the harmonic mean between potential consumers and producers (Eq. 1):

$$MES = 2 \times \frac{P \times C}{P + C} \quad (1)$$

Where P is the number of potential producers and C is the number of potential consumers of a given metabolite. Note that MES will be zero if a metabolite is only produced or only consumed but not exchanged among microorganisms.

The specific metabolites for which cross-feeding partners were significantly lost were identified with a Kruskal–Wallis test comparing diseased phenotypes against the healthy population. The Bonferroni method was used to account for multiple tests (0.05 as target alpha, divided by the number of tests), and only metabolites present in at least 50 individuals, including at least 15 diseased subjects, were included in the analyses. Water and oxygen were excluded from the analyses. For a simplified graphical representation (Fig. 2c), metabolites were selected for display if they showed a significant reduction in the number of cross-feeding partners, and if they were in the top 5 metabolites with the highest difference in MES in any disease. Barplots were generated and coloured according to the metabolite Sub Class defined in the Human Metabolome Database[59] using the *ggplot2* R package[60]. An additional word cloud including up to 100 metabolites with significant MES differences between healthy and diseased was generated with the *wordcloud* R package[61].

Species diversity effects

To estimate taxonomic diversity, the metagenome reads were mapped to the 955 species-level MAGs with KMA v.1.3.13[56]. Shannon index and species richness (total number of species in each sample, according to the reads mapping result) were used to quantify alpha-diversity, and compared between healthy and diseased microbiomes using the Wilcoxon test (holm method to account for multiple comparisons). Species richness were then used as a measure of species diversity for downstream analyses.

Differences in the slopes between species diversity and consumer or producer correlations were assessed on the entire dataset (including healthy and diseased microbiomes) by fitting a linear model (lm) in R, considering the interaction between number of producers and consumers with their category (producer or consumer). The statistical significance for the difference between slopes was corrected for multiple comparisons using the Bonferroni method.

Nutritional interactions in the microbiome associated with Crohn's disease

We selected a case-control study for an in-depth analysis that demonstrates how our framework can be applied to identify promising therapeutic targets. Given that the completeness of metagenome-assembled genomes is optimized by co-binning large datasets[22], we opted to select a case-control study from our quality-controlled dataset to take advantage of the large number of high-quality MAGs used to model community-wide metabolism. A total of 84 samples from the study of He and colleagues[33]—the largest CD study within our dataset—passed our quality control and were included in our analyses, including 46 patients with Crohn's disease and 38 healthy controls. The specific metabolites for which cross-feeding partners were lost were identified with a Kruskal–Wallis test, using only metabolites observed in over half of the samples and adjusting for multiple tests with a Bonferroni correction.

The flux of H_2S, estimated in millimoles per hour per gram of dry weight, was multiplied by species abundances to obtain the total H_2S production and consumption exchanged among microorganisms. Fluxes were $\log_2$-transformed for the statistical tests and graphical representation. Differences between the diversity of H_2S producers and consumers, ratios of producers to consumers, and their fluxes was evaluated with Kruskal–Wallis tests. The H_2S predicted to be exported to medium was used to estimate the excess H_2S production by the microbiome.

We used a nested linear model to account for the confounding effects of species diversity on the associations between number or flux of producers/consumers and disease status. Samples containing less than 99 species (the minimum number of species in the healthy cohort) were excluded from this analysis ($n = 58$ samples remaining), ensuring a linear relationship between species diversity and number of H_2S consumers or producers.

To better understand the genetic basis of H_2S production and consumption in MAGs observed within the CD case-control study, we performed a Hidden Markov Model (HMM) survey of 74 genes involved in H_2S cycling[36] with HMMer v.3.3.2[62], using trusted cutoff scores to ensure homology. We used a linear model to test if these genes were differentially distributed between healthy and CD individuals, using only samples with at least 100 species and genes observed in at least 10 samples. Analyses were performed considering both MAGs abundance (by multiplying gene counts by spp. abundance) and prevalence (using species presence/absence, which would be more informative when relatively rare taxa are responsible for a large proportion of the production and consumption of H_2S). Data was offset by 0.1 to avoid infinity upon log-transformation, species diversity was used as a confounding variable and the Bonferroni correction was used to account for multiple comparisons.

In order to identify species that may be promising targets of microbiome therapy in CD, we weighted in their flux of H_2S and relative abundances within CD and healthy cohorts. Specifically, weighted H_2S fluxes of each microbial species was estimated by multiplying their H_2S fluxes by their relative abundances. The weighted sum of H_2S fluxes was calculated as the sum of all weighted fluxes within healthy or diseased cohorts. Differences in the weighted sum of H_2S between healthy and CD cohorts pointed to the key H_2S producers and consumers associated with Crohn's disease. The Crohn's disease cohort contained more individuals than the healthy one, therefore, eight random samples were excluded to ensure the same number of individuals (38) in healthy and diseased categories. The metabolic model of *Roseburia intestinalis*, one key H_2S consumer, was visualized with Fluxer[63] using best *k*-shortest paths to visualize pathways between H_2S intake and cell growth.

To better understand how the modelling framework compare to more traditional composition-based analyses, we visualized the community beta diversity using a PCA plot of CLR-normalized species abundances with mixOmics[64], using the balanced dataset from He and colleagues[33] described above. We then performed a random-forest analysis[65] where 70% of the samples were randomly selected for training the model and the remaining 30% were used to test the classifier. Feature importance (mean decrease in Gini) was used to rank the species that most explained the variation between healthy and CD-associated microbiomes.

Statistics and reproducibility

The statistical tests applied here are described within their relevant section above using R. For reproducibility, we provide the R scripts in our Zenodo repository[24]. Data exclusion was performed based on quality/sequencing depth of metagenomes and completeness of the metadata (see 'Global survey of gut metagenomes and quality control section'). No statistical method was used to predetermine sample size.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The data used in this study is publicly available in the European Nucleotide Archive (ENA). All assemblies and MAGs reconstructed in this study have been deposited in ENA under project PRJEB63093. BioSample IDs for the raw sequence data and assembly IDs for the assemblies performed in this study are provided in Supplementary Data 1. ENA sample accessions for all metagenome bins reconstructed in this study are provided in Supplementary Data 12, and the ENA analysis ID for the 955 species-level MAGs are provided in Supplementary Data 2. All high-quality MAGs are also available in Zenodo[24] [https://zenodo.org/record/8223163]. Metabolite classes were inferred from the Human Metabolome Database HMDB 4.0 [https://hmdb.ca].

Code availability

The code developed to run the metabolic modelling analysis, perform statistical tests and to produce the graphs presented here, along with a step-by-step description of the analysis workflow, are available in Zenodo[24]: https://zenodo.org/record/8223163 (repository v.1.2.2), and in GitHub: https://github.com/vrmarcelino/MetaModels.

References

1. Wang, T., Goyal, A., Dubinkina, V. & Maslov, S. Evidence for a multilevel trophic organization of the human gut microbiome. *PLOS Comput. Biol.* **15**, e1007524 (2019).
2. Fischbach, M. A. & Sonnenburg, J. L. Eating for two: how metabolism establishes interspecies interactions in the gut. *Cell Host Microbe* **10**, 336–347 (2011).
3. Gralka, M., Szabo, R., Stocker, R. & Cordero, O. X. Trophic interactions and the drivers of microbial community assembly. *Curr. Biol.* **30**, R1176–R1188 (2020).
4. Goyal, A., Wang, T., Dubinkina, V. & Maslov, S. Ecology-guided prediction of cross-feeding interactions in the human gut microbiome. *Nat. Commun.* **12**, 1335 (2021).
5. Coyte, K. Z., Schluter, J. & Foster, K. R. The ecology of the microbiome: networks, competition, and stability. *Science* **350**, 663–666 (2015).
6. Moya, A. & Ferrer, M. Functional redundancy-induced stability of gut microbiota subjected to disturbance. *Trends Microbiol.* **24**, 402–413 (2016).
7. Tian, L. et al. Deciphering functional redundancy in the human microbiome. *Nat. Commun.* **11**, 6217 (2020).
8. Fassarella, M. et al. Gut microbiome stability and resilience: elucidating the response to perturbations in order to modulate gut health. *Gut* **70**, 595–605 (2021).
9. Sung, J. et al. Global metabolic interaction network of the human gut microbiota for context-specific community-scale analysis. *Nat. Commun.* **8**, 15393 (2017).
10. Fang, X., Lloyd, C. J. & Palsson, B. Ø. Reconstructing organisms in silico: genome-scale models and their emerging applications. *Nat. Rev. Microbiol.* **18**, 731–743 (2020).
11. Heinken, A., Basile, A., Hertel, J., Thinnes, C. & Thiele, I. Genome-scale metabolic modeling of the human microbiome in the era of personalized medicine. *Annu. Rev. Microbiol.* **75**, 199–222 (2021).
12. Freilich, S. et al. Competitive and cooperative metabolic interactions in bacterial communities. *Nat. Commun.* **2**, 589 (2011).
13. Levy, R. & Borenstein, E. Metabolic modeling of species interaction in the human microbiome elucidates community-level assembly rules. *Proc. Natl Acad. Sci.* **110**, 12804–12809 (2013).
14. Machado, D., Andrejev, S., Tramontano, M. & Patil, K. R. Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. *Nucleic Acids Res.* **46**, 7542–7553 (2018).
15. Magnúsdóttir, S. et al. Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. *Nat. Biotechnol.* **35**, 81–89 (2017).
16. Heinken, A. et al. Genome-scale metabolic reconstruction of 7302 human microorganisms for personalized medicine. *Nat. Biotechnol.* **41,** 1320–1331 (2023).
17. Zelezniak, A. et al. Metabolic dependencies drive species co-occurrence in diverse microbial communities. *Proc. Natl Acad. Sci.* **112**, 6449–6454 (2015).
18. Diener, C., Gibbons, S. M. & Resendis-Antonio, O. MICOM: metagenome-scale modeling to infer metabolic interactions in the gut microbiota. *mSystems* **5**, e00606–e00619 (2020).
19. Zorrilla, F., Buric, F., Patil, K. R. & Zelezniak, A. metaGEM: reconstruction of genome scale metabolic models directly from metagenomes. *Nucleic Acids Res.* **49**, e126–e126 (2021).
20. Heinken, A., Hertel, J. & Thiele, I. Metabolic modelling reveals broad changes in gut microbial metabolism in inflammatory bowel disease patients with dysbiosis. *Npj Syst. Biol. Appl.* **7**, 19 (2021).
21. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).
22. Nissen, J. N. et al. Improved metagenome binning and assembly using deep variational autoencoders. *Nat. Biotechnol.* **39**, 555–560 (2021).
23. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
24. Marcelino, V. R. et al. Code for community-wide metabolic modelling, calculation of metabolite exchange scores (MES) and statistical tests. version 1.2.2. https://doi.org/10.5281/zenodo.8223163 (2023).
25. Uebanso, T., Shimohata, T., Mawatari, K. & Takahashi, A. Functional roles of B-vitamins in the gut and gut microbiome. *Mol. Nutr. Food Res.* **64**, 2000426 (2020).
26. Mortensen, P. B., Holtug, K. & Rasmussen, H. S. Short-chain fatty acid production from mono- and disaccharides in a fecal incubation system: implications for colonic fermentation of dietary fiber in humans. *J. Nutr.* **118**, 321–325 (1988).
27. Baker, H. et al. Inability of chronic alcoholics with liver disease to use food as a source of folates, thiamin and vitamin B6. *Am. J. Clin. Nutr.* **28**, 1377–1380 (1975).
28. Tallaksen, C. M. E., Bell, H. & Bøhmer, T. The concentration of thiamin and thiamin phosphate esters in patients with alcoholic liver cirrhosis. *Alcohol. Alcohol.* **27**, 523–530 (1992).
29. Costantini, A. & Pala, M. I. Thiamine and fatigue in inflammatory bowel diseases: an open-label pilot study. *J. Altern. Complement. Med.* **19**, 704–708 (2013).
30. Tsuruya, A. et al. Ecophysiological consequences of alcoholism on human gut microbiota: implications for ethanol-related pathogenesis of colon cancer. *Sci. Rep.* **6**, 27923 (2016).

31. Mottawea, W. et al. Altered intestinal microbiota–host mitochondria crosstalk in new onset Crohn's disease. *Nat. Commun.* **7**, 13419 (2016).
32. Dordević, D., Jančíková, S., Vítězová, M. & Kushkevych, I. Hydrogen sulfide toxicity in the gut environment: meta-analysis of sulfate-reducing and lactic acid bacteria in inflammatory processes. *J. Adv. Res.* **27**, 55–69 (2021).
33. He, Q. et al. Two distinct metacommunities characterize the gut microbiota in Crohn's disease patients. *GigaScience* **6**, 1–11 (2017).
34. Roediger, E. W. & Millard, S. Reducing sulfur compounds of the colon impair coionocyte nutrition: implications for ulcerative colitis. *Gastroenterology* **104**, 802–809 (1993).
35. Braccia, D. J., Jiang, X., Pop, M. & Hall, A. B. The capacity to produce hydrogen sulfide (H_2S) via cysteine degradation is ubiquitous in the human gut microbiome. *Front. Microbiol.* **12**, 705583 (2021).
36. Wolf, P. G. et al. Diversity and distribution of sulfur metabolic genes in the human gut microbiome and their association with colorectal cancer. *Microbiome* **10**, 64 (2022).
37. Perez-Sanchez, C. et al. POS0394 NAD^+ boosters reestablish the altered NAD^+ metabolism of leukocytes from rheumatoid arthritis patients improving their oxidative, apoptotic and inflammatory status. *Ann. Rheum. Dis.* **80**, 426.2–426 (2021).
38. Mehmel, M., Jovanović, N. & Spitz, U. Nicotinamide riboside—the current state of research and therapeutic uses. *Nutrients* **12**, 1616 (2020).
39. LoConte, N. K., Brewster, A. M., Kaur, J. S., Merrill, J. K. & Alberg, A. J. Alcohol and cancer: a statement of the American Society of Clinical Oncology. *J. Clin. Oncol.* **36**, 83–93 (2018).
40. Louis, P., Hold, G. L. & Flint, H. J. The gut microbiota, bacterial metabolites and colorectal cancer. *Nat. Rev. Microbiol.* **12**, 661–672 (2014).
41. Watson, A. R. et al. Metabolic independence drives gut microbial colonization and resilience in health and disease. *Genome Biol.* **24**, 78 (2023).
42. Veseli, I. et al. Microbes with higher metabolic independence are enriched in human gut microbiomes under stress. *eLife.* **12**, RP89862 (2023).
43. Blachier, F. et al. Luminal sulfide and large intestine mucosa: friend or foe? *Amino Acids* **39**, 335–347 (2010).
44. Gemici, B. & Wallace, J. L. Anti-inflammatory and cytoprotective properties of hydrogen sulfide. in *Methods in Enzymology* Vol. 555, 169–193 (Elsevier, 2015).
45. Wallace, J. L., Motta, J.-P. & Buret, A. G. Hydrogen sulfide: an agent of stability at the microbiome-mucosa interface. *Am. J. Physiol. Gastrointest. Liver Physiol.* **314**, G143–G149 (2018).
46. Blachier, F., Beaumont, M. & Kim, E. Cysteine-derived hydrogen sulfide and gut health: a matter of endogenous or bacterial origin. *Curr. Opin. Clin. Nutr. Metab. Care* **22**, 68–75 (2019).
47. Zampieri, G., Campanaro, S., Angione, C. & Treu, L. Metatranscriptomics-guided genome-scale metabolic modeling of microbial communities. *Cell Rep. Methods* **3**, 100383 (2023).
48. Thiele, I. et al. Personalized whole-body models integrate metabolism, physiology, and the gut microbiome. *Mol. Syst. Biol.* **16**, e8982 (2020).
49. Pasolli, E. et al. Accessible, curated metagenomic data through ExperimentHub. *Nat. Methods* **14**, 1023–1024 (2017).
50. Gupta, V. K. et al. A predictive index for health status using species-level gut microbiome profiling. *Nat. Commun.* **11**, 4635 (2020).
51. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
52. Salazar, V. W. et al. Metaphor—a workflow for streamlined assembly and binning of metagenomes. *GigaScience* **12**, giad055 (2022).
53. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
54. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.* **11**, 2864–2868 (2017).
55. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925–1927 (2019).
56. Clausen, P. T. L. C., Aarestrup, F. M. & Lund, O. Rapid and precise alignment of raw reads against redundant databases with KMA. *BMC Bioinformatics* **19**, 307 (2018).
57. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, W293–W296 (2021).
58. Noronha, A. et al. The Virtual Metabolic Human database: integrating human and gut microbiome metabolism with nutrition and disease. *Nucleic Acids Res.* **47**, D614–D624 (2019).
59. Wishart, D. S. et al. HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res.* **46**, D608–D617 (2018).
60. Wickham, H. *ggplot2:* Elegant graphics for data analysis. *Springer-Verlag New York* (2016).
61. Fellows, I. wordcloud : Word Clouds. *R package version 2*, 331 (2018).
62. Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
63. Hari, A. & Lobo, D. Fluxer: a web application to compute, analyze and visualize genome-scale metabolic flux networks. *Nucleic Acids Res.* **48**, W427–W435 (2020).
64. Rohart, F., Gautier, B., Singh, A. & Lê Cao, K.-A. mixOmics: an R package for 'omics feature selection and multiple data integration. *PLOS Comput. Biol.* **13**, e1005752 (2017).
65. Liaw, A. & Wiener, M. Classification and regression by randomForest. *R News* **2**, 18–22 (2002).

Acknowledgements

This work was supported by the Australian Research Council (DP190101504) and the Australian National Health and Medical Research Council (APP1181105 and APP1186371). V.R.M. is supported by an Australian Research Council DECRA Fellowship (DE220100965), C.G. is supported by an National Health & Medical Research Council EL2 Fellowship (APP1178715), and S.C.F. is supported by a CSL Centenary Fellowship. S.M.G. and C.D. were supported by the National Institute of Diabetes and Digestive and Kidney Diseases of the National Institutes of Health (R01DK133468). The authors acknowledge the Monash eResearch Centre for access to computational resources and expertise and the support of the Victorian Government's Operational Infrastructure Support Program. We thank Dr Paul Harrison and Dr Jamie Gearing for statistical and bioinformatics advice, and Dr Lucas Schiffer for help with curatedMetagenomicData. We also thank the stool donors and researchers who made their metadata publicly available and the reviewers of this manuscript for their constructive feedback. Open access charges funded by the Hudson Institute of Medical Research.

Author contributions

V.R.M. and S.C.F. designed the study. V.R.M. and R.B.Y. identified samples and curated the metadata. V.R.M. conducted the metabolic modelling analyses. C.D. and S.M.G. assisted with data analysis and interpretation. C.W. and C.G. performed the survey of H_2S genes. E.L.G., E.L.R., and R.B.Y. contributed with bacterial microbiology expertise, and E.M.G contributed with clinical expertise in IBD. All authors contributed to the results interpretation and manuscript writing.

Competing interests

S.C.F. is an inventor on patents and has acted as an advisor to BiomeBank and Microbiotica. R.B.Y. has acted as an advisor to BiomeBank. All other authors have no competing interests to declare.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-023-42112-w.

Correspondence and requests for materials should be addressed to Vanessa R. Marcelino or Samuel C. Forster.

Peer review information *Nature Communications* thanks Francisco Zorrilla and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.